Evaluating LLM-as-a-Judge for Medical Term Simplification

Ioana Buhnila[1,2] **Aman Sinha**[1,3,4] **Rohit Agarwal**[5]
Dilip K. Prasad[5] **Mathieu Constant**[1]

[1]ATILF, University of Lorraine - CNRS, France
[2]Center for Data Science in Humanities, Chosun University, South Korea
[3]IECL, University of Lorraine, France
[4]Institut Strauss, Strasbourg, France [5]UiT Tromso, Norway
Correspondence: ioana.buhnila@chosun.ac.kr

Abstract

Highly technical medical terms are difficult for patients to understand during fast-paced hospital consultations, leading them to rely on Large Language Models (LLMs) for simplified explanations. However, LLMs can produce inaccurate or false information. Since expert evaluation is costly and time-consuming, LLM-as-a-Judge (LaaJ) approach is increasingly adopted to assess the quality of LLM-generated text. In this paper, we investigate the reliability and robustness of LaaJ for specialized medical knowledge by evaluating six LLMs for their judgment capabilities on three dimensions: correctness, readability, and completeness. We utilized three judgment setups: Vanilla, Epistemic, and Bias to probe robustness, and assess them against human expert annotations to measure alignment. To address the lack of specialized medical benchmarks, we introduce BrainCancerDB, an English dataset of 219 brain cancer terms with 23,652 annotations. Our findings indicate that while LLM-Judges and humans display similar trends in ranking simplified explanations, LLM-Judges tend to be more lenient on correctness, which may have serious implications in medical setting. Additionally, we observe that hallucinations in LaaJ setups can be mitigated by epistemic markers.

1 Introduction

Large Language Models (LLMs) are used for many tasks in real world, such as question-answering, abstract writing, or recommendations (Acharya et al. 2023; Wu et al. 2024; Lyu et al. 2024). Out of many use cases, health related ones require specific caution, as people tend to use LLMs instead of real doctors for medical information or advice (Zada et al., 2025). Medical consultations are often too short to allow patients to fully interact with doctors, or to get explanations for the medical jargon (Umapathi et al., 2023). Complex medical terms such as *oligodendroglioma* or *temozolomide* are difficult to understand for patients with different levels of medical literacy. Not understanding technical medical terms can hinder the treatment of the disease, and in cases of diseases very complex to cure, such as brain cancer, simplifying medical terms for patients is of utmost importance (Ivchenko and Grabar, 2022).

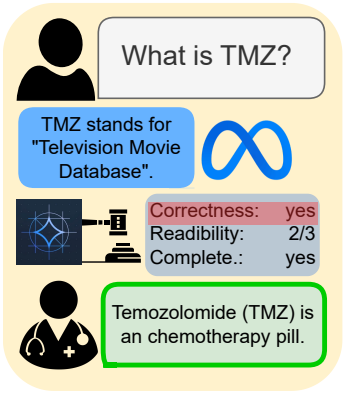

Figure 1: Illustration of limitation of LLM-as-a-Judge (Gemma-1b) evaluation in the medical domain.

There is a high risk for LLMs giving inaccurate answers that can have dangerous consequences (misinformation, erroneous scientific content, bias) (Bang et al. 2025; Huang et al. 2023; Zhang et al. 2023). LLM-as-a-Judge (LaaJ) prompting method was proposed as a solution to evaluate LLMs answers (Li et al., 2024). However, this method can fail in evaluating LLM-generated answers or it can be prone to *self-preference bias* (Wataoka et al., 2024), data contamination issues (Li et al., 2025), or lack of robustness when using positive of negative epistemic markers (eg., *I am confident*, *I am not sure*) (Lee et al., 2025).

In this study, we aimed to answer two main research questions: (RQ1) *Is LaaJ a reliable automatic evaluation method for medical term simplification?*, and (RQ2) *Are LLM-Judges aligned with human evaluation for medical term simplification?* In this paper, we used *LLM-as-a-Judge* (LaaJ) to

Proceedings of the 25th Workshop on Biomedical Language Processing (BioNLP 2026), pages 687–694

refer to the evaluation method, and *LLM-Judge* to talk about the LLM used within a Chain-of-Thought prompting framework for an evaluation task. To test the reliability and robustness of the LaaJ method, we conducted a fine-grained evaluation of the capacity of six LLMs in evaluating medical text simplification of definitions and explanations of highly medical brain cancer related terms, in three zero-shot settings: Vanilla, Bias, and Epistemic. This paper's contribution is twofold:

1. We propose `BrainCancerDB`, an English medical dataset of 219 brain cancer terms compiled from real-life hospital consultations and medical online databases. We extended the `BrainCancerDB` dataset with scientific and human simplified definitions for all 219 terms, along with 23,652 LLM-generated annotations. Moreover, we share the code of our `fine-grained evaluation pipeline` with the NLP and medical community to serve as an evaluation framework for the LLM-as-a-Judge method for the medical term simplification task[1].

2. We conducted a `two step fine-grained evaluation` of open-source LaaJ's performance in evaluating simple explanations for highly technical medical terms for three dimensions, `correctness`, `readability` and `completeness`. We compared LaaJ against human expert evaluation and SARI, a text simplification metric (Xu et al., 2016).

2 Related Work

As human expert evaluation of LLM generated content is costly and time-consuming, LLM-as-a-Judge evaluation became a convenient, cheap and fast tool (Li et al. 2024; Gu et al. 2024; Zheng et al. 2023). Nevertheless, there are few studies on the reliability and robustness of LLM-as-a-Judge as an evaluation method for very technical medical knowledge. While many studies were conducted on brain cancer MRI images or reports (Rashed et al. 2025; Kanzawa et al. 2024), there are very few research studies on brain cancer related simplification for patients. Concurrent studies worked on simplifying brain magnetic resonance imaging (MRI) reports for patients (Xu and Wang, 2024), or on summarizing brain tumour forums (Muasher-Kerwin et al., 2025). However, these studies do not share their datasets on brain cancer for further reproducibility.

[1] github.com/ATILF-UMR7118/BrainCancerDB-Corpus

medical_term: MGMT methylation

sci_definition: MGMT promoter methylation is related to the increased sensitivity of tumour tissue to chemotherapy with temozolomide (TMZ) and thus to improved patient survival.

simple_definition: MGMT methylation is an indicator of how sensitive the tumour cells are to cancer drugs treatment.

Figure 2: Instance of BrainCancerDB dataset. The `medical_term` was extracted from the brain cancer related terms collected from hospital visits, while the `sci_definition` represents the scientific definition extracted from medical websites. The `simple_definition` was manually written by a human annotator.

Our LaaJ prompting method was proposed as a solution to identify hallucinations in LLMs answers (Li et al., 2024). However, this method can be prone to *self-preference bias* (Wataoka et al., 2024) or data contamination issues, when the same family of models are used for generation and evaluation (Li et al., 2025). Benchmarks for judging the LLM-Jugdes were proposed by (Tan et al., 2024) for knowledge, reasoning, math, and coding tasks, nevertheless, not extended to the medical domain. Szymanski et al. (2025) evaluated human agreement between human experts and LLM judges, and showed 64-68% agreement for the dietetics domain and mental health. A concurrent study, Diekmann et al. (2025) showed that LaaJ achieved high accuracy of generated medical answers in terms of scientific and grammatical correctness, but they were less efficient on evaluating empathy or extent of harm.

In a recent study, Lee et al. (2025) analyzed LLM-as-a-Judge robustness when using epistemic markers in the generated outputs. Their study showed that LLMs, even proprietary ones like GPT-4o, showed lack of robustness when asked to use uncertain (*I am not sure*) epistemic markers. The authors evaluated the LLMs on a proposed benchmark, EMBER, on the general domain in English. Our study proposed a new benchmark to extend the LLM-as-a-Judge robustness evaluation to a very technical domain, such as brain cancer.

3 Dataset: BrainCancerDB

For this study, we introduced a new English dataset of 219 technical terms related to brain can-

LLM family	LLM size	LLM-JUGDES Abbreviation	LLM-STUDENTS L1B	L3B	L8B	G1B	G4B	A8B
Llama	`Llama-3.2-1B-Instruct`	**L1B**	B	–	–	EV	–	–
	`Llama-3.2-3B-Instruct`	**L3B**	B	B	–	EV	–	–
	`Llama-3.1-8B-Instruct`	**L8B**	B	B	B	EV	EV	EV
Gemma	`gemma-3-1b-it`	**G1B**	EV	–	–	B	–	–
	`gemma-3-4b-it`	**G4B**	EV	EV	–	B	B	–
Cohere	`aya-expanse-8b`	**A8B**	EV	EV	EV	EV	EV	B

Table 1: Families and sizes of LLM-Judges and different LLM-as-a-Judge / LLM-Student configurations used in our study (V: Vanilla, B: Bias, E: Epistemic).

cer treatment and disease, `BrainCancerDB`. This dataset was composed of terms collected by researchers during hospital consultations between neuro-oncologists and brain cancer patients, as well as gathered from specialized cancer websites, in consultation with cancer specialists. The dataset resulted in a list of main brain cancer terms that can be difficult to understand for patients, such as *astrocytome, PET-dopa*, or unknown drugs names like *temozolomide* or *levetiracetam*.

We extend our dataset with two set of definitions: (i) generated definitions from six different LLMs from three open-source family models: Gemma3 (Team et al., 2025), Llama3 (Grattafiori et al., 2024), and Cohere-Aya (Dang et al., 2024) (see LLM-Students in Table 1) (ii) curated definition from medical internet sources[2], and (iii) human simplified definitions for a subset of the dataset (50 terms) (example of the three data types in Figure 2).

Additionally, we conducted a manual annotation task of a subset of the dataset (50 terms) to analyze how efficient LLM-Judges are with respect to expert human judgment. For this, a human biomedical NLP expert annotator was given the task to evaluate medical definitions or explanations for the selected subset, and thus take the role of a Human-Judge (more details in Section 3.1).

3.1 Human Expert Annotation Details

Since it is very costly to conduct human medical expert annotation, we manually evaluated only 300 simple explanations, for 50 terms in 6 LLM-Student configurations. The human experts annotation followed the same three evaluation dimensions presented in Table 2. The annotation resulted in a total of 900 human medical expert annotations. Note that the LLMs were anonymized to human experts to mitigate biased judgment.

[2]We collected technical definitions from medical websites for 50 brain cancer terms. We provided the source website for each scientific definition in our dedicated GitHub page.

4 Methodology

Our paper analyzed the quality of the LaaJ evaluation method and its alignment with human annotation when used to simplify brain cancer terms. In our paper, *LLM-generated output* or *LLM answers* refer to LLM-Student[3] generated simple definitions or explanations. *LLM-Judges evaluation* refers to LLM-as-a-Judge evaluation of the LLM-Student-generated simple definitions. We detailed our method below.

Task Description. We constructed a general LLM-as-a-Judge prompt (see Table 4 in Appendix A) that guides the LLMs to generate a judgment on whether the generated outputs are scientifically correct, simple to read, and complete.

Experimental Setup. We tested three experimental setups of judgment: *Vanilla* setup (V), where the LLM-Student definitions are judged by the same-sized or bigger LLM models from the other LLM family. Secondly, *Epistemic* (E), that takes into account positive, uncertain, and negative *epistemic markers*, such as *I am confident*, *Likely*, *I'm not sure*, within the vanilla setup. We chose the most common epistemic markers used in similar LLM studies (Lee et al., 2025). Finally, *Bias* (B), that considers the same-sized or bigger LLM models from the same LLM family as a judge. We designed the biased setup to evaluate self-preference *bias*. Table 1 shows each of the configurations that are accounted for in each of the three scenarios. For the 219-term dataset, we generated through zero-shot inference a total of 7,884 LLM-Judge evaluations for each dimension, with a total of **23,652 annotations** for the three evaluation dimensions[4].

[3]LLM-Student should not be confused with Teacher-Student paradigms in knowledge distillation.

[4]Breakdown : $219\times[13_{(V)} + 13_{(E)} + 10_{(B)}] = 7884$ annotations; $7884\times3_{(\text{Correctness, Readability, Completeness})}=23562$ annotations

	Correctness	Completeness
1	text encompasses the correct medical knowledge and is in language of term	text represents a full answer, meaning the language model generated a concise answer
0	if above condition is not met	if above condition is not met
	Readability	
1	text is fluent, grammatically correct and easy to understand for laypeople	
2	text is quite difficult to understand	
3	text is very difficult to understand	

Table 2: Human expert annotation guidelines.

S(→)	L1B	L3B	L8B	G1B	G4B	A8B
			(a) Correctness (↑)			
V	$0.89_{0.31}$	$0.99_{0.10}$	$1.00_{0.00}$	$0.91_{0.28}$	$0.99_{0.10}$	$0.98_{0.14}$
E	$0.90_{0.31}$	$0.99_{0.10}$	$1.00_{0.00}$	$0.90_{0.31}$	$0.99_{0.10}$	$1.00_{0.00}$
B	$0.79_{0.41}$	$0.98_{0.14}$	$1.00_{0.00}$	$0.89_{0.32}$	$1.00_{0.00}$	$1.00_{0.00}$
H	$\mathbf{0.56}_{0.50}$	$\mathbf{0.86}_{0.35}$	$\mathbf{0.84}_{0.37}$	$\mathbf{0.64}_{0.48}$	$\mathbf{0.86}_{0.35}$	$\mathbf{0.84}_{0.37}$
			(b) Readability (↓)			
V	$1.65_{0.56}$	$1.52_{0.52}$	$1.12_{0.33}$	$\mathbf{1.65}_{0.55}$	$1.09_{0.29}$	$1.14_{0.35}$
E	$1.74_{0.54}$	$1.54_{0.50}$	$1.08_{0.34}$	$\mathbf{1.70}_{0.58}$	$1.06_{0.24}$	$1.06_{0.24}$
B	$1.77_{0.62}$	$1.56_{0.50}$	$1.04_{0.20}$	$2.04_{0.40}$	$1.96_{0.20}$	$1.02_{0.14}$
H	$\mathbf{1.06}_{0.24}$	$\mathbf{1.00}_{0.00}$	$1.12_{0.33}$	$2.04_{0.20}$	$1.06_{0.24}$	$1.06_{0.24}$
			(c) Completeness (↑)			
V	$0.81_{0.39}$	$0.98_{0.14}$	$1.00_{0.00}$	$0.64_{0.48}$	$0.96_{0.20}$	$0.98_{0.14}$
E	$0.82_{0.38}$	$0.97_{0.17}$	$1.00_{0.00}$	$0.70_{0.46}$	$0.95_{0.22}$	$0.94_{0.24}$
B	$0.49_{0.50}$	$0.63_{0.49}$	$0.98_{0.14}$	$0.52_{0.50}$	$0.92_{0.27}$	$1.00_{0.00}$
H	$0.74_{0.44}$	$0.96_{0.20}$	$0.98_{0.14}$	$\mathbf{0.82}_{0.39}$	$\mathbf{1.00}_{0.00}$	$1.00_{0.00}$
			(d) SARI metric (↑)			
MEAN	48.50	48.27	48.02	48.47	47.59	48.39

Table 3: Evaluation performance for three LLM-as-a-Judge setups: Vanilla (V), Epistemic (E), and Bias (B) against human judgments (H); and one automatic metric SARI metric (MEAN) for 50 medical terms.

Evaluation Metrics. We consider three evaluation dimensions inspired from Buhnila et al. (2024): `correctness`, `readability`, `completeness`. In the Table 2 we show the evaluation criteria used for the human expert annotation. We evaluated three dimensions: `correctness`: the generated text encompasses the correct medical knowledge in the expected language (score [1] if the two conditions are fulfilled, [0] if not); `readability`: evaluates how simple the generated answer is (scored from [1] to [3], where [1] means that the generated text is easy to understand for laypeople, [2] the text is quite difficult to understand, and [3] the text is very difficult to understand.); and `completeness`: the generated text represents a concise and full answer (score [1] if yes, [0] if no).

Additionally, we compared LLM-as-a-Judge with a traditional automatic simplification metric, SARI (Xu et al., 2016), that has been the most adapted metric for text simplification evaluation up to date (Guo et al., 2024). SARI scores the simplification level of a sentence comparing human gold annotation and LLM generated text with a technical definition reference. The simplicity level is scored from 0 to 100 (100 being the simplest).

5 Results and Discussion

Table 3 presents the evaluation of LLM-as-a-Judge setup's evaluation performance against human expert annotators for 50 medical terms. Additionally, we present in Appendix B, Table 6 the LaaJ evaluation performance across three model families on the full dataset. Our findings are as follows:

LLM-Judges are more lenient than Humans with medical correctness, but stricter with readability and completeness. On average, LLM judges achieve correctness scores of 0.95 in the Vanilla setup, 0.94 in the Epistemic setup, and 0.92 in the Bias setup, with at least 19.5% (in Bias) higher than human experts. In contrast, LLMs perceive the outputs as less readable and less complete than humans, showing a notable divergence in judgment. The largest difference in readability occurs in the Bias setup (34.4%), followed by Epistemic (21.3%) and Vanilla (18.9%). Similarly, completeness scores differ most in the Bias setup (25%), followed by Vanilla (8.7%) and Epistemic (7.6%). These results demonstrate a clear misalignment between LLM and human judgment: *LLMs tend to overestimate correctness while being stricter on readability and completeness*.

Smaller LLM-Judges hallucinate more. Hallucinations originate almost exclusively from the smaller LLM-judges (=1B), whereas larger judges (>1B) do not hallucinate, except for a few isolated cases. Among smaller judges, G1B produces 40.7%, 57.8%, and 75.8% hallucinations in the Bias, Vanilla, and Epistemic setups, respectively, while L1B shows 57.6%, 41.7%, and 23.2% hallucinations. The addition of epistemic markers reduces hallucinations by 69.3% for L1B and 27.9% for G1B, suggesting that epistemic setups are more effective in mitigating hallucinations for L1B.

Bias LLM-Judges are relatively more hallucinated. The Bias setup exhibits the highest rate of hallucinations compared to the Vanilla and Epistemic configurations. Overall, it hallucinates 11.5% of responses, whereas the Vanilla and Epistemic setups hallucinate 9.2% and 5.1%, respectively. This pattern remains consistent across all three evaluation dimensions (Figure 3), with the Bias setup showing hallucination rates of 12.4%, 8.8%, and 13.2% for correctness, readability, and complete-

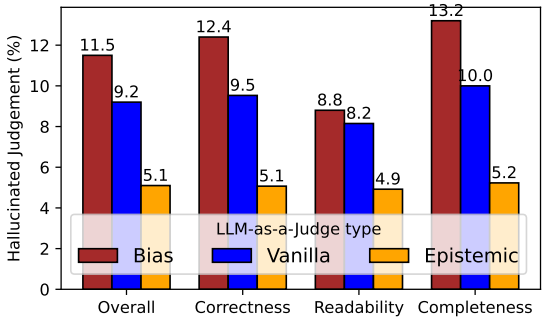

Figure 3: Percentage of hallucination in generated judgment with LLM-as-a-Judge.

ness, respectively, surpassing both Vanilla and Epistemic setups across dimensions.

Epistemic LLM-Judges are less prone to hallucinations. Across the 1,950 total annotations (650 per evaluation dimension[5]), hallucinations account for roughly 9.2% of responses in the vanilla setup, but drop to about 5.1% under the epistemic setup. This decline is consistent across all three evaluation dimensions (correctness, readability, and completeness) indicating that epistemic markers reduce hallucinations (Figure 3). Among the non-hallucinated cases, there are 580, 587, and 577 paired judgments (vanilla and epistemic both valid) for the above stated evaluation dimensions. Non-agreement between vanilla and epistemic judgments is 3.8% for correctness, 5.7% for completeness, and 15.7% for readability. Thus, epistemic markers do not substantially alter judgments for correctness or completeness but introduce a notable shift in readability evaluations. Moreover, out of the non-agreement cases, the proportion of vanilla and epistemic judgments that align with human annotations remains between 40–60% across all three dimensions, indicating that epistemic markers do not enhance human–LLM alignment.

LLM-Judges are mutually aligned. Across all three setups (Vanilla, Epistemic, and Bias) LLM-Judges demonstrated strong consistency in their evaluations, with no statistically significant differences observed in their average scores across the three evaluation dimensions (Appendix B, Table 6). This alignment is particularly evident when assessing outputs from larger student models, where judgments remain comparable across setups. However, within the Gemma family of models, the Bias setup rates readability significantly worse than the Vanilla and Epistemic setups. Similarly, for relatively smaller Llama students (<=3B), the Bias setup assigns significantly lower completeness scores, suggesting that bias-sensitive judgment may influence specific evaluation dimensions despite overall agreement among judges.

Humans are still better at simplification tasks. We compared automatic metric SARI and Humans evaluations (H) on the same 50 terms and 300 LLM generated answers (Table 3). Spearman rank correlation analysis reveals that the SARI metric aligns more closely with readability than with correctness. Across models, higher readability consistently corresponds to higher SARI scores, particularly for the Vanilla (ρ=0.93) and Epistemic (ρ=0.75) models. In contrast, correctness and completeness generally show negative correlations with SARI, with some reaching statistical significance (e.g., correctness (V) ρ=–0.90; Bias correctness (B) ρ=–0.82), indicating that SARI may undervalue medical correctness or completeness while favoring fluent outputs. These results suggest that SARI is more sensitive to surface-level readability than to correctness.

6 Conclusion

To conclude, in this work we introduce `BrainCancerDB`, a dataset of 219 brain cancer–related medical terms with both scientific and simplified definitions, which we will share with the research community. Further, our evaluation study shows that while LLM-as-a-Judge is a convenient and affordable method for evaluating simplified medical explanations, it is not fully reliable for highly technical terms. LLM-Judges tend to be more lenient on medical correctness and do not consistently align with human expert judgments regarding scientific content. Different evaluation configurations, Vanilla, Bias, and Epistemic demonstrate that even with adjustments, caution is needed when relying on LLM-Judges as the sole assessment metric. To ensure robust evaluation, LLM-based judgment should be complemented with automatic metrics such as SARI and with human expert evaluation.

Ethical Considerations

The dataset used in this study was built using only medical terms. During consultations, private or sensitive information was not collected. Patients gave their consent to the presence of the researcher during consultations. The simple definitions and

[5]650 = 13(E) x 50 medical terms.

explanations were written by human linguists annotators, as part of their involvement in the research.

Limitations

This study was conducted on English only, meaning that results can vary across multiple languages, especially on lower resource languages. For our experiments we used only general language LLMs, and not medical LLMs, as our main research goal was to evaluate the simplification performance of LLM-Students, as well as the evaluation ability of general language LLM-Judges. The dataset of our in depth analysis is small, as human brain cancer expert annotation is costly and difficult to implement, due to doctors' demanding hospital workload. Similar studies on bigger datasets or using bigger LLMs can yield different results. This study was designed to be reproducible by a wide range of research laboratories across the globe, regardless of their computational infrastructure. Therefore we tested only open source LLMs of relatively small size, that can be implemented without high computational costs.

References

Arkadeep Acharya, Brijraj Singh, and Naoyuki Onoe. 2023. Llm based generation of item-description for recommendation system. In *Proceedings of the 17th ACM conference on recommender systems*, pages 1204–1207.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. HalluLens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria. Association for Computational Linguistics.

Ioana Buhnila, Aman Sinha, and Matthieu Constant. 2024. Retrieve, generate, evaluate: A case study for medical paraphrases generation with small language models. In *Proceedings of the 1st Workshop on Towards Knowledgeable Language Models (KnowLLM 2024)*, pages 189–203.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, and 1 others. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Yella Diekmann, Chase Fensore, Rodrigo Carrillo-Larco, Eduard Castejon Rosales, Sakshi Shiromani, Rima Pai, Megha Shah, and Joyce Ho. 2025. Llms as medical safety judges: Evaluating alignment with human annotation in patient-facing qa. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 217–224.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jiawei Gu, Xuhui Jiang, Zhichao Shi, Hexiang Tan, Xuehao Zhai, Chengjin Xu, Wei Li, Yinghan Shen, Shengjie Ma, Honghao Liu, and 1 others. 2024. A survey on llm-as-a-judge. *arXiv preprint arXiv:2411.15594*.

Yue Guo, Tal August, Gondy Leroy, Trevor Cohen, and Lucy Lu Wang. 2024. Appls: Evaluating evaluation metrics for plain language summarization. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing. Conference on Empirical Methods in Natural Language Processing*, volume 2024, page 9194.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and 1 others. 2023. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *arXiv preprint arXiv:2311.05232*.

Oksana Ivchenko and Natalia Grabar. 2022. Impact of the text simplification on understanding. In *Challenges of Trustable AI and Added-Value on Health*, pages 634–638. IOS Press.

Jun Kanzawa, Koichiro Yasaka, Nana Fujita, Shin Fujiwara, and Osamu Abe. 2024. Automated classification of brain mri reports using fine-tuned large language models. *Neuroradiology*, pages 1–7.

Dongryeol Lee, Yerin Hwang, Yongil Kim, Joonsuk Park, and Kyomin Jung. 2025. Are llm-judges robust to expressions of uncertainty? investigating the effect of epistemic markers on llm-based evaluation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8962–8984.

Dawei Li, Renliang Sun, Yue Huang, Ming Zhong, Bohan Jiang, Jiawei Han, Xiangliang Zhang, Wei Wang, and Huan Liu. 2025. Preference leakage: A contamination problem in llm-as-a-judge. *arXiv preprint arXiv:2502.01534*.

Haitao Li, Qian Dong, Junjie Chen, Huixue Su, Yujia Zhou, Qingyao Ai, Ziyi Ye, and Yiqun Liu. 2024. Llms-as-judges: a comprehensive survey on llm-based evaluation methods. *arXiv preprint arXiv:2412.05579*.

Hanjia Lyu, Song Jiang, Hanqing Zeng, Yinglong Xia, Qifan Wang, Si Zhang, Ren Chen, Chris Leung, Jiajie Tang, and Jiebo Luo. 2024. Llm-rec: Personalized

recommendation via prompting large language models. In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 583–612.

Christy Muasher-Kerwin, M Courtney Hughes, Michelle L Foster, Ibrahim Al Azher, and Hamed Alhoori. 2025. Exploring large language models for summarizing and interpreting an online brain tumor support forum. *Digital Health*, 11:20552076251337345.

Essam A. Rashed, Walayat Hussain, Mohammed Mousa, and Mohammad al Shatouri. 2025. Automatic generation of brain tumor diagnostic reports from multimodality mri using large language models. In *2025 IEEE 22nd International Symposium on Biomedical Imaging (ISBI)*, pages 1–5.

Annalisa Szymanski, Noah Ziems, Heather A Eicher-Miller, Toby Jia-Jun Li, Meng Jiang, and Ronald A Metoyer. 2025. Limitations of the llm-as-a-judge approach for evaluating llm outputs in expert knowledge tasks. In *Proceedings of the 30th International Conference on Intelligent User Interfaces*, pages 952–966.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Y Tang, Alejandro Cuadron, Chenguang Wang, Raluca Ada Popa, and Ion Stoica. 2024. Judgebench: A benchmark for evaluating llm-based judges. *arXiv preprint arXiv:2410.12784*.

Gemma Team, A Kamath, J Ferret, S Pathak, N Vieillard, R Merhej, S Perrin, T Matejovicova, A Ramé, M Rivière, and 1 others. 2025. Gemma 3 technical report. arxiv 2025. *arXiv preprint arXiv:2503.19786*.

Logesh Kumar Umapathi, Ankit Pal, and Malaikannan Sankarasubbu. 2023. Med-halt: Medical domain hallucination test for large language models. *arXiv preprint arXiv:2307.15343*.

Koki Wataoka, Tsubasa Takahashi, and Ryokan Ri. 2024. Self-preference bias in llm-as-a-judge. *arXiv preprint arXiv:2410.21819*.

Likang Wu, Zhi Zheng, Zhaopeng Qiu, Hao Wang, Hongchao Gu, Tingjia Shen, Chuan Qin, Chen Zhu, Hengshu Zhu, Qi Liu, and 1 others. 2024. A survey on large language models for recommendation. *World Wide Web*, 27(5):60.

Min Xu and Yiwen Wang. 2024. Assessing the feasibility of using ai models to simplify brain imaging reports for patients: A comparative analysis of four large language models. In *International Workshop on Human Brain and Artificial Intelligence*, pages 396–406. Springer.

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.

You are an expert medical evaluator. Rate the medical definitions, explanations and paraphrases of the medical terms using specific criteria. Respond in English only with the corresponding numerical ratings in brackets [] and include an epistemic marker (e.g., I am confident, likely, I'm not sure) in round brackets () for each criterion. **Question**: [term] **Generated Answer**: [ ] Evaluate each criterion. For each, choose exactly one option from the given choices, respond with it in brackets [], and add an epistemic marker indicating your confidence in round brackets (). **Correctness**: the generated text encompasses the correct medical knowledge and it is in English (score [1] if the two conditions are fulfilled, [0] if not). **Readability**: scored from [1] to [3], where [1] means that the generated text is fluent, grammatically correct and easy to understand for laypeople, [2] that the text is quite difficult to understand, and [3] if the text is very difficult to understand. **Completeness**: the generated text represents a full answer, meaning the language model generated a concise answer (score [1] if the text respects this condition, [0] if not).

Table 4: The prompts used in this study for LLM-as-a-Judge evaluation. Text is blue denotes additional instruction for incorporating epistemic markers.

Troy Zada, Natalie Tam, Francois Barnard, Marlize Van Sittert, Venkat Bhat, Sirisha Rambhatla, and 1 others. 2025. Medical misinformation in ai-assisted self-diagnosis: Development of a method (eval-prompt) for analyzing large language models. *JMIR Formative Research*, 9(1):e66207.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, and 1 others. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, and 1 others. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in neural information processing systems*, 36:46595–46623.

A Prompts

We provide the prompts for the LLM-Student (Table 5) and the LLM-as-a-Judge setup (Table 4).

B Performance on the Full Dataset

Table 6 shows the evaluation performance of LLM-as-a-Judge on the full 219 terms dataset.

You are a helpful assistant that explains complex medical terms in simple, clear language.
Your task is to explain the following term so that a family member with no medical or technical background can easily understand it.

Term: "TERM"

Guidelines:
- Use plain, everyday language.
- Keep the explanation short (1 sentence only).
- Avoid jargon, and if you must use it, explain it simply.

Now explain the term above.

Table 5: The prompts used in this study for LLM-Student for generating medical term explanations.

S(→)	L1B	L3B	L8B	G1B	G4B	A8B
(a) Correctness (↑)						
V	$0.92_{0.27}$	$0.99_{0.11}$	$1.00_{0.00}$	$0.94_{0.25}$	$0.98_{0.13}$	$0.96_{0.19}$
E	$0.93_{0.25}$	$0.98_{0.13}$	$1.00_{0.00}$	$0.92_{0.27}$	$0.98_{0.14}$	$0.96_{0.19}$
B	$0.84_{0.37}$	$0.98_{0.14}$	$1.00_{0.07}$	$0.91_{0.28}$	$1.00_{0.00}$	$1.00_{0.07}$
(b) Readability (↓)						
V	$1.67_{0.56}$	$1.56_{0.53}$	$1.10_{0.32}$	$1.63_{0.57}$	$1.09_{0.33}$	$1.10_{0.31}$
E	$1.69_{0.52}$	$1.54_{0.51}$	$1.05_{0.24}$	$1.66_{0.56}$	$1.08_{0.29}$	$1.08_{0.31}$
B	$1.73_{0.59}$	$1.56_{0.51}$	$1.06_{0.24}$	$2.05_{0.35}$	$1.98_{0.15}$	$1.10_{0.38}$
(c) Completeness (↑)						
V	$0.85_{0.36}$	$0.98_{0.13}$	$1.00_{0.00}$	$0.66_{0.48}$	$0.97_{0.18}$	$0.95_{0.21}$
E	$0.84_{0.36}$	$0.96_{0.20}$	$1.00_{0.00}$	$0.69_{0.46}$	$0.95_{0.22}$	$0.92_{0.27}$
B	$0.50_{0.50}$	$0.63_{0.48}$	$0.98_{0.13}$	$0.66_{0.48}$	$0.94_{0.24}$	$1.00_{0.07}$

Table 6: Evaluation performance for LLM-as-a-Judge setups: Vanilla (V), Epistemic (E), and Bias (B).