European Journal of Radiology Artificial Intelligence 5 (2026) 100066

Contents lists available at ScienceDirect

European Journal of Radiology Artificial Intelligence

journal homepage: www.sciencedirect.com/journal/european-journal-of-radiology-artificial-intelligence

PARROT, an open multilingual radiology reports dataset

Bastien Le Guellec [a,b,*], Kokou Adambounou [c], Lisa C. Adams [d], Thibault Agripnidis [e], Sung Soo Ahn [f], Radhia Ait Chalal [g], Tugba Akinci D'Antonoli [h,i], Philippe Amouyel [j,k], Henrik Andersson [l], Raphaël Bentegeac [j,k], Claudio Benzoni [m], Antonino Andrea Blandino [n], Felix Busch [d], Elif Can [o], Riccardo Cau [p], Armando Ugo Cavallo [q], Christelle Chavihot [r], Erwin Chiquete [s], Renato Cuocolo [t], Eugen Divjak [u], Barbara Dziadkowiec-Macek [v], Armel Elogne [w], Salvatore Claudio Fanni [x], Carlos Ferrarotti [y], Claudia Fossataro [z], Federica Fossataro [aa], Katarzyna Fułek [ab], Michał Fułek [ac], Paweł Gać [ad], Martyna Gachowska [ab], Ignacio García-Juárez [ae], Marco Gatti [af], Natalia Gorelik [ag], Alexia Maria Goulianou [ah], Aghiles Hamroun [j,k], Nicolas Herinirina [ai], Quentin Holay [aj], Gordana Ivanac [u], Felipe Kitamura [ak,al], Michail E. Klontzas [am,1], Anna Kompanowska [an], Rafał Kompanowski [ao], Krzysztof Kraik [ap], Dominik Krupka [ap], Alexandre Lefèvre [a], Tristan Lemke [d], Maximilian Lindholz [aq], Piotr Macek [ac], Marcus Makowski [d], Luigi Mannacio [ar], Aymen Meddeb [as], Lukas Müller [at], Antonio Natale [au], Béatrice Nguema Edzang [av], Adriana Ojeda [aw], Yae Won Park [f], Federica Piccione [af], Andrea Ponsiglione [ar], Małgorzata Poręba [ax], Rafał Poręba [ad], Philipp Prucker [d], Jean-Pierre Pruvo [a,b], Rosa alba Pugliesi [n,ay], Feno Hasina Rabemanorintsoa [az], Vasileios Rafailidis [ba], Katarzyna Resler [ab], Jan Rotkegel [ap], Luca Saba [p], Ezann Siebert [bb], Arnaldo Stanzione [ar], Ali Fuat Tekin [bc], Liz Toapanta-Yanchapaxi [af], Matthaios Triantafyllou [ah], Ekaterini Tsaoulia [ba], Szymon Urban [bd], Evangelia Vassalou [ah], Federica Vernuccio [n], Weilang Wang [be], Johan Wassélius [l], Adrian Włodarczak [bd], Szymon Włodarczak [bd], Andrzej Wysocki [ad], Lina Xu [aq], Tomasz Zatoński [ab], Shuhang Zhang [be], Sebastian Ziegelmayer [d], Grégory Kuchcinski [a,b], Keno K. Bressem [d]

[a] *Department of Neuroradiology, Lille University Hospital, Salengro Hospital, Lille, France*
[b] *U1172–LilNCog-Lille Neuroscience & Cognition, Université Lille, Lille, France*
[c] *Campus University Hospital Centre, Department of Radiology and Medical Imaging, Lome, Togo*
[d] *Department of Diagnostic and Interventional Radiology, Klinikum rechts der Isar, TUM University Hospital, Technical University of Munich, Munich, Germany*
[e] *Interventional Radiology Section, Department of Medical Imaging, University Hospital Timone, AP-HM, 264 Rue Saint-Pierre, Marseille 13005, France*
[f] *Department of Radiology and Research Institute of Radiological Science and Center for Clinical Imaging Data Science,Yonsei University College of Medicine, Seoul, South Korea*
[g] *Department of Radiology, Bab El-Oued University Hospital, Algiers, Algeria*
[h] *Department of Diagnostic and Interventional Neuroradiology, University Hospital Basel, France*
[i] *Department of Pediatric Radiology, University Children's Hospital Basel, France*
[j] *U1167 RID-AGE, Pasteur Institute of Lille, Inserm, Lille University, Lille, France*
[k] *Public Health – Epidemiology Department, Lille University Hospital Center, Lille, France*
[l] *Department of Medical Imaging and Physiology, Skåne University Hospital, Lund 221 85, Sweden*
[m] *Institute of AI and Informatics in Medicine (AIIM), TUM University Hospital, Technical University of Munich, Munich, Germany*
[n] *Section of Radiology, Department of Biomedicine, Neuroscience and Advanced Diagnostics (BiND), University of Palermo, Via del Vespro 129, Palermo 90127, Italy*
[o] *Department of Diagnostic and Interventional Radiology, Medical Center – University of Freiburg, Faculty of Medicine, University of Freiburg, Germany*
[p] *Department of Radiology, Azienda Ospedaliero Universitaria (A.O.U.), di Cagliari - Polo di Monserrato s.s. 554 Monserrato, Cagliari, Italy*
[q] *Division of Radiology, Istituto Dermopatico dell'Immacolata (IDI) IRCCS, Rome, Italy*

* Corresponding author at: Department of Neuroradiology, Lille University Hospital, Salengro Hospital, Lille, France.
E-mail address: bastien.leguellec@chu-lille.fr (B. Le Guellec).

https://doi.org/10.1016/j.ejrai.2025.100066
Received 30 September 2025; Received in revised form 12 December 2025; Accepted 19 December 2025
Available online 22 December 2025

[r] Department of Radiology, Hôpital Instruction des Armées, Libreville, Gabon
[s] Department of Neurology, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, Mexico City, Mexico
[t] Department of Medicine, Surgery and Dentistry, University of Salerno, Baronissi, Italy
[u] Department of Diagnostic and Interventional Radiology, University Hospital Dubrava, Avenija Gojka Suska 6, Zagreb 10000, Croatia
[v] Department of Physiology and Pathophysiology, Wroclaw Medical University, Wroclaw, Poland
[w] Department of Radiology, Hôpital militaire d'Abidjan, Abidjan, Côte d'Ivoire 28 BP 1303, France
[x] Department of Translational Research, Academic Radiology, University of Pisa, Pisa, Italy
[y] Department of Diagnostic Imaging, Center for Medical Education and Clinical Research "Norberto Quirno" (CEMIC), Autonomous City of Buenos Aires, Argentina
[z] Department of Ophthalmology, Catholic University "Sacro Cuore", Rome, Italy
[aa] Department of Ophthalmology, ASST Fatebenefratelli Sacco, Milan, Italy
[ab] Department and Clinic of Otolaryngology, Head and Neck Surgery, Wroclaw Medical University, Wroclaw 50-556, Poland
[ac] Department and Clinic of Diabetology, Hypertension and Internal Diseases, Institute of Internal Diseases, Wroclaw Medical University, Wroclaw, Poland
[ad] Department of Radiology and Diagnostic Imaging, 4th Military Hospital, Wroclaw, Poland
[ae] Department of Gastroenterology, Instituto Nacional de Ciencias Médicas y Nutrición Salvador Zubirán, México DF, Mexico
[af] Department of Surgical Sciences, Radiology Unit, University of Turin, Turin, Italy
[ag] Department of Radiology, McGill University Health Center, Quebec, Montreal, Canada
[ah] Department of Medical Imaging, University Hospital of Heraklion, Crete, Voutes 71110, Greece
[ai] Department of Radiology, Tanambao University Hospital, Antsiranana, Madagascar
[aj] Department of Radiology, Sainte-Anne Teaching military Hospital, Toulon, France
[ak] Bunkerhill Health, San Francisco, CA, United States
[al] Department of Diagnostic Imaging, Universidade Federal de São Paulo (UNIFESP), São Paulo, Brazil
[am] Artificial Intelligence and Translational Imaging (ATI) Lab, Department of Radiology, School of Medicine, University of Crete, Heraklion, Greece
[an] Department of Pediatrics, Klodzko County Hospital, Klodzko, Poland
[ao] Orthopedics and Traumatology Department of the Musculoskeletal System, Specialist Medical Centre, Polanica-Zdroj, Poland
[ap] Faculty of medicine, Wroclaw Medical University, Wroclaw, Poland
[aq] Department of Radiology, Charité University Hospital Berlin, Berlin, Germany
[ar] Department of Advanced Biomedical Sciences, University of Naples Federico II, Naples, Italy
[as] Department of Neuroradiology, Charité University Hospital Berlin, Berlin, Germany
[at] Department of Diagnostic and Interventional Radiology, University Medical Center of the, Johannes Gutenberg-University Mainz, Langenbeckst. 1, Mainz 55131, Germany
[au] Department of Radiological and Hematological Sciences, Section of Radiology, Università Cattolica del Sacro Cuore, L.go Francesco Vito 1, Rome 00168, Italy
[av] Department of Radiology, Hôpital Instruction des Armées, Libreville, Gabon
[aw] Departamento de Neuroradiología, DMO, Rosario, Argentina
[ax] Department of Paralympic Sport Wroclaw University of Health and Sport Sciences, Wroclaw, Poland
[ay] Division of Health Care Sciences Center for Clinical Research and Management Education Dresden, Dresden International University, Dresden, Germany
[az] Department of Radiology, Morafeno Toamasina University Hospital, Toamasina, Madagascar
[ba] Department of Radiology, Aristotle University of Thessaloniki, School of Medicine, AHEPA University General Hospital, Thessaloniki, Greece
[bb] Department of Ophthalmology, Sir Charles Gairdner Hospital, Perth, Western Australia, Australia
[bc] Department of Radiology, Basaksehir Cam & Sakura City Hospital, Istanbul 34480, Turkey
[bd] Department of Cardiology, The Copper Health Center, Lubin, Poland
[be] Department of Radiology, Zhongda Hospital, Southeast University, 87 Ding, Nanjing, Jiaqiao 210009, PR China

A R T I C L E I N F O

Keywords:
ChatGPT
Large Language Models
Dataset
Multilingual
Artificial Intelligence

A B S T R A C T

Aims: To develop and validate PARROT (Polyglottal Annotated Radiology Reports for Open Testing), a multicentric, open-access dataset of fictional radiology reports spanning multiple languages for testing natural language processing applications in radiology.
Methods: From May to September 2024, radiologists were invited to contribute fictional radiology reports following their standard reporting practices. Contributors provided at least 20 reports with associated metadata including anatomical region, imaging modality, clinical context, and for non-English reports, English translations. All reports were assigned ICD-10 codes. A human vs. AI report differentiation study was conducted with 154 participants (radiologists, healthcare professionals, and non-healthcare professionals) assessing whether reports were human-authored or AI-generated.
Results: The dataset comprises 2658 radiology reports from 76 authors across 21 countries and 13 languages. Reports cover multiple imaging modalities (CT: 36.1 %, MRI: 22.8 %, radiography: 19.0 %, ultrasound: 16.8 %) and anatomical regions, with chest (19.9 %), abdomen (18.6 %), head (17.3 %), and pelvis (14.1 %) being most prevalent. In the differentiation study, participants achieved 53.9 % accuracy (95 % CI: 50.7 %-57.1 %) in distinguishing between human and AI-generated reports, with radiologists performing significantly better (56.9 %, 95 % CI: 53.3 %-60.6 %, $p < 0.05$) than other groups.
Conclusion: PARROT represents the largest open multilingual radiology report dataset, enabling testing and validation of natural language processing applications across linguistic, geographic, and clinical boundaries without privacy constraints.

1. Introduction

Radiology reports are the communication bridge between radiologists and clinicians, capturing complex visual findings in actionable text that directly informs patient care decisions [1]. The quality and clarity of these reports significantly impact diagnostic accuracy and treatment outcomes, with communication breakdowns representing a notable source of medical errors [2–4].

Large language models (LLMs) have demonstrated remarkable

[1] Michail Klontzas serves as Advisory Editor for European Journal of Radiology Artificial Intelligence (EJR AI). He was involved in the review or selection process for this article.Michail Klontzas serves as Advisory Editor for European Journal of Radiology Artificial Intelligence (EJR AI). He was involved in the review or selection process for this article.

capabilities in enhancing radiology workflows [5–7]. Recent studies show these models can automatically structure free-text findings into standardized formats suitable for data sharing [8,9], simplify complex radiological terminology for non-specialists [10], and detect inconsistencies and errors with performance comparable to expert radiologists [11]. Such capabilities hold substantial promise for improving communication efficiency and report quality in clinical practice.

However, a considerable barrier to deploying LLMs is their English-language centricity. Widely used datasets such as MIMIC-IV and MIMIC-CXR contain thousands of radiology reports but represent only English-language practice [12,13]

Privacy concerns and stringent data-sharing regulations constrain studies involving non-English radiology reports to proprietary datasets that remain inaccessible to the wider research community [9]. This absence of openly available multilingual datasets undermines scientific reproducibility and comparability across different healthcare systems [14].The consequences of this linguistic bias extend beyond simple language barriers. Radiology reporting practices exhibit substantial diversity worldwide, not just in language but in reporting styles, preferred formats, terminology, and conventions. These variations are so pronounced that evaluation metrics developed for one region often fail to generalize across reports from different countries [15]]. These challenges are compounded by inherent limitations in LLMs that make even advanced models show performance disparities across languages, with capabilities in low-resource languages lagging behind those in English [15,16].

Addressing these gaps requires developing truly multilingual resources that reflect authentic, native-language radiology practices. Preserving language-specific medical terminology is crucial for clinical accuracy [17], while capturing diverse reporting conventions ensures AI tools align with existing clinical workflows. Furthermore, with AI-generated reports becoming increasingly prevalent, it is crucial to assess if human-authored reports can be distinguished from AI-generated content, as training or evaluating language models on AI-generated reports may inadvertently propagate subtle inaccuracies.

In this context, we introduce PARROT (Polyglottal Annotated Radiology Reports for Open Testing), a collaborative initiative designed to overcome multilingual barriers in radiology AI research. By bringing together radiologists and AI researchers to curate expert-authored reports from diverse linguistic and clinical environments, PARROT provides an open benchmark for testing inclusive radiology AI tools that serve global healthcare needs effectively and equitably.

2. Methods

The PARROT initiative was launched in May 2024 and ran until September 2024. The project was advertised through multiple channels including social media platforms, professional radiological societies, and direct networking. To ensure diverse global representation, particularly from underrepresented regions, we leveraged the established network of the COMFORT AI survey [18], which facilitated connections with radiologists across diverse geographic areas. Because the dataset consists entirely of fictional reports with no real patient information, institutional review board approval was deemed unnecessary and formally waived. Each contributor attested that all submitted reports were fabricated. This fictional nature creates unique advantages for the dataset, enabling unrestricted sharing, modification, and augmentation under the Creative Commons license CC-4.0-NC-SA. Unlike datasets containing real patient data, PARROT can be freely distributed across institutional and national boundaries, facilitating collaborative research without privacy constraints.

2.1. Submission requirements

Contributors had to submit at least 20 fictional reports with no upper limit. Each submission required the complete radiological report following the contributor's typical reporting style for their language and region. Contributors were granted the freedom to choose clinical scenarios and pathologies based on their area of expertise, with no constraint of specific diseases or templates. This approach was chosen to ensure the reports reflected their natural reporting style and clinical reality. Report metadata included anatomical region, imaging modality (CT, MRI, US, XR, etc.), brief clinical context/indication and report language. For non-English reports, contributors provided an English translation to facilitate cross-linguistic analysis. Contributors were given freedom to choose specialties and clinical scenarios according to their expertise. They were specifically instructed to create plausible but non-specific clinical scenarios, include typical incidental findings appropriate for the demographic, and incorporate normal anatomic variants at realistic frequencies. This approach allowed for capturing authentic reporting styles while ensuring the dataset represents a realistic distribution of normal and pathological findings.

All submitted reports were assigned ICD-10 codes either by the contributor, or through a human-in-the-loop workflow, where the lead author (BLG) reviewed, verified, and finalized preliminary codes proposed by the o3-mini-high language model (OpenAI). Discrepancies were resolved through communication with contributors or, when necessary, adjustment by the review team with documentation of changes.

2.2. Dataset organization and processing

The final dataset was organized in a standardized structure to facilitate research use (.jsonl file). The entire dataset was version-controlled using Git, with the initial public release tagged as v1.0. The repository includes documentation and usage examples to facilitate adoption. This version control approach enables ongoing collaboration and contribution, allowing the dataset to grow and improve over time while maintaining traceability of changes. Future contributors can submit additional reports following the established protocols, enhancing the linguistic and clinical diversity of the resource. The project is accessible at https://github.com/PARROT-reports.

2.3. Human vs. AI report differentiation study

To validate the authenticity and distinctive characteristics of PARROT reports compared to AI-generated content, we conducted a discrimination study. We generated AI reports using GPT-o1 (OpenAI), using the prompt “Create 5 radiology reports in [language], encompassing different diseases and modalities. They have to look absolutely real, like they could have been written by a [country] radiologist.” A comparable number of PARROT reports were selected for comparison. Participants recruited through social media (LinkedIn) and professional networks (e.g., European Society of Medical Imaging Informatics) were asked to identify whether reports were human-authored or AI-generated. The study was conducted on English, German, Italian, French, Greek and Polish reports via Google Docs, where participants selected their language of practice, reviewed and voted on 10 reports, with 5 being AI-generated and 5 human-authored from PARROT. Reports included MRI, CT, X-ray and echographic examinations from multiple anatomical regions. We specifically targeted a mix of radiologists, computer scientists, and general participants to assess whether professional expertise influenced the ability to distinguish authentic radiological reporting from AI-generated content.

This evaluation was motivated by concerns that researchers might

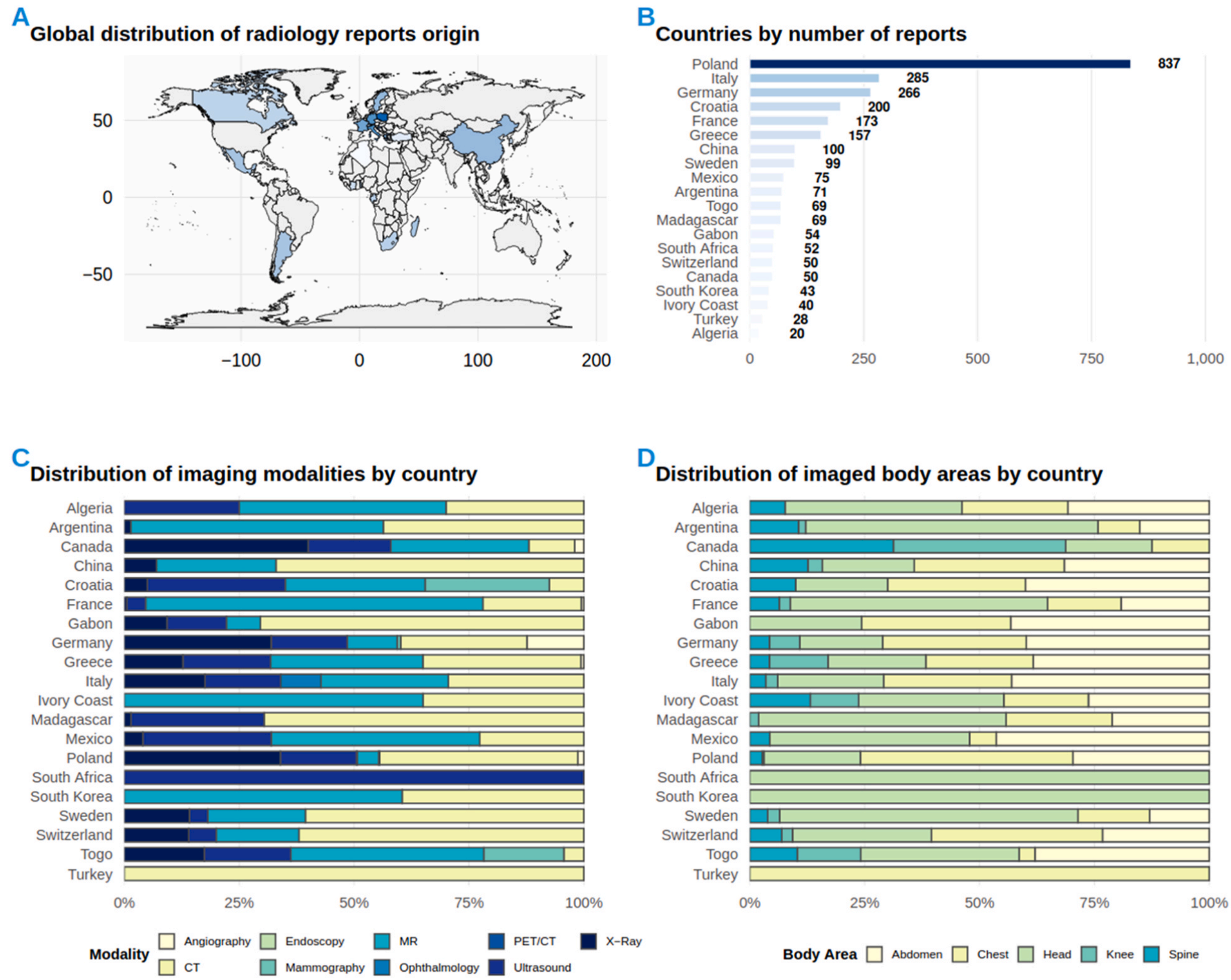

Fig. 1. Geographic distribution and characteristics of the PARROT dataset. A) World map showing the distribution of radiology reports by country of origin, with darker blue indicating higher numbers of reports. B) Bar chart displaying the count of reports by country, arranged in descending order. C) Distribution of imaging modalities by country, showing the percentage of each modality within countries. D) Distribution of five most imaged body areas by country, revealing regional preferences in anatomical focus.

use language models to generate synthetic radiology reports for testing purposes, potentially introducing subtle but important deviations from authentic clinical documentation. We hypothesized that those reports might be difficult to differentiate from human-generated content, but that domain experts such as radiologists would better recognize these differences than non-experts, highlighting the value of professionally authored content in the PARROT dataset.

2.4. Statistical analysis

A descriptive analysis was performed to characterize the dataset. Country of origin and frequency of imaging modalities were summarized, and body regions were categorized to assess the diversity of the contributed scenarios. Preliminary maps of geographic origins were generated to visualize the global participation in PARROT. All statistical analyses and visualizations were carried out using R version 4.3.1. Key analytical steps included summarizing submission counts per country and language, enumerating modality usage across the dataset, grouping body regions under standardized terms, analyzing the distribution of normal versus abnormal findings, and categorizing reports by ICD-10 classification. The analysis focused on demonstrating the breadth and diversity of the dataset across linguistic, geographic, and clinical dimensions to establish its utility as a representative multilingual resource for radiology AI research. Differences in performance for the differentiation study were assessed with a X^2 test, with a significance level of 0.05 [19].

3. Results

3.1. Reports origin and linguistic diversity

The PARROT dataset comprises 2658 radiology reports from 21 countries across four continents. Poland contributed the largest number of reports (837, 31.5 %), followed by Germany (316, 11.9 %), Italy (285, 10.7 %), Croatia (200, 7.5 %), and France (173, 6.5 %) as shown in Fig. 1**A**. The geographical distribution reveals significant representation from European countries, with emerging contributions from Asia, South America, and Africa (Fig. 1**A**). Countries from the Global South including Argentina (71, 2.7 %), China (100, 3.8 %), and Mexico (75, 2.8 %) collectively accounted for approximately 19 % of the dataset.

The PARROT dataset encompasses reports in 13 different languages. The distribution of languages reflects the geographical representation, with Polish, German, Italian, and French being the most prevalent. Notably, French-language reports originated from multiple countries including Algeria, France, Ivory Coast, Madagascar, Togo, and Canada, highlighting regional variations in reporting practices within the same language. Similarly, Spanish-language reports were contributed from both Argentina and Mexico, capturing different regional medical terminologies and reporting conventions.

3.2. Reports content and structure

Analysis of report content revealed diverse imaging modalities, with CT representing 36.1 % (n = 989) of the dataset, followed by MRI (22.8 %, n = 625), conventional radiography (XR, 19.0 %, n = 520), and ultrasound (US, 16.8 %, n = 461) as displayed in Fig. 1**C.**

The anatomical distribution analysis demonstrated a predominance of reports covering chest (677, 19.9 %), abdomen (631, 18.6 %), head (588, 17.3 %), and pelvis (480, 14.1 %) as detailed in Table 1 and visualized in Fig. 2. Less commonly represented regions included extremities and specialized areas such as orbit or pituitary gland. The relationship between modalities and anatomical regions is shown in Supplementary Figure S1.

The median report length varied across languages. The longest reports were in Turkish (median length 382 words, IQR: 154.75 words), followed by Spanish (196, IQR 136.5) and French (172, IQR: 130) with the shortest reports being in Afrikaans with a median word count of 36.5 (IQR: 12.75). Further details on word count are provided in Table 2.

3.3. Human vs. AI report differentiation study

To evaluate the authenticity of the human-authored reports in PARROT compared to AI-generated content, we conducted a differentiation study with 154 participants from diverse backgrounds. Overall, participants achieved a mean accuracy of 53.9 % (95 % CI: 50.7 %-57.1 %) in distinguishing between human-authored and AI-generated reports, indicating performance only slightly above chance level (Fig. 2**A**).

Analysis by occupation revealed radiologists performed better (56.9 %, 95 % CI: 53.3 %-60.6 %) than non-healthcare professionals (49.7 %, 95 % CI: 41.4 %-58.0 %) and other healthcare professionals (48.3 %, 95 % CI: 40.1 %-56.5 %), as shown in Fig. 2**B-C**. This difference was statistically significant ($p < 0.05$), suggesting domain expertise provides some advantage in discerning report authenticity.

Self-reported confidence showed a weak association with actual performance. Participants reporting the highest confidence levels ("Very confident" and "Completely confident") achieved modestly higher accuracy (61.2 % and 60.0 % respectively) compared to those with lower confidence ratings (Fig. 2**D**). However, the correlation between confidence and accuracy was relatively weak ($r = 0.07$), indicating poor metacognition in this task.

Analysis of response bias revealed a slight tendency across all professional groups to classify reports as human-authored rather than AI-generated. Radiologists classified 43.8 % of reports as AI-generated, while non-healthcare professionals and other healthcare professionals classified 42.3 % and 41.8 % respectively (Fig. 2**B**). Logistic regression analysis confirmed that professional background significantly influenced accuracy, with non-healthcare professionals (estimate: −0.328, $p = 0.018$) and other healthcare professionals (estimate: −0.360, $p = 0.008$) performing worse than radiologists (Fig. 3).

Table 1
Distribution of body areas.

Body Area	Count	Percentage
Chest	677	19.9
Abdomen	631	18.6
Head	588	17.3
Pelvis	480	14.1
Breast	176	5.2
Neck	138	4.1
Lumbar Spine	102	3.0
Spine	95	2.8
Orbit	79	2.3
Cervical	71	2.1
Knee	60	1.8
Lower Extremity	50	1.5
Shoulder	49	1.4
Thoracic Spine	46	1.4
Upper Extremity	30	0.9
Foot	22	0.6
Wrist	21	0.6
Ankle	20	0.6
Face	15	0.4
Pituitary Gland	12	0.4
Elbow	10	0.3
Testis	8	0.2
Whole Body	8	0.2
Ear	4	0.1
Hip	3	0.1
Nose	2	0.1

3.4. ICD-10 code distribution

The dataset encompasses pathologies across all major ICD-10 chapters, with notable predominance of codes from Chapter IX (Diseases of the circulatory system), Chapter X (Diseases of the respiratory system), Chapter XI (Diseases of the digestive system), and XIII (Diseases of the musculoskeletal system and connective tissue) collectively accounting for 53.9 % of all coded reports. Among specific three-digit codes, I70 (Atherosclerosis) and I63 (Infarction) represented the most frequent pathology category with 133 and 109 reports respectively. Table 3 provides an overview of the distribution of codes across chapters. Fig. 4 is a treemap representation, demonstrating the hierarchical distribution of codes grouped by chapter and three-digit classifications. A detailed overview of ICD-10 chapters, three-digit codes as well as full ICD-10 codes is provided in Table S1.

4. Discussion

In this work, we present PARROT, the largest openly available multilingual radiology report dataset to date. Comprising 2658 fictional yet realistic reports contributed by 76 radiologists from 21 countries covering 13 languages, PARROT spans multiple imaging modalities (X-ray, ultrasound, CT, MRI) and nearly all major body regions. The collection's synthetic nature enables open sharing without privacy concerns while preserving authentic clinical language and style. By capturing diverse reporting practices across languages, modalities, and anatomical regions, PARROT serves as a resource for evaluating NLP applications in a multilingual context.

PARROT facilitates the benchmarking of various NLP tasks in a multilingual context, including Named Entity Recognition (NER) [9], automatic structuring [8], abbreviation disambiguation, error detection [11], translation, and automated ICD-10 coding. Currently, the validation of LLMs for tasks such as error detection is limited in non-English settings due to a scarcity of appropriate datasets [11,20]. For instance, recent work by Salam and colleagues required the ad hoc creation of fictional reports with post-hoc error injection to evaluate this specific task in German. PARROT addresses this bottleneck by providing a ready-to-use multilingual resource, thereby streamlining the testing of NLP applications and eliminating the need for researchers to replicate the labor-intensive report generation process for every new study [8,20].

Several open-access datasets for natural language processing have been published in recent years, yet most remain limited in linguistic and geographic scope. The MIMIC-IV clinical database and its derivative MIMIC-CXR corpus represent among the largest collections of radiology reports, but they exclusively reflect English-language, U.S.-based practice [12,13]. This monolinguistic bias affects research development, with a systematic review finding that among 164 radiology NLP studies, 142 (86.6 %) focused exclusively on English reports, with minimal representation of other languages [15].

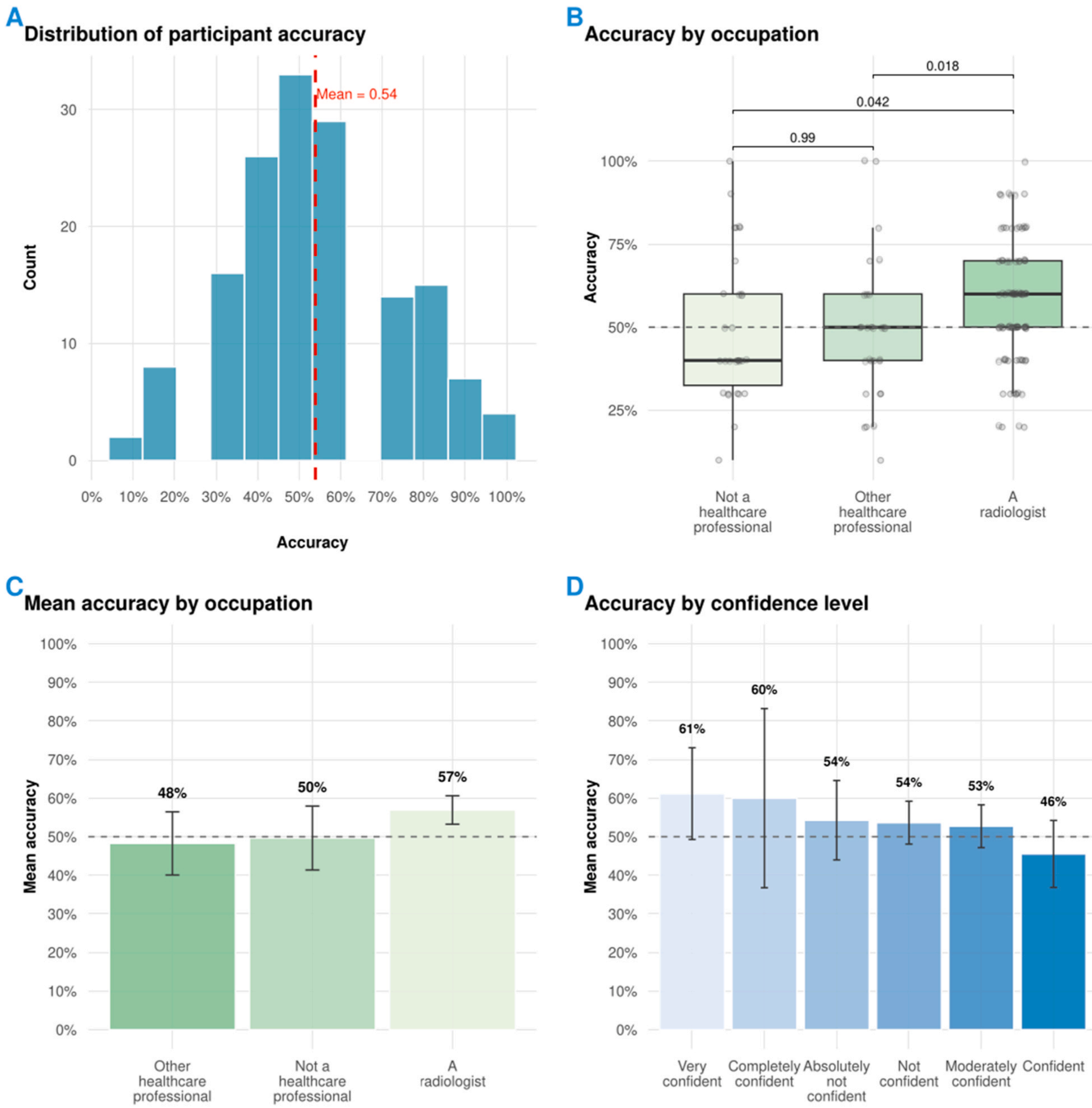

Fig. 2. Results of the human vs. AI report differentiation study A) Distribution of participant accuracy in distinguishing between human-authored and AI-generated radiology reports, with dashed line representing chance level (50 %). B) Box plots showing accuracy by occupation group, with radiologists performing significantly better than other groups. C) Effect sizes from logistic regression model showing factors associated with accuracy in the differentiation task. D) Accuracy by self-reported confidence level, revealing modest correlation between confidence and performance.

Beyond language, openness and geographic diversity are relevant considerations for dataset development. Many publicly available radiology text datasets originate from a single country or healthcare system. The aforementioned collections, for instance, derive from U.S. hospital data [13]. The privacy regulations governing clinical text data, including HIPAA in the United States and GDPR in Europe, necessitate extensive de-identification protocols that often slow down or limit international data sharing. These constraints may lead to AI models being overfitted to the phrasing, style, and clinical patterns of a limited context. Language and regional biases can potentially limit algorithm utility when confronted with different populations or reporting styles [21]. PARROT aims to address these challenges through its foundation on fictional cases created by domain experts, offering an approach to data accessibility issues.

Yet, the ongoing development of increasingly potent LLMs, such as Generative Pretrained Transformer (GPT) presents an alternative approach to generating synthetic radiology reports. By prompting frontier models, like GPT-o1 [22], researchers could potentially create artificial datasets tailored to specific use cases. However, this approach may introduce subtle errors that can be difficult to detect, especially for those without domain expertise. Shumailov and colleagues recently showed that AI-generated synthetic data contains subtle differences with genuine human-generated content that induces to models to collapse after recursive training [23]. Our differentiation study examined this potential issue on radiology reports, assessing the existence of potential domain-specific differences between synthetic and human-generated radiology reports. Participants achieved 53.9 % accuracy overall in distinguishing between human-authored PARROT reports and AI-generated alternatives. Radiologists performed statistically significantly better (56.9 %, $p < 0.05$) than other healthcare professionals

Table 2
Word count of reports across languages.

Language	Median length (Interquartile Range)
Afrikaans	36.5 (IQR: 12.75)
Chinese	182.5 (IQR: 137)
Croatian	67 (IQR: 41)
Dutch	39.5 (IQR: 15.25)
French	172 (IQR: 130)
German	71.5 (IQR: 79.25)
Greek	86 (IQR: 72)
Italian	94 (IQR: 87)
Korean	43 (IQR: 28)
Polish	112 (IQR: 145)
Spanish	196 (IQR: 136.5)
Swedish	73 (IQR: 73)
Turkish	328 (IQR: 154.75)

(48.3 %) or non-healthcare participants (49.7 %), meaning domain expertise is necessary to detect subtle stylistic changes in AI-generated reports. This observation has methodological implications, as computer scientists, who likely conduct most research on natural language processing, might not identify domain-specific errors in AI-generated reports [24]. PARROT provides reports crafted by experts who incorporate clinical reasoning patterns that current language models may not fully replicate. Because radiologists performed only marginally better than chance in differentiating AI- and human-generated reports, LLMs seem already be able to mimic the reporting style of radiologists sufficiently well to pass for radiologists style wise. However, whether the LLMs are able to integrate medical authenticity and reasoning is not known. A recent study by Hager et al. [25] showed that LLMs had limitations in medical reasoning. As such, the value of the human authorship of PARROT may reside in the medical logic contained in the report, in addition to their phrasing and reporting style.

Still, PARROT has several limitations that warrant consideration. A primary limitation is the current imbalance in geographic representation, with a predominance of reports originating from Europe, including 31.5 % of reports from Poland, and comparatively sparse representation from Africa, South America, and extensive regions of Asia. Approximately 19 % of reports originate from the Global South. This geographic distribution affects the dataset's coverage of global diversity in languages, dialects, and regional reporting practices. Specifically, the high prevalence of central European reports generates imbalance and may bias the aggregate evaluation metrics. As such, we advise for stratified evaluation when using PARROT, including evaluation metrics per language. Additionally, while advantageous for addressing privacy barriers, the fictional nature of the reports means they lack connection to real patient imaging data and clinical outcomes, preventing its use in applications requiring image-text correlation or multimodal model training. We acknowledge that the fictional nature of PARROT reports may also distance them from the reality of clinical reports, particularly regarding noise and dictation errors. Consequently, PARROT reports could rather serve as a clean baseline rather than an exact representation of the reports written in a clinical context. Recruitment strategies relying on social media platforms and professional networks may also introduce selection bias, potentially favoring contributions from individuals active in academic circles, which could influence the range of clinical scenarios, terminology, and reporting styles represented within the dataset.

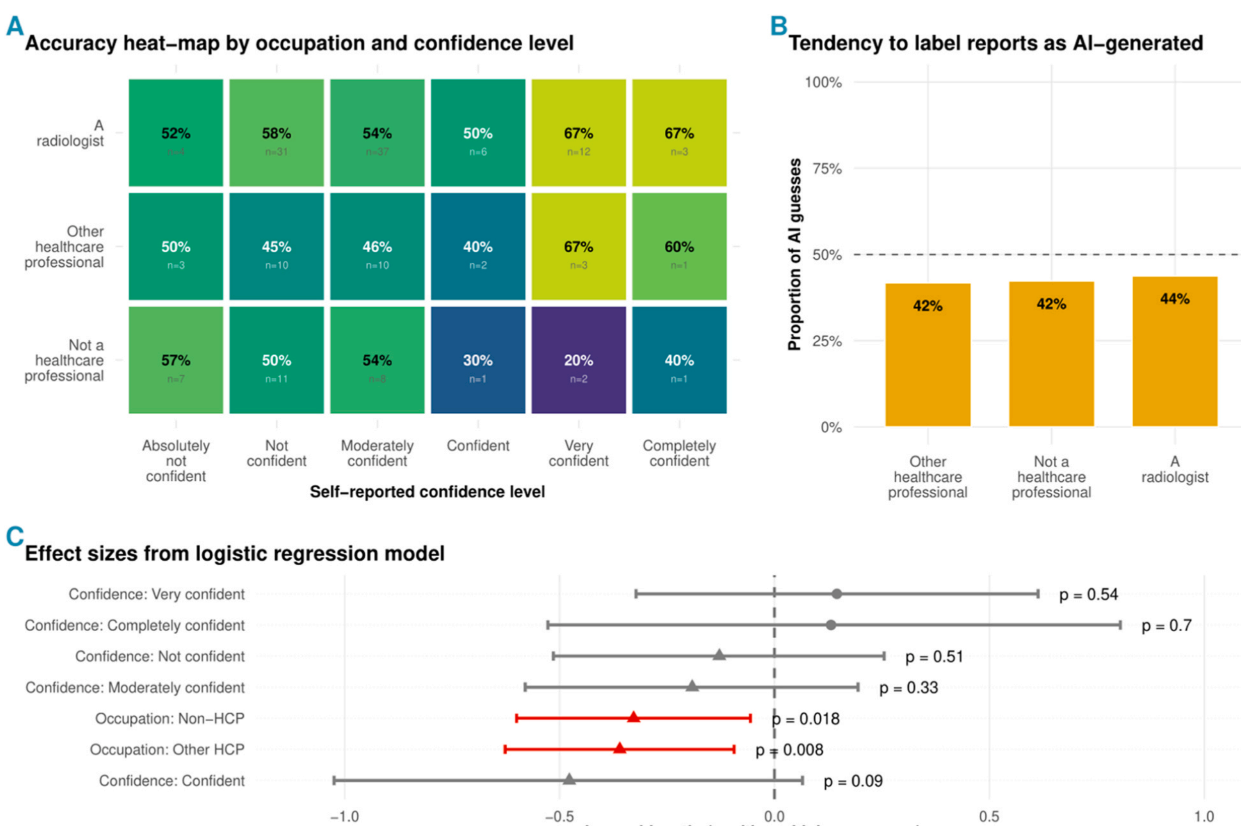

Fig. 3. Analysis of human vs. AI report differentiation performance A) Accuracy heat-map by occupation and confidence level, showing percentage accuracy for each combination with sample size indicated. Radiologists generally achieved higher accuracy across confidence levels compared to other groups. B) Tendency to label reports as AI-generated by occupation group, showing similar bias patterns across all groups with approximately 42–44 % of reports identified as AI-generated, indicating a slight bias towards labeling a report as human generated. C) Effect sizes from logistic regression model analyzing factors associated with discrimination accuracy. Non-healthcare professionals (Non-HCP) and other healthcare professionals (Other HCP) showed significantly lower accuracy compared to radiologists ($p = 0.018$ and $p = 0.008$ respectively), while confidence levels had varying but generally non-significant effects on performance.

Table 3
Overview of ICD-10 codes.

Chapter	% of Codes	3 Digit ICD-10 Codes
I	0.4	A15, A18, A40, A41, B00, B16, B25, B37, B44, B58, B69, B96
II	7.7	C16, C18, C19, C20, C22, C23, C24, C25, C32, C34, C37, C41, C45, C50, C53, C54, C61, C64, C67, C69, C70, C71, C73, C74, C77, C78, C79, C81, C85, C90, C91, C92, D05, D10, D11, D13, D16, D17, D18, D24, D25, D28, D30, D31, D32, D33, D34, D35, D37, D41, D43, D45, D48
III	0.4	D50, D57, D62, D64, D73, D76, D82, D86
IV	1.5	E03, E04, E06, E07, E11, E14, E22, E23, E27, E34, E46, E66, E78, E84, E85, E87
IX	17.9	I07, I10, I11, I12, I15, I21, I25, I26, I27, I28, I31, I33, I34, I35, I36, I37, I40, I42, I46, I48, I49, I50, I51, I60, I61, I62, I63, I64, I65, I66, I67, I69, I70, I71, I72, I73, I74, I77, I78, I80, I81, I82, I83, I87, I89, I95
V	0.1	F01, F03, F10, F33, F41, F79
VI	1.5	G00, G06, G20, G23, G25, G30, G31, G35, G36, G37, G40, G41, G43, G45, G50, G51, G54, G56, G57, G70, G81, G83, G90, G91, G93, G96, G97
VII	1.3	H02, H04, H11, H20, H21, H27, H33, H35, H43, H44, H47, H53, H54
VIII	0.2	H61, H65, H66, H70, H71, H83, H90, H92, H93
X	13.8	J01, J11, J12, J15, J18, J20, J21, J31, J32, J33, J34, J35, J36, J38, J39, J40, J42, J43, J44, J45, J47, J68, J69, J80, J81, J84, J86, J90, J92, J93, J94, J96, J98
XI	8.9	K02, K04, K08, K09, K11, K21, K22, K25, K26, K31, K35, K38, K40, K41, K42, K43, K44, K50, K51, K52, K55, K56, K57, K58, K59, K60, K63, K65, K66, K70, K74, K75, K76, K80, K81, K82, K83, K85, K86, K91, K92
XII	0.0	L59, L72
XIII	9.8	M06, M07, M11, M13, M15, M16, M17, M18, M19, M21, M22, M23, M24, M25, M34, M35, M40, M41, M43, M45, M46, M47, M48, M50, M51, M53, M54, M61, M62, M65, M67, M70, M71, M75, M76, M77, M79, M80, M81, M84, M85, M87, M89, M93, M94, M96, M99
XIV	7.4	N10, N11, N13, N15, N17, N18, N19, N20, N21, N28, N30, N32, N39, N40, N41, N42, N43, N44, N50, N60, N61, N63, N64, N80, N82, N83, N84, N85, N89, N91, N92, N93, N94, N97
XIX	5.2	S00, S02, S03, S05, S06, S12, S13, S20, S22, S26, S27, S32, S33, S36, S42, S43, S52, S53, S62, S70, S72, S73, S82, S83, S90, S92, S93, T17, T18, T44, T78, T80, T81, T82, T83, T88, T91, T92
XV	0.1	O00, O07, O43, O45
XVI	0.2	P07, P21, P22, P24, P29, P35, P36, P78
XVII	2.2	Q00, Q01, Q02, Q04, Q07, Q17, Q18, Q21, Q24, Q25, Q26, Q27, Q28, Q39, Q40, Q43, Q44, Q53, Q61, Q62, Q63, Q65, Q67, Q74, Q75, Q76, Q78, Q79, Q82, Q85, Q86, Q89, Q90
XVIII	11.7	R00, R04, R05, R06, R07, R10, R11, R13, R16, R17, R18, R19, R20, R21, R22, R23, R25, R26, R27, R29, R30, R31, R33, R35, R40, R41, R42, R45, R47, R50, R51, R52, R55, R56, R58, R59, R60, R63, R64, R70, R74, R90, R93, R94
XX	0.0	V99
XXI	9.5	Z00, Z01, Z03, Z04, Z08, Z09, Z31, Z34, Z42, Z45, Z47, Z72, Z80, Z85, Z90, Z92, Z93, Z94, Z95, Z96, Z98, Z99
XXII	0.1	U07
NA	0.0	U09

This recruitment bias may also explain the lack of nuclear medicine reports, which may be a priority for the future additions. Another limitation is that reports were not peer-reviewed in their original language and only their translation was assessed during the ICD-10 coding review. Their authenticity is thus ensured by the credentials of their writers, who are all board-certified physicians. Finally, the only medical ontology included in the current version of PARROT is ICD-10. Future studies could assess how domain-specific ontologies such as RadLex could be included in PARROT.

In conclusion, PARROT provides an open-access, multilingual collection of radiology reports addressing some of the linguistic and accessibility constraints in the evaluation of NLP tools for non-English contexts. By offering fictional yet radiologist-authored reports across multiple languages, PARROT serves as a research resource for testing AI applications that can function across different healthcare systems and languages.

CRediT authorship contribution statement

Philipp Prucker: Writing – review & editing, Investigation, Data curation. **Michał Fułek:** Writing – review & editing, Investigation, Data curation. **Katarzyna Fułek:** Writing – review & editing, Investigation, Data curation. **Federica Fossataro:** Writing – review & editing, Investigation, Data curation. **Claudia Fossataro:** Writing – review & editing, Investigation, Data curation. **Carlos Ferrarotti:** Writing – review & editing, Investigation, Data curation. **Salvatore Claudio Fanni:** Writing – review & editing, Investigation, Data curation. **Armel Elogne:** Writing – review & editing, Investigation, Data curation. **Barbara Dziadkowiec-Macek:** Writing – review & editing, Investigation, Data curation. **Andrea Ponsiglione:** Writing – review & editing, Investigation, Data curation. **Federica Piccione:** Writing – review & editing, Investigation, Data curation. **Yae Won Park:** Writing – review & editing, Investigation, Data curation. **Adriana Ojeda:** Writing – review & editing, Investigation, Data curation. **Béatrice Nguema Edzang:** Writing – review & editing, Investigation, Data curation. **Antonio Natale:** Writing – review & editing, Investigation, Data curation. **Eugen Divjak:** Writing – review & editing, Investigation, Data curation. **Lukas Müller:** Writing – review & editing, Investigation, Data curation. **Renato Cuocolo:** Writing – review & editing, Investigation, Data curation. **Aymen Meddeb:** Writing – review & editing, Investigation, Data curation. **Erwin Chiquete:** Writing – review & editing, Investigation, Data curation. **Rafał Poręba:** Writing – review & editing, Investigation, Data curation. **Małgorzata Poręba:** Writing – review & editing, Investigation, Data curation. **Liz Toapanta-Yanchapaxi:** Writing – review & editing, Investigation, Data curation. **Ali Fuat Tekin:** Writing – review & editing, Investigation, Data curation. **Ezann Siebert:** Writing – review & editing, Investigation, Data curation. **Luca Saba:** Writing – review & editing, Investigation, Data curation. **Jan Rotkegel:** Writing – review & editing, Investigation, Data curation. **Katarzyna Resler:** Writing – review & editing, Investigation, Data curation. **Vasileios Rafailidis:** Writing – review & editing, Investigation, Data curation. **Feno Hasina Rabemanorintsoa:** Writing – review & editing, Investigation, Data curation. **Rosa alba Pugliesi:** Writing – review & editing, Investigation, Data curation. **Martyna Gachowska:** Writing – review & editing, Investigation, Data curation. **Jean-Pierre Pruvo:** Writing – review & editing, Investigation, Data curation. **Paweł Gać:** Writing – review & editing, Investigation, Data curation. **Arnaldo Stanzione:** Writing – review & editing, Investigation, Data curation. **Lina Xu:** Writing – review & editing, Investigation, Data curation. **Radhia Ait Chalal:** Writing – review & editing, Investigation, Data curation. **Andrzej Wysocki:** Writing – review & editing, Investigation, Data curation. **Sung Soo Ahn:** Writing – review & editing, Investigation, Data curation. **Szymon Włodarczak:** Writing – review & editing, Investigation, Data curation. **Agripnidis Thibaut:** Writing – review & editing, Investigation, Data curation. **Adams Lisa C:** Writing – review & editing, Investigation, Data curation. **Aambounou Kokou:** Writing – review & editing, Investigation, Data curation. **LE GuelleC Bastien:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Adrian Włodarczak:** Writing – review & editing, Investigation, Data curation. **Nicolas Herinirina:** Writing – review & editing, Investigation, Conceptualization. **Johan Wassélius:** Writing – review & editing, Investigation, Data curation. **Aghiles Hamroun:** Writing – review & editing, Investigation, Data curation. **Weilang Wang:** Writing – review & editing, Investigation, Data curation. **Alexia Maria Goulianou:** Writing – review & editing, Investigation, Data curation. **Federica Vernuccio:** Writing – review & editing, Investigation, Data curation. **Natalia Gorelik:** Writing – review & editing,

Fig. 4. Treemap of ICD-10 Codes in the PARROT dataset, grouped by the chapter number and 3 digit ICD-10 code. Refer to Table S1 for a detailed overview of each code and its subcodes.

Investigation, Conceptualization. **Evangelia Vassalou:** Writing – review & editing, Investigation, Data curation. **Marco Gatti:** Writing – review & editing, Investigation, Data curation. **Szymon Urban:** Writing – review & editing, Investigation, Data curation. **Ignacio García-Juárez:** Writing – review & editing, Investigation, Data curation. **Ekaterini Tsaoulia:** Writing – review & editing, Investigation, Data curation. **Matthaios Triantafyllou:** Writing – review & editing, Investigation, Data curation. **Klontzas Michail E:** Writing – review & editing, Investigation, Data curation. **Felipe Kitamura:** Writing – review & editing, Investigation, Data curation. **Gordana Ivanac:** Writing – review & editing, Investigation, Data curation. **Quentin Holay:** Writing – review & editing, Investigation, Data curation. **Christelle Chavihot:** Writing – review & editing, Investigation, Data curation. **Armando Ugo Cavallo:** Writing – review & editing, Investigation, Data curation. **Riccardo Cau:** Writing – review & editing, Investigation, Data curation. **Elif Can:** Writing – review & editing, Investigation, Data curation. **Felix Busch:** Writing – review & editing, Investigation, Data curation. **Antonino Andrea Blandino:** Writing – review & editing, Investigation, Data curation. **Claudio Benzoni:** Writing – review & editing, Investigation, Data curation. **Maximilian Lindholz:** Writing – review & editing, Investigation, Data curation. **Tristan Lemke:** Writing – review & editing, Investigation, Data curation. **Alexandre Lefèvre:** Writing – review & editing, Investigation, Data curation. **Bressem Keno K:** Writing – review & editing, Writing – original draft, Visualization, Validation, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Dominik Krupka:** Writing – review & editing, Investigation, Data curation. **Grégory Kuchcinski:** Writing – review & editing, Supervision, Investigation, Data curation, Conceptualization. **Krzysztof Kraik:** Writing – review & editing, Investigation, Data curation. **Raphaël Bentegeac:** Writing – review & editing, Investigation, Data curation. **Sebastian Ziegelmayer:** Writing – review & editing, Investigation, Data curation. **Rafał Kompanowski:** Writing – review & editing, Investigation, Data curation. **Henrik Andersson:** Writing – review & editing, Investigation, Data curation. **Shuhang Zhang:** Writing – review & editing, Investigation, Data curation. **Anna Kompanowska:** Writing – review & editing, Investigation, Data curation. **Philippe Amouyel:** Writing – review & editing, Investigation, Data curation. **Tomasz Zatoński:** Writing – review & editing, Investigation, Data curation. **Tugba Akinci D′Antonoli:** Writing – review & editing, Investigation, Data curation. **Luigi Mannacio:** Writing – review & editing, Investigation, Data curation. **Marcus Makowski:** Writing – review & editing, Investigation, Data curation. **Piotr Macek:** Writing – review & editing, Investigation, Data curation.

Ethics in publishing statement

I testify on behalf of all co-authors that our article submitted followed ethical principles in publishing.

All authors agree that:

This research presents an accurate account of the work performed, all data presented are accurate and methodologies detailed enough to permit others to replicate the work.

This manuscript represents entirely original works and or if work and/or words of others have been used, that this has been appropriately cited or quoted and permission has been obtained where necessary.

This material has not been published in whole or in part elsewhere.

The manuscript is not currently being considered for publication in another journal.

That generative AI and AI-assisted technologies have not been utilized in the writing process or if used, disclosed in the manuscript the use of AI and AI-assisted technologies and a statement will appear in the published work.

That generative AI and AI-assisted technologies have not been used to create or alter images unless specifically used as part of the research design where such use must be described in a reproducible manner in the methods section.

All authors have been personally and actively involved in substantive work leading to the manuscript and will hold themselves jointly and individually responsible for its content.

Funding

no specific funding

Declaration of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

Appendix A. Supporting information

Supplementary data associated with this article can be found in the online version at doi:10.1016/j.ejrai.2025.100066.

Data availability

https://github.com/PARROT-reports/PARROT_v1.0/

References

[1] M.P. Hartung, I.C. Bickle, F. Gaillard, J.P. Kanne, How to create a great radiology report, RadioGraphics 40 (2020) 1658–1670, https://doi.org/10.1148/rg.2020200020.

[2] C.M. Hawkins, S. Hall, B. Zhang, A.J. Towbin, Creation and implementation of department-wide structured reports: an analysis of the impact on error rate in radiology reports, J. Digit Imaging 27 (2014) 581–587, https://doi.org/10.1007/s10278-014-9699-7.

[3] J.M. Nobel, K. Van Geel, S.G.F. Robben, Structured reporting in radiology: a systematic review to explore its potential, Eur. Radio. 32 (2022) 2837–2854, https://doi.org/10.1007/s00330-021-08327-5.

[4] L.H. Schwartz, D.M. Panicek, A.R. Berk, Y. Li, H. Hricak, Improving communication of diagnostic radiology findings through structured reporting, Radiology 260 (2011) 174–181, https://doi.org/10.1148/radiol.11101913.

[5] P. Keshavarz, S. Bagherieh, S.A. Nabipoorashrafi, H. Chalian, A.A. Rahsepar, G.H. J. Kim, et al., ChatGPT in radiology: a systematic review of performance, pitfalls, and future perspectives, Diagn. Interv. Imaging 105 (2024) 251–265, https://doi.org/10.1016/j.diii.2024.04.003.

[6] A. Lecler, L. Duron, P. Soyer, Revolutionizing radiology with GPT-based models: current applications, future possibilities and limitations of ChatGPT, Diagn. Inter. Imaging 104 (2023) 269–274, https://doi.org/10.1016/j.diii.2023.02.003.

[7] B. Le Guellec, C. Bruge, N. Chalhoub, V. Chaton, E. De Sousa, Y. Gaillandre, et al., Comparison between multimodal foundation models and radiologists for the diagnosis of challenging neuroradiology cases with text and images, Diagn. Interv. Imaging (2025), https://doi.org/10.1016/j.diii.2025.04.006.

[8] L.C. Adams, D. Truhn, F. Busch, A. Kader, S.M. Niehues, M.R. Makowski, et al., Leveraging GPT-4 for Post Hoc transformation of free-text radiology reports into structured reporting: a multilingual feasibility study, Radiology 230725 (2023), https://doi.org/10.1148/radiol.230725.

[9] B. Le Guellec, A. Lefèvre, C. Geay, L. Shorten, C. Bruge, L. Hacein-Bey, et al., Performance of an open-source large language model in extracting information from free-text radiology reports, Radiol. Artif. Intell. (2024) e230364, https://doi.org/10.1148/ryai.230364.

[10] K.S. Amin, M.A. Davis, R. Doshi, A.H. Haims, P. Khosla, H.P. Forman, Accuracy of ChatGPT, google bard, and microsoft bing for simplifying radiology reports, Radiology 309 (2023) e232561, https://doi.org/10.1148/radiol.232561.

[11] R.J. Gertz, T. Dratsch, A.C. Bunck, S. Lennartz, A.-I. Iuga, M.G. Hellmich, et al., Potential of GPT-4 for detecting errors in radiology reports: implications for reporting accuracy, Radiology 311 (2024) e232714, https://doi.org/10.1148/radiol.232714.

[12] A.E.W. Johnson, T.J. Pollard, S.J. Berkowitz, N.R. Greenbaum, M.P. Lungren, C. Deng, et al., MIMIC-CXR, a de-identified publicly available database of chest radiographs with free-text reports, Sci. Data 6 (2019) 317, https://doi.org/10.1038/s41597-019-0322-0.

[13] A.E.W. Johnson, L. Bulgarelli, L. Shen, A. Gayles, A. Shammout, S. Horng, et al., MIMIC-IV, a freely accessible electronic health record dataset, Sci. Data 10 (1) (2023), https://doi.org/10.1038/s41597-022-01899-x.

[14] J. Wu, X. Liu, M. Li, W. Li, Z. Su, S. Lin, et al., Clinical text datasets for medical artificial intelligence and large language models — a systematic review, NEJM AI 1 (2024), https://doi.org/10.1056/AIra2400012.

[15] O. Banerjee, A. Saenz, K. Wu, W. Clements, A. Zia, D. Buensalido, et al., ReXamine-Global: A Framework for Uncovering Inconsistencies in Radiology Report Generation Metrics. Biocomputing 2025, Kohala Coast, Hawaii, USA: World scientific, 2024, pp. 185–198, https://doi.org/10.1142/9789819807024_0014.

[16] M. Al-Duwais, H. Al-Khalifa, A. Al-Salman, A benchmark evaluation of multilingual large language models for arabic cross-lingual named-entity recognition, Electronics 13 (2024) 3574, https://doi.org/10.3390/electronics13173574.

[17] L. Campos, V. Pedro, F. Couto, Impact of translation on named-entity recognition in radiology texts, Database 2017 (2017), https://doi.org/10.1093/database/bax064.

[18] F. Busch, L. Hoffmann, L. Xu, L. Zhang, B. Hu, I. García-Juárez, et al., Multinatl. Attitudes Towards AI Healthc. Diagn. Hosp. Patients (2024), https://doi.org/10.1101/2024.09.01.24312016.

[19] E. O'Shaughnessy, E. Detrinidad, P. Soyer, A. Lecler, An introductory guide to statistics for the radiologist, Diagn. Interv. Imaging 106 (2025) 49–52, https://doi.org/10.1016/j.diii.2024.11.003.

[20] B. Salam, C. Stüwe, S. Nowak, A.M. Sprinkart, M. Theis, D. Kravchenko, et al., Large language models for error detection in radiology reports: a comparative analysis between closed-source and privacy-compliant open-source models, Eur. Radio. 35 (2025) 4549–4557, https://doi.org/10.1007/s00330-025-11438-y.

[21] T. Suenghataiphorn, N. Tribuddharat, P. Danpanichkul, N. Kulthamrongsri, Bias Large Lang. Models Across Clin. Appl. A Syst. Rev. (2025), https://doi.org/10.48550/arXiv.2504.02917.

[22] OpenAI o1 System Card n.d. ⟨https://openai.com/index/openai-o1-system-card/⟩ (accessed December 11, 2024).

[23] I. Shumailov, Z. Shumaylov, Y. Zhao, N. Papernot, R. Anderson, Y. Gal, AI models collapse when trained on recursively generated data, Nature 631 (2024) 755–759, https://doi.org/10.1038/s41586-024-07566-y.

[24] P. Hager, F. Jungmann, R. Holland, K. Bhagat, I. Hubrecht, M. Knauer, et al., Evaluation and mitigation of the limitations of large language models in clinical decision-making, Nat. Med. 30 (2024) 2613–2622, https://doi.org/10.1038/s41591-024-03097-1.

[25] J. Wang, H. Deng, B. Liu, A. Hu, J. Liang, L. Fan, et al., Systematic evaluation of research progress on natural language processing in medicine over the past 20 years: bibliometric study on pubMed, J. Med. Internet Res. 22 (2020) e16816, https://doi.org/10.2196/16816.