UNICA

UNIVERSITÀ
DEGLI STUDI
DI CAGLIARI

Ph.D. DEGREE IN

MATHEMATICS AND COMPUTER SCIENCE

Cycle XXXV

TITLE OF THE Ph.D. THESIS

Natural Language Processing for Motivational Interviewing Counselling: Addressing Challenges in Resources, Benchmarking and Evaluation

Scientific Disciplinary Sector

INF/01

Ph.D. Student:	Zixiu Wu
Supervisor	Prof. Diego Reforgiato Recupero (UniCa)
Co-Supervisor	Prof. Daniele Riboni (UniCa)

Final exam. Academic Year 2021/2022
Thesis defence: October 2023 Session

Abstract

Motivational interviewing (MI) is a counselling style often used in healthcare to improve patient health and quality of life by promoting positive behaviour changes. Natural language processing (NLP) has been explored for supporting MI use cases of insights/feedback generation and therapist training, such as automatically assigning behaviour labels to therapist/client utterances and generating possible therapist responses.

Despite the progress of NLP for MI applications, significant challenges remain. The most prominent one is the lack of publicly available and annotated MI dialogue corpora due to privacy constraints. Consequently, there is also a lack of common benchmarks and poor reproducibility across studies. Furthermore, human evaluation for therapist response generation is expensive and difficult to scale due to its dependence on MI experts as evaluators. In this thesis, we address these challenges in 4 directions: low-resource NLP modelling, MI dialogue dataset creation, benchmark development for real-world applicable tasks, and laypeople-experts human evaluation study.

First, we explore zero-shot binary empathy assessment at the utterance level. We experiment with a supervised approach that trains on heuristically constructed empathy vs. non-empathy contrast in non-therapy dialogues. While this approach has better performance than other models without empathy-aware training, it is still suboptimal and therefore highlights the need for a well-annotated MI dataset.

Next, we create `AnnoMI`, the first publicly available dataset of expert-annotated MI dialogues. It contains MI conversations that demonstrate both high- and low-quality counselling, with extensive annotations by domain experts covering key MI attributes. We also conduct comprehensive analyses of the dataset.

Then, we investigate two `AnnoMI`-based real-world applicable tasks: predicting current-turn therapist/client behaviour given the utterance, and forecasting next-turn therapist behaviour given the dialogue history. We find that language models (LMs) perform well on predicting therapist behaviours with good generalisability to new dialogue topics. However, LMs have suboptimal forecasting performance, which reflects therapists' flexibility where multiple optimal next-turn actions are possible.

Lastly, we ask both laypeople and experts to evaluate the generation of a crucial type of therapist responses — *reflection* — on a key quality aspect: coherence and context-consistency. We find that laypeople are a viable alternative to experts, as laypeople show good agreement with each other and correlation with experts. We also find that a large LM generates mostly coherent and consistent reflections.

Overall, the work of this thesis broadens access to NLP for MI significantly as well as presents a wide range of findings on related natural language understanding/generation tasks with a real-world focus. Thus, our contributions lay the groundwork for the broader NLP community to be more engaged in research for MI, which will ultimately improve the quality of life for recipients of MI counselling.

Acknowledgements

As I start writing this special section of my thesis, I cannot help but recall how this incredible journey began. It was the summer of 2019, when I met my future supervisor Rim at the ACL conference in Florence. I was immediately intrigued by her introduction of this Marie Curie PhD programme, and I would later move from the UK to the Netherlands for it. What makes my encounter with Rim even more remarkable is that it almost did not happen — I had nearly given up on going to Florence after a chaotic flight cancellation, but eventually I managed to go. Now, looking back, I am really happy with the way things turned out, because otherwise I would not have worked with so many talented people on all those exciting projects.

Many people have helped me along the way. First of all, I would like to express my deepest gratitude to my supervisors Dr. Rim Helaoui, Prof. Diego Reforgiato Recupero and Prof. Daniele Riboni for their guidance and support. As my industrial supervisor, Rim has inspired me deeply to focus on creating both scientific and real-world value in my research. As my academic supervisors, Diego and Daniele have always provided insightful feedback to help me make my work more rigorous and robust. The brainstorming sessions with Rim, Diego and Daniele were some of my favourite parts of this journey, as they greatly expanded my perspective and helped me develop creative ideas into solid research plans. In addition, I really appreciate the supervisors' constant encouragement through the ups and downs of these years, and I am also grateful for their help with various kinds of logistics that made my time at Philips and UniCa a smooth ride.

For my academic research, I am extremely fortunate to have collaborated with Simone Balloccu and Vivek Kumar, who are my fellow PhD students of the PhilHumans programme, and Simone's supervisor Prof. Ehud Reiter. The collaboration was some of the best teamwork I have ever had, and I thoroughly enjoyed exchanging with them a ton of great ideas and feedback. For my industrial projects at Philips, I had the privilege of working with Dr. Aki Härmä, Kateryna, Abhinay, Eva Z. and Eva M. among others, and some of them have indeed become my good friends. In particular, I would like to express my appreciation to Aki, my manager and co-coordinator of PhilHumans, who gave me significant academic freedom, valuable advice and considerable support with logistics.

I would like to also thank my other PhilHumans PhD colleagues: Allmin, Vadim, Ivan, Asfand and Marharyta. I really enjoyed the time we spent together, such as at the PhilHumans summer school in Vienna and workshop in Eindhoven. Also, many thanks to colleagues at UniCa's GARR lab and beyond: Danilo, Mirko, Giacomo B., Nino, Simone A., Andrea, Giacomo M., Ludovico, Valentino and Vittoria, without whom my secondment in Cagliari would not have been as rewarding and enjoyable. Furthermore, I am grateful to my other Philips colleagues, e.g., Heereen, Erno, Francesca, Arlette, Milan and Dr. Mark Aloia, for their support, especially Heereen who has given me so much advice and help.

Special thanks to those who helped me with administrative matters. At UniCa, they are: Prof. Michele Marchesi, Prof. Roberto Tonelli and Prof. Antonio Iannizzotto of the Mathematics and Computer Science Department, and Ms. Roberta Leu of the PhD and Master's Office. At Philips, they are: Mr. Ernst Hermens, Odette, Marita and Di.

I am also very grateful to my thesis reviewers — Prof. Ehud Reiter and Prof. Giovanni Farinella — for their insightful feedback that helped me improve the quality of this thesis.

Finally, of course, a huge dose of thanks to my dear friends Na, Nancy, Tom, Julia and Kaixin, as well as my parents, for always believing in me and having my back. It is reassuring knowing I can always count on you.

Thank you all. 谢谢. Bedankt. Grazie.

Funding

All of the original research presented in this thesis was funded by the European Commission in the H2020 Marie Skłodowska-Curie PhilHumans project, contract no. 812882. Zixiu Wu acknowledges the EC for this financial support.

PHILHUMANS

PHILIPS

UNIVERSITAS STVDIORVM CARALITANA

European Commission

Contents

Abstract **i**

Acknowledgements **iii**

Funding **v**

Contents **vii**

List of Abbreviations **xi**

List of Figures **xiii**

List of Tables **xv**

1 Introduction **1**

1.1 Motivational Interviewing (MI) 1
1.2 NLP for MI 2
1.3 Research Questions 3
1.4 Thesis Outline and Contributions 4
1.5 List of Publications 6

2 Background & Related Work **9**

2.1 General NLP Background 9
2.1.1 Transformer Basics 9
2.1.2 Transformer Pretraining 10
2.1.3 Transformer Fine-Tuning & NLP Tasks 10
2.1.4 Large Language Models (LLMs) 11
2.1.5 NLU & NLG Performance Evaluation 11
2.2 NLP & Speech Processing for MI 12
2.2.1 MI Resources 12
2.2.2 Automated Therapist Empathy Assessment for MI 14
2.2.3 Automated Utterance Coding for MI 15
2.2.4 Reflection Generation for MI 17
2.3 NLP for Empathetic Non-Counselling Dialogue 20

2.3.1 Overview . . . 20
2.3.2 Dialogue Analysis . . . 21
2.3.3 Dialogue Generation . . . 22

3 Low-Resource Real-Time Therapist Empathy Assessment 25
3.1 Introduction . . . 26
3.2 Data . . . 28
3.2.1 Non-Counselling Conversations . . . 29
3.2.2 MI Dialogues . . . 31
3.3 Zero-Shot Empathy Prediction . . . 32
3.3.1 Task Definition . . . 32
3.3.2 Supervised: OOD Training . . . 34
3.3.3 Unsupervised: NLI-Based . . . 35
3.4 Experiments . . . 36
3.4.1 Implementation . . . 36
3.4.2 Results . . . 38
3.5 Clinical Application & Impact . . . 40
3.6 Summary . . . 41

4 Creation of Counselling Dialogue Dataset 43
4.1 Introduction . . . 44
4.2 Creating `AnnoMI` . . . 44
4.2.1 MI Demonstration Videos . . . 45
4.2.2 Transcription . . . 47
4.2.3 Expert Annotators & Workload Assignment . . . 47
4.2.4 `AnnoMI` & “Real-World” MI . . . 49
4.2.5 Annotation Scheme . . . 50
4.3 IAA Results . . . 54
4.3.1 Default Measure: Fleiss’ Kappa at Utterance-Level . . . 54
4.3.2 Results of Default IAA Measure . . . 55
4.3.3 Supplementary IAA Measure: Intraclass Correlation . . . 56
4.3.4 Dataset Release . . . 56
4.4 Dataset Analysis . . . 57
4.4.1 Overall **(Main) Behaviour** & **Talk Type** Distributions . . . 57
4.4.2 Posterior **(Main) Behaviour** & **Talk Type** Distributions . . . 58
4.4.3 **(Main) Behaviour** & **Talk Type** as Conversation Proceeds 60
4.4.4 Utterance Length Distributions . . . 61
4.4.5 Frequent 3-Grams . . . 64
4.4.6 Utterance Embedding Distribution . . . 65
4.5 Discussion . . . 66
4.6 Summary . . . 67

5 Utterance-Level Behaviour Prediction & Forecasting 69
5.1 Introduction . . . 70
5.2 Current-Turn Therapist/Client Behaviour Prediction . . . 71
5.2.1 Tasks and Setup . . . 71
5.2.2 Overall Performance . . . 73
5.2.3 Topic-Specific Performance and Generalisability to New Topics 77
5.2.4 Discussion & Summary . . . 79
5.3 Next-Turn Therapist Behaviour Forecasting . . . 79
5.3.1 Task Setup . . . 79
5.3.2 Modelling Choices . . . 80
5.3.3 Results & Analysis . . . 82
5.3.4 Discussion & Summary . . . 86

6 Human Evaluation for Therapist Response Generation 87
6.1 Introduction . . . 88
6.2 Developing Coherence Evaluation Scheme . . . 91
6.2.1 Generating Synthetic Reflections . . . 92
6.2.2 Collecting Incoherence Error Descriptions . . . 94
6.2.3 Establishing Error Categories . . . 97
6.3 Human Evaluation: Laypeople & Experts . . . 99
6.3.1 Generating Synthetic Reflections . . . 99
6.3.2 Evaluation Procedure . . . 101
6.3.3 Agreement Among Annotators . . . 103
6.3.4 Cross-Stage Annotation Differences . . . 105
6.3.5 Case Study . . . 110
6.4 Summary . . . 110

7 Conclusion 113
7.1 Contributions and Limitations . . . 113
7.2 Future Work . . . 115

Bibliography 119

A Human Evaluation of Reflections - *Supplement* 145
A.1 Reflection Sampling for Annotation & Inadequacy of BART . . . 145
A.1.1 Reflection Sampling Procedure . . . 146
A.1.2 Lack of Diversity Among BART Reflections . . . 146
A.1.3 Label Distribution for BART Reflections . . . 147
A.2 Waiting Period Between Stages . . . 148
A.3 Label Distribution for Differently Generated Reflections . . . 149
A.4 Full Annotation Flow . . . 151
A.5 Annotation Interface . . . 152
A.6 Limitations . . . 152

A.7 Data Use & Creation . . . 154
A.8 Ethics Statement . . . 154
A.8.1 Ethical Review . . . 154
A.8.2 Risks . . . 154
A.8.3 Information and Consent . . . 155
A.8.4 Demographic Information of Annotators . . . 155
A.8.5 Remuneration . . . 155
A.8.6 Data Anonymisation . . . 156

List of Abbreviations

AI	Artificial Intelligence
BERT	Bidirectional Encoder Representations from Transformers
BLEU	Bilingual Evaluation Understudy
CNN	Convolutional Neural Network
CV	Cross Validation
DL	Deep Learning
DNN	Deep Neural Network
DUI	Driving Under the Influence
GPT	Generative Pretrained Transformer
GPU	Graphics Processing Unit
GRU	Gated Recurrent Units
IAA	Inter-Annotator Agreement
ICC	Intraclass Correlation
LLM	Large Language Model
LM	Language Model
LSTM	Long Short-Term Memory
MCC	Matthews Correlation Coefficient
MI	Motivational Interviewing
MISC	Motivational Interviewing Skill Code
MITI	Motivational Interviewing Treatment Integrity
ML	Machine Learning
NLG	Natural Language Generation
NLI	Natural Language Inference
NLP	Natural Language Processing
NLU	Natural Language Understanding
OOD	Out of Domain
RL	Reinforcement Learning
RLHF	Reinforcement Learning from Human Feedback
RNN	Recurrent Neural Network
RoBERTa	Robustly Optimized BERT Pretraining Approach
ROUGE	Recall-Oriented Understudy for Gisting Evaluation
SD	Standard Deviation
Seq2Seq	Sequence-to-Sequence

T-SNE	t-Distributed Stochastic Neighbor Embedding

List of Figures

3.1 Overview of supervised approach: OOD empathy contrast training on non-counselling conversations and in-domain testing on MI dialogues. **Listener/Therapist** utterance is required; preceding Speaker/Client utterance as context is optional. 27

3.2 Overview of unsupervised approach based on NLI. 27

4.1 Distribution of dialogue lengths (number of utterances per dialogue). 46

4.2 Annotator survey on whether `AnnoMI` reflects real-world counselling. . 49

4.3 (Main) Behaviour distributions in high- & low-quality MI (©2022 IEEE) 57

4.4 Talk Type distributions in high- & low-quality MI (©2022 IEEE) . . 58

4.5 Distribution of next-turn client talk types (Y-axis) given different therapist behaviours in the current turn (X-axis). 59

4.6 Distribution of next-turn therapist behaviours (Y-axis) given different client talk types in the current turn (X-axis). 59

4.7 Proportions of therapist behaviours in different conversation stages in high- and low-quality MI. 60

4.8 Proportions of client talk types in different conversation stages in high- and low-quality MI. 61

4.9 Lengths (number of words) of therapist & client utterances in high- & low-quality MI. 62

4.10 Utterance lengths (number of words) of different therapist behaviours in high- & low-quality MI. 62

4.11 Ratios between the length (number of words) of the next-turn therapist response (broken down into 4 types of therapist behaviours) and that of the current-turn client utterance. 63

4.12 T-SNE of therapist utterance embeddings. 66

4.13 T-SNE of client utterance embeddings. 66

5.1 Confusion matrix of BERT w/ Adapters for (main) therapist behaviour prediction in the Original Unbalanced setting. Normalised by row. 75

5.2 Confusion matrix of BERT w/ Adapters for client talk type prediction in the Original Unbalanced setting. Normalised by row. 76
5.3 Impact of dialogue history length and data augmentation 83
5.4 Confusion matrices of 1- and *max*-utterance baselines 84
5.5 Impact of inserting therapist/client utterance labels in dialogue history on model performance 85
5.6 Impact of contrasting high- & low-quality MI 85

6.1 Human evaluation overview. The same human reflections are included in both evaluation stages, mixed with GPT-2 reflections in the GPT-2 stage and with GPT-3 reflections in the GPT-3 stage. . . 90
6.2 Annotation flow for candidate reflections of one context, during collection of incoherence error descriptions. 95
6.3 Annotation interface for one reflection candidate, during collection of incoherence error descriptions. 96
6.4 Prompting formats. 100
6.5 Labels distribution on human and synthetic reflections in the GPT-2 stage and GPT-3 stage. *Incoherent* labels are broken down into fine-grained error categories. 106

A.1 Full annotation flow for one batch. In this work, we do not investigate annotations w.r.t. empathy assessment or the most evident error category. 151
A.2 Annotation interface when the annotator annotates a reflection as coherent & consistent. 152
A.3 Annotation interface when the annotator annotates a reflection as incoherent/inconsistent and chooses one error category. 153
A.4 Annotation interface when the annotator annotates a reflection as incoherent/inconsistent and chooses multiple error categories. 153

List of Tables

3.1 Overview of statistics of filtered *ReddItConvs* and *EmpDial*. #(Dialogues): number of dialogues. Avg#(Speaker/Listener Utts): average number of speaker/listener utterances per dialogue. Avg(Speaker/Listener Utt Len): average speaker/listener utterance length, namely number of tokens. 29

3.2 Example dialogues from *ReddItConvs* and *EmpDial*. The dialogue lengths reflect the Avg#(Speaker/Listener Utts) shown in Table 3.1. . 30

3.3 Overview of `DemoMI` and `AnnoDemoMI`. #(Dialogues): number of dialogues. Total#(Therapist/Client Utts): total number of therapist/client utterances. Avg#(Therapist/Client Utts): average number of therapist/client utterances per dialogue. Avg#(Therapist/Client Utt Len): average therapist/client utterance length, namely number of tokens. %(Empathetic Therapist Utts): percentage of empathetic therapist utterances. 31

3.4 Example dialogue snippets from `AnnoDemoMI`. Underlined: empathetic. The snippets also show minimal transcription error ("weight"→"way") and a de-identification placeholder (⟨date/time⟩). . 33

3.5 Illustration of NLI. Example from `https://paperswithcode.com/task/natural-language-inference`. 35

3.6 Overview of NLI-based models as zero-shot empathy classifiers. . . . 35

3.7 Overview of OOD models' performance on their OOD test sets (not on `AnnoDemoMI`). The metric is MCC, which ranges from -1 to 1. . . . 38

3.8 Overview of OOD models' performance on `AnnoDemoMI`. The metric is MCC, which ranges from -1 to 1. **Bold**: best performance. 39

3.9 NLI performance overview. The metric is MCC, which ranges from -1 to 1. **Bold**: best performance. 40

4.1 `AnnoMI` overview. (©2022 IEEE) 45

4.2 High- and low-quality MI conversation snippets, where the goal is smoking cessation/reduction. ***T***: therapist; ***C***: client. 46

4.3 Transcription quality comparison between AnnoMI and [1]. Red: incorrectly transcribed word; Blue: omitted words/phrases; Orange: words from the other interlocutor that should have started a new utterance; ~~Strikethrough~~: incorrectly transcribed word within such a misplaced utterance; {C}/{T}: missing client/therapist utterance. . . 48
4.4 Top 10-topics in AnnoMI in terms of (1) number/percentage of dialogues and (2) total number/percentage of utterances. 50
4.5 Utterance-level multi-choice annotation scheme. [+] implies presence of utterance attribute (e.g., "Simple reflection" label entails that **Reflection** exists in utterance), while [−] indicates absence thereof (e.g., "No reflection" label implies **Reflection** is not present in utterance). . 51
4.6 Example labelling for therapist **Question** from the dataset. Detailed Open/Closed question types (e.g., "Number") are illustrative only and not annotated. 52
4.7 Example Labelling for therapist **Input** from the dataset. 53
4.8 Example labelling for therapist **Reflection** from [2]. 53
4.9 Example labelling for client **Talk Type** from the dataset. 54
4.10 IAA on utterance-level annotations, in Fleiss' kappa. Orange, blue, cyan and green indicate fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80) and almost perfect (0.80-1.00) agreement, respectively. 55
4.11 ICC as IAA. 56
4.12 Most frequent 3-grams of therapist utterances of different **(Main) Behaviours** in high- and low-quality MI. Numbers of occurrences are shown in parentheses. 64
4.13 Most frequent 3-grams of client utterances of different **Talk Types** in high- and low-quality MI. Numbers of occurrences are shown in parentheses. 65

5.1 Distribution of (Main) Behaviour and Talk Type after de-duplication. Overall, (Main) Behaviour has 3796 unique examples and Talk Type has 3685. 72
5.2 Macro F1 and per-class F1's of (main) therapist behaviour prediction. All results averaged from 5-fold CV. ↑/↓: performance increase/decrease by using Augmented Balanced compared to using Original Unbalanced. 74
5.3 Macro F1 and per-class F1's of client talk type prediction. All results averaged from 5-fold CV. ↑/↓: performance increase/decrease by using Augmented Balanced compared to using Original Unbalanced. 76
5.4 Topic-specific and cross-topic performance (macro F1) in Original Unbalanced for 1) Therapist Behaviour Prediction and 2) Client Talk Type Prediction. ↑/↓: cross-topic performance is higher/lower than topic-specific performance. 77

5.5 A 3-turn dialogue history from a high-quality MI conversation 80

6.1 A dialogue context from `AnnoMI` ([3], Chapter 4) about reducing substance use, together with its **human reflection** and two examples of **synthetic reflections**. Self-repetitions and mid-sentence changes (e.g., "it-it-whatever") are characteristics of spoken dialogue that are faithfully transcribed. 89
6.2 A 3-turn context and its gold-standard human reflection from an MI dialogue in `AnnoMI`. 92
6.3 Perplexity of each response generator on human reflections. 94
6.4 Illustrative examples for error categories. 98
6.5 Overview of Annotation Workload. 101
6.6 Global agreement on *Coherent*/*Incoherent* binary choice (intra-group agreement). 103
6.7 Per-label majority agreement ratios on *Coherent* and *Incoherent* separately (intra-group agreement). 104
6.8 Per-label majority agreement ratios for error categories (intra-group agreement). *Italic*: less than 10 reflections are given this error category by any annotator. 105
6.9 Inter-group correlations between laypeople- and experts-based coherence scores. $p <$ 1e-7 for all 4 values. 105
6.10 *Coherent*/*Incoherent* label distributions for human reflections. We report how often (%) the annotators annotate a human reflection as *Coherent*. **Bold**: significant (chi-squared test, $p < 0.05$) cross-stage shift. 108
6.11 Overview of how often each layperson (L1∼L9) and each expert (E1∼E9) consistently assigns the same coherence label (Yes/No) to a **recurring** human reflection in both stages. **N/A**: annotator has no recurring human reflections in workload. †: annotator is not always consistent. 109
6.12 Overview of how often each layperson (L1 - L9) and each expert (E1 - E9) annotates a **non-recurring** human reflection as *Coherent* in each stage. ↑/↓: increase/decrease in GPT-3 stage compared to GPT-2 stage. 109
6.13 Complete dialogue context of Table 6.1 and annotations on reflection examples. **L1/L2/.../L9**: 9 laypeople. **E1/E2/.../E9**: 9 experts. Red: incoherence error category. ‡: Annotator annotated the human reflection in both stages. 111

A.1 Overview of lexical (Self-BLEU) & semantic (average pairwise embedding similarity) diversity among reflections generated by different models. Lower values indicate more diversity. 147
A.2 Label distribution for BART-generated reflections. 147

A.3 BART-generated reflection for the case study dialogue (Table 6.13) and its annotations. **L1/L2/.../L9**: 9 laypeople. **E1/E2/.../E9**: 9 experts. Red: incoherence error category. 148
A.4 Waiting period lengths (number of days) between the two stages. . . 148
A.5 Answers given to the post-annotation question "While you were annotating in Phase 2 (i.e., GPT-3 stage), did you remember seeing any response candidate that you had seen in Phase 1 (i.e., GPT-2 stage)?". 149
A.6 Label distribution on synthetic reflections from GPT-2 and GPT-3 under different generation settings. **L**: laypeople. **E**: experts. All GPT-3 reflections are generated with nucleus decoding. 150

Chapter 1

Introduction

Motivational interviewing (MI) is a highly effective counselling practice in healthcare. Recent years have seen progress in natural language processing (NLP) for supporting/automating MI use cases, but significant challenges remain, which we address in this thesis. In this introduction, we first introduce MI (§1.1) and provide an overview of NLP for MI and its challenges (§1.2). Then, we detail our related research questions (§1.3) and outline how they are addressed by the chapters of this thesis (§1.4). We present the list of publications in §1.5.

1.1 Motivational Interviewing (MI)

Recent studies have shown that populations in many parts of the world are becoming less healthy and suffering more from chronic health conditions than previous generations (e.g., [4, 5]), to which unhealthy lifestyles such as inactivity can be a major contributor [6]. Accordingly, behaviour changes such as smoking cessation and alcohol use reduction can considerably improve health outcomes [7, 8]. However, people often struggle with adherence to recommended behaviour changes, especially over a long period of time [9].

Thus, to promote adoption of and adherence to positive behaviour changes, MI [10, 11] was developed as a client-centred counselling style that evokes motivation for change from the client themselves. MI has proved effective for various behaviour change treatments, such as smoking cessation and physical activity [12].

To generate insights and evaluate therapists, MI sessions — each session being a therapist-client conversation — can be annotated, which is usually done through post-session behaviour observation and coding [13, 14, 15]: upon reviewing the complete session, trained annotators 1) give ratings w.r.t. session-level qualities such as therapist empathy; and 2) assign a label (often referred to as "code") to each utterance (turn of conversation[1]). The label of a therapist utterance indicates a

[1] We discuss utterance definition in more detail in §3.2.

therapist skill/behaviour such as question, and similarly the label of a client utterance indicates a client behaviour such as change talk (i.e., a statement that favours behaviour change). The coding process is typically based on an established protocol, such as the Motivational Interviewing Skill Code (MISC) [16] and the Motivational Interviewing Treatment Integrity Code (MITI) [2].

1.2 NLP for MI

NLP is a branch of Artificial Intelligence (AI) that enables computers to understand and generate natural language, focusing mostly on texts. In recent years, NLP as a research field has grown exponentially, thanks to increasing computational power (e.g., [17, 18]) and various model architectures (e.g., [19, 20, 21, 22]) based on deep neural networks (DNNs) [23]. In particular, state-of-the-art NLP systems have outperformed human baselines on various natural language understanding (NLU) benchmarks (e.g., [24, 25]) and shown impressive human-like natural language generation (NLG) capabilities (e.g., [26, 27, 28]).

Accordingly, in recent years, researchers have explored NLP-based use cases for MI, mostly for insights/feedback generation and therapist training. One major direction is NLP-powered transcript analysis for complete MI sessions, including automatic empathy assessment and utterance coding ([29, 30, 31, 32, 33, 34], *inter alia*), motivated by the fact that manual behaviour coding is costly and time-consuming. Beyond post-session analysis, progress has also been made in providing automated guidance to the therapist in real-time, such as 1) what action to take in the next dialogue turn [13] with an NLU model, and 2) what to say next to the client [35, 36, 37] with an NLG model (i.e., therapist response generation).

Despite the growth of NLP for MI, significant challenges remain in this domain. We list below 3 major challenges that are the most relevant to this thesis.

Challenge 1 The most prominent challenge is the lack of publicly available MI dialogue data, which is a problem in counselling data sharing in general due to concerns for client privacy and related regulatory barriers [38, 39]. This challenge alone has considerably limited access to this research area, in particular for researchers who are not affiliated or in collaboration with institutions that hold such data.

Challenge 2 The lack of data sharing has also led to the common practice in NLP for MI where each research group collects its own MI dialogue data, conducts internal data annotation by MI experts, and runs NLP experiments on the annotated data. This in turn means 1) published NLP-for-MI results lack reproducibility, and 2) results from different research groups are often not comparable.

Challenge 3 In therapist response generation for MI, human evaluation is challenging due to its dependence on MI experts such as professional therapists to read and assess model-generated texts. In particular, the high costs of hiring experts and the lack of expert availability make such evaluations difficult to scale.

1.3 Research Questions

In this thesis, we focus on addressing the 3 challenges of NLP for MI laid out in §1.2. Specifically, we raise 4 research questions and investigate them in our work:

Research Question 1 (RQ1): How to approach real-time empathy assessment for MI in a low-resource setting, where there is little to no MI dialogue data with empathy-related annotations?

This research question is related to **Challenge 1**.

Empathy assessment is important for understanding counselling quality, but it has so far been almost exclusively modelled as rating therapist empathy for a complete session (e.g., [29, 30, 31]), rather than as delivering real-time feedback at the utterance level in an ongoing MI dialogue.

In particular, given the lack of empathy-annotated MI dialogue data in the context of Challenge 1, it is crucial to investigate if real-time empathy assessment is possible without needing (considerable) training data. A positive outcome would ease the low-resource constraint on this and similar tasks. A negative outcome would emphasise the need for a well-annotated publicly available MI dialogue dataset.

Research Question 2 (RQ2): How to create a publicly available and expert-annotated dataset of MI dialogues to benefit the research community?

This research question is related to **Challenge 1**.

A publicly available and well-annotated MI dialogue dataset would alleviate Challenge 1 considerably and broaden access to the research area of NLP for MI. To ensure maximum usefulness, such a dataset should be

- reflective of real-world MI, so that models trained on the dataset do not have a domain gap when applied to real-world use cases;
- annotated by MI experts such as professional therapists, so that models trained on the dataset can learn from insights given by experts;
- explicitly compliant with regulations related to privacy and consent, so that the dataset can be used by researchers without causing legal implications.

Research Question 3 (RQ3): How to leverage the dataset of RQ2 to create benchmark tasks and models with potential for real-world application?

This research question is related to **Challenge 2**.

Once the dataset of RQ2 is created, it is essential to use it as a benchmark for real-world applicable tasks. In particular, tasks explored in prior work (e.g., [34, 13]) with real-world relevance should have priority. Experiments with those tasks on the new dataset would provide the first reproducible results that facilitate comparison with future studies, thus addressing Challenge 2.

Considering that prior studies mainly used older models like long short-term memory networks (LSTMs) [19], the new experiments should use more advanced (transformer-based) language models[2] (LMs) and investigate 1) the impact of different modelling choices on performance; 2) practical concerns such as model generalisability to new behaviour change topics.

Research Question 4 (RQ4): In therapist response generation for MI, what criteria should human evaluators meet to ensure effective evaluation?

This research question is related to **Challenge 3**.

In therapist response generation, expert human evaluators are critical because of their understanding of the complex and sensitive domain of counselling. Nevertheless, a generated response has to first "make sense" in the dialogue context before it can be evaluated against MI principles. Thus, human evaluation can be split into two steps: 1) checking if the response makes sense in context, and 2) assessing if it meets psychotherapy standards. We argue that laypeople (non-experts) can perform the first step, saving time and resources and thus addressing Challenge 3.

We note that assessing whether generated text makes sense in context is crucial in light of the recent development of large language models (LLMs) [17, 28], as these powerful models can still struggle with basic errors like hallucination [40, 41, 42], a phenomenon where the output is unfaithful/ungrounded w.r.t. the input.

1.4 Thesis Outline and Contributions

The rest of the thesis is structured as follows:

Chapter 2: We introduce the background and related work of the research presented in this thesis.

[2] In this thesis, we use "language models" to exclusively refer to pretrained transformers, unless otherwise specified.

Chapter 3: Focusing on **RQ1**, our objective is zero-shot binary empathy assessment at the utterance level. We propose two methods: 1) a supervised approach that utilises heuristically constructed empathy vs. non-empathy contrast in non-counselling (thus out-of-domain) conversations, and 2) an unsupervised method that uses natural language inference as a proxy task for empathy prediction. Our findings indicate that empathy vs. non-empathy contrast leads to the best performance, although it is not sufficiently high. The benefit of the contrast becomes clear when it is compared to the unsupervised method and control-group supervised models trained without empathy contrast. This chapter is based on [43].

Chapter 4: To address **RQ2**, we present `AnnoMI`, the first publicly available dataset of professionally transcribed and expert-annotated counselling dialogues. It comprises 133 dialogues demonstrating high- and low-quality MI, with annotations by domain experts covering essential MI attributes. We detail the data collection process, including dialogue selection, transcription, and annotation. Based on the expert annotations, we conduct comprehensive analyses of `AnnoMI` at the utterance, dialogue, and corpus levels. We also discuss possible applications of the dataset. This chapter is based on [3, 44].

Chapter 5: To approach **RQ3**, we investigate two `AnnoMI`-based tasks with potential real-world applications: current-turn therapist/client behaviour prediction and next-turn therapist behaviour forecasting. Prediction identifies the behaviour label of the current turn given the turn's utterance, while forecasting anticipates next-turn behaviour using the dialogue history up to that point. For prediction, we find that LMs have better results on therapist behaviours than on client behaviours, both in overall performance and generalisability to new topics. For forecasting, we observe suboptimal performance despite various NLP modelling choices, which reflects therapists' flexibility where often multiple optimal next-turn actions are possible. This chapter is based on [44, 45].

Chapter 6: To explore **RQ4**, we focus on one type of therapist response — *reflection* — and human evaluation of generated reflections. Reflection is a critical skill in MI where the therapist conveys their interpretation of the client's words. Our investigation centres on whether laypeople can be an alternative to experts in evaluating the fundamental quality aspect of coherence and context-consistency, i.e., "Does this reflection make sense (in context)?". We first develop an evaluation scheme based on laypeople's descriptions of incoherence/inconsistency errors in generated reflections. Then, using this scheme, we ask both laypeople and experts to annotate both generated and human reflections. Our results show that laypeople are capable of this evaluation, based on their agreement with each other and correlation with experts' results. Furthermore, we find that an LLM (GPT-3) generates predominantly coherent and consistent reflections, and we also examine how evaluation results change when the source of generated reflections changes from a less powerful LM (GPT-2) to a powerful LLM (GPT-3). This chapter is based on [46, 47].

Chapter 7: We conclude this thesis and discuss directions for future work.

Overall, our work substantially improves access (Chapters 4, 6) to the NLP-for-MI research field as well as presents extensive findings on related NLU (Chapters 3, 5) and NLG (Chapter 6) tasks with real-world applicability. Thus, this thesis establishes a foundation for the wider NLP community to be involved in research for MI, which will ultimately benefit MI clients.

1.5 List of Publications

This thesis incorporates the following publications:

- [43] Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Towards low-resource real-time assessment of empathy in counselling. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 204–216, Online, June 2021. Association for Computational Linguistics

- [3] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-mi: A dataset of expert-annotated counselling dialogues. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6177–6181. IEEE, 2022. (©2022 IEEE. Reprinted, with permission.)

- [44] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3), 2023

- [45] Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Towards automated counselling decision-making: Remarks on therapist action forecasting on the annomi dataset. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 1906–1910. ISCA, 2022. DOI: 10.21437/Interspeech.2022-506

- [46] Zixiu Wu, Simone Balloccu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Towards in-context non-expert evaluation of reflection generation for counselling conversations. In *Proceedings of the 2nd Workshop* *on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 116–124, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics

- [47] Zixiu Wu, Simone Balloccu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Are experts needed? on human evaluation of counselling reflection generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6906–6930, Toronto, Canada, July 2023. Association for Computational Linguistics

The publications listed below are related to the research presented in Chapters 3, 4, 5, but they are not the focus of this thesis:

- [48] Zixiu Wu, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. Towards detecting need for empathetic response in motivational interviewing. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 497–502, 2020

- [49] Vivek Kumar, Simone Balloccu, Zixiu Wu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Data augmentation for reliability and fairness in counselling quality classification. In *Proceedings of the 1st Workshop on Scarce Data in Artificial Intelligence for Healthcare - SDAIH*, pages 23–28. INSTICC, SciTePress, 2023

Chapter 2

Background & Related Work

In this chapter, we introduce the background and related work of our research. First, we provide relevant NLP background knowledge in §2.1. Then, we present a literature review of related work in NLP and speech processing for MI in §2.2. Finally, we give an overview of recent research in NLP for empathetic non-counselling dialogues in §2.3, since it is relevant to our work in Chapter 3 and also to NLP for MI as a research field.

2.1 General NLP Background

In this section, we provide a high-level overview of general NLP concepts that are most relevant to our research. For a more comprehensive and detailed introduction to NLP, we refer the reader to resources such as [50]. We note that the structures of §2.1.1, §2.1.2, §2.1.3 and §2.1.4 are inspired by [50].

2.1.1 Transformer Basics

The transformer architecture [20] is the mainstream for a wide variety of NLP tasks. A transformer consists of an encoder and a decoder: given an ⟨input, output⟩ text pair, the encoder produces contextualised token representations of the input text, which provide grounding for the decoder to generate the output text token by token. Notably, representation learning in a transformer is entirely based on the attention mechanism [51].

A transformer encoder/decoder consists of stacked encoder/decoder blocks, where the input to one block is the output from the block below. In an encoder block, the self-attention mechanism builds the representation of each token by attending to all the others tokens in the sequence, so that the token representations are grounded in both left- and right-contexts. In a decoder block, there are two attention steps: masked self-attention and cross attention. Masked self-attention is self-attention without attending to future (i.e., yet to be generated) tokens, thus allowing for unidirectional text generation. Then, cross attention enriches decoder-side output token representations by attending to encoder-side input token representations.

In our work, we use transformer-based models in Chapters 3, 5 and 6.

2.1.2 Transformer Pretraining

Transformer pretraining is a self-supervised training technique that has been a main reason for the success of transformers at various NLP tasks. In pretraining, a transformer learns to reconstruct deliberately masked parts of the input, and does so on a large corpus of raw texts. As the reconstruction capability of a transformer grows, it gradually learns structural and semantic aspects of the language. Once pretraining is complete, the model can be directly used (e.g., [17]) or fine-tuned (e.g., [21]) for a specific downstream task.

Pre-trained transformers can be divided into encoder-only, decoder-only and encoder-decoder models. An encoder-only model is a transformer encoder, often pretrained to predict randomly masked tokens in the input (e.g., [21]). A decoder-only model (often referred to as a language model, or LM) is a transformer decoder, whose prevailing pretraining objective is language modelling, i.e., generating the entire input token-by-token (e.g., [22]). An encoder-decoder model is a full transformer, which can be pretrained by adding noise to the encoder input and training the decoder to reconstruct the original input (e.g., [52]).

In this thesis, we use pretrained encoder-only models in Chapters 3 (BERT [21]) and 5 (BERT, RoBERTa [53]), decoder-only models (GPT-2 [54], GPT-3 [17]) in Chapter 6, and an encoder-decoder model (BART [52]) in Chapters 3 and 6.

2.1.3 Transformer Fine-Tuning & NLP Tasks

Given a downstream NLP task, a pretrained transformer is often fine-tuned into a specialised model on the task-specific data. NLP tasks have 2 general categories: natural language understanding (NLU) and natural language generation (NLG). NLU tasks focus on classification at the token (e.g., part-of-speech tagging) or sequence level (e.g., sentiment classification), while NLG tasks are often in the form of generating a target text given a source text (e.g., summarisation, dialogue generation).

An encoder-only model is mostly fine-tuned for NLU by adding a fully-connected layer over top-level token representations [21, 53, 55]. Notably, this method can also be used for decoder-only and encoder-decoder models on NLU tasks [22, 52]. On the other hand, for NLG, fine-tuning decoder-only and encoder-decoder models is more common. Usually, a decoder-only model is fed the input text and generates the target text to the right [22, 54], while an encoder-decoder model receives the input at the encoder and generates the target text at the decoder [52, 56].

Notably, some recent studies have proposed lightweight fine-tuning methods that do not require updating the entire pretrained model. One such example is the adapter design [57], which inserts trainable lightweight adapter modules into a pretrained model. During fine-tuning, only the adapters need to be updated, which proved to be less resource-intensive and cause minimal performance loss [57].

In our research, we fine-tune pretrained transformers for NLU in Chapters 3, 5 and for NLG in Chapter 6. In particular, we explore adapters in Chapter 5.

2.1.4 Large Language Models (LLMs)

Recently, large language models (LLMs) such as GPT-3 [17] have shown superior performance in various NLP tasks, especially in zero- and few-shot settings. Those models are typically large-scale (billions of parameters) decoder-only transformers that have been pretrained on large corpora of raw text crawled from the internet. Notably, LLMs are often used with in-context learning, where the model generates the target output (e.g., class label) given a "prompt" that contains task instructions and several illustrative input-output examples followed by the test input.

In the latest NLP research, LLM performance has been further improved by instruction tuning [58] and reinforcement learning (RL) from human feedback (RLHF [59]). Specifically, instruction tuning fine-tunes an LLM on a wide range (e.g., hundreds) of diverse NLP tasks whose examples share the unified ⟨instruction, target text⟩ format, so that the LLM can generalise well to new tasks. RLHF uses RL to optimise LLMs in such a way that the generated texts align with human preference, instead of simply minimising the language modelling loss. As a result, RLHF-trained LLMs have achieved better performance in terms of human preference [60].

In our work, we use GPT-3 in Chapter 6 for reflection generation.

2.1.5 NLU & NLG Performance Evaluation

NLU performance evaluation is usually straightforward when ground-truth class labels are available. Common metrics include precision, recall, F1, and matthews correlation coefficient (MCC [61]).

On the other hand, NLG evaluation is often less clear, especially for tasks like dialogue generation where there is not a unique optimal target text [62]. Nevertheless, automatic metrics still often compare generated texts with the reference text w.r.t. lexical overlap (e.g., BLEU [63], ROUGE [64]) or word-/sequence-level semantic similarity (e.g., BERTScore [65]), although some recent works have explored reference-free automatic evaluation [66, 67].

In the rest of this sub-section, we provide an overview of human evaluation for dialogue generation and expert/non-expert evaluation for NLG, since these two topics are relevant to our human evaluation research in Chapter 6.

Human Evaluation for Dialogue Generation

In most studies of dialogue generation, human evaluation is considered the ultimate benchmark, since it can assess quality aspects like interestingness and safety [68, 62, 69] that may elude automatic metrics. Typically, the human annotator rates model-generated responses in an interactive or static setup.

In an interactive setting, the human converses with the model and evaluates its responses as good/bad (e.g., [70]) or selects applicable attributes like knowledgeable and engaging (e.g., [71]). In a static setup, the human evaluates responses or entire dialogues on a Likert scale for an attribute (e.g., [72, 73]) or compares responses from different models through ranking or A/B testing (e.g., [74, 75]).

Despite their popularity, standard human evaluation protocols suffer from various issues. One such example is subjectivity [76, 77], in particular in the context of Likert scales. Other issues include the lack of reproducibility across studies and the influence of evaluation instructions [78, 79, 80].

Expert and Non-Expert Evaluation for NLG

Whether to use experts for NLG evaluation generally depends on the domain. For example, open-domain dialogue generation mostly involves non-experts to assess attributes like engaging-ness and human-ness (e.g., [81, 71]), while dialogue generation for specialised domains like mental health [82] and clinical dialogue [83] is largely evaluated by domain experts.

Some human evaluation studies have compared expert and non-expert NLG evaluations, including for summarisation [84, 85], machine translation [86], story generation [87] and others (e.g., [88]). Many of these works reveal considerable gaps between assessments from experts and those from crowdworkers. In particular, Freitag et al. [86] found that automatic metrics can outperform crowdworkers in terms of correlation with expert judgement.

2.2 NLP & Speech Processing for MI

In this section, we present a literature review of MI-related studies that are relevant to the contributions of our research, focusing on 4 topics: resources (§2.2.1), automated therapist empathy assessment (§2.2.2), automated utterance coding (§2.2.3) and reflection generation (§2.2.4). Where applicable, we include both NLP and speech/multimodal approaches to better contextualise our research.

2.2.1 MI Resources

We provide below an overview of MI resources created before and after the release of our `AnnoMI` dialogue dataset, in order to provide the background for Chapter 4 which details the creation of `AnnoMI`.

Before AnnoMI

MI conversation resources are scarce. As real-world counselling often contains sensitive topics and information, counselling dialogues are mostly privately owned or proprietary (e.g., counselling transcripts from Alexander Street[1]). Such MI dialogues [89, 90, 32, 91] are often transcripts/recordings of MI sessions from clinical trials on topics such as alcohol- and drug-related issues, and they are generally coded following established protocols like MISC/MITI. For example, Lee et al. [89] created a corpus of intervention sessions for 783 university students, where the goal is to reduce high-risk drinking during spring breaks, and collected MITI-based session ratings given by trained coders. Notably, the dataset from [92] also includes sources such as wellness coaching phone calls and students' counselling sessions from a graduate level MI course. NLP-for-MI studies have leveraged those datasets for tasks such as utterance-level code prediction (e.g., [93]) and session-level empathy prediction (e.g., [94]), but it is hard to replicate or build on those studies, since the datasets remain publicly inaccessible.

Prior to the publication of AnnoMI, to the best of our knowledge, the only freely and publicly available MI dialogue corpus was from Pérez-Rosas et al. [1], who created it based on automatic transcripts of MI videos on YouTube/Vimeo demonstrating high- and low-quality MI. In the same study, the authors also collected annotations w.r.t. reflections and questions for the corpus and conducted related analyses, but those annotations are not available at the time of writing. Also, considerable transcription errors from automatic captioning are present in the corpus (§4.2.2), thus limiting the quality of the dataset.

After AnnoMI

There have also been MI-related dataset creation studies since the publication of AnnoMI, of which we detail several below. We also include some non-dialogue MI datasets, since such resources are increasingly being developed and utilised.

In terms of dialogues, Shah et al. [95] and Welivita et al. [96] collected short conversations from online peer-to-peer counselling/support forums and annotated them with MITI-based schemes. In both studies, conversations are one-to-one between a speaker and a listener, where the listener shows support and empathy to the speaker's situation. Listeners in [95] are volunteers (not professional therapists) who have undergone brief MI-related training, while those in [96] are ordinary Reddit[2] users with no counselling experience.

There have also been recent efforts in curating non-dialogue MI resources. For example, Min et al. [97] compiled a dataset of ⟨client statement, therapist reflection⟩ pairs. The client statements come from MI teaching materials and cover diverse

[1] https://alexanderstreet.com/products/counseling-therapy-video-library

[2] Reddit (https://www.reddit.com/) is an online platform comprised of sub-forums (known as subreddits), each with a specific topic for Reddit users to discuss.

topics, while the therapist reflections are written by both MI specialists and crowd workers to emulate high, medium and low reflection quality for comparison. In terms of non-English resources, Meyer et al. [98] collected client statements from a German-language weight loss forum and annotated them as change/neutral/sustain talks based on MISC. Notably, the labels are fine-grained for change and sustain talks, showing sub-types including reason (desire/ability/need), commitment and taking steps as defined in MISC.

While those resources are useful for MI research, we note that, unlike `AnnoMI`, none of those datasets contains long, multi-turn dialogues between a professional therapist and a client. Furthermore, at the time of writing, [95] and [98] are not publicly available. Therefore, `AnnoMI` remains a unique and valuable resource.

2.2.2 Automated Therapist Empathy Assessment for MI

In counselling, therapist empathy is known to be a crucial component and enables better outcomes [99, 100]. Accordingly, both MISC and MITI include session-level therapist empathy rating on a Likert scale, which shows the degree of empathy displayed by the therapist throughout an MI session. To date, automated therapist empathy assessment has been explored from both the NLP (transcript-based) and speech/multimodal (recording-based) perspectives to alleviate the need for trained coders to rate session-level empathy as feedback for the therapist. Notably, studies in this direction generally convert the Likert scale into a binary label space to facilitate model training, e.g., considering 5/6/7 on a 7-point Likert scale as a "high empathy" label and 1/2/3/4 as a "low empathy" label [30].

This sub-section serves as an overview of related work in this direction, in order to provide context for Chapter 3 which focuses on the novel task of low-resource utterance-level therapist empathy assessment.

Text-Based Approaches

Xiao et al. [29] proposed one of the earliest approaches using an N-gram LM. This method is unique in that it first trains an N-gram LM on more than 7K individual MI utterances annotated as empathetic/non-empathetic, and then uses the LM to extract N-gram-based features for a Bayesian linear regression model that predicts session-level empathy. To the best of our knowledge, this work is the only study that has approached utterance-level empathy prediction, which is also our objective in Chapter 3. We note that our work differs in its focus on zero-shot (low-resource), use of advanced transformer LMs, and exploration of empathy in non-counselling dialogues for the objective.

Also based on classical machine learning (ML), the work of Gibson et al. [30] leveraged both N-gram and psycholinguistic norm features. Specifically, those norms are computed at the word level to indicate the similarity of the word to a range of categories that reflect affective (e.g., arousal) and cognitive (e.g., concreteness/abstractness) processes. The analysis shows that N-gram and norm features are complementary in contributing to better performance.

More recently, a two-step deep learning (DL) framework was introduced in [31], which requires less feature engineering. At the first step, an LSTM [19] is trained to predict the MISC codes of individual utterances. Then, the trained LSTM is used to produce an encoding for each utterance in an MI session, and the utterance encodings are fed to a DNN that predicts the empathy label of the entire session. This method outperforms an ablative model that predicts the final label from the utterances directly, which shows that learning utterance-level dialogue dynamics (utterance code prediction) contributes to global empathy prediction.

Speech-Based and Multimodal Approaches

Speech-based methods generally use classical ML models with extracted speech features as the input. Notably, those studies tend to adopt a leave-one-therapist-out or leave-one-session-out setup to test the model on voices of unseen therapists.

An early work is from Xiao et al. [101], who explored vocal entrainment, i.e., the increasing pitch alignment between therapist and client as an MI session develops. The experiments show that features derived from vocal entrainment are significantly correlated with therapist empathy ratings and lead to better-than-chance prediction in a binary high-vs.-low setting.

A later study [102] investigated the relationship between empathy and prosody (the non-verbal elements of speech), focusing on 5 categories of prosodic features — pitch, shimmer, jitter, utterance duration and vocal energy. Aside from outperforming the model of [101] in accuracy, this approach shows that empathetic therapists tend to have lower pitch and vocal energy. Notably, a follow-up study [103] combined features derived from both prosody and speech rate entrainment (i.e., therapist and client becoming similar in how fast they talk), reaching higher accuracy than when features from either source are ablated.

Lastly, we note that some studies have also proposed multimodal approaches that utilise both speech and linguistic features. An example is [94], which showed that multimodal models are comparable to linguistic-features-only models in overall performance but show better results on low-empathy dialogues.

2.2.3 Automated Utterance Coding for MI

Like session-level empathy rating, utterance-level behaviour coding is an integral part of major coding frameworks including MISC and MITI. Utterance codes (e.g., question/reflection/... for a therapist utterance) are particularly useful when they are aggregated to produce session-level insights, such as open question percentage and reflection/question ratio, which are essential to gauging therapist adherence to MI guidelines [16].

Since manual utterance coding is laborious, automated utterance coding has garnered considerable research interest. In a similar vein to automated empathy assessment, research works in this area can be divided into text-based — only using transcripts — and speech/multimodal — solely or additionally using recordings. In this sub-section, we provide an overview of those works to better contextualise Chapter 5, which approaches the tasks of current-turn therapist/client behaviour prediction and next-turn therapist behaviour forecasting.

Text-Based Approaches

Early approaches in this direction were based on classical ML models. For example, Can et al. [104] proposed a maximum entropy Markov model trained for reflection utterance classification. The model utilises features such as N-grams and similarity between the therapist utterance and its preceding client utterances, since reflections tend to "reflect back" what the client said. On the other hand, Atkins et al. [105] leveraged a labelled topic model, which is a variant of topic model that also learns from non-text data. In this case, the non-text data consists of utterance codes, allowing for the model to associate certain N-grams with certain codes and hence act as an automated coder.

As neural networks grew in popularity, studies began to adopt them for modelling. Tanana et al. [32] proposed to use recursive neural networks [106] to process utterances as dependency trees, where each leaf node is the GloVe embedding [107] of the corresponding word. Thus, the neural representation of the root node is passed to a multinomial regression model for code prediction. In contrast, Xiao et al. [33] leveraged bi-directional LSTMs and gated recurrent units (GRUs [108]) to process an utterance as a sequence of word embeddings, so that the LSTM/GRU hidden state of the final time step is processed by a dense layer to yield a code prediction. In particular, the study found that better performance is achieved with word embeddings trained on an in-domian psychotherapy corpus, instead of general GloVe [107] embeddings.

More recent works have developed more elaborate DL models. Gibson et al. [34] introduced a multi-task LSTM-based framework, where a single model is trained to predict both utterance-level codes for MI dialogues and session-level codes for cognitive behaviour therapy dialogues. Also, the study shows that its multi-label modelling benefits prediction of less frequent codes. Furthermore, Cao et al. [13] established two distinct classification tasks: 1) code categorisation, which predicts the code of a known utterance given also its preceding utterances as the context; 2) code forecasting, which predicts the code of the unknown upcoming utterance given the dialogue history. As both tasks use multi-turn dialogue context as the input, the study proposed hierarchical LSTMs to process the context, with word-level attention in the lower LSTM and utterance-level attention in the upper one.

Notably, Flemotomos et al. [15] developed an automatic rating tool for MI sessions, predicting a range of session-level scores (e.g., w.r.t. empathy and acceptance) in addition to utterance-level codes. The tool utilises speech technologies such as speaker diarisation and automatic speech recognition to generate a rich, role-labelled transcript. Then, each utterance is processed by a bi-LSTM with attention to yield an utterance code, while tf-idf features are extracted from a session transcript and fed to a support vector regressor to produce a score for each session-level attribute (e.g., empathy). Finally, those predictions are aggregated as insights in a feedback report for the therapist.

Speech-Based and Multimodal Approaches

Among speech-based approaches to utterance coding, few utilise speech features only. One such example is the work of Singla et al. [109], which uses a forced aligner to segment the speech signals of an utterance into words and then trains a Speech2Vec [110] encoder to generate speech features for each word, thus converting the utterance into a sequence of word-level speech features. The sequence is fed to a bi-directional LSTM with self-attention that eventually yields a code prediction. The analysis shows that this approach achieved competitive performance w.r.t. transcript-based methods, without requiring transcription.

Most other speech-based methods are multimodal, using both language and speech features. For example, Singla et al. [111] proposed to fuse, for each word in an utterance, its word embeddings and prosodic features (e.g., pitch, jitter, pause, loudness). Thus, an utterance becomes a sequence of word-level multimodal features that are fed to a bi-LSTM for code prediction (same as [109]), and the results show that the feature fusion leads to performance improvement w.r.t. the text-only baseline. Chen et al. [112] adopted a similar multimodal framework, but it exploits class confusion to improve code prediction, using the confusion matrix of a baseline prediction model. Concretely, for each group of codes frequently mis-classified as each other, a group-specific classifier is trained to further distinguish between them. Thus, when the baseline classifier is not sufficiently confident about its prediction, the group-specific classifier is used in addition to produce the final prediction, which is shown to improve performance in the experiments.

2.2.4 Reflection Generation for MI

Given a dialogue history, MI reflection generation is the task of generating an appropriate context-aware reflection to the last client utterance. As learning effective reflection requires considerable training time and expert supervision [113, 114], high-quality reflection generation can assist junior therapists in their training. As a result, reflection generation has gained more research focus in recent years, although many works (e.g., [115]) are formulated as retrieving scripted reflections and/or filling pre-defined reflection templates, which arguably lack flexibility considering the context-sensitivity of reflections, as opposed to free-form generation.

In order to provide more context for Chapter 6, which approaches free-form reflection generation and its human evaluation, we present below an overview of free-form MI reflection generation studies, focusing on the modelling and evaluation of the proposed reflection generators.

Modelling

One of the first works on free-form reflection generation is from Shen et al. [35], which introduced a GPT-2 [54]-based model with retrieval of potential reflection candidates to improve generation. Given an input dialogue history, the method first retrieves the most similar conversation from a corpus of MI transcripts using tf-idf features. Then, for each reflection in the retrieved conversation, a fine-tuned GPT-2-based classifier predicts whether it can be a reflection to the input dialogue history. Thus, the reflection with the highest classifier score is chosen as a "reflection candidate", which is combined with the given dialogue history as input for training a GPT-2-based reflection generator.

A more recent study [36] proposed a BART-based model enriched by commonsense and domain knowledge for better reflection generation. Given a dialogue history, relevant external knowledge is both retrieved and generated to serve as additional context for better reflection generation. For knowledge retrieval, medical and commonsense knowledge triples [116] are verbalised into natural language and then compared against the dialogue history w.r.t. sentence embedding similarity, so that the most relevant triples are selected. For knowledge generation, a generative commonsense knowledge model [117] is adopted to directly generate relevant triples for salient phrases in the dialogue history.

Among the first attempts at few-shot reflection generation with LLMs is [37]. However, in this work, the input is not strictly a dialogue history but a $\langle$situation, client statement$\rangle$ pair, where an ideal output is a situation-aware reflection to the client statement. This work adopts GPT-2 [54] and GPT-3 for few-shot learning and additionally fine-tunes GPT-2 for comparison, while exploring various hyperparameters, e.g., number of few-shot examples and top-p/k [118, 119] values during decoding. The study shows that GPT-3 has superior performance overall, although GPT-2 improves considerably after applying a post-hoc low-quality reflection filter.

We note that those works all use no more than 5 preceding utterances as the dialogue history, which is arguably limiting for reflection generation, since reflections are heavily based on conversation context [11, 16].

Evaluation

Evaluation of generated reflections can be divided into automatic and human: the former uses automatic NLG metrics for scores, while the latter asks for assessments from human evaluators who are mostly domain experts like professional therapists.

In terms of automatic evaluation, Shen et al. [35] considered both similarity to and difference from gold-standard reflections, as good reflections should contain elements of the gold-standard high-quality reflections without being their duplicates. The similarity metrics include ROUGE [64], which quantifies lexical overlap, and word/sentence-level embedding similarity, which measures semantic resemblance. The difference metric is the proportion of N-grams in generated reflections that are not present in the gold-standard reflection. A similar automatic evaluation setup is adopted in [36], but it differs in 1) its additional use of BLEU [63] and METEOR [120] for lexical overlap; 2) its adoption of BERTScore [65] for embedding similarity; 3) its diversity metric as the proportion of distinct N-grams among generated reflections instead of w.r.t. the gold standard.

Like for other response generation tasks, automatic evaluation for reflection generation can be challenging due to factors such as the one-to-many nature of dialogue (i.e., there is more than one optimal response). For example, Shen et al. [35] found that the similarity metrics correlate weakly with human judgement, and that simply measuring reflection lengths can have better correlations in comparison.

For human evaluation, Shen et al. [35] sampled 50 dialogue histories and asked human evaluators with domain knowledge to rate the generated and gold-standard reflections on Likert scales w.r.t. 3 attributes: relevance to dialogue, displayed level of understanding of the client, and overall grammaticality/fluency. The same framework is used in [36], along with additional setups such as asking the evaluators to choose the better reflection between generated ones and the gold-standard. On the other hand, Ahmed et al. [37] sampled 369 ⟨situation, client statement, generated reflection⟩ triples and asked an professional therapist to evaluate the reflections in a binary setting, i.e., whether the reflection adheres to MI guidelines, as well as to propose changes to the non-adherent reflections to improve their adherence.

While human evaluation for reflection generation is more holistic and flexible than automatic evaluation, it still has many issues. For example, as shown in [35], the agreement between human evaluators can be poor and therefore make it challenging to reach a definitive conclusion on reflection quality.

An important limitation of the human evaluation frameworks listed above is the lack of consideration for hallucination, especially since they use very short ($\leq$ 5 turns) dialogue histories for generation and evaluation. While a low rating on "relevance to dialogue" may suggest that the reflection is off-topic, it is only one type of hallucination. Ji et al. [40] defined a response to be "intrinsic" hallucination if it contradicts the input (e.g., [121, 122]) and "extrinsic" hallucination if it cannot be verified based on the input (e.g., [123, 81]). Hence, a hallucinating reflection can be on-topic but contradict the context (intrinsic) or be unverifiable w.r.t. the context (extrinsic). Since reflective listening is highly context-sensitive, we argue that a hallucinating (and thus unnatural-sounding) reflection can cause quick client disengagement, and therefore hallucination should be an important evaluation aspect — we address this in Chapter 6.

2.3 NLP for Empathetic Non-Counselling Dialogue

2.3.1 Overview

Counselling dialogues such as MI are an extremely low-resource domain, since they are subject to strict privacy-related restrictions. In comparison, empathetic non-counselling dialogues have far fewer constraints on resource creation and sharing, and as a result they have garnered significant research interest in recent years. At the same time, people seeking support are increasingly turning to some forms of empathetic non-counselling dialogue, such as peer-support conversation, which provides a therapeutic and beneficial experience [95, 124, 125]. Notably, researchers have compared empathetic counselling and non-counselling dialogues to investigate the extent to which peer and expert behaviours align (e.g., [96]). In our work, we utilise empathy/non-empathy of non-counselling dialogues to approach therapist empathy assessment for MI dialogues (Chapter 3).

In the rest of this section, we present an overview of recent research in NLP for empathetic non-counselling dialogues. In particular, we consider 2 main dialogue types:

- General Open-Domain Dialogues: open-domain conversations about daily life, typically between a speaker, who initiates the dialogue to share an experience/situation/feeling, and a listener who shows empathy in their replies. Sources of such conversations include many sub-forums on Reddit for chitchatting (e.g., r/Happy) and curated datasets. A popular benchmark dataset is EmpatheticDialogues [72], which contains empathetic conversations between crowdworkers and grounded in specific speaker emotions and situations.

- Peer-Support Dialogues: typically between a “support seeker”, who initiates the dialogue to share their struggles and raw feelings, and a “peer supporter” who offers empathy and emotional support in their replies [82]. Sources of such dialogues are mostly online communities (including some sub-forums of Reddit, e.g., r/depression) and curated datasets. A popular benchmark dataset is ESConv ([126]), which contains peer-support dialogues between crowdworkers, where each peer supporter utterance is annotated with its peer-supporting strategy, e.g., “self-disclosure” and “providing suggestions”.

For both dialogue types, we review recent studies on dialogue analysis and dialogue generation: the former explores dialogue patterns and dynamics, while the latter builds a response generator that plays the listener (/peer supporter) role and produces an empathetic response given a dialogue history of preceding utterances.

2.3.2 Dialogue Analysis

General Open-Domain Dialogue

Some studies in this area focus on emotion dynamics in such dialogues. For example, Li et al. [127] created links between utterances in each conversation based on emotion-related commonsense [117]. Thus, a graph transformer [128] processes the utterances as nodes and the links as edges, and its final output is used to predict the emotion of each utterance. Relatedly, the work of [129] investigated the task of predicting the emotion of the upcoming utterance given the dialogue history so far, using commonsense-enriched utterance embeddings as well as LSTM gates to modulate the speaker's emotion drift caused by the listener.

In terms of empathy itself, Welivita et al. [130] annotated the listener utterances in EmpatheticDialogues based on a taxonomy of fine-grained empathy intents, such as questioning, acknowledging and agreeing. The annotations reveal many empathy patterns in this dataset, e.g., acknowledging and questioning are the most prominent intents overall. Relatedly, Svikhnushina et al. [131] built a taxonomy for empathetic questions by tagging each question with an act and an intent, where the act indicates the communicative purpose, e.g., "request information", while the intent indicates the intended emotional impact on the speaker, e.g., "amplify speaker's pride". On the other hand, Xie et al. [132] clustered a corpus of empathetic dialogues into a large-scale graph of inter-connected utterances that captures the dynamics of such conversations, and they showed that the graph is effective for building a retrieval-based empathetic bot.

Peer-Support Dialogue

In terms of empathy classification, Khanpour et al. [133] built binary empathy classifiers for messages exchanged on a network of cancer survivors, using LSTMs and convolutional neural networks (CNNs [134]). Relatedly, Hosseini et al. [135] collected dialogues from the same network and provided fine-grained sentence-level annotations of empathy seeking and providing. They also showed that empathy-providing considerably contributes to positive mood shifts of support seekers.

Notably, Sharma et al. [136] proposed an empathy analysis framework for conversations on a mental health platform, dividing empathetic replies into 3 categories: emotional reaction to seeker's post, interpretation of seeker's feelings, and exploration of seeker's experience not explicitly mentioned in the post. Each category is combined with an intensity label, e.g., strong exploration and weak interpretation. On a related note, the work of [137] analysed engagement patterns in post threads on the same platform, using both attention-based indicators (number of posts and peer supporters in a thread) and interaction-based ones (posting frequency and seeker-supporter interaction levels within a thread). The findings also provide suggestions on designing such platforms for better retention of seekers and supporters.

For Reddit, Zhou et al. [138] analysed condolence/distress-related sub-forums, showing that elements of effective empathetic responses in those online channels have notable differences when compared to face-to-face settings. On the other hand, Welivita et al. [139] developed an extensive conversation graph on a large corpus of distress-related Reddit dialogues, in order to represent the various flows in such conversations. The types of nodes in this graph include stressors (e.g., suicidal thoughts) and speaker feedback types (e.g., agreeing/suggesting/...), etc..

2.3.3 Dialogue Generation

General Open-Domain Dialogue

Research in this area has mostly focused on emotion-aware dialogue modelling to generate more empathetic responses.

Earlier approaches were generally based on recurrent neural networks (RNNs [140]). For example, Zhou et al. [141] trained conditional variational autoencoders to generate appropriately emotion-conditioned tweet responses, while Lubis et al. [142] built a hierarchical sequence-to-sequence (seq2seq) model to produce replies that evoke positive speaker emotions, using dialogue sources such as movie subtitles and Wizard-of-Oz conversations.

Since the release of datasets such as EmpatheticDialogues and their transformer-based dialogue generation baselines, recent studies have mostly adopted transformer-based models and benchmarked them on those new datasets. In those studies, emotions are utilised in various ways. For example, Lin et al. [143] used speaker emotion detection as an auxiliary objective to response generation in a multi-task setup, while the work of [144] introduced a mixture-of-experts framework, where each speaker emotion has its separate response generator and the outputs of all those models are aggregated to yield a final response. Furthermore, Zeng et al. [145] showed that simply fusing the dialogue context embedding and speaker/listener emotion embedding during decoding can lead to competitive performance.

Notably, some recent works have explored emotions beyond the surface level. As an example, Kim et al. [75] proposed to recognise the words in a speaker utterance that cause its emotion, so that the generator can include those words in its response. Furthermore, Wang et al. [146] extracted dialogue history parts that contain emotion causes, and used them to retrieve relevant triples from a commonsense knowledge graph [116] to modulate next-token distribution during decoding. Relatedly, Sabour et al. [147] explored both emotion- and situation-related commonsense knowledge from a generative commonsense model [117] as auxiliary input.

Peer-Support Dialogue

Thanks to datasets such as ESConv that contain utterance annotations w.r.t. peer-supporting strategies, researchers have developed empathetic response generators guided by strategy planning. For example, Cheng et al. [148] proposed forecasting of strategies in future steps and consequent feedback from the support seeker. Thus, at the next step, the model chooses a strategy that will optimise support seeker feedback over a longer time window. On the other hand, the work of [149] proposed to explicitly derive the support seeker's persona based on the dialogue context, and the persona is then incorporated into a strategy-informed decoding approach to produce more personalised empathetic responses.

Some studies have also utilised external knowledge as auxiliary input to the generator. In [150], the support seeker's utterances are fed to a commonsense model [117] to deduce the emotion transition of the support seeker, and it is used together with soft (instead of one-hot) strategy selection for more effective and natural empathetic responses. Relatedly, Deng et al. [151] used the same commonsense model to enhance seeker utterances, so that they can serve as queries to retrieve relevant emotional knowledge from an extensive graph of distress-related dialogue dynamics.

On a different but related front, Sharma et al. [82] investigated the task of empathetic re-writing. Given a dialogue history and a peer supporter reply of low empathy, the goal is to edit the reply to increase its level of empathy. The work adopts RL to improve a fine-tuned DialoGPT [152]-based response generator, rewarding edits that lead to higher levels of empathy, fluency, coherence and context-specificity. In a follow-up work, Sharma et al. [153] deployed the re-writer as a smartphone application to assist peer supporters in a randomised controlled trial, where the system proved capable of helping peer supporters write considerably more empathetic messages.

Chapter 3

Low-Resource Real-Time Therapist Empathy Assessment

This chapter is based on:

Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Towards low-resource real-time assessment of empathy in counselling. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 204–216, Online, June 2021. Association for Computational Linguistics

Gauging therapist empathy in MI is an important component of understanding counselling quality. While ML-based post-session empathy assessment at the session level has been investigated extensively, it relies on relatively large amounts of MI dialogue data with empathy annotations. Also, real-time empathy assessment has largely been overlooked in the past, despite its promising applications in delivering real-time feedback for the human therapist. In this chapter, we focus on the task of zero-shot utterance-level binary empathy assessment. We develop 1) a supervised method that leverages heuristically constructed empathy vs. non-empathy contrast in non-counselling conversations, and 2) an unsupervised method that formulates natural language inference as a proxy task for empathy prediction. Our results show that the empathy vs. non-empathy contrast enables the best performance, even though it is not sufficiently high. Upon probing, we find that the benefit of the contrast becomes clear when it is compared to the unsupervised approach and control-group supervised models without empathy contrast training. We note that this chapter is an improved version of our paper [43].

3.1 Introduction

In counselling, therapist empathy is known to be a crucial component and enables better outcomes [99, 100]. In particular, "listening with empathy" is considered a guiding principle in MI [11]. Therefore, gauging therapist empathy is essential to assessing MI integrity [2].

Conventionally, empathy assessment for MI has been conducted manually by trained annotators, which requires extensive annotator training and transcript review. Since such a time-consuming and costly process is difficult to scale, recent years have seen attempts at automating the process with ML, including transcript-based [29, 30, 31], speech-based [102, 103], and multimodal [94] methods (§2.2.2). However, those works are limited in that

- They nearly always only assess therapist empathy at the session level after a session is completed, rather than at the utterance level while a session is ongoing.

- They are all based on sizeable privately-owned MI dialogue corpora with empathy annotations at the session level, but in reality such well-annotated data is often very limited and unavailable publicly due to privacy constraints.

- They use classical ML with heuristic feature engineering, while recent DL frameworks have not been utilised for this purpose.

- They do not explore the link between empathy in non-counselling conversation and empathy in MI.

In this work, we make the first attempt at addressing those limitations by exploring supervised and unsupervised zero-shot therapist empathy prediction for MI at the utterance level. We utilise pre-trained LMs such as BERT [21] for text-based binary empathy/non-empathy prediction on a therapist utterance, optionally taking the preceding client utterance as additional input for more context.

Our supervised approach (Figure 3.1) learns from the empathy vs. non-empathy contrast (referred to as **empathy contrast** for brevity) in non-counselling conversations, in other words out-of-domain (OOD) data. To this end, we leverage publicly available datasets of non-counselling conversations with heuristic empathy labels [72, 154] for OOD empathy contrast training. In doing so, our assumption is that there are links between therapist empathy and empathy in non-counselling dialogues, and we investigate whether a good classifier of empathy in non-counselling dialogues can also perform well on therapist empathy.

Our unsupervised approach (Figure 3.2) leverages models fine-tuned for natural language inference (NLI), the task of predicting the logical relationship between a premise and a hypothesis. Specifically, we reformulate binary empathy prediction as a proxy NLI task in 2 ways: empathy-as-hypothesis and empathy-as-alignment.

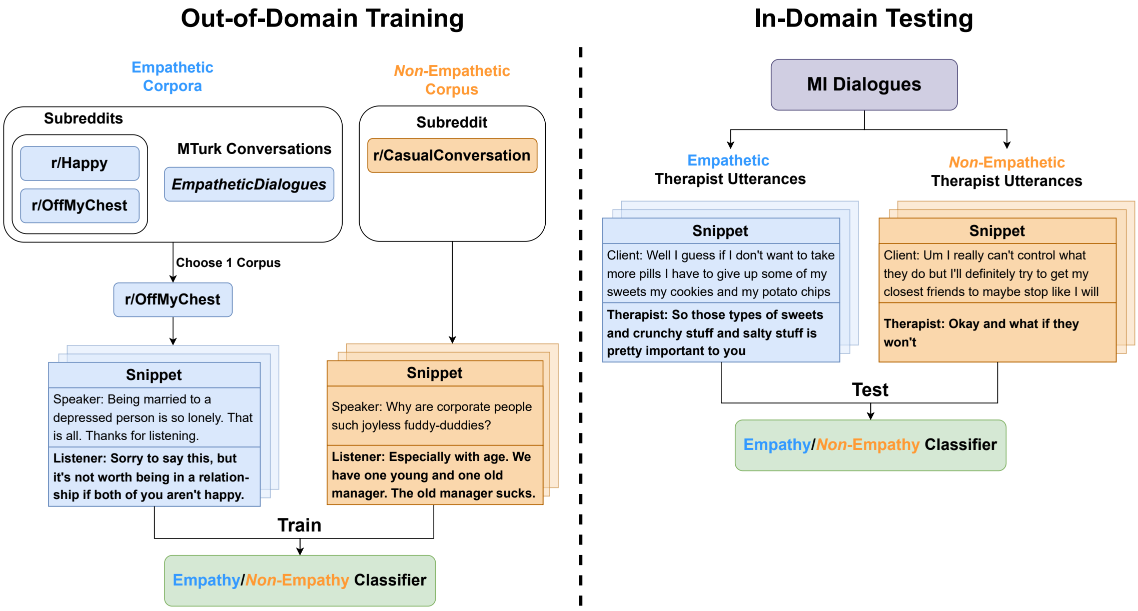

Out-of-Domain Training
Empathetic Corpora
Non-Empathetic Corpus
Subreddits
r/Happy
r/OffMyChest
MTurk Conversations
EmpatheticDialogues
Subreddit
r/CasualConversation
Choose 1 Corpus
r/OffMyChest
Snippet
Speaker: Being married to a depressed person is so lonely. That is all. Thanks for listening.
Listener: Sorry to say this, but it's not worth being in a relationship if both of you aren't happy.
Snippet
Speaker: Why are corporate people such joyless fuddy-duddies?
Listener: Especially with age. We have one young and one old manager. The old manager sucks.
Train
Empathy/Non-Empathy Classifier
In-Domain Testing
MI Dialogues
Empathetic Therapist Utterances
Non-Empathetic Therapist Utterances
Snippet
Client: Well I guess if I don't want to take more pills I have to give up some of my sweets my cookies and my potato chips
Therapist: So those types of sweets and crunchy stuff and salty stuff is pretty important to you
Snippet
Client: Um I really can't control what they do but I'll definitely try to get my closest friends to maybe stop like I will
Therapist: Okay and what if they won't
Test
Empathy/Non-Empathy Classifier

Figure 3.1: Overview of supervised approach: OOD empathy contrast training on non-counselling conversations and in-domain testing on MI dialogues. **Listener/Therapist** utterance is required; preceding Speaker/Client utterance as context is optional.

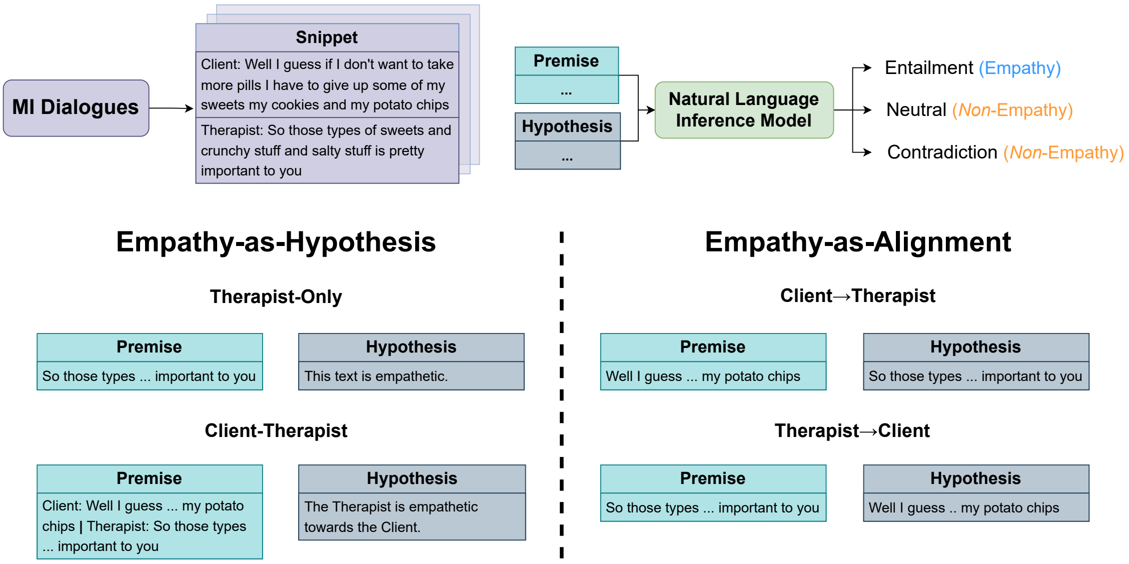

MI Dialogues
Snippet
Client: Well I guess if I don't want to take more pills I have to give up some of my sweets my cookies and my potato chips
Therapist: So those types of sweets and crunchy stuff and salty stuff is pretty important to you
Premise
...
Hypothesis
...
Natural Language Inference Model
Entailment (Empathy)
Neutral (Non-Empathy)
Contradiction (Non-Empathy)
Empathy-as-Hypothesis
Therapist-Only
Premise
So those types ... important to you
Hypothesis
This text is empathetic.
Client-Therapist
Premise
Client: Well I guess ... my potato chips | Therapist: So those types ... important to you
Hypothesis
The Therapist is empathetic towards the Client.
Empathy-as-Alignment
Client→Therapist
Premise
Well I guess ... my potato chips
Hypothesis
So those types ... important to you
Therapist→Client
Premise
So those types ... important to you
Hypothesis
Well I guess .. my potato chips

Figure 3.2: Overview of unsupervised approach based on NLI.

Empathy-as-hypothesis directly asks the NLI model whether the therapist utterance shows empathy, by verbalising the empathy class as the hypothesis, e.g., “this text is empathetic”. Empathy-as-alignment tests client-therapist language alignment through logical entailment, as an empathetic therapist tends to acknowledge the difficulties and feelings of the client [11].

To evaluate the models on MI dialogues, we manually annotate utterance-level therapist empathy for a subset of a publicly available corpus of transcribed dialogues demonstrating high- and low-quality MI [1]. Our experiments show that models trained on OOD empathy contrast have the best performance, but they are not sufficiently accurate classifiers of MI empathy/non-empathy, likely due to the domain gap between non-counselling conversation and therapy. Nevertheless, the benefit of such training becomes clear when compared to A) training on OOD data without empathy contrast and B) the unsupervised NLI-based approach. We also show that the best-performing empathy contrast setup is where the empathetic examples involve deeply emotional experiences, which may be because those examples are relatively more similar to client-therapist interactions.

3.2 Data

We leverage two types of data[1]: non-counselling conversations and MI dialogues, both of which are in the form of two conversation participants (also known as interlocutors) taking turns to speak to each other. For both types of data — and in this thesis in general — we define an **utterance** as *"everything said by an interlocutor in their turn"*, which is the most widely used definition of utterance in the literature of DL for conversational AI.

We note that the definition of utterance in this thesis differs from some definitions of utterance in the counselling literature. For example, an "utterance" in this work is identical to a "volley" as defined in [16], while an "utterance" in [16] is "a complete thought that ends either when one thought is completed or a new thought begins with the same speaker, or by an utterance from the other speaker".

[1] Identifiable information (e.g., names, dates) is replaced with placeholders in our experiments.

Table 3.1: Overview of statistics of filtered *RedditConvs* and *EmpDial*. #(Dialogues): number of dialogues. Avg#(Speaker/Listener Utts): average number of speaker/listener utterances per dialogue. Avg(Speaker/Listener Utt Len): average speaker/listener utterance length, namely number of tokens.

RedditConvs - **Empathetic & Non-Empathetic**

	Empathetic						**Non-Empathetic**		
	r/Happy			*r/OffMyChest*			*r/CasualConv*		
	Train	Dev	Test	Train	Dev	Test	Train	Dev	Test
#(Dialogues)	114K	14K	16K	94K	12K	12K	530K	67K	67K
Avg#(Speaker Utts)	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.1	1.1
Avg#(Listener Utts)	1.0	1.0	1.0	1.0	1.0	1.0	1.1	1.1	1.1
Avg(Speaker Utt Len)	30.8	30.2	30.4	48.9	51.0	47.8	42.8	42.9	43.2
Avg(Listener Utt Len)	13.3	13.5	13.3	15.7	15.7	15.6	16.9	16.8	16.8

EmpDial - **Empathetic**

	Train	Dev	Test
#Dialogues	18K	3K	3K
Avg#(Speaker Utts)	2.2	2.3	2.2
Avg#(Listener Utts)	2.1	2.1	2.1
Avg(Speaker Utt Len)	17.6	19.4	21.2
Avg(Listener Utt Len)	13.7	14.3	14.5

3.2.1 Non-Counselling Conversations

For a 2-person non-counselling conversation, we consider the initiator of the dialogue as the **speaker** and the other as the **listener**. We use two non-counselling conversation datasets: *Persona-based Empathetic Conversation* [154] (referred to as ***RedditConvs*** for brevity) and *EmpatheticDialogues* [72] (referred to as ***EmpDial*** for brevity). Table 3.1 presents an overview of the data statistics and Table 3.2 shows example dialogues.

RedditConvs consists of conversations crawled from 3 subreddits: *r/Happy*[2], *r/OffMyChest*[3], and *r/CasualConversation*[4] (referred to as ***r/CasualConv*** for brevity). Reddit users exchange happy experiences and thoughts in *r/Happy*, share deeply emotional stories that cannot be told easily in *r/OffMyChest*, and simply talk casually in *r/CasualConv*. We filter *RedditConvs* to remove a conversation if A) it has more than two participants and/or B) it is effectively a subset of another conversation, such as a 3-turn conversation that is actually the first 3 turns of a 5-turn conversation. Thus, the filtered *RedditConvs* aligns with the 2-person — therapist

[2] https://www.reddit.com/r/happy/
[3] https://www.reddit.com/r/offmychest/
[4] https://www.reddit.com/r/CasualConversation

Table 3.2: Example dialogues from *RedditConvs* and *EmpDial*. The dialogue lengths reflect the Avg#(Speaker/Listener Utts) shown in Table 3.1.

r/Happy	*r/OffMyChest*	*r/CasualConv*
Speaker: Can't believe how happy this photo makes me! Got to marry this gorgeous man! **Listener**: You both look amazing! Congrats!	**Speaker**: Being married to a depressed person is so lonely. That is all. Thanks for listening. **Listener**: Sorry to say this, but it's not worth being in a relationship if both of you aren't happy.	**Speaker**: Why are corporate people such joyless fuddy-duddies? **Listener**: Especially with age. We have one young and one old manager. The old manager sucks.

EmpDial
Speaker: I really broke down when I heard my mom was sick. **Listener**: I'm so sorry. You feel so helpless when someone you love is ill. **Speaker**: Yeah, I cried and cried ... but I believe she will be OK soon. **Listener**: Oh, I'm so glad to hear that! I hope for her speedy recovery!

and client — nature of therapy dialogue, and it accounts for 56% of the conversations in the original *RedditConvs*.

EmpDial contains 23.1K non-counselling conversations between pairs of crowd-source workers on Amazon Mechanical Turk[5]. Each dialogue is conditioned on an emotion label, e.g., "Devastated", and a situation involving the emotion, e.g., "I really broke down when I heard my mom was sick". Given the emotion and situation, the speaker initiates a conversation about this situation with a listener.

Heuristic Empathy Labelling

Using heuristics, we label all listener utterances in *r/Happy*, *r/OffMyChest* and *EmpDial* to be empathetic, and all listener utterances in *r/CasualConv* to be non-empathetic.

For *RedditConvs*, the heuristics is based on the annotator ratings [154] which show that listener utterances in *r/Happy* and *r/OffMyChest* are significantly more empathetic than those in *r/CasualConv*, and the inter-annotator agreement (IAA) on this is substantial as measured by Fleiss' kappa [155]. For *EmpDial*, the heuristics is straightforward, since the authors explicitly instructed the crowdworkers to respond as empathetic listeners during data collection.

We note that our heuristic labelling for *RedditConvs* and *EmpDial* is based on the corpus-level empathy labels given by the creators of the datasets, thus it may

[5] Amazon Mechanical Turk (`https://www.mturk.com/`) is a crowdsourcing website.

Table 3.3: Overview of DemoMI and AnnoDemoMI. #(Dialogues): number of dialogues. Total#(Therapist/Client Utts): total number of therapist/client utterances. Avg#(Therapist/Client Utts): average number of therapist/client utterances per dialogue. Avg#(Therapist/Client Utt Len): average therapist/client utterance length, namely number of tokens. %(Empathetic Therapist Utts): percentage of empathetic therapist utterances.

	DemoMI		AnnoDemoMI	
MI Quality	**High**	**Low**	**High**	**Low**
#(Dialogues)	152	101	7	14
Total#(Therapist Utts)	3928	1534	185	181
Total#(Client Utts)	3808	1460	185	181
Avg#(Therapist Utts)	25.8	15.2	26.4	12.9
Avg#(Client Utts)	25.1	14.5	26.4	12.9
Avg(Therapist Utt Len)	33.5	31.1	33.9	33.7
Avg(Client Utt Len)	28.5	20.6	23.5	21.5
%(Empathetic Therapist Utts)	n/a	n/a	38.4%	3.3%

not be completely accurate at the utterance or sentence level. We nevertheless utilise the heuristic utterance labels for our experiments and leave more fine-grained annotation to future work.

3.2.2 MI Dialogues

Our MI conversations are from [1], a publicly available dataset of counselling dialogues. The conversations are 152 demonstrations of high-quality (i.e., MI adherent) counselling and 101 of low-quality (i.e., MI non-adherent) counselling from video-sharing platforms (YouTube/Vimeo), transcribed using YouTube automatic captioning. We refer to this dataset as DemoMI, and show its overview in Table 3.3.

Manual Empathy Annotation

We manually annotate empathy as a binary choice at the utterance level (i.e., whether or not a therapist utterance shows empathy) for a subset of DemoMI to build a benchmark dataset — AnnoDemoMI — for our models. The annotation guideline follows the definition of high empathy in MISC [16]: "*Therapists high on the empathy scale show an active interest in making sure they understand what the client is saying, including the client's perceptions, situation, meaning, and feelings.*"

As machine-transcribed spoken dialogues, DemoMI is more noisy and less clean than *RedditConvs* and *EmpDial* which contain exclusively written chats. Therefore, we ensure that the DemoMI transcripts chosen for annotation are clean and have minimal transcription noise. Under these criteria, we select 7 high-quality MI dialogues and 14 low-quality ones (Table 3.3). For those dialogues, we only consider the therapist utterances that have a preceding client utterance, as the MI principle of "listening with empathy" [11] means empathy is shown in the therapist's reply. Overall, the selected high- and low-quality MI dialogues are balanced, with 185 and 181 therapist utterances in total, respectively.

Two human annotators conduct binary empathy annotation on the therapist utterances. One annotator is a senior researcher who has received formal MI training in the past, and the other is a PhD student who has read MI literature in depth. We instruct the annotators to mark an utterance as **empathetic** if it shows MISC-defined high empathy, otherwise as **non-empathetic**, which ranges from neutrality to apathy. When annotating each therapist utterance, the annotators also state their confidence in the annotation on a 5-point Likert scale, with 1 being "not confident at all" and 5 being "totally confident". Notably, we instruct the annotators to mark an utterance as empathetic if at least a part of the utterance shows empathy, even if the other parts do not. This is because `DemoMI` transcripts do not have punctuation, which makes it non-trivial to separate an utterance into parts that show empathy and parts that do not. More fine-grained annotation would be possible with punctuated utterances, which we leave to future work.

Overall, the empathy annotations show a substantial IAA of 0.68 measured by Cohen's kappa [156]. For any therapist utterance that is annotated differently by the annotators, we resolve the annotations if a confidence criterion is met. Concretely, if an utterance is annotated as empathetic (/non-empathetic) by one annotator with a confidence of 4 or more and as non-empathetic (/empathetic) by the other annotator with a confidence of 2 or less, we consider the utterance empathetic (/non-empathetic). After this step, 90.1% of all the annotated therapist utterances have a unique final empathy label, and we refer to them and their respective preceding client utterances as `AnnoDemoMI`.

As Table 3.3 shows, 38.4% of the therapist utterances in the high-quality MI conversations in `AnnoDemoMI` are empathetic, while it is only 3.3% for low-quality MI dialogues. This suggests a marked difference between high- and low-quality MI in empathy. Table 3.4 presents example dialogue snippets from high- and low-quality MI in `AnnoDemoMI`, with empathetic therapist utterances highlighted.

3.3 Zero-Shot Empathy Prediction

In this section, we first define the task of zero-shot binary empathy prediction, then lay out both the supervised OOD training approach and the unsupervised method based on NLI.

3.3.1 Task Definition

Given an MI dialogue dataset $D^{MI} = \{\langle u_i^C,\ u_i^T,\ e_i \rangle\}_{i=1}^{N}$ of N examples where

Table 3.4: Example dialogue snippets from AnnoDemoMI. Underlined: empathetic. The snippets also show minimal transcription error ("weight"→"way") and a de-identification placeholder (⟨date/time⟩).

High-Quality MI
...
Therapist: Tell me what you know about the consequences of having high blood sugars that are untreated
Client: Well I know you know in reading on the Internet and doing my research certainly could end up blind or on dialysis and I know some people have even lost limbs that's really scary to me I don't want that either
Therapist: So on the one hand there are some things that really scare you about having diabetes that's uncontrolled and on the other hand it's been difficult for you I know we've talked in the past about working on controlling your your diet and looking at your **way** so where does that leave you at this point
Client: Well I guess if I don't want to take more pills I have to give up some of my sweets my cookies and my potato chips
Therapist: So those types of sweets and crunchy stuff and salty stuff is is pretty important to you
...
Low-Quality MI
...
Therapist: And that's going to be your goal and that's great you know that's really why we're here is to talk about that and make sure that you come to that decision um so I guess can you quit **⟨date/time⟩** then
Client: Yeah yeah definitely I'll just not have a drink
Therapist: Okay and then what about your friends who are still drinking
Client: Um i really can't control what they do but I'll definitely try to get my closest friends to maybe stop like I will
Therapist: Okay and what if they won't
...

- u_i^C is a client utterance;
- u_i^T is the therapist's reply utterance to u_i^C;
- $e_i \in \{\texttt{emp}, \texttt{non-emp}\}$ is the label indicating whether u_i^T shows empathy.

Our benchmark task is to predict e_i given u_i^T, optionally using u_i^C as additional input for more context. In practice, we use AnnoDemoMI as D^{MI}.

3.3.2 Supervised: OOD Training

We explore learning from empathy contrast in OOD data (Figure 3.1). Specifically, as described in §3.2.1, we utilise all listener utterances in *r/Happy*, *r/OffMyChest* and *EmpDial* as empathetic examples and all listener utterances in *r/CasualConv* as non-empathetic examples. Our assumption is that there are parallels between therapist empathy and empathy in non-counselling conversations, and that those parallels can be leveraged for this task.

Thus, we create **3 OOD pairs with empathy contrast** from non-counselling conversations in the form of ⟨Positive vs. Negative⟩, where empathetic examples fall into the positive class and non-empathetic ones fall into the negative class.

- ⟨*r/Happy* vs. *r/CasualConv*⟩
- ⟨*r/OffMyChest* vs. *r/CasualConv*⟩
- ⟨*EmpDial* vs. *r/CasualConv*⟩

We note that we use "positive" and "negative" simply as the names of two classes that are distinguished from each other, similar to "1" and "0". Therefore, "positive"/"negative" is NOT an alias for "empathy"/"non-empathy".

We also add a control group of **3 OOD pairs without empathy contrast**, where both the positive and negative examples are empathetic:

- ⟨*EmpDial* vs. *r/Happy*⟩
- ⟨*EmpDial* vs. *r/OffMyChest*⟩
- ⟨*r/OffMyChest* vs. *r/Happy*⟩

For each OOD pair — with or without empathy contrast — we sample an equal number of positive and negative examples to construct a non-counselling conversation dataset $D^{NC} = \{\langle u_j^S, u_j^L, l_j \rangle\}_{j=1}^{M}$ of M examples where

- u_j^S is a speaker utterance;
- u_j^L is the listener's reply utterance to u_j^S;
- $l_j \in \{\texttt{pos}, \texttt{neg}\}$ is the label indicating whether u_j^L is a positive or negative example.

Our sampling ensures that the positive and negative classes in each OOD pair are balanced. For each OOD pair, we train 2 OOD classifiers:

- **listener-only**: predict l_j given u_j^L,
- **speaker-listener**: predict l_j given $\{u_j^S, u_j^L\}$;

Table 3.5: Illustration of NLI. Example from `https://paperswithcode.com/task/natural-language-inference`.

Premise: A senior is waiting at the window of a restaurant that serves sandwiches.	
Hypothesis	**NLI Label**
A person waits to be served his food.	Entailment
A man is looking to order a grilled cheese sandwich.	Neutral
A man is waiting in line for the bus.	Contradiction

Table 3.6: Overview of NLI-based models as zero-shot empathy classifiers.

NLI Label	**Empathy Label**
Entailment	Empathy
Neutral	Non-Empathy
Contradiction	Non-Empathy

Model Type	**Model**	**Premise**	**Hypothesis**
Empathy-as-Hypothesis	therapist-only	u_i^T	“This text is empathetic.”
	client-therapist	“Client: u_i^C | Therapist: u_i^T”	“The Therapist is empathetic towards the Client.”
Empathy-as-Alignment	client→therapist	u_i^C	u_i^T
	therapist→client	u_i^T	u_i^C

Once the training is complete, we test the 2 classifiers directly on D^{MI}, treating the therapist as the listener and the client as the speaker, i.e., u_i^C as u_j^S and u_i^T as u_j^L. In order to calculate model performance, we treat `emp` as `pos` and `non-emp` as `neg`, but we reiterate that “positive”/“negative” examples are not equivalent to “empathetic”/“non-empathetic” examples — for instance, the negative examples are empathetic in the 3 OOD pairs without empathy contrast.

3.3.3 Unsupervised: NLI-Based

NLI [157] is an NLU task on sequence-level logical relationship. Given one sequence as the premise and another as the hypothesis, an NLI model predicts whether the premise entails, contradicts, or is neutral w.r.t. the hypothesis (Table 3.5). Inspired by [158] where NLI models prove effective as off-the-shelf zero-shot sequence classifiers, we formulate empathy prediction as an NLI task. We do so in two ways: **empathy-as-hypothesis** and **empathy-as-alignment**. An overview is given in Figure 3.2 as well as Table 3.6.

Empathy-as-hypothesis frames the hypothesis as a positive statement about therapist empathy. Specifically, we experiment with two models as detailed below:

- **therapist-only**: Using the therapist utterance (u_i^T) as the premise and the statement "This text is empathetic" as the hypothesis.

- **client-therapist**: Concatenating the therapist utterance (u_i^T) and its preceding client utterance (u_i^C) into "Client: u_i^C|Therapist: u_i^T" as the premise, and the statement "The Therapist is empathetic towards the Client" as the hypothesis.

For both therapist-only and client-therapist, we consider an Entailment prediction to be a proxy prediction of empathy for the therapist utterance, and similarly we treat neutral and contradiction as equivalent to non-empathy.

Empathy-as-alignment tests therapist-client alignment. This is based on both the literature [16, 11] and our observation of DemoMI that an empathetic therapist tends to acknowledge the client's difficulties and feelings, e.g., through reflections, which makes the therapist's language align with the client's. Thus, we explore

- **client→therapist**: Using the preceding client utterance (u_i^C) as the premise and the therapist utterance (u_i^T) as the hypothesis.

- **therapist→client**: Using the therapist utterance (u_i^T) as the premise and its preceding client utterance (u_i^C) as the hypothesis.

For both client→therapist and therapist→client, we consider an Entailment prediction to indicate therapist-client alignment and thus a proxy prediction of empathy for the therapist utterance. Similarly, we treat neutral and contradiction as equivalent to non-empathy.

3.4 Experiments

3.4.1 Implementation

Considering the empathy/non-empathy class imbalance of AnnoDemoMI (e.g., only 3.3% of therapist utterances in low-quality MI are empathetic), we choose MCC [61] as the metric because of its robustness to class imbalance. MCC ranges between -1 to 1, where -1 represents total disagreement between prediction and observation, 0 means no better than random prediction, and 1 indicates perfect prediction.

We leverage pre-trained LMs for all our experiments, using the HuggingFace framework[6] [159]. We implement model evaluation/testing with scikit-learn[7] [160].

[6] https://github.com/huggingface/transformers
[7] https://scikit-learn.org/stable/

OOD Training & Testing

For OOD training both with and without empathy contrast, we keep the original train/dev/test splits of *RedditConvs* and *EmpDial*. Since the two datasets in each OOD pair can be vastly different in size (e.g., *EmpDial* has only 17.8K training examples whereas *r/CasualConv* has 530.2K), we always sample the positive and negative examples such that they are identical in size to *EmpDial*, the smallest dataset. This ensures that A) the two classes are balanced in each pair, and B) different OOD models are trained with equal amounts of data and their performances are hence comparable.

In order to account for sampling bias, for each OOD pair, we sample OOD data 5 times to create 5 different splits of class-balanced {train, dev, set}. Thus, we effectively train 5 different models for each OOD pair using those 5 splits.

We choose BERT [21] (`BERT-BASE-UNCASED`) as the backbone of our OOD models. We add a fully connected layer on top of the classification token (`[CLS]`) position of the LM to implement a binary classifier. Thus, we train the entire model end-to-end on the OOD pairs.

During OOD training, we use a learning rate of 1e-5 and a batch size of 32. We evaluate every 500 steps on the dev set, and we stop training if dev set performance (measured by MCC) has not improved in the most recent 10 evaluations. Then, we select the best checkpoint w.r.t. the dev set and test it on both the OOD test set and `AnnoDemoMI`.

The input format to each OOD model is as follows:

- listener-only: {`[CLS]` u_j^L `[SEP]`} during OOD training/validation/testing and {`[CLS]` u_i^T `[SEP]`} during testing on `AnnoDemoMI`.
- speaker-listener: {`[CLS]` u_j^S `[SEP]` u_j^L `[SEP]`} during OOD training/validation/testing and {`[CLS]` u_i^C `[SEP]` u_i^T `[SEP]`} during testing on `AnnoDemoMI`.

NLI Setup

For the backbone of the NLI models, we use a BART [52] (`facebook/bart-large-mnli`) model that has been fine-tuned on MultiNLI [161], a large-scale NLI dataset that includes both transcripts and written texts.

The input format to each NLI model is as follows:

- empathy-as-hypothesis, therapist-only: {`[CLS]` u_i^T `[SEP]` This text is empathetic. `[SEP]`}
- empathy-as-hypothesis, client-therapist: {`[CLS]` Client: u_i^C | Therapist: u_i^T `[SEP]` The Therapist is empathetic towards the Client. `[SEP]`}
- empathy-as-alignment, client→therapist: {`[CLS]` u_i^C `[SEP]` u_i^T `[SEP]`}
- empathy-as-alignment, therapist→client: {`[CLS]` u_i^T `[SEP]` u_i^C `[SEP]`}

Table 3.7: Overview of OOD models' performance on their OOD test sets (not on `AnnoDemoMI`). The metric is MCC, which ranges from -1 to 1.

Positive	**Negative**	**Model**	**Mean**	**SD**
		With Empathy Contrast		
EmpDial	*r/CasualConv*	listener-only	0.816	0.006
		speaker-listener	0.919	0.006
r/Happy	*r/CasualConv*	listener-only	0.705	0.003
		speaker-listener	0.896	0.005
r/OffMyChest	*r/CasualConv*	listener-only	0.562	0.011
		speaker-listener	0.784	0.011
		Without Empathy Contrast		
EmpDial	*r/Happy*	listener-only	0.850	0.003
		speaker-listener	0.949	0.004
EmpDial	*r/OffMyChest*	listener-only	0.790	0.003
		speaker-listener	0.938	0.005
r/OffMyChest	*r/Happy*	listener-only	0.595	0.004
		speaker-listener	0.864	0.008

3.4.2 Results

OOD Models

To verify that the positive and negative examples in each OOD pair have distinct semantic features that can be captured by the OOD models, we first inspect the performance of these models (Table 3.7) on their OOD test sets rather than `AnnoDemoMI`. Since each OOD model is trained and tested 5 times using different train/dev/test data from random sampling, we report the mean and standard deviation of the 5 test-set results. For all OOD pairs, the speaker-listener model has considerably higher ($\Delta > 0.1$) performance than its listener-only counterpart, which is unsurprising since the speaker-listener model has more context in its input. Overall, regardless of with or without empathy contrast, speaker-listener OOD models are reliable classifiers on their test sets, with mean MCCs ranging from 0.784 to 0.949 and low performance variation from using different splits (standard deviation ≤ 0.011).

However, when the OOD models are tested on `AnnoDemoMI`, the performance drops significantly (Table 3.8). For example, the speaker-listener model trained on ⟨*r/OffMyChest* vs. *r/CasualConv*⟩ is the best-performing classifier but its mean MCC is only 0.176, compared to its OOD test-set performance of 0.784. Therefore,

Table 3.8: Overview of OOD models' performance on `AnnoDemoMI`. The metric is MCC, which ranges from -1 to 1. **Bold**: best performance.

Positive	Negative	Model	Mean	SD
With Empathy Contrast				
EmpDial	*r/CasualConv*	listener-only	0.131	0.018
		speaker-listener	0.121	0.047
r/Happy	*r/CasualConv*	listener-only	0.078	0.043
		speaker-listener	-0.030	0.023
r/OffMyChest	*r/CasualConv*	listener-only	0.120	0.009
		speaker-listener	**0.176**	0.031
Without Empathy Contrast				
EmpDial	*r/Happy*	listener-only	0.061	0.040
		speaker-listener	0.044	0.010
EmpDial	*r/OffMyChest*	listener-only	0.103	0.016
		speaker-listener	0.041	0.079
r/OffMyChest	*r/Happy*	listener-only	-0.017	0.035
		speaker-listener	0.035	0.017

no OOD models are reliable classifiers on `AnnoDemoMI`, which shows the wide domain gap between therapist empathy and empathy in non-counselling dialogues. The OOD models also have larger performance variation on `AnnoDemoMI` when trained on different randomly sampled training data. For example, the speaker-listener ⟨*r/OffMyChest* vs. *r/CasualConv*⟩ model has a standard deviation of 0.031 on `AnnoDemoMI`, compared to 0.011 on its OOD test data. This pattern is present in all OOD models, which shows their brittleness in empathy prediction for MI.

While the OOD models have low performance in general, the ones with empathy contrast do often outperform the ones without empathy contrast. Under the listener-only setup, the highest-scoring model with empathy contrast (⟨*EmpDial* vs. *r/CasualConv*⟩) leads the best-performing model without empathy contrast (⟨*EmpDial* vs. *r/OffMyChest*⟩) by 0.028, and the gap grows to 0.132 in the speaker-listener setting (between ⟨*r/OffMyChest* vs. *r/CasualConv*⟩ and ⟨*EmpDial* vs. *r/Happy*⟩). This shows that the benefit of learning from OOD empathy contrast is minor but does exist, and it is more obvious when more conversation context is available to the models (i.e., speaker-listener mode).

As for the choice between listener-only and speaker-listener models, the effects are mixed. Specifically, ⟨*EmpDial* vs. *r/CasualConv*⟩ and ⟨*r/Happy* vs.

Table 3.9: NLI performance overview. The metric is MCC, which ranges from -1 to 1. **Bold**: best performance.

Model Type	Model	MCC
Empathy-as-Hypothesis	therapist-only	**0.104**
	client-therapist	0.039
Empathy-as-Alignment	client→therapist	0.091
	therapist→client	-0.027

r/CasualConv⟩ both have better performance under the listener-only setting, while the reverse is true for ⟨*r/OffMyChest* vs. *r/CasualConv*⟩. In fact, as the speaker-listener ⟨*r/OffMyChest* vs. *r/CasualConv*⟩ model has higher performance than any other setup, it could be because a client talks more about negative experiences in a therapy session, which is to some extent similar to how a speaker shares emotional stories in *r/OffMyChest*. In contrast, speakers in *r/Happy* are more likely to recount positive experiences, which could explain why ⟨*r/Happy* vs. *r/CasualConv*⟩ has a relatively large performance drop ($\Delta > 0.1$) when including the speaker utterance (speaker-listener) compared to when it does not (listener-only).

NLI Models

As shown in Table 3.9, the NLI models overall have suboptimal performance (MCC ≤ 0.104). Notably, the NLI models are clearly outperformed by ⟨*EmpDial* vs. *r/CasualConv*⟩ ($\Delta \geq 0.027$) and ⟨*r/OffMyChest* vs. *r/CasualConv*⟩ ($\Delta \geq 0.072$), which further shows the benefit of OOD empathy contrast training.

On `AnnoDemoMI`, empathy-as-hypothesis scores 0.104 in the therapist-only setting but only 0.039 in client-therapist. Therefore, while knowledge gained from general NLI tasks is not sufficient for reasoning about complex concepts such as empathy, it is even more challenging to reason about empathetic interaction in client-therapist than to reason about the empathy of a single turn in therapist-only.

As for empathy-as-alignment, client→therapist clearly outperforms ($\Delta > 0.1$) therapist→client. We postulate that this is because the client→therapist model better captures cases where the therapist shows empathy by acknowledging the client, as the NLI entailment score is likely to be higher when the hypothesis (therapist utterance) acknowledges the premise (client utterance).

3.5 Clinical Application & Impact

The motivation for this zero-shot work was to minimise the annotation effort needed for effective utterance-level prediction of therapist empathy/non-empathy in MI.

Overall, our results show that the supervised method with OOD empathy contrast and the unsupervised NLI-based approach do not produce sufficiently accurate predictions, which limits their application in clinical settings. Compared to supervised learning of session-level empathy on sizeable corpora of well-annotated MI conversations (e.g., [31]), the task of utterance-level empathy prediction without in-domain fine-tuning is inherently more challenging, and accordingly our models have lower performance. As discussed, 1) the domain gap between therapist empathy and empathy in non-counselling dialogues is the main cause for the suboptimal performance of OOD empathy contrast models; 2) therapist empathy prediction is likely not amenable to NLI modelling as a proxy task.

We also note that several data quality issues are present in this study, some of which are alluded to in §3.2. First, the heuristic empathy labels of *RedditConvs* and *EmpDial* may not always be accurate at the utterance level. Also, `AnnoDemoMI` is a small-scale benchmark dataset and contains minor transcription noise. Furthermore, while the human annotators of `AnnoDemoMI` are familiar with the topic of therapist empathy and show substantial IAA (§3.2.2), they are not counselling professionals and thus may not always assign empathy labels the way an experienced therapist would. Overall, these data quality issues highlight the need for a clean and larger-scale MI dialogue dataset with fine-grained annotations given by experts, which we present in Chapter 4.

Nevertheless, we believe that this work is a meaningful step towards low-resource real-time assessment of empathy in counselling, and that the idea of utilising pre-trained LMs for low-resource scenarios related to clinical psychology is still relevant. With smoothed domain gaps and more fine-grained annotations (e.g., true utterance labels instead of corpus-level labels used heuristically as utterance labels), future work can still use pre-trained LMs to leverage parallels between therapist empathy and empathy in non-counselling dialogues. For example, knowledge of empathy contrast learned from well-annotated non-counselling conversations can serve as a pre-training step for empathy contrast training on a minimal amount of well-annotated therapy dialogues. The motivation is that a small to modest amount of therapy dialogue data is sometimes available for a specialised domain like MI, and therefore OOD empathy knowledge can be leveraged as a good starting point for in-domain fine-tuning and thus maximise the benefit of OOD empathy training.

3.6 Summary

In this work, we made the first attempt at zero-shot binary prediction of utterance-level therapist empathy for MI. We proposed 1) a supervised method that trains BERT on heuristically constructed empathy vs. non-empathy contrast in non-counselling conversations; and 2) an unsupervised method that formulates NLI as a proxy task for therapist empathy prediction. Our results showed that those zero-shot approaches are not sufficiently accurate, but we found that the empathy vs.

non-empathy contrast enables the best performance. Our analysis showed that the benefit of this contrast is clear when it is compared to control-group supervised OOD models without empathy contrast and the unsupervised approach.

Future work may investigate higher-quality and more fine-grained empathy annotations, for example at the sentence level, where we expect less noise and more accurate predictions. Another direction worth exploring is few-shot methods for therapist empathy prediction with OOD empathy contrast training as a pre-training step.

Ethics & Privacy

Empathy often involves deeply personal circumstances such as distress and struggles. Therefore, DL studies in this area warrant ethical consideration. The greatest ethical risk of this work is privacy implications, as the dialogue data we used could contain large amounts of sensitive identifiable information. To mitigate this risk, we worked exclusively with de-identified data where mentions of information like name, date, and location were replaced with placeholders. We argue that this study has considerable benefit as the first investigation of using knowledge of empathy from non-counselling dialogues to support low-resource computational analysis of therapist empathy, and the findings can inspire future efforts in this important research direction.

Chapter 4

Creation of Counselling Dialogue Dataset

This chapter is based on:

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-mi: A dataset of expert-annotated counselling dialogues. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6177–6181. IEEE, 2022. (©2022 IEEE. Reprinted, with permission.)

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3), 2023

Research on NLP for MI has seen substantial development in recent years, but access to this area remains extremely limited due to the lack of publicly available well-annotated MI conversations. In this chapter, we introduce `AnnoMI`, the first publicly and freely accessible dataset of professionally transcribed and expert-annotated MI dialogues. It consists of 133 conversations that demonstrate high- and low-quality MI, with rich annotations by domain experts covering key MI attributes. We detail the data collection process including dialogue selection, transcription and annotation. Based on the expert annotations on key MI aspects, we carry out thorough analyses of `AnnoMI` with respect to counselling-related properties at the levels of utterance, dialogue and corpus. We also discuss potential applications of this dataset.

4.1 Introduction

Recent years have seen significant interest in the research of linguistic and statistical MI analysis (§2.2.2, §2.2.3). The first computational model for identifying reflection, a key skill in MI, was introduced by Can et al. [104]. More broadly, the modelling of MI-related aspects such as therapist empathy and utterance-level therapist/client behaviour codes [16, 2] has been approached with methods based on classical ML [29, 105, 30] (e.g., support vector machines) and DL [31, 33, 34, 13] (e.g., RNNs).

Despite the progress, NLP for MI is still an extremely low-resource domain (§2.2.1), owing to privacy-related restrictions. As research in this field has been conducted primarily on private/undisclosed annotated MI dialogues, it has been challenging to replicate and further develop previous work. Prior to the publication of this work, to the best of our knowledge, the only publicly available dataset of MI conversations was created by Pérez-Rosas et al. [1] through automatic captioning of YouTube/Vimeo videos that demonstrated high- and low-quality MI. However, its transcript quality is compromised by the considerable transcription errors from automatic captioning that can make the transcripts difficult to understand (§4.2.2). In the same study, the authors also analysed two MI therapist behaviour codes — reflection and question — based on the dataset annotations from trained students, but those annotations are unavailable at the time of writing.

To address the lack of publicly available expert-annotated MI dialogues and improve access to MI-related NLP research, we present `AnnoMI`, a dataset[1] of 133 high- and low-quality MI conversations that were

- Professionally transcribed from MI demonstrations on video-sharing platforms;
- Obtained through explicit consent from the video owners that permits dataset creation, release to the public, and use for research purposes; and
- Annotated by experienced MI practitioners based on a scheme covering key MI aspects.

We describe our MI video acquisition, transcription, annotator recruitment and annotation scheme in §4.2. We show the results of IAA in §4.3. We present thorough analyses of `AnnoMI` in §4.4 and discussions over the creation and application of `AnnoMI` in §4.5, before this chapter is concluded in §4.6.

4.2 Creating `AnnoMI`

Considering the scarcity/absence of publicly available conversation datasets of real-life MI and their privacy-related legal and ethical restrictions, we rely instead on

[1] Available at `https://github.com/uccollab/annomi` under the Public Domain license.

Table 4.1: AnnoMI overview. (©2022 IEEE)

	High-Quality MI	Low-Quality MI
Number of Dialogues	110 (82.7%)	23 (17.3%)
Number of Utterances	8839 (91.1%)	860 (8.9%)

demonstrations of MI-adherent and non-adherent counselling from online video-sharing platforms, in a similar vein to [1]. With explicit consent from the video owners, we obtain professional transcripts of the demonstrations and recruit MI experts to annotate the transcripts following a scheme covering key MI elements[2].

4.2.1 MI Demonstration Videos

As a trade-off between counselling authenticity and privacy preservation, we leverage MI demonstrations on video-sharing websites (YouTube and Vimeo). We identified 346 videos that demonstrate high- and low-quality MI, using key phrases including "good motivational interviewing" and "bad MI counselling". According to the literature [10], high-quality MI is centred on the client and conducted with empathy, whereas low-quality MI is characterised by frequent instructions and suggestions.

We label each video to be high-quality MI or low-quality using its title (e.g., "Motivational Interviewing - Good example"/"The Ineffective Physician: Non-Motivational Approach") as well as descriptions and narrator comments (e.g., "Demonstration of the motivational interviewing approach in a brief medical encounter"). We consider such labelling to be verified automatically, as the video uploaders are professional MI practitioners and organisations focused on healthcare and behaviour change. We also note that the definition of high- and low-quality MI is clear in the literature ([10, 11], *inter alia*), therefore the high/low MI quality divide is consistent across different institutions/therapists and different demonstrations.

We gained explicit permission from the content owners for us to use their videos to create, analyse and publicly release a transcript-based MI dialogue dataset. Where applicable, we also gained consent of the individuals appearing in those videos. We eventually obtained permission to use 119[3] of those videos, which contained 133 complete conversations (a video may contain multiple dialogues). 110 of the dialogues showcase high-quality MI and the other 23 low-quality MI (Table 4.1). As shown in Figure 4.1, high-quality MI dialogues are generally longer than low-quality ones, with several surpassing 200 utterances in length, but most dialogues have less than 100 utterances. A pair of high- and low-quality MI session excerpts, both about smoking cessation/reduction, are presented in Table 4.2.

[2] Prior to our experiments, the materials and methodology of our study underwent ethical review by our institution's Ethics Board, and the study was subsequently approved. Informed consent was obtained from all subjects involved in the study.

[3] Overlap with [1]: 42 videos.

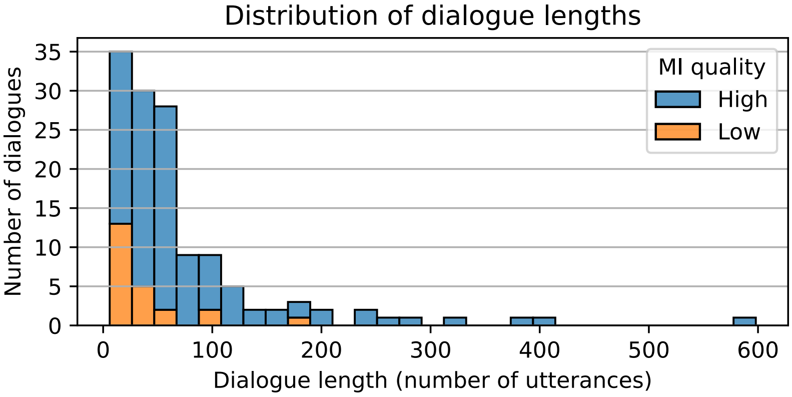


Figure 4.1: Distribution of dialogue lengths (number of utterances per dialogue).

Table 4.2: High- and low-quality MI conversation snippets, where the goal is smoking cessation/reduction. ***T***: therapist; ***C***: client.

High-Quality MI
T: Um, I did wanna talk to you though. I'm a little bit concerned looking through his chart of how many ear infections he's had recently. And I-I noticed that you had checked the box that someone's smoking in the home. So I was wondering if you can tell me a little more about that. ***C***: Well, um, It's just me and him and I do smoke. Um, I try really hard not to smoke around him, but I-I've been smoking for 10 years except when I was pregnant with him. But it– everything is so stressful being a single mom and-and my having a full-time job. And so it's just– that's why I started smoking again. ***T***: You have a lot of things going on and smoking's kind of a way to relax and destress. ***C***: Yeah. Some people have a glass of wine. I have a cigarette. ***T***: Sure. And it sounds like you're trying not to smoke around him. Why did you make that decision?
Low-Quality MI
T: Well, now's the time to quit. It's really gotten to the point where you can't keep smoking. Not only for him, like I said, but also for you. You're putting yourself at risk for lung cancer, for emphysema, for oral cancers, for heart disease, for all kinds of things- ***C***: I know, I know. I've heard– People have told me before, I've heard all that. I just don't know how to do it. How am I supposed to quit? It's-it's so hard. ***T***: Well, there's all kinds of things you can use now. It's not as hard as it used to be. You can use nicotine replacement. There's patches, there's lozenges, there's gum, there's the inhaler, there's nasal spray. We can talk about medications. You can try Chantix, you can try Zyban, there's quit smoking groups you can go to, there's hotlines you can call. ***C***: I just don't have time for any of that.

The imbalance w.r.t. high- and low-quality MI dialogue volumes can be attributed to A) fewer low-quality MI video owners responded to our request or consented; B) low-quality MI videos are relatively scarce on Youtube/Vimeo, possibly because MI-adherence demonstrations are deemed more valuable as "good examples" and thus filmed and uploaded more.

4.2.2 Transcription

Using a professional transcription service[4], we collected fluent and accurately transcribed MI conversations from the videos, whereas the transcripts of [1] were produced by automatic captioning. While a step of verifying video content-caption matching is reported in [1], in practice we find considerable incorrectly transcribed words/phrases and mismatched speaker (therapist/client) labels in the corpus of [1] that can significantly hinder text understanding. Table 4.3 presents transcript snippets from [1] and `AnnoMI` of the same video to exemplify the marked difference in transcription quality between the two datasets. `AnnoMI` is also free from other noises such as narrations but retains context-relevant details, including “hmm”, “right” and speaker sentiment/emotion [162, 163, 164] indicators such as “[laugh]”.

4.2.3 Expert Annotators & Workload Assignment

Since MI annotation requires specialised knowledge, we rely on experienced MI practitioners to annotate the transcripts. Specifically, we recruited 10 therapists through professional networks, in particular the Motivational Interviewing Network of Trainers[5], an international organisation of MI trainers and a widely recognised authority in MI. The annotators had high proficiency in English and prior experience in practising/coding MI. We also collected informed consent from all the annotators.

Overall, each expert annotated 19 to 20 transcripts with total lengths around 144 minutes in terms of the total duration of the original 19 to 20 videos. To facilitate computation of IAA, we selected 7 common transcripts to be annotated by all experts, based on 3 criteria: 1) they should add up to about 1/3 (45 minutes) of the workload of each annotator; 2) they should cover diverse topics (6 out of the 7 transcripts have distinct topics); 3) they should cover both high- and low-quality demonstrations (5 showing high-quality MI and 2 showing low-quality). We tried various combination of transcripts before we found one combination that satisfied the criteria above. During the annotation process, no expert was aware that a part of their workload would be used to compute the IAA. Each of the 126 (133-7) non-IAA transcripts was annotated by one expert due to our budget limit.

We note that the IAA results of `AnnoMI` are not directly comparable with those of other annotated MI corpora, since the former are calculated based on the annotations from 10 experts while the latter often come from much fewer (e.g., 2 or 3) annotators, and it is usually less likely to reach the same or higher level of IAA with more annotators. This also means that the attributes of `AnnoMI` that do have good IAAs are indeed reliably annotated.

[4] `https://gotranscript.com/`

[5] `https://motivationalinterviewing.org/trainer-listing`

Table 4.3: Transcription quality comparison between `AnnoMI` and [1]. Red: incorrectly transcribed word; Blue: omitted words/phrases; Orange: words from the other interlocutor that should have started a new utterance; ~~Strikethrough~~: incorrectly transcribed word within such a misplaced utterance; {C}/{T}: missing client/therapist utterance.

`AnnoMI`
C: Right. Well, it would be good if I knew, you know, that my kids are taken care of too- ***T***: Yeah. ***C***: - so I'm not worried about them while I'm at work. ***T***: Right. Yeah. Because you're- you're the kind of parent that wants to make sure your kids are doing well. ***C***: Right. ***T***: Yeah. Um, so tell me, what would it take to get you to like a five in confidence, to feel a little bit more confident about getting work? ***C***: Well, I mean, being able to make the interviews would be the priority. ***T***: Okay, Yeah. ***C***: Um, so chi- you know, taking care, having some childcare, having- ***T***: Mm-hmm. ***C***: - having someone I trust that I can call when I know I've got an interview. ***T***: Yeah. Because you definitely need to go to an interview in order to get the job. ***C***: Right. Yeah. ***T***: So having taken care of that part, having some reliable childcare would definitely help. ***C***: Yeah.
[1]
C: one it would be good if I knew you know that my kids are taking care of ("too") - yeah so I'm not worried about them law in the work right yeah ***T***: because you're you're the kind of parent that wants to make sure your kids are doing well great ({C}) yeah um so tell me what would it take to get you to like a five in confidence to feel a little bit more confident ("about") getting work ***C***: well I mean being able to make the interviews would be the priority again ({T}) um so try you know taking care having some child care I mean having ({T}) someone I trust that I can call when I you know what that interview because you definitely need to go to an interview ~~of~~ in order to get ~~three~~ ("the job") ***T***: ~~yeah~~ yeah so having taken care of that part having some reliable child care ("would definitely help") ***C***: yeah definitely not

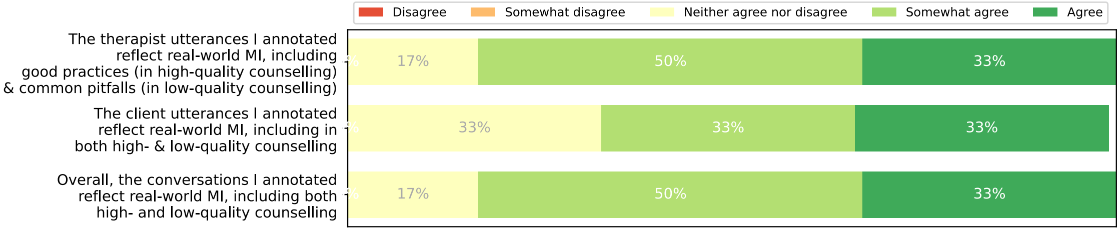


Figure 4.2: Annotator survey on whether AnnoMI reflects real-world counselling.

4.2.4 AnnoMI & "Real-World" MI

For AnnoMI to be useful for real-world applications, it is crucial that its dialogues reflect both high- and low-quality MI in the real world. Therefore, we surveyed the 10 annotators after they completed their tasks, asking them whether they felt the AnnoMI dialogues resembled real-world MI, and we eventually received responses from 6 annotators. As shown in Figure 4.2, 83% of the responses "agree" or "somewhat agree" that the therapist utterances and the dialogues overall reflect real-world MI, and the figure is 66% for the client utterances. The clear majority in each case shows that AnnoMI indeed sufficiently captures the characteristics of real-world MI, even though the dialogue sources are demonstrations.

We note that researchers in the field of NLP for counselling are faced with a very challenging legal and regulatory landscape, due to privacy-related concerns and rules in different jurisdictions. Therefore, a dataset like ours can be used significantly more broadly, since it does not have any privacy implications or legal issues concerning different jurisdictions.

Table 4.4: Top 10-topics in AnnoMI in terms of (1) number/percentage of dialogues and (2) total number/percentage of utterances.

Topic	#Dialogues	Topic	#Utterances
Reducing alcohol consumption	28 (21.1%)	Reducing alcohol consumption	1914 (19.7%)
Smoking cessation	21 (15.8%)	Reducing recidivism	1303 (13.4%)
Weight loss	9 (6.8%)	Smoking cessation	1106 (11.4%)
Taking medicine	9 (6.8%)	Diabetes management	709 (7.3%)
More exercise	9 (6.8%)	Reducing drug use	578 (6.0%)
Reducing drug use	8 (6.0%)	Taking medicine	574 (5.9%)
Reducing recidivism	7 (5.3%)	More exercise	525 (5.4%)
Compliance with rules	5 (3.8%)	Weight loss	396 (4.1%)
Asthma management	5 (3.8%)	Avoiding DUI	394 (4.1%)
Diabetes management	5 (3.8%)	Changing approach to disease	315 (3.2%)
Other	33 (24.8%)	Other	2107 (21.7%)

4.2.5 Annotation Scheme

We design a detailed annotation scheme to study therapist and client behaviours, based on the MI literature, existing coding protocols (MISC/MITI), and feedback from therapists. At the conversation level, we ask the annotators to briefly describe the dialogue goal, e.g., “smoking cessation”. Thus, we summarise in Table 4.4 the top-10 topics in terms of A) the number of conversations that have those topics, and B) the total number of utterances in those conversations. Table 4.5 shows the utterance-level annotation scheme, which contains 4 therapist utterance attributes and 1 client utterance attribute. When annotating an utterance, an annotator could also see the preceding and subsequent utterances for more context.

Therapist Utterance Attribute 1: (Main) Behaviour

In MI, three fundamental yet crucial skills to achieve effective counselling are: Asking, Informing and Listening [11]. In view of this principle and related components of mainstream coding schemes for MI, we consider **Question**, **Input** and **Reflection** as major therapist behaviours that correspond to Asking, Informing and Listening, respectively. In cases where more than one behaviour is present in an utterance, e.g., a question after an input, the expert is asked to further select the **Main Behaviour**. **Other** is listed as the fourth and default option where no Question, Input, or Reflection appears in the utterance.

We also list **Question**, **Input** and **Reflection** as separate attributes of therapist utterances — detailed as “Therapist Utterance Attribute 2/3/4” in the rest of this sub-section — in order to investigate their sub-types.

Table 4.5: Utterance-level multi-choice annotation scheme. [+] implies presence of utterance attribute (e.g., "Simple reflection" label entails that **Reflection** exists in utterance), while [−] indicates absence thereof (e.g., "No reflection" label implies **Reflection** is not present in utterance).

Therapist Utterance Attributes	**Label**
(Main) Behaviour	**Question** **Input** **Reflection** **Other**
Question	Open question [+] Closed question [+] No question [−]
Input	Information [+] Advice [+] Options [+] Negotiation/Goal-Setting [+] No input [−]
Reflection	Simple reflection [+] Complex reflection [+] No reflection [−]
Client Utterance Attribute	**Label**
Talk Type	Change Neutral Sustain

We note that this work is more focused on the use of Asking, Informing and Listening in the AnnoMI dialogues[6], therefore it does not seek to compare directly with previous work that uses the complete MISC/MITI for annotation.

Therapist Utterance Attribute 2: Question

Therapists use Asking to develop an understanding of the client and their problems. Therefore, we include **Question** as a therapist behaviour and define any question as *open* or *closed* in accordance with mainstream MI coding conventions. An open question allows a wide range of possible answers and may seek information, invite the client's perspective or encourage self-exploration, while a closed question implies

[6] For the same reason, the original annotation scheme was more ambitious and had several non-MITI/MISC annotation fields, but they are not included in this paper due to their very low IAAs (Fleiss' kappa), and thus the annotation scheme presented in this section may look like a subset/regrouping of MISC to some readers.

Table 4.6: Example labelling for therapist **Question** from the dataset. Detailed Open/Closed question types (e.g., "Number") are illustrative only and not annotated.

Utterance	**Question Type**
Do you have children in your house?	Closed (Yes/No answer)
How much does it actually cost you a week?	Closed (Number)
Okay. What kind of alcohol do you drink at parties?	Closed (Specific fact)
So what is a typical week for you as far as your alcohol use is concerned?	Open (Seek information)
Okay. So how do you feel about being here today?	Open (Invite client's perspective)
So, when you think about what you like and don't like about your drinking, where do you wanna go from here?	Open (Encourage self-exploration)

a short answer such as Yes/No, a specific fact, a number, etc. [16]. Some examples are given in Table 4.6.

Therapist Utterance Attribute 3: Input

Informing is the primary manner of communicating knowledge and recommendations/advice to the client. Inspired by coding protocols (e.g., [16]) and insights from a professional therapist regarding the patterns of Informing in the AnnoMI transcripts, we use the term **Input** to include a wide range of conveyed knowledge and consider 4 subtypes: providing information, giving advice, presenting options and setting goals (negotiation). Some examples are given in Table 4.7. When an utterance contains more than one type of Input, the annotators choose the "main" type of Input to make the labels mutually exclusive and facilitate utterance-level NLP applications.

Therapist Utterance Attribute 4: Reflection

Reflection is an essential way of Listening. In using reflections, the therapist shows that they are listening to and understanding the client, which is effective in helping people change. Following MISC, we consider two reflection types: *simple* & *complex*. A simple reflection shows an understanding of what the client said explicitly, for example through rephrasing, but does not go much further. In comparison, a complex reflection conveys a deeper interpretation/exploration of the client's viewpoint and experience, using techniques such as metaphors and exaggeration [16]. Two illustrative pairs of contrasting simple and complex reflections to the same client statement are presented in Table 4.8.

Table 4.7: Example Labelling for therapist **Input** from the dataset.

Utterance	Input Type
You're not alone in feeling that way. Binge drinking can feel normal to some people.	Information
So that's a hormone that allows you to utilise sugar in your body.	Information
I want you to be healthy. And I don't want to see you coming back in here for something else. So I'm really gonna recommend that you try to cut down to that amount.	Advice
That's why I recommend that all my adolescent patients not drink at all.	Advice
So, what have you looked into about, um, you know, advocacy in that area or expungement or anything like that?	Options
Okay. So, exploring some yoga classes. Is doing yoga in your living room appealing to you at all?	Options
So for you being in your class, when that bell rings, then you know, this is the goal.	Goal-setting
Do you think you could go two months without drinking?	Goal-setting

Table 4.8: Example labelling for therapist **Reflection** from [2].

Scenario 1

Speaker	Utterance	Reflection Type
Client	This is her third speeding ticket in three months. Our insurance is going to go through the roof. I could just kill her. Can't she see we need that money for other things?	
Therapist 1	You're furious about this.	Simple
Therapist 2	This is the last straw for you.	Complex

Scenario 2

Speaker	Utterance	Reflection Type
Client	My mother is driving me crazy. She says she wants to remain independent, but she calls me 4 times a day with trivial questions. Then she gets mad when I give her advice.	
Therapist 1	Things are very stressful with your mother.	Simple
Therapist 2	You're having a hard time figuring out what your mother really wants.	Complex

Client Utterance Attribute: Talk Type

According to the MI literature [11], clients usually feel ambivalent about adopting positive behaviour change, and thus the desirable outcome of MI is for the client to pick up pro-change arguments and talk themselves into changing, provided that it aligns with their aspirations and values. This type of talks that favour change are known as "change talks". Conversely, a "sustain talk" conveys resistance to behaviour change and favours the status quo, so it is also desirable in MI to reduce sustain talks [165]. On the other hand, a "neutral talk" indicates no preference for or against change. Hence, we name **Change Talk**, **Sustain Talk**, and **Neutral Talk** as the three types of the client **Talk Type** attribute. Table 4.9 presents some examples of those talk types in different scenarios such as reducing alcohol consumption.

Table 4.9: Example labelling for client **Talk Type** from the dataset.

Utterance	**Talk Type**
Yeah, I just want to do what's right.	Change
Well, that was fine until I came here, um, but now that I know about the health risk, um, I have something I gotta think about.	Change
Um, I mean, the 10 drinks seems like not a lot for me and my tolerance.	Sustain
Yeah, whatever. I know you got to do your job, but I don't care.	Sustain
Yeah, I would like to play soccer in college.	Neutral
And um, I think she used to look after me because she used to do the cooking and stuff like that.	Neutral

4.3 IAA Results

4.3.1 Default Measure: Fleiss' Kappa at Utterance-Level

We use Fleiss' kappa [155] as the default measure for calculating utterance-level IAA over the annotations on the 7 common transcripts. We consider 3 ways of calculation: ALL, ALL(STRICT), and BINARY. ALL applies to all the utterance attributes, while ALL(STRICT) and BINARY apply to Input, Reflection and Question only.

Specifically, since Input, Reflection and Question have a default "abscence" option (i.e., No input, No reflection and No question, as shown in Table 4.5), we compute a two-class presence-vs.-absence (i.e., BINARY) IAA for them in addition to the fine-grained all-class IAA (i.e., ALL). For example, when computing ALL-IAA for Question, we consider the original label space: {Open question [+], Closed question [+], No question [−]}, where [+] means there is a question in the utterance and [−] means there is not. Conversely, we only consider the presence-vs.-absence {[+], [−]} space when calculating BINARY-IAA.

We also calculate ALL(STRICT)-IAA, which computes IAA within the original label space but on a more challenging subset of utterances, motivated by the ob-

Table 4.10: IAA on utterance-level annotations, in Fleiss' kappa. Orange, blue, cyan and green indicate fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80) and almost perfect (0.80-1.00) agreement, respectively.

Therapist Utterance Attribute	**IAA Setting**	**IAA**
Input	All(Strict)	0.34
	All	0.51
	Binary	0.64
Reflection	All(Strict)	0.32
	All	0.50
	Binary	0.66
Question	All(Strict)	0.54
	All	0.74
	Binary	0.87
(Main) Behaviour	All	0.74
Client Utterance Attribute	**IAA Setting**	**IAA**
Talk type	All	0.47

servation that it is substantially more difficult to distinguish between the presence [+] labels than between presence [+] and absence [−]. For example, differentiating between "Simple reflection [+]" and "Complex reflection [+]" is harder than between Reflection and No-Reflection. Therefore, we compute All(Strict) on the utterances where at least one annotator chose a presence [+] option. For example, for Reflection, we calculate All(Strict)-IAA on the utterances where at least one annotator selected "Simple reflection [+]" or "Complex reflection [+]".

4.3.2 Results of Default IAA Measure

All Fleiss'-kappa-based IAAs are listed in Table 4.10. Following [166], we group the IAAs into slight (0.01-0.20), fair (0.21-0.40), moderate (0.41-0.60), substantial (0.61-0.80) and almost perfect (0.80-1.00) agreement. We consider an attribute **predictable** if its IAA shows moderate or better agreement.

We notice that for the utterance attributes where Binary and All(Strict) are applicable, the order of agreements is, without exception, All(Strict)-IAA < All-IAA < Binary-IAA, which proves the challenge of the subset used for computing All(Strict)-IAA as well as the ease of annotating the absence/presence of a particular utterance attribute.

The annotators show fair agreement on Input and Reflection under All(Strict), which reveals the difficulty of annotating those attributes despite their inclusion in MISC/MITI, particularly when their presence in an utterance cannot be easily ruled out. Nevertheless, the IAA jumps to substantial agreement

Table 4.11: ICC as IAA.

(Main) Therapist Behaviour	ICC
Input	0.975
Reflection	0.991
Question	0.997
Other	0.996
Client Talk Type	ICC
Change	0.916
Neutral	0.986
Sustain	0.890

for Input and Reflection under the BINARY setting, which suggests the presence of distinguishable linguistic features unique to those two attributes.

Encouragingly, Question, (Main) Behaviour and Talk Type all record moderate or better IAAs under all settings, which shows the text-based predictability and therefore the existence of distinct linguistic features of those attributes.

4.3.3 Supplementary IAA Measure: Intraclass Correlation

Following MITI, we also use Intraclass Correlation (ICC) to analyse (Main) Behaviour and Talk Type at the label level to gain more insights and facilitate comparison with other studies. Specifically, ICC describes how much of the total variation in the label counts is due to differences among annotators. For each label, we count how many times the label is used by each annotator for the utterances of each session. Thus, each of the 10 annotators has 7 label counts corresponding to the 7 IAA transcripts. Also following MITI, we compute the ICC scores using a two-way mixed model with absolute agreement and average measures [2].

As Table 4.11 presents, all the (Main) Behaviour and Talk Type labels have excellent (0.75-1) [167] agreement scores, which shows the reliability of these annotations. Nevertheless, Change Talk and Sustain Talk have slightly lower ICCs — around 0.9 — compared to the other ICCs that are almost 1.0, which echoes the lower Fleiss'-kappa-based IAA of Talk Type compared to that of (Main) Behaviour.

4.3.4 Dataset Release

We release `AnnoMI` in full, which has the following attributes:

- **Question**: {Open question, Closed question, No question}
- **Input**: {Information, Advice, Options, Negotiation/Goal-Setting, No input}
- **Reflection**: {Simple reflection, Complex reflection, No reflection}

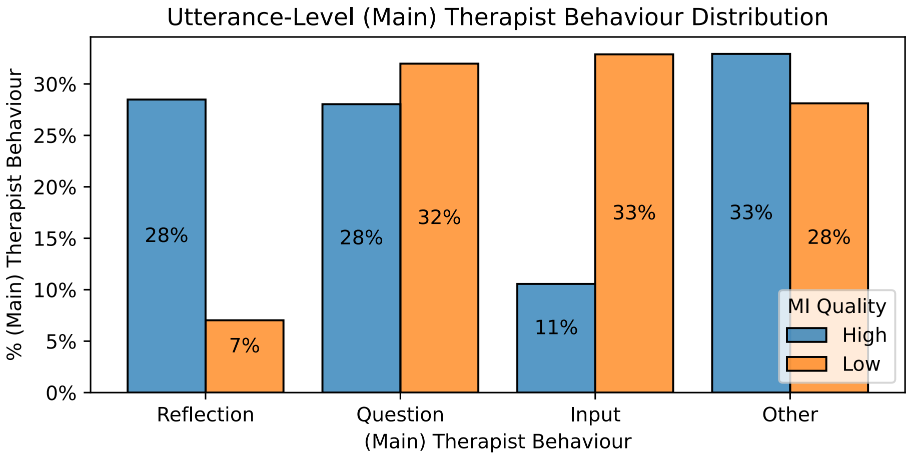


Figure 4.3: (Main) Behaviour distributions in high- & low-quality MI (©2022 IEEE)

- **(Main) Behaviour**: {Question, Input, Reflection, Other}
- **Talk Type**: {Change, Neutral, Sustain}

For the 7 IAA transcripts that are annotated multiple times, we release the annotations from each expert.

4.4 Dataset Analysis

We analyse the annotations[7] via visualisations. Unless otherwise specified, (Main) Behaviour represents the behaviour of an utterance. For example, if a therapist utterance consists of a reflection and a question but Reflection is annotated as the main behaviour, we consider the utterance to be a reflection instead of a question, in order to facilitate further analysis.

We also note that while there are clear correlations between utterance attribute distribution and MI quality in some cases, they do not necessarily point to causation, especially given the relatively low amount of data and potential sampling bias.

4.4.1 Overall (Main) Behaviour & Talk Type Distributions

As Figure 4.3 demonstrates, the most marked contrast between therapist behaviours in MI-adherent and non-adherent counselling lies in the proportions of Reflection and Input. The average MI-adherent therapist employs Reflection in 28% of their utterances whereas it is only 7% in non-adherent counselling, echoing the MI requirement of trying to understand the client's perspective and communicating it. On the other hand, Input is given 33% of the time in low-quality MI but only 11%

[7] For the IAA transcripts, we use majority-voted utterance labels during dataset analysis.

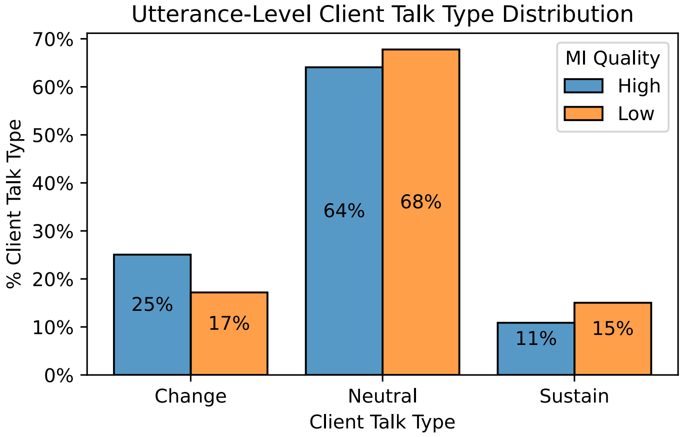


Figure 4.4: Talk Type distributions in high- & low-quality MI (©2022 IEEE)

in high-quality MI, which, together with the statistics of Reflection, conforms to the observation [11] that high-quality MI emphasises understanding the client as opposed to speaking from the therapist's own point of view. The relationship between MI quality and the share of Question and Other is relatively weak.

As for Talk Type (Figure 4.4), Change Talk is more frequent in high-quality MI — 25% vs. 17%, whereas Sustain Talk has a stronger presence in low-quality MI — 11% vs. 15%. Nevertheless, those contrasts are less obvious than those found in Reflection and Input. Possible explanations include A) some clients in low-quality MI could adopt tepid change-talk-like speech such as "Yeah, maybe" only to end the counselling quickly; and B) some clients in high-quality MI are simply more reluctant to change but the therapist still respects that, as is recommended in MI. On the other hand, most (64%-68%) client utterances belong to the Neutral Talk category regardless of MI quality, to which we empirically find the prevalence of short utterances like "Mhmm" and "Uh huh" to be a major contributing factor.

4.4.2 Posterior (Main) Behaviour & Talk Type Distributions

MI guidelines have specific recommendations on how a therapist should respond when the client talks in certain ways, and a client may also react to the therapist in particular patterns. We therefore probe the posterior distributions of next-turn therapist behaviours(/client talk types) given the current-turn client talk type(/therapist behaviour). Denoting u_t^T as the therapist utterance at turn (time step) t and u_{t+1}^C as the client reply in the following turn, the posterior distribution of client talk types can be represented as $p(\mathit{Talk\_Type}(u_{t+1}^C) \mid \mathit{Behaviour}(u_t^T))$. Similarly, the posterior distribution of therapist behaviours can be formulated as $p(\mathit{Behaviour}(u_{t+1}^T) \mid \mathit{Talk\_Type}(u_t^C))$

Figure 4.5 presents the posterior distribution of client talk types, i.e., $p(\mathit{Talk\_Type}(u_{t+1}^C) \mid \mathit{Behaviour}(u_t^T))$. While Neutral Talk is clearly the ma-

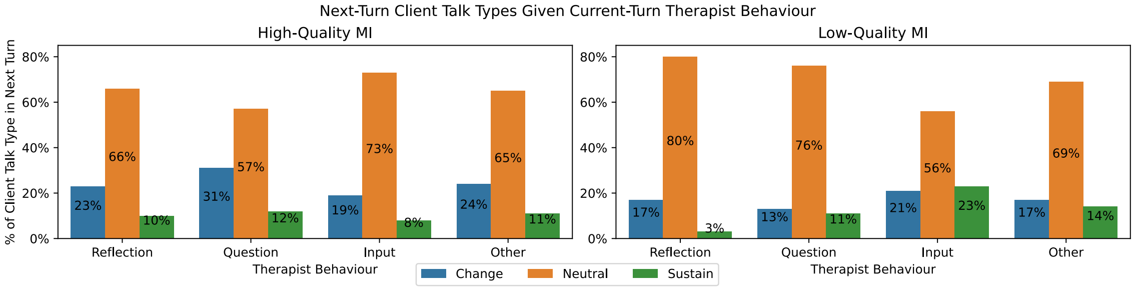


Figure 4.5: Distribution of next-turn client talk types (Y-axis) given different therapist behaviours in the current turn (X-axis).

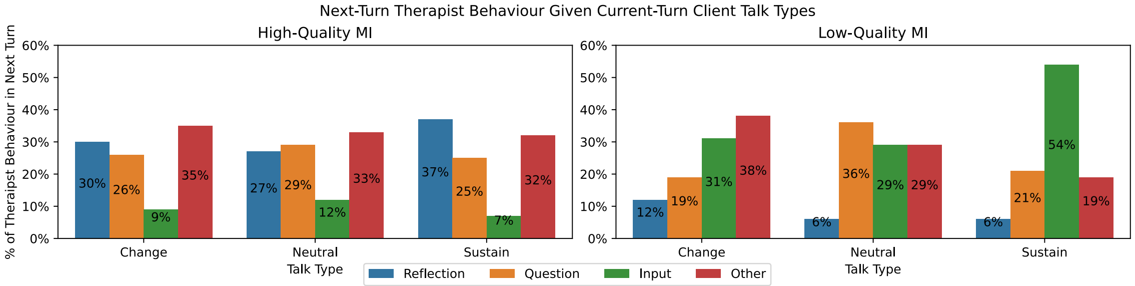


Figure 4.6: Distribution of next-turn therapist behaviours (Y-axis) given different client talk types in the current turn (X-axis).

jority talk type of the client response, in most cases $p(\mathit{Talk\_Type}(u^C_{t+1}) = \text{Change} \mid \mathit{Behaviour}(u^T_t))$ is substantially larger in high-quality MI than in low-quality MI regardless of $\mathit{Behaviour}(u^T_t)$, which shows that an MI-adherent therapist is more likely to evoke Change Talk from the client, irrespective of specific therapist behaviours. On a more granular level, Question is the most likely (31%) therapist behaviour in high-quality MI to evoke Change Talk, which may be because some therapist questions lead to Change Talk more often, such as asking the client what steps they could take towards a behaviour change or how confident they are about adopting a change. Interestingly, Input triggers more Change Talk (21%) than any other therapist behaviour in low-quality MI, but it is also the therapist behaviour that prompts the most (23%) Sustain Talk, which suggests that the effect of frequent input — characteristic of low-quality MI (Figure 4.3) — is far from certain in terms of evoking change talk and reducing sustain talk.

Figure 4.6 shows the posterior distribution of therapist behaviours, i.e., $p(\mathit{Behaviour}(u^T_{t+1}) \mid \mathit{Talk\_Type}(u^C_t))$. One can observe that MI-adherent therapists in general use considerably more reflections than non-adherent therapists do — 30% vs. 12% — in response to Change Talk, which confirms that high-quality MI utilises Reflection to reinforce willingness to change. On the other hand, the most commonly shown therapist behaviour in response to Sustain Talk in high-quality MI is Reflection (37%), while the dominant pattern of reacting to Sustain Talk in low-quality

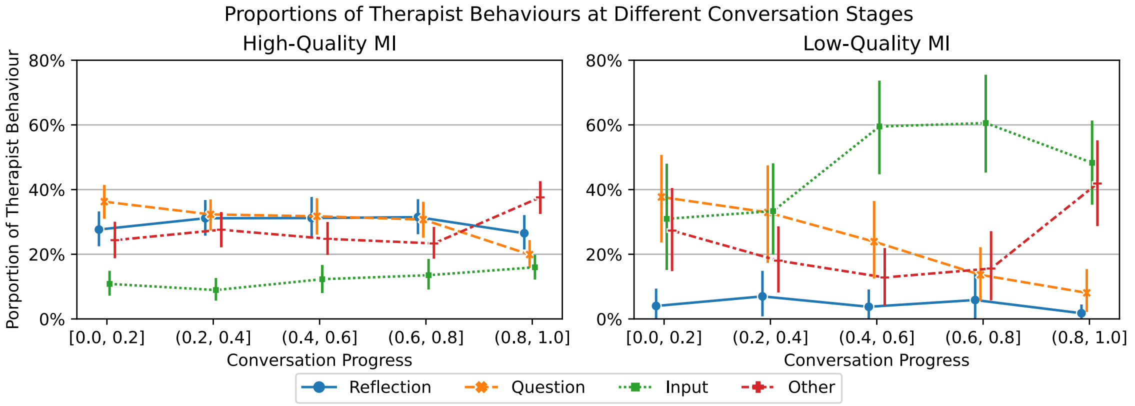


Figure 4.7: Proportions of therapist behaviours in different conversation stages in high- and low-quality MI.

MI is Input (54%). This contrast serves as a strong evidence that MI-adherent counselling focuses more on showing empathy and trying to understand the client through Reflection when faced with resistance, whereas a non-adherent therapist is more likely to try to challenge, correct or persuade the client through more Input — a common mistake in MI non-adherent counselling [11].

4.4.3 (Main) Behaviour & Talk Type as Conversation Proceeds

Following [1], we divide each conversation into 5 parts: $[0.0, 0.2]$, $(0.2, 0.4]$, $(0.4, 0.6]$, $(0.6, 0.8]$ and $(0.8, 1.0]$, in order to probe conversational properties at different dialogue stages. Specifically, we examine the distributions of different therapist behaviours and client talk types at those stages.

As shown by Figure 4.7[8], in both high- and low-quality MI, the proportion of Question gradually decreases while the conversation develops, as the therapist gathers more and more information about the client. The amount of Reflection, on the other hand, generally fluctuates within a small interval throughout a dialogue in both high- (26% - 31%) and low-quality MI (2% - 7%), which means Reflection is common throughout a high-quality MI session and rare throughout a low-quality one. Finally, the proportion of Input rises during the middle stages ($(0.4, 0.8]$) in both high- and low-quality MI, but the increase is substantially more pronounced in low-quality MI sessions (from ∼33% to ∼61%) than in high-quality ones (from ∼9% to ∼14%), which indicates that a non-adherent therapist tends to talk from their own perspective more as the conversation develops.

The trends of different client talk types are displayed in Figure 4.8. A clear

[8] In all the line charts, the "marked" data points are the sample means and the error bars around them are calculated using bootstrapping with a 95% confidence interval.

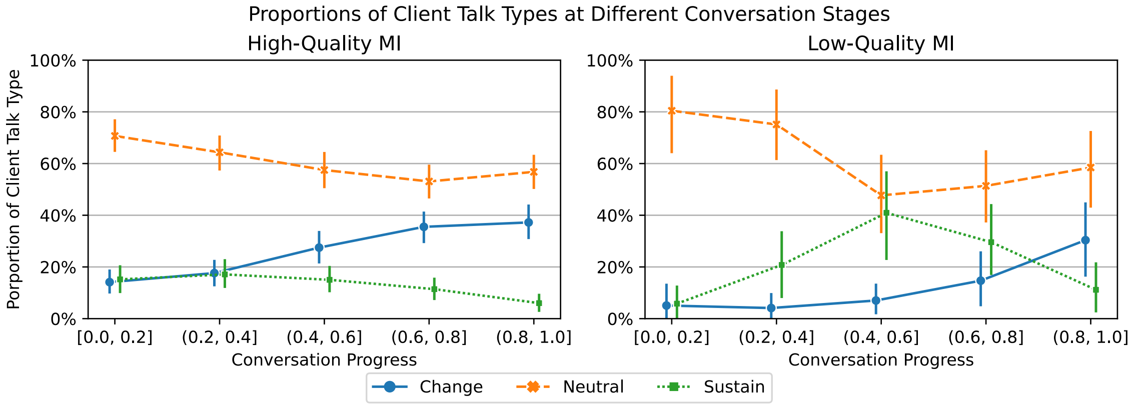


Figure 4.8: Proportions of client talk types in different conversation stages in high- and low-quality MI.

shift is shown in high-quality MI: there are similar amounts of Change Talk and Sustain Talk at the beginning of a conversation, but Change Talk becomes more present steadily and eventually reaches around 37% at the end of a dialogue, while the share of Sustain Talk diminishes gradually at the same time and drops to around 6%. In other words, the desired effects of MI-adherent counselling, namely change talk evocation and sustain talk reduction, become increasingly prominent with the progress of a session. In contrast, in low-quality MI, during the early & middle conversation stages (i.e., $[0.0, 0.6]$) the proportion of Sustain Talk soars from approximately 6% to a little over 40% while the number for Change Talk remains under 10%. Interestingly, the later stages (i.e., $(0.6, 1.0]$) show the opposite trend, as the growing share of Change Talk surpasses the declining proportion of Sustain Talk, finishing at around 30% and 11% respectively at the end. Nevertheless, the absolute $\%\{\text{Change Talk}\} - \%\{\text{Sustain Talk}\}$ difference is clearly larger at the end of high-quality MI sessions.

4.4.4 Utterance Length Distributions

Following [1], we study the lengths (number of words) of utterances of different types. To better represent the distribution of individual utterance lengths, we opt for violin plots to render a kernel density estimation of each underlying distribution, with the first, second and third quartiles marked as dashed lines. This applies to Figures 4.9, 4.10 and 4.11, although Figure 4.11 shows the distribution of utterance length ratios instead of absolute lengths.

Figure 4.9 shows the therapist and client utterance length distributions in high- and low-quality MI. It is clear that the client utterance length distributions are similar in MI-adherent and non-adherent sessions whereas therapist utterances are generally shorter in high-quality MI than in low-quality MI, which is another indicator that an MI-adherent therapist takes more time to actively listen to and

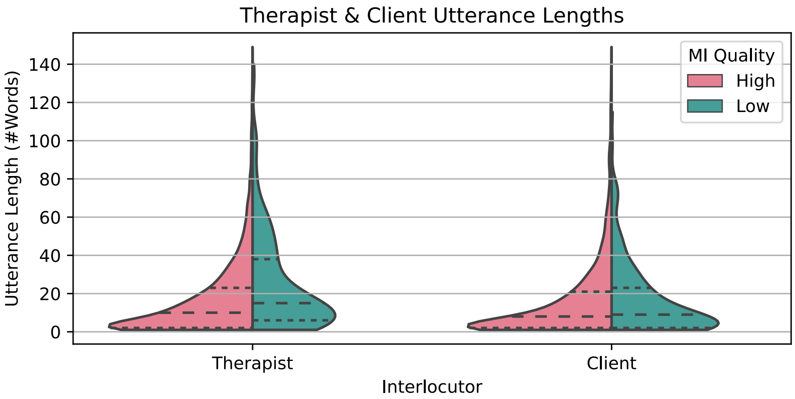


Figure 4.9: Lengths (number of words) of therapist & client utterances in high- & low-quality MI.

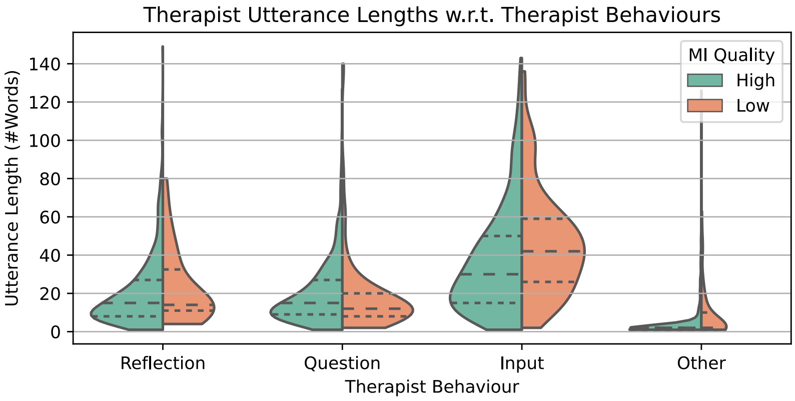


Figure 4.10: Utterance lengths (number of words) of different therapist behaviours in high- & low-quality MI.

understand their client.

Figure 4.10 shows a more fine-grained therapist utterance length distribution w.r.t. each therapist behaviour. For Reflection, the median utterance length is about the same in high- and low-quality MI, but the proportion of shorter utterances is clearly larger in the former. In terms of Question, an MI-adherent therapist tends to pose slightly longer questions than their non-adherent counterpart, which may suggest that an MI-adherent therapist more often asks tailored and nuanced questions. Input is substantially longer in both high- and low-quality MI, but input from an MI-non-adherent therapist is generally 10 words or more longer than that from an adherent therapist, indicating the relatively larger degree to which an MI-non-adherent therapist talks from their own perspective. Finally, the fact that utterances of the Other behaviour are generally short (no more than a few words) shows that those utterances mostly carry little meaning and are often simply used to facilitate the conversation.

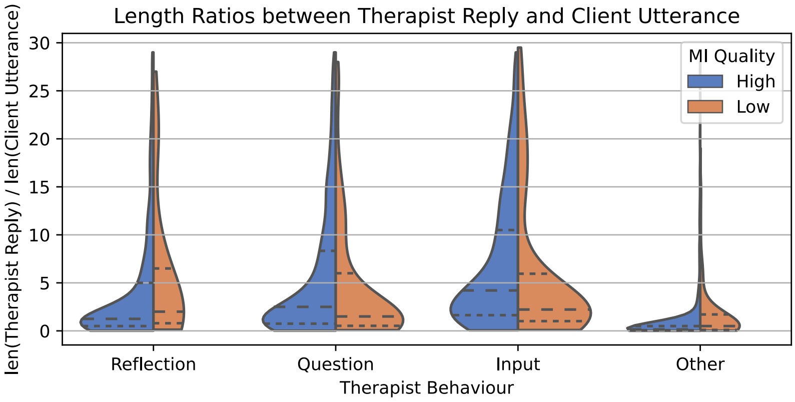


Figure 4.11: Ratios between the length (number of words) of the next-turn therapist response (broken down into 4 types of therapist behaviours) and that of the current-turn client utterance.

We also investigate the length ratio between a therapist reply and its immediately preceding client utterance, which shows how much longer the therapist "talks in return". As illustrated in Figure 4.11, Reflection has smaller length ratios in high-quality MI than in low-quality MI while Question shows the opposite, both of which are in line with the previous observation of the absolute lengths of Reflection and Question utterances in Figure 4.10. However, the Input length ratios are generally larger in high-quality MI sessions than in low-quality ones, which could be attributed to some utterances in high-quality MI annotated as Question where the therapist asks for permission to provide input. For example, the therapist might say "So, can I share with you some information on alcohol use?", and the client would simply say "Yes" or "Sure, why not", before the therapist replies with a substantially longer Input utterance, thus leading to a larger utterance length ratio.

Table 4.12: Most frequent 3-grams of therapist utterances of different **(Main) Behaviours** in high- and low-quality MI. Numbers of occurrences are shown in parentheses.

	High-Quality MI	**Low-Quality MI**
Reflection	“it sounds like” (78)	“’re gon na” (4)
	“sounds like you” (56)	‘you do n’t” (3)
	“a little bit” (51)	“you ’re here” (3)
	“you do n’t” (43)	“you ’re gon” (3)
	“a lot of” (39)	‘you ’ve already” (2)
Question	“do you think” (91)	“do you think” (11)
	“a little bit” (62)	“you think you” (6)
	“me a little” (35)	“a lot of” (5)
	“little bit about” (33)	“that you ’re” (5)
	“I ’m wondering” (30)	‘you ’re not” (5)
Input	“a lot of” (32)	“a lot of” (15)
	“a little bit” (27)	“you need to” (11)
	“one of the” (16)	“that you ’re” (9)
	“that you ’re” (13)	“’s gon na” (8)
	“you ’d be” (13)	“that you ’ve” (7)
Other	“for coming in” (12)	“that ’s certainly” (2)
	“that you ’re” (8)	“so it ’s” (2)
	“a little bit” (8)	“you ’re not” (2)
	“coming in today” (7)	“you ’re still” (2)
	“I do n’t” (7)	“be able to” (2)

4.4.5 Frequent 3-Grams

Tables 4.12 and 4.13 list the most frequent 3-grams in the therapist utterances of each behaviour and in the client utterances of each talk type, respectively. It is clear from the table that an MI-adherent therapist tends to use “it sounds like” to initiate a reflection — a common way of doing so in MI [11] — more often than a non-adherent therapist. Otherwise, however, the frequent 3-grams reveal little about the characteristics of utterances of different types or MI qualities. This suggests that utterance-level semantic differences are more nuanced and contextualised.

Table 4.13: Most frequent 3-grams of client utterances of different **Talk Types** in high- and low-quality MI. Numbers of occurrences are shown in parentheses.

	High-Quality MI	**Low-Quality MI**
Change Talk	"I do n't" (188) "do n't know" (68) "I 'm not" (42) "do n't want" (30) "I think I" (30)	"I do n't" (8) "I guess I" (6) "I think I" (5) "do n't know" (4) "I-I guess I" (4)
Neutral Talk	"I do n't" (261) "do n't know" (142) "I 'm not" (53) "do n't really" (47) "I did n't" (27)	"I do n't" (27) "I 'm not" (7) "do n't know" (7) "I 've been" (6) "I have n't" (5)
Sustain Talk	"I do n't" (135) "do n't know" (57) "I 'm not" (28) "it 's not" (23) "do n't really" (23)	"I do n't" (14) "I 'm not" (8) "do n't know (5) "It 's just" (5) "I just need" (4)

4.4.6 Utterance Embedding Distribution

To further investigate the semantic-level differences between utterances of different types, we probe clustering of utterance embeddings. Specifically, we obtain the utterance embeddings using an LM[9] as a sequence-level encoder. Through t-SNE[10] [169] — an unsupervised, non-linear technique for visualising high-dimensional data — Figure 4.12 and Figure 4.13 show that there is no obvious clustering of utterances of the same therapist **(Main) Behaviour** or client **Talk Type**, which is evidence that more advanced ML-based methods are needed to distinguish between utterances of different types.

[9] `sentence-transformers/all-MiniLM-L6-v2` [168] on Hugging Face. It is lightweight and performs well on sentence embedding tasks (`https://www.sbert.net/docs/pretrained_models.html`).

[10] maximum 1000 iterations, perplexity = 30.

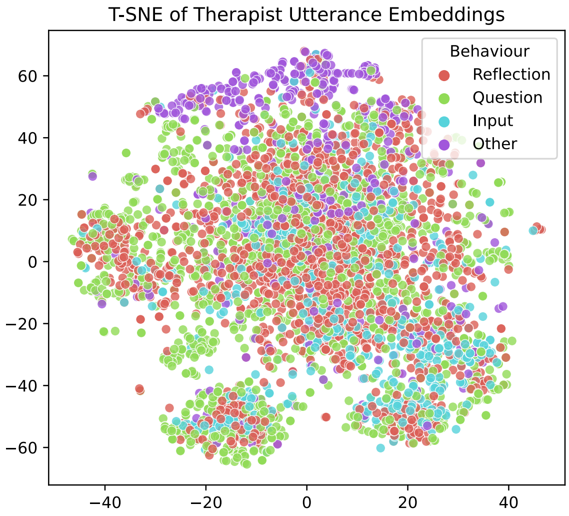


Figure 4.12: T-SNE of therapist utterance embeddings.

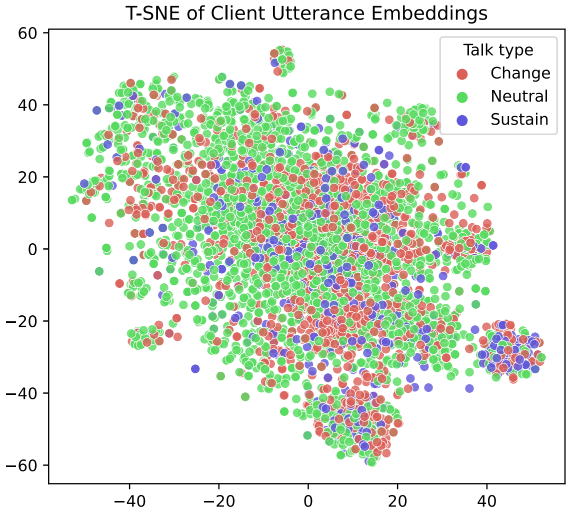


Figure 4.13: T-SNE of client utterance embeddings.

4.5 Discussion

While `AnnoMI` contains transcripts of MI demonstrations instead of real counselling sessions, we believe that it is the closest approximation possible without privacy violations, while the accurate transcription and the accompanying expert annotations further make it more reliable and versatile than similar datasets (e.g. [1]). We note that the source videos are from professional therapists and research organisations/institutes dedicated to relevant topics (e.g., reducing substance use), therefore the realism of the demonstrated client-therapist interaction can be considered reliable, as confirmed by the survey responses[11] from the professional annotators. Nevertheless, it could be interesting to explore possible domain gaps between the corpus and undisclosed real-world counselling datasets. For example, as the average duration of the source videos is 7 minutes and thus shorter than usual real-world counselling sessions [170], in future work we will replicate our analysis on other corpora with longer sessions and then compare the results with those obtained based on AnnoMI.

For applications, AnnoMI can be readily used to develop NLP/ML models for MI fidelity assessment, such as generating feedback to help train and supervise therapists. Example use cases of this kind include 1) categorising current-turn therapist behaviour and/or client talk, and 2) forecasting next-turn client talk type and/or MI-adherent therapist behaviour, both of which we explore in Chapter 5. Apart from those NLU settings, AnnoMI can also be used for NLG to assist human therapists, such as providing suggestions on what a therapist could say next (i.e., response generation) given the past utterances of an ongoing session, which we investigate in Chapter 6. Beyond our works detailed in this thesis, AnnoMI is also being used in other studies, such as [49] which approaches counselling quality classification from a perspective of fairness and bias mitigation.

4.6 Summary

We released AnnoMI [3], a dataset of professionally transcribed and expert-annotated conversations that demonstrate high- and low-quality MI. Based on the rich annotations by experienced therapists, we thoroughly analysed various counselling-related properties at the levels of utterance, dialogue and corpus.

AnnoMI represents a powerful resource for research in the important direction of counselling-related NLP. For future work, we will explore applications of AnnoMI with real-world impact, in particular those laid out in §4.5.

[11] We note that the high-quality MI dialogues could be closer to real-world good practices than the low-quality ones are to real-world pitfalls, but we cannot verify it based on the survey results alone. A different survey asking the annotators about the realism of high- and low-quality MI dialogues separately could have obtained more insights in this regard. We leave this to future work.

Chapter 5

Utterance-Level Behaviour Prediction & Forecasting

This chapter is based on:

Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3), 2023

Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Towards automated counselling decision-making: Remarks on therapist action forecasting on the annomi dataset. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 1906–1910. ISCA, 2022. DOI: 10.21437/Interspeech.2022-506

Significant progress has been made recently in NLP for MI dialogue analysis, but there lacks a common benchmark for tasks in this domain due to privacy-related constraints on data sharing. Fortunately, the introduction of the publicly available `AnnoMI` greatly reduces this obstacle. As a first step towards utilising `AnnoMI` as a benchmark, we explore 2 types of tasks with potential for real-life application based on this dataset: current-turn therapist/client behaviour prediction and next-turn therapist behaviour forecasting. Prediction is focused on identifying the behaviour label of the current turn given the single utterance, while forecasting aims to forecast the behaviour of the upcoming turn given the dialogue history. For prediction, we find that LMs such as BERT achieve higher performance on therapist behaviour than on client behaviour, and encouragingly those models generalise well to new topics for therapist behaviour prediction. For forecasting, we use LMs and explore a range of NLP modelling choices such as dialogue history length and contrastive training examples, and our results show that model performance is suboptimal irrespective of modelling choices, which reflects the broad latitude of therapists in counselling where there is often not a unique optimal next-turn action to take. With our findings, we hope to provide insights on these tasks and inspire future efforts in AnnoMI-based counselling dialogue analysis.

5.1 Introduction

Dialogue-related NLP research has seen significant development recently, driven by increasingly powerful LMs. In domains with counselling-like elements such as peer support dialogue, progress has been made (§2.3) ranging from empathy detection ([136, 138, 135], *inter alia*) to empathetic response generation (e.g., [82]). NLP for counselling dialogue analysis, however, has not been developed to the same extent, mostly due to privacy constraints on using real therapy conversation data (§2.2.1). Nevertheless, recent works have looked into topics such as examining therapist strategies [171] and providing real-time counsellor evaluation [172].

For MI-related NLP (§2.2), in particular, researchers have explored applications such as therapist empathy modelling [29, 30, 31, 48, 43], reflection generation [35, 36], automatic MISC/MITI coding of a single turn of therapist/client utterance [105, 33, 34, 13], as well as forecasting of the MISC/MITI code of the next-turn therapist/client utterance based on dialogue context [13]. The methods in these studies range from classical ML (e.g., support vector machines) with linguistic features to DL (e.g., RNNs). Despite this growth in NLP for MI, there lacks a common benchmark for these tasks, due to considerable privacy-related constraints on sharing MI dialogue datasets. Fortunately, with the introduction of the publicly available AnnoMI dataset [3] (Chapter 4), researchers can build on this expert-annotated MI conversation corpus.

As a first step of leveraging AnnoMI as a benchmark, we focus on 2 types of tasks: **current-turn therapist/client behaviour**[1] **prediction** and **next-turn therapist behaviour forecasting**[2]. Current-turn behaviour prediction is similar to automatic coding [105, 33, 34] in that it predicts the utterance label (i.e., therapist/client behaviour) of a known turn given its utterance, which accelerates post-session dialogue analysis and provides feedback and evaluation. Next-turn therapist behaviour forecasting is analogous to next-turn MISC/MITI code forecasting [13] in that it forecasts the therapist behaviour of the unknown upcoming therapist utterance given a dialogue history, which can assist the therapist by offering recommended next-turn actions to take in an ongoing session. Thus, both types of tasks have strong potential for real-world application, although we note that utter-

[1] For brevity, we use "therapist/client behaviour" to refer to both therapist (Main) Behaviour and client Talk Type.

[2] To avoid ambiguity, we use the term "prediction" exclusively when the target utterance is the current turn, and similarly exclusively "forecasting" when the target utterance is the next turn.

ance labels in AnnoMI follow a MISC/MITI-inspired scheme (§4.2.5) rather than the original MISC/MITI.

Our investigation and findings on these 2 task types are summarised below:

Current-Turn Therapist/Client Behaviour Prediction We experiment with various ML-based models, ranging from classical ML methods (random forest) to advanced LMs (BERT [21]). In particular, we also explore two aspects that have not been studied in previous work: 1) we investigate the performance impact of artificial class balance via data augmentation; 2) we examine the performance of models on different topics and their generalisability to new topics. We find that predicting therapist behaviours has higher performance than predicting client behaviours, which matches the difference between the two in terms of IAA on ground-truth labels. Our results also show, among other findings, that 1) artificial class balance hurts overall performance; 2) BERT models are generalisable to new topics for therapist behaviour prediction.

Next-Turn Therapist Behaviour Forecasting We utilise an advanced LM (RoBERTa [53]) to approach this task, unlike previous work [13] which only used GRUs. We also creatively probe a range of modelling choices and analyse their effects on forecasting performance, including 1) varying dialogue history length, 2) using data augmentation to expand training data, 3) inserting therapist and/or client behaviour labels in the dialogue history, and 4) contrasting next-turn therapist behaviours in high- and low-quality MI. Our experiments show that the baseline of using the original dialogue histories without special processing achieves the best performance, with minor contribution from dialogue history length. The modelling choices explored are mostly not conducive to better performance, which we posit is due to the noise introduced in the process. The best-performing model does not produce highly accurate forecasting if only the top-1 result is used, which reflects the latitude of a therapist in responding to the client — there is often more than one optimal next-turn action to take.

We detail and discuss the current-turn behaviour prediction experiments in §5.2 and the next-turn therapist behaviour forecasting experiments in §5.3.

5.2 Current-Turn Therapist/Client Behaviour Prediction

5.2.1 Tasks and Setup

We focus on two current-turn utterance label prediction tasks based on AnnoMI:

- **Therapist Behaviour Prediction:** Given a therapist utterance, predict its (Main) Behaviour.

Table 5.1: Distribution of (Main) Behaviour and Talk Type after de-duplication. Overall, (Main) Behaviour has 3796 unique examples and Talk Type has 3685.

(Main) Behaviour				**Talk Type**		
Reflection	Question	Input	Other	Change Talk	Neutral Talk	Sustain Talk
34%	36%	16%	14%	29%	57%	14%

- **Client Talk Type Prediction:** Given a client utterance, predict its Talk Type.

From a practical point of view, an accurate prediction model of therapist behaviour and client talk type can automatically label utterances and thus facilitate post-session analysis and insights/feedback generation for the therapist, ultimately improving counselling quality.

Each task allows a single utterance as the input and requires a class label as the output. We experiment with 4 ML-based models, as listed below. We implement the BERT variants with AdapterHub[3] [173] (in turn based on HuggingFace [159]), the CNN models with Keras [4], and the other models with Scikit-learn [160].

- **BERT w/o Adapters**: BERT-base-uncased [21] fine-tuned on `AnnoMI`.
- **BERT w/ Adapters**: BERT-base-uncased with adapters ([57, 173], §2.1.3) fine-tuned on `AnnoMI`.
- **CNN**: CNNs initialised with word2vec embeddings [174] and fine-tuned on `AnnoMI`.
- **Random Forest**: random forest with tf-idf features.

We also use 2 random baseline classifiers for comparison:

- **Prior**: random prediction based on the class distribution in the training set.
- **Uniform**: random prediction based on the uniform distribution of the classes.

Since duplicate utterances are present in `AnnoMI`, especially in the categories of Other and Neutral Talk (e.g., "Uh-huh" and "OK"), we perform de-duplication as a preprocessing step. Specifically, if multiple identical utterances have the same label, we randomly select one of them to keep and remove the others. The distribution of (Main) Behaviour and Talk Type after this step is shown in Table 5.1.

We first inspect overall model performance with cross validation (CV) in §5.2.2, and then examine the performance on different topics and model generalisability to new topics in §5.2.3. We use Macro F1 as the metric, since it is commonly used for classification tasks and robust to class imbalance.

[3] `https://github.com/Adapter-Hub/adapter-transformers`

[4] `https://keras.io/`

5.2.2 Overall Performance

Considering the relatively small size of `AnnoMI`, we conduct 5-fold CV at the utterance level with stratification w.r.t. utterance labels, so that A) the class distribution in each fold is close to being identical, and B) each time 4 folds are used as training + validation data and 1 fold is used as test data. The training to validation data ratio is 9:1, and we select the best-performing checkpoint (for CNN and BERT) based on performance on the validation set, so that we can test the checkpoint on the test set. For Prior and Uniform whose outputs are random, we run the models 1000 times on the test data and calculate the average performance. Therefore, each of the 6 models listed in §5.2.1 eventually has 5 performances from 5-fold CV, and we take the mean as the final performance of the model.

To address class imbalance between different (Main) Behaviours and between different Talk Types, we introduce two versions for each training set:

- **Original Unbalanced**: keeping the original data in each CV training set.
- **Augmented Balanced**: augmenting ([175]) the non-majority classes, so that each class in Augmented Balanced is equal in size to the majority class in Original Unbalanced. Based on a preliminary study, we use an off-the-shelf Pegasus [176]-based neural paraphrasing model[5] to generate utterance paraphrases as augmentations.

Therapist Behaviour Prediction

As shown in Table 5.2, the BERT variants score the highest with macro F1s at 0.72, followed by CNN at 0.6 and Random Forest at approximately 0.5. Compared to the random baselines (Prior & Uniform) with macro F1s below 0.25, the trained models, especially the BERT variants, have clearly learned contextualised semantics. No substantial difference exists between the results of BERT w/o Adapters and BERT w/ Adapters, which shows the efficacy of adapters.

We also observe that the effects of data augmentation are minor and universally negative for the BERT variants and CNN. We postulate that this is attributable to the altered semantics in some paraphrase utterances caused by hallucination [40]. Specifically, upon closer inspection of the paraphrase utterances, we notice that while most paraphrases carry the same meaning as the original utterances, semantic alterations do exist. For example, a Question utterance “What else besides drinking helps you relax and unwind in the evenings?” is paraphrased into “In the evening, drinking helps you relax.”, which would be a Reflection utterance instead. Thus, those semantic alterations likely introduced noise in the training examples and consequently confused the classifiers.

The order of per-class F1 by the best-performing models (BERT variants) is generally Input ≈ Other < Reflection < Question, which largely aligns with the

[5] https://huggingface.co/tuner007/pegasus_paraphrase

Table 5.2: Macro F1 and per-class F1's of (main) therapist behaviour prediction. All results averaged from 5-fold CV. ↑/↓: performance increase/decrease by using Augmented Balanced compared to using Original Unbalanced.

(Main) Therapist Behaviour Prediction

Result Format: **Original Unbalanced [Augmented Balanced]**

Model	`F1-Macro`	Reflection	Question	Input	Other
BERT w/ Adapters	.72 [.70↓]	.77 [.75↓]	.86 [.84↓]	.63 [.60↓]	.64 [.62↓]
BERT w/o Adapters	.72 [.70↓]	.77 [.75↓]	.85 [.85]	.63 [.60↓]	.64 [.62↓]
CNN	.60 [.58↓]	.64 [.63↓]	.70 [.70]	.50 [.48↓]	.56 [.52↓]
Random Forest	.50 [.50]	.56 [.53↓]	.58 [.54↓]	.41 [.45↑]	.46 [.46]
Prior	.25 [.24↓]	.34 [.29↓]	.36 [.30↓]	.16 [.20↑]	.14 [.18↑]
Uniform	.24 [.24]	.29 [.29]	.30 [.29↓]	.20 [.20]	.18 [.18]

order of proportions of those labels in the task (Table 5.1). The performance gap between different classes is not reduced in the Augmented Balanced scenario, which shows that this issue cannot be resolved by simple paraphrasing-based class-wise data augmentation. Interestingly, Question shows better (e.g., $\Delta = 0.09$ $F1$ for BERT w/ Adapters) performance than Reflection despite having similar amounts of examples, which may be because 1) Question utterances generally have syntactic cues such as question marks and are hence easier to classify, 2) Question has higher IAA than Reflection (Table 4.10) and its ground-truth labels are thus less noisy.

By inspecting the confusion matrix (Figure 5.1) of BERT w/ Adapters in the Original Unbalanced setting, we observe that Input and Other utterances are most frequently misclassified into Reflection. Since Input and Other are less than half in size compared to Reflection (Table 5.1), this imbalance may have contributed to the instances of misclassification. On the other hand, Reflection and Question are similarly sized but Question contributes less to the misclassifications of Input and Other, which may again be linked to the syntactic cues and less noisy labels of Question as mentioned before.

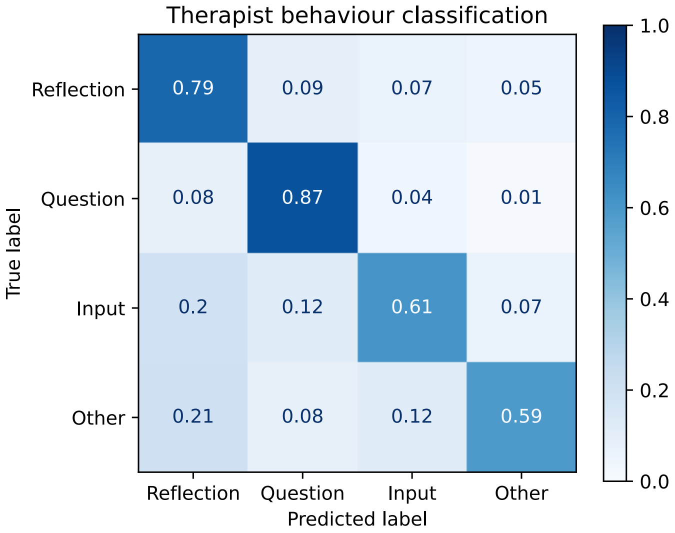


Figure 5.1: Confusion matrix of BERT w/ Adapters for (main) therapist behaviour prediction in the Original Unbalanced setting. Normalised by row.

Client Talk Type Prediction

As shown in Table 5.3, this task records universally lower scores than Therapist Behaviour Prediction for all the trained models — the best BERT-variant performances are around 0.55 macro F1 while CNN and Random Forest score below 0.47, irrespective of data augmentation. Two factors likely responsible for the performance gap between the two tasks are **dialogue context** and **annotation noise**. In terms of dialogue context, in some cases, the talk type of a client utterance can only be determined with context grounding. For example, "Yeah, for sure" as a reply to "So you work out every day?" is Neutral Talk, but it should be Change Talk when it follows "So do you think you could smoke less for the sake of your kids?". As for annotation noise, the Fleiss'-kappa-based IAA for client talk type is around 0.47 while it is 0.74 for therapist behaviour (§4.3.2), which suggests that annotating talk type is more challenging and therefore more noise is present in the ground-truth labels and in turn makes it harder to optimise the trainable models.

Among the talk types, Neutral Talk has the best performance, followed by Change Talk and Sustain Talk, which matches the class distribution (Table 5.1), similar to the finding in Therapist Behaviour Prediction (§5.2.2). Interestingly, in some cases, data augmentation reduces performance gap between classes. For example, for BERT w/ Adapters, the gap between Change Talk and Neutral Talk is 0.23 F1 in Original Unbalanced but 0.14 F1 in Augmented Balanced, although it is mostly attributable to the performance decrease of 0.07 F1 on Neutral Talk under Augmented Balanced. Considering that the boundaries between Change Talk and Neutral Talk may not always be clear-cut (shown by lower IAA on Talk Type), we postulate that the augmentations of Change Talk training examples that resemble

Table 5.3: Macro F1 and per-class F1's of client talk type prediction. All results averaged from 5-fold CV. ↑/↓: performance increase/decrease by using Augmented Balanced compared to using Original Unbalanced.

Client Talk Type Prediction

Result Format: **Original Unbalanced [Augmented Balanced]**

Model	`F1-Macro`	Change Talk	Neutral Talk	Sustain Talk
BERT W/ ADAPTERS	.55 [.53↓]	.51 [.53↑]	.74 [.67↓]	.39 [.37↓]
BERT W/O ADAPTERS	.53 [.52↓]	.49 [.51↑]	.71 [.67↓]	.39 [.39]
CNN	.47 [.46↓]	.45 [.44↓]	.65 [.63↓]	.31 [.31]
RANDOM FOREST	.39 [.44↑]	.38 [.40↑]	.71 [.65↓]	.10 [.26↑]
PRIOR	.33 [.31↓]	.29 [.31↑]	.57 [.42↓]	.14 [.20↑]
UNIFORM	.31 [.31]	.31 [.31]	.42 [.42]	.20 [.20]

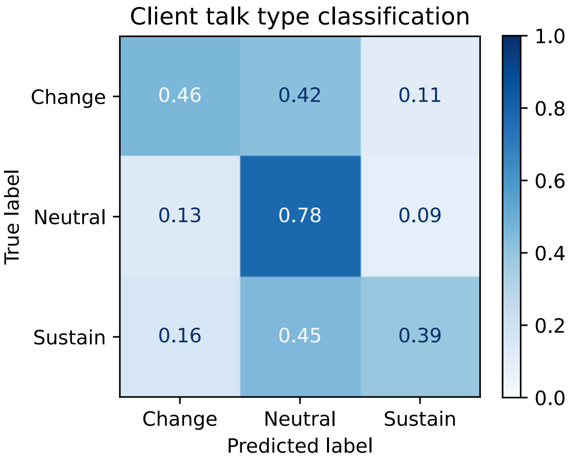


Figure 5.2: Confusion matrix of BERT W/ ADAPTERS for client talk type prediction in the Original Unbalanced setting. Normalised by row.

Neutral Talk make the classifiers less certain about Neutral Talk and in turn cause the performance drop on Neutral Talk.

As can be seen in the confusion matrix (Figure 5.2), both Change Talk and Sustain Talk are frequently misclassified as Neutral Talk even by the best performing model BERT W/ ADAPTERS. Using dialogue context as an additional input may reduce misclassification to a certain extent as hypothesised before, but class imbalance may ultimately become the bottleneck for performance improvement [13]. We leave further probing to future work.

Table 5.4: Topic-specific and cross-topic performance (macro F1) in Original Unbalanced for 1) Therapist Behaviour Prediction and 2) Client Talk Type Prediction. ↑/↓: cross-topic performance is higher/lower than topic-specific performance.

Result Format: **Topic-Specific → Cross-Topic**			
Topic	*Reducing Alcohol Consumption*	*Reducing Recidivism*	*Smoking Cessation*
(Main) Therapist Behaviour Prediction			
BERT w/ Adapters	.74 → .74	.63 → .62↓	.70 → .72↑
BERT w/o Adapters	.72 → .75↑	.65 → .66↑	.72 → .70↓
CNN	.59 → .55↓	.50 → .52↑	.64 → .60↓
Random Forest	.49 → .49	.40 → .36↓	.53 → .48↓
Client Talk Type Prediction			
BERT w/ Adapters	.55 → .52↓	.41 → .43↑	.56 → .51↓
BERT w/o Adapters	.54 → .52↓	.41 → .42↑	.55 → .50↓
CNN	.47 → .45↓	.39 → .39	.50 → .43↓
Random Forest	.42 → .38↓	.33 → .34↑	.37 → .32↓

5.2.3 Topic-Specific Performance and Generalisability to New Topics

Apart from CV performance over the entire `AnnoMI`, we also explore **topic-specific performance**, i.e., how well the models perform on conversations of different topics, as we hypothesise that some topics may be more challenging for certain models on particular tasks. Since reliable models with real-world impact should generalise well to topics unseen during training, we also probe **cross-topic model performance** by training on data of all but one topic and testing on examples from that topic.

Based on the topic coverage of `AnnoMI` (Table 4.4), we select three topics — *reducing alcohol consumption*, *reducing recidivism*, and *smoking cessation* — for probing the topic-specific and cross-topic performance of all the trained models on the two tasks defined in §5.2.1, since between 10% and 20% of the utterances in `AnnoMI` belong to conversations of each of these topics. We focus on the Original Unbalanced setting, as the performance in Augmented Balanced is similar. The results are shown in Table 5.4.

Topic-Specific Performance

To obtain the performance on topic T_i, we re-use the 5-fold CV models for the two tasks (§5.2.2), but we test each model on a T_i-specific subset of the corresponding test fold. Specifically, the subset consists entirely of utterances that are originally from conversations that have T_i as the only topic. By averaging the performances of the 5 models on their respective T_i-specific test-fold subsets, this method covers all T_i-utterances and thus yields a reliable measure of the T_i-specific performance of each model type.

Generally, it is clear from Table 5.4 that the model performances, especially those of the BERT variants, follow the topic-wise ordering below:

- **Therapist Behaviour Prediction:** reducing alcohol consumption > smoking cessation > reducing recidivism
- **Client Talk Type Prediction:** reducing alcohol consumption ≈ smoking cessation > reducing recidivism

One contributing factor to the performance gaps between different topics could be topic coverage, namely the number of utterances from sessions of a particular topic, since better coverage entails more data used for training. For example, reducing alcohol consumption has more utterances than reducing recidivism (Table 4.4) and correspondingly also better performance.

However, it is also clear that the performance on reducing-recidivism conversations is considerably lower than on smoking cessation, despite the slightly larger coverage of reducing recidivism. This is more likely because the utterances of the topic themselves are more semantically challenging for the task, and it also shows the necessity to include a wide range of topics in a counselling dialogue dataset.

Cross-Topic Performance

It is often important for trained models to generalise to unseen domains. While conversations of different topics are not completely different domains, the topic-specific performance shows that models indeed have varying levels of performance depending on the topic. Hence, to complement the topic-specific setting where models trained on dialogues of **all** topics are examined for their performance on certain topics, we probe the generalisasbility of a model by removing a topic T_i from its training set completely and then analysing its performance on a T_i-only test set.

Concretely, we adopt a leave-1-topic-out approach by training on all the `AnnoMI` utterances from conversations that do **not** have topic T_i and testing on all the `AnnoMI` utterances from dialogues that **only** have topic T_i. Conversations with multiple topics that include T_i are not present during training or testing. We note that this setup and the topic-specific setting have effectively the same test set, therefore the cross-topic and topic-specific performances can be compared fairly.

As Table 5.4 shows, for therapist behaviour prediction, the performance of the BERT models remains stable moving from topic-specific to cross-topic, which shows 1) the models are generalisable to new topics for this task, 2) therapist language is relatively consistent in conversations of different topics.

For client talk type prediction, on the other hand, consistent and more noticeable (as much as 0.05 F1 for the BERT models) performance drops can be seen for reducing alcohol consumption and smoking cessation. While this may indicate that client language varies more across topics, we note that client talk type prediction generally has lower performance than therapist behaviour prediction (§5.2.2), and thus it may be a more challenging task in general and need more training data irrespective of topic.

5.2.4 Discussion & Summary

We experimented with a range of ML models to approach the tasks of current-turn therapist/client behaviour prediction, and we found that the BERT models have the best performance for both tasks overall. Our results showed that, although class imbalance leads to lower performance on minority classes, artificial class balance via data augmentation generally hurts overall performance.

Overall, client behaviour prediction has lower performance than therapist behaviour prediction, which we postulated is partially attributable to the more noisy ground-truth labels. We also posited that better performance could be achieved by including dialogue context such as preceding utterances as auxiliary input.

Furthermore, we found that model performance is higher on some dialogue topics than on others, but generally the BERT models have similar performance on therapist behaviour prediction when tested on topics unseen during training time, which shows their generalisability to new topics.

For future work, we plan to explore rich dialogue context as auxiliary input for both therapist behaviour and client talk type prediction, and in particular doing so in a few-shot setting with LLMs such as GPT-3 [17] and GPT-4 [28]. Future work may also probe possible gaps between performance on high-quality MI utterances and performance on low-quality ones, as well as how those gaps may be bridged.

5.3 Next-Turn Therapist Behaviour Forecasting

5.3.1 Task Setup

We explore the task of next-turn therapist behaviour forecasting, which forecasts the behaviour of the upcoming utterance, given the most recent turns as the dialogue history. From a practical point of view, an accurate forecasting model can assist the human therapist in a live counselling session by suggesting the optimal next-turn action to take, which can be particularly helpful for training junior therapists.

Problem Statement

At time step t in a counselling dialogue, we take the most recent N turns (one utterance per turn) as the dialogue history, namely $H_t^N = \{u_{t-N}, \cdots, u_{t-1}\}$ where u_{t-1} is a client utterance, and the goal is to forecast the therapist behaviour label

Table 5.5: A 3-turn dialogue history from a high-quality MI conversation

Utterance	Role	Label	Text
u_{t-3}	**Client**	**Neutral**	I guess I'm not paying attention to it.
u_{t-2}	**Therapist**	**Reflection**	Yeah. There's certainly been– There's no problems and you, as you said, only have had it for a short time.
u_{t-1}	**Client**	**Neutral**	Mm-hmm.

y_t^T of the immediate therapist response u_t, given H_t^N as the input. $y_t^T \in Y^T$ where Y^T is a predefined set of therapist behaviour labels.

Data Description

We base our experiments on the AnnoMI dataset [3] (Chapter 4). We form the training/validation/test sets for this task using the 110 dialogues showcasing high-quality MI with over 8.8K utterances in total (Table 4.1), since the goal is to emulate next-turn behaviours of a **good** therapist. Nevertheless, we do utilise low-quality MI dialogues as part of the training data in some setups (§5.3.2).

For $\{y_t^T\}$, we use the therapist **(Main) Behaviour** annotations of AnnoMI, which means $Y^T = \{\text{Reflection, Question, Input, Other}\}$. We also use the client **Talk Type** annotations in some setups as auxiliary input (§5.3.2).

General Input Format

To distinguish between the utterances in a dialogue history, we insert interlocutor labels and utterance separators as plain text. For example, the 3-turn dialogue history H_t^3 in Table 5.5 is converted into a single sequence as input to the model as follows:

"*⟨client⟩I guess I'm not paying attention to it.|⟨therapist⟩Yeah. There's certainly been– There's no problems and you, as you said, only have had it for a short time.|⟨client⟩Mm-hmm.*"

5.3.2 Modelling Choices

Fine-tuning LMs on AnnoMI for the task, we explore several modelling choices. While recent work (e.g., [17]) has shown the superior few-shot performance of LLMs, we leave the probing of few-shot learning for this task to future work.

Dialogue History Length

More knowledge of the dialogue exchanges that have taken place may lead to better next-turn therapist behaviour suggestion. Thus, we vary the dialogue history length (i.e., number of utterances) to probe its effects. In particular, for every other modelling choice in §5.3.2, we combine it with different dialogue history lengths for deeper insights.

Implementation Details: We consider 6 dialogue history length options: {1, 3, 5, 7, 9, *max*}, using the most recent 1/3/5/7/9 turns or using as much dialogue history as possible (*max*). Where applicable, we left-truncate the input to keep the 512 tokens representing the most recent context.

Data Augmentation

The relatively small scale of `AnnoMI` motivates the use of data augmentation. Similar to in current-turn therapist/client behaviour prediction (§5.2), we opt for paraphrasing as the means of augmenting dialogue histories.

Implementation Details: We adopt the same neural paraphrasing model used in §5.2 to generate utterance paraphrases, and we combine them to create dialogue history paraphrases. Specifically, we generate 10 paraphrases $\{\hat{u}_i^m\}_{m=1}^{10}$ for each utterance $u_i \in$ `AnnoMI`. Then, for each original dialogue history $H_t^N = \{\ldots, u_i, \ldots\}$, we obtain 5 augmentations $\{\tilde{H}_t^{N,o}\}_{o=1}^5$ where $\tilde{H}_t^{N,o} = \{\ldots, \tilde{u}_i^o, \ldots\}$ and $\tilde{u}_i^o$ is randomly sampled from $\{u_i\} \cup \{\hat{u}_i^m\}_{m=1}^{10}$. Thus, we use $\{\tilde{H}_t^{N,o}\}_{o=1}^5$ during training to effectively train on 5x amount of data, while keeping the validation and test sets unchanged.

Inserting Utterance Labels in Dialogue History

As utterance-level labels offer MI-relevant details, incorporating them in the dialogue history may positively impact task performance. We explore 3 options of inserting utterance labels as plain text:

- **Therapist Only:** prepending the label of each therapist utterance to the utterance as plain text.

- **Client Only:** prepending the label of each client utterance to the utterance as plain text.

- **Therapist & Client:** prepending the label of each therapist and client utterance to the utterance as plain text.

Since ground-truth utterance labels are not available during inference time, we experiment with 3 label sources:

- **Oracle:** using ground-truth utterance labels.

- **Predicted:** training a current-turn utterance label classifier[6] (only current-turn utterance as input) and using its predicted label for each utterance in the dialogue history.

[6] Similar to §5.2 but independently trained using `roberta-base` due to practical reasons.

- **Random:** using randomly sampled utterance labels, for comparison with **Oracle** and **Predicted**.

As an example, the dialogue history H_t^3 in Table 5.5 is converted into the following using **Therapist & Client** and **Oracle**:

"*⟨client⟩~⟨neutral⟩I guess I'm not paying attention to it.|⟨therapist⟩~⟨reflection⟩ Yeah. There's certainly been– There's no problems and you, as you said, only have had it for a short time.|⟨client⟩~⟨neutral⟩Mm-hmm.*"

Contrasting High- & Low-Quality MI

Inspired by recent work using plain-text control codes to influence LM output (e.g., [177]), we probe contrasting high- & low-quality MI with plain-text MI quality labels, since low-quality MI as negative examples may improve decision boundaries.

Specifically, we prepend the MI quality label of the conversation from which a dialogue history H_t^N is taken, while the other parts of the input remain unchanged. For the high-quality MI H_t^3 shown in Table 5.5, the input becomes:

> "*[high][SEP]⟨client⟩I guess I'm not paying attention to it.|⟨therapist⟩Yeah. There's certainly been– There's no problems and you, as you said, only have had it for a short time.|⟨client⟩Mm-hmm.*",

where *[SEP]* is a model-specific separator for a pair of texts. The models are trained on contrasting high- & low-quality MI dialogues and then tested on high-quality MI conversations only.

Due to the imbalance between high- and low-quality MI dialogue volumes (Table 4.1), we explore two variants of contrast where examples from high- and low-quality MI are mixed:

- **Mixed-Unbalanced:** using the original unbalanced high- and low-quality MI dialogue histories.

- **Mixed-Aug-Balanced:** using augmented low-quality MI dialogue histories to achieve balance, where the augmentations are generated in the same way as described in the **Data Augmentation** modelling choice.

5.3.3 Results & Analysis

Considering the relatively small scale of `AnnoMI`, 5-fold CV is used in all our experiments, and the dialogues in the training, validation and test sets are mutually exclusive. With 5-fold CV, we obtain 5 test-set scores from the 5 splits for each training setup. Therefore, in Figures 5.3, 5.5 and 5.6, we show the mean score (line) with a 95% confidence interval (error band) based on the 5 test-set scores of each

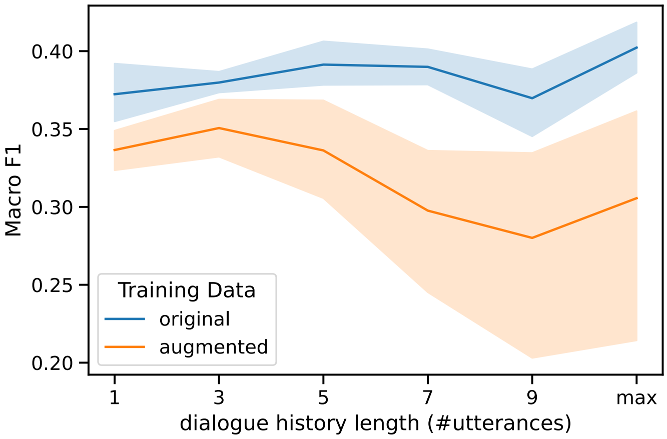


Figure 5.3: Impact of dialogue history length and data augmentation

setup, which offers insights on cross-split performance variation. Unless otherwise specified, we use macro F1 as the metric, following [13].

All our models are based on the HuggingFace [159] implementation of `roberta-base`[7], using an AdamW [178] optimiser with linear learning rate decay from an initial 2e-5. The batch size is 8, and the maximum input length is 512 tokens.

Dialogue History Length & Data Augmentation

Training on unaugmented high-quality MI dialogues, the performance improves with longer dialogue histories and reaches 0.39 macro F1 under the 5-utterance dialogue history setting (Figure 5.3). Afterwards, it steadily decreases as the dialogue history grows further but rebounds to 0.4 when using maximum history. Overall, the performance does not vary substantially across the splits.

Figure 5.4 shows the confusion matrices of the 1- and *max*-utterance models, where we observe that both models forecast Question most correctly, followed by Other, Reflection and Input. In particular, increasing the dialogue history length from 1 to *max* benefits the forecasting of Input and Reflection the most.

The best score of 0.4 macro F1 is insufficient for real-world deployment, echoing [13] where the best code forecasting score is 0.31, though the results are not directly comparable since [13] is based on an undisclosed counselling dialogue dataset that is annotated differently. The low score is likely linked to the latitude of therapists in counselling, as sometimes there are multiple good actions to take. Just as the confusion matrices show the model uncertainty between Reflection and Question, the therapist may reflect or pose a question after, for example, the client explains their personal circumstances.

[7] We also explored `roberta-large`, but it had consistently lower scores and larger cross-split performance variation in CV.

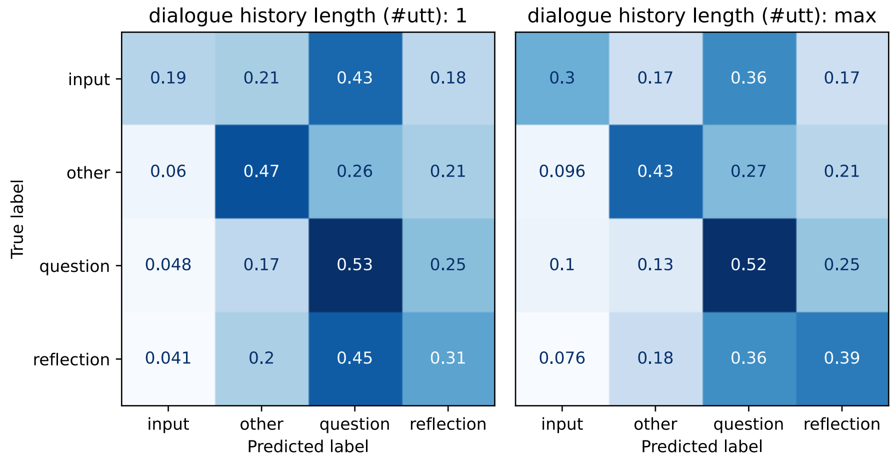


Figure 5.4: Confusion matrices of 1- and *max*-utterance baselines

Training on augmented dialogue history yields consistently lower performance and considerable cross-split performance variation. As the dialogue history grows longer — thus more paraphrases used in the input — the performance generally worsens and larger cross-split performance variation occurs.

Thus, we hypothesise that therapist behaviour forecasting is sensitive to conversation semantics, and that the performance is thus negatively impacted when some paraphrases contain semantic alterations caused by hallucination of the paraphrasing model, which echoes §5.2.2. For example, the client utterance "But I'm healthy. What health problems are you talking about?", which shows the client's defensiveness when they are told about potential health problems caused by excessive drinking, is paraphrased into "I'm good. What health problems do you have?", which carries a completely different meaning.

Inserting Utterance Labels in Context

Figure 5.5 shows the model performances when therapist and/or client utterance labels are incorporated in the dialogue history, where the label sources are Oracle, Predicted, or Random.

Overall, using only Oracle therapist utterance labels slightly (≤ 0.01 macro F1) outperforms the label-less baseline in most settings, while using Oracle labels for client utterances only and for both therapist and client utterances shows mixed results and larger cross-split performance variation. This difference likely points to the closer alignment between therapist utterances and their Oracle labels as evidenced by the higher IAA (Table 4.10), which enables useful additional training signals.

Using Predicted and Random utterance labels mostly underperforms the baseline and shows larger cross-split performance variation, with Random suffering slightly more. While Random introduces considerable noise in the input, which unsurpris-

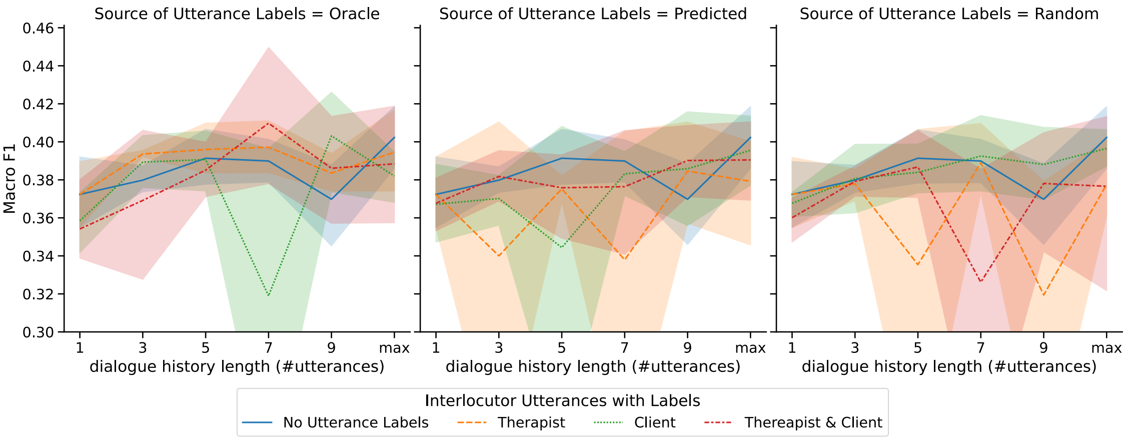


Figure 5.5: Impact of inserting therapist/client utterance labels in dialogue history on model performance

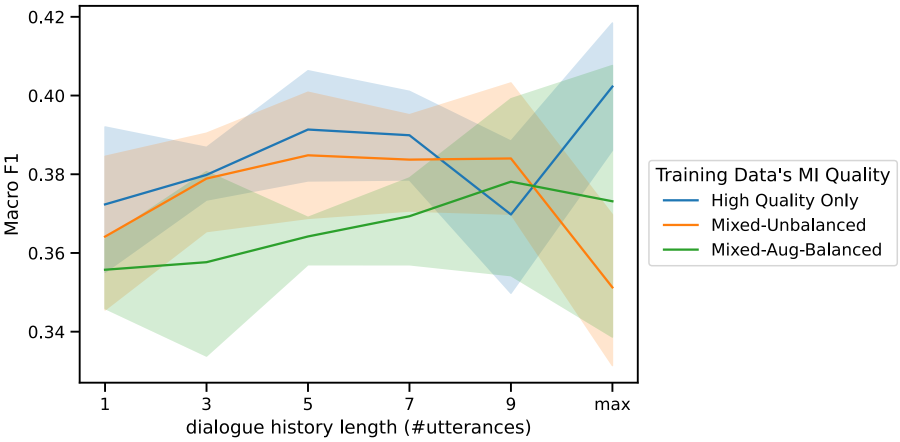


Figure 5.6: Impact of contrasting high- & low-quality MI

ingly harms performance, the low performance of Predicted is unexpected. One possible explanation is that the label-less baseline already understands the dialogue history relatively well without slightly noisy predicted utterance labels, especially in setups with longer dialogue histories where incorrect predicted labels are more likely to occur.

Contrasting High- & Low-Quality MI

The baseline trained only on high-quality MI dialogues mostly surpasses the models trained on mixed-quality MI conversations (Figure 5.6). In particular, Mixed-Aug-Balanced generally yields the lowest scores, echoing our previous finding that training on paraphrased dialogues harms performance. As for Mixed-Unbalanced, a plausible reason for its underperforming the baseline is that the therapist behaviours in the low-quality MI dialogues may not adequately represent the "mistakes" of the baseline, and hence the contrast is not effective enough to improve the decision boundaries, while the low-quality MI dialogues at the same time introduce more uncertainty into the ground truth behaviour labels during training.

5.3.4 Discussion & Summary

We experimented with various modelling choices for next-turn therapist behaviour forecasting, including dialogue history length, data augmentation, inserting utterance labels into the dialogue history, and contrasting high/low-quality MI dialogues. Generally, the baseline using plain dialogue history without particular NLP techniques achieves the best results, with relatively minor impact from dialogue history length. The techniques explored in this work proved to mostly introduce noise and hurt performance.

Overall, the strong baseline is not an ideal forecaster if only the top-1 forecast is used. Since the ground-truth labels are well-defined and annotated with a substantial agreement, a likely explanation for the low performance is that it is linked to the latitude/flexibility of therapists in their response. Future work may probe this flexibility further by asking professional therapists to annotate each dialogue history with alternative optimal next-turn therapist behaviours, where applicable. Future work could also investigate probabilistic formulations of the task to accommodate this flexibility. Also worth exploring are 1) suggesting what action(s) the therapist should **not** take next and 2) detecting worsening of counselling quality in real-time, where the contrast between high- and low-quality MI dialogues is likely more beneficial.

Chapter 6

Human Evaluation for Therapist Response Generation

This chapter is based on:

Zixiu Wu, Simone Balloccu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Towards in-context non-expert evaluation of reflection generation for counselling conversations. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 116–124, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics

Zixiu Wu, Simone Balloccu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Are experts needed? on human evaluation of counselling reflection generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6906–6930, Toronto, Canada, July 2023. Association for Computational Linguistics

Reflection is a crucial counselling skill where the therapist conveys to the client their interpretation of what the client has said. LMs have recently been used to generate reflections automatically, but human evaluation is challenging, particularly due to the cost of hiring experts. Laypeople-based evaluation is less expensive and easier to scale, but its quality is unknown for reflections. Therefore, we explore whether laypeople can be an alternative to experts in evaluating a fundamental quality aspect: coherence and context-consistency. To do so, we first conduct a preliminary study to create an evaluation scheme, based on free-text descriptions by laypeople about incoherence/inconsistency errors in generated synthetic reflections[1].

[1] In this chapter, we refer to generated reflections as "synthetic reflections" for better distinction from "human reflections".

Then, using this scheme, we ask a group of laypeople and a group of experts to annotate both synthetic reflections from models and human reflections from actual therapists. We find that both laypeople and experts are reliable annotators and that they have moderate-to-strong inter-group correlation, which shows that laypeople can be trusted for such evaluations. We also discover that GPT-3 mostly produces coherent and consistent reflections, and we explore changes in evaluation results when the source of synthetic reflections is switched from the less powerful GPT-2 to the more powerful GPT-3.

6.1 Introduction

In MI, reflective listening is a crucial strategy of showing empathy, where the therapist conveys a brief conversational summary of how they understand what the client has said [16, 11], as shown in the example in Table 6.1. Learning effective reflective listening requires considerable training time and expert supervision [113, 114]. Therefore, recent studies have used LMs as automatic reflection generators (§2.2.4) to aid therapist training [35, 36, 37], where the LM receives a dialogue context as the input and outputs a reflection (Table 6.1).

Human evaluation of reflection generation is crucial, since automatic metrics are often not robust [62]. For such evaluations (§2.2.4), experts (professional therapists) are used due to their deep understanding of the complex and sensitive domain of counselling dialogue. However, expert evaluation is costly and difficult to scale, and previous human evaluation studies often adopted over-simplified annotation schemes (e.g., good vs. bad reflection) or worked with short dialogue contexts ($\leq$ 5 turns). Evaluation with laypeople (such as crowdworkers) tends to be less expensive [179], but to the best of our knowledge its reliability for reflections is unknown.

In this work, we investigate if laypeople are a viable alternative to experts for human evaluation of **coherence and context-consistency** (referred to as **coherence** for brevity). This is a weakness of recent generative models [40] and also a fundamental quality aspect of reflection generation, because a reflection has to first "make sense" in the context before it can be evaluated against counselling principles. Furthermore, since coherence is related to common sense rather than psychology/psychotherapy expertise, we hypothesise that laypeople and experts are both capable of coherence evaluation.

Therefore, we first conduct a preliminary study (§6.2) to devise a coherence evaluation scheme. We do so by 1) collecting free-text descriptions from laypeople about incoherence/inconsistency errors in synthetic reflections and 2) identifying common patterns in the error descriptions and summarising them into discrete categories using thematic analysis [180], similar to recent works (e.g., [181]) that adopt bottom-up designs of text error annotation schemes. We use longer contexts — 14 turns on average — to better ground reflection generation and error assessment by laypeople. Thus, we establish {Malformed, Off-Topic, Dialogue-Contradicting, Parroting,

Table 6.1: A dialogue context from AnnoMI ([3], Chapter 4) about reducing substance use, together with its **human reflection** and two examples of **synthetic reflections**. Self-repetitions and mid-sentence changes (e.g., "it-it-whatever") are characteristics of spoken dialogue that are faithfully transcribed.

Context
Client: Well, I'm here because my mom wants me to be here. **Therapist**: Mm-hmm. **Client**: I don't really wanna be here, but it-it– whatever. **Therapist**: Got it. **Client**: Um, she-she found my stash- **Therapist**: Uh-huh. **Client**: -and she freaked out, and she's going crazy over it. Um, I don't why she was going through stuff in the first place, but whatever, so, now I'm here. · · · (intermediate turns) **Therapist**: And, uh, sounds like you're-you're pretty upset with your mom for-for doing that? **Client**: I am. **Therapist**: Yeah. **Client**: I mean, it's my stuff, I don't know why she's–
Reflection Candidates
Therapist (Human): Right. It's like your private place and, you know, it's– that's– it's your stuff.
GPT-2: It's a very sad thing.
GPT-3: It sounds like you're really upset with her because she invaded your privacy.

On-Topic But Unverifiable} as the error categories, and a synthetic reflection may suffer from one or more categories of error. Most of these categories require a deeper understanding of dialogue context, and we note that {Dialogue-Contradicting, Parroting, On-Topic But Unverifiable} have not been explicitly included in previous human evaluation studies for reflection generation.

Then, we carry out human evaluation based on this scheme (§6.3). We recruit a group of MI experts and a group of laypeople as annotators and analyse their evaluation[2] quality (Figure 6.1). The workload of each annotator consists of mixed human reflections from actual therapists and synthetic reflections produced by LMs (GPT-2 [54] and GPT-3 [17][3]), and the annotator is not informed of the source of any reflection. For each reflection, the annotator decides whether it is coherent as a Yes/No binary choice. If "No" is chosen, the annotator proceeds to select one

[2] Data available at https://github.com/uccollab/expert_laypeople_reflection_annotation.

[3] We also conducted human evaluation of reflections generated by BART [52], but the results are not included in the main body as the model failed to generate sufficiently diverse reflections (Appendix A.1)

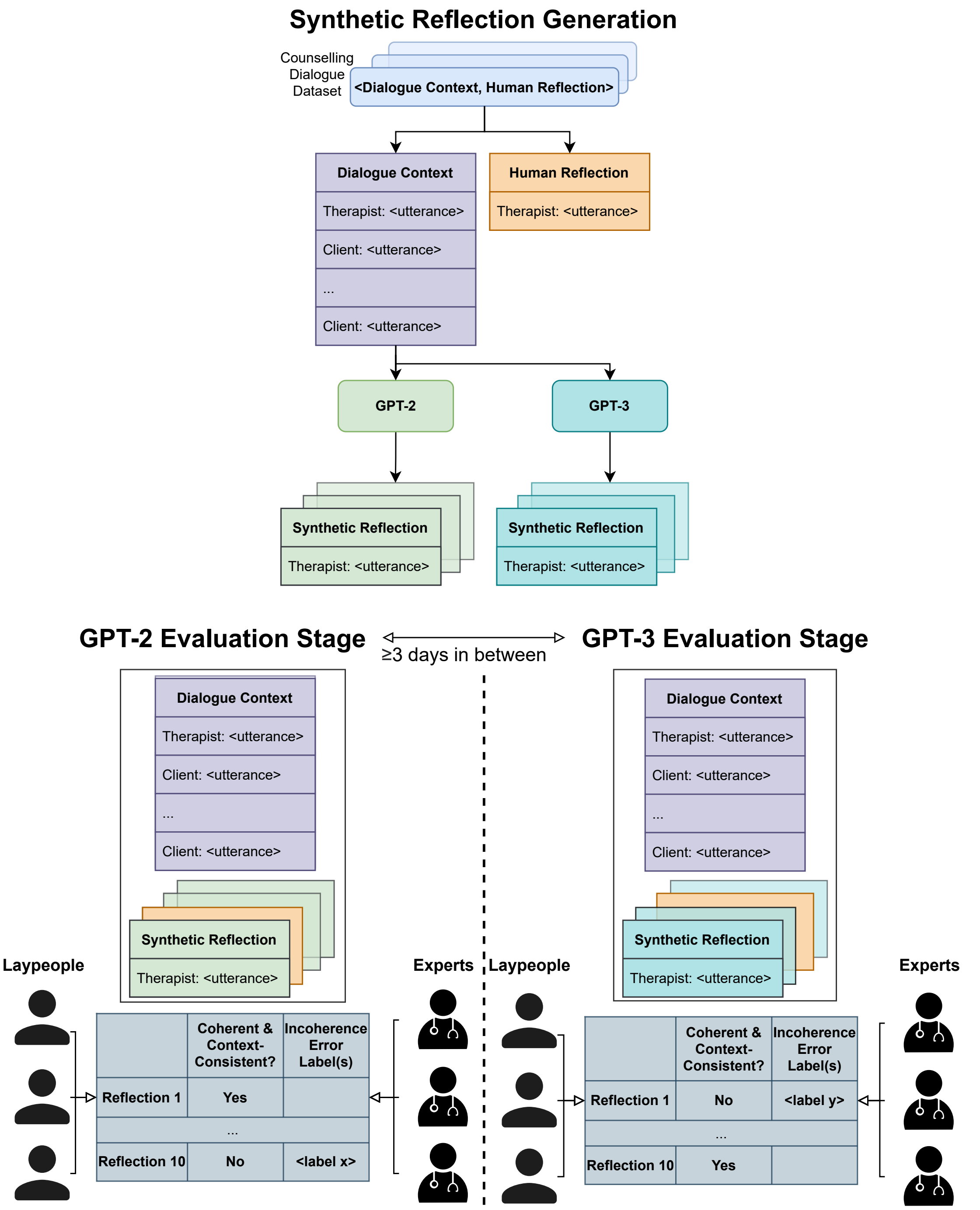


Figure 6.1: Human evaluation overview. The same human reflections are included in both evaluation stages, mixed with GPT-2 reflections in the GPT-2 stage and with GPT-3 reflections in the GPT-3 stage.

or more applicable incoherence error categories. In doing so, our evaluation goes beyond a binary Yes/No scheme and sheds light on the types of incoherence errors made by reflection generators, especially in settings with long dialogue contexts.

Based on the human evaluation results, we conduct in-depth analysis of intra-group agreement among laypeople and among experts, as well as the inter-group correlation between laypeople and experts. We also explore whether more powerful LMs produce more coherent synthetic reflections and how they affect evaluation of human reflections. We find that:

I Both laypeople and experts are reliable annotators based on their intra-group agreements on binary coherence evaluation. They also show moderate to strong inter-group correlation.

II Human reflections are more often annotated as coherent than GPT-2 reflections, but it is not the case with the more powerful GPT-3. Interestingly, both laypeople and experts are less likely to annotate a human reflection as coherent when it is mixed with GPT-3 reflections (than when it is mixed with GPT-2 reflections), though experts are relatively more consistent in this regard.

I represents the first evidence that laypeople are capable of coherence evaluation for reflection generation. II poses an interesting research question on whether synthetic reflections from large LMs can match or outperform human reflections on dimensions deeper than coherence, such as empathy.

6.2 Developing Coherence Evaluation Scheme

We develop the coherence evaluation scheme in 3 steps:

1. Generating synthetic reflections with LMs (§6.2.1).

2. Collecting free-text descriptions from laypeople about incoherence errors in synthetic reflections (§6.2.2).

3. Summarising free-text descriptions into discrete error categories (§6.2.3).

For Step 2, we consistently use the terms “annotations”, “annotators”, and “annotate” to describe the results, participants, and action of the laypeople, in order to be in line with the literature. However, this step is not and should not be confused with the human evaluation by laypeople and experts detailed in §6.3 where we also use these terms.

Table 6.2: A 3-turn context and its gold-standard human reflection from an MI dialogue in `AnnoMI`.

Context
Client: The baby was up all night and I'm exhausted. **Therapist**: So, what you're saying is you've had a rough night? **Client**: Yes. She was up every three hours to eat, I don't understand it.

Gold-Standard Human Response (Therapist Reflection)
So, she needed to eat every three hours last night and that was really frustrating for you?

6.2.1 Generating Synthetic Reflections

Counselling Dialogue Data: `AnnoMI`

We utilise `AnnoMI` ([3], Chapter 4) to train therapist response generators. Aiming at generating responses that a **good** therapist would say, we leverage the 110 conversations of high-quality MI with 8839 utterances in total (Table 4.1), where 28% (1256) of the 4441 therapist utterances are reflections (Figure 4.3). We refer to the 4441 gold-standard therapist utterances in those conversations as **human responses** and generated therapist utterances as **synthetic responses**.

Input/Output Format

Like most open-domain dialogue models, our response generators output a response given an N-turn dialogue history (i.e., context) where the last turn comes from the client. An illustrative 3-turn context and its human response from the dataset are shown in Table 6.2.

In practice, for each human response, we concatenate its preceding utterances and keep the rightmost (i.e., temporally most recent) 384 tokens as the **context**, which contains 14 previous turns on average. Notably, this is 3 times the context size used in previous work ($\leq$ 5 turns), as we assume richer context enables better response generation. For each human response, we keep the first 128 tokens. Thus, we construct 4441 $\langle$context, human response$\rangle$ pairs, of which 1256 are $\langle$context, human reflection$\rangle$ pairs.

As the volume of `AnnoMI` reflections is relatively small, we fine-tune LMs on all $\langle$context, human response$\rangle$ pairs instead of only on $\langle$context, human reflection$\rangle$ pairs, so that the models are trained to be general-purpose therapist response generators. The underlying assumption is that this will enable more training data for the language modelling of therapy dialogue while better shaping the boundaries of reflections in the latent semantic space.

Specifically, we train the models to generate all types of therapist responses (i.e., Reflection, Question, Input, Other) by using the type of the human response as a plain-text conditioning code, inspired by recent work (e.g., [177]) of similar approaches. Concretely, we construct the input as a sequence of context utterances with interlocutor labels and utterance separators, appended by the human response type for conditioning. For example, the context in Table 6.2 would become[4]:

> "⟨*client*⟩ *The baby was up all night and I'm exhausted.*|⟨*therapist*⟩ *So, what you're saying is you've had a rough night?*|⟨*client*⟩ *Yes. She was up every three hours to eat, I don't understand it.*|⟨*therapist*⟩~⟨*listening*⟩"

while the human response is simply

> *"So, she needed to eat every three hours last night and that was really frustrating for you?"*

Thus, a training/validation/test example is simply a ⟨context, human response⟩ pair representing the $\langle input, output \rangle$.

Training Response Generator

We train similarly sized `gpt2-medium` ([54], 355M parameters) and `bart-large` ([52], 406M parameters) into response generators. We do not use pre-trained dialogue models like DialoGPT [152] since they are mostly pre-trained on written conversations with only a few turns as the context, whereas therapy dialogues are spoken and long, which entails a large domain gap.

For training, we first divide the 110 high-quality MI dialogues into 10 folds, which means each ⟨context, human response⟩ pair is effectively assigned to a fold based on the dialogue it is from. In this process, we also minimise the size difference between different folds w.r.t. the quantity of ⟨context, human response⟩ pairs included in a fold (rather than w.r.t. the quantity of dialogues).

Then, we fine-tune the same pre-trained model (GPT-2/BART) 10 times independently to generate synthetic responses for the pairs in each test fold. Each time when fine-tuning a model, we use 8 folds as training data, 1 as validation data and 1 as test data. Since ⟨context, human response⟩ pairs from the same dialogue are always in the same fold, there is no overlap between training/validation/test data.

Our experiments are based on the HuggingFace package[5]. We use 2e-5 as the learning rate for training, based on a hyperparameter search over different learning rates where the metric is perplexity (the lower the better). The other hyperparameters are fixed, including 8 as the batch size and 42 as the random seed. The fine-tuning stops when perplexity has not improved on validation data for 3 epochs. We ran the fine-tuning on an NVIDIA V100 GPU (16GB). In total, the fine-tuning and inference took under 50 GPU hours.

[4] In practice, we use "⟨asking⟩", "⟨informing⟩", "⟨listening⟩", "⟨other⟩" as the plain-text control codes for Question, Input, Reflection and Other, respectively.

[5] `https://huggingface.co/`

Table 6.3: Perplexity of each response generator on human reflections.

Model	GPT-2	BART
Perplexity	17.36	13.29

Following most recent studies ([69], [70], *inter alia*) in response generation, we report in Table 6.3 the perplexity of each model on reflections, which quantifies how uncertain a model is about generating the human reflections in the test data. We do not compare these numbers with other studies because 1) achieving state-of-the-art is not our focus, 2) the dataset and task are unique and have no comparable state-of-the-art, and 3) to the best of our knowledge, there is no study on the usefulness of perplexity as a metric for reflection generation or counselling dialogue modelling. We also experimented with paraphrasing-based data augmentation, but it did not lead to significant improvement.

Test-Time Reflection Generation

Once the models are trained, we use them to generate synthetic reflections for the context in each ⟨context, human reflection⟩ pair, by conditioning the output using the ⟨*listening*⟩ code as detailed before.

Following recent work (e.g., [182]) on hallucination in dialogue generation, we experiment with a range of decoding strategies, in order to capture a broad spectrum of potential errors in synthetic reflections. For both GPT-2 and BART, we use

- Greedy decoding
- 5-Beam decoding, using all of the 5 decoded sequences at the final time step
- Nucleus decoding [119], $p \in \{0.4, 0.6, 0.8, 0.95\}$, 5 sequences sampled for each p

6.2.2 Collecting Incoherence Error Descriptions

Since we hypothesise that incoherence errors can be spotted by non-experts, we survey laypeople for their own descriptions of incoherence errors in synthetic reflections, so that we can later summarise those descriptions into discrete categories (§6.2.3).

Annotators We recruit 6 volunteers with high proficiency in English and no prior experience in NLP or psychology/psychotherapy.

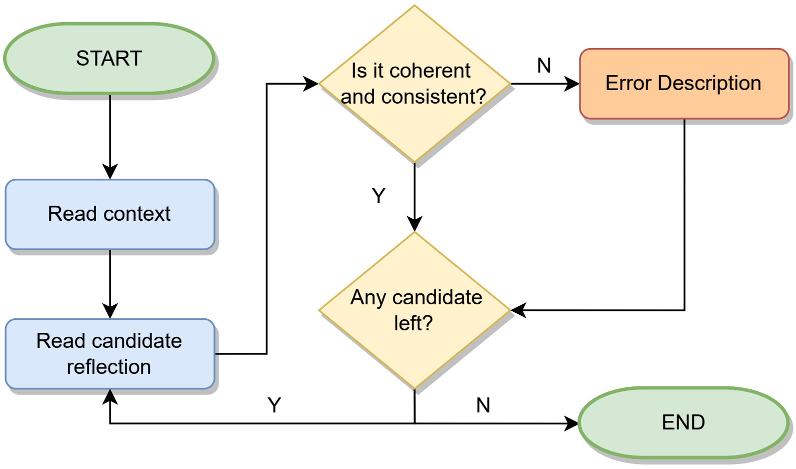


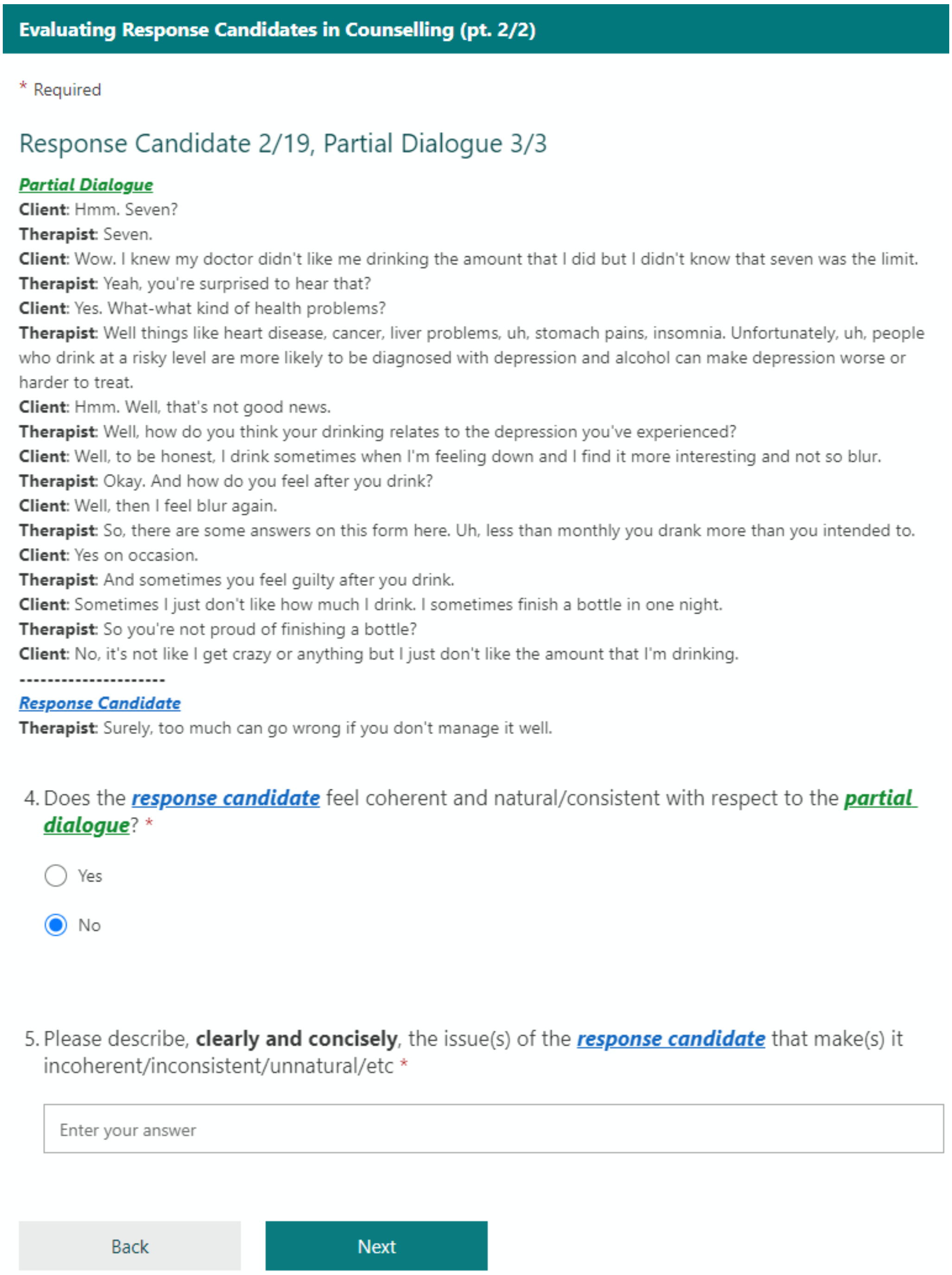
Figure 6.2: Annotation flow for candidate reflections of one context, during collection of incoherence error descriptions.

Annotation Workload We sample 3 ⟨context, human reflection⟩ pairs from 3 different `AnnoMI` dialogues and use their human and synthetic reflections for annotation. Overall, the annotation workload consists of 60 reflections in total for the 3 contexts, and the workload is the same for each annotator.

Annotation Procedure The procedure is illustrated in Figure 6.2, and the annotation interface is presented in Figure 6.3. The annotators are shown one $\langle context, reflection \rangle$ pair and first need to answer whether the reflection feels coherent/consistent given the context. If they choose "No", they are asked to describe the error(s) of the reflection that cause(s) incoherence/inconsistency, otherwise they will proceed to the next example. We note that we do not define "incoherent" or "inconsistent" and instead leave it to the discretion of the annotators, in order to gather more natural insights on incoherence errors. For the same reason, we use the phrase "response candidate" in the annotation interface instead of the more complex term "reflection". We do not mention the source of any response candidate.

IAA The IAA on the "Coherent & Consistent?" question is 0.37 in terms of Fleiss' kappa [155], which is in the "fair agreement" range (0.2-0.4) [166] but close to the "moderate agreement" threshold of 0.4 [166]. We attribute the relatively low IAA to two factors:

- We purposefully did not provide a strict definition of "coherence" or "consistency" to the annotators, which led some of them to consider issues like "intimidating tone" as reasons for incoherence/inconsistency, but those problems are about "appropriateness" and are thus not related to incoherence/inconsistency. Instead, the "appropriateness" dimension should be left to experts-based human evaluation on the overall quality of reflections.



Evaluating Response Candidates in Counselling (pt. 2/2)

* Required

Response Candidate 2/19, Partial Dialogue 3/3

Partial Dialogue
Client: Hmm. Seven?
Therapist: Seven.
Client: Wow. I knew my doctor didn't like me drinking the amount that I did but I didn't know that seven was the limit.
Therapist: Yeah, you're surprised to hear that?
Client: Yes. What-what kind of health problems?
Therapist: Well things like heart disease, cancer, liver problems, uh, stomach pains, insomnia. Unfortunately, uh, people who drink at a risky level are more likely to be diagnosed with depression and alcohol can make depression worse or harder to treat.
Client: Hmm. Well, that's not good news.
Therapist: Well, how do you think your drinking relates to the depression you've experienced?
Client: Well, to be honest, I drink sometimes when I'm feeling down and I find it more interesting and not so blur.
Therapist: Okay. And how do you feel after you drink?
Client: Well, then I feel blur again.
Therapist: So, there are some answers on this form here. Uh, less than monthly you drank more than you intended to.
Client: Yes on occasion.
Therapist: And sometimes you feel guilty after you drink.
Client: Sometimes I just don't like how much I drink. I sometimes finish a bottle in one night.
Therapist: So you're not proud of finishing a bottle?
Client: No, it's not like I get crazy or anything but I just don't like the amount that I'm drinking.

Response Candidate
Therapist: Surely, too much can go wrong if you don't manage it well.

4. Does the ***response candidate*** feel coherent and natural/consistent with respect to the ***partial dialogue***? *

○ Yes

● No

5. Please describe, **clearly and concisely**, the issue(s) of the ***response candidate*** that make(s) it incoherent/inconsistent/unnatural/etc *

Enter your answer

Figure 6.3: Annotation interface for one reflection candidate, during collection of incoherence error descriptions.

- 6 annotators are involved in the annotation process rather than just 2 to 3 as is commonly done for human evaluation of generated reflections [35, 36], and it is usually less likely to get higher agreement with more annotators.

6.2.3 Establishing Error Categories

We use thematic analysis [180] to manually and systematically identify common patterns in the annotators' incoherence error descriptions and summarise those patterns into the following error categories:

- **Malformed**: a response that "feels broken" because 1) it has unclear references, 2) it is incomprehensibly ungrammatical, and/or 3) its sentences are issue-free on their own but confusing when combined.
- **Dialogue-Contradicting**: a response that contradicts the context, either partially or fully.
- **Parroting**: a response that repeats a certain part of the context in an unnatural way.
- **Off-Topic**: a reply that has little to no relevance to the dialogue.
- **On-Topic But Unverifiable**: an on-topic reply that cannot be verified based on the context alone.

For illustrative examples of the categories, see Table 6.4.

Other Considerations Good reflections sometimes repeat something that the client has said, for example to affirm it, but those are natural and good practices rather than unnatural repetition (Parroting). Also, broadly speaking, Dialogue-Contradicting, Off-Topic, and On-Topic But Unverifiable reflections are all unfaithful and ungrounded w.r.t. the context, making them all manifestations of hallucination. Finally, we note that a small percentage ($\approx 8\%$) of error descriptions do not contain sufficient information (e.g., "Doesn't feel like a natural response") and are therefore excluded from the thematic analysis.

Table 6.4: Illustrative examples for error categories.

Context
Therapist: Mm. So, embraces this life of the picket fence and the, you know, the little house and having the baby in this- **Client**: Yeah, suburbs. Yeah. **Therapist**: So, you don't wanna be that- that- that woman that's just sort of living in this house with the baby carriage, and at the same time, that world's kind of saying it's behind you. **Client**: Yeah, I-I guess. Like, I don't wanna- I don't wanna lose my friends because of this. And I don't wanna put pressure on them to change because they're, I mean, they're good people. They're just– **Therapist**: You care about them. **Client**: Yeah. **Therapist**: It's just the scene that they're-they're still in because their life maybe hasn't taken that change that your life has made. And so now, you're here and you're not sure. You want the baby, but you're not sure about the suburbs and the picket fence thing. You're not sure– **Client**: I don't want that. **Therapist**: You don't want that. **Client**: No. **Therapist**: You want this baby. **Client**: Yeah. **Therapist**: And you wanna be, um, take care of your baby as the baby's growing. **Client**: Like, I wanna- I wanna be responsible for this thing. **Therapist**: You don't wanna let your baby down. **Client**: No.
Gold-Standard Human Reflection (Coherent & Consistent)
And at the same time, it's hard to let go of the past, but on the other hand, you're thinking, "I don't have a choice. I've got to look this way." Is that kind of what it feels like?
Malformed
You-you wanna be the one that, um–you're the one who's gonna pick up the phone, uh, uh–you wanna pick up and, uh –or call them, um –to-to-unquote-the-mother and-to tell them that your-your daughter is going to come home and she's, uh...
Dialogue-Contradicting
OK, so what I'm hearing you say is that taking care of a baby would be too much responsibility for you right now, so you are not ready to have a baby.
Parroting
Mm-hmm. So, you don't want to let your baby down.
Off-Topic
So you do not wanna allow your kids to go across the street without you.
On-Topic But Unverifiable
You've really wanted to be a mom since you were a little girl.

6.3 Human Evaluation: Laypeople & Experts

Based on the evaluation scheme developed in §6.2, we carry out human evaluation of reflection coherence with both laypeople and experts. We detail how we generate synthetic reflections in §6.3.1, describe the human evaluation process in §6.3.2, and conduct thorough analyses of evaluation results in §6.3.3, §6.3.4 and §6.3.5.

To be in line with the literature, we consistently use the terms "annotations", "annotators", and "annotate" to describe the results, participants, and action of human evaluation, but the evaluation process is not and should not be confused with the previous step of collecting incoherence error descriptions from laypeople (§6.2.2) where we also used these terms.

6.3.1 Generating Synthetic Reflections

We use both GPT-2 and GPT-3[6] to generate synthetic reflections for human evaluation. The use of GPT-3 is in light of the impressive generative capabilities of LLMs shown recently ([26, 27], *inter alia*) including for reflection generation [37]. For GPT-2, we reuse the synthetic reflections from §6.2.1. For GPT-3, we use few-shot prompting to generate synthetic reflections, as detailed in the rest of this section (§6.3.1).

To generate from GPT-3, We use the default temperature (1.0) and $p \in \{0.4, 0.6, 0.8, 0.95\}$ for decoding. We model our prompt as asking GPT-3 to read a series of ⟨context, human reflection⟩ pairs (learning examples) and then to complete a final dialogue context where the reflection is missing (test example).

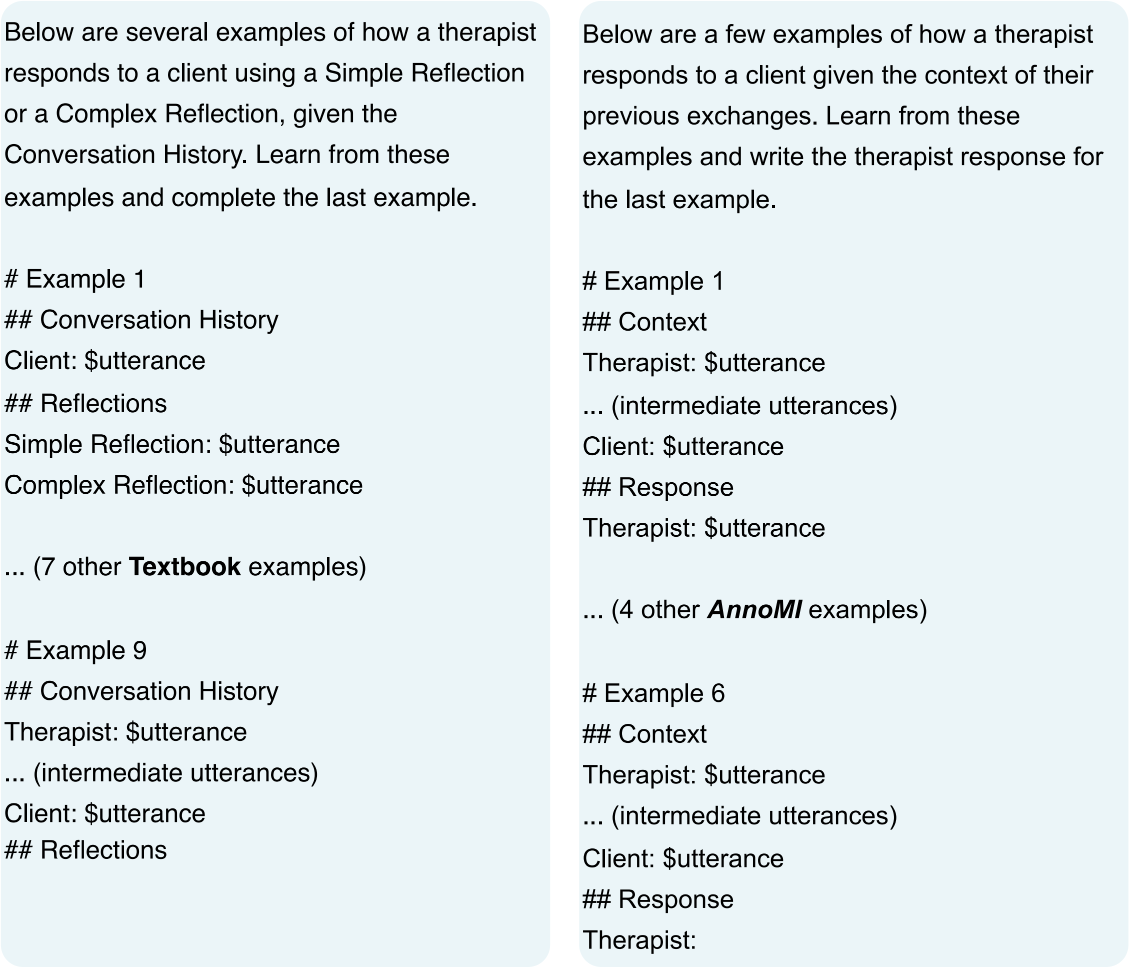
The test example is always a dialogue context from `AnnoMI`, but we explore two sources of learning examples — **textbook** and `AnnoMI` — to diversify the generation. **Textbook** examples (Figure 6.4a) are taken from the MITI coding manual [183], while `AnnoMI` examples (Figure 6.4b) are simply the ⟨context, human reflection⟩ pairs that we constructed previously. Each textbook example consists of a client statement — which we use as dialogue context — along with a simple reflection and a complex one, where the complex reflection adds more meaning/emphasis to the client statement than the simple one [16] (§4.2.5).

Prompting with Textbook Examples

As learning examples, textbook examples are different from `AnnoMI` examples in that 1) textbook examples are written texts instead of transcripts like `AnnoMI`, and 2) the context in a textbook example is considerably shorter than the average `AnnoMI` context which contains 14 utterances.

A prompt (Figure 6.4a) begins with an instruction, followed by 8 textbook examples and the test example placed at the end. Thus, the model is prompted to

[6] We use `text-davinci-002`, the largest GPT-3 model (175B parameters) at the time of experiment. The total cost of generation was 23.68 US Dollars.



Below are several examples of how a therapist responds to a client using a Simple Reflection or a Complex Reflection, given the Conversation History. Learn from these examples and complete the last example.

```
# Example 1
## Conversation History
Client: $utterance
## Reflections
Simple Reflection: $utterance
Complex Reflection: $utterance

... (7 other Textbook examples)

# Example 9
## Conversation History
Therapist: $utterance
... (intermediate utterances)
Client: $utterance
## Reflections
```

(a) Using textbook examples.

Below are a few examples of how a therapist responds to a client given the context of their previous exchanges. Learn from these examples and write the therapist response for the last example.

```
# Example 1
## Context
Therapist: $utterance
... (intermediate utterances)
Client: $utterance
## Response
Therapist: $utterance

... (4 other AnnoMI examples)

# Example 6
## Context
Therapist: $utterance
... (intermediate utterances)
Client: $utterance
## Response
Therapist:
```

(b) Using AnnoMI examples.

Figure 6.4: Prompting formats.

generate 2 synthetic reflections, one simple and the other complex. Considering recent studies (e.g., [184]) about the impact of few-shot example ordering on the output, we create 3 prompts to generate 3 different sets of {simple reflection, complex reflection}, where the textbook examples in each prompt are identical but in different random orders.

Prompting with AnnoMI Examples

In this prompting method, we do not take simple/complex reflection into account because, while human reflections in AnnoMI do have such labels, the IAA on reflection types is not sufficiently high (Table 4.10). Similar to prompting with textbook examples, we construct 3 prompts for each test example to obtain diverse GPT-3-generated reflections. The difference, however, is that we create those 3 prompts by

Table 6.5: Overview of Annotation Workload.

Each batch contains	1 dialogue context, 1 human reflection, N synthetic reflections
GPT-2 stage	
Each layperson/expert has	5 batches
Each reflection annotated by	3 laypeople, 3 experts
Synthetic reflections per batch (N)	7.13 on average
Total batches	15
Total human reflections	15
Total synthetic reflections	107
GPT-3 stage	
Each layperson/expert has	5 batches
Each reflection annotated by	3 laypeople, 3 experts
Synthetic reflections per batch (N)	9 (except one batch with 7)
Total batches	15
Total human reflections	15
Total synthetic reflections	133

sampling 3 different sets of learning examples instead of shuffling. Therefore, the learning example set in each of the 3 prompts is unique. We note that the learning examples in a prompt come from different dialogues, and, to ensure fairness, no learning example is from the dialogue that the test example is from.

6.3.2 Evaluation Procedure

We recruit 2 groups of annotators:

- 9 laypeople known to us and with no experience in MI, none of whom participated in the evaluation scheme development (§6.2).
- 9 experts found through professional networks, in particular the Motivational Interviewing Network of Trainers[7], an international organisation of MI trainers and a widely recognised MI authority.

Workload

Table 6.5 presents the annotation workload overview.

[7] https://motivationalinterviewing.org/

To create annotation materials, we randomly sample 15 ⟨context, human reflection⟩ pairs from 15 `AnnoMI` dialogues. For the context in each pair, we generate (§6.3.1) 9 semantically diverse synthetic reflections[8] with GPT-3 and 7.13 on average[9] with GPT-2. Thus, for each ⟨context, human reflection⟩ pair, we create 2 annotation batches that each contain the context, the human reflection and synthetic reflections. The two batches differ in that the synthetic reflections in one batch come from GPT-2 while those in the other batch are from GPT-3.

Each annotator is first randomly assigned 5 batches where the synthetic reflections are from GPT-2 (**GPT-2 stage**). After completion of these batches and then a waiting period of at least 3 days (Appendix A.2), the annotator is randomly assigned 5 more batches where the synthetic reflections are from GPT-3 (**GPT-3 stage**). The task ends when the annotator has finished all 10 batches. Overall, each batch is randomly assigned to 3 laypeople and 3 experts, resulting in each reflection being annotated 3 times by laypeople and 3 times by experts.

Annotating One Batch

When annotating one batch (Figure 6.1), the annotator first reads the context and then sequentially and iteratively annotates all the reflections. The reflections in each batch are shuffled, and the annotator is not informed of the source of any reflection.

For each reflection, the annotator chooses Yes/No regarding whether it is coherent. If the answer is No, the annotator selects one or more applicable error categories from the error annotation scheme that we developed previously in §6.2.3. Prior to annotation, the annotator reads a mandatory tutorial about coherence and consistency with examples for each error category, and it remains accessible throughout the annotation process.

Cross-Stage Human Reflection Recurrence

Due to random batch assignment, an annotator may annotate one batch in the GPT-2 stage and another batch in the GPT-3 stage where the two batches share the same ⟨context, human reflection⟩. For the annotator in such cases, the shared human reflection is **recurring** across stages, and hence the annotator annotates it twice. To make it less likely that an annotator annotates a recurring human reflection in the GPT-3 stage based on how they recall annotating it in the GPT-2 stage, each annotator waits for at least 3 days[10] between completing their

[8] There is one context with 7 instead of 9 GPT-3 reflections due to lack of semantic diversity among generated candidates.

[9] In practice, GPT-2 and BART reflections were evaluated together, and their combined size is the same as GPT-3 reflections'. Thus, there are fewer GPT-2 reflections than GPT-3 reflections. We exclude BART reflections from the GPT-2 stage for fairness considerations. More details in Appendix A.1.

[10] More details in Appendix A.2.

Table 6.6: Global agreement on *Coherent*/*Incoherent* binary choice (intra-group agreement).

	Laypeople		**Experts**	
	GPT-2 Stage	*GPT-3 Stage*	*GPT-2 Stage*	*GPT-3 Stage*
Fleiss' kappa	0.42	0.23	0.44	0.04
Randolph's kappa	0.42	0.30	0.45	0.42

last batch in the GPT-2 stage and starting their first batch in the GPT-3 stage.

6.3.3 Agreement Among Annotators

Once the human evaluation is completed, we measure the agreement among the annotators in 2 ways:

- **Intra-Group Agreement** among laypeople and among experts: how much the annotators of the same group (laypeople/experts) agree with each other.

- **Inter-Group Correlation**: correlation between laypeople and experts w.r.t. their annotations.

Intra-Group Agreement on *Coherent* and *Incoherent*

We first analyse the intra-group agreement on the global binary Yes/No (*Coherent*/*Incoherent*) annotation. We adopt both the classical Fleiss' kappa [155] and Randolph's fixed-marginal kappa [185], because 1) Fleiss' kappa is known to be overly penalising when the marginal label distribution is imbalanced [186, 187] and 2) Randolph's kappa is preferable when the annotators have no prior knowledge of the expected label distribution [188].

As Table 6.6 shows, Fleiss' kappa in the GPT-2 stage indicates moderate agreement [166] for both annotator groups, but in the GPT-3 stage it drops to fair agreement for laypeople and almost zero for experts. The drop may appear to suggest a drastic change in agreement, but deeper analysis reveals a considerable cross-stage change of marginal label distribution that may skew Fleiss' kappa — for example, experts annotate GPT-3 reflections as *Coherent* 82% of the time (§6.3.4) as opposed to 38% for GPT-2 reflections. As an evidence, Randolph's kappa, which is not influenced by marginal label distribution, still shows (Table 6.6) fair agreement among laypeople and moderate agreement among experts in the GPT-3 stage.

Beyond global agreement, we conduct more granular analysis on which one of {*Coherent*, *Incoherent*} is easier to agree upon. Specifically, we follow [186] to calculate the **per-label majority agreement ratio** (referred to as **agreement ratio** for brevity) for *Coherent* and *Incoherent* separately. For a label l, its agreement

Table 6.7: Per-label majority agreement ratios on *Coherent* and *Incoherent* separately (intra-group agreement).

	Laypeople		**Experts**	
	GPT-2 Stage	*GPT-3 Stage*	*GPT-2 Stage*	*GPT-3 Stage*
Coherent	0.69	0.76	0.66	0.90
Incoherent	0.71	0.51	0.75	0.25

ratio $A^M(l)$ is:

$$A^M(l) = \frac{\#(\text{reflections assigned } l \text{ by 2 annotators})}{\#(\text{reflections assigned } l \text{ by} \geq 1 \text{ annotators})}$$

For example, the agreement ratio of *Coherent* is the number of reflections annotated as *Coherent* by 2 out of 3 annotators (hence majority agreement) divided by the number of reflections annotated as *Coherent* by any annotator.

As Table 6.7 shows, the agreement ratio of *Incoherent* has a minor lead over that of *Coherent* in the GPT-2 stage. In the GPT-3 stage, however, *Coherent* shows substantially higher agreement ratio than *Incoherent*. Therefore, as the LM grows in power (GPT-2→GPT-3), it becomes easier for annotators to agree on what is *Coherent* than on what is not, and this applies to both laypeople and experts.

We note that, in [186], an example is given label l if the agreement ratio of l is above 0.3 and a majority of annotators assign l to the example. Our results show that both laypeople and experts have agreement ratios that are almost always comfortably higher than the 0.3 threshold, particularly w.r.t. *Coherent* (0.66∼0.90). Thus, also considering the global agreement results (Table 6.6), both laypeople and experts appear to be reliable annotators, and a reflection should be considered *Coherent* if a majority of annotators deem it so.

Intra-Group Agreement on Error Categories

To further investigate intra-group agreement, we also measure agreement ratio for each error category to inspect whether some errors are easier than others for annotators to agree upon.

Based on Tables 6.7 and 6.8, one can observe that agreement ratio is generally higher for *Incoherent* than for any error category. While it may be inherently more challenging to annotate an error category than to annotate *Coherent*/*Incoherent* due to the label space size difference (5 vs. 2), this is still a strong indication that it is easier for annotators to agree that a reflection is *Incoherent* than to agree upon any specific incoherence problem.

Interestingly, Parroting has a clearly higher agreement ratio among laypeople than among experts in both stages, which means some experts are more tolerant of Parroting than others but laypeople are similar to each other in this regard.

Table 6.8: Per-label majority agreement ratios for error categories (intra-group agreement). *Italic*: less than 10 reflections are given this error category by any annotator.

	Laypeople		**Experts**	
	GPT-2 Stage	*GPT-3 Stage*	*GPT-2 Stage*	*GPT-3 Stage*
Parroting	*0.38*	0.45	*0.00*	0.11
Malformed	0.47	*0.00*	0.37	0.00
Off-Topic	0.35	0.00	0.55	*0.00*
Dialogue-Contradicting	0.34	0.16	0.24	0.30
On-Topic But Unverifiable	0.20	0.23	0.29	0.12

Table 6.9: Inter-group correlations between laypeople- and experts-based coherence scores. $p <$ 1e-7 for all 4 values.

	Spearman	**Pearson**
GPT-2 Stage	0.741	0.742
GPT-3 Stage	0.444	0.446

Inter-Group Correlation

We measure inter-group correlation based on **coherence scores**: Given a reflection and the 3 annotators to whom it was assigned, its coherence score is the number of annotators that annotated it as *Coherent*. Thus, a coherence score has a range of $\{0, 1, 2, 3\}$, and each reflection has one score from laypeople and one from experts.

As Table 6.9 shows, inter-group correlation is strong in the GPT-2 stage and moderate in the GPT-3 stage [189]. Combined with our previous findings on the intra-group agreement on coherence, this is further evidence that laypeople can be a viable alternative to experts for scaled-up reflection coherence evaluation. In particular, a binary *Coherent*/*Incoherent* setup may be more suitable, since per-label majority agreement ratios are clearly higher on *Coherent* and *Incoherent* than on the error categories, as we found previously. Nevertheless, the weaker inter-group correlation in the GPT-3 stage does suggest experts-laypeople differences (we probe them further in §6.3.4), and it also shows that laypeople-based evaluation is relatively more challenging when the reflections come from powerful LLMs.

6.3.4 Cross-Stage Annotation Differences

We further investigate how reflections — both human and synthetic ones — are annotated differently in different stages. In particular, we focus on the distribution of *Coherent*/*Incoherent* labels and error labels based on the results in Figure 6.5.

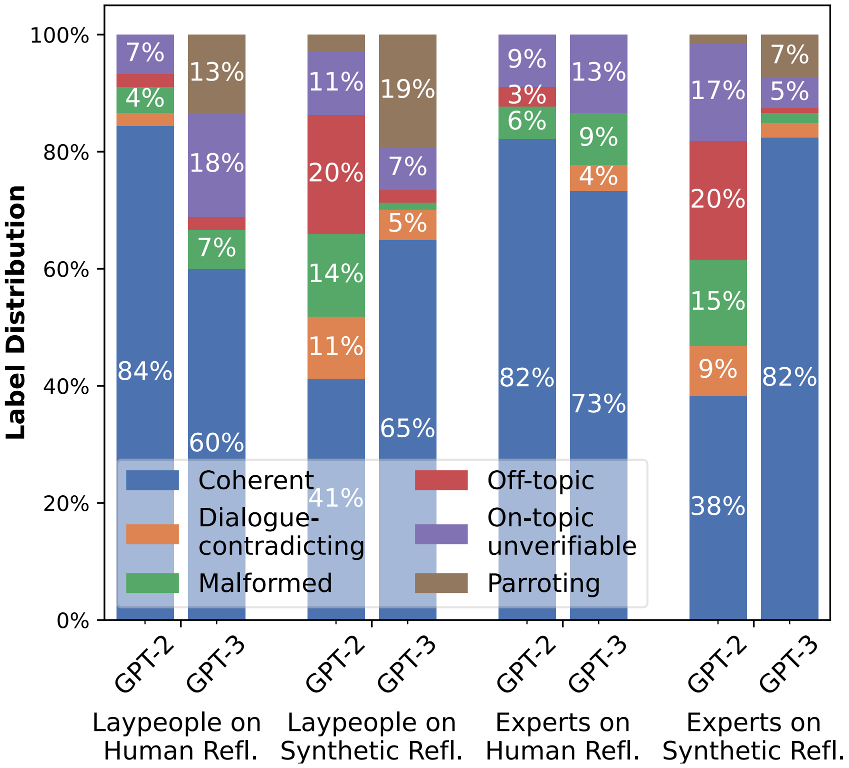


Figure 6.5: Labels distribution on human and synthetic reflections in the GPT-2 stage and GPT-3 stage. *Incoherent* labels are broken down into fine-grained error categories.

Human vs. Synthetic in *Coherent* Rate

We compare human and synthetic reflections w.r.t. proportion of *Coherent* labels[11]. As shown in Figure 6.5, human reflections are annotated as *Coherent* significantly (chi-squared test, $p < 0.05$) more often than synthetic reflections by both laypeople and experts in the GPT-2 stage. This is not unexpected, since human reflections are considered the gold standard. However, the trend is reversed in the GPT-3 stage, even though the lead of GPT-3 over human reflections is not significant. This shows that GPT-3 is capable of producing coherent reflections, and it can even sometimes match or outperform human reflections. It also raises interesting research questions on whether GPT-3 (and more advanced LLMs, e.g., GPT-4 [28]) can compete with human reflections on aspects deeper than coherence, such as empathy and adherence to counselling principles.

[11] We do not compare at the granular error-category-level due to the different scales of human and synthetic reflections.

Cross-Stage Differences on Synthetic Reflections

As shown in Figure 6.5, GPT-3 reflections are significantly (chi-squared test, $p < 0.05$) more often annotated as *Coherent* than GPT-2 ones by both laypeople and experts, which is not surprising given that GPT-3 is considerably more powerful. Interestingly, while laypeople and experts are similar in *Coherent*/*Incoherent* label distribution for GPT-2, experts are significantly (chi-squared test, $p < 0.05$) more likely than laypeople to annotate GPT-3 reflections as *Coherent*.

Upon further analysis, we notice that much of the laypeople-experts divide on GPT-3's *Coherent* rate can be attributed to Parroting, which is used 19% of the time by laypeople but only 7% by experts. For the other 4 error categories, laypeople are experts behave similarly: the proportion of each category is substantially lower in the GPT-3 stage. This shows that GPT-3 makes most types of incoherence errors less often than GPT-2.

Overall, it is clear that experts are less strict than laypeople about Parroting on synthetic reflections in both stages. This is likely because a reflection summarises what the client has said, which may sometimes appear repetitive to a layperson when an expert may consider it good practice. As further evidence, we note that human reflections, which showcase good practice, are not annotated as Parroting by experts in either stage, while laypeople do so in the GPT-3 stage.

Cross-Stage Shifts of Annotator Groups on Human Reflections

As Figure 6.5 shows, both laypeople and experts annotate human reflections as *Coherent* less often in the GPT-3 stage than in the GPT-2 stage. Therefore, we analyse the distribution of *Coherent* and *Incoherent* labels given to human reflections and examine whether the cross-stage distribution shift is significant. We do so under 2 settings:

- **All**: Taking into account all the *Coherent* and *Incoherent* labels.

- **Recurrence-Free**: Removing the labels from an annotator for a human reflection if the reflection is recurring (§6.3.2) for the annotator (i.e., the annotator annotated the reflection in both stages) and thereby removing potential recurrence-caused annotator bias.

As shown in Table 6.10, under both All and Recurrence-Free, both laypeople and experts less often annotate human reflections as *Coherent* in the GPT-3 stage. Notably, the shift of laypeople is significant, while the shift of experts is not.

Besides the distribution of *Coherent* and *Incoherent* labels, we also inspect cross-stage shift w.r.t. coherence scores (defined in §6.3.3 to compute inter-group correlation) of human reflections. With the paired Wilcoxon signed-rank test, we have a similar finding: laypeople-based coherence scores are significantly ($p < 0.05$) lower in the GPT-3 stage than in the GPT-2 stage, but it is not the case for experts.

Table 6.10: *Coherent*/*Incoherent* label distributions for human reflections. We report how often (%) the annotators annotate a human reflection as *Coherent*. **Bold**: significant (chi-squared test, $p < 0.05$) cross-stage shift.

	All		**Recurrence-Free**	
	GPT-2 Stage	*GPT-3 Stage*	*GPT-2 Stage*	*GPT-3 Stage*
Laypeople	84%	**60%**	87%	**58%**
Experts	82%	73%	83%	77%

Also shown in Figure 6.5, human reflections are clearly more likely ($\Delta \geq 11\%$) to be annotated by laypeople as Parroting and On-Topic But Unverifiable in the GPT-3 stage than in the GPT-2 stage. In comparison, error annotations by experts for human reflections are more consistent across stages, with minor ($\Delta \leq 4\%$) increases in On-Topic But Unverifiable, Malformed and Dialogue-Contradicting.

Therefore, compared to experts, laypeople are overall more influenced by synthetic reflections when annotating human reflections. This annotation fluidity is a potential concern for laypeople-based scaled-up coherence evaluation.

Cross-Stage Shifts of Individual Annotators on Human Reflections

Beyond annotator-group-level annotation shifts, we also inspect whether or how each layperson/expert has shifting annotations on human reflections across stages. Since the workload of each annotator consists of **non-recurring** human reflections (appearing in only one stage) and sometimes also **recurring** human reflections (appearing in both stages), we probe the shift of each annotator on these two types of human reflections separately.

We first examine how often each annotator consistently assigns the same coherence label (Yes/No) to a **recurring** human reflection in both stages. As shown in Table 6.11, 8 laypeople and 8 experts have recurring human reflections in their workload. Among those annotators, 3 laypeople and 4 experts fail to consistently assign coherence labels to all (100%) recurring human reflections. Overall, laypeople and experts consistently assign coherence labels to recurring human reflections 71% and 73% of the time, respectively. Those similar numbers are evidence that the laypeople-experts difference in the between-phase waiting period duration (Appendix A.2) is not critical.

Then, we investigate whether each annotator annotates **non-recurring** human reflections more, equally, or less often as *Coherent* in the GPT-3 stage than in the GPT-2 stage. As Table 6.12 shows, 5 laypeople assign *Coherent* labels less often, 1 does so more often, while the other 3 stay at the same level in both stages. Among the experts, 4 give *Coherent* annotations less often, 2 do so more often, while the remaining 3 do not show cross-stage frequency change, which is a similar distribution compared to laypeople. Considering that laypeople and experts have different

Table 6.11: Overview of how often each layperson (L1∼L9) and each expert (E1∼E9) consistently assigns the same coherence label (Yes/No) to a **recurring** human reflection in both stages. **N/A**: annotator has no recurring human reflections in workload. **†**: annotator is not always consistent.

Each Annotator's Cross-Stage Annotation Consistency on *Recurring* Human Reflections			
L1	100%	**E1**	100%
L2	100%	**E2**	100%
L3	100%	**E3**	50%†
L4	N/A	**E4**	0%†
L5	100%	**E5**	50%†
L6	50%†	**E6**	N/A
L7	33%†	**E7**	100%
L8	100%	**E8**	100%
L9	50%†	**E9**	67%†
Overall	71%	**Overall**	73%

Table 6.12: Overview of how often each layperson (L1 - L9) and each expert (E1 - E9) annotates a **non-recurring** human reflection as *Coherent* in each stage. ↑/↓: increase/decrease in GPT-3 stage compared to GPT-2 stage.

How Often Each Annotator Annotates a *Non-Recurring* Human Reflection as *Coherent*					
	GPT-2 Stage	*GPT-3 Stage*		*GPT-2 Stage*	*GPT-3 Stage*
L1	100%	50%↓	**E1**	100%	100%
L2	100%	100%	**E2**	100%	75%↓
L3	100%	50%↓	**E3**	100%	67%↓
L4	100%	20%↓	**E4**	75%	100%↑
L5	100%	100%	**E5**	67%	67%
L6	67%	67%	**E6**	80%	60%↓
L7	100%	50%↓	**E7**	75%	100%↑
L8	60%	25%↓	**E8**	67%	67%
L9	67%	100%↑	**E9**	100%	50%↓
Overall	87%	58%↓	**Overall**	83%	77%↓

levels of overall cross-stage shift on non-recurring human reflections — -29% for laypeople and -6% for experts — we posit that laypeople and experts differ less in the proportion of "shifting" annotators but more in the magnitude of shifts displayed by individual annotators.

6.3.5 Case Study

To gain qualitative insights into the annotations, we show a case study in Table 6.13 which presents the annotations on the reflections shown in Table 6.1.

While the human reflection is annotated as *Coherent* by every layperson in the GPT-2 stage, it is flagged by 2 laypeople as Parroting in the GPT-3 stage, which may be because those 2 laypeople found the human reflection to be repeating of the last client utterance (e.g., "it's your stuff" in the human reflection compared to "it's my stuff" in the client utterance). In particular, Layperson 7 (L7) shows a *Coherent*→Parroting shift when going GPT-2 stage→GPT-3 stage. Notably, this example echoes the overall trend that human reflections are more likely (0%→13%) to be flagged by laypeople as Parroting in the GPT-3 stage (§6.3.4).

On the other hand, the human reflection is annotated as *Coherent* by every expert in the GPT-2 stage, but it is flagged by 1 expert as Malformed in the GPT-3 stage. We postulate that the fluency of GPT-3 reflections may make the human reflection appear less fluent to some annotators. This may be particularly true when there are faithfully transcribed self-repetitions and mid-sentence changes ("it's-that's-it's your stuff") in the human reflection, even though we explicitly informed the annotators that those are normal.

For comparison, we also analyse the annotations on the examples of GPT-2 and GPT-3 synthetic reflections. The GPT-2 reflection roughly matches the mood of the client but is also generic, and it is annotated as Off-Topic by 1 layperson and 2 experts. On the other hand, the GPT-3 reflection is fluent and more specific to the dialogue, and unsurprisingly it is annotated as *Coherent* by all 6 annotators. While those two reflections cannot cover all of the variety of synthetic reflections, their qualitative difference w.r.t. the human reflection is a good example for showing why annotators may be influenced by the surrounding synthetic reflections when they are annotating a human reflection.

6.4 Summary

In this work, we probed whether laypeople can be used as an alternative to experts in evaluating coherence and context-consistency for counselling reflection generation. As the first step, we conducted a preliminary study to create a coherence/consistency evaluation scheme for reflections, based on free-text error descriptions given by laypeople. Then, based on this scheme, we asked both laypeople and experts to annotate synthetic reflections from LMs and human reflections from actual therapists. We found that both laypeople and experts are reliable annotators and that they also show moderate to strong inter-group correlation, which is the first concrete evidence that laypeople are capable of this type of evaluation, although laypeople are relatively less aligned with experts on GPT-3 reflections. Furthermore, we found that GPT-3 is mostly able to generate coherent and con-

Table 6.13: Complete dialogue context of Table 6.1 and annotations on reflection examples. **L1/L2/.../L9**: 9 laypeople. **E1/E2/.../E9**: 9 experts. Red: incoherence error category. ‡: Annotator annotated the human reflection in both stages.

Context
Client: Well, I'm here because my mom wants me to be here.
Therapist: Mm-hmm.
Client: I don't really wanna be here, but it-it– whatever.
Therapist: Got it.
Client: Um, she-she found my stash-
Therapist: Uh-huh.
Client: -and she freaked out, and she's going crazy over it. Um, I don't why she was going through stuff in the first place, but whatever, so, now I'm here.
Therapist: Mm-hmm.
Client: Um, I've been hanging out with a new cool crowd of people that I really like.
Therapist: Mm-hmm.
Client: Uh, a-and-and that's-that's basically it.
Therapist: Yeah. So-so you've got this new group of friends and-and, um, you-you actually kind of like where you're at with things right now. And your mom was going through your stuff and found your stash, and it's just turned into a, you know, all of this.
Client: Yeah.
Therapist: Yeah.
Client: Yeah, basically.
Therapist: Yeah.
Client: Mm-hmm.
Therapist: And, uh, sounds like you're-you're pretty upset with your mom for-for doing that?
Client: I am.
Therapist: Yeah.
Client: I mean, it's my stuff, I don't know why she's–

Therapist (Human): Right. It's like your private place and, you know, it's– that's– it's your stuff.		
	GPT-2 Stage Annotation	
L2: *Coherent*	**L3**: *Coherent*	**L7‡**: *Coherent*
E2: *Coherent*	**E7**: *Coherent*	**E8**: *Coherent*
	GPT-3 Stage Annotation	
L1: *Coherent*	**L4**: Parroting	**L7‡**: Parroting
E3: *Coherent*	**E4**: *Coherent*	**E5**: Malformed

GPT-2: It's a very sad thing.		
L2: *Coherent*	**L3** *Coherent*	**L7**: Off-Topic
E2: *Coherent*	**E7**: Off-Topic	**E8**: Off-Topic

GPT-3: It sounds like you're really upset with her because she invaded your privacy.		
L1: *Coherent*	**L4**: *Coherent*	**L7**: *Coherent*
E3: *Coherent*	**E4**: *Coherent*	**E5**: *Coherent*

sistent reflections, and we also explored the annotation shift on human reflections when the source of synthetic reflections changes from a smaller LM (GPT-2) to the more powerful GPT-3.

For future work, we plan to mix, in each batch, synthetic reflections from models of different scales, and investigate how the resulting human evaluations might differ. Another direction worth exploring is alternative ways of coherence annotation, such as ranking, for more nuanced human evaluation results. Future work may also re-examine and modify the error categories to increase IAA on error annotations. We also leave potentially IAA-improving annotation procedures to future work, such as using a warm-up exercise task before actual annotation and allowing annotators to discuss with each other to resolve their differences.

Chapter 7

Conclusion

In this thesis, we focused on addressing 3 major challenges in NLP for MI (§1.2): lack of publicly available data, lack of common benchmarks and reproducibility, and dependence on experts in human evaluation for therapist response generation. Specifically, we investigated low-resource NLP methods for MI analysis, created an expert-annotated MI dialogue dataset, explored real-life applicable tasks on the dataset, and showed the feasibility of laypeople-based evaluation of a key quality aspect for therapist response generation.

In this chapter, we first review the contributions and limitations of our work (§7.1) in the context of the 4 research questions laid out in §1.3, and then discuss potential avenues for future work (§7.2).

7.1 Contributions and Limitations

RQ1: How to approach real-time empathy assessment for MI in a low-resource setting, where there is little to no MI dialogue data with empathy-related annotations?

For RQ1, we approached the novel task of zero-shot binary prediction of therapist empathy at the utterance level (Chapter 3). We proposed 1) a supervised method that trains BERT on heuristically constructed empathy vs. non-empathy contrast in non-counselling conversations; and 2) an unsupervised method that utilises NLI as a proxy task for empathy prediction. Although our results indicated that those zero-shot methods are not sufficiently accurate, we found that empathy vs. non-empathy contrast yields the best performance. We also showed that the benefit of this contrast is more apparent when compared to the unsupervised approach and control-group supervised models without empathy contrast training.

The main limitations of this work are the noise and low granularity of labels. The heuristic empathy labels for utterances in non-counselling conversations are based on corpus-level empathy labels assigned by the creators of the original dialogue datasets, so the heuristic labels may not be entirely accurate at the sentence or utterance level. In addition, our annotation of utterance-level empathy for the MI dialogues (in-domain test data) can sometimes be coarse-grained because some utterances have both empathetic and non-empathetic parts, and the absence of sentence-level punctuation made it non-trivial to separate those parts. We postulated that more fine-grained empathy labels may lead to better performance in future work.

RQ2: How to create a publicly available and expert-annotated dataset of MI dialogues to benefit the research community?

For RQ2, we made publicly available `AnnoMI` (Chapter 4), an expert-annotated dataset of professionally transcribed conversations that demonstrate high- and low-quality MI. We performed a series of thorough analyses of MI-related properties at the levels of utterance, dialogue, and corpus, utilising the extensive annotations provided by experienced therapists. `AnnoMI` represents a valuable resource for advancing research in the vital field of NLP for counselling and social good.

The main limitation of this work is that `AnnoMI` consists of transcripts of MI demonstrations rather than actual counselling sessions. Nonetheless, we consider it to be the most accurate representation possible of real counselling without compromising client privacy. To ensure the resemblance of `AnnoMI` to real-life counselling, we sourced videos from professional therapists and research institutions. In addition, the feedback from expert annotators confirms the realism of `AnnoMI`.

RQ3: How to leverage the dataset of RQ2 to create benchmark tasks and models with potential for real-world application?

For RQ3, we examined two `AnnoMI`-based tasks with real-world applicability (Chapter 5): current-turn therapist/client behaviour prediction and next-turn therapist behaviour forecasting. Prediction identifies the behaviour label of the current turn by using its utterance, while forecasting anticipates next-turn behaviour using the dialogue history so far.

For prediction, we found that LMs outperform other models and achieve better results on therapist behaviours than on client behaviours, in terms of overall performance and generalisability to new topics. For forecasting, we observed suboptimal performance of the baseline LM with mostly no improvement conferred by the various NLP modelling choices, which suggests counselling is often flexible and allows for multiple optimal next-turn actions.

For prediction, the main limitation is the lack of contextualisation for the input utterance, and we postulated that using dialogue context as auxiliary input may lead to performance gains. For forecasting, the main limitation is the assumption of a single optimal next-turn behaviour in our experiments, and we posited that probabilistic modelling may be a better alternative. We leave the experimentation of these ideas to future work.

RQ4: In therapist response generation for MI, what criteria should human evaluators meet to ensure effective evaluation?

For RQ4, we zoomed in on *reflection*, an essential type of therapist response in MI, and examined human evaluation of generated reflections (Chapter 6). Our objective was to investigate whether laypeople can be an alternative to experts in evaluating the fundamental quality aspect of coherence and context-consistency, namely "Does this reflection make sense (in context)?". To this end, we created an evaluation scheme based on laypeople's descriptions of incoherence and inconsistency errors in generated reflections. Subsequently, we asked both laypeople and experts to annotate both human and generated reflections using this scheme. Our results revealed that laypeople are capable of this evaluation, as shown by their agreement with each other and correlation with experts. Additionally, we found that the powerful GPT-3 generates predominantly coherent and consistent reflections. We also explored how the evaluation results change when the source of generated reflections is switched from the less powerful GPT-2 to the more powerful GPT-3.

The main limitation of this work is the relatively small number of human reflections, which is considerably lower than the quantities of GPT-2 and GPT-3 reflections. With more annotated human reflections, we could confirm various potential findings, such as whether GPT-3 reflections have a statistically significant lead in coherence/consistency compared to human reflections.

Impact: In addressing the 4 research questions, we considerably expanded access (Chapters 4, 6) to NLP-for-MI research as well as presented extensive findings on related NLU (Chapters 3, 5) and NLG (Chapter 6) tasks with real-world relevance. Thus, our work paves the way for greater participation in research for MI from the broader NLP community, which will ultimately benefit clients of MI counselling.

7.2 Future Work

There are still many research opportunities in the field of NLP for MI, especially in the context of rapidly developing LLMs [17, 28]. As a first step, future work could start with addressing the limitations of our work, in particular those noted in §7.1. Furthermore, in order for our research to be more useful for therapists, future work should conduct clinical studies to gain more real-world insights, which should cover both the effectiveness of the technologies we proposed and practical concerns (e.g., using federated learning [190] to improve models without compromising data privacy). Additionally, there are numerous NLP-for-MI tasks that have yet to be explored. We conclude this thesis by discussing a few potential avenues for future research from an NLP perspective.

LLMs for Few-Shot NLU One possible direction for future research is to model the NLU tasks approached in this work (Chapters 3 and 5) in a few-shot setting, as in-context learning with LLMs has seen considerable progress recently for classification/prediction tasks [17]. In particular, a few-shot approach would considerably alleviate the need for many training examples in fine-tuning smaller LMs while possibly achieving competitive if not better performance, which is meaningful for low-resource domains like NLP for counselling. Nevertheless, future research in this area will need to take into account in-context learning considerations, such as the ordering [184] and selection [191] of few-shot examples.

LLMs for Data Augmentation Another potential use of LLMs for this field is data augmentation. Considering the small size of `AnnoMI`, data augmentation is useful for generating more training examples to fine-tune smaller LMs. While we mostly did not observe performance gains when using augmented examples from a non-LLM neural paraphrasing model (Chapter 5), LLMs are considerably more powerful and would likely produce higher-quality augmentations. A simple first step would be to use LLMs to create augmentations by paraphrasing existing examples, as recent studies have shown the efficacy of this approach [192]. Future research may also investigate using existing examples (e.g., utterances/dialogues) in a few-shot setup to prompt LLMs into generating completely new examples, which have proved to be of high quality in recent work [193, 194, 195].

LLMs as NLG Evaluators A third promising application of LLMs is as quality evaluators for therapist response generation. Very recent studies have shown that LLMs such as GPT-4 [28] have become state-of-the-art automatic quality evaluators in NLG tasks like dialogue generation, summarisation and machine translation [196, 197, 198], in terms of correlation with human judgement. Therefore, future research could examine whether LLM-produced evaluation of model-generated therapist responses correlates with the judgement of MI experts, both in terms of surface-level aspects like coherence/consistency and more complex dimensions like compliance with counselling principles. A particularly interesting experiment would be to ask an LLM-based evaluator to produce free-text explanations for its evaluation, and then ask human experts to assess the validity of such explanations.

Multilingual NLP for MI Lastly, future research should more actively investigate NLP for MI beyond English, especially for low-resource languages, which is important for achieving NLP for social good [199] in general. A starting point could be to manually or automatically translate `AnnoMI` into different languages and then test and improve the performance of language-specific (e.g., [200]) and multilingual (e.g., [201]) LMs on MI-related NLU and NLG tasks. Considering that state-of-the-art LLMs have shown increasingly narrow performance gaps between English and low-resource languages [28], future research may leverage LLMs for multilingual NLP-for-MI tasks directly and work towards improving the performance in low-resource languages on that basis.

Bibliography

[1] Verónica Pérez-Rosas, Xinyi Wu, Kenneth Resnicow, and Rada Mihalcea. What makes a good counselor? learning to distinguish between high-quality and low-quality counseling conversations. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 926–935, Florence, Italy, July 2019. Association for Computational Linguistics.

[2] Theresa B Moyers, Lauren N Rowell, Jennifer K Manuel, Denise Ernst, and Jon M Houck. The motivational interviewing treatment integrity code (miti 4): rationale, preliminary reliability and validity. *Journal of substance abuse treatment*, 65:36–42, 2016.

[3] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Ehud Reiter, Diego Reforgiato Recupero, and Daniele Riboni. Anno-mi: A dataset of expert-annotated counselling dialogues. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2022, Virtual and Singapore, 23-27 May 2022*, pages 6177–6181. IEEE, 2022. (©2022 IEEE. Reprinted, with permission.).

[4] Laure Carcaillon-Bentata, Noémie Soullier, Nathalie Beltzer, and Joël Coste. Alteration in perceived health status of those aged 55 to 65 between 2010 and 2017 in france: role of socioeconomic determinants. *BMC public health*, 21:1–11, 2021.

[5] Ilse Reinders, Natasja M Van Schoor, Dorly JH Deeg, Martijn Huisman, and Marjolein Visser. Trends in lifestyle among three cohorts of adults aged 55–64 years in 1992/1993, 2002/2003 and 2012/2013. *The European Journal of Public Health*, 28(3):564–570, 2018.

[6] Frank W Booth, Christian K Roberts, John P Thyfault, Gregory N Ruegsegger, and Ryan G Toedebusch. Role of inactivity in chronic diseases: evolutionary insight and pathophysiological mechanisms. *Physiological reviews*, 2017.

[7] Michael Roerecke, Janusz Kaczorowski, Sheldon W Tobe, Gerrit Gmel, Omer SM Hasan, and Juergen Rehm. The effect of a reduction in alcohol

consumption on blood pressure: a systematic review and meta-analysis. *The Lancet Public Health*, 2(2):e108–e120, 2017.

[8] Jane Wu and Don D Sin. Improved patient outcome with smoking cessation: when is it too late? *International journal of chronic obstructive pulmonary disease*, pages 259–267, 2011.

[9] Kathryn R Middleton, Stephen D Anton, and Michal G Perri. Long-term adherence to health behavior change. *American journal of lifestyle medicine*, 7(6):395–404, 2013.

[10] William R Miller and Stephen Rollnick. *Motivational interviewing: Helping people change.* Guilford press, 2012.

[11] Stephen Rollnick, William R Miller, and Christopher Butler. *Motivational interviewing in health care: helping patients change behavior.* Guilford Press, 2008.

[12] David R Thompson, Sek Y Chair, Sally W Chan, Felicity Astin, Patricia M Davidson, and Chantal F Ski. Motivational interviewing: a useful approach to improving cardiovascular health? *Journal of clinical nursing*, 20(9-10):1236–1244, 2011.

[13] Jie Cao, Michael Tanana, Zac E. Imel, Eric Poitras, David C. Atkins, and Vivek Srikumar. Observing dialogue in therapy: Categorizing and forecasting behavioral codes. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5599–5611. Association for Computational Linguistics, 2019.

[14] Roger Bakeman and Vicenç Quera. Behavioral observation. In *APA handbook of research methods in psychology, Vol 1: Foundations, planning, measures, and psychometrics.*, APA handbooks in psychology®., pages 207–225. American Psychological Association, Washington, DC, US, 2012.

[15] Nikolaos Flemotomos, Victor R Martinez, Zhuohao Chen, Karan Singla, Victor Ardulov, Raghuveer Peri, Derek D Caperton, James Gibson, Michael J Tanana, Panayiotis Georgiou, et al. Automated evaluation of psychotherapy skills using speech and language technologies. *Behavior Research Methods*, 54(2):690–711, 2022.

[16] William R Miller, Theresa B Moyers, Denise Ernst, and Paul Amrhein. Manual for the motivational interviewing skill code (misc). *Unpublished manuscript. Albuquerque: Center on Alcoholism, Substance Abuse and Addictions, University of New Mexico*, 2003.

[17] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[18] Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. Palm: Scaling language modeling with pathways. *CoRR*, abs/2204.02311, 2022.

[19] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[20] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008, 2017.

[21] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In Jill Burstein, Christy Doran, and Thamar Solorio, editors, *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers)*, pages 4171–4186. Association for Computational Linguistics, 2019.

[22] Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

[23] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[24] Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[25] Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. Superglue: A stickier benchmark for general-purpose language understanding systems. In Hanna M. Wallach, Hugo Larochelle, Alina Beygelzimer, Florence d'Alché-Buc, Emily B. Fox, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 3261–3275, 2019.

[26] Adithya Bhaskar, Alexander R. Fabbri, and Greg Durrett. Prompted opinion summarization with GPT-3.5. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 9282–9300. Association for Computational Linguistics, 2023.

[27] Tanya Goyal, Junyi Jessy Li, and Greg Durrett. News summarization and evaluation in the era of GPT-3. *CoRR*, abs/2209.12356, 2022.

[28] OpenAI. GPT-4 technical report. *CoRR*, abs/2303.08774, 2023.

[29] Bo Xiao, Dogan Can, Panayiotis G. Georgiou, David C. Atkins, and Shrikanth S. Narayanan. Analyzing the language of therapist empathy in motivational interview based psychotherapy. In *Asia-Pacific Signal and Information Processing Association Annual Summit and Conference, APSIPA 2012, Hollywood, CA, USA, December 3-6, 2012*, pages 1–4. IEEE, 2012.

[30] James Gibson, Nikolaos Malandrakis, Francisco Romero, David C. Atkins, and Shrikanth S. Narayanan. Predicting therapist empathy in motivational interviews using language features inspired by psycholinguistic norms. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 1947–1951. ISCA, 2015.

[31] James Gibson, Dogan Can, Bo Xiao, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. A deep learning approach to modeling empathy in addiction counseling. In Nelson Morgan, editor, *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 1447–1451. ISCA, 2016.

[32] Michael Tanana, Kevin A Hallgren, Zac E Imel, David C Atkins, and Vivek Srikumar. A comparison of natural language processing methods for automated coding of motivational interviewing. *Journal of substance abuse treatment*, 65:43–50, 2016.

[33] Bo Xiao, Dogan Can, James Gibson, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Behavioral coding of therapist language in addiction counseling using recurrent neural networks. In Nelson Morgan, editor, *Interspeech 2016, 17th Annual Conference of the International Speech Communication Association, San Francisco, CA, USA, September 8-12, 2016*, pages 908–912. ISCA, 2016.

[34] James Gibson, David Atkins, Torrey Creed, Zac Imel, Panayiotis Georgiou, and Shrikanth Narayanan. Multi-label multi-task deep learning for behavioral coding. *IEEE Transactions on Affective Computing*, 2019.

[35] Siqi Shen, Charles Welch, Rada Mihalcea, and Verónica Pérez-Rosas. Counseling-style reflection generation using generative pretrained transformers with augmented context. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 10–20, 2020.

[36] Siqi Shen, Verónica Pérez-Rosas, Charles Welch, Soujanya Poria, and Rada Mihalcea. Knowledge enhanced reflection generation for counseling dialogues. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3096–3107, 2022.

[37] Imtihan Ahmed. *Automatic Generation and Detection of Motivational-Interviewing-Style Reflections for Smoking Cessation Therapeutic Conversations Using Transformer-based Language Models.* PhD thesis, University of Toronto, 2022.

[38] Adela Grando, Julia Ivanova, Megan Hiestand, Hiral Soni, Anita Murcko, Michael Saks, David Kaufman, Mary Jo Whitfield, Christy Dye, Darwyn Chern, et al. Mental health professional perspectives on health data sharing: Mixed methods study. *Health informatics journal*, 26(3):2067–2082, 2020.

[39] Michael J Saks, Adela Grando, Anita Murcko, and Chase Millea. Granular patient control of personal health information: federal and state law considerations. *Jurimetrics*, 58(4):411, 2018.

[40] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation. *ACM Comput. Surv.*, 55(12), mar 2023.

[41] Mingyu Zong and Bhaskar Krishnamachari. a survey on gpt-3. *arXiv preprint arXiv:2212.00857*, 2022.

[42] Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A Smith, and Yejin Choi. Is gpt-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, 2022.

[43] Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Towards low-resource real-time assessment of empathy in counselling. In *Proceedings of the Seventh Workshop on Computational Linguistics and Clinical Psychology: Improving Access*, pages 204–216, Online, June 2021. Association for Computational Linguistics.

[44] Zixiu Wu, Simone Balloccu, Vivek Kumar, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Creation, analysis and evaluation of annomi, a dataset of expert-annotated counselling dialogues. *Future Internet*, 15(3), 2023.

[45] Zixiu Wu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Towards automated counselling decision-making: Remarks on therapist action forecasting on the annomi dataset. In Hanseok Ko and John H. L. Hansen, editors, *Interspeech 2022, 23rd Annual Conference of the International Speech Communication Association, Incheon, Korea, 18-22 September 2022*, pages 1906–1910. ISCA, 2022. DOI: 10.21437/Interspeech.2022-506.

[46] Zixiu Wu, Simone Balloccu, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Towards in-context non-expert evaluation of reflection generation for counselling conversations. In *Proceedings of the 2nd Workshop on Natural Language Generation, Evaluation, and Metrics (GEM)*, pages 116–124, Abu Dhabi, United Arab Emirates (Hybrid), December 2022. Association for Computational Linguistics.

[47] Zixiu Wu, Simone Balloccu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Are experts needed? on human evaluation of counselling reflection generation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6906–6930, Toronto, Canada, July 2023. Association for Computational Linguistics.

[48] Zixiu Wu, Rim Helaoui, Vivek Kumar, Diego Reforgiato Recupero, and Daniele Riboni. Towards detecting need for empathetic response in motivational interviewing. In *Companion Publication of the 2020 International Conference on Multimodal Interaction*, pages 497–502, 2020.

[49] Vivek Kumar, Simone Balloccu, Zixiu Wu, Ehud Reiter, Rim Helaoui, Diego Reforgiato Recupero, and Daniele Riboni. Data augmentation for reliability and fairness in counselling quality classification. In *Proceedings of the*

1st Workshop on Scarce Data in Artificial Intelligence for Healthcare - SDAIH, pages 23–28. INSTICC, SciTePress, 2023.

[50] Christopher Manning. Cs224n: Natural language processing with deep learning. https://web.stanford.edu/class/cs224n/, 2023. Accessed: 4 August 2023.

[51] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. Neural machine translation by jointly learning to align and translate. In Yoshua Bengio and Yann LeCun, editors, *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, 2015.

[52] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. BART: denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 7871–7880. Association for Computational Linguistics, 2020.

[53] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

[54] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. Language models are unsupervised multitask learners. *OpenAI Blog*, 1(8):9, 2019.

[55] Zhenzhong Lan, Mingda Chen, Sebastian Goodman, Kevin Gimpel, Piyush Sharma, and Radu Soricut. ALBERT: A lite BERT for self-supervised learning of language representations. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net, 2020.

[56] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21:140:1–140:67, 2020.

[57] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for nlp. In *International Conference on Machine Learning*, pages 2790–2799. PMLR, 2019.

[58] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416, 2022.

[59] Paul F. Christiano, Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. In Isabelle Guyon, Ulrike von Luxburg, Samy Bengio, Hanna M. Wallach, Rob Fergus, S. V. N. Vishwanathan, and Roman Garnett, editors, *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4299–4307, 2017.

[60] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[61] Brian W Matthews. Comparison of the predicted and observed secondary structure of t4 phage lysozyme. *Biochimica et Biophysica Acta (BBA)-Protein Structure*, 405(2):442–451, 1975.

[62] Chia-Wei Liu, Ryan Lowe, Iulian Vlad Serban, Mike Noseworthy, Laurent Charlin, and Joelle Pineau. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2122–2132, 2016.

[63] Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, July 6-12, 2002, Philadelphia, PA, USA*, pages 311–318. ACL, 2002.

[64] Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain, July 2004. Association for Computational Linguistics.

[65] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*, 2019.

[66] Shikib Mehri and Maxine Eskenazi. Usr: An unsupervised and reference free evaluation metric for dialog generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 681–707, 2020.

[67] Shikib Mehri and Maxine Eskenazi. Unsupervised evaluation of interactive dialog with dialogpt. In *Proceedings of the 21st Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 225–235, 2020.

[68] Jan Deriu, Álvaro Rodrigo, Arantxa Otegi, Guillermo Echegoyen, Sophie Rosset, Eneko Agirre, and Mark Cieliebak. Survey on evaluation methods for dialogue systems. *Artif. Intell. Rev.*, 54(1):755–810, 2021.

[69] Romal Thoppilan, Daniel De Freitas, Jamie Hall, Noam Shazeer, Apoorv Kulshreshtha, Heng-Tze Cheng, Alicia Jin, Taylor Bos, Leslie Baker, Yu Du, et al. Lamda: Language models for dialog applications. *arXiv preprint arXiv:2201.08239*, 2022.

[70] Kurt Shuster, Jing Xu, Mojtaba Komeili, Da Ju, Eric Michael Smith, Stephen Roller, Megan Ung, Moya Chen, Kushal Arora, Joshua Lane, et al. Blenderbot 3: a deployed conversational agent that continually learns to responsibly engage. *arXiv preprint arXiv:2208.03188*, 2022.

[71] Mojtaba Komeili, Kurt Shuster, and Jason Weston. Internet-augmented dialogue generation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 8460–8478. Association for Computational Linguistics, 2022.

[72] Hannah Rashkin, Eric Michael Smith, Margaret Li, and Y-Lan Boureau. Towards empathetic open-domain conversation models: A new benchmark and dataset. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 5370–5381. Association for Computational Linguistics, 2019.

[73] Qintong Li, Hongshen Chen, Zhaochun Ren, Pengjie Ren, Zhaopeng Tu, and Zhumin Chen. Empdg: Multi-resolution interactive empathetic dialogue generation. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 4454–4466, 2020.

[74] Yubo Xie and Pearl Pu. Empathetic dialog generation with fine-grained intents. In *Proceedings of the 25th Conference on Computational Natural Language Learning*, pages 133–147, 2021.

[75] Hyunwoo Kim, Byeongchang Kim, and Gunhee Kim. Perspective-taking and pragmatics for generating empathetic responses focused on emotion causes. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 2227–2240, 2021.

[76] Margaret Li, Jason Weston, and Stephen Roller. ACUTE-EVAL: improved dialogue evaluation with optimized questions and multi-turn comparisons. *CoRR*, abs/1909.03087, 2019.

[77] David M. Howcroft and Verena Rieser. What happens if you treat ordinal ratings as interval data? human evaluations in NLP are even more underpowered than you think. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 8932–8939. Association for Computational Linguistics, 2021.

[78] Anya Belz, Craig Thomson, and Ehud Reiter. Missing information, unresponsive authors, experimental flaws: The impossibility of assessing the reproducibility of previous human evaluations in NLP. In *The Fourth Workshop on Insights from Negative Results in NLP*, pages 1–10, Dubrovnik, Croatia, May 2023. Association for Computational Linguistics.

[79] Jessica Huynh, Jeffrey Bigham, and Maxine Eskénazi. A survey of nlp-related crowdsourcing hits: what works and what does not. *CoRR*, abs/2111.05241, 2021.

[80] Eric Michael Smith, Orion Hsu, Rebecca Qian, Stephen Roller, Y-Lan Boureau, and Jason Weston. Human evaluation of conversations is an open problem: comparing the sensitivity of various methods for evaluating dialogue agents. In Bing Liu, Alexandros Papangelis, Stefan Ultes, Abhinav Rastogi, Yun-Nung Chen, Georgios Spithourakis, Elnaz Nouri, and Weiyan Shi, editors, *Proceedings of the 4th Workshop on NLP for Conversational AI, ConvAI@ACL 2022, Dublin, Ireland, May 27, 2022*, pages 77–97. Association for Computational Linguistics, 2022.

[81] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Eric Michael Smith, Y-Lan Boureau, and Jason Weston. Recipes for building an open-domain chatbot. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 300–325, Online, April 2021. Association for Computational Linguistics.

[82] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Towards facilitating empathic conversations in online mental health support:

A reinforcement learning approach. In *Proceedings of the Web Conference 2021*, pages 194–205, 2021.

[83] Juliana Miehle, Nadine Gerstenlauer, Daniel Ostler, Hubertus Feußner, Wolfgang Minker, and Stefan Ultes. Expert evaluation of a spoken dialogue system in a clinical operating room. In Nicoletta Calzolari, Khalid Choukri, Christopher Cieri, Thierry Declerck, Sara Goggi, Kôiti Hasida, Hitoshi Isahara, Bente Maegaard, Joseph Mariani, Hélène Mazo, Asunción Moreno, Jan Odijk, Stelios Piperidis, and Takenobu Tokunaga, editors, *Proceedings of the Eleventh International Conference on Language Resources and Evaluation, LREC 2018, Miyazaki, Japan, May 7-12, 2018*. European Language Resources Association (ELRA), 2018.

[84] Dan Gillick and Yang Liu. Non-expert evaluation of summarization systems is risky. In Chris Callison-Burch and Mark Dredze, editors, *Proceedings of the 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk, Los Angeles, USA, June 6, 2010*, pages 148–151. Association for Computational Linguistics, 2010.

[85] Alexander R. Fabbri, Wojciech Kryscinski, Bryan McCann, Caiming Xiong, Richard Socher, and Dragomir R. Radev. Summeval: Re-evaluating summarization evaluation. *Trans. Assoc. Comput. Linguistics*, 9:391–409, 2021.

[86] Markus Freitag, George F. Foster, David Grangier, Viresh Ratnakar, Qijun Tan, and Wolfgang Macherey. Experts, errors, and context: A large-scale study of human evaluation for machine translation. *Trans. Assoc. Comput. Linguistics*, 9:1460–1474, 2021.

[87] Marzena Karpinska, Nader Akoury, and Mohit Iyyer. The perils of using mechanical turk to evaluate open-ended text generation. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 7-11 November, 2021*, pages 1265–1285. Association for Computational Linguistics, 2021.

[88] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Y. Ng. Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks. In *2008 Conference on Empirical Methods in Natural Language Processing, EMNLP 2008, Proceedings of the Conference, 25-27 October 2008, Honolulu, Hawaii, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 254–263. ACL, 2008.

[89] Peter Roy-Byrne, Kristin Bumgardner, Antoinette Krupski, Chris Dunn, Richard Ries, Dennis Donovan, Imara I West, Charles Maynard, David C

Atkins, Meredith C Graves, et al. Brief intervention for problem drug use in safety-net primary care settings: a randomized clinical trial. *Jama*, 312(5):492–501, 2014.

[90] Christine M Lee, Clayton Neighbors, Melissa A Lewis, Debra Kaysen, Angela Mittmann, Irene M Geisner, David C Atkins, Cheng Zheng, Lisa A Garberson, Jason R Kilmer, et al. Randomized controlled trial of a spring break intervention to reduce high-risk drinking. *Journal of consulting and clinical psychology*, 82(2):189, 2014.

[91] Dogan Can, David C. Atkins, and Shrikanth S. Narayanan. A dialog act tagging approach to behavioral coding: a case study of addiction counseling conversations. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 339–343. ISCA, 2015.

[92] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. Building a motivational interviewing dataset. In *Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology*, pages 42–51, 2016.

[93] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, Lawrence An, Kathy J Goggin, and Delwyn Catley. Predicting counselor behaviors in motivational interviewing encounters. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1128–1137, 2017.

[94] Verónica Pérez-Rosas, Rada Mihalcea, Kenneth Resnicow, Satinder Singh, and Lawrence An. Understanding and predicting empathic behavior in counseling therapy. In Regina Barzilay and Min-Yen Kan, editors, *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1426–1435. Association for Computational Linguistics, 2017.

[95] Raj Sanjay Shah, Faye Holt, Shirley Anugrah Hayati, Aastha Agarwal, Yi-Chia Wang, Robert E Kraut, and Diyi Yang. Modeling motivational interviewing strategies on an online peer-to-peer counseling platform. *Proceedings of the ACM on Human-Computer Interaction*, 6(CSCW2):1–24, 2022.

[96] Anuradha Welivita and Pearl Pu. Curating a large-scale motivational interviewing dataset using peer support forums. In *Proceedings of the 29th International Conference on Computational Linguistics*, pages 3315–3330, Gyeongju, Republic of Korea, October 2022. International Committee on Computational Linguistics.

[97] Do June Min, Verónica Pérez-Rosas, Kenneth Resnicow, and Rada Mihalcea. PAIR: Prompt-aware margIn ranking for counselor reflection scoring in motivational interviewing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 148–158, Abu Dhabi, United Arab Emirates, December 2022. Association for Computational Linguistics.

[98] Selina Meyer and David Elsweiler. Glohbcd: A naturalistic german dataset for language of health behaviour change on online support forums. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2226–2235, 2022.

[99] William J Reynolds and Brain Scott. Empathy: a crucial component of the helping relationship. *Journal of psychiatric and mental health nursing*, 6(5):363–370, 1999.

[100] Frans Derksen, Jozien Bensing, and Antoine Lagro-Janssen. Effectiveness of empathy in general practice: a systematic review. *British journal of general practice*, 63(606):e76–e84, 2013.

[101] Bo Xiao, Panayiotis G Georgiou, Zac E Imel, David C Atkins, and Shrikanth S Narayanan. Modeling therapist empathy and vocal entrainment in drug addiction counseling. In *Interspeech*, pages 2861–2865, 2013.

[102] Bo Xiao, Daniel Bone, Maarten Van Segbroeck, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Modeling therapist empathy through prosody in drug addiction counseling. In Haizhou Li, Helen M. Meng, Bin Ma, Engsiong Chng, and Lei Xie, editors, *INTERSPEECH 2014, 15th Annual Conference of the International Speech Communication Association, Singapore, September 14-18, 2014*, pages 213–217. ISCA, 2014.

[103] Bo Xiao, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Analyzing speech rate entrainment and its relation to therapist empathy in drug addiction counseling. In *INTERSPEECH 2015, 16th Annual Conference of the International Speech Communication Association, Dresden, Germany, September 6-10, 2015*, pages 2489–2493. ISCA, 2015.

[104] Doğan Can, Panayiotis G Georgiou, David C Atkins, and Shrikanth S Narayanan. A case study: Detecting counselor reflections in psychotherapy for addictions using linguistic features. In *Thirteenth Annual Conference of the International Speech Communication Association*, 2012.

[105] David C Atkins, Mark Steyvers, Zac E Imel, and Padhraic Smyth. Scaling up the evaluation of psychotherapy: evaluating motivational interviewing fidelity via statistical text classification. *Implementation Science*, 9(1):1–11, 2014.

[106] Richard Socher, Alex Perelygin, Jean Wu, Jason Chuang, Christopher D. Manning, Andrew Y. Ng, and Christopher Potts. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, EMNLP 2013, 18-21 October 2013, Grand Hyatt Seattle, Seattle, Washington, USA, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1631–1642. ACL, 2013.

[107] Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543, 2014.

[108] Kyunghyun Cho, Bart van Merrienboer, Çaglar Gülçehre, Dzmitry Bahdanau, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. Learning phrase representations using RNN encoder-decoder for statistical machine translation. In Alessandro Moschitti, Bo Pang, and Walter Daelemans, editors, *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing, EMNLP 2014, October 25-29, 2014, Doha, Qatar, A meeting of SIGDAT, a Special Interest Group of the ACL*, pages 1724–1734. ACL, 2014.

[109] Karan Singla, Zhuohao Chen, David C. Atkins, and Shrikanth Narayanan. Towards end-2-end learning for predicting behavior codes from spoken utterances in psychotherapy conversations. In Dan Jurafsky, Joyce Chai, Natalie Schluter, and Joel R. Tetreault, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 3797–3803. Association for Computational Linguistics, 2020.

[110] Yu-An Chung and James R. Glass. Speech2vec: A sequence-to-sequence framework for learning word embeddings from speech. In B. Yegnanarayana, editor, *Interspeech 2018, 19th Annual Conference of the International Speech Communication Association, Hyderabad, India, 2-6 September 2018*, pages 811–815. ISCA, 2018.

[111] Karan Singla, Zhuohao Chen, Nikolaos Flemotomos, James Gibson, Dogan Can, David C Atkins, and Shrikanth S Narayanan. Using prosodic and lexical information for learning utterance-level behaviors in psychotherapy. In *Interspeech*, pages 3413–3417, 2018.

[112] Zhuohao Chen, Karan Singla, James Gibson, Dogan Can, Zac E. Imel, David C. Atkins, Panayiotis G. Georgiou, and Shrikanth S. Narayanan. Improving the prediction of therapist behaviors in addiction counseling by exploiting class confusions. In *IEEE International Conference on Acoustics, Speech and Signal Processing, ICASSP 2019, Brighton, United Kingdom, May 12-17, 2019*, pages 6605–6609. IEEE, 2019.

[113] Erik Rautalinko, Hans-Olof Lisper, and Bo Ekehammar. Reflective listening in counseling: effects of training time and evaluator social skills. *American journal of psychotherapy*, 61(2):191–209, 2007.

[114] Sean J Tollison, Christine M Lee, Clayton Neighbors, Teryl A Neil, Nichole D Olson, and Mary E Larimer. Questions and reflections: the use of motivational interviewing microskills in a peer-led brief alcohol intervention for college students. *Behavior therapy*, 39(2):183–194, 2008.

[115] Fahad Almusharraf, Jonathan Rose, and Peter Selby. Engaging unmotivated smokers to move toward quitting: design of motivational interviewing–based chatbot through iterative interactions. *Journal of Medical Internet Research*, 22(11):e20251, 2020.

[116] Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In Satinder Singh and Shaul Markovitch, editors, *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA*, pages 4444–4451. AAAI Press, 2017.

[117] Antoine Bosselut, Hannah Rashkin, Maarten Sap, Chaitanya Malaviya, Asli Celikyilmaz, and Yejin Choi. COMET: commonsense transformers for automatic knowledge graph construction. In Anna Korhonen, David R. Traum, and Lluís Màrquez, editors, *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28-August 2, 2019, Volume 1: Long Papers*, pages 4762–4779. Association for Computational Linguistics, 2019.

[118] Angela Fan, Mike Lewis, and Yann N. Dauphin. Hierarchical neural story generation. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 889–898. Association for Computational Linguistics, 2018.

[119] Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. The curious case of neural text degeneration. In *International Conference on Learning Representations*, 2019.

[120] Michael Denkowski and Alon Lavie. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the ninth workshop on statistical machine translation*, pages 376–380, 2014.

[121] Nouha Dziri, Ehsan Kamalloo, Kory Mathewson, and Osmar R Zaiane. Evaluating coherence in dialogue systems using entailment. In *Proceedings of the*

2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pages 3806–3812, 2019.

[122] Margaret Li, Stephen Roller, Ilia Kulikov, Sean Welleck, Y-Lan Boureau, Kyunghyun Cho, and Jason Weston. Don't say that! making inconsistent dialogue unlikely with unlikelihood training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4715–4728, 2020.

[123] Sabrina J Mielke, Arthur Szlam, Emily Dinan, and Y-Lan Boureau. Reducing conversational agents' overconfidence through linguistic calibration. *Transactions of the Association for Computational Linguistics*, 10:857–872, 2022.

[124] Amit Baumel. Online emotional support delivered by trained volunteers: users' satisfaction and their perception of the service compared to psychotherapy. *Journal of mental health*, 24(5):313–320, 2015.

[125] Ruben Fukkink. Peer counseling in an online chat service: A content analysis of social support. *Cyberpsychology, behavior, and social networking*, 14(4):247–251, 2011.

[126] Siyang Liu, Chujie Zheng, Orianna Demasi, Sahand Sabour, Yu Li, Zhou Yu, Yong Jiang, and Minlie Huang. Towards emotional support dialog systems. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3469–3483, 2021.

[127] Jiangnan Li, Zheng Lin, Peng Fu, and Weiping Wang. Past, present, and future: Conversational emotion recognition through structural modeling of psychological knowledge. In Marie-Francine Moens, Xuanjing Huang, Lucia Specia, and Scott Wen-tau Yih, editors, *Findings of the Association for Computational Linguistics: EMNLP 2021, Virtual Event / Punta Cana, Dominican Republic, 16-20 November, 2021*, pages 1204–1214. Association for Computational Linguistics, 2021.

[128] Yunsheng Shi, Zhengjie Huang, Shikun Feng, Hui Zhong, Wenjing Wang, and Yu Sun. Masked label prediction: Unified message passing model for semi-supervised classification. In Zhi-Hua Zhou, editor, *Proceedings of the Thirtieth International Joint Conference on Artificial Intelligence, IJCAI 2021, Virtual Event / Montreal, Canada, 19-27 August 2021*, pages 1548–1554. ijcai.org, 2021.

[129] Dayu Li, Xiaodan Zhu, Yang Li, Suge Wang, Deyu Li, Jian Liao, and Jianxing Zheng. Enhancing emotion inference in conversations with commonsense knowledge. *Knowl. Based Syst.*, 232:107449, 2021.

[130] Anuradha Welivita and Pearl Pu. A taxonomy of empathetic response intents in human social conversations. In Donia Scott, Núria Bel, and Chengqing Zong, editors, *Proceedings of the 28th International Conference on Computational Linguistics, COLING 2020, Barcelona, Spain (Online), December 8-13, 2020*, pages 4886–4899. International Committee on Computational Linguistics, 2020.

[131] Ekaterina Svikhnushina, Iuliana Voinea, Anuradha Welivita, and Pearl Pu. A taxonomy of empathetic questions in social dialogs. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 2952–2973. Association for Computational Linguistics, 2022.

[132] Yubo Xie, Junze Li, and Pearl Pu. AFEC: A knowledge graph capturing social intelligence in casual conversations. *CoRR*, abs/2205.10850, 2022.

[133] Hamed Khanpour, Cornelia Caragea, and Prakhar Biyani. Identifying empathetic messages in online health communities. In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 246–251, 2017.

[134] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[135] Mahshid Hosseini and Cornelia Caragea. It takes two to empathize: One to seek and one to provide. In *Thirty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2021, Thirty-Third Conference on Innovative Applications of Artificial Intelligence, IAAI 2021, The Eleventh Symposium on Educational Advances in Artificial Intelligence, EAAI 2021, Virtual Event, February 2-9, 2021*, pages 13018–13026. AAAI Press, 2021.

[136] Ashish Sharma, Adam S. Miner, David C. Atkins, and Tim Althoff. A computational approach to understanding empathy expressed in text-based mental health support. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 5263–5276. Association for Computational Linguistics, 2020.

[137] Ashish Sharma, Monojit Choudhury, Tim Althoff, and Amit Sharma. Engagement patterns of peer-to-peer interactions on mental health platforms. In Munmun De Choudhury, Rumi Chunara, Aron Culotta, and Brooke Foucault Welles, editors, *Proceedings of the Fourteenth International AAAI Conference*

on Web and Social Media, ICWSM 2020, Held Virtually, Original Venue: Atlanta, Georgia, USA, June 8-11, 2020, pages 614–625. AAAI Press, 2020.

[138] Naitian Zhou and David Jurgens. Condolence and empathy in online communities. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 609–626. Association for Computational Linguistics, 2020.

[139] Anuradha Welivita and Pearl Pu. HEAL: A knowledge graph for distress management conversations. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11459–11467. AAAI Press, 2022.

[140] Paul J. Werbos. Backpropagation through time: what it does and how to do it. *Proc. IEEE*, 78(10):1550–1560, 1990.

[141] Xianda Zhou and William Yang Wang. Mojitalk: Generating emotional responses at scale. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics, ACL 2018, Melbourne, Australia, July 15-20, 2018, Volume 1: Long Papers*, pages 1128–1137. Association for Computational Linguistics, 2018.

[142] Nurul Lubis, Sakriani Sakti, Koichiro Yoshino, and Satoshi Nakamura. Eliciting positive emotion through affect-sensitive dialogue response generation: A neural network approach. In Sheila A. McIlraith and Kilian Q. Weinberger, editors, *Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence, (AAAI-18), the 30th innovative Applications of Artificial Intelligence (IAAI-18), and the 8th AAAI Symposium on Educational Advances in Artificial Intelligence (EAAI-18), New Orleans, Louisiana, USA, February 2-7, 2018*, pages 5293–5300. AAAI Press, 2018.

[143] Zhaojiang Lin, Peng Xu, Genta Indra Winata, Farhad Bin Siddique, Zihan Liu, Jamin Shin, and Pascale Fung. Caire: An end-to-end empathetic chatbot. In *The Thirty-Fourth AAAI Conference on Artificial Intelligence, AAAI 2020, The Thirty-Second Innovative Applications of Artificial Intelligence Conference, IAAI 2020, The Tenth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2020, New York, NY, USA, February 7-12, 2020*, pages 13622–13623. AAAI Press, 2020.

[144] Zhaojiang Lin, Andrea Madotto, Jamin Shin, Peng Xu, and Pascale Fung. MoEL: Mixture of empathetic listeners. In *Proceedings of the 2019 Conference*

on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 121–132, Hong Kong, China, November 2019. Association for Computational Linguistics.

[145] Chengkun Zheng, Guanyi Chen, Chenghua Lin, Ruizhe Li, and Zhi Chen. Affective decoding for empathetic response generation. In Anya Belz, Angela Fan, Ehud Reiter, and Yaji Sripada, editors, *Proceedings of the 14th International Conference on Natural Language Generation, INLG 2021, Aberdeen, Scotland, UK, 20-24 September, 2021*, pages 331–340. Association for Computational Linguistics, 2021.

[146] Jiashuo Wang, Wenjie Li, Peiqin Lin, and Feiteng Mu. Empathetic response generation through graph-based multi-hop reasoning on emotional causality. *Knowledge-Based Systems*, 233:107547, 2021.

[147] Sahand Sabour, Chujie Zheng, and Minlie Huang. CEM: commonsense-aware empathetic response generation. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 11229–11237. AAAI Press, 2022.

[148] Yi Cheng, Wenge Liu, Wenjie Li, Jiashuo Wang, Ruihui Zhao, Bang Liu, Xiaodan Liang, and Yefeng Zheng. Improving multi-turn emotional support dialogue generation with lookahead strategy planning. In Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, editors, *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing, EMNLP 2022, Abu Dhabi, United Arab Emirates, December 7-11, 2022*, pages 3014–3026. Association for Computational Linguistics, 2022.

[149] Jiale Cheng, Sahand Sabour, Hao Sun, Zhuang Chen, and Minlie Huang. PAL: persona-augmented emotional support conversation generation. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 535–554. Association for Computational Linguistics, 2023.

[150] Quan Tu, Yanran Li, Jianwei Cui, Bin Wang, Ji-Rong Wen, and Rui Yan. MISC: A mixed strategy-aware model integrating COMET for emotional support conversation. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 308–319. Association for Computational Linguistics, 2022.

[151] Yang Deng, Wenxuan Zhang, Yifei Yuan, and Wai Lam. Knowledge-enhanced mixed-initiative dialogue system for emotional support conversations. In Anna Rogers, Jordan L. Boyd-Graber, and Naoaki Okazaki, editors, *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 4079–4095. Association for Computational Linguistics, 2023.

[152] Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. DIALOGPT : Large-scale generative pre-training for conversational response generation. In Asli Çelikyilmaz and Tsung-Hsien Wen, editors, *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: System Demonstrations, ACL 2020, Online, July 5-10, 2020*, pages 270–278. Association for Computational Linguistics, 2020.

[153] Ashish Sharma, Inna W Lin, Adam S Miner, David C Atkins, and Tim Althoff. Human–ai collaboration enables more empathic conversations in text-based peer-to-peer mental health support. *Nature Machine Intelligence*, 5(1):46–57, 2023.

[154] Peixiang Zhong, Chen Zhang, Hao Wang, Yong Liu, and Chunyan Miao. Towards persona-based empathetic conversational models. In Bonnie Webber, Trevor Cohn, Yulan He, and Yang Liu, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing, EMNLP 2020, Online, November 16-20, 2020*, pages 6556–6566. Association for Computational Linguistics, 2020.

[155] Joseph L Fleiss. Measuring nominal scale agreement among many raters. *Psychological bulletin*, 76(5):378, 1971.

[156] Jacob Cohen. Weighted kappa: nominal scale agreement provision for scaled disagreement or partial credit. *Psychological bulletin*, 70(4):213, 1968.

[157] Samuel R. Bowman, Gabor Angeli, Christopher Potts, and Christopher D. Manning. A large annotated corpus for learning natural language inference. In Lluís Màrquez, Chris Callison-Burch, Jian Su, Daniele Pighin, and Yuval Marton, editors, *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 632–642. The Association for Computational Linguistics, 2015.

[158] Wenpeng Yin, Jamaal Hay, and Dan Roth. Benchmarking zero-shot text classification: Datasets, evaluation and entailment approach. In Kentaro Inui, Jing Jiang, Vincent Ng, and Xiaojun Wan, editors, *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the*

9th International Joint Conference on Natural Language Processing, EMNLP-IJCNLP 2019, Hong Kong, China, November 3-7, 2019, pages 3912–3921. Association for Computational Linguistics, 2019.

[159] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. Transformers: State-of-the-art natural language processing. In Qun Liu and David Schlangen, editors, *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics, 2020.

[160] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830, 2011.

[161] Adina Williams, Nikita Nangia, and Samuel Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 1112–1122. Association for Computational Linguistics, 2018.

[162] Diego Reforgiato Recupero and Erik Cambria. Eswc'14 challenge on concept-level sentiment analysis. In Valentina Presutti, Milan Stankovic, Erik Cambria, Iván Cantador, Angelo Di Iorio, Tommaso Di Noia, Christoph Lange, Diego Reforgiato Recupero, and Anna Tordai, editors, *Semantic Web Evaluation Challenge - SemWebEval 2014 at ESWC 2014, Anissaras, Crete, Greece, May 25-29, 2014, Revised Selected Papers*, volume 475 of *Communications in Computer and Information Science*, pages 3–20. Springer, 2014.

[163] Amna Dridi and Diego Reforgiato Recupero. Leveraging semantics for sentiment polarity detection in social media. *International Journal of Machine Learning and Cybernetics*, 10:2045–2055, 2019.

[164] Diego Reforgiato Recupero, Mehwish Alam, Davide Buscaldi, Aude Grezka, and Farideh Tavazoee. Frame-based detection of figurative language in tweets [application notes]. *IEEE Comput. Intell. Mag.*, 14(4):77–88, 2019.

[165] Timothy R Apodaca, Kristina M Jackson, Brian Borsari, Molly Magill, Richard Longabaugh, Nadine R Mastroleo, and Nancy P Barnett. Which

individual therapist behaviors elicit client change talk and sustain talk in motivational interviewing? *Journal of substance abuse treatment*, 61:60–65, 2016.

[166] J Richard Landis and Gary G Koch. The measurement of observer agreement for categorical data. *biometrics*, pages 159–174, 1977.

[167] Domenic V Cicchetti and Sara A Sparrow. Developing criteria for establishing interrater reliability of specific items: applications to assessment of adaptive behavior. *American journal of mental deficiency*, 1981.

[168] Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. Minilm: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. *Advances in Neural Information Processing Systems*, 33:5776–5788, 2020.

[169] Laurens Van der Maaten and Geoffrey Hinton. Visualizing data using t-sne. *Journal of machine learning research*, 9(11), 2008.

[170] Sune Rubak, Annelli Sandbæk, Torsten Lauritzen, and Bo Christensen. Motivational interviewing: a systematic review and meta-analysis. *British journal of general practice*, 55(513):305–312, 2005.

[171] Justine Zhang and Cristian Danescu-Niculescu-Mizil. Balancing objectives in counseling conversations: Advancing forwards or looking backwards. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5276–5289, 2020.

[172] Anqi Li, Jingsong Ma, Lizhi Ma, Pengfei Fang, Hongliang He, and Zhenzhong Lan. Towards automated real-time evaluation in text-based counseling. *arXiv preprint arXiv:2203.03442*, 2022.

[173] Jonas Pfeiffer, Andreas Rücklé, Clifton Poth, Aishwarya Kamath, Ivan Vulić, Sebastian Ruder, Kyunghyun Cho, and Iryna Gurevych. Adapterhub: A framework for adapting transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 46–54, 2020.

[174] Tomás Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. In Yoshua Bengio and Yann LeCun, editors, *1st International Conference on Learning Representations, ICLR 2013, Scottsdale, Arizona, USA, May 2-4, 2013, Workshop Track Proceedings*, 2013.

[175] Bohan Li, Yutai Hou, and Wanxiang Che. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90, 2022.

[176] Jingqing Zhang, Yao Zhao, Mohammad Saleh, and Peter Liu. Pegasus: Pre-training with extracted gap-sentences for abstractive summarization. In *International Conference on Machine Learning*, pages 11328–11339. PMLR, 2020.

[177] Hannah Rashkin, David Reitter, Gaurav Singh Tomar, and Dipanjan Das. Increasing faithfulness in knowledge-grounded dialogue with controllable features. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 704–718, 2021.

[178] Ilya Loshchilov and Frank Hutter. Decoupled weight decay regularization. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net, 2019.

[179] Neslihan Iskender, Tim Polzehl, and Sebastian Möller. Best practices for crowd-based evaluation of German summarization: Comparing crowd, expert and automatic evaluation. In *Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems*, pages 164–175, Online, November 2020. Association for Computational Linguistics.

[180] Virginia Braun and Victoria Clarke. Using thematic analysis in psychology. *Qualitative research in psychology*, 3(2):77–101, 2006.

[181] Craig Thomson and Ehud Reiter. A gold standard methodology for evaluating accuracy in data-to-text systems. In *Proceedings of the 13th International Conference on Natural Language Generation*, pages 158–168, 2020.

[182] Sashank Santhanam, Behnam Hedayatnia, Spandana Gella, Aishwarya Padmakumar, Seokhwan Kim, Yang Liu, and Dilek Hakkani-Tür. Rome was built in 1776: A case study on factual correctness in knowledge-grounded response generation. In *EMNLP 2021 Workshop on NLP for Conversational AI*, 2021.

[183] TB Moyers, JK Manuel, and D Ernst. Motivational interviewing treatment integrity coding manual 4.1 (miti 4.1). *Unpublished manual*, 2014.

[184] Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In Marina Meila and Tong Zhang, editors, *Proceedings of the 38th International Conference on Machine Learning, ICML 2021, 18-24 July 2021, Virtual Event*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR, 2021.

[185] Justus J Randolph. Free-marginal multirater kappa (multirater κfree): An alternative to fleiss' fixed-marginal multirater kappa. In *Presented at the Joensuu Learning and Instruction Symposium*, volume 2005, 2005.

[186] Adam Tsakalidis, Federico Nanni, Anthony Hills, Jenny Chim, Jiayu Song, and Maria Liakata. Identifying moments of change from longitudinal user text. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022*, pages 4647–4660. Association for Computational Linguistics, 2022.

[187] Alvan R. Feinstein and Domenic V. Cicchetti. High agreement but low kappa: I. the problems of two paradoxes. *Journal of Clinical Epidemiology*, 43(6):543–549, 1990.

[188] Stavros Assimakopoulos, Rebecca Vella Muskat, Lonneke van der Plas, and Albert Gatt. Annotating for hate speech: The MaNeCo corpus and some input from critical discourse analysis. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 5088–5097, Marseille, France, May 2020. European Language Resources Association.

[189] Susan Prion and Katie Anne Haerling. Making sense of methods and measurement: Spearman-rho ranked-order correlation coefficient. *Clinical Simulation in Nursing*, 10(10):535–536, 2014.

[190] Brendan McMahan, Eider Moore, Daniel Ramage, Seth Hampson, and Blaise Agüera y Arcas. Communication-efficient learning of deep networks from decentralized data. In Aarti Singh and Xiaojin (Jerry) Zhu, editors, *Proceedings of the 20th International Conference on Artificial Intelligence and Statistics, AISTATS 2017, 20-22 April 2017, Fort Lauderdale, FL, USA*, volume 54 of *Proceedings of Machine Learning Research*, pages 1273–1282. PMLR, 2017.

[191] Jiachang Liu, Dinghan Shen, Yizhe Zhang, William B Dolan, Lawrence Carin, and Weizhu Chen. What makes good in-context examples for gpt-3? In *Proceedings of Deep Learning Inside Out (DeeLIO 2022): The 3rd Workshop on Knowledge Extraction and Integration for Deep Learning Architectures*, pages 100–114, 2022.

[192] Saket Sharma, Aviral Joshi, Namrata Mukhija, Yiyun Zhao, Hanoz Bhathena, Prateek Singh, Sashank Santhanam, and Pritam Biswas. Systematic review of effect of data augmentation using paraphrasing on named entity recognition. In *NeurIPS 2022 Workshop on Synthetic Data for Empowering ML Research*, 2022.

[193] Chujie Zheng, Sahand Sabour, Jiaxin Wen, Zheng Zhang, and Minlie Huang. AugESC: Dialogue augmentation with large language models for emotional support conversation. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 1552–1568, Toronto, Canada, July 2023. Association for Computational Linguistics.

[194] Kang Min Yoo, Dongju Park, Jaewook Kang, Sang-Woo Lee, and Woomyoung Park. Gpt3mix: Leveraging large-scale language models for text augmentation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2225–2239, 2021.

[195] Gaurav Sahu, Pau Rodriguez, Issam Laradji, Parmida Atighehchian, David Vazquez, and Dzmitry Bahdanau. Data augmentation for intent classification with off-the-shelf large language models. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 47–57, 2022.

[196] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. Gpteval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*, 2023.

[197] Tom Kocmi and Christian Federmann. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*, 2023.

[198] Jiaan Wang, Yunlong Liang, Fandong Meng, Haoxiang Shi, Zhixu Li, Jinan Xu, Jianfeng Qu, and Jie Zhou. Is chatgpt a good nlg evaluator? a preliminary study. *arXiv preprint arXiv:2303.04048*, 2023.

[199] Zhijing Jin, Geeticka Chauhan, Brian Tse, Mrinmaya Sachan, and Rada Mihalcea. How good is nlp? a sober look at nlp tasks through the lens of social impact. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3099–3113, 2021.

[200] Jón Guðnason and Hrafn Loftsson. Pre-training and evaluating transformer-based language models for icelandic. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 7386–7391, 2022.

[201] Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, 2020.

[202] Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. In Hugo Larochelle, Marc'Aurelio Ranzato, Raia Hadsell, Maria-Florina Balcan, and Hsuan-Tien Lin, editors, *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*, 2020.

[203] Yaoming Zhu, Sidi Lu, Lei Zheng, Jiaxian Guo, Weinan Zhang, Jun Wang, and Yong Yu. Texygen: A benchmarking platform for text generation models.

In Kevyn Collins-Thompson, Qiaozhu Mei, Brian D. Davison, Yiqun Liu, and Emine Yilmaz, editors, *The 41st International ACM SIGIR Conference on Research & Development in Information Retrieval, SIGIR 2018, Ann Arbor, MI, USA, July 08-12, 2018*, pages 1097–1100. ACM, 2018.

[204] Tulika Saha, Saichethan Reddy, Anindya Das, Sriparna Saha, and Pushpak Bhattacharyya. A shoulder to cry on: Towards a motivational virtual assistant for assuaging mental agony. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2436–2449, Seattle, United States, July 2022. Association for Computational Linguistics.

[205] Aziliz Le Glaz, Yannis Haralambous, Deok-Hee Kim-Dufor, Philippe Lenca, Romain Billot, Taylor C Ryan, Jonathan Marsh, Jordan Devylder, Michel Walter, Sofian Berrouiguet, et al. Machine learning and natural language processing in mental health: systematic review. *Journal of Medical Internet Research*, 23(5):e15708, 2021.

[206] Sean A Dennis, Brian M Goodson, and Christopher A Pearson. Online worker fraud and evolving threats to the integrity of mturk data: A discussion of virtual private servers and the limitations of ip-based screening procedures. *Behavioral Research in Accounting*, 32(1):119–134, 2020.

Appendix A

Human Evaluation of Reflections - *Supplement*

In this appendix, we include additional details of the laypeople- and experts-based human evaluation study (§6.3).

A.1 Reflection Sampling for Annotation & Inadequacy of BART

As mentioned briefly in Chapter 6 (Footnotes 3 and 9), human evaluation in the GPT-2 stage included both GPT-2 reflections and BART reflections in practice, since we wanted to diversify synthetic reflections from smaller LMs in the GPT-2 stage.

Overall, for the context in each of the 15 sampled ⟨context, human reflection⟩ pairs, we generated 26 synthetic reflections in total with GPT-2, 26 with BART and 36 with GPT-3. Like with GPT-2 reflections, we effectively reused the BART reflections from §6.2.1. In order to ensure smaller LMs and large LMs were equally present in the human evaluation of synthetic reflections, we randomly sampled (Appendix A.1.1) 9 semantically distinct reflections from the 52 GPT-2/BART reflections and also 9 from the 36 GPT-3 reflections for human evaluation.

Thus, for each ⟨context, human reflection⟩ pair, we created 2 annotation batches that each contained the context, the human reflection and 9 synthetic reflections. The two batches differed in that the synthetic reflections in one batch came from GPT-2 and BART while those in the other batch were from GPT-3. Both batches were later annotated (§6.3.2). In other words, GPT-2 and BART reflections were annotated together in the GPT-2 stage. However, BART reflections were vastly outnumbered by GPT-2 and GPT-3 reflections because they were sampled less frequently due to a lack of diversity (Appendix A.1.2), so we reported only GPT-2 and GPT-3 in the main body for fairness.

Nevertheless, we analyse the annotations on BART-generated synthetic reflections in Appendix A.1.3, but we note that it is limited by the small quantity of BART reflections and therefore in particular should not be used to compare with the findings w.r.t. GPT-2 and GPT-3 reflections.

A.1.1 Reflection Sampling Procedure

We grouped reflections through semantic clustering based on their embeddings[1], so that the reflections in each cluster were semantically almost identical. For example, if two reflections were identical except that one had a "Hmm." at the beginning while the other did not, they were grouped into the same cluster. Afterwards, we randomly sampled 9 clusters from all the GPT-2 and BART reflection clusters, and we similarly sampled 9 GPT-3 reflection clusters. Finally, we drew from each cluster the reflection with the most tokens — deeming it as semantically the richest — for human evaluation.

A.1.2 Lack of Diversity Among BART Reflections

While we generated the same number (26) of GPT-2 and BART reflections for sampling, in practice there was a considerable lack of diversity among BART reflections that led to them being grouped into fewer clusters and therefore less frequently sampled. Specifically, GPT-2 reflections outnumbered BART reflections 4:1, which means the overall BART:GPT-2:GPT-3 reflection quantity ratio was 1:4:5. Therefore, to ensure fairness, we only reported GPT-2 and GPT-3 reflections in the main body, considering their similar quantities.

To illustrate the lack of diversity among BART reflections, we measure the lexical diversity of synthetic reflections from GPT-2/BART/GPT-3 with Self-BLEU [203], and we use average pairwise embedding similarity to measure semantic diversity.

Self-BLEU is based on BLEU [63] which measures the lexical similarity between two sentences at the n-gram level ($n \in \{1, 2, 3, \cdots\}$). Self-BLEU takes all pairs of generated texts (in our case, reflections for the same context), calculates the BLEU score for each pair, and averages the pairwise BLEU scores. Thus, lower Self-BLEU indicates higher diversity among the generated texts. We follow [203] in reporting 2-, 3-, 4-, and 5-gram-level Self-BLEU[2] for BART, GPT-2 and GPT-3 reflections in Table A.1. Clearly, BART reflections are substantially more homogeneous than those from GPT-2 and GPT-3. For example, the Self-BLEU-4 of BART is at 40.70, compared to the drastically lower 4.49 of GPT-2 and 12.02 of GPT-3.

To compute average pairwise embedding similarity, we first A) compute the cosine similarity between the embeddings (from the same embedding model used for clustering) of the two sequences in each pair of generated reflections for the same context, and then B) average the similarities of all pairs. As shown in Table A.1, the

[1] We used the SentenceTransformers package (`https://www.sbert.net/`) and `all-mpnet-base-v2` [202] as the embedding model.

[2] We calculate Self-BLEU based on NLTK's (`https://www.nltk.org/`) BLEU implementation.

Table A.1: Overview of lexical (Self-BLEU) & semantic (average pairwise embedding similarity) diversity among reflections generated by different models. Lower values indicate more diversity.

	BART	**GPT-2**	**GPT-3**
Lexical Diversity			
Self-BLEU-2	48.63	8.44	17.74
Self-BLEU-3	44.36	5.77	14.10
Self-BLEU-4	40.70	4.49	12.02
Self-BLEU-5	37.38	3.75	10.55
Semantic Diversity			
Average Pairwise Embedding Similarity	0.6952	0.3034	0.4666

Table A.2: Label distribution for BART-generated reflections.

	Laypeople	**Experts**
Coherent	38.1%	77.4%
Dialogue-Contradicting	1.8%	3.6%
Malformed	1.8%	0.6%
Off-Topic	3.0%	2.4%
On-Topic But Unverifiable	13.7%	3.6%
Parroting	41.7%	12.5%

semantic similarity between BART reflections is also considerably higher compared to GPT-2 and GPT-3.

A.1.3 Label Distribution for BART Reflections

We show in Table A.2 the distribution of labels given to BART reflections. Notably, laypeople and experts show considerable difference ($\Delta = 39\%$) in the proportion of *Coherent* labels, which is substantially higher compared to GPT-2 ($\Delta = 3\%$) and GPT-3 ($\Delta = 17\%$) shown in Figure 6.5.

Upon a closer inspection, it is clear that most of the laypeople-experts divide in coherence annotation can be attributed to Parroting, which is used considerably more ($\Delta = 29\%$) by laypeople than experts. This again echoes the observation in §6.3.4 that laypeople are more strict about Parroting than experts.

Qualitatively, Table A.3 shows the BART reflection for the case study dialogue (Table 6.13), which clearly mirrors the last client utterance. Matching the trend discussed above, the reflection is annotated by every layperson as Parroting but by every expert as *Coherent*.

This finding, together with the low diversity among BART reflections (Ap-

Table A.3: BART-generated reflection for the case study dialogue (Table 6.13) and its annotations. **L1/L2/.../L9**: 9 laypeople. **E1/E2/.../E9**: 9 experts. Red: incoherence error category.

Context		
··· (previous turns)		
Therapist: And, uh, sounds like you're-you're pretty upset with your mom for-for doing that? **Client**: I am. **Therapist**: Yeah. **Client**: I mean, it's my stuff, I don't know why she's–		
BART: Okay. So, it's your stuff.		
L1: Parroting **E3**: *Coherent*	**L4**: Parroting **E4**: *Coherent*	**L7**: Parroting **E5**: *Coherent*

Table A.4: Waiting period lengths (number of days) between the two stages.

	Mean	Standard Deviation
Laypeople	39.1	7.8
Experts	6.9	3.1

pendix A.1.2), shows that BART tends to repeat or rephrase a part of the dialogue context and does not show considerable deviation from this pattern under different decoding parameters. Empirically, this is also our observation of BART reflections in general.

A.2 Waiting Period Between Stages

Initially, we conducted the ⟨Laypeople, GPT-2 stage⟩ setup. We then collected GPT-3-generated reflections and invited the same laypeople for GPT-3 stage annotation. As those two stages were not planned together, there was about a one-month period in between.

Upon discovering the shifting human reflection annotations (§6.3.4) from the laypeople's results, we recruited the experts to investigate whether the phenomenon was limited to laypeople. Due to time constraint, we were only able to enforce a minimum waiting period of 3 days between the two stages for the experts.

The mean and standard deviation of the waiting period lengths of each annotator group are shown in Table A.4. On average, laypeople had a 39-day gap between the two stages while experts had 7 days.

To probe whether the waiting period difference had an effect, we requested the annotators to fill out a post-annotation questionnaire, where we asked the question "While you were annotating in Phase 2 (i.e., GPT-3 stage), did you remember see-

Table A.5: Answers given to the post-annotation question "While you were annotating in Phase 2 (i.e., GPT-3 stage), did you remember seeing any response candidate that you had seen in Phase 1 (i.e., GPT-2 stage)?".

	Yes	No	Maybe
Laypeople	3	1	3
Experts	3	3	1

ing any response candidate that you had seen in Phase 1 (i.e., GPT-2 stage)?". We received 7 valid responses from the 8 laypeople who had annotated recurring human reflections, and similarly 7 from the 8 experts that had had recurring human reflections in their workload. Their answers are shown in Table A.5.

Clearly, the same number (3) of experts and laypeople remembered seeing recurring human reflections in the GPT-3 stage, but 3 experts answered "No" while 3 laypeople answered "Maybe", which is not surprising since the longer waiting period may have caused more laypeople not to be able to recall exactly. Nevertheless, the fact that the same number of experts and laypeople were positive about seeing recurring human reflections shows that the waiting period for experts was not overly short and may have in fact been sufficient. This is further evidenced by the finding (§6.3.4) that laypeople and experts are similarly consistent in annotating recurring human reflections.

A.3 Label Distribution for Differently Generated Reflections

Table A.6 shows the distribution of *Coherent* and error labels for synthetic reflections from GPT-2 and GPT-3 under different generation settings.

For GPT-2 reflections, larger p values in nucleus decoding lead to less coherent reflections, especially when $p \in \{0.8, 0.95\}$. This is unsurprising, since larger p's give the model more freedom in generation and thus also make it more prone to errors.

For GPT-3, reflections generated through textbook-based in-context learning are overall less coherent than reflections generated through `AnnoMI`-based in-context learning. This is not surprising, since test examples themselves are from `AnnoMI`, which means learning examples from `AnnoMI` are more useful in helping the model learn to produce coherent reflections for long dialogue contexts.

Among reflections from GPT-3 (textbook), simple reflections are overall more often annotated as Parroting than complex ones, especially by laypeople. This is likely because simple reflections mostly repeat/rephrase what the client said, which may appear repetitive to a layperson when an expert would more likely consider it good practice (§6.3.4).

Table A.6: Label distribution on synthetic reflections from GPT-2 and GPT-3 under different generation settings. **L**: laypeople. **E**: experts. All GPT-3 reflections are generated with nucleus decoding.

GPT-2 Using Greedy and Beam Decoding

	Greedy		Beam Search	
	L	**E**	**L**	**E**
Coherent	50.0%	66.7%	50.0%	44.4%
Dialogue-Contradicting	16.7%	0.0%	27.8%	27.8%
Malformed	8.3%	0.0%	0.0%	11.1%
Off-Topic	25.0%	0.0%	2.8%	0.0%
On-Topic But Unverifiable	0.0%	33.3%	8.3%	0.0%
Parroting	0.0%	0.0%	11.1%	16.7%

GPT-2 Using Nucleus Decoding

	$p = 0.4$		$p = 0.6$		$p = 0.8$		$p = 0.95$	
	L	**E**	**L**	**E**	**L**	**E**	**L**	**E**
Coherent	56.1%	54.5%	52.4%	54.8%	31.8%	21.2%	22.2%	18.5%
Dialogue-Contradicting	12.1%	7.6%	6.5%	11.3%	8.1%	6.1%	11.8%	4.9%
Malformed	4.5%	5.3%	6.5%	3.6%	21.0%	25.8%	27.9%	26.5%
Off-Topic	12.9%	16.7%	13.7%	8.3%	28.0%	25.0%	30.3%	37.7%
On-Topic But Unverifiable	10.6%	14.4%	16.1%	21.4%	9.6%	22.0%	7.7%	12.3%
Parroting	3.8%	1.5%	4.8%	0.6%	1.5%	0.0%	0.0%	0.0%

Simple Reflections From GPT-3 - Textbook Examples for In-Context Learning

	$p = 0.4$		$p = 0.6$		$p = 0.8$		$p = 0.95$	
	L	**E**	**L**	**E**	**L**	**E**	**L**	**E**
Coherent	38.5%	74.4%	40.5%	66.7%	37.5%	72.9%	59.5%	83.3%
Dialogue-Contradicting	5.1%	3.8%	7.1%	0.0%	4.2%	0.0%	2.4%	0.0%
Malformed	5.1%	2.6%	2.4%	0.0%	0.0%	1.0%	1.2%	0.0%
Off-Topic	10.3%	0.0%	2.4%	2.4%	4.2%	1.0%	4.8%	0.0%
On-Topic But Unverifiable	0.0%	1.3%	4.8%	7.1%	2.1%	4.2%	13.1%	7.1%
Parroting	41.0%	17.9%	42.9%	23.8%	52.1%	20.8%	19.0%	9.5%

Complex Reflections From GPT-3 - Textbook Examples for In-Context Learning

	$p = 0.4$		$p = 0.6$		$p = 0.8$		$p = 0.95$	
	L	**E**	**L**	**E**	**L**	**E**	**L**	**E**
Coherent	73.3%	75.6%	66.7%	82.2%	57.8%	80.0%	47.6%	90.5%
Dialogue-Contradicting	12.2%	11.1%	2.2%	0.0%	2.2%	3.3%	0.0%	0.0%
Malformed	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	4.8%	0.0%
Off-Topic	2.2%	0.0%	0.0%	0.0%	2.2%	0.0%	0.0%	0.0%
On-Topic But Unverifiable	3.3%	6.7%	15.6%	11.1%	13.3%	4.4%	9.5%	4.8%
Parroting	8.9%	6.7%	15.6%	6.7%	24.4%	12.2%	38.1%	4.8%

Reflections From GPT-3 - `AnnoMI` Examples for In-Context Learning

	$p = 0.4$		$p = 0.6$		$p = 0.8$		$p = 0.95$	
	L	**E**	**L**	**E**	**L**	**E**	**L**	**E**
Coherent	75.6%	86.7%	85.7%	92.9%	95.2%	95.2%	82.1%	89.7%
Dialogue-Contradicting	6.7%	4.4%	2.4%	0.0%	0.0%	0.0%	3.8%	0.0%
Malformed	0.0%	4.4%	0.0%	4.8%	0.0%	4.8%	0.0%	2.6%
Off-Topic	0.0%	2.2%	0.0%	0.0%	0.0%	0.0%	2.6%	2.6%
On-Topic But Unverifiable	8.9%	2.2%	7.1%	2.4%	0.0%	0.0%	2.6%	2.6%
Parroting	8.9%	0.0%	4.8%	0.0%	4.8%	0.0%	9.0%	2.6%

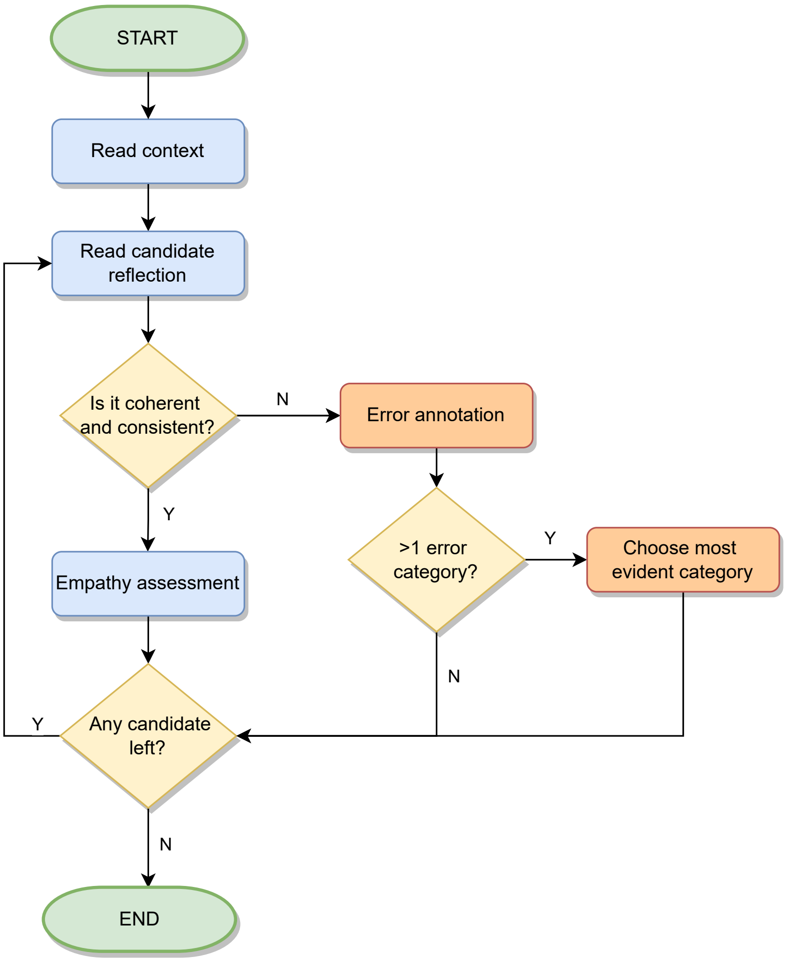


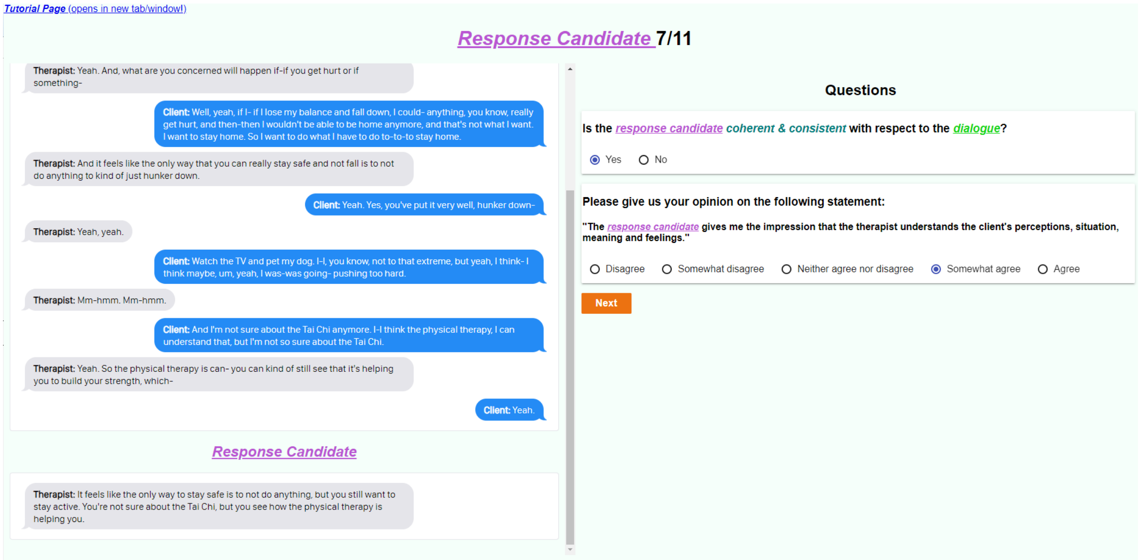
Figure A.1: Full annotation flow for one batch. In this work, we do not investigate annotations w.r.t. empathy assessment or the most evident error category.

Finally, we note that the *Coherent* rates of GPT-3 reflections can vary considerably under different nucleus decoding p's but without a clear trend, which we leave to future work to probe.

A.4 Full Annotation Flow

In practice, each annotation batch contained some parts that are not investigated in this study, which are therefore not shown in the main body. The complete annotation flow is detailed below.

As shown in Figure A.1, a batch starts with the annotator reading the context. Then, the annotator reads one reflection and chooses Yes/No regarding whether it is coherent and context-consistent. If the answer is Yes, the annotator assesses the level of empathy displayed in the reflection. If the answer is No, the annotator selects



Tutorial Page (opens in new tab/window!)

Response Candidate 7/11

Therapist: Yeah. And, what are you concerned will happen if-if you get hurt or if something-

Client: Well, yeah, if I- if I lose my balance and fall down, I could- anything, you know, really get hurt, and then-then I wouldn't be able to be home anymore, and that's not what I want. I want to stay home. So I want to do what I have to do to-to-to stay home.

Therapist: And it feels like the only way that you can really stay safe and not fall is to not do anything to kind of just hunker down.

Client: Yeah. Yes, you've put it very well, hunker down-

Therapist: Yeah, yeah.

Client: Watch the TV and pet my dog. I-I, you know, not to that extreme, but yeah, I think- I think maybe, um, yeah, I was-was going- pushing too hard.

Therapist: Mm-hmm. Mm-hmm.

Client: And I'm not sure about the Tai Chi anymore. I-I think the physical therapy, I can understand that, but I'm not so sure about the Tai Chi.

Therapist: Yeah. So the physical therapy is can- you can kind of still see that it's helping you to build your strength, which-

Client: Yeah.

Response Candidate

Therapist: It feels like the only way to stay safe is to not do anything, but you still want to stay active. You're not sure about the Tai Chi, but you see how the physical therapy is helping you.

Questions

Is the *response candidate* *coherent & consistent* with respect to the *dialogue*?

● Yes ○ No

Please give us your opinion on the following statement:

"The *response candidate* gives me the impression that the therapist understands the client's perceptions, situation, meaning and feelings."

○ Disagree ○ Somewhat disagree ○ Neither agree nor disagree ● Somewhat agree ○ Agree

Figure A.2: Annotation interface when the annotator annotates a reflection as coherent & consistent.

one or more error categories that apply, and in the case of multiple selected errors the annotator further pinpoints the most evident one. Afterwards, the annotator proceeds to annotate the next reflection in the same steps, and the batch ends when all its reflections have been annotated.

A.5 Annotation Interface

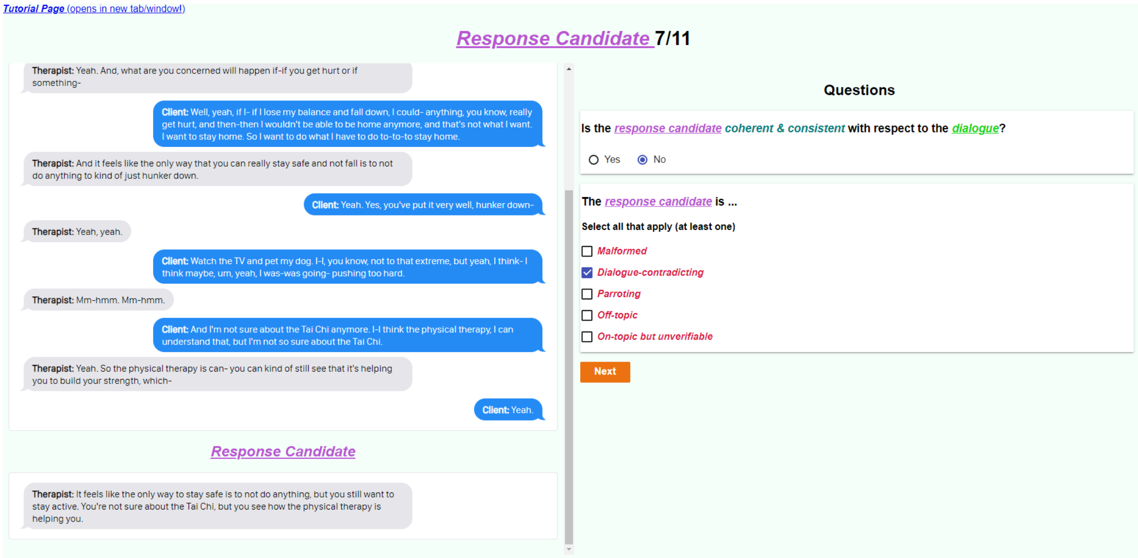
The annotation process takes place in the Mechanical Turk Sandbox[3]. Details of the annotation interface are shown in Figures A.2, A.3 and A.4. We note that there is a purposefully off-topic reflection in each batch as an anti-scam mechanism, which is why there appear to be 11 reflections instead of 10 to annotate in those figures.

A.6 Limitations

The main limitation of this work is the quantity of annotated human reflections. Overall, 15 human reflections are annotated, which are outnumbered more than 7:1 by GPT-2 reflections and 9:1 by GPT-3 reflections. If there were more human reflections annotated, we may be able to confirm, among other potential findings, that GPT-3 reflections were indeed statistically significantly more often annotated as *Coherent* compared to human reflections.

We also note that the laypeople had a longer between-stage waiting period than the experts, because we could not enforce a similarly long waiting period for the

[3] `https://workersandbox.mturk.com/`



Tutorial Page (opens in new tab/window!)

Response Candidate 7/11

Therapist: Yeah. And, what are you concerned will happen if-if you get hurt or if something-

Client: Well, yeah, if I- if I lose my balance and fall down, I could- anything, you know, really get hurt, and then-then I wouldn't be able to be home anymore, and that's not what I want. I want to stay home. So I want to do what I have to do to-to-to stay home.

Therapist: And it feels like the only way that you can really stay safe and not fall is to not do anything to kind of just hunker down.

Client: Yeah. Yes, you've put it very well, hunker down-

Therapist: Yeah, yeah.

Client: Watch the TV and pet my dog. I-I, you know, not to that extreme, but yeah, I think- I think maybe, um, yeah, I was-was going- pushing too hard.

Therapist: Mm-hmm. Mm-hmm.

Client: And I'm not sure about the Tai Chi anymore. I-I think the physical therapy, I can understand that, but I'm not so sure about the Tai Chi.

Therapist: Yeah. So the physical therapy is can- you can kind of still see that it's helping you to build your strength, which-

Client: Yeah.

Response Candidate

Therapist: It feels like the only way to stay safe is to not do anything, but you still want to stay active. You're not sure about the Tai Chi, but you see how the physical therapy is helping you.

Questions

Is the *response candidate* *coherent & consistent* with respect to the *dialogue*?

○ Yes ● No

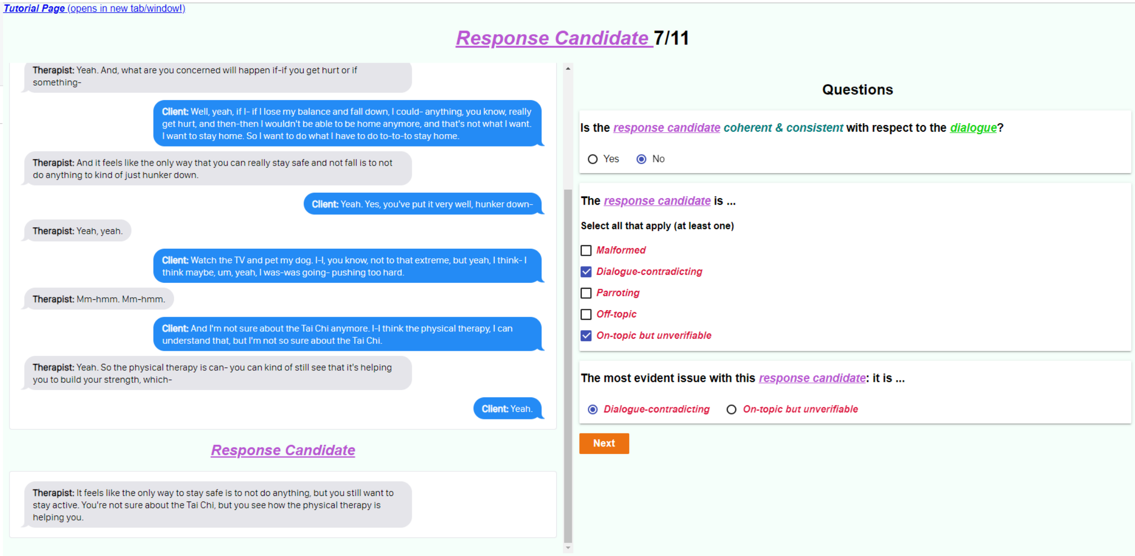
The *response candidate* is ...

Select all that apply (at least one)

☐ *Malformed*
☑ *Dialogue-contradicting*
☐ *Parroting*
☐ *Off-topic*
☐ *On-topic but unverifiable*

Next

Figure A.3: Annotation interface when the annotator annotates a reflection as incoherent/inconsistent and chooses one error category.



Tutorial Page (opens in new tab/window!)

Response Candidate 7/11

Therapist: Yeah. And, what are you concerned will happen if-if you get hurt or if something-

Client: Well, yeah, if I- if I lose my balance and fall down, I could- anything, you know, really get hurt, and then-then I wouldn't be able to be home anymore, and that's not what I want. I want to stay home. So I want to do what I have to do to-to-to stay home.

Therapist: And it feels like the only way that you can really stay safe and not fall is to not do anything to kind of just hunker down.

Client: Yeah. Yes, you've put it very well, hunker down-

Therapist: Yeah, yeah.

Client: Watch the TV and pet my dog. I-I, you know, not to that extreme, but yeah, I think- I think maybe, um, yeah, I was-was going- pushing too hard.

Therapist: Mm-hmm. Mm-hmm.

Client: And I'm not sure about the Tai Chi anymore. I-I think the physical therapy, I can understand that, but I'm not so sure about the Tai Chi.

Therapist: Yeah. So the physical therapy is can- you can kind of still see that it's helping you to build your strength, which-

Client: Yeah.

Response Candidate

Therapist: It feels like the only way to stay safe is to not do anything, but you still want to stay active. You're not sure about the Tai Chi, but you see how the physical therapy is helping you.

Questions

Is the *response candidate* *coherent & consistent* with respect to the *dialogue*?

○ Yes ● No

The *response candidate* is ...

Select all that apply (at least one)

☐ *Malformed*
☑ *Dialogue-contradicting*
☐ *Parroting*
☐ *Off-topic*
☑ *On-topic but unverifiable*

The most evident issue with this *response candidate*: it is ...

● *Dialogue-contradicting* ○ *On-topic but unverifiable*

Next

Figure A.4: Annotation interface when the annotator annotates a reflection as incoherent/inconsistent and chooses multiple error categories.

experts due to practical reasons (Appendix A.2). While an ideal setup would keep the same waiting period duration, Appendix A.2 (survey results) and §6.3.4 (laypeople and experts are overall similarly consistent when annotating recurring human reflections) show that the duration difference is not critical.

Furthermore, we adopted sequential annotation for reflections within a batch to make the interface easier to navigate for the human annotators, but this also means that the early samples in a batch might indirectly affect the annotation of the later samples. We leave more investigation of this to future work.

A.7 Data Use & Creation

We leveraged `AnnoMI`, a dataset available under the Public Domain license. We used it for research purposes, which is consistent with its intended use. While `AnnoMI` contains therapy dialogues, the data does not reveal personal information since the dialogues are transcripts of professionally produced MI demonstrations. The dataset does not reveal demographic information, but the dialogues are in English and we observe that the dialogues seem to be set in English-speaking countries.

Based on `AnnoMI`, we created a dataset of human annotations w.r.t. coherence of reflections, and we released it[4] under the CC BY-NC license, which is also compatible with the access conditions of `AnnoMI`. The human annotations do not reveal any information of the laypeople or experts, and we use L1∼9 to represent the 9 laypeople and E1∼9 to represent the 9 experts. We discuss the demographic information of the annotators in Appendix A.8.4.

A.8 Ethics Statement

In this section, we briefly discuss the ethical aspects of our experiments.

A.8.1 Ethical Review

Prior to the experiments, our methodology and materials underwent ethical review by our institution's Ethics Board. The proposal was considered ethically compliant and accepted without major revisions.

A.8.2 Risks

Our work inspects the annotation differences between laypeople and experts in the counselling domain (MI and reflections in particular). With these premises, it could be seen as a message that therapy can be fully automated, or that laypeople can replace therapists in creating such systems and generative models could act as "virtual counsellors". We acknowledge that past work inspected similar options (e.g., [204]), but we take distance from it. Our work is framed as modelling technological advancements that are solely directed at therapist training. We foresee the use of neural NLG as promising in counselling, but only for supporting trainees. We also point out that previous work has shown why replacing mental health practices with LMs (or AI in general) should not be considered [205].

[4] Available at `https://github.com/uccollab/expert_laypeople_reflection_annotation`.

A.8.3 Information and Consent

Prior to starting the annotation, both laypeople and experts received an electronic information sheet containing details on the task, purpose of research, workload and pay. This also included the fact that data would be made available for future research, in accordance with data anonymisation requirements.

Upon starting the annotation, annotators were prompted with a mandatory consent form to confirm their understanding of the terms and conditions and their willingness to take part in the annotation. Annotators were also given an email contact in case of problems during the annotation or any other query. Annotators were automatically prevented from doing the annotation if they did not provide consent.

A.8.4 Demographic Information of Annotators

All annotators were highly proficient in English, which is the language of the dialogues. 5 out of the 9 laypeople were based in the Netherlands while the other 4 resided in Italy. Among the experts, 4 were based in the UK, 1 in the Netherlands, 1 in Hungary, 1 in Italy and 2 in Sweden.

We recruited laypeople who were known to us, as this allowed active monitoring of the annotation task, hence ensuring high annotation quality. While this approach is different from other standard ones (such as using crowdsourcing platforms), we argue that the focus of this work is to understand if fully committed laypeople can be valid annotators, which can be challenging considering the annotation quality issues that crowdsourcing platforms suffer from [206].

We also note that the group of laypeople is diverse in demographics and educational backgrounds. Specifically, the group includes people of 5 nationalities in their 20s, 30s and 40s who range from bachelor's student to professional with a PhD.

To verify the generalisability of our laypeople-based evaluation, future work may replicate our setup on crowdworkers and compare the resulting annotations with ours.

A.8.5 Remuneration

The annotation workload was made explicit in the task (a total of 5 annotation batches in each stage, with a detailed description of what a batch contained). Annotators were given 30 minutes to complete each annotation batch: laypeople received 19.5 USD/hour, while experts received 21.6 USD/hour. This difference was motivated by the generally higher hourly pay of experts. The remuneration was considerably ($> 50\%$) higher than the minimum wage levels of the countries of residence of the annotators. It also took most annotators much less than 30 minutes (e.g., 10 to 15 minutes) to complete a batch, so the effective hourly remuneration was higher than 19.5/21.6 USD.

A.8.6 Data Anonymisation

No personal data about the annotators was kept stored at the end of the experiment. During the annotation process, no annotator ever got in touch with anyone involved in the experiments except for the researchers.