How Far Can a Lightweight Multilingual Model Go on Clinical Summarization?

DACHausa at CLEF 2026

Ibrahim Rabiu Abdullahi[1,*], Idris Abdulmumin[2]

[1]Division of Agricultural Colleges, Ahmadu Bello University, Nigeria
[2]University of Pretoria, South Africa

Abstract

This paper explores the strengths and limitations of a compact multilingual model for clinical case summarization. The system, submitted to the MultiClinSum-2 shared task (Task 3 of the BioASQ workshop at CLEF 2026), fully fine-tunes mT5-small (300M parameters) to summarize medical case reports across 15 languages, trained on 334,024 case-summary pairs, without the cost of billion-parameter models. Across 14 evaluated languages, the system achieves mean ROUGE-1 of 0.325 (range: 0.272-0.372), ROUGE-2 of 0.125 (0.070-0.160), ROUGE-L of 0.237 (0.191-0.312), and BERTScore F1 of 0.834 (0.822-0.847). The organizers' official clinical-quality evaluation yields mean faithfulness of 0.632, completeness of 0.520, fluency of 0.688, and consistency of 0.336. Romance languages consistently outperform Germanic and Slavic languages on lexical metrics, while BERTScore remains stable across all language families (= 0.007) (=0.007), showing that even a lightweight model sustains consistent semantic fidelity across typologically diverse languages. Russian is identified as the most challenging language, with competitive ROUGE-L (0.312) but the lowest clinical consistency score (0.268). The correlation between faithfulness and completeness (r = 0.70 , p < 0.01) (r=0.70,p<0.01) suggests these dimensions capture overlapping quality aspects. All hyperparameters and reproducibility measures are documented.

Keywords

clinical summarization, multilingual NLP, mT5, lightweight models, low-resource languages

1. Introduction and Motivation

Automatic summarisation of clinical text has become increasingly important as the volume of medical literature and patient records continues to grow exponentially [1]. The volume of clinical records and biomedical literature has grown rapidly, driven by contributions from clinicians and biomedical researchers worldwide [2]. Clinicians face overwhelming information loads, spending excessive time reviewing documents rather than attending to patients [3]. Biomedical information from diverse documents provides clinicians and researchers with essential knowledge to track advances, formulate and test hypotheses, conduct experiments, and interpret results [4]. However, this abundance of information creates significant documentation burdens. Electronic health records, while improving information access, increase documentation workload and contribute to clinician burnout [5]. Clinicians face a critical summarisation problem: they spend excessive time documenting patient information—including diagnostic reports, progress notes, and treatment histories from multiple specialists [6]. This documentation burden increases the risk of errors and communication failures, ultimately compromising patient care and contributing to clinician burnout [7]. Effective automatic summarisation could alleviate this burden by distilling key clinical information from lengthy case reports [8].

Unlike traditional summarisation benchmarks, clinical summarisation requires not only lexical fidelity but also factual accuracy, completeness, and medical coherence—dimensions that automatic metrics alone cannot fully capture [9]. Therefore, a comprehensive evaluation framework must combine lexical, semantic, and qualitative assessments.

CLEF 2026 Working Notes, 21 – 24 September 2026, Jena, Germany
*Corresponding author.
abunimra0@gmail.com (I. R. Abdullahi); idris.abdulmumin@up.ac.za (I. Abdulmumin)
https://abumafrim.github.io (I. Abdulmumin)
0000-0002-3795-8381 (I. Abdulmumin)

The MultiClinSum-2 shared task (Task 3 of the BioASQ workshop at CLEF 2026) addresses these challenges by providing a multilingual corpus of case-summary pairs across 15 languages, tasking participants with generating concise, clinically relevant summaries of medical case reports [10].

In this paper, the evaluation methodology for the submitted system to MultiClinSum-2 is presented. The contributions are:

1. Evidence that a lightweight 300M-parameter model achieves consistent semantic fidelity across 15 languages, with the strongest lexical agreement in higher-resourced languages;
2. A full fine-tuning pipeline for mT5-small on 15-language clinical summarization;
3. A file-indexed streaming dataset that loads text on the fly, keeping host-memory usage constant regardless of corpus size;
4. A cross-lingual analysis of the official shared-task metrics (ROUGE, BERTScore, and clinical-quality scores) across 14 languages and language families;
5. Comprehensive documentation of hyperparameters and reproducibility measures;
6. Language-tagged input conditioning for improved cross-lingual transfer.

2. Related Work

This section reviews prior work in automatic summarisation evaluation, covering traditional metrics (ROUGE), semantic metrics (BERTScore), LLM-as-a-judge approaches, and clinical summarisation evaluation.

2.1. Automatic Summarisation Metrics

ROUGE [11] ignores readability and grammar, treats extractive and abstractive approaches equally, and depends on n-gram matching rather than semantics. Consequently, it evaluates summarisation quality poorly.

BERTScore [12] uses contextualised embeddings from BERT to overcome this drawback by determining semantic similarity at the token level. It has been expanded to multilingual situations using mBERT and correlates more favorably with human judgment than ROUGE for abstractive tasks [13].

2.2. LLM-as-a-Judge

Recent work has explored using large language models as evaluators [14]. GPT-4 has demonstrated a strong association with human assessments in a number of areas [15]. However, it is not feasible for iterative development due to API charges and latency [16]. The LLM-as-a-judge paradigm typically involves using a powerful language model to evaluate the quality of generated text across multiple dimensions, often correlating well with human judgments while offering scalability advantages over traditional human evaluation.

According to Fraile Navarro et al. [17], ChatGPT showed the highest scores on UniEval (coherence 0.955, consistency 0.736, fluency 0.958, relevance 0.944, overall 0.898) and clinician evaluations (coherency 0.973, consistency 0.908, fluency 0.96, clinical use 0.862), but it received the lowest scores on ROUGE, suggesting that ROUGE metrics may be unreliable for assessing clinical dialogue summarisation.

2.3. Clinical Summarisation Evaluation

For the MultiClinSum-2 shared task, Vachharajani [18] proposed an automated prompt optimisation approach that used a "judge worker" LLM paradigm to fine-tune a multilingual Qwen model via LoRA, obtaining ROUGE-L F1 scores of 0.281–0.292 across four languages and a BERTScore F1 of 0.864 in English.

Asgari et al. [19] developed a clinical safety framework to assess LLM-generated medical notes and discovered a 3.45% omission rate and a 1.47% hallucination rate across 12,999 annotated sentences.

Iterative prompt engineering reduced significant errors below previously documented human note-taking rates.

According to Sloan et al. [20], despite transformer-based architectures and multimodal inputs being state of the art, traditional NLP metrics like BLEU-4 (~ 0.25), ROUGE (~ 0.42), and METEOR (~ 0.22) correlate poorly with clinical quality.

Despite these advances, significant gaps remain. Most prior work focuses on English or fewer than five languages. Few efforts have fine-tuned a single multilingual model across a 15-language clinical summarization task. Additionally, reproducibility documentation—including fixed random seeds, complete hyperparameters, and hardware specifications—is rarely comprehensive. These gaps are addressed through a full 15-language fine-tuning pipeline for mT5-small, a streaming dataset architecture for memory efficiency, and complete reproducibility documentation.

3. System Architecture and Dataset

3.1. Dataset Statistics

The MultiClinSum-2 training set [21] was employed, consisting of 371,145 case-summary pairs across 15 languages. A per-language stratified random split (seed 42) yields 90% train (334,024) and 10% validation (37,121); evaluation is performed on the official test set of 15,000 cases (1,000 per language).

A file-indexed streaming dataset was implemented that saves only file paths and loads text on-the-fly during *getitem* to handle the entire dataset within host-memory constraints, ensuring consistent memory usage regardless of corpus size.

3.2. Base Model Selection

As the foundation model, mT5-small [22] was selected—a 300M-parameter encoder-decoder architecture (300.6M parameters; $d_model = 512$, 8 encoder and 8 decoder layers)—chosen for its 101-language pretraining on mC4, demonstrated cross-lingual transfer, and compatibility with T5 text-to-text summarization.

3.3. Full Fine-Tuning

All 300.6M parameters of mT5-small were fully fine-tuned in 32-bit floating-point precision. Training was distributed across two NVIDIA RTX A6000 GPUs (49 GB each) using PyTorch DistributedDataParallel. No adapters, quantization, or other parameter-efficient methods were employed; every model weight was updated during training.

3.4. Input Processing and Language Conditioning

A language identifier was prepended to every source document to facilitate explicit cross-lingual transfer:

`[lang] case_report_text`

Example: `[es] Paciente masculino de 45 anos presenta...`

The identifiers `[en]`, `[es]`, `[fr]`, `[pt]`, `[it]`, `[ru]`, `[ca]`, `[no]`, `[da]`, `[ro]`, `[de]`, `[el]`, `[nl]`, `[cs]`, `[sv]` were passed as plain-text prefixes and tokenized by the existing mT5 SentencePiece vocabulary (250,115 sub-word units); no special tokens were added and the embedding matrix was not resized.

3.5. Tokenization

Using the mT5 SentencePiece tokenizer with padding tokens masked for loss computation, inputs were tokenized with maximum lengths of 512 (source) and 256 (target) tokens.

4. Training Configuration

The model was fully fine-tuned using the AdamW optimizer (adamw_torch) with a learning rate of 3×10^{-5}, linear decay, and no warmup. The effective training batch size was 8 (4 per device across two GPUs under DistributedDataParallel, with no gradient accumulation). Regularization comprised weight decay (0.01) and gradient clipping to a maximum norm of 1.0; training ran in full fp32 precision with a fixed random seed of 42. Three epochs were planned; the run completed two epochs before terminating during epoch-3 evaluation (see Section 6), and the best checkpoint was selected by validation BERTScore-F1. Generation used beam search with four beams and a maximum output length of 256 tokens. Training reached 83,506 optimizer steps on two NVIDIA RTX A6000 GPUs.

5. Results

5.1. Training and Validation Dynamics

Table 1 reports the held-out validation metrics recorded during training (37,121 validation cases). Validation ROUGE and BERTScore-F1 improved from epoch 1 to epoch 2, and the epoch-2 checkpoint (step 83,506) was selected as the best model by validation BERTScore-F1. The validation BERTScore-F1 is computed by the training pipeline and is not directly comparable to the organizers' test-set BERTScore (Table 2), which uses a different scoring configuration.

Table 1
Training/validation metrics per epoch (held-out 37,121 cases)

Epoch	Step	Eval loss	ROUGE-1	ROUGE-2	ROUGE-L	BERTScore-F1
1	41,753	2.264	0.296	0.109	0.233	0.666
2 (best)	83,506	2.170	0.320	0.122	0.242	0.684

5.2. Overall Performance (Official Test Results)

The fine-tuned mT5-small model achieved consistent performance across all evaluated languages. Table 2 presents the complete per-language results returned by the official MultiClinSum-2 evaluation on 1,000 test samples per language.

5.3. Analysis of Lexical Metrics (ROUGE)

ROUGE scores reveal substantial variation across language families. Romance languages consistently outperform Germanic and Slavic languages as shown in Table 3.

ROUGE-1 and ROUGE-L are fairly high in Russian and Greek, whereas ROUGE-2 is much lower (0.073 and 0.070, respectively). This probably reflects the morphological complexity of these languages: even when unigram content is kept, bigram overlap is reduced by Greek's compound word creation and Russian's sophisticated inflectional system.

Although its ROUGE-2 (0.097) is similar to other Germanic languages, German has the lowest ROUGE-1 (0.272), which is consistent with its propensity for longer compound nouns that complicate unigram matching.

5.4. Semantic Evaluation (BERTScore)

The benefit of multilingual embeddings for cross-lingual semantic evaluation is demonstrated by the very constant BERTScore F1 scores (mean 0.834, $\sigma = 0.007$) across languages. Regardless of the source

Table 2
Per-language evaluation results (official test metrics)

Language	R-1	R-2	R-L	BERTScore	Faith.	Comp.	Flu.	Cons.
Catalan (ca)	0.367	0.160	0.247	0.840	0.658	0.689	0.694	0.338
Spanish (es)	0.372	0.159	0.254	0.837	0.630	0.507	0.690	0.361
French (fr)	0.370	0.159	0.244	0.841	0.649	0.519	0.694	0.312
Portuguese (pt)	0.343	0.141	0.238	0.835	0.609	0.472	0.685	0.351
English (en)	0.338	0.153	0.243	0.837	0.638	0.529	0.681	0.483
Romanian (ro)	0.335	0.145	0.239	0.845	0.634	0.604	0.691	0.294
Russian (ru)	0.319	0.073	0.312	0.829	0.628	0.425	0.677	0.268
Dutch (nl)	0.317	0.121	0.222	0.826	0.620	0.444	0.685	0.346
Italian (it)	0.316	0.128	0.220	0.835	0.592	0.445	0.693	0.327
Greek (el)	0.309	0.070	0.304	0.847	0.654	0.578	0.695	0.282
Czech (cs)	0.299	0.102	0.192	0.835	0.632	0.453	0.678	0.288
Swedish (sv)	0.295	0.118	0.206	0.827	0.640	0.576	0.692	0.364
Danish (da)	0.294	0.118	0.208	0.826	0.631	0.582	0.680	0.354
German (de)	0.272	0.097	0.191	0.822	0.625	0.456	0.691	0.340
Mean	**0.325**	**0.125**	**0.237**	**0.834**	**0.632**	**0.520**	**0.688**	**0.336**
Std Dev	**0.031**	**0.030**	**0.036**	**0.007**	**0.017**	**0.077**	**0.006**	**0.053**

Table 3
Performance by language family

Language Family	Languages	Mean R-1	Mean R-2	Mean R-L
Romance	ca, es, fr, pt, ro, it	0.351	0.149	0.240
Germanic	en, nl, sv, da, de	0.303	0.129	0.214
Slavic	cs, ru	0.309	0.088	0.252
Other	el	0.309	0.070	0.304

language, the fine-tuning method produces consistent semantic fidelity, as seen by the small standard deviation.

German has the lowest BERTScore (0.822) while Greek has the highest (0.847). Semantic similarity is more stable than lexical overlap across typologically varied languages, as this 0.025 range is far smaller than the ROUGE range.

5.5. Clinical Quality Assessment (Official Metrics)

The clinical-quality scores in Table 2 were computed by the MultiClinSum-2 organizers using an LLM-as-a-judge evaluation framework, where a large language model assessed generated summaries across four dimensions: faithfulness, completeness, fluency, and consistency.

Table 4
Clinical quality dimensions summary (official metrics, 0–1 scale)

Dimension	Mean	Min (Language)	Max (Language)	Range
Faithfulness	0.632	0.592 (it)	0.658 (ca)	0.066
Completeness	0.520	0.425 (ru)	0.689 (ca)	0.264
Fluency	0.687	0.677 (ru)	0.695 (el)	0.018
Consistency	0.336	0.268 (ru)	0.483 (en)	0.215

While completeness differs most ($\sigma = 0.077$), with Russian rating the lowest (0.425) and Catalan scoring the highest (0.689), fluency is the most consistent dimension across languages ($\sigma = 0.006$). English (0.483) significantly outperforms other languages in consistency, which is the weakest dimension overall (mean 0.336). This is probably due to mT5's English-heavy pretraining. Despite competitive

ROUGE-L (0.312), Russian is the most difficult language, scoring lowest on completeness (0.425) and consistency (0.268), indicating fluent but clinically insufficient summaries.

5.6. Correlation Analysis

Table 5 presents inter-metric correlations computed across the 14 evaluated languages from the official per-language scores.

Table 5
Inter-metric correlations across 14 languages

Metric Pair	Pearson r	p-value	Interpretation
ROUGE-1 ↔ BERTScore	0.600	0.024*	Moderate positive
ROUGE-L ↔ BERTScore	0.490	0.079	Moderate positive (marginal)
BERTScore ↔ Faithfulness	0.400	0.156	Weak positive (n.s.)
BERTScore ↔ Completeness	0.460	0.103	Moderate positive (n.s.)
Faithfulness ↔ Completeness	0.700	0.005**	Strong positive
ROUGE-2 ↔ Consistency	0.510	0.060	Moderate positive (marginal)

Note: $^{*}p < 0.05$, $^{**}p < 0.01$

The strongest association is between faithfulness and completeness ($r = 0.70, p < 0.01$), suggesting partial overlap in these clinical dimensions. ROUGE-1 also correlates significantly with BERTScore ($r = 0.60, p < 0.05$). The remaining pairs—including ROUGE-2 with consistency ($r = 0.51$)—show positive but non-significant trends at $n = 14$, so they should be read as indicative rather than conclusive.

6. Discussion and Conclusions

This paper has explored the strengths and limitations of a compact multilingual model for clinical case summarization. The findings demonstrate that a 300M-parameter model, when fully fine-tuned on 334,024 case-summary pairs across 15 languages, can achieve consistent semantic fidelity across typologically diverse languages. The key contributions of this work include the full fine-tuning pipeline for mT5-small, the memory-efficient streaming dataset architecture, and comprehensive documentation of hyperparameters and reproducibility measures.

The results reveal several important insights. First, lexical metrics (ROUGE) show substantial variation across language families, with Romance languages consistently outperforming Germanic and Slavic languages. This variation is likely attributable to morphological complexity, as evidenced by the particularly low ROUGE-2 scores for Russian and Greek, despite competitive ROUGE-L scores. Second, semantic evaluation via BERTScore proves remarkably stable across all languages ($\sigma = 0.007$), suggesting that multilingual embeddings effectively capture meaning regardless of surface-level lexical differences. Third, the clinical-quality assessment reveals that completeness is the most variable dimension across languages, while fluency remains consistently high. The low consistency scores across all languages (mean 0.336) indicate that factual consistency remains a significant challenge for multilingual clinical summarization.

The correlation analysis reveals that faithfulness and completeness are strongly correlated ($r = 0.70, p < 0.01$), suggesting these clinical quality dimensions capture overlapping aspects of summary quality. This finding has implications for evaluation design: if these dimensions are highly correlated, future tasks might consider consolidating or refining these metrics to better distinguish distinct quality aspects.

The limitations of this work should be acknowledged. Norwegian was inadvertently excluded from training due to a language code mismatch (the training driver referenced `no`, while the corpus stores Norwegian under `nb`), though predictions were still submitted for Norwegian. This highlights the importance of careful data preparation pipeline validation. Future work should also verify whether other

languages were similarly affected by data preparation issues, particularly given the automated nature of the corpus construction. Additionally, training did not complete three epochs due to termination during epoch-3 evaluation, and the reported results use the epoch-2 checkpoint, which was the best by validation BERTScore-F1. The data preparation pipeline validation process should be strengthened in future work to prevent similar language code mismatches.

Future work could explore several directions. First, a brief comparison against zero-shot or prompt-only baselines using larger instruction-tuned models would help contextualize how competitive a 300M fully fine-tuned model is relative to larger inference-only approaches under similar or greater resource constraints. Second, investigating whether morphological complexity consistently predicts summarization difficulty across more languages would be valuable. Third, exploring parameter-efficient fine-tuning methods (e.g., LoRA, adapters) could potentially reduce the computational cost while maintaining performance. Fourth, the moderate correlations between ROUGE and clinical quality metrics suggest that developing more clinically-aligned automatic evaluation metrics remains an important research direction.

In conclusion, this work demonstrates that lightweight multilingual models can achieve clinically meaningful performance across a wide range of languages, though challenges remain for morphologically complex languages and for factual consistency. The comprehensive documentation of all hyperparameters and reproducibility measures contributes to the broader goal of reproducible research in multilingual clinical NLP.

Acknowledgements

The authors thank the organizers of MultiClinSum-2 and CLEF 2026 for providing the dataset and evaluation framework.

Declaration on Generative AI

The authors did not use generative AI tools to produce the system, experiments, or results reported in this paper. The clinical-quality scores (faithfulness, completeness, fluency, consistency) were computed by the MultiClinSum-2 organizers as part of the official shared-task evaluation of the submitted predictions. The author(s) take(s) full responsibility for the publication's content.

References

[1] T. Searle, Z. Ibrahim, J. Teo, R. J. B. Dobson, Discharge summary hospital course summarisation of inpatient electronic health record text with clinical concept guided deep pre-trained transformer models, Journal of Biomedical Informatics 141 (2023) 104358. doi:10.1016/j.jbi.2023.104358.

[2] M. Moradi, N. Ghadiri, Text summarization in the biomedical domain, arXiv (2019). URL: http://arxiv.org/abs/1908.02285, arXiv:1908.02285.

[3] M. Arnold, M. Goldschmitt, T. Rigotti, Dealing with information overload: a comprehensive review, Frontiers in Psychology 14 (2023). doi:10.3389/fpsyg.2023.1122200.

[4] J. Xu, C. Yu, J. Xu, V. I. Torvik, J. Kang, M. Sung, M. Song, Y. Bu, Y. Ding, Pubmed knowledge graph 2.0: Connecting papers, patents, and clinical trials in biomedical science, Scientific Data 12 (2025). doi:10.1038/s41597-025-05343-8.

[5] M. Alkhalaf, P. Yu, M. Yin, C. Deng, Applying generative ai with retrieval augmented generation to summarize and extract key clinical information from electronic health records, Journal of Biomedical Informatics 156 (2024) 104662. doi:10.1016/j.jbi.2024.104662.

[6] J. J. W. Ng, E. Wang, X. Zhou, K. X. Zhou, C. X. L. Goh, G. Z. N. Sim, H. K. Tan, S. S. N. Goh, Q. X. Ng, Evaluating the performance of artificial intelligence-based speech recognition for clinical

documentation: a systematic review, BMC Medical Informatics and Decision Making 25 (2025). doi:10.1186/s12911-025-03061-0.

[7] S. Nijor, G. Rallis, N. Lad, E. Gokcen, Patient safety issues from information overload in electronic medical records, Journal of Patient Safety 18 (2022) e999–e1003. doi:10.1097/PTS.0000000000001002.

[8] J. D. Oliveira, H. D. P. Santos, A. H. D. P. S. Ulbrich, J. C. Couto, M. Arocha, J. Santos, M. M. Costa, D. Faccio, F. O. Tabalipa, R. F. Nogueira, Development and evaluation of a clinical note summarization system using large language models, Communications Medicine 5 (2025). doi:10.1038/s43856-025-01091-3.

[9] L. Bednarczyk, D. Reichenpfader, C. Gaudet-Blavignac, A. K. Ette, J. Zaghir, Y. Zheng, A. Bensahla, M. Bjelogrlic, C. Lovis, Scientific evidence for clinical text summarization using large language models: Scoping review, Journal of Medical Internet Research 27 (2025) e68998. doi:10.2196/68998.

[10] M. Rodríguez-Ortega, E. Rodríguez-López, S. Lima-López, C. Escolano, A. Mash, M. Melero, L. Pratesi, L. Vigil-Gimenez, L. Fernandez-Lopez, E. Farré-Maduell, M. Krallinger, Overview of multiclinsum task at bioasq 2025: Evaluation of clinical case summarization strategies for multiple languages: Data, evaluation, resources and results, CEUR Workshop Proceedings 4038 (2025) 19–40.

[11] M. Barbella, G. Tortora, Rouge metric evaluation for text summarization techniques, SSRN Electronic Journal (2022). doi:10.2139/ssrn.4120317.

[12] A. Alam Falaki, R. Gras, A novel unsupervised fine-tuning method for text summarization, and highlighting the limitations of rouge score, Machine Learning with Applications 20 (2025) 100666. doi:10.1016/j.mlwa.2025.100666.

[13] A. Mukherjee, V. Hassija, V. Chamola, K. K. Gupta, A detailed comparative analysis of automatic neural metrics for machine translation: Bleurt & bertscore, IEEE Open Journal of the Computer Society 6 (2025) 658–668. doi:10.1109/OJCS.2025.3560333.

[14] Z. Luo, Q. Xie, S. Ananiadou, Factual consistency evaluation of summarization in the era of large language models, Expert Systems with Applications 254 (2024) 124456. doi:10.1016/j.eswa.2024.124456.

[15] R. Mao, G. Chen, X. Zhang, F. Guerin, E. Cambria, Gpteval: A survey on assessments of chatgpt and gpt-4, in: Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024), 2024, pp. 7844–7866.

[16] A. R. Sadik, S. Brulin, M. Olhofer, Coding by design: Gpt-4 empowers agile model driven development, in: Proceedings of the 12th International Conference on Model-Driven Engineering and Software Development (MODELSWARD), 2024, pp. 149–156. doi:10.5220/0012356100003645.

[17] D. Fraile Navarro, E. Coiera, T. W. Hambly, Z. Triplett, N. Asif, A. Susanto, A. Chowdhury, A. Azcoaga Lorenzo, M. Dras, S. Berkovsky, Expert evaluation of large language models for clinical dialogue summarization, Scientific Reports 15 (2025) 1195. doi:10.1038/s41598-024-84850-x.

[18] P. Vachharajani, pjmathematician at multiclinsum 2025: A novel automated prompt optimization framework for multilingual clinical summarization, in: CEUR Workshop Proceedings, volume 4038, 2025, pp. 642–650.

[19] E. Asgari, N. Montaña Brown, M. Dubois, S. Khalil, J. Balloch, J. A. Yeung, D. Pimenta, A framework to assess clinical safety and hallucination rates of llms for medical text summarisation, npj Digital Medicine 8 (2025). doi:10.1038/s41746-025-01670-7.

[20] P. Sloan, P. Clatworthy, E. Simpson, M. Mirmehdi, Automated radiology report generation: A review of recent advances, IEEE Reviews in Biomedical Engineering 18 (2025) 368–387. doi:10.1109/RBME.2024.3408456.

[21] L. Xue, N. Constant, A. Roberts, M. Kale, R. Al-Rfou, A. Siddhant, A. Barua, C. Raffel, mt5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT 2021), 2021, pp. 483–498. doi:10.18653/v1/2021.naacl-main.41.

[22] BioASQ MultiClinSum-2 Organizers, Multiclinsum-2: Multilingual clinical summarization (task 3,

bioasq at clef 2026), https://temu.bsc.es/multiclinsum2/, 2026. Accessed: 2026-05-29.