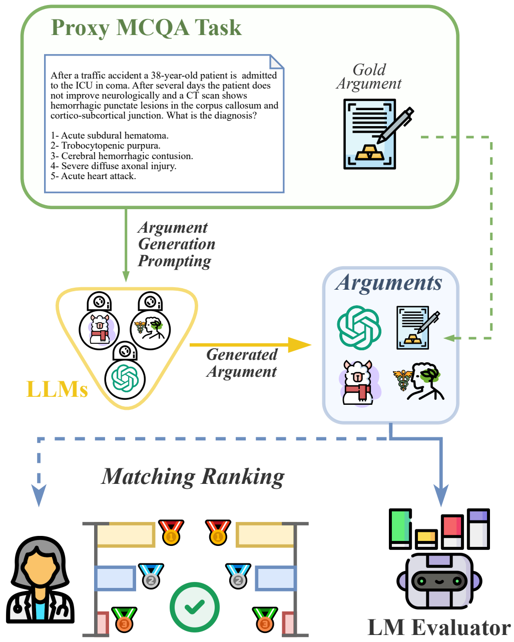
Ranking Over Scoring: Towards Reliable and Robust Automated Evaluation of LLM-Generated Medical Explanatory Arguments

Iker De la Iglesia[1]*, Iakes Goenaga[1]*, Johanna Ramirez-Romero[2], Jose Maria Villa-Gonzalez[2], Josu Goikoetxea[1], Ander Barrena[1]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU,
[2]Cruces University Hospital, (Barakaldo, Biscay, Spain)
{iker.delaiglesia, iakes.goenaga, ander.barrena}@ehu.eus

Abstract

Evaluating LLM-generated text has become a key challenge, especially in domain-specific contexts like the medical field. This work introduces a novel evaluation methodology for LLM-generated medical explanatory arguments, relying on Proxy Tasks and rankings to closely align results with human evaluation criteria, overcoming the biases typically seen in LLMs used as judges. We demonstrate that the proposed evaluators are robust against adversarial attacks, including the assessment of non-argumentative text. Additionally, the human-crafted arguments needed to train the evaluators are minimized to just one example per Proxy Task. By examining multiple LLM-generated arguments, we establish a methodology for determining whether a Proxy Task is suitable for evaluating LLM-generated medical explanatory arguments, requiring only five examples and two human experts. The Proxy Tasks, LM evaluators, and the code are available for reproducibility[1].



After a traffic accident a 38-year-old patient is admitted to the ICU in coma. After several days the patient does not improve neurologically and a CT scan shows hemorrhagic punctate lesions in the corpus callosum and cortico-subcortical junction. What is the diagnosis?

1- Acute subdural hematoma.
2- Trobocytopenic purpura.
3- Cerebral hemorrhagic contusion.
4- Severe diffuse axonal injury.
5- Acute heart attack.

Figure 1: Graphical abstract illustrating the key elements of our approach. Synthetic arguments are first generated by prompting multiple LLMs, which are then ranked alongside gold-standard arguments by both our trained LM evaluator and a human expert. Our results show the LM evaluator aligns with human preferences.

1 Introduction

The field of Natural Language Processing (NLP) has undergone a transformative evolution with the advent of Language Models (LMs) and Large Language Models (LLMs). The results in the medical domain have been particularly notable, with LLMs achieving remarkable accuracy in solving medical exams (Singhal et al., 2023; Strong et al., 2023; Liu et al., 2024). This success is driving the ongoing development of these models to further enhance support for Evidence-Based Medicine (EBM) which involves the conscientious, explicit, and thoughtful use of present best medical evidence in making medical decisions (Sackett et al., 1996). With the advent of large autoregressive generative models, decoder-only architectures such as GPT (Radford and Narasimhan, 2018) and Llama (Touvron et al., 2023) have been increasingly used for pre-training on medical text data, leading to notable improvements in the coherence and relevance of generated medical explanations. However, the evaluation of such explanatory arguments remains a considerable challenge (Chang et al., 2023), particularly within the medical domain, where obtaining meaningful datasets and assessing accuracy is inherently difficult. The high-entropy nature of language allows for multiple valid responses, complicating the evaluation of relevance, coherence, and factual accuracy. This complexity is further exacerbated by the challenge of objectively quantifying these factors while also accounting for human preferences.

[*] Equal Contribution.

[1] https://github.com/hitz-zentroa/Ranking-Over-Scoring-COLING-2025

Proceedings of the 31st International Conference on Computational Linguistics, pages 9456–9471
January 19–24, 2025. ©2025 Association for Computational Linguistics

Despite the long tradition of automatic long-text evaluation, particularly in Machine Translation, it remains an unresolved challenge. Metrics like BLEU for translation (Papineni et al., 2002), ROUGE for summarization (Lin, 2004), and embedding-based scores such as BERTScore (Zhang* et al., 2020), BLEURT (Sellam et al., 2020), and COMET (Rei et al., 2020) have been widely used, but they present major issues for evaluating explanatory arguments. These metrics rely on reference texts, which are difficult to obtain in the medical domain, and often overestimate irrelevant differences due to the high entropy of valid arguments. This problem extends beyond explanatory argumentation to all long-text evaluations, as noted in the literature (Liu et al., 2023; Sulem et al., 2018; Sun et al., 2022).

More recent approaches have tried to address these problems using LLMs as Judges (Zheng et al., 2024; Li et al., 2024; Kocmi and Federmann, 2023; Shen et al., 2023). Despite the increasing popularity of such evaluators, it has been observed that these models often exhibit their own biases; *self-enhancement bias*, tending to recognize and favor their own outputs over those generated by other models, *positional bias*, namely, the propensity to favor certain positions over other and *verbosity bias*, prioritizing lengthier, more verbose responses, even when they fall short in clarity, quality, or accuracy compared to more concise options (Panickssery et al., 2024; Zheng et al., 2024).

Given this perspective, it is clear that the most reliable way to evaluate the responses generated by LLMs is through human evaluation. Nonetheless, this method has significant drawbacks especially when it comes to highly specialized domains like the medical domain. Finding experts capable of accurately evaluating the responses is very difficult. Added to this, human evaluation is very expensive and time-consuming and in the case of long-text evaluation it is difficult to assess quality guidelines properly. This is particularly evident in extreme cases, where multiple correct responses make the differences too subtle to evaluate, or when the generated texts are incorrect, making it challenging to assess the results objectively, for example, in cases where two argumentative explanations use different but completely non-sense evidence.

Considering that explanatory arguments are intended to assist medical decision-making within an EBM framework, the present proposal aims to measure the adequacy of the explanatory arguments in making medical decisions, modeled using proxy tasks along the lines of (Tan et al., 2024). This way, the following *Research Questions* arise:

RQ1 Can we develop a discriminative LM evaluator that reliably aligns with human preferences when assessing medical explanatory arguments in EBM, while avoiding the biases commonly seen in generative LLM-based evaluators?

RQ2 Are LM evaluators built upon Proxy Tasks that model EBM suitable for evaluating the goodness of LLM-generated medical arguments and robust against adversarial attacks?

RQ3 Are all Proxy Tasks equally valuable in evaluating LLM-generated medical arguments, and what factors influence their suitability for reliable argument ranking?

RQ4 How consistent are human evaluations across different Proxy Tasks, and can this consistency be used as an indicator of the suitability of a task for automatic argument evaluation?

RQ5 Does a higher Proxy Task score correlate with better alignment to human criteria when ranking medical arguments, or is there a better approach to achieve this alignment?

By exploring these research questions, we aim to introduce a fast and cost-effective automatic evaluation method to evaluate medical explanatory argumentation provided by LLMs in the framework of EBM. To do so, we propose to use a discriminative LM for evaluating the arguments generated by LLMs, rather than evaluating these arguments against a reference gold-standard text. In our approach the discriminative LM evaluator will indirectly evaluate how helpful and informative the generated arguments are via three Proxy Tasks, namely, Medical Question Answering, Misinformation, and Natural Language Inference in clinical trials. More specifically, we will compute the results for different types of arguments across the three Proxy Tasks to rank their contributions to the performance in these tasks. Subsequently, we will analyze how the rankings produced by this discriminative LM evaluator align with those conducted by expert physicians. We also analyze the adequacy of each Proxy Task for evaluating accurate explanatory arguments. Overall, our proposal eliminates the need for evaluations by subject matter experts and the presence of a reference gold-standard explanatory argument, while also minimizing certain biases of generative LLM judges.

2 Related Work

The development of LLMs in the medical domain focuses nowadays on scaling up pre-training data and model parameters or adapting general-purpose LLMs to the medical domain. Notable examples include Med-PaLM 2 and Meditron. Med-PaLM 2, achieved 86.5% accuracy on MedQA (US Medical Licensing Exam-style questions), surpassing the previous state-of-the-art (Singhal et al., 2023), while Meditron integrates diverse medical information for comprehensive insights and high-quality medical argumentation (Chen et al., 2023). However, we will not focus on directly assessing the capability of LLMs to solve tasks but rather evaluating the informativeness of LLM-generated explanatory arguments in the medical domain, which is doubly challenging.

To address the issue of long-text evaluation in a general domain, Tan et al. (2024) propose using a QA task as a proxy to assess the helpfulness and relevance of content generated by LLMs. Their system comprises two key components: meta-questions and proxy-questions. Meta-questions prompt LLMs to generate comprehensive, factually correct text requiring a full understanding of the topic, while proxy-questions evaluate the quality of the generated content by assessing whether it includes sufficient relevant and accurate information. For example, if the meta-question asks about the First Industrial Revolution, a proxy-question might be, "True or False: The steam engine played a crucial role in the First Industrial Revolution." The authors compare their Proxy-QA evaluator with human evaluators and LLM-as-judges, using GPT-as-Judge. They randomly sample ten meta-questions and use four LLMs to generate long-texts. These 4 generated text candidates are then evaluated through pairwise comparisons between three evaluators (their Proxy-QA evaluator, human evaluators, and GPT-as-judge) using the win rate measure[2].

The primary goal of this pairwise comparison in win rate is to determine how closely ProxyQA correlates with human judgment compared to LLM-based evaluations.

The authors found that ProxyQA's evaluations were highly correlated with human preferences, whereas GPT-as-judge tended to overestimate the quality of the text generated by GPT models. ProxyQA showed a balanced and reliable evaluation, reflecting human preferences more closely than GPT evaluators, which were biased towards outputs from GPT-based models. Scalability and domain adaptation is one of the main pitfalls of this method, creating and maintaining high-quality meta-questions is human-intensive. Additionally, the results lack of comparison between the performance of the systems with or without the generated long-text making it difficult to assess the real impact that adding the generated long-text has on solving the Proxy Task. Our proposal does not require building meta-questions and we include a Naive version of every system where there is no explanatory argument included to solve the Proxy Tasks. We also extend the number of Proxy Tasks.

Yao et al.'s work also explores long-text evaluation, with a particular emphasis on human-annotated natural language explanations to assess whether they consistently enhance machine learning models in NLP. Especially relevant for this work is their analysis of how human-annotated explanations show varying levels of helpfulness, depending on the task and dataset used. The study evaluates five large-scale datasets (e.g., CoS-E, e-SNLI) using two NLP models (T5 and BART) to assess explanation quality. Their findings show that explanations in ECQA are highly beneficial, while CoS-E explanations, although noisy, still offer improvements in model predictions. This suggests that explanation evaluation should focus on task-specific performance rather than treating all explanations as equally valuable. While they introduce a metric to assess the helpfulness of long texts, they neither compare different explanations nor verify if their metric aligns with human preferences.

To summarize, there is an urgent need for an objective system for independent evaluation of modern LLMs' medical argument generation abilities. To address this, we have developed a medical argumentation evaluation method based on Proxy Tasks that aligns with the assessments of medical experts. Our evaluation method allows us to assess medical argumentations quickly, efficiently, and cost-effectively.

[2] Win Rate Calculation: The win rate is calculated based on pairwise comparisons of the reports. If one model's output is preferred over another, it wins that comparison. This win rate measures how often one model's report is rated better than others by the evaluators.

3 Experimental Setup

In this study, we developed an experimental framework to investigate the alignment between LM evaluator systems and human preferences in assessing explanatory arguments. Argument quality is indirectly estimated by its impact on Proxy Task performance. These tasks are handled by LMs trained to perform the original task. These LMs also serve as evaluators when incorporating explanatory arguments as additional input, by ranking the incorporated arguments based on the task score.

The departure point of our approach is the generation of explanatory arguments. Indeed, they comprise the base of our approach, and they will be generated by humans or LLMs. On the one hand, each task will have high-quality arguments written by human experts that we will consider as the gold standard. On the other hand, we will generate diverse arguments for each task using different LLMs. The main focus of the evaluation approach presented in this paper is focused on these two kinds of arguments, termed *Primary Arguments*.

Regarding the Proxy Tasks, we employ a diverse set, including Medical Multiple Choice Question Answering (MMCQA), Medical Misinformation Detection, and Natural Language Inference (NLI) in clinical trials. These tasks are selected because they represent different contexts where explanatory argumentation is helpful, each task requiring specific types of arguments. By employing a diverse set of Proxy Tasks rather than relying on a single one, we aim to explore which tasks are most relevant and suitable for evaluation purposes (addressing RQs 2 and 3).

We will also have two types of evaluators: human evaluators and LM ones. For the latter, we train discriminative LMs on the Proxy Tasks to function as evaluators, as mentioned previously. The evaluators thus provide an indirect assessment by leveraging task performance metrics to differentiate between arguments, thereby addressing the potential biases associated with LLMs as Judges-based evaluation methods (RQ1 and RQ2). Alongside these Proxy Tasks, human experts independently estimate the quality of arguments within the context of these Proxy Tasks, providing a standard against which the evaluators can be compared, addressing RQs 3 and 4.

Therefore, we will have human and LM evaluators, and, essentially, the core of our analysis focuses on examining which of the latter aligns most closely with the former. We analyze the degree to which the rankings generated by the LM evaluators reflect human preferences, thereby assessing not just task performance but also the meaningfulness of the rankings in the context of human-aligned argument evaluation. We also examine whether the LM evaluator with maximized overall Proxy Task score is the one with the closest ranking alignment with human criteria (RQ5).

To further test the robustness and ability to discern the quality of the arguments of our LM evaluators, we introduce a second set of arguments, termed *Control Cases*, which complement *Primary Arguments*. Through this approach, LM evaluators are tested with four adversarial scenarios during inference, designed to assess their ability to distinguish meaningful arguments from irrelevant or misleading content, as detailed in subsubsection 3.2.2.

Through this experimental setup, we aim to thoroughly investigate the effectiveness of Proxy Task-based evaluators in modeling human judgment and the relative value of different Proxy Tasks.

3.1 Proxy Tasks & Proxy Task LM Evaluators

3.1.1 Proxy Tasks Benchmarks

We repurposed three diverse benchmarks as Proxy Tasks, each selected to capture distinct types of argumentation, offering a broad evaluation across a range of complex scenarios. One of the reasons for selecting these three datasets is that they include a complementary gold-standard argument supporting the correct label, which is unnecessary for performing the base task, as the tasks were originally designed to be performed without arguments. In this paper, we incorporate these complementary arguments into a broader set of Primary Arguments and Control Cases, testing them individually to analyze their contribution to task behavior. Detailed examples of instances from each dataset can be found in Appendix D.

Medical Multiple Choice QA Benchmark We employed the English translation of the *CasiMedicos* dataset (Goenaga et al., 2023), which assesses models' ability to answer medical multiple-choice questions. Each instance includes a question with a clinical case, possible answers, and a gold-standard explanation supporting the correct choice. The original split distribution was kept. To reduce label prior bias, ensuring the model predicts correct answers based on content rather than answer order, we preprocessed the dataset by creating multiple versions of each instance, varying the position of the correct answer. Additionally, we modified the gold-standard explanations by removing statements that explicitly identified the correct answer and replaced references to the answer's position with the answer's text.

Misinformation Detection Benchmark We employed a subset of the English version of the *HealthFC* dataset (Vladika et al., 2024), which focuses on indicating whether health-related claims are supported, refuted, or lack enough information. The dataset contains 742 instances, which we stratified and split into 70% (518) for training, 15% (111) for development, and 15% (112) for testing. As mentioned in subsection 3.3, instances labeled "Not Enough Evidence" in the test split were excluded from both human and automatic rankings when calculating the final scores. We termed this subset the *Misinformation With Evidence dataset*.

NLI Benchmark The *NLI4CT* clinical trial dataset (Jullien et al., 2023) contains clinical trial records (CTR), including a medical statement and a label indicating whether the CTR supports or contradicts the statement. Unlike the other tasks, the arguments in this case are extracted directly from the CTR. While the original dataset incorporates instances involving two clinical trials, we focused solely on those involving a single trial. The distribution is 1035 instances (74%) for training, 140 instances (10%) for development, and 229 instances (16%) for testing.

3.1.2 Discriminative LM Evaluators

This section introduces our key contribution: the discriminative LM evaluators, designed to systematically rank medical explanatory arguments without direct human involvement. These evaluators will be compared with expert assessments to see which approach aligns most closely with human judgments. Our method uses discriminative language models trained on Proxy Tasks, avoiding the bias that generative LLMs introduce when acting as evaluators. Generative models tend to favor arguments similar to those they generate, whereas discriminative models focus purely on task performance. This ensures a more objective ranking based on how effectively the arguments improve Proxy Task outcomes.

We developed three evaluators, all based on the EriBERTa encoder model (De la Iglesia et al., 2023), each trained with different types of arguments. Table 5 in Appendix B outlines the training inputs for each evaluator based on each Proxy Task. We used the *train* and *dev* splits for training and tuning, and the *test* split for the final ranking.

Baseline Evaluator It serves as the simplest model in this work. It is the original classification task, which means, training without the complementary arguments.

Expert-Trained Evaluator Trained using human-crafted gold-standard arguments. This evaluator is expected to align most closely with human judgment, as the training data comes directly from domain experts.

LLM-Trained Evaluator One key contribution of this study, an LM trained exclusively with synthetic arguments generated by LLMs (detailed in subsubsection 3.2.1). Each training instance includes an argument randomly selected from various LLMs, ensuring a balanced representation. This approach allows the evaluator to learn diverse argument styles, reducing favoritism toward any specific LLM-generated argument and improving its neutrality and robustness in assessing argument quality. We trained three models with different argument sets to minimize bias and variability.

3.2 Primary Arguments and Control Cases

Primary Arguments and *Control Cases* are two main components of our evaluation framework. *Primary Arguments* are central to our research and include both gold-standard arguments crafted by domain experts and synthetic arguments generated by various LLMs. These arguments are the only ones also evaluated by human experts, providing a benchmark for comparing the performance of our automated LM evaluators. In contrast, *Control Cases* are designed to test the robustness of the LM evaluators by incorporating misleading or irrelevant content (see Figure 2).

3.2.1 LLM-Generated Synthetic Arguments

To evaluate automated medical argumentation, we generated synthetic arguments using three LLMs: GPT-4o[3] (OpenAI, 2024), known for its strong general reasoning abilities; OpenBioLLM[4] (Ankit Pal, 2024), a model fine-tuned on large-scale biomedical datasets for high accuracy in medical text generation; and Llama3-70B-Instruct[5] (Meta, 2024),

[3] GPT-4o-2024-05-13

[4] aaditya/Llama3-OpenBioLLM-70B

[5] meta-llama/Meta-Llama-3-70B-Instruct

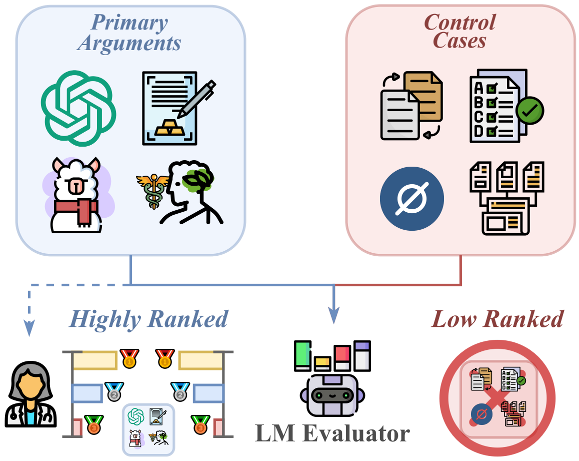


Figure 2: A graphical abstract illustrating the system's main components and behavior. The proposed LM evaluator prioritizes ranking primary arguments first and placing control cases last.

which we instruction-tuned for each Proxy Task to optimize its performance in the medical domain (see Appendix C for an example). The OpenBioLLM and Llama 3 models have been quantized to 4 bits to enhance computational efficiency and reduce memory usage during argument generation.

For consistency and minimum human intervention, we used a single example for one-shot prompting for each Proxy Task. For MMCQA and Misinformation Detection tasks, the LLMs generated free-style explanatory arguments, while for NLI tasks, they extracted evidence from the CTR. Identical generation parameters were used across all models: maximum token length of 256, sampling enabled, temperature of 0.9, and top-p at 0.85.

3.2.2 Control Cases

Control Cases are designed to assess the LM evaluators' ability to differentiate between meaningful arguments and irrelevant or misleading content. *Control Cases* serve as adversarial attacks for evaluators (Jia and Liang, 2017). This section outlines the construction of these cases and their role in our broader experimental framework.

No Argument This case includes no medical argumentation, testing whether evaluators actually rely on arguments when making predictions. Evaluators trained to evaluate medical arguments are expected to struggle, as they rely on the presence of explanations for predictions.

Label-Only Input The correct answers to the Proxy Tasks are provided but without any supporting argumentation. The purpose is to see if evaluators penalize the lack of argumentation, despite having the correct answers. We expect evaluators trained on medical argumentation to prioritize explanations and perform worse compared to instances with proper arguments.

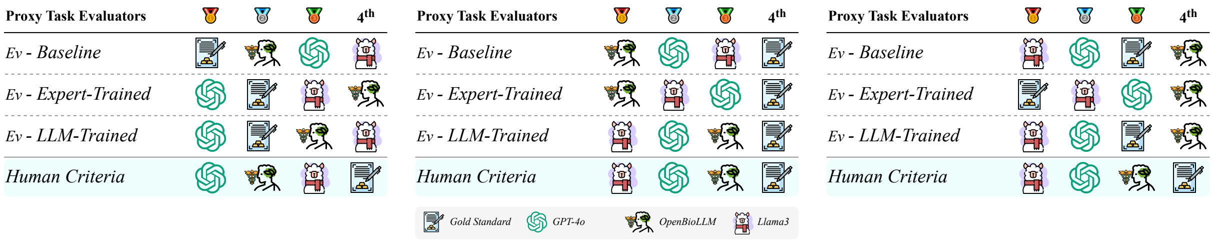
Noise Argument In this scenario, medical arguments are present but irrelevant to the instance, having been randomly selected from unrelated examples. We anticipate that well-trained evaluators will recognize the mismatch and perform poorly, as the arguments do not align with the instance.

IR Passages In this test, we use passages from the WikiMed corpus (Vashishth et al., 2021), retrieved via an Information Retrieval (IR) system. While these passages contain medical information, they do not necessarily constitute coherent or valid arguments. This case is designed to challenge evaluators in distinguishing between structured medical arguments and mere informative text. Passages were retrieved by indexing full documents with FAISS (Johnson et al., 2019) using the Deka et al. (2022) embedding model, querying each instance's text, and extracting the top five documents. These were split into 300-character chunks and reranked using ColBERTv2 (Santhanam et al., 2022), with the top three passages fed to the evaluator.

3.3 Human and Automatic Ranking

We engaged two clinicians with prior experience in medical annotation and system evaluation, utilizing the *test* split of the datasets. After a preliminary round, 5 examples were ranked for each task to calculate the Inter-Annotator Agreement (ITA). Experts ranked the four *Primary Arguments* on a scale of 1 to 5, with 5 assigned to clearly incorrect arguments, and ties were allowed when arguments were of equal quality. We used Krippendorff's alpha (Krippendorff, 2011) to calculate ITA, achieving the following scores: MMCQA=0.72, Misinformation=0.61, and NLI=0.44.

In the ITA phase, we noticed significant disagreement in the Misinformation Detection task, particularly for instances labeled as "Not enough evidence". To address this, we removed those instances and recalculated ITA using 14 new examples, improving the alpha to 0.73. After this adjustment, the clinicians ranked the arguments independently: 61 instances for MMCQA, 39 for Misinformation, and 98 for NLI.



Proxy Task Evaluators	🥇	🥈	🥉	4th
Ev - Baseline	Gold Standard	OpenBioLLM	GPT-4o	Llama3
Ev - Expert-Trained	GPT-4o	Gold Standard	Llama3	OpenBioLLM
Ev - LLM-Trained	GPT-4o	Gold Standard	OpenBioLLM	Llama3
Human Criteria	GPT-4o	OpenBioLLM	Llama3	Gold Standard

(a) MMCQA

Proxy Task Evaluators	🥇	🥈	🥉	4th
Ev - Baseline	OpenBioLLM	GPT-4o	Llama3	Gold Standard
Ev - Expert-Trained	OpenBioLLM	Llama3	GPT-4o	Gold Standard
Ev - LLM-Trained	Llama3	GPT-4o	OpenBioLLM	Gold Standard
Human Criteria	Llama3	GPT-4o	OpenBioLLM	Gold Standard

Gold Standard *GPT-4o* *OpenBioLLM* *Llama3*

(b) Misinformation With Evidence

Proxy Task Evaluators	🥇	🥈	🥉	4th
Ev - Baseline	Llama3	GPT-4o	Gold Standard	OpenBioLLM
Ev - Expert-Trained	Gold Standard	Llama3	GPT-4o	OpenBioLLM
Ev - LLM-Trained	Llama3	GPT-4o	Gold Standard	OpenBioLLM
Human Criteria	Llama3	GPT-4o	OpenBioLLM	Gold Standard

(c) NLI

Figure 3: Ranking of the *Primary Arguments*. Each row corresponds to a distinct evaluator: the first three rows correspond to our proposed Proxy Task evaluators based on discriminative classification models, while the last row reflects the human criteria, obtained by having experts directly rank the arguments.

For both human and LM evaluator rankings, we calculated the average rank for each system and used a Friedman non-parametric test (Friedman, 1937) ($\alpha = 5\%$) to assess significant differences. In the human rankings, the NLI task was the only one that failed the Friedman test ($p = 0.561$), reflecting its low ITA. In contrast, the other two tasks, which had relatively high ITA given the difficulty of argument ranking, passed the test.

4 Results

This section will first present the main results, comparing the proposed automatic evaluators to human criteria. Finally, we will examine the *Control Cases* to demonstrate the automatic evaluators' ability to discard non-argumentative inputs.

4.1 Automatic Evaluations Results

Figure 3 presents the main results of this study. Setting aside evaluator accuracy scores, we focus on the rankings produced by the proposed three discriminative evaluators and human criteria. For the MMCQA task (a), the rankings demonstrate that, in the absence of the gold standard, the LLM-trained evaluator aligns with human criteria when ranking LLM-generated synthetic arguments, with GPT-4o being the top choice in 3 out of 4 rankings. The lower ranking of gold-standard arguments by human experts stems from their design for last-year medical students. These arguments prioritize straightforwardness, highlighting only key elements needed to discern the correct answer, assuming prior knowledge. However, in the context of analyzing clinical cases rather than exam preparation, a higher degree of contextualization is preferred. For the NLI task (c), a similar pattern emerges, but here the finetuned Llama3 model ranks first. In the misinformation task (b), the LLM-trained evaluator perfectly matches human criteria, ranking Llama3-generated arguments first.

Regarding the evaluators, the lack of argumentation during training causes the baseline evaluator to produce rankings that do not align with human criteria. In contrast, the expert-trained evaluator improves upon the baseline. However, the LLM-trained approach proves to be the winning strategy, demonstrating that we can effectively evaluate LLM-generated argumentation by using synthetic data and training discriminative evaluators, without relying on human-generated arguments.

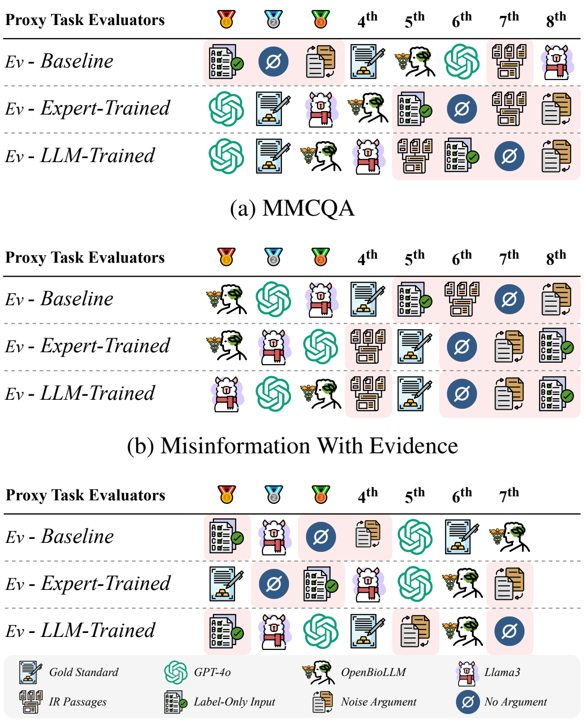
As mentioned, LLMs acting as judges tend to overestimate self-generated text and show a preference for longer responses. Our approach addresses the first issue by using an EriBERTa encoder. We also observed that the longest text in the MMCQA task was generated by OpenBioLLM, in the misinformation task by Llama3, and again by OpenBioLLM in the NLI task. The rankings provided by the LLM-trained evaluator in Figure 3 demonstrate that this length bias is absent in our approach for MMCQA and NLI tasks. In the case where the bias appears, such as Llama3 in the misinformation task, human evaluators also ranked it first.

As previously mentioned, the best evaluator does not necessarily produce the highest Proxy Task scores. The left side of Table 1 shows the average dataset scores for each evaluator. While the expert-trained evaluator produces the highest scores, the LLM-trained evaluator is the one most aligned with human judgment (see Figure 4 and Tables 2, 3 and 4 for details). On the right side of Table 1, when examining the scores of the best system for each evaluator[6], we observe the same pattern.

[6] The MCQA column represents the scores for the gold-standard, GPT-4o, and GPT-4o. For Misinformation: OpenBioLLM, OpenBioLLM, and Llama3. For NLI: Llama3, the gold standard, and Llama3.

	Dataset Average			Best System Per Evaluator		
LM Evaluators	MMCQA	Misinfo	NLI	MMCQA	Misinfo	NLI
Baseline	36.00	44.56	61.12	41.18	48.30	61.50
Expert Trained	**72.83**	**58.67**	**62.61**	**82.91**	**61.22**	**67.62**
LLM Trained	70.85	39.74	58.02	78.90	49.43	61.12

Table 1: The left side shows the average dataset scores for *Primary Arguments* across three Proxy Tasks. While the right side displays the best system per evaluator for each LM evaluator. The highest score is marked in bold, and the second best is underlined.



(a) MMCQA

(b) Misinformation With Evidence

(c) NLI

Figure 4: Ranking of the *Primary Arguments* and *Control Cases* by the Proxy Task evaluators for each Proxy Task. Each row represents a distinct evaluator, and the columns represent the evaluated arguments. This table highlights the evaluators' ability to differentiate between proper and improper arguments.

4.2 Control Cases

We have already demonstrated that the LLM-trained evaluator aligns with human criteria when ranking LLM-generated arguments. Figure 4 presents an enhanced ranking that includes *Control Cases*, which serve as a form of adversarial attack. Ideally, a robust evaluator should rank all *Control Cases* in the lowest positions. (a) In the MMCQA task, both the Expert and LLM-trained evaluators prefer argumentations over *Control Cases*, while the baseline evaluator is misled by 3 out of 4 *Control Cases*. (b) For the misinformation task, all evaluators perform well, ranking argumentations first and *Control Cases* last. (c) In the NLI task, all models are misled by the *Control Cases*, with the LLM-trained evaluator proving to be the most resilient against control case attacks.

Note that, depending on the task, each control case behaves differently. In MMCQA and NLI, the label-only control case is the most effective attacker, while in the misinformation task, passage retrieval proves to be the strongest.

5 Discussion

This study offers critical insights into the effectiveness of discriminative LM evaluators in assessing LLM-generated medical arguments in the EBM context. We addressed five key questions, focusing on alignment with human judgment, robustness to adversarial inputs, and the value of Proxy Tasks. Using a discriminative LM evaluator and minimal hand-labeled data, we aimed to establish an evaluation framework closely aligned with human preferences. In the following paragraphs, we delve deeper into our results, systematically addressing each research question and discussing the implications of our findings for future research and practice.

RQ1 - Alignment with Human Preferences In our study, we demonstrate that even with a limited amount of hand-labeled data (utilizing only one argumentation per Proxy Task) we can effectively evaluate LLM-generated text while aligning closely with human judgment by training an LLM-trained discriminative LM-evaluator. By leveraging a discriminative model pre-trained on external corpora, we minimized evaluation biases. Although we did not conduct direct comparisons with LLM-as-a-judge approaches, our experiments show that our method avoids specific issues, such as verbosity bias. Furthermore, by employing a discriminative LM trained on diverse sources, our evaluators inherently avoid self-reinforcing biases, hallucinations, and other issues typical of generative models.

Our results revealed a slight discrepancy in rankings between human evaluators and automated evaluators for gold-standard arguments, which were ranked lower by humans. This highlights the limitations of metrics like BLEU and ROUGE, reliant on gold-standard references which human evaluators often rank lower. Consequently, we argue that such metrics can be misleading and misaligned with human preferences.

RQ2 - Robustness Against Adversarial Inputs This is the first study to test *Control Cases* for evaluating robustness to adversarial attacks. The baseline and expert-trained models were frequently misled by these adversarial inputs, unable to distinguish between meaningful arguments and misleading content. In contrast, the LLM-trained LM evaluator effectively filtered out non-argumentative or irrelevant inputs. This highlights the importance of adversarial robustness in evaluating automated systems, beyond simply optimizing Proxy Task scores.

RQ3 - Value of Proxy Tasks Proxy Tasks, validated across three datasets, provide a scalable alternative to extensive human argument ranking in different contexts. However, the nature of certain tasks affects the reliability of argument ranking if not correctly handled. For example, tasks involving ambiguous labels or evidence-deficient arguments, as seen in NLI and original misinformation tasks, make rankings challenging for both humans and models. Therefore, while some tasks offer a solid framework for assessing argument quality, like MMCQA and misinformation detection (under clear evidence conditions), not all datasets are equally valuable for evaluating LLM-generated arguments.

RQ4 - Consistency of Human Evaluations Inconsistent human evaluations across Proxy Tasks correlate with their suitability for automated evaluation. Using just two annotators and five examples, inter-tagger agreement (ITA) effectively indicated the viability of datasets for Proxy Tasks. Moreover, high ITA correlated with a better discriminative LM-evaluator performance in discarding *Control Cases* and aligning rankings with human judgment.

RQ5 - Score Performance and Ranking Correlation We prioritized relative rankings over numerical scores to better evaluate argument quality. Our findings challenge the assumption that expert-trained models, often treated as the upper bound of evaluation performance (Alonso et al., 2024; Yao et al., 2023), provide the best benchmarks. While these models achieve high scores, they fail to align with human preferences. In contrast, the LLM-trained evaluator, despite lower scores, aligned better with human judgment and was more robust to adversarial inputs, highlighting the limitations of score-centric evaluation.

6 Conclusions

In this work, we show that across three distinct Proxy Task scenarios, the automatic evaluation of medical explanatory arguments closely aligns with human judgment. Beyond standard MCQA tasks, we broaden our scope to include Misinformation Detection and NLI, providing a more comprehensive assessment. We present a novel approach that moves beyond traditional score maximization to prioritize improved ranking capabilities, addressing the inherent biases in LLMs when used as judges. Our LLM-trained evaluator aligns closely with human preferences and demonstrates resilience to adversarial attacks. Remarkably, only one hand-labeled example per task is needed to generate the synthetic arguments to develop the LLM-trained evaluator that best resembles human criteria. Additionally, we demonstrate that just five examples ranked by two human experts are enough to validate the chosen Proxy Task, confirming the practicality of our evaluation method.

7 Limitations

Our approach has the next limitations. First, the discriminative LM model used in this study has a token limit of 512, which may restrict the model's ability to fully process longer, more complex arguments. However, current advances in expanding language models' context size will mitigate this constraint. Finally, we do not focus explicitly on measuring hallucinations, factual accuracy, or coherence in the generated arguments.

Acknowledgments

This work has been partially supported by the HiTZ Center and the Basque Government (IXA excellence research group funding IT-1570-22 and IKER-GAITU project), as well as by the Spanish Ministry of Universities, Science and Innovation MCIN/AEI/10.13039/501100011033 by means of the projects: Proyectos de Generación de Conocimiento 2022 (EDHER-MED/EDHIA PID2022-136522OB-C22), DeepKnowledge (PID2021-127777OB-C21), DeepMinor (CNS2023-144375), and EU NextGeneration EU/PRTR (DeepR3 TED2021-130295B-C31). And also by an FPU grant (Formación de Profesorado Universitario) from the Spanish Ministry of Science and Innovation (MCINN) to the first author.

References

Iñigo Alonso, Maite Oronoz, and Rodrigo Agerri. 2024. Medexpqa: Multilingual benchmarking of large language models for medical question answering. *Artificial Intelligence in Medicine*, 155:102938.

Malaikannan Sankarasubbu Ankit Pal. 2024. Openbiollms: Advancing open-source large language models for healthcare and life sciences. https://huggingface.co/aaditya/OpenBioLLM-Llama3-70B.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, Wei Ye, Yue Zhang, Yi Chang, Philip S. Yu, Qiang Yang, and Xing Xie. 2023. A survey on evaluation of large language models. *Preprint*, arXiv:2307.03109.

Zeming Chen, Alejandro Hernández-Cano, Angelika Romanou, Antoine Bonnet, Kyle Matoba, Francesco Salvi, Matteo Pagliardini, Simin Fan, Andreas Köpf, Amirkeivan Mohtashami, Alexandre Sallinen, Alireza Sakhaeirad, Vinitra Swamy, Igor Krawczuk, Deniz Bayazit, Axel Marmet, Syrielle Montariol, Mary-Anne Hartley, Martin Jaggi, and Antoine Bosselut. 2023. Meditron-70b: Scaling medical pretraining for large language models. *Preprint*, arXiv:2311.16079.

Iker De la Iglesia, Aitziber Atutxa, Koldo Gojenola, and Ander Barrena. 2023. Eriberta: A bilingual pretrained language model for clinical natural language processing. *arXiv*, 2306.07373.

Pritam Deka, Anna Jurek-Loughrey, and P Deepak. 2022. Improved methods to aid unsupervised evidence-based fact checking for online health news. *Journal of Data Intelligence*, 3(4):474–504.

Milton Friedman. 1937. The use of ranks to avoid the assumption of normality implicit in the analysis of variance. *Journal of the american statistical association*, 32(200):675–701.

Iakes Goenaga, Aitziber Atutxa, Koldo Gojenola, Maite Oronoz, and Rodrigo Agerri. 2023. Explanatory argument extraction of correct answers in resident medical exams. *arXiv preprint arXiv:2312.00567*.

Robin Jia and Percy Liang. 2017. Adversarial examples for evaluating reading comprehension systems. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2021–2031, Copenhagen, Denmark. Association for Computational Linguistics.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.

Maël Jullien, Marco Valentino, Hannah Frost, Paul O'Regan, Dónal Landers, and André Freitas. 2023. NLI4CT: multi-evidence natural language inference for clinical trial reports. In *EMNLP*, pages 16745–16764. Association for Computational Linguistics.

Tom Kocmi and Christian Federmann. 2023. Large language models are state-of-the-art evaluators of translation quality. *arXiv preprint arXiv:2302.14520*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

Zhen Li, Xiaohan Xu, Tao Shen, Can Xu, Jia-Chen Gu, and Chongyang Tao. 2024. Leveraging large language models for nlg evaluation: A survey. *arXiv preprint arXiv:2401.07103*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Mingxin Liu, Tsuyoshi Okuhara, XinYi Chang, Ritsuko Shirabe, Yuriko Nishiie, Hiroko Okada, and Takahiro Kiuchi. 2024. Performance of chatgpt across different versions in medical licensing examinations worldwide: Systematic review and meta-analysis. *J Med Internet Res*, 26:e60807.

Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: NLG evaluation using gpt-4 with better human alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2511–2522, Singapore. Association for Computational Linguistics.

AI Meta. 2024. Introducing meta llama 3: The most capable openly available llm to date. *Meta AI*.

OpenAI. 2024. Hello gpt-4o. https://openai.com/index/hello-gpt-4o/.

Arjun Panickssery, Samuel R Bowman, and Shi Feng. 2024. Llm evaluators recognize and favor their own generations. *arXiv preprint arXiv:2404.13076*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing 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.

Alec Radford and Karthik Narasimhan. 2018. Improving language understanding by generative pre-training.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for MT evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702, Online. Association for Computational Linguistics.

David L Sackett, William M C Rosenberg, J A Muir Gray, R Brian Haynes, and W Scott Richardson. 1996. Evidence based medicine: what it is and what it isn't. *BMJ*, 312(7023):71–72.

Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2022. Colbertv2: Effective and efficient retrieval via lightweight late interaction. In *NAACL-HLT*, pages 3715–3734. Association for Computational Linguistics.

Thibault Sellam, Dipanjan Das, and Ankur P Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of ACL*.

Chenhui Shen, Liying Cheng, Xuan-Phi Nguyen, Yang You, and Lidong Bing. 2023. Large language models are not yet human-level evaluators for abstractive summarization. *arXiv preprint arXiv:2305.13091*.

Karan Singhal, Tao Tu, Juraj Gottweis, Rory Sayres, Ellery Wulczyn, Le Hou, Kevin Clark, Stephen Pfohl, Heather Cole-Lewis, Darlene Neal, et al. 2023. Towards expert-level medical question answering with large language models. *arXiv preprint arXiv:2305.09617*.

Eric Strong, Alicia DiGiammarino, Yingjie Weng, Andre Kumar, Poonam Hosamani, Jason Hom, and Jonathan H. Chen. 2023. Chatbot vs Medical Student Performance on Free-Response Clinical Reasoning Examinations. *JAMA Internal Medicine*, 183(9):1028–1030.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018. Bleu is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*.

Tianxiang Sun, Junliang He, Xipeng Qiu, and Xuan-Jing Huang. 2022. Bertscore is unfair: On social bias in language model-based metrics for text generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3726–3739.

Haochen Tan, Zhijiang Guo, Zhan Shi, Lu Xu, Zhili Liu, Yunlong Feng, Xiaoguang Li, Yasheng Wang, Lifeng Shang, Qun Liu, and Linqi Song. 2024. ProxyQA: An alternative framework for evaluating long-form text generation with large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6806–6827, Bangkok, Thailand. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Shikhar Vashishth, Denis Newman-Griffis, Rishabh Joshi, Ritam Dutt, and Carolyn P. Rosé. 2021. Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. *J. Biomed. Informatics*, 121:103880.

Juraj Vladika, Phillip Schneider, and Florian Matthes. 2024. HealthFC: Verifying health claims with evidence-based medical fact-checking. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8095–8107, Torino, Italy. ELRA and ICCL.

Bingsheng Yao, Prithviraj Sen, Lucian Popa, James Hendler, and Dakuo Wang. 2023. Are human explanations always helpful? towards objective evaluation of human natural language explanations. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14698–14713.

Tianyi Zhang*, Varsha Kishore*, Felix Wu*, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2024. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems*, 36.

A Experiment Results

			Primary Arguments				Control Cases			
			Gold	**GPT4**	**OpenBioLLM**	**Llama3**	**No Argument**	**Noise**	**Correct Label**	**IR**
Baseline		**Mean ± Std**	37.26 ± 4.57	34.73 ± 4.57	36.98 ± 4.18	35.01 ± 3.52	40.62 ± 3.78	41.18 ± 5.18	40.62 ± 2.55	35.01 ± 2.86
		run 1	*31.09*	*28.57*	*31.09*	*30.25*	*40.34*	*40.34*	*40.34*	*31.09*
		run 2	*42.02*	*36.13*	*39.50*	*36.13*	*45.38*	*47.90*	*43.87*	*36.13*
		run 3	*38.66*	*39.50*	*40.34*	*38.66*	*36.13*	*35.29*	*37.65*	*37.82*
Expert-Trained		**Mean ± Std**	77.59 ± 1.05	82.91 ± 0.79	64.99 ± 1.73	65.83 ± 1.58	37.82 ± 3.63	31.37 ± 1.98	39.67 ± 3.77	38.38 ± 0.40
		run 1	*78.99*	*82.35*	*63.03*	*64.71*	*32.77*	*28.57*	*35.63*	*38.66*
		run 2	*77.31*	*82.35*	*67.23*	*68.07*	*41.18*	*32.77*	*44.71*	*37.82*
		run 3	*76.47*	*84.03*	*64.71*	*64.71*	*39.50*	*32.77*	*38.66*	*38.66*
LLM-Trained		**Mean ± Std**	72.64 ± 2.27	78.90 ± 3.37	66.86 ± 2.10	64.99 ± 1.19	34.45 ± 3.69	33.80 ± 2.88	35.89 ± 2.37	38.39 ± 3.50
	Mixture 1	**Mean ± Std**	73.11 ± 3.14	77.87 ± 4.41	66.67 ± 1.05	64.71 ± 1.82	37.25 ± 0.40	35.85 ± 2.78	37.59 ± 0.94	38.94 ± 4.47
		run 1	*76.47*	*84.03*	*68.07*	*66.39*	*37.82*	*35.29*	*37.98*	*38.66*
		run 2	*68.91*	*75.63*	*66.39*	*65.55*	*36.97*	*39.50*	*38.49*	*44.54*
		run 3	*73.95*	*73.95*	*65.55*	*62.18*	*36.97*	*32.77*	*36.30*	*33.61*
	Mixture 2	**Mean ± Std**	71.99 ± 1.05	78.15 ± 2.38	66.95 ± 1.58	64.99 ± 0.40	35.01 ± 1.43	33.61 ± 2.47	36.19 ± 2.04	37.31 ± 2.65
		run 1	*72.27*	*81.51*	*68.07*	*65.55*	*34.45*	*31.09*	*33.45*	*33.61*
		run 2	*73.11*	*76.47*	*68.07*	*64.71*	*36.97*	*36.97*	*38.32*	*39.66*
		run 3	*70.59*	*76.47*	*64.71*	*64.71*	*33.61*	*32.77*	*36.81*	*38.66*
	Mixture 3	**Mean ± Std**	72.83 ± 1.43	80.67 ± 0.69	66.95 ± 2.86	65.27 ± 0.40	31.09 ± 3.82	31.93 ± 0.69	33.89 ± 1.74	38.94 ± 1.98
		run 1	*72.27*	*79.83*	*63.03*	*64.71*	*31.93*	*31.93*	*32.27*	*40.34*
		run 2	*74.79*	*81.51*	*68.07*	*65.55*	*35.29*	*32.77*	*36.30*	*40.34*
		run 3	*71.43*	*80.67*	*69.75*	*65.55*	*26.05*	*31.09*	*33.11*	*36.13*

Table 2: This table includes the accuracy obtained by the different automatic evaluators on selected tests in QA proxy task.

			Primary Arguments				Control Cases			
			Gold	**GPT4**	**OpenBioLLM**	**Llama3**	**No Argument**	**Noise**	**Correct Label**	**IR**
Baseline		**Mean ± Std**	40.82 ± 4.41	44.90 ± 1.67	48.30 ± 4.9	44.22 ± 5.36	35.37 ± 2.55	30.61 ± 1.67	39.46 ± 2.55	37.41 ± 9.48
		run 1	*34.69*	*42.86*	*42.86*	*36.73*	*38.78*	*32.65*	*38.78*	*24.49*
		run 2	*42.86*	*44.90*	*48.98*	*46.94*	*32.65*	*28.57*	*36.73*	*40.82*
		run 3	*44.90*	*46.94*	*53.06*	*48.98*	*34.69*	*30.61*	*42.86*	*46.94*
Expert-Trained		**Mean ± Std**	53.06 ± 1.67	59.86 ± 3.85	61.22 ± 3.33	60.54 ± 2.55	46.26 ± 5.36	13.61 ± 0.96	0.68 ± 0.96	56.46 ± 0.96
		run 1	*55.10*	*65.31*	*65.31*	*61.22*	*38.78*	*14.29*	*0.00*	*57.14*
		run 2	*51.02*	*57.14*	*57.14*	*57.14*	*48.98*	*12.24*	*0.00*	*55.10*
		run 3	*53.06*	*57.14*	*61.22*	*63.27*	*51.02*	*14.29*	*2.04*	*57.14*
LLM-Trained		**Mean ± Std**	25.85 ± 4.79	42.86 ± 6.04	40.82 ± 5.68	49.43 ± 4.42	17.01 ± 10.36	11.56 ± 5.59	5.22 ± 4.80	38.10 ± 8.35
	Mixture 1	**Mean ± Std**	23.81 ± 6.94	38.78 ± 2.89	44.90 ± 2.89	46.94 ± 3.33	13.61 ± 8.55	11.56 ± 5.36	0.00 ± 0.00	43.54 ± 6.31
		run 1	*14.29*	*36.73*	*42.86*	*51.02*	*2.04*	*4.08*	*0.00*	*34.69*
		run 2	*30.61*	*36.73*	*42.86*	*46.94*	*16.33*	*16.33*	*0.00*	*46.94*
		run 3	*26.53*	*42.86*	*48.98*	*42.86*	*22.45*	*14.29*	*0.00*	*48.98*
	Mixture 2	**Mean ± Std**	25.85 ± 1.92	45.58 ± 2.55	36.05 ± 3.47	47.62 ± 0.96	22.45 ± 10.41	6.80 ± 0.96	8.84 ± 0.96	29.25 ± 3.47
		run 1	*28.57*	*48.98*	*40.82*	*46.94*	*36.73*	*6.12*	*8.16*	*30.61*
		run 2	*24.49*	*44.90*	*32.65*	*46.94*	*12.24*	*8.16*	*10.20*	*24.49*
		run 3	*24.49*	*42.86*	*34.69*	*48.98*	*18.37*	*6.12*	*8.16*	*32.65*
	Mixture 3	**Mean ± Std**	27.89 ± 0.96	44.22 ± 7.51	41.50 ± 5.09	53.74 ± 3.47	14.97 ± 7.70	16.33 ± 2.89	6.80 ± 4.19	41.50 ± 3.85
		run 1	*26.53*	*44.90*	*42.86*	*57.14*	*4.08*	*12.24*	*2.04*	*46.94*
		run 2	*28.57*	*53.06*	*46.94*	*55.10*	*20.41*	*18.37*	*6.12*	*38.78*
		run 3	*28.57*	*34.69*	*34.69*	*48.98*	*20.41*	*18.37*	*12.24*	*38.78*

Table 3: This table includes the accuracy obtained by the different automatic evaluators on selected tests in Misinformation Detection proxy task.

			Primary Arguments				Control Cases		
			Gold	**GPT4**	**OpenBioLLM**	**Llama3**	**No Argument**	**Noise**	**Correct Label**
Baseline		**Mean ± Std**	60.47 ± 4.10	60.39 ± 3.68	60.63 ± 0.73	61.50 ± 3.96	60.32 ± 5.02	60.80 ± 3.45	62.33 ± 1.97
		run 1	*54.75*	*55.53*	*59.60*	*55.91*	*53.38*	*56.09*	*59.62*
		run 2	*64.20*	*64.43*	*61.16*	*64.12*	*65.07*	*64.24*	*63.12*
		run 3	*62.45*	*61.20*	*61.14*	*64.48*	*62.50*	*62.07*	*64.25*
Expert-Trained		**Mean ± Std**	67.62 ± 1.68	61.89 ± 1.33	61.36 ± 0.48	63.06 ± 0.81	65.10 ± 1.22	60.62 ± 2.08	64.58 ± 1.55
		run 1	*69.96*	*63.76*	*62.03*	*64.20*	*66.83*	*58.37*	*64.23*
		run 2	*66.09*	*61.14*	*61.07*	*62.46*	*64.20*	*60.10*	*62.89*
		run 3	*66.81*	*60.77*	*60.97*	*62.52*	*64.28*	*63.38*	*66.63*
LLM-Trained		**Mean ± Std**	57.34 ± 2.66	58.22 ± 2.04	54.73 ± 2.86	60.58 ± 2.48	54.33 ± 3.02	56.15 ± 3.29	61.12 ± 2.51
	Mixture 1	**Mean ± Std**	55.95 ± 1.90	60.44 ± 1.05	53.18 ± 1.67	62.68 ± 1.77	53.50 ± 2.26	55.91 ± 2.59	60.46 ± 0.68
		run 1	*57.66*	*61.81*	*55.54*	*63.34*	*52.87*	*59.10*	*59.93*
		run 2	*53.30*	*60.26*	*51.98*	*64.43*	*51.09*	*55.89*	*61.43*
		run 3	*56.90*	*59.26*	*52.03*	*60.26*	*56.53*	*52.75*	*60.03*
	Mixture 2	**Mean ± Std**	59.04 ± 2.88	56.47 ± 1.29	56.99 ± 1.42	59.62 ± 0.66	54.81 ± 2.53	56.15 ± 0.68	59.62 ± 1.91
		run 1	*55.12*	*54.96*	*55.12*	*58.71*	*51.23*	*55.19*	*61.73*
		run 2	*60.03*	*56.33*	*57.29*	*60.27*	*56.53*	*56.67*	*60.03*
		run 3	*61.97*	*58.12*	*58.55*	*59.87*	*56.66*	*56.59*	*57.11*
	Mixture 3	**Mean ± Std**	57.03 ± 1.42	57.74 ± 0.37	54.02 ± 3.01	59.46 ± 2.51	54.68 ± 3.43	56.39 ± 4.65	63.27 ± 2.32
		run 1	*56.49*	*57.23*	*57.52*	*55.94*	*49.84*	*49.83*	*60.37*
		run 2	*55.63*	*57.92*	*50.17*	*61.61*	*56.93*	*59.96*	*66.05*
		run 3	*58.97*	*58.08*	*54.38*	*60.82*	*57.28*	*59.39*	*63.40*

Table 4: This table includes the micro F-score obtained by the different automatic evaluators on selected tests in NLI proxy task.

B Automatic Evaluator's Inpunts

EVALUATORS	**INPUTS**		
	QA	**Missinformation Detection**	**NLI**
Naive Evaluator	Question Clinical Case Possible Answers Correct Answer	Question Label	Statement Full Section Label
Clinician Lined Up Evaluator	Question Clinical Case Possible Answers **Gold Argumentation** Correct Answer	Question **Gold Argumentation** Label	Statement **Gold Evidences** Label
LLMs Lined Up Evaluators	Question Clinical Case Possible Answers **LLMs Argumentation** Correct Answer	Question **LLMs Argumentation** Label	Statement **LLMs Evidences** Label

Table 5: This table includes the inputs used for each automatic evaluator depending on the proxy task.

C Instruction Tuning Example

	Instruction Tuning Example Used For QA
Instruction	<s>[INST] A 52-year-old man with no concomitant diseases comes to the emergency department for melena of 24 hours evolution without hemodynamic repercussions. He denies taking nonsteroidal anti-inflammatory drugs. The hematocrit is 33% and the rest of the laboratory tests are normal. The upper gastrointestinal endoscopy performed urgently 6 hours after admission shows a normal stomach, without blood or hematic debris and an excavated ulcer of 8 mm in diameter in the anterior face of the duodenal bulb with "visible vessel" at its base and without active bleeding. Which of the following statements is true? The possible answers are: 1- In the initial endoscopy it is indicated to apply an endoscopic therapy and subsequently to establish endovenous treatment with high doses of a proton pump inhibitor. This strategy has been shown to reduce the risk of hemorrhagic recurrence and mortality. 2- In the initial endoscopy, given the absence of active bleeding, endoscopic therapy is not indicated. Subsequently, to reduce the risk of hemorrhagic recurrence, intravenous treatment with high doses of a proton pump inhibitor should be started. 3- n initial endoscopy, endoscopic therapy is indicated. Subsequent treatment with high doses of a proton pump inhibitor has not been shown to be of any additional benefit. 4- Since this is a complicated ulcer (hemorrhage) the best therapeutic option, once the hemorrhagic episode is resolved, is a vagotomy and pyloroplasty. [/INST]
Output	This is a grade IIa ulcer (Forrest classification), with a high risk of recurrence. Therefore, endoscopic treatment and hospitalization with intravenous treatment with PPIs (it is usually omeprazole) for at least 72 hours is clearly indicated. </s>

Table 6: This is an example of an instance used to fine-tune Llama3 for QA proxy task applying IT.

D Benchmarks Examples

	Example of an instance from NLI Benchmark
Statement	There were 7 more cases of Anaemia and 1 more case of Disseminated intravascular coagulation in cohort 1 of the primary trial compared to cohort 2.
Gold Evidences	Adverse Events 1: ** Total: 158/482 (32.78%) ** Anaemia 7/482 (1.45%) ** Disseminated intravascular coagulation 1/482 (0.21%) ** Adverse Events 2: ** Total: 37/238 (15.55%) ** Anaemia 2/238 (0.84%) ** Disseminated intravascular coagulation 0/238 (0.00%)
Full Document	INTERVENTION 1: ** Everolimus + Exemestane ** Everolimus 10 mg daily in combination with exemestane 25 mg daily ** INTERVENTION 2: ** Placebo + Exemestane ** Placebo of everolimus in combination with exemestane 25 mg daily ** Inclusion Criteria: ** Adult women (18 years of age) with metastatic or locally advanced breast cancer not amenable to curative treatment by surgery or radiotherapy. ** Histological or cytological confirmation of estrogen-receptor positive (ER+) breast cancer ** Postmenopausal women. ** Disease refractory to non steroidal aromatase inhibitors (NSAI), ** Radiological or clinical evidence of recurrence or progression on or after the last systemic therapy prior to randomization. ** Patients must have at least one lesion that can be accurately measured or bone lesions in the absence of measurable disease as defined above. ** Exclusion Criteria: ** HER2-overexpressing patients ** Patients with only non-measurable lesions other than bone metastasis (e.g. pleural effusion, ascites etc.). ** Patients who received more than one chemotherapy line for Advanced Breast Cancer. ** Previous treatment with exemestane or mTOR inhibitors. ** Known hypersensitivity to mTOR inhibitors, e.g. sirolimus (rapamycin). ** Radiotherapy within four weeks prior to randomization ** Currently receiving hormone replacement therapy, ** Other protocol-defined inclusion/exclusion criteria may apply ** Outcome Measurement: ** Progression-free Survival (PFS) Based on Local Radiology Review of Tumor Assessments. ** Progression-free survival, the primary endpoint in this study, is defined as the time from the date of randomization to the date of first documented radiological progression or death due to any cause. Disease progression was based on the tumor assessment by the local radiologist or investigator using RECIST 1.0 criteria. If a patient did not progress or known to have died at the date of the analysis cut-off or start of another antineoplastic therapy, the PFS date was censored to the date of last adequate tumor assessment prior to cut-off date or start of antineoplastic therapy. For patients with lytic or mixed (lytic+sclerotic) bone lesions, the following is considered progression: appearance of 1 new lytic lesions in bone; the appearance of new lesions outside of bone and unequivocal progression of existing bone lesions. ** Time frame: date of randomization to the date of first documented tumor progression or death from any cause, whichever occurs first, reported between day of first patient randomized up to about 19 months ** Results 1: ** Arm/Group Title: Everolimus + Exemestane ** Arm/Group Description: Everolimus 10 mg daily in combination with exemestane 25 mg daily ** Overall Number of Participants Analyzed: 485 ** Median (95% Confidence Interval) ** Unit of Measure: months 6.93 (6.44 to 8.05) ** Results 2: ** Arm/Group Title: Placebo + Exemestane ** Arm/Group Description: Placebo of everolimus in combination with exemestane 25 mg daily ** Overall Number of Participants Analyzed: 239 ** Median (95% Confidence Interval) ** Unit of Measure: months 2.83 (2.76 to 4.14) ** Adverse Events 1: ** Total: 158/482 (32.78%) ** Anaemia 7/482 (1.45%) ** Disseminated intravascular coagulation 1/482 (0.21%) ** Lymphadenopathy 0/482 (0.00%) ** Neutropenia 0/482 (0.00%) ** Thrombocytopenia 2/482 (0.41%) ** Anaemia 28/482 (1.66%) ** Disseminated intravascular coagulation 21/482 (0.21%) ** Febrile neutropenia 21/482 (0.21%) ** Lymphadenopathy 20/482 (0.00%) ** Neutropenia 20/482 (0.00%) ** Adverse Events 2: ** Total: 37/238 (15.55%) ** Anaemia 2/238 (0.84%) ** Disseminated intravascular coagulation 0/238 (0.00%) ** Lymphadenopathy 1/238 (0.42%) ** Neutropenia 1/238 (0.42%) ** Thrombocytopenia 0/238 (0.00%) ** Anaemia 22/238 (0.84%) ** Disseminated intravascular coagulation 20/238 (0.00%) ** Febrile neutropenia 21/238 (0.42%) ** Lymphadenopathy 21/238 (0.42%) ** Neutropenia 21/238 (0.42%)
Full Section	Adverse Events 1: ** Total: 158/482 (32.78%) ** Anaemia 7/482 (1.45%) ** Disseminated intravascular coagulation 1/482 (0.21%) ** Lymphadenopathy 0/482 (0.00%) ** Neutropenia 0/482 (0.00%) ** Thrombocytopenia 2/482 (0.41%) ** Anaemia 28/482 (1.66%) ** Disseminated intravascular coagulation 21/482 (0.21%) ** Febrile neutropenia 21/482 (0.21%) ** Lymphadenopathy 20/482 (0.00%) ** Neutropenia 20/482 (0.00%) ** Adverse Events 2: ** Total: 37/238 (15.55%) ** Anaemia 2/238 (0.84%) ** Disseminated intravascular coagulation 0/238 (0.00%) ** Lymphadenopathy 1/238 (0.42%) ** Neutropenia 1/238 (0.42%) ** Thrombocytopenia 0/238 (0.00%) ** Anaemia 22/238 (0.84%) ** Disseminated intravascular coagulation 20/238 (0.00%) ** Febrile neutropenia 21/238 (0.42%) ** Lymphadenopathy 21/238 (0.42%) ** Neutropenia 21/238 (0.42%)
Label	Entailment

Table 7: An instance example from NLI Benchmark.

	Example of an instance from Missinformation Detection Benchmark
Question	Can filtering out blue light using blue filter glasses or night mode settings on smartphone, tablet or laptop screens have a beneficial effect on sleep?
Gold Argumentation	In previous studies, it makes no noticeable difference to sleep when the blue light component of display screen devices is filtered out in the evening. However, the results are not well validated because the studies are of low quality and usually only examined a few people.
Label	Refuted

Table 8: An instance example from Missinformation Detection Benchmark.

Example of a document from the Preprocessed Antidote CasiMedicos Dataset	
C	A 45-year-old man undergoes a truncal vagotomy and antrectomy with Billroth II reconstruction for chronic peptic ulcer disease with pyloro-duodenal stricture. Six weeks after the surgery she reports that shortly after (less than half an hour) after ingestions she presents nausea, asthenia and sweating, dizziness and abdominal cramps usually accompanied by diarrhea.
Q	Which of the following is the most appropriate approach for her initial management?
P	**(1)** Apply treatment with a somatostatin inhibitor (octreotide). **(2)** Follow specific dietary measures. **(3)** Trial treatment with a benzodiazepine. **(4)** Search for a probable neuroendocrine tumor (e.g. carcinoid). **(5)** Indicate surgical treatment to perform an antiperistaltic Roux-en-Y gastrojejunostomy.
E	Answers 1, 2 and 5 are appropriate treatments for dumping syndrome or postgastrectomy, but the question is focused on initial management, so the most appropriate answer seems to be 2.
NE	Applying treatment with a somatostatin inhibitor (octreotide), following specific dietary measures and indicating surgical treatment to perform an antiperistaltic Roux-en-Y gastrojejunostomy are appropriate treatments for dumping syndrome or postgastrectomy, but the question is focused on initial management, so the most appropriate approach seems to be following specific dietary measures.

Table 9: Example of a document in the Preprocessed Antidote CasiMedicos dataset with the explanation about the correct answer manually neutralized. **C**: Clinical Case; **Q**: Question; **P**: Possible Answers; **E**: Correct Answer Explanation. The *Clinical Case*, *Question*, *Possible Answers*, *Correct Answer Explanation* sections are the original annotations of the Antidote CasiMedicos dataset. The preprocessing of the medical doctors' explanations (**NE**) is part of this work.

E Prompts For Medical Argumentation Generation

	Prompt used to generate medical argumentation for QA
"role": "system", "content":	You are a medical student and given a medical case, a question and five possible answers, tell me which is the correct answer and argument in favor of it. Example: A medical case and a question related to it <casequestion> After a traffic accident a 38-year-old patient is admitted to the ICU in coma. After several days the patient does not improve neurologically and a CT scan shows hemorrhagic punctate lesions in the corpus callosum and cortico-subcortical junction. What is the diagnosis? <\casequestion> And five possible answers: <ans>1- Acute subdural hematoma.<\ans> <ans>2- Trobocytopenic purpura.<\ans> <ans>3- Cerebral hemorrhagic contusion.<\ans> <ans>4- Severe diffuse axonal injury.<\ans> <ans>5- Acute heart attack.<\ans> The argument for the correct answer without mentioning the options and focusing exclusively on the arguments is: Diffuse axonal injury produces an early and sustained deterioration of the level of consciousness (as mentioned in the case statement) without a lesion on CT scan to justify the picture. Sometimes, punctate hemorrhages at the level of the corpus callosum, corticosubcortical junction and dorsolateral portion of the brainstem are evidenced in this imaging test.
"role": "user", "content":	Given this new case and the question related to it: <casequestion> {case_question} <\casequestion> And five possible answers: <ans> {ans1} <\ans> <ans> {ans2} <\ans> <ans> {ans3} <\ans> <ans> {ans4} <\ans> <ans> {ans5} <\ans> The argument for the correct answer without mentioning the options and focusing exclusively on the arguments is:

Table 10: This is the prompt we used to generate medical argumentation for QA, where {case_question} is a new clinical case and a question related to it from the dataset, and {ans1-5} are the possible answer options for the question. The same prompt has been used on GPT-4o, OpenBioLLM and Llama3.

	Prompt used to generate medical argumentation for Missinformation Detection
"role": "system", "content":	You are a medical student. Given a medical question, you must answer the question and include the arguments you use to reach your answer. Example: A question <question> Can taking the enzyme diamino oxidase prevent alcohol-related hangover symptoms? <\question> The argument for the correct answer and focusing exclusively on the arguments is: Such an effect is not likely, nor do clinical studies exist on this issue.
"role": "user", "content":	Given this new question: <question> {question} <\question> The argument for the correct answer and focusing exclusively on the arguments is:

Table 11: This is the prompt we used to generate medical argumentation for Missinformation Detection, where {question} is a new question from the dataset. The same prompt has been used on GPT-4o, OpenBioLLM and Llama3.

	Prompt used to extract medical argumentation for NLI
"role": "system", "content":	You are a medical student. Given a medical hypothesis and evidences separated by **, extract the evidences that supports or contradicts the hypothesis without adding any other words. Remember, do not generate any new text. Extract only the relevant parts exactly as they appear in the given text. Example: A hypothesis <hypothesis> Patients with significantly elevated ejection fraction are excluded from the primary trial, but can still be eligible for the secondary trial if they are 55 years of age or over. <\hypothesis> A list of possible evidences <evidences> Inclusion criteria: ** Inclusion Criteria: ** Female patients age 18 years or older ** Histologically proven breast cancer after failure or relapse of no more than three lines of chemotherapy including adjuvant, irrespective of prior hormone therapy metastatic disease (stage IV); ** HER2-negative patients (HER2 1+ or negative, or HER2 2+ and FISH negative) ** At least one measurable tumour lesion (RECIST); ** Exclusion criteria: ** Exclusion Criteria: ** Active infectious disease ** Gastrointestinal disorders that may interfere with the absorption of the study drug or chronic diarrhoea ** Serious illness, concomitant non-oncological disease or mental problems considered by the investigator to be incompatible with the protocol ** Active/symptomatic brain metastases ** Cardiac left ventricular function with resting ejection fraction < 50% (below upper limit of normal) ** ANC less than 1500/mm3 platelet count less than 100 000/mm3 ** Bilirubin greater than 1.5 mg /dl (>26 and 61549 mol /L, SI unit equivalent) ** AST and ALT greater than 2.5 times the upper limit of normal or greater 5 times the upper limit of normal in case of known liver metastases ** Serum creatinine greater than 1.5 mg/dl (>132 and 61549 mol/L, SI unit equivalent) ** Patients who are sexually active and unwilling to use a medically acceptable method of contraception ** Pregnancy or breast-feeding ** Concomitant treatment with other investigational drugs or other anti-cancer-therapy during this study and/or during the past two/four weeks, prior to the first treatment with the trial drug. Concurrent treatment with biphosphonates is allowed ** Previous treatment with trastuzumab, EGFR-, or EGFR/HER2-inhibitors patients unable to comply with the protocol ** Active alcohol or drug abuse ** Other malignancy within the past 5 years' 'Premenopausal women 55 years of age or younger with regular menstrual cycles (at least four cycles in the last six months). Women with fewer than 4 menses in the last 6 months or who have had a hysterectomy with ovaries intact will be considered premenopausal if FSH level < 20. ** Women with breast density 25% (scattered fibroglandular densities or greater) are eligible. ** Prior Treatment ** Patients who are currently receiving hormone replacement therapy (estrogen or progesterone); or are taking tamoxifen or raloxifene are not eligible. Women who have taken these medications must have stopped for at least 4 months prior to study entry. ** Topical estrogen (eg, transdermal patches and vaginal estrogens) is allowed. ** Patients with a diagnosis of osteoporosis with physician recommendation for treatment of low bone mass are not eligible. ** Patients known to have hyperparathyroid disease or other serious disturbances of calcium metabolism requiring intervention in the past 5 years are not eligible. ** Patients with a history of kidney stones (unless documented not to have been a calcium stone) are not eligible. ** Patients participating in a concurrent breast cancer chemoprevention trial are not eligible. ** Required initial laboratory values - Calcium < 10.5 mg/dL' <\evidences> The evidences that supports or contradicts the hypothesis without adding any other words are: Cardiac left ventricular function with resting ejection fraction < 50% (below upper limit of normal). ** Premenopausal women 55 years of age or younger with regular menstrual cycles (at least four cycles in the last six months).
"role": "user", "content":	Given this new hypothesis: <hypothesis> {statement} <\hypothesis> And given this new list of possible evidences <evidences> {evidences} <\evidences> The evidences that supports or contradicts the hypothesis without adding any other words are:

Table 12: This is the prompt we used to extract medical argumentation for NLI, where {statement} is a new hypothesis from the dataset and {evidences} is a new list of evidences from the dataset related to the hypothesis. The same prompt has been used on GPT-4o, OpenBioLLM and Llama3.