Judging LLM-as-a-Judge: Concerning Rubric Artifacts in LLM-based Automated Text Generation Evaluation

Anonymous ACL submission

Abstract

LLM-as-a-Judge pipelines are increasingly used to evaluate AI-generated text, based on the assumption that judgments arise from reasoning over candidate responses with respect to a rubric. We show that this assumption warrants further scrutiny. Classifiers trained only on rubric text, without access to any evaluated response, achieve nontrivial predictive performance on judge outputs. This suggests that rubric formulations encode recoverable evaluative signals, allowing scores to be partially anticipated independently of model outputs. We further find that this signal is not explained by superficial artifacts. The geometry of the rubrics' embedding indicates that the signal is distributed and semantically structured. Our findings raise concerns about the validity of rubric-based LLM evaluation and highlight the need for further methodological study into automated evaluation via LLMs.

1 Introduction

LLM-based evaluators are now widely used to benchmark and validate generative systems, despite growing evidence of systematic biases such as position effects, verbosity preferences, and sensitivity to prompt framing (Shi et al., 2025; Chen et al., 2024; Li et al., 2026; Gu et al., 2024; Tan et al., 2025). These issues raise a broader concern: whether current evaluation pipelines reliably measure response quality, or instead reflect artifacts of the evaluation setup.

Building on critiques of LLMs as judges, there are many evaluation settings, including direct scoring, pairwise comparison, and reference-based scoring, which vary in the structure of inputs and the guidance provided (Gu et al., 2024). Among these, rubric-based evaluation is widely adopted, especially in high-stakes domains, because it provides explicit criteria intended to standardize judgments and improve transparency (Croxford et al., 2025). This reliance rests on the assumption that rubrics constrain models without introducing an independent evaluative signal. We test this assumption by proposing a novel controlled experiment (Figure 1) that asks whether an evaluative signal can be recovered from the rubric alone, without access to the candidate response or its context. This formulation isolates the rubric's contribution and removes systemic biases, such as position effects, that arise in response-based evaluation.

Using HEALTHBENCH (Arora et al., 2025), we find that rubric text alone predicts judge outputs, suggesting a previously underexamined coupling between rubric formulation and evaluation outcomes. Our embedding-geometry analysis further suggests that rubric text conveys an evaluative signal. This motivates closer scrutiny of the rubric design in LLM-based evaluation.

2 Related Work

Prior work on LLM-as-a-Judge has shown that automated evaluation is sensitive to position effects, scoring bias, prompt framing, and other non-semantic factors, raising concerns about the reliability of model-based judgments (Zheng et al., 2023; Shi et al., 2025; Chen et al., 2024; Li et al., 2026; Tan et al., 2025; Gu et al., 2024). In benchmark settings such as HealthBench, where expert-authored rubrics define the evaluation criterion for medical responses, the rubric is typically treated as an instruction rather than a potential source of signal (Arora et al., 2025). Our work differs by probing whether the rubric text itself encodes recoverable evaluative priors, independent of the candidate response. This connects to shortcut-learning research showing that language models often exploit superficial lexical or structural cues rather than intended semantic signal (Zhou et al., 2024; Ong Ly et al., 2024).

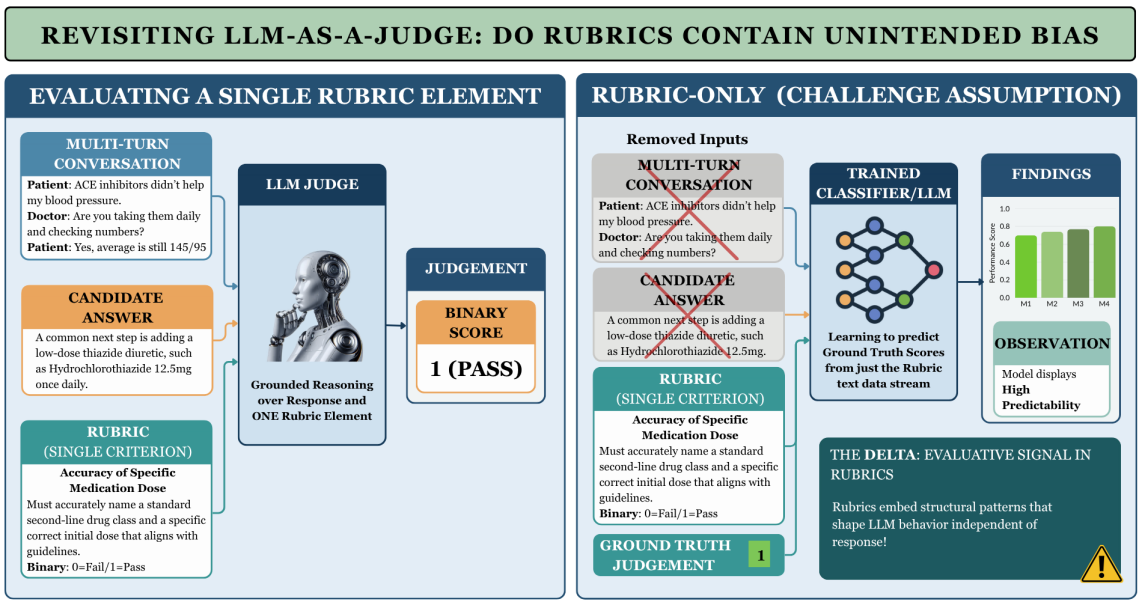

Figure 1: Rubric Probe: Given a rubric alone, we train a classifier to predict the judge label, testing whether rubric text carries recoverable evaluative signal beyond chance.

3 LLM-as-a-Judge Probe Design

A central assumption in rubric-based evaluation is that judgments arise from grounded reasoning over the candidate response with respect to the rubric, which is treated as an evaluative instruction. We instead investigate whether rubric formulations encode latent priors correlated with evaluation labels, such that a model can partially infer the expected judgment from the rubric text alone (Figure 2).

Formulation

Let r denote a rubric, z the input context, x a candidate response, and y the judge label. LLM-as-a-Judge aims to model

$$p(y \mid r, z, x),$$

where the rubric conditions how responses are evaluated. For evaluation to be valid, the rubric should guide judgment without independently determining the label. In particular, y should not be predictable from r alone:

$$p(y \mid r) \approx 0.5.$$

We empirically test whether

$$p(y \mid r) > 0.5,$$

which indicates that the rubric text contains a standalone predictive signal about y that is independent of the response x and the context z.

Beyond predictive performance, we analyze the structure of the rubric signal using generalizability studies and the geometry of the embedding space.

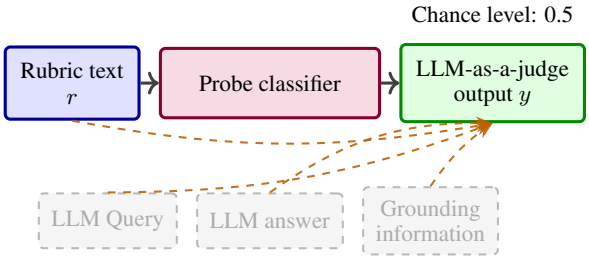

Figure 2: Probe overview. The classifier receives only rubric text and predicts the LLM-as-a-judge output, while the evaluated answer, contextual metadata, and grounding information are excluded.

4 LLM-as-a-Judge Probe Experiment

We start by describing the details of our experiment and setup, followed by the probe results.

4.1 Experiment Details

We evaluate our probe on HEALTHBENCH and its variants, HEALTHBENCH-EVAL and HEALTHBENCH-HARD, which provide conversation-specific rubrics for judging open-ended healthcare responses.

Dataset Source: We chose HEALTHBENCH because it is a comprehensive healthcare dataset, comprising complex healthcare-related conversations, with a unique feature: each conversation has a case-specific rubric rather than an overall rubric. More importantly, the rubrics are provided by clinicians and are NOT generated by an LLM. These fea-

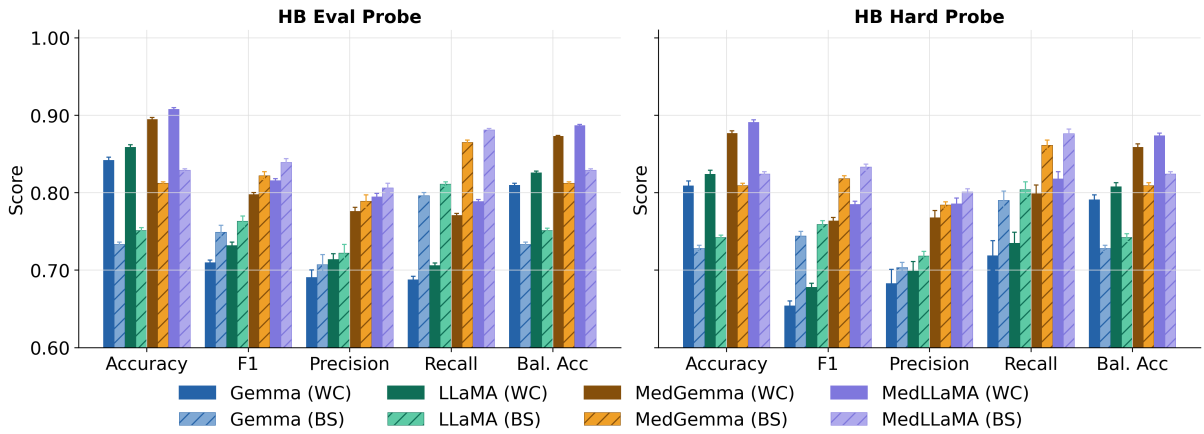

Figure 3: Per-metric comparison across models and evaluation splits. BERT classifier trained on LLM judge labels with rubric input. Error bars represent std. dev. The weighted classification (WC) and balanced subsampling (BS) cohorts are presented here to address class imbalance.

tures, which follow best-practice benchmarking, made HEALTHBENCH a strong candidate for the analysis. In the traditional setting, an LLM would evaluate the conversation and provide an answer. This answer would be weighed against the rubric and assigned satisfiability (0/1) by the LLM-as-a-Judge. We have 5k such conversations paired with an expected answer and case-specific rubrics.

Probing Dataset: We first run the candidate LLM model and obtain the responses across various models. We then apply the LLM-as-a-Judge model and obtain the class labels (0/1). We use Qwen (Qwen2.5-7B-Instruct (Qwen et al., 2025)) as an LLM-as-a-Judge on LLM outputs from LlaMA (meta-llama/Llama-3.1-8B-Instruct (Grattafiori et al., 2024)), MedLlaMA (MMed-Llama-3-8B (Qiu et al., 2024)), Gemma (google/gemma-7b-it (Gemma et al., 2024)), and MedGemma (google/medgemma-1.5-4b-it (Sellergren et al., 2026)). Then, we construct four classifier datasets, using the rubric as input and the LLM-as-a-Judge output as the label for each model. HEALTHBENCH-EVAL-PROBE comprises approximately ≈43k rubrics, while HEALTHBENCH-HARD-PROBE contains around ≈11k rubrics.The final experiment evaluates eight cohorts across these two datasets. Each classifier is trained using an 80/20 train–test split.

Classifier Setup: We use the PubMedBERT (Gu et al., 2021) based classifier for our main experiments. For generalization, we test cross-dataset transfer between HEALTHBENCH-EVAL-PROBE and HEALTHBENCH-HARD-PROBE. [1]

[1]Code available here: Repository Link

4.2 Rubric Probe Results

In this study, we evaluate the PubMedBERT-based classifier's performance against the four cohorts (Gemma, MedGemma, LlaMA, MedLLaMa), across the HEALTHBENCH-EVAL-PROBE and HEALTHBENCH-HARD-PROBE using standard metrics while explicitly accounting for class imbalance through weighted classification and balanced subsampling strategies (Figure 3). We observe that the classifier outputs are **> 0.5 and, in some cases, exceeding 0.8**, even after accounting for dataset imbalance. These results indicate that the rubric text contains a structured signal that is predictive of judge labels beyond random chance. Additional analyses on these classifier experiments are included in the Appendix A.

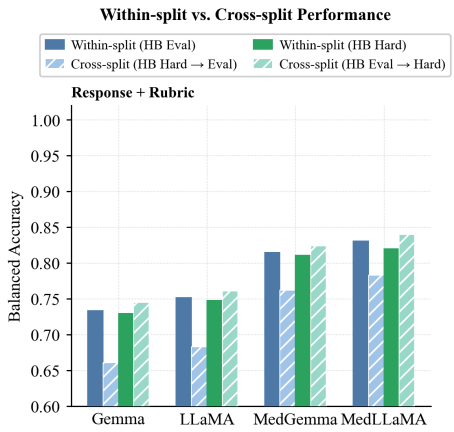

Figure 4: Cross-dataset transfer: within-split vs. cross-split balanced accuracy. Solid bars = within-split (train = test). Hatched bars = cross-split (train on one split, test on the other) Lower gap = better generalization.

4.3 Generalization Checks

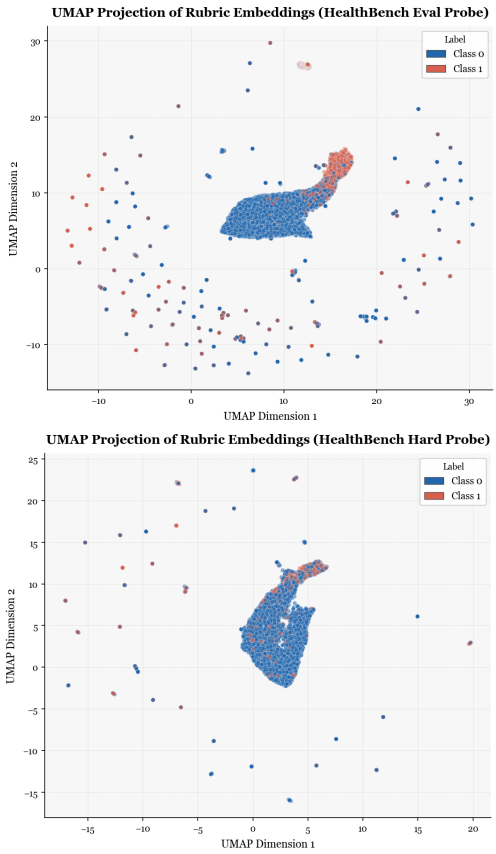

Figure 5: The UMAP projections of HEALTHBENCH-EVAL-PROBE and HEALTHBENCH-HARD-PROBE

To ensure we are not drawing conclusions from a single run of the experiments, we have performed several sanity checks to assess generalizability. First, we performed the basic cross-validation check. Second, we conducted an experiment in which we trained on one subset, say HEALTHBENCH-EVAL-PROBE, and tested on the other. We analyze the effect of such a scheme on our metrics (Figure 4). Finally, we have detailed reports on experiments with the LLM-as-a-judge, answering the rubric alone, as well as studies using simple classifiers, such as logistic regression on TF-IDF, in Appendix B. Notably, performance remains above the 0.5 chance level even under this harder cross-split setting, suggesting that the model learns transferable signals rather than relying purely on chance.

4.4 Semantic Analysis of Rubrics

We next probe whether the rubrics themselves, independent of the LLM-as-a-Judge output, encode semantic cues that may drive the observed label structure. With UMAP (McInnes et al., 2018), we visualize BERT embeddings to test whether Eval and Hard rubrics form separable clusters (Figure 5). Additional semantic probes are included in Appendix C.

The correct label is defined by the rubric, the context, and the candidate response, and should not be recoverable from the rubric alone. Yet we observe consistent correlations between rubric text and judge labels even without the response. Moreover, HEALTHBENCH-EVAL rubrics are more linearly separable with respect to judge labels, suggesting that some rubric formulations are systematically associated with particular outcomes. Together, these findings indicate that rubric text encodes label-correlated signal, raising concerns that evaluation may reflect rubric properties independently of the response.

4.5 Discussion

Our probe shows that rubric text alone predicts LLM-as-a-Judge outputs above chance, demonstrating the presence of a recoverable rubric-conditioned signal independent of the candidate response. This signal is distributed across semantic and topical structure rather than being reducible to keyword artifacts. These patterns indicate that evaluation difficulty is not solely driven by the task, but also by variability and ambiguity in rubric formulation. This is not a model capacity limitation: scaling to larger or more capable judges does not remove the signal already encoded in the rubric. Similarly, our findings are unlikely to be explained by dataset-specific confounding alone, since the probe never observes the response or context, and the effect persists across HEALTHBENCH-PROBE splits. While we do not establish causality, we identify a consistent association between rubric text and judge outputs. This motivates systematic validation of rubric design, including stress-testing rubric sensitivity, auditing judge rationales (DeLucia et al., 2026), and complementing rubric-based evaluation with more robust metrics (Shailya et al., 2025; Ito et al., 2025; Croxford et al., 2025).

5 Conclusion

Rubric-conditioned evaluation may inadvertently encode predictive structure correlated with judge outputs. To make LLM-as-a-Judge reliable, benchmarks should aim to eliminate recoverable rubric priors. We hope this work encourages more careful analysis of rubric design in automated evaluation.

6 Limitations

Our probe intentionally isolates rubric text, thereby capturing potential rubric priors rather than end-to-end judge behavior in full evaluation settings. We focus on HealthBench and its variants, so the findings may not transfer directly to other domains, rubric styles, or languages. We also rely on a small set of probing methods and a binary formulation, which may understate or distort more subtle rubric effects. Finally, while the results are consistent with shortcut learning, they do not identify the full causal mechanism behind rubric-conditioned judgments. We posit that despite these limitations, we present our work as an initial probe to invite research into the space of designing automated text evaluation.

Acknowledgements

We acknowledge the usage of AI for writing assistance in terms of grammatical corrections and minor re-writing.

References

Rahul K Arora, Jason Wei, Rebecca Soskin Hicks, Preston Bowman, Joaquin Quiñonero-Candela, Foivos Tsimpourlas, Michael Sharman, Meghan Shah, Andrea Vallone, Alex Beutel, and 1 others. 2025. Healthbench: Evaluating large language models towards improved human health. *arXiv preprint arXiv:2505.08775*.

Guiming Hardy Chen, Shunian Chen, Ziche Liu, Feng Jiang, and Benyou Wang. 2024. Humans or llms as the judge? a study on judgement bias. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 8301–8327, Miami, Florida, USA. Association for Computational Linguistics.

Emma Croxford, Yanjun Gao, Elliot First, Nicholas Pellegrino, Miranda Schnier, John Caskey, Madeline Oguss, Graham Wills, Guanhua Chen, Dmitriy Dligach, and 1 others. 2025. Evaluating clinical ai summaries with large language models as judges. *npj Digital Medicine*, 8(1):640.

Alexandra DeLucia, Heyuan Huang, Sonal Joshi, Mahsa Yarmohammadi, Ahmed Hassoon, and Mark Dredze. 2026. Same verdict, different reasons: Llm-as-a-judge and clinician disagreement on medical chatbot completeness. *arXiv preprint arXiv:2604.16383*.

Gemma, Thomas Mesnard, Team, Cassidy Hardin, Robert Dadashi, Surya Bhupatiraju, Shreya Pathak, Laurent Sifre, Morgane Rivière, Mihir Sanjay Kale, Juliette Love, and 1 others. 2024. Gemma: Open models based on gemini research and technology. *arXiv preprint arXiv:2403.08295*.

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*.

Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.

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. *The Innovation*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Trans. Comput. Healthcare*, 3(1).

Takumi Ito, Kees van Deemter, and Jun Suzuki. 2025. Reference-free evaluation metrics for text generation: A survey. *arXiv preprint arXiv:2501.12011*.

Qingquan Li, Shaoyu Dou, Kailai Shao, Chao Chen, and Haixiang Hu. 2026. Evaluating scoring bias in llm-as-a-judge. In *International Conference on Database Systems for Advanced Applications*, pages 19–34. Springer.

Leland McInnes, John Healy, and James Melville. 2018. Umap: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*.

Cathy Ong Ly, Balagopal Unnikrishnan, Tony Tadic, Tirth Patel, Joe Duhamel, Sonja Kandel, Yasbanoo Moayedi, Michael Brudno, Andrew Hope, Heather Ross, and 1 others. 2024. Shortcut learning in medical ai hinders generalization: method for estimating ai model generalization without external data. *NPJ digital medicine*, 7(1):124.

Pengcheng Qiu, Chaoyi Wu, Xiaoman Zhang, Weixiong Lin, Haicheng Wang, Ya Zhang, Yanfeng Wang, and Weidi Xie. 2024. Towards building multilingual language model for medicine. *Nature Communications*, 15(1):8384.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Andrew Sellergren, Chufan Gao, Fereshteh Mahvar, Timo Kohlberger, Fayaz Jamil, Madeleine Traverse, Alberto Tono, Bashir Sadjad, Lin Yang, Charles Lau, and 1 others. 2026. Medgemma 1.5 technical report. *arXiv preprint arXiv:2604.05081*.

Krithi Shailya, Shreya Rajpal, Gokul S Krishnan, and Balaraman Ravindran. 2025. Lext: Towards evaluating trustworthiness of natural language explanations. In *Proceedings of the 2025 ACM Conference on Fairness, Accountability, and Transparency*, pages 1565–1587.

Lin Shi, Chiyu Ma, Wenhua Liang, Xingjian Diao, Weicheng Ma, and Soroush Vosoughi. 2025. Judging the judges: A systematic study of position bias in llm-as-a-judge. In *Proceedings of the 14th International Joint Conference on Natural Language Processing and the 4th Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics*, pages 292–314.

Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Tang, Alejandro Cuadron, Chenguang Wang, Raluca Popa, and Ion Stoica. 2025. Judgebench: A benchmark for evaluating llm-based judges. In *International Conference on Learning Representations*, volume 2025, pages 63277–63303.

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.

Yuqing Zhou, Ruixiang Tang, Ziyu Yao, and Ziwei Zhu. 2024. Navigating the shortcut maze: A comprehensive analysis of shortcut learning in text classification by language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2586–2614.

A Rubric Classification Exploration

Apart from the main classification results, we now present deeper analysis in precision and recall through confusion matrices for an illustration (Figures 7, 8, 9, 10). We evaluated models on the OpenAI HealthBench dataset, which is released under the **MIT license** with strict safeguards under OpenAI Usage Policies and standard HuggingFace Content Policy.

A.1 Models Explored

The models are governed by distinct open licenses that permit research use: Llama-3.1-8B-Instruct is licensed under the **Llama 3.1 Community License Agreement**, Gemma-7b-it is governed by the **Gemma Terms of Use**, and MedGemma model is licensed under the **Health AI Developer Foundations License** but the repository is licensed under the **Apache 2.0 License**. MedLLaMA is licensed under **CC-BY-NC-ND**. Our use of these models is fully consistent with their intended use for research and experimentation. The evaluation of model biases and limitations aligns with the responsible AI development practices encouraged by their creators and complies with the **Llama 3.1 Acceptable Use Policy**, the **Gemma Prohibited Use Policy**, the **MedLlaMA Acceptable Use Policy** and the **Health AI Developer Foundations Terms of Use**.

A.2 Computing Requirements

The experimental pipeline was implemented in Python 3.10. Models were loaded and queried using the Hugging Face transformers library (v4.38.2) with the PyTorch (v2.1) backend. All experiments were executed on a GPU cloud instance with NVIDIA A100 GPUs to accelerate inference. Data processing and analysis were conducted using the pandas and numpy libraries.

A total of 5,000 Question Answering operations, along with 43,000 Yes/No answers for LLM-as-a-judge answers across 4 model generations were performed. The total computational budget is estimated to be approximately 30-35 GPU hours on the specified hardware.

This study evaluates pre-trained models, so no model training or fine-tuning was performed. The key hyperparameters relate to the text generation (decoding) process. To ensure a fair and consistent comparison across all models, a fixed set of decoding parameters was used for every query detailed in Table 1.

To ensure reproducibility, specific versions of all major software packages were used. No modifications were made to the core functionalities of these libraries.

- Core ML/DL Libraries: transformers (v4.38.2), torch (v2.1).
- Data Handling: pandas (v2.0.3), numpy (v1.25.2).
- Geospatial Calculations: geopy (v2.4.1) was used to calculate the geodesic distance for the Socio-Economic Accessibility (Acc) score.

B Simple Classification Exploration

To further probe whether the evaluative signal is recoverable from the rubric text alone, we project the rubric inputs into a TF-IDF vector space and evaluate two simple classifiers (multinomial Naïve Bayes and L2-regularized logistic regression) alongside a majority-class dummy baseline.

	Predicted	
Actual	1043	904
	757	5982

(a) HB Eval Probe

	Predicted	
Actual	153	150
	167	1772

(b) HB Hard Probe

Figure 7: Gemma Model

	Predicted	
Actual	1114	833
	392	6347

(a) HB Eval Probe

	Predicted	
Actual	164	139
	256	1683

(b) HB Hard Probe

Figure 8: LLaMA Model

	Predicted	
Actual	1311	636
	277	6462

(a) HB Eval Probe

	Predicted	
Actual	215	88
	188	1751

(b) HB Hard Probe

Figure 9: MedGemma Model

	Predicted	
Actual	1372	575
	224	6515

(a) HB Eval Probe

	Predicted	
Actual	225	78
	166	1773

(b) HB Hard Probe

Figure 10: MedLLaMA Model

Parameter	Value
temperature	1.0
repetition penalty	1.15
top_p	0.95
num_return_sequences	1

Table 1: Decoding hyperparameters used for all model queries. All the other parameters were default values.

Figure 6 reports cross-validated accuracy, macro-F1, weighted-F1, and ROC-AUC across four embedding models used as judges: Gemma, LLaMA, MedGemma, and MedLLaMA. Across all four metrics, logistic regression consistently outperforms both the Naïve Bayes baseline and the majority-class dummy, confirming that rubric-derived lexical features carry a discriminative signal beyond class imbalance. This pattern holds across embedding families, though performance varies by judge, suggesting that different models impose distinct evaluative surfaces on the same rubric text. These results strengthen the case that rubric-based LLM evaluation pipelines warrant closer methodological scrutiny.

C Rubric Semantics Exploration

C.1 Keywords

To examine whether rubric text encodes lexically recoverable evaluative priors, we conduct two complementary analyses at the word level: binary lexical cue profiling across judge labels, and con-

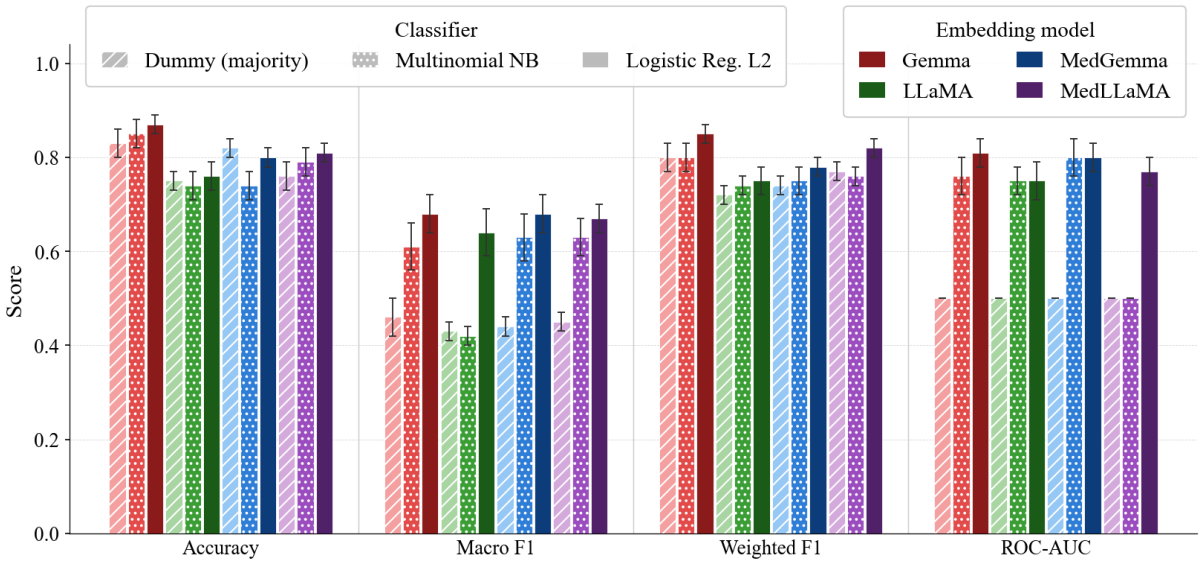

Figure 6: Rubric-Only Classifiers Achieve Nontrivial Performance across Judge Models and Evaluation Metrics

trastive token attribution stratified by class.

Figure 11 shows that several surface tokens exhibit pronounced frequency asymmetries across judge labels. The terms fail, pass, and not appear substantially more often in rubrics assigned label 0, while label 1 rubrics show comparatively sparse usage of these markers. The word should is elevated in both classes but skews toward label 0, suggesting prescriptive and negation-heavy language is a consistent surface correlate of negative judgments.

Figure 12 presents the top logistic regression TF-IDF weights for each class, displayed as a diverging chart. Rubrics associated with negative judgments (label 0) are dominated by failure-marking and negation vocabulary, with justification also carrying strong negative weight, suggesting that rubrics that foreground what must be justified tend to predict failure. Positive-judgment rubrics (label 1) show the mirror pattern: the highest-weighted tokens are epistemic and affirmative verbs, reflecting rubric language that describes what a correct response should exhibit rather than what it must not do. The two vocabularies are entirely disjoint, indicating that failure-predictive language is encoded more strongly and distinctively in rubric text than success-predictive language.

C.2 Word Length Analysis

Figure 13 shows the distribution of rubric word counts stratified by judge label. Label 0 rubrics are systematically longer than label 1 rubrics. Both the distributions overlap substantially, so word count alone does not cleanly separate the two classes. The many outliers in both labels suggest that a small number of very long rubrics exist in each group.

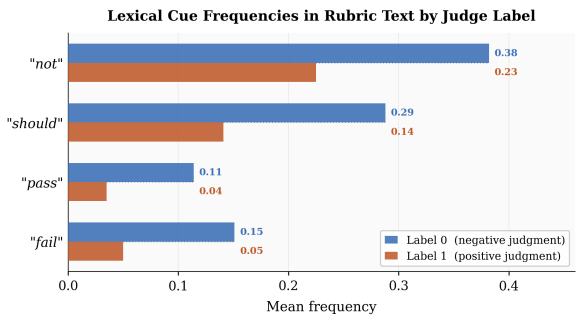

Figure 11: Lexical Cue Frequencies in Rubric Text. Mean frequency of selected binary lexical features across rubrics stratified by judge label .

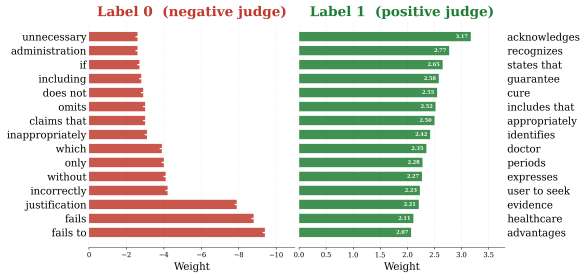

Figure 12: Top Logistic Regression Weights by Judge Label. Rubric phrasing encodes a class-discriminative lexical signal

C.3 BERTopic Modeling

To move beyond individual tokens and examine the sentence-level semantic structure of rubric text, we apply BERTopic (Grootendorst, 2022) conditioned on the judge label. Rather than jointly discovering topics across all rubrics, we fit separate topic models for each class, allowing us to characterize the evaluative concerns prototypical of positive and negative judgments independently.

The top five topics per class, shown in Figure 14, reveal a semantically coherent contrast. Class 0 rubrics cluster around failure modes and clinical edge cases (symptom justification requirements, verbose or repetitive responses, and scenarios involving vaccines, emergency services, and medication interactions) consistent with the negation-heavy lexical profile identified earlier. Class 1 rubrics organize around affirmative clinical competencies: general medical advice, dosage correctness, vaccination guidance, postpartum mental health, and antibiotic stewardship.

The two topic sets are largely disjoint, with no direct thematic overlap in the top five, confirming

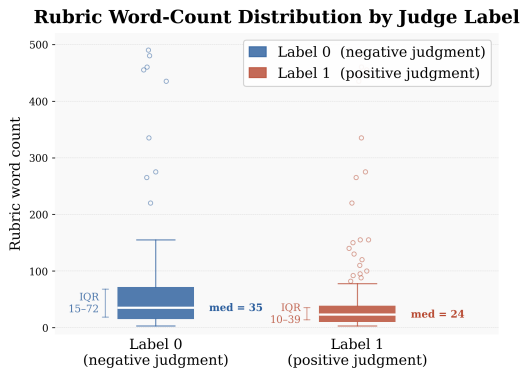

Figure 13: Rubric Word-Count Distribution by Judge Label. Label 0 rubrics are longer on average; distributions overlap substantially.

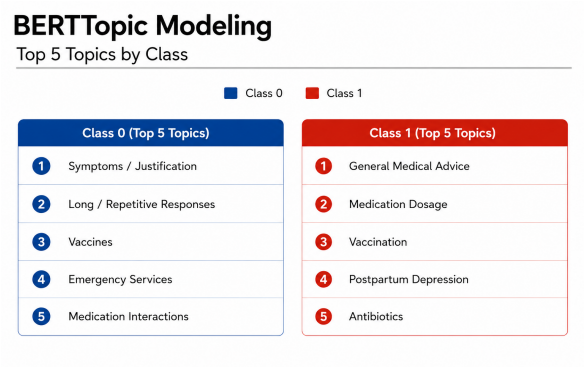

Figure 14: BERTopic Model results on HEALTHBENCH-HARD and HEALTHBENCH-EVAL

that positive and negative judgment rubrics occupy distinct semantic regions. That these regions correspond to coherent clinical subfields rather than generic discourse patterns demonstrates that the evaluative signal in rubric text is semantically structured at the sentence level, not merely a surface lexical artifact.