

LAMP-MedQA: A Lightweight Multi-Agent System for Patient-Oriented Medical Question Answering

Jack Johnson¹, Meghali Banerjee¹, Joseph Crawford², James Welch²,
Jim Davies³ and Tingyan Wang^{1,3}

¹Nuffield Department of Medicine, University of Oxford, UK

²Big Data Institute, Oxford Population Health, University of Oxford, UK

³NIHR Oxford Biomedical Research Centre, Oxford, UK

Correspondence: jack.johnson@ndm.ox.ac.uk, tingyan.wang@ndm.ox.ac.uk

Abstract

Patient health literacy is critical to health outcomes, yet medical discharge summaries remain inaccessible to many patients due to jargon and complex language. Large language models (LLMs) offer a promising means of bridging this gap, but their deployment in resource-constrained hospital environments demands lightweight, privacy-preserving solutions. We evaluate a range of open- and closed-source LLMs on the MeDiSumQA dataset, comprising real patient discharge summaries paired with lay questions and clinician-verified answers, and demonstrate that larger open-source models achieve accuracy and semantic similarity performance comparable to GPT-5. We then introduce LAMP-MedQA, a lightweight multi-agent framework for patient-oriented medical question answering. The framework decomposes the task into two sequential stages: question-relevant evidence extraction and patient-facing answer simplification. Each stage is governed by an automated, metric-driven feedback loop that enables iterative self-correction without human-in-the-loop supervision. Using Qwen2.5-7B-Instruct for generation agents and Phi-3.5-Mini-Instruct for reviewer/verifier agents, it achieves significantly lower FKGL than zero-shot GPT-5, indicating better readability, and obtains the highest simplification quality (SARI) among all evaluated models, while remaining broadly competitive on accuracy and semantic similarity. This competitiveness is further improved by an offline medical glossary, which narrows the gap in n-gram overlap and contextual-similarity metrics. These results suggest that collaborative lightweight agents represent a viable approach to improving health literacy in clinical settings. Our code is available at: <https://github.com/JackJ3636/LAMP-MedQA>

1 Introduction

Patient health literacy helps ensure that treatments are accessible to the wider public. Studies have

found that fewer than half of patients entering general internal medicine units had adequate health literacy, and patients with low health literacy had a higher probability of emergency department revisits (Shahid et al., 2022). Similarly, lack of adequate health literacy is a key barrier to patients being able to cope with chronic illness (Köppen et al., 2018). Research has stressed the role of digital communication tools in improving health literacy (Fitzpatrick, 2023). Consequently, developing tools that translate medical reports into more understandable forms is crucial for enhancing patients' understanding of disease and improving health outcomes.

While previous studies have applied agentic large language model (LLM) pipelines to the simplification of published medical literature and medical question answering (Rezaei et al., 2025; Yang et al., 2025; Wang et al., 2024), little work has explored the use of LLMs to ensure that answers to patient queries are both medically grounded and readable. Recently, patient discharge summaries have presented themselves as useful documents for summarisation and improvement of patient engagement. Medical notes often contain jargon, disrupting patient communication and potentially leading to poor adherence to treatment. Thus, improving health literacy is critical in the medical domain. LLMs have the potential to extract relevant information from patient discharge summaries and present an answer to the patient that is both understandable to a non-specialist audience, and maintains the clinical accuracy of the original summary.

The MeDiSumQA dataset provides matched pairs of real patient discharge summaries and associated lay questions from a patient's perspective, and a short (typically one or two sentences) response from a medical expert (Dada et al., 2025a,b; Goldberger et al., 2000). This offers a unique opportunity to assess LLM efficacy on discharge summary extraction and simplification for patient

question-answering tasks. In this study, we first examined a number of open- and closed-source LLMs and evaluated their performance on the MeDiSumQA dataset. We then developed a multi-agent system based on lightweight open-source models, with an aspiration for use in medical settings with restricted infrastructure.

The main contributions of this study are:

- We demonstrate that larger open-source LLMs can achieve performance comparable to closed-source systems on a health literacy task through a comprehensive evaluation of both state-of-the-art open- and closed-source models.
- We introduce LAMP-MedQA, a lightweight multi-agent system comprising a dedicated *extractor*, an *accuracy reviewer*, a patient-friendly *simplifier*, and a *readability verifier*. Implemented with Qwen2.5-7B-Instruct for generation agents and Phi-3.5-Mini-Instruct for reviewer/verifier agents, LAMP-MedQA substantially improves readability and simplification quality over single-agent lightweight baselines.
- We show that LAMP-MedQA achieves the best readability (FKGL) and simplification quality (SARI) among all evaluated systems, including GPT-5, while maintaining competitive accuracy and semantic similarity, particularly when augmented with an offline medical glossary.

2 Related Work

Early multi-stage approaches demonstrated that tackling clinical note processing through discrete transformation steps could improve accessibility for lay readers (Doppalapudi et al., 2022). Named Entity Recognition has been applied to free-text clinical notes, particularly through Transformer-based models such as BioBERT and ClinicalBERT (Fraile Navarro et al., 2023). However, such technologies have proved hard to roll out to clinical deployment (Fraile Navarro et al., 2023). LLMs have demonstrated promise in medical text simplification, including the simplification of vascular patient educational materials (Xu et al., 2026). Lyu et al. proposed the "Society of Medical Simplifiers", applying multiple instances of GPT-3.5-Turbo-1106 in a 5-agent system to simplify biomedical literature (Lyu and Pergola, 2024). While strong read-

ability scores were attained, this approach relied on heavy closed-source models. This raises the question of whether lighter, open-source models would suffice.

3 Methodology

3.1 MeDiSumQA Dataset

The MeDiSumQA Dataset was created from MIMIC-IV discharge summaries (Dada et al., 2025a,b; Goldberger et al., 2000). Patient question and clinician answer pairs were then generated from the summaries using Llama-3-70B-Instruct. Question-and-answer (QA) candidates were then manually vetted by a physician who selected high-quality examples based upon factual correctness, safety and complexity, resulting in human-verified QA pairs after manual curation. We sampled 20 reports as a development set and held out 200 reports as a test set.

3.2 Initial Model Selection

We evaluate a range of open- and closed-source LLMs of varying sizes on the development set (Appendix A, Table A1). In accordance with the PhysioNet Credentialed Data Use Agreement (Goldberger et al., 2000), GPT models were accessed in a secure research environment with no data retention; our motivating goal is to develop techniques applicable to private medical datasets where data cannot leave the host environment.

3.3 Metrics

Models were assessed across accuracy, semantic similarity, simplification quality, and readability. For **accuracy and semantic similarity**, we use SBERT cosine similarity (all-mpnet-base-v2) (Reimers and Gurevych, 2019), Token F1 (Rajpurkar et al., 2016), BLEU (Papineni et al., 2002), ROUGE-L (Lin, 2004), and BERTScore (distilbert-base-uncased) (Zhang et al., 2020), and LLM-as-a-Judge scoring, where GPT-4 rates semantic similarity on a 1–5 scale as an expert medical QA judge (see Appendix B, Table B1). For **simplification quality**, we use SARI, which balances word retention, addition and deletion against the summary and reference answer (Xu et al., 2016). For **readability**, we use Flesch-Kincaid Grade Level (FKGL), with lower scores indicating simpler and more readable text (Kincaid et al., 1975).

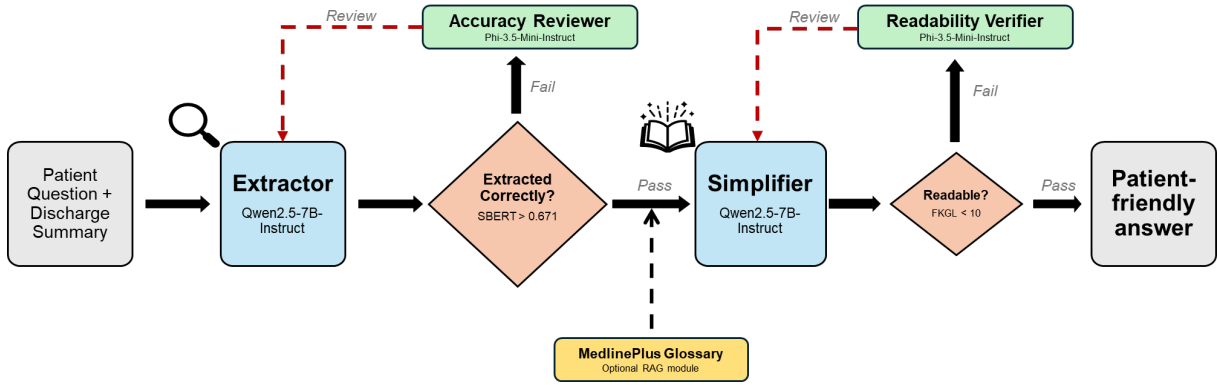


Figure 1: LAMP-MedQA pipeline. The curated response must pass an extraction-relevance gate (SBERT cosine similarity between the candidate answer and the patient’s question; threshold $\tau = 0.671$) and a readability gate (Flesch–Kincaid Grade Level < 10), with a maximum of three revisions per loop. The extraction threshold was tuned on the 20-sample development set by balanced-accuracy maximisation; the readability threshold was set from strong single-model performance on the same development set. An optional offline MedlinePlus glossary is provided to improve domain-specific model knowledge.

3.4 LAMP-MedQA

The patient-facing QA task involves two stages: accurate information extraction from long medical text, then simplification into patient-friendly language. We propose LAMP-MedQA, a multi-agent framework that decomposes the task into these stages with two corresponding iterative feedback loops, in which an *accuracy reviewer* and a *readability verifier* assess intermediate outputs against metric thresholds and trigger refinement (Figure 1). This enables iterative self-improvement without human supervision. The system is implemented using two lightweight LLMs orchestrated via LangChain, testing whether collaborative lightweight models can match larger models.

Models Model A (Qwen2.5-7B-Instruct) serves as both *extractor* and *simplifier*, while Model B (Phi-3.5-Mini-Instruct, 3.8B parameters) acts as an *accuracy reviewer* and *readability verifier* that facilitates targeted feedback. The model pipeline operates with the following stages and loops, with prompts shown in Appendix C, Table C2:

1. **Information Extraction:** Given a patient question and discharge summary, the *extractor* generates a short, accurate, and relevant answer (1–2 sentences) according to a prompt template (Appendix C, Table C2).
2. **Extraction Review:** The candidate answer is evaluated using a relevance signal computed from SBERT cosine similarity with the patient’s question, with acceptance determined by a threshold $\tau = 0.671$, tuned on the 20-sample development set. If the threshold is not met, the *accuracy reviewer* (Model B) re-

turns 1–2 sentences of actionable feedback identifying specific content gaps, and the *extractor* retries (up to 3 attempts).

3. **Text Simplification:** The *simplifier* rewrites the validated extraction in plain-English, targeting an 8th-grade reading level. In the **tool-augmented** variant, the *simplifier* first identifies medical jargon in the extraction, which is looked up in an offline MedlinePlus glossary (1,926 health topic entries pre-downloaded from the MedlinePlus Health Topics XML (U.S. National Library of Medicine, 2024)). The resulting plain-English definitions are provided to the *simplifier* alongside the extraction to guide the simplification.
4. **Readability Verification:** The simplified output is checked against a Flesch-Kincaid Grade Level (FKGL) threshold of < 10 , set from strong single-model performance on the development set (Appendix C, Table C1). If it fails, the *readability verifier* provides readability-specific feedback (e.g. shorter sentences or simpler vocabulary), and the *simplifier* retries (up to 3 attempts).

We evaluate two variants on the held-out test set: a baseline without tool augmentation and a tool-augmented variant in which the *simplifier* accesses the MedlinePlus glossary, against zero-shot GPT-5 and zero-shot/one-shot Qwen2.5-32B-Instruct baselines.

3.5 Statistical Significance Tests

Statistical significance was assessed using per-sample paired bootstrap resampling (10,000 resam-

Model	SARI \uparrow	FKGL \downarrow	SBERT \uparrow	Token F1 \uparrow	BLEU \uparrow	ROUGE-L \uparrow	BERTScore \uparrow	LLM-Judge \uparrow
GPT-5	49.81	10.91	66.67	28.65	4.27	22.05	82.27	3.98
Qwen2.5-32B (zero-shot)	48.83	11.57*	65.79	31.86*	6.39*	25.58*	82.89	3.85
Qwen2.5-32B (one-shot)	47.85*	12.90*	65.25	31.53*	6.46*	25.50*	82.48	3.72
LAMP-MedQA (no tools)	50.96	7.24*	61.38*	25.76*	4.22	20.27	81.15*	3.32*
LAMP-MedQA (with tools)	51.59	8.25*	63.60*	27.59	4.06	21.16	81.92	3.72*

Table 1: Mean scores are reported over the 200-sample test set for each system. Paired significance tests were computed only on examples with valid outputs from both systems in each comparison. Scores scaled to a 0–100 range (except FKGL and LLM-Judge (1–5)). Arrows indicate whether higher (\uparrow) or lower (\downarrow) values are better. Best score per column is **bolded**. * denotes statistically significant difference vs. GPT-5 after Bonferroni correction ($\alpha = 0.05$, $m = 90$ tests, paired bootstrap with 10,000 resamples). Significance is computed on per-sample paired differences, so a small mean gap with consistent per-sample direction can be significant while a larger but noisier mean gap may not be.

ples) and Wilcoxon signed-rank tests, with Bonferroni correction across $m = 90$ pairwise tests ($\alpha = 0.05$). Significance markers in the tables refer to paired bootstrap p -values after Bonferroni correction; Wilcoxon tests were used as a robustness check. Blank predictions receive worst-case scores.

4 Results

On the development set, closed-source and open-source models showed comparable performance, with Qwen2.5-72B-Instruct, Llama-3.3-70B-Instruct and Qwen2.5-32B-Instruct rivalling GPT-5 across SARI, FKGL, SBERT, BERTScore and LLM-Judge (Appendix C, Table C1). Qwen2.5-32B-Instruct, particularly strong on Token F1, ROUGE-L and BERTScore, was chosen as the open-source baseline.

On the 200-sample test set (Table 1), zero- and one-shot Qwen2.5-32B-Instruct showed comparable accuracy and semantic similarity to GPT-5 but significantly worse readability.

LAMP-MedQA without tools achieved substantially lower FKGL than GPT-5 (7.24 vs. 10.91; Bonferroni-corrected $p < 0.001$). It also achieved slightly higher SARI than GPT-5 (50.96 vs. 49.81), although this difference was not statistically significant. The trade-off appeared in semantic-similarity metrics: SBERT, Token F1, BERTScore, and LLM-Judge were significantly lower than GPT-5 after correction, whereas BLEU and ROUGE-L differences were not statistically significant (Table 1). This reflects a patient-facing simplification trade-off: paraphrasing into plain English reduces surface-level overlap with reference answers that retain clinical phrasing, even when the clinical content is preserved (Appendix D, Tables D1–D4).

Augmenting the simplifier with the offline MedlinePlus glossary (LAMP-MedQA with tools) closes much of this gap (Table 1): SBERT, To-

ken F1 and BERTScore are improved when compared to the without-tools variant, rising toward GPT-5’s levels. The highest SARI (51.59) of the systems evaluated is achieved by LAMP-MedQA (with tools). The cost is a small readability reduction (FKGL 8.25 vs. 7.24).

After Bonferroni correction, only FKGL, LLM-Judge and SBERT remain significantly different from GPT-5; BERTScore and Token F1 do not. LAMP-MedQA (with tools) outperforms GPT-5 in readability (FKGL 8.25 vs 10.91).

The glossary encourages slightly longer, more definition-rich phrasing, and the iterative validation loop catches cases where the no-tools variant locks onto a surface-prominent topic that does not match the patient’s question (Appendix D, Table D4).

To assess the contribution of each component, we conducted a series of ablation studies (Table 2). Removing the *readability verifier* significantly increases FKGL (7.24 \rightarrow 8.02; $p < 0.001$ after Bonferroni correction) and slightly raises LLM-Judge (3.32 \rightarrow 3.42; $p_{Bonf} = 0.04$), but leaves the n-gram overlap and contextual-similarity metrics statistically indistinguishable from the full pipeline — consistent with its role as a readability gate rather than a content gate. Removing only the *accuracy reviewer* produced no statistically significant change on any metric: based on the relevance signal derived from the patient’s question, the reviewer’s contribution to first-pass extraction quality appears limited. Removing both validation loops yields significantly worse readability than the full pipeline (FKGL 8.16 vs. 7.24; $p < 0.001$), a small but significant SBERT increase (62.99 vs. 61.38; $p < 0.001$), and a significant LLM-Judge rise (3.48 vs. 3.32; $p_{Bonf} = 0.018$), but no significant change on the remaining metrics.

We also evaluated each constituent model in isolation. Qwen2.5-7B-Instruct alone scores significantly higher on every n-gram overlap and

Configuration	SARI \uparrow	FKGL \downarrow	SBERT \uparrow	Token F1 \uparrow	BLEU \uparrow	ROUGE-L \uparrow	BERTScore \uparrow	LLM-Judge \uparrow
LAMP-MedQA (no tools)	50.96	7.24	61.38	25.76	4.22	20.27	81.15	3.32
without readability verifier	51.20	8.02*	62.05	26.59	4.49	21.06	81.47	3.42*
without accuracy reviewer	51.21	7.34	62.38	26.09	4.22	20.76	81.22	3.42
without accuracy reviewer or readability verifier	51.19	8.16*	62.99*	26.72	4.44	21.42	81.56	3.48*
Qwen2.5-7B-Instruct (zero-shot)	46.96*	11.90*	63.93*	30.12*	6.38*	23.95*	82.20*	3.64*
Phi-3.5-Mini-Instruct (zero-shot)	43.54*	13.04*	60.38	26.00	4.62	20.47	80.43	3.36

Table 2: Ablation studies for LAMP-MedQA (no tools). The first block removes validation components; the second reports each constituent model as a zero-shot baseline. Means are over the 200-sample test set; paired tests use examples with valid outputs from both systems. Scores are scaled to 0–100 except FKGL and LLM-Judge (1–5). Best scores are **bolded**. * indicates significance vs. full LAMP-MedQA after Bonferroni correction ($\alpha = 0.05$, $m = 90$; paired bootstrap, 10,000 resamples).

semantic-similarity metric than the full LAMP-MedQA (no tools) pipeline, but at the cost of significantly worse readability (FKGL 11.90 vs. 7.24) and significantly lower SARI (46.96 vs. 50.96). Phi-3.5-Mini-Instruct alone has significantly worse SARI and FKGL than the full pipeline and is not significantly different on the other metrics. Together, these results indicate that the readability verifier is the dominant driver of the pipeline’s readability, and that neither constituent model alone reproduces the readability/simplification balance of LAMP-MedQA.

5 Discussion

The close performance between open-source models and closed-source GPT-5 across both the development and test sets is promising for private hospital datasets. Qwen2.5-32B-Instruct in particular rivals GPT-5 on semantic similarity and simplification accuracy, but fails to produce sufficiently readable answers, limiting its patient-facing utility; one-shot prompting did not improve over zero-shot, motivating an agentic approach over prompt engineering.

LAMP-MedQA achieves strong simplification quality and the best readability of any system evaluated, while keeping semantic content largely intact through the iterative validation loops, all without requiring substantial human oversight. This is particularly promising for hospital settings where smaller models offer improved computational efficiency and edge deployment. The pipeline outperforms each constituent model in isolation on the readability–simplification balance (Tables 1 and 2): Qwen2.5-7B-Instruct alone is more verbose and clinically dense, and Phi-3.5-Mini-Instruct alone underperforms broadly. The lower n-gram overlap and SBERT scores relative to GPT-5 reflect a plain-English trade-off rather than a loss of content: simplified outputs share fewer tokens with

reference answers that retain clinical phrasing.

The addition of access to the offline MedlinePlus glossary mitigates this trade-off by supplying the simplifier with patient-facing definitions, raising n-gram overlap and contextual-similarity metrics towards GPT-5’s at a small readability cost (FKGL 8.25 vs. 7.24; still well below GPT-5’s 10.91). The two variants thus offer a tunable readability–accuracy trade-off: no-tools for maximum readability, and with-tools when clinical specificity matters more.

The ablation study reveals that each validation loop plays a focused, complementary role: the readability verifier delivers the pipeline’s significant FKGL improvement while preserving its content-quality profile, and the accuracy reviewer functions as a content safety net that catches occasional off-topic extractions before they propagate to the simplifier (Appendix D, Table D4). Both are retained because, in patient-facing clinical deployment, even rare lapses in readability or faithfulness are worse failure modes than the modest overhead they add.

6 Conclusion

Overall, our lightweight agentic pipeline, LAMP-MedQA, applies automated metric thresholds to iteratively refine intermediate outputs without human supervision. The system significantly improves readability in patient-facing question answering and achieves the highest SARI score of any system evaluated. At the same time, it remains broadly competitive with larger open- and closed-source models on accuracy and semantic similarity, especially when augmented with an offline medical glossary. Future work could explore alternative lightweight model configurations with agentic workflows for medical information extraction and simplification, with potential resource savings in hospital settings.

Limitations

Limited computational resources precluded a more extensive evaluation of LLMs. Future work should consider a broader range of models, as well as exploring medical glossaries beyond MedlinePlus. Additionally, while vetted and verified by a human, the MeDiSumQA dataset still used Llama-3-70B to generate QA pairs, and thus all models are still being compared against LLM-generated references. Finally, our evaluation framework could be further strengthened through human validation to better align outputs with patient needs.

Acknowledgements

T.W. acknowledges support from the National Institute for Health and Care Research (NIHR) through an Advanced Fellowship. The authors acknowledge support from the Thames Valley and Surrey Secure Data Environment (TVS SDE), hosted by Oxford University Hospitals NHS Foundation Trust. This study has been delivered through the NIHR Oxford Biomedical Research Centre (BRC). The views expressed are those of the author(s) and not necessarily those of the NIHR or the Department of Health and Social Care. The authors thank Ellie Barnes for support. The authors also thank members of the Healthcare Data Science and Intelligence Lab for their discussions and feedback on this work. The authors acknowledge the University of Oxford AI & Machine Learning Competency Centre for providing access to OpenAI API credits and infrastructure support. The authors would like to acknowledge the use of the University of Oxford Advanced Research Computing (ARC) facility in carrying out this work (<https://doi.org/10.5281/zenodo.22558>).

Conflicts of Interest

The authors have no conflicts of interest to declare.

References

- A. Dada, O. Koras, M. Bauer, A. Butler, K. Smith, J. Kleesiek, and J. Friedrich. 2025a. **Medisumqa: Patient-oriented question-answer generation from discharge letters**. *arXiv preprint arXiv:2502.03298*.
- Amin Dada, Osman Alperen Koras, Marie Bauer, Amanda Butler, Kaleb Smith, Jens Kleesiek, and Julian Friedrich. 2025b. **MeDiSumQA: Patient-Oriented Question-Answer Generation from Discharge Letters**. *PhysioNet*. Version 1.0.0.
- Saritha Doppalapudi, Tao Wang, and Robin Qiu. 2022. **Transforming unstructured digital clinical notes for improved health literacy**. *Digital Transformation and Society*, 1(1):9–28.
- P. J. Fitzpatrick. 2023. **Improving health literacy using the power of digital communications to achieve better health outcomes for patients and practitioners**. *Frontiers in Digital Health*, 5:1264780.
- David Fraile Navarro, Kiran Ijaz, Dana Rezagadegan, Hania Rahimi-Ardabili, Mark Dras, Enrico Coiera, and Shlomo Berkovsky. 2023. **Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review**. *International Journal of Medical Informatics*, 177:105122.
- Ary L. Goldberger, Luis A. N. Amaral, Leon Glass, Jeffrey M. Hausdorff, Plamen Ch. Ivanov, Roger G. Mark, others, and H. Eugene Stanley. 2000. **Physiobank, physiotoolkit, and physionet: Components of a new research resource for complex physiologic signals**. *Circulation*, 101(23):e215–e220. RRID:SCR_007345.
- J. Peter Kincaid, Robert P. Jr. Fishburne, Richard L. Rogers, and Brad S. Chissom. 1975. **Derivation of new readability formulas (automated readability index, fog count and flesch reading ease formula) for navy enlisted personnel**. Technical Report Research Branch Report 8-75, Naval Technical Training Command, Millington, TN.
- Philipp Johannes Köppen, Thomas Ernst Dorner, Katharina Viktoria Stein, Judit Simon, and Richard Crevenna. 2018. **Health literacy, pain intensity and pain perception in patients with chronic pain**. *Wiener klinische Wochenschrift*, 130(1-2):23–30.
- Chin-Yew Lin. 2004. **ROUGE: A package for automatic evaluation of summaries**. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.
- Chen Lyu and Gabriele Pergola. 2024. **Society of medical simplifiers**. In *Proceedings of the Third Workshop on Text Simplification, Accessibility and Readability (TSAR 2024)*, pages 61–68, Miami, Florida, USA. Association for Computational Linguistics.
- Kishore Papineni, Salim Roukos, Todd Ward, and Weijing Zhu. 2002. **Bleu: a method for automatic evaluation of machine translation**. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, USA. Association for Computational Linguistics.
- Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. **SQuAD: 100,000+ questions for machine comprehension of text**. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2383–2392, Austin, Texas. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. [Sentence-BERT: Sentence embeddings using Siamese BERT-networks](#). 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 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Mohammad Reza Rezaei, Reza Saadati Fard, Jayson Lee Parker, Rahul G Krishnan, and Milad Lankarany. 2025. Agentic medical knowledge graphs enhance medical question answering: Bridging the gap between LLMs and evolving medical knowledge. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 12682–12701, Suzhou, China. Association for Computational Linguistics.

Rabia Shahid, Muhammad Shoker, Luan Manh Chu, Ryan Frehlick, Heather Ward, and Punam Pahwa. 2022. [Impact of low health literacy on patients’ health outcomes: a multicenter cohort study](#). *BMC Health Services Research*, 22:1148.

U.S. National Library of Medicine. 2024. [MedlinePlus health topics](#). Accessed: 2026-03-18.

Ziqing Wang, Chengsheng Mao, Xiaole Wen, Yuan Luo, and Kaize Ding. 2024. AMANDA: Agentic medical knowledge augmentation for data-efficient medical visual question answering. *arXiv preprint arXiv:2510.02328*.

Melissa Chen Xu, Nicola Habash, Justin Do, Jiaying Bi, Dawn Salvatore, Paul Dimuzio, Michael Nooromid, and Babak Abai. 2026. [Large language models improve readability of patient education materials on vascular conditions](#). *JVS-Vascular Insights*, 4:100318.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Hang Yang, Hao Chen, Hui Guo, Yineng Chen, Ching-Sheng Lin, Shu Hu, Jinrong Hu, Xi Wu, and Xin Wang. 2025. LLM-MedQA: Enhancing medical question answering through case studies in large language models. *arXiv preprint arXiv:2501.05464*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. [BERTscore: Evaluating text generation with BERT](#). In *Proceedings of the 8th International Conference on Learning Representations (ICLR)*.

A Model Details

Table A1 summarises all models evaluated in this study. All open-source models were run locally with 8-bit quantisation for large models and a maximum output of 2,000 new tokens. Closed-source

GPT models were accessed via API in a secure research workspace with no data retention.

B LLM-as-a-Judge Prompt

Semantic similarity between predicted and reference answers was scored by GPT-4 using the prompt template shown in Table B1. The model was called with temperature = 0 and a maximum of 150 new tokens.

C Preliminary Model Selection and Prompt Templates

Preliminary model selection results are shown in Table C1, whilst model prompts are detailed in Table C2.

D Qualitative Examples

Examples of model pipeline responses are given in Tables D1, D2, D3, and D4.

Model	Source	Params	Quant.	Domain	Role in this study
<i>Closed-source models</i>					
GPT-3.5-Turbo	Closed	—	—	General	Preliminary eval
GPT-4	Closed	—	—	General	Preliminary eval; LLM-Judge
GPT-4-Turbo	Closed	—	—	General	Preliminary eval
GPT-5	Closed	—	—	General	Preliminary eval; Primary baseline
<i>Open-source models — general purpose</i>					
Qwen2.5-72B-Instruct	Open	72B	8-bit	General	Preliminary eval
Llama-3.3-70B-Instruct	Open	70B	8-bit	General	Preliminary eval
Qwen3-30B-A3B-Instruct-2507	Open	30B	8-bit	General	Preliminary eval
Qwen2.5-32B-Instruct	Open	32B	8-bit	General	Preliminary eval; Open-source baseline
Phi-4	Open	14B	fp16	General	Preliminary eval
Qwen2.5-7B-Instruct	Open	7B	fp16	General	Preliminary eval; LAMP-MedQA: extractor & simplifier
Phi-3.5-Mini-Instruct	Open	3.8B	fp16	General	Preliminary eval; LAMP-MedQA: reviewer & verifier
Llama-3.1-8B-Instruct	Open	8B	fp16	General	Preliminary eval
<i>Open-source models — biomedical</i>					
Llama3-OpenBioLLM-70B	Open	70B	8-bit	Biomedical	Preliminary eval
Meditron-7B	Open	7B	8-bit	Medical	Preliminary eval [†]

Table A1: Models evaluated in this study. Params: approximate parameter count; Quant.: quantisation level applied during inference; —, not publicly disclosed. [†]Meditron-7B produced incomprehensible, unstructured outputs on our task, resulting in extreme FKGL scores (see Appendix C, Table C1).

LLM-as-a-Judge Prompt Template

You are an expert medical QA judge. Rate the semantic similarity of the predicted answer and the reference answer.

Reference Answer (gold standard): {ref}

Predicted Answer (to evaluate): {pred}

Rate the semantic similarity on this scale:

- 1 – Not semantically similar at all
- 2 – Not very semantically similar
- 3 – Somewhat semantically similar
- 4 – Quite semantically similar
- 5 – Very semantically similar

Return ONLY a single JSON object with no markdown formatting: {"similarity": <score>}

Table B1: Prompt template used for LLM-as-a-Judge evaluation. GPT-4 was called with temperature = 0 and max tokens = 150.

Model	SARI ↑	FKGL ↓	SBERT ↑	Token F1 ↑	BLEU ↑	ROUGE-L ↑	BERTScore ↑	LLM-Judge ↑
GPT-3.5-Turbo	50.16	10.604	68.43	33.39	7.49	27.66	83.32	3.70
GPT-4	47.30	9.246	64.66	29.47	4.66	25.99	79.12	3.90
GPT-4-Turbo	48.29	9.962	69.44	33.07	3.83	28.56	83.44	3.93
GPT-5	49.43	10.973	70.83	32.70	4.70	25.90	83.76	4.35
Qwen2.5-72B-Instruct	50.38	9.735	66.02	34.50	9.19	29.47	80.32	3.95
Llama-3.3-70B-Instruct	48.14	11.443	66.24	29.53	6.22	24.32	79.15	3.75
Phi-4	44.50	11.566	67.83	30.29	6.32	25.40	82.79	3.85
Qwen2.5-32B-Instruct	48.61	10.623	69.40	35.80	6.38	29.95	84.30	3.98
Qwen2.5-7B-Instruct	45.94	11.706	66.45	30.90	4.88	25.01	82.58	3.95
Qwen3-30B-A3B-Instruct-2507	46.76	10.897	67.05	29.03	5.06	23.22	82.49	4.05
Llama3-OpenBioLLM-70B	37.68	12.131	44.34	8.95	0.11	6.93	70.43	2.10
Phi-3.5-Mini-Instruct	44.15	13.474	62.00	29.07	4.53	25.27	82.04	3.35
Llama-3.1-8B-Instruct	41.70	13.020	65.22	22.67	2.80	19.52	79.79	4.05
Meditron-7B	43.42	48.720	15.23	1.20	0.03	1.34	56.07	1.00

Table C1: Preliminary model selection: automatic evaluation metrics on 20 randomly sampled reports, rescaled to a 0–100 range (except FKGL and LLM-Judge (1–5)). Arrows indicate whether higher (↑) or lower (↓) values are better. Best score per column is bolded.

Model/Setup	Stage	Role	Prompt (abridged)
<i>Baselines (GPT-5, Qwen2.5-32B)</i>			
All baselines	System	System	You are a medical text summariser. Identify key info from a discharge summary and answer in plain-English (1–2 sentences). No jargon.
GPT-5, Qwen (0-shot)	QA	User	Question: {q} Discharge summary: {ds} Answer:
Qwen (1-shot)	QA	User	Example: {exemplar} Question: {q} Discharge summary: {ds} Answer:
<i>LAMP-MedQA Pipeline (Qwen2.5-7B-Instruct + Phi-3.5-Mini)</i>			
Model A	1. Extract	System	Medical summariser: extract relevant info and produce a short (1–2 sentence) plain-English answer.
Model A	1. Extract	User	Question: {q} Discharge summary: {ds} Answer:
Model A	1b. Retry	User	Prev answer: {a}, Feedback: {f} Improve answer (1–2 sentences, plain-English).
Model B	Feedback	System	Quality reviewer: provide actionable feedback on extraction errors.
Model B	Feedback	User	Question, discharge summary, current answer, failed gate report → give 1–2 sentence feedback (reference answer is <i>not</i> provided).
Model A	2. Simplify	System	Plain-language editor: rewrite in simple terms (1–2 sentences), no new info.
Model A	2. Simplify	User	Medical answer: {a} → Plain-English rewrite
Model A	2b. Retry	User	Prev simplification + feedback → improve (8th-grade level).
Model A	2t. Tool	System	Plain-language editor with glossary: replace jargon using definitions.
Model A	2t. Tool	User	Answer + glossary → rewrite using definitions.
Model A	2t. Retry	User	Prev output + glossary + feedback → improved rewrite.
Model B	Readability FB	System	Readability reviewer: suggest how to lower FKGL.
Model B	Readability FB	User	Text + FKGL score → give simplification feedback.

Table C2: Comparison of prompting strategies across baseline and LAMP-MedQA models. Prompts are abbreviated for space; all enforce short (1–2 sentence), plain-English outputs without medical jargon.

Question	What changes were made to my Clonazepam and Prozac during my hospital stay?
Reference	We adjusted your medication by decreasing your Clonazepam to 0.5mg twice daily as needed, and increasing your Prozac to 30mg daily.
GPT-5	Your clonazepam was cut down from 1 mg twice a day as needed to 0.5 mg once a day as needed. Your Prozac (fluoxetine) was increased from 20 mg daily to 30 mg daily.
Qwen-32B (zero-shot)	During your hospital stay, your Clonazepam dose was reduced, and your Fluoxetine (Prozac) dose was increased.
Qwen-32B (one-shot)	During your hospital stay, your Clonazepam dose was reduced to 0.5 mg daily as needed for anxiety, and your Fluoxetine dose was increased to 30 mg daily.
LAMP-MedQA (no tools)	During your hospital stay, you were given less Clonazepam—only 0.5 mg twice a day when needed for anxiety. Your Prozac dose was raised to 30 mg every day. <i>Extraction attempt 1:</i> During your hospital stay, your Clonazepam dose was reduced to 0.5 mg twice daily as needed for anxiety, and your Prozac (Fluoxetine) dose was increased to 30 mg daily. <i>Simplification attempt 1:</i> During your hospital stay, you were given less Clonazepam—only 0.5 mg twice a day when needed for anxiety. Your Prozac dose was raised to 30 mg every day.
LAMP-MedQA (with tools)	During your hospital stay, your doctor lowered the amount of Clonazepam you take to 0.5 milligrams twice a day when you feel anxious. They also raised the amount of Prozac (which is the same as Fluoxetine) to 30 milligrams each day to help with your mood. <i>Extraction attempt 1:</i> During your hospital stay, your Clonazepam dose was reduced to 0.5 mg twice daily as needed for anxiety, and your Prozac (Fluoxetine) dose was increased to 30 mg daily. <i>Simplification attempt 1:</i> During your hospital stay, your doctor lowered the amount of Clonazepam you take to 0.5 milligrams twice a day when you feel anxious. They also raised the amount of Prozac (which is the same as Fluoxetine) to 30 milligrams each day to help with your mood.

Table D1: Qualitative comparison: Example 1 (Index 30).

Question	Which of my medications were stopped, and why were they discontinued?
Reference	We discontinued your medications, including Ciprofloxacin, linezolid, and Voltaren. We stopped Ciprofloxacin and linezolid because you no longer needed them, and we stopped Voltaren because it could potentially harm your kidneys.
GPT-5	Ciprofloxacin and linezolid were stopped because you developed a new kidney problem, likely triggered by ciprofloxacin. We also stopped your Voltaren (an anti-inflammatory pain pill) to avoid further strain on your kidneys.
Qwen-32B (zero-shot)	The medications that were stopped are Linezolid and Ciprofloxacin because they may have caused your kidney problems.
Qwen-32B (one-shot)	We stopped your Linezolid and Ciprofloxacin because they may have caused your kidney problems. We also asked you to hold your NSAID, Voltaren, until your next kidney check.
LAMP-MedQA (no tools)	<p>Your doctor stopped using Ciprofloxacin, linezolid, and Voltaren because you had kidney problems that might be from an allergy. This was done to help protect your kidneys.</p> <p><i>Extraction attempt 1:</i> The medications linezolid and ciprofloxacin were stopped because you developed acute renal failure, likely due to an allergic reaction to these antibiotics.</p> <p><i>Extraction attempt 2:</i> Ciprofloxacin and linezolid were stopped because you developed acute renal failure, possibly due to an allergic reaction. Voltaren was also discontinued to avoid further kidney problems.</p> <p><i>Extraction attempt 3:</i> Ciprofloxacin and linezolid were stopped because you developed acute renal failure, likely due to an allergic reaction. Voltaren was also discontinued to protect your kidneys.</p> <p><i>Simplification attempt 1:</i> Your doctor stopped using Ciprofloxacin, linezolid, and Voltaren because you had kidney problems that might be from an allergy. This was done to help protect your kidneys.</p>
LAMP-MedQA (with tools)	<p>Your kidneys stopped working well because of an allergic reaction. To help them recover, we stopped using ciprofloxacin, linezolid, and Voltaren.</p> <p><i>Extraction attempt 1:</i> The medications linezolid and ciprofloxacin were stopped because you developed acute renal failure, likely due to an allergic reaction to these antibiotics.</p> <p><i>Extraction attempt 2:</i> Ciprofloxacin and linezolid were stopped because you developed acute renal failure, possibly due to an allergic reaction. Voltaren was also discontinued to avoid further kidney problems.</p> <p><i>Extraction attempt 3:</i> Ciprofloxacin and linezolid were stopped because you developed acute renal failure, likely due to an allergic reaction. Voltaren was also discontinued to protect your kidneys.</p> <p><i>Simplification attempt 1:</i> Your kidneys stopped working well because of an allergic reaction. To help them recover, we stopped using ciprofloxacin, linezolid, and Voltaren.</p>

Table D2: Qualitative comparison: Example 2 (Index 180).

Question	What did the doctors think was causing my symptoms, and what test results led them to think that?
Reference	We initially suspected that you had a mini-stroke, and the blocked artery in your brain supported this suspicion.
GPT-5	The team thought your confusion and the crash were mainly due to your long-standing memory problems (dementia), possibly made worse by dehydration, not a new stroke, heart attack, or injury. They felt this way because your brain scan showed shrinkage but no bleeding, your blood/urine/chest tests showed no infection, your heart tests and pacemaker check showed no heart attack, and other scans found chronic issues (severely blocked neck arteries and a weaker heart).
Qwen-32B (zero-shot)	The doctors thought your symptoms might be caused by cognitive issues like dementia, given your memory problems and confusion. They also considered possible temporary loss of control due to conditions like carotid stenosis or dehydration, but tests like the CT scan did not show any new brain injuries or acute issues.
Qwen-32B (one-shot)	The doctors initially thought your symptoms might be related to a motor vehicle accident or possible syncope due to conditions like carotid stenosis or heart issues. However, tests like the CT scan and echo showed no acute injuries or significant cardiac problems, and your symptoms were attributed to your baseline cognitive state due to dementia.
LAMP-MedQA (no tools)	<p>Your symptoms might be from a mini-stroke because a blood vessel in your brain is blocked. This is shown by a test that found your left carotid artery is completely blocked. Your high salt levels and weak heart also played a part.</p> <p><i>Extraction attempt 1:</i> The doctors thought your symptoms might be due to cognitive issues, such as dementia, and possibly carotid stenosis. Your test results showed high sodium levels (hyponatremia) and an echocardiogram indicating a moderately depressed ejection fraction. These findings suggested you had cognitive deficits and heart problems that could explain why you lost control of your car.</p> <p><i>Extraction attempt 2:</i> The doctors thought your symptoms might be due to a mini-stroke caused by a blocked artery in your brain. This is supported by the carotid ultrasound showing a completely blocked left carotid artery. Your high sodium levels and reduced heart pumping efficiency also contributed to your condition.</p> <p><i>Simplification attempt 1:</i> Your symptoms might be from a mini-stroke because a blood vessel in your brain is blocked. This is shown by a test that found your left carotid artery is completely blocked. Your high salt levels and weak heart also played a part.</p>
LAMP-MedQA (with tools)	<p>The doctors think your symptoms might be from a mini-stroke caused by a blocked artery in your brain. This is shown by a test called a carotid ultrasound, which revealed a completely blocked left carotid artery. Your high salt levels and weak heart also played a role in your condition.</p> <p><i>Extraction attempt 1:</i> The doctors thought your symptoms might be due to cognitive issues, such as dementia, and possibly carotid stenosis. Your test results showed high sodium levels (hyponatremia) and an echocardiogram indicating a moderately depressed ejection fraction. These findings suggested you had cognitive deficits and heart problems that could explain why you lost control of your car.</p> <p><i>Extraction attempt 2:</i> The doctors thought your symptoms might be due to a mini-stroke caused by a blocked artery in your brain. This is supported by the carotid ultrasound showing a completely blocked left carotid artery. Your high sodium levels and reduced heart pumping efficiency also contributed to your condition.</p> <p><i>Simplification attempt 1:</i> The doctors think your symptoms might be from a mini-stroke caused by a blocked artery in your brain. This is shown by a test called a carotid ultrasound, which revealed a completely blocked left carotid artery. Your high salt levels and weak heart also played a role in your condition.</p>

Table D3: Qualitative comparison: Example 3 (Index 45).

Question	What additional health issue did the doctors discover while I was in the hospital, and what follow-up care do I need for it?
Reference	We discovered a small cyst, about 2 centimeters in size, near your right ovary on the CT scan, and we recommend that you have a follow-up pelvic ultrasound to further evaluate it.
GPT-5	They also found a small cyst near your right ovary; if you've gone through menopause, you should have a follow-up ultrasound to check it. They noted a low blood count (anemia) as well, which needs follow-up blood tests with your primary doctor.
Qwen-32B (zero-shot)	While you were in the hospital, doctors found a small cyst on your right ovary, which might need further checking if you are past menopause. For now, the main issue was a skin infection, and you need to take the prescribed antibiotics (Ciprofloxacin and Clindamycin) for 12 days to treat it.
Qwen-32B (one-shot)	While you were in the hospital, doctors discovered that you had a staphylococcal soft tissue abscess, which required treatment with antibiotics. You will need to continue taking the prescribed antibiotics (Ciprofloxacin and Clindamycin) for 12 days and follow up with your primary care provider for further management of your diabetes and heart conditions.
LAMP-MedQA (no tools)	<p>You have an infection called a staph abscess. You need to take two medicines, clindamycin and ciprofloxacin, for 12 days. After that, see your doctor to check how you're doing and get more instructions.</p> <p><i>Extraction attempt 1:</i> The doctors discovered a staphylococcal soft tissue abscess during your stay. You need to continue taking clindamycin and ciprofloxacin for 12 days as prescribed, and follow up with your doctor after completing the antibiotics.</p> <p><i>Extraction attempt 2:</i> During your hospital stay, doctors diagnosed a staphylococcal soft tissue abscess. You need to take clindamycin and ciprofloxacin for 12 days as prescribed, and follow up with your doctor afterward.</p> <p><i>Extraction attempt 3:</i> During your hospital stay, doctors diagnosed a staphylococcal soft tissue abscess. You need to take clindamycin and ciprofloxacin for 12 days and schedule a follow-up appointment with your doctor to check your recovery and adjust your medications as needed.</p> <p><i>Simplification attempt 1:</i> You have an infection called a staph abscess. You need to take two medicines, clindamycin and ciprofloxacin, for 12 days. After that, see your doctor to check how you're doing and get more instructions.</p>
LAMP-MedQA (with tools)	<p>The doctors saw a small sac filled with fluid near your right ovary in the scan. They suggest you get another scan to check it out.</p> <p><i>Extraction attempt 1:</i> The doctors discovered a staphylococcal soft tissue abscess during your stay. You will need to take clindamycin and ciprofloxacin for 12 days to treat this infection.</p> <p><i>Extraction attempt 2:</i> The doctors found a small cyst near your right ovary on the CT scan. They recommend you get a follow-up pelvic ultrasound to further assess it.</p> <p><i>Simplification attempt 1:</i> The doctors saw a small sac filled with fluid near your right ovary in the scan. They suggest you get another scan to check it out.</p>

Table D4: Qualitative comparison: Example 4 (Index 108). The patient question asks specifically about an *additional* finding and its follow-up care. The no-tools pipeline locks onto the staphylococcal abscess (the primary diagnosis discussed in the discharge summary) across all three retry attempts and never addresses the question; LLM-Judge scores it 1/5. The with-tools variant rejects the first extraction at the relevance gate, and on retry the extractor surfaces the small ovarian cyst the question is actually asking about, recommending the follow-up pelvic ultrasound that matches the reference; LLM-Judge scores it 5/5.