CONCORD: Disagreement-Aware Alignment of Medical LLM Judges

Anonymous ACL submission

Abstract

Large Language Models (LLMs) are increasingly used as automatic evaluators, yet reliable assessment remains difficult for open-ended medical reasoning tasks. In medicine, expert judgments are often costly, limited, and characterized by substantial inter-clinician disagreement, making deterministic supervision unreliable. We introduce CONCORD, a disagreement-aware framework for aligning medical LLM judges under limited and uncertain expert supervision. Rather than treating disagreement as annotation noise, our approach explicitly incorporates annotation uncertainty during both supervised fine-tuning (SFT) and reinforcement learning (RL) through agreement-weighted, uncertainty-aware, and ordinal-aware optimization strategies. We additionally study a meta-evaluation setting where a stronger frozen LLM provides reward signals without requiring gold annotations. Experiments on medical reasoning evaluation show that our adapted 4B-parameter judge substantially improves agreement with clinicians and approaches the performance of much larger proprietary evaluators such as Gemini-2.5-Pro. Our results highlight the potential of explicitly modeling annotation disagreement when aligning medical LLM judges.

1 Introduction

Reliable evaluation of natural language generation remains a major challenge in NLP (Sai et al., 2022), particularly for tasks involving complex reasoning and multiple valid responses. The instruction-following capabilities of generative LLMs further increase the diversity of acceptable outputs (Chang et al., 2024), making evaluation based solely on reference overlap increasingly insufficient. Traditional reference-based metrics such as BLEU or ROUGE (Papineni et al., 2002; Lin, 2004) often correlate poorly with human judgments, especially for reasoning-intensive tasks (Sai et al., 2022; Yim et al., 2026). While multiple-choice benchmarks partially simplify evaluation (Hendrycks et al., 2021), they fail to capture the free-form reasoning required in many real-world LLM applications (Zheng et al., 2023).

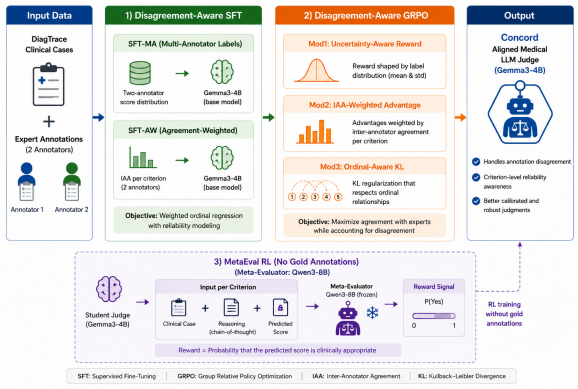

Figure 1: Overview of CONCORD framework for disagreement-aware medical LLM judgment alignment.

These limitations become particularly critical in medicine, where evaluating generated responses often depends on the quality, consistency, and clinical validity of the underlying reasoning rather than similarity to a single reference answer (Kim et al., 2025). Moreover, clinical evaluation itself is inherently uncertain. Physicians frequently disagree on diagnoses, treatment decisions, and reasoning quality due to differences in experience, specialties, clinical uncertainty, and interpretation strategies (Pavlick and Kwiatkowski, 2019). As a result, expert annotations for medical reasoning evaluation are both expensive to obtain and intrinsically heterogeneous.

LLM-as-a-judge (Zheng et al., 2023) has emerged as a promising alternative to costly expert evaluation. However, existing general-purpose judges often correlate poorly with domain experts in specialized medical settings (Gu et al., 2024; Li et al., 2024). Current solutions typically rely on increasingly larger proprietary models or ensembles of evaluators (Bedi et al., 2026), improving performance at the cost of scalability, reproducibil-

ity, and accessibility. More importantly, most existing judge-alignment approaches implicitly assume clean and consistent supervision signals, an assumption that rarely holds in realistic clinical scenarios.

In this work, we address this challenge through CONCORD, a disagreement-aware framework for aligning medical LLM judges under limited and inconsistent expert supervision (Figure 1). Rather than treating annotator disagreement as noise to discard, we explicitly model disagreement as a structured uncertainty signal during both supervised fine-tuning (SFT) and reinforcement learning (RL). Our approach introduces agreement-weighted supervision, uncertainty-aware reward shaping, and ordinal-aware policy optimization for criterion-level medical reasoning evaluation.

Beyond expert-supervised alignment, we further investigate whether reliable medical judges can be trained entirely without gold annotations through meta-evaluation, where a stronger frozen LLM provides reward signals for RL. This setting explores a scalable alternative for domains where expert annotation is prohibitively expensive.

We investigate the following research questions:

- ***RQ1:*** *Can small LLM judges be reliably aligned for medical reasoning evaluation under limited and disagreement-prone expert supervision?*
- ***RQ2:*** *Does explicitly modeling inter-annotator disagreement improve judge alignment and uncertainty calibration compared to standard supervised and reinforcement learning approaches?*
- ***RQ3:*** *Can strong LLM evaluators provide sufficiently informative reward signals to train medical judges without any gold annotations?*

Our main contributions are summarized as follows: **(i)** We introduce CONCORD, a disagreement-aware framework for aligning medical LLM judges through SFT and RL under limited and inconsistent expert supervision. **(ii)** We study meta-evaluation for medical reasoning assessment, investigating whether smaller judges can be trained entirely from reward signals generated by stronger frozen LLM evaluators in low-resource annotation setting.

We publicly release all code, trained models, and evaluation resources to support reproducibility and future research.[1]

[1] https://anonymous.4open.science/r/Concord-F714/README.md

2 Related Work

2.1 LLM-as-a-judge

The use of LLMs as evaluators of open-ended text has become a popular paradigm in NLP evaluation (Li et al., 2025; Zheng et al., 2023; Gu et al., 2024) to take advantage of scalability and cost advantages over human evaluation, though mostly relying on large commercial LLMs to ensure quality (Huang et al., 2025). To reduce dependence on large commercial models, a parallel line of work has trained dedicated open-weight judge models, such as the Prometheus family (Pombal et al., 2025). In medical NLP, automated evaluation is particularly challenging due to the high stakes and specificity of clinical judgment. Bedi et al. (2026) and Grolleau et al. (2025) experimented with using a jury of LLM judges to evaluate free-form texts with a high level of correlation to a panel of clinicians, but Rouillard et al. (2026) cautioned that a frontier-level LLM jury without any calibration scored systematically lower than a panel of clinicians given the same task. Despite promising results, reliance on proprietary LLMs remains challenging in healthcare due to patient privacy constraints and limited institutional resources. These limitations motivate approaches that adapt compact LLM judges to specialized medical domains rather than depending exclusively on large general-purpose models.

2.2 Model Alignment for Specialized Judgment task

Most LLM-as-a-judge systems utilize SFT to align smaller models to judgment datasets created by larger models or human annotators. The Prometheus family (Pombal et al., 2025) and JudgeLM (Zhu et al., 2025) relied on GPT-4 as a judgment generator, while Wang et al. (2024) used human preference data. Belmadani et al. (2026) further improved SFT alignment using GRPO (Shao et al., 2024) in the medical domain. However, these approaches assume clean and consistent supervision signals, an assumption that rarely holds in specialized clinical settings where expert annotations are scarce, noisy, and inherently subjective (Gingerich et al., 2017; Pavlick and Kwiatkowski, 2019). The problem of learning under annotator disagreement has a long history, tracing back to probabilistic models of observer error (Dawid and

Skene, 1979), and more recently to loss adjustment techniques for noisy supervision (Song et al., 2022; Jamison and Gurevych, 2015). Our framework builds on this tradition and is further inspired by reinforcement learning with meta-evaluation (RLME) (Rentschler and Roberts, 2026a), adapting both the SFT and GRPO learning signals to the scarce and noisy annotation regime.

3 Experimental Framework

3.1 Task Definition

We address the problem of automated expert medical reasoning evaluation: given a clinical case and a predicted reasoning chain produced by an LLM, a judge model must assign criterion-level scores reflecting the quality of the reasoning. This setting corresponds to the *LLM-as-a-Judge* paradigm, yet unlike typical evaluation settings that focus primarily on final answers or QA correctness, the judge here must assess the reasoning process leading to the diagnosis rather than the diagnosis alone.

3.2 Dataset

We conduct all experiments on *DiagTrace* (Bazoge et al., 2026), a bilingual benchmark designed to evaluate LLM diagnostic reasoning on real-world clinical cases. The dataset comprises 180 clinical cases drawn from multiple medical specialties, each presented in both French and their English translation. A total of 1,800 prediction instances are sourced from five LLMs of varying capability and medical specialization: GPT-4-Turbo, Llama-3.1-405B-Instruct, DeepSeek-R1, o3, and BioMistral, with 360 predictions per generator.

Each prediction is independently annotated by two medical doctors across six evaluation criteria. Table 1 summarizes the evaluation dimensions and score ranges used in the benchmark.

We split the dataset at the case level to prevent data leakage: 90 cases (900 examples) for training, 10 cases (100 examples) for validation, and 80 cases (800 examples) for testing. Score distributions are balanced across splits, with mean total scores of 14.38, 14.15, and 14.39, respectively.

3.3 Inter-Annotator Agreement Analysis

To better characterize supervision reliability, we analyze inter-annotator agreement (IAA) using Krippendorff's α, computed between the dataset's two expert annotators across all evaluation criteria (Figure 2). IAA is measured using ordinal distance for polytomous criteria and nominal distance for the binary *differential diagnosis* criterion. We observe substantial heterogeneity across evaluation criteria, with *final diagnosis* and *logical structure* exhibiting the strongest agreement, while *external validity* remains substantially more subjective. Agreement patterns remain relatively stable across splits. The largest language discrepancy is observed for *expression*, where English annotations achieve substantially higher agreement, likely because annotations were performed by French-speaking clinicians who can evaluate stylistic and linguistic quality more consistently in their native language and tend to apply stricter standards in French than in English. Overall, several criteria remain below the commonly used $\alpha = 0.6$ threshold, highlighting the noisy and heterogeneous nature of medical reasoning evaluation. This observation directly motivates CONCORD, which explicitly incorporates inter-annotator uncertainty during both supervised and reinforcement learning optimization.

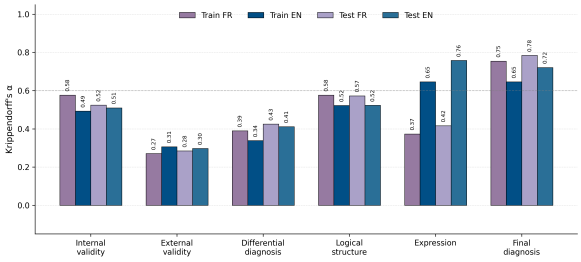

Figure 2: Krippendorff's alpha values of each criterion between the two annotators.

3.4 Baselines

We evaluate ten off-the-shelf LLMs as zero-shot judge baselines across proprietary, general-purpose, medical-specialized, and judge-specialized settings. Proprietary models include GPT-5.4[2] and Gemini-2.5-Pro (Comanici et al., 2025). Open general-purpose models include Gemma3-4b-it (Team et al., 2025), Qwen3-8B (Yang et al., 2025a), Qwen2.5-7B (Yang et al., 2025b), and Llama-3.1-8B-Instruct[3]. Medical-specialized models include MedGemma-4b-it (Sellergren et al., 2025), Aloe-Beta-8B (Garcia-Gasulla et al., 2026) and II-Medical-8B (Internet, 2025), while judge-specialized models include M-Prometheus-7B (Pombal et al., 2025). Models span comparable parameter scales (∼4–8B), enabling controlled

[2] https://openai.com/fr-FR/index/introducing-gpt-5-4/

[3] https://ai.meta.com/blog/meta-llama-3-1/

comparisons across model families and specialization strategies.

3.5 Prompting Strategies

All experiments rely on a unified criterion-level medical reasoning evaluation framework. We investigate two prompting settings throughout both evaluation and training: *with-ref*, where the evaluator can use gold expert diagnosis and reasoning as a reference to evaluate; and *no-ref*, where evaluation is performed solely from the clinical case and generated reasoning chain. The *no-ref* setting reflects a realistic clinical deployment scenario, whereas the *with-ref* setting provides an approximate upper bound on evaluation performance.

Prompts are bilingual and automatically adapted to the language of the evaluated clinical case. The same prompting structure is reused consistently across zero-shot evaluation and training. Detailed prompting templates are provided in Appendix B.

4 CONCORD Framework

Given the subjectivity of medical reasoning evaluation (Gingerich et al., 2017), we investigate explicitly modeling inter-annotator variability during training, rather than treating it as noise. Figure 2 reveals substantial IAA heterogeneity across evaluation criteria on DiagTrace, indicating uneven supervision reliability across dimensions.

To address this issue, we propose **CONCORD**, a framework for training robust medical LLM judges under annotation disagreement. It explicitly incorporates inter-annotator uncertainty during both SFT and RL, allowing optimization to adapt dynamically to the reliability of different criteria and examples. We use Gemma3-4B-it as the backbone model throughout all experiments and perform all updates using LoRA (Hu et al., 2021) adapters. Hyperparameter details of training strategies are provided in Appendix G.

4.1 Supervised Fine-Tuning

We propose two disagreement-aware SFT strategies.

SFT-MA (Multi-Annotator). Each annotation is treated as an independent supervision target; disagreed examples thus generate multiple training instances, increasing the representation of ambiguous cases.

SFT-AW (Agreement-Weighted). We additionally explore disagreement-aware loss weighting. For training example i, the token-level cross-entropy loss is scaled by an agreement weight w_i:

$$\mathcal{L}_{\text{SFT-AW}} = \frac{1}{N}\sum_{i=1}^{N} w_i \cdot \frac{1}{T_i}\sum_{t=1}^{T_i} \text{CE}(\hat{y}_{i,t}, y_{i,t})$$

where:

$$w_i = \frac{1}{C}\sum_{c=1}^{C}\left(1 - \frac{|s_{c,e_1}^{(i)} - s_{c,e_2}^{(i)}|}{s_{c,\max}}\right)$$

with C evaluation criteria, expert annotations e_1, e_2, and criterion maximum score $s_{c,\max}$. This formulation reduces the influence of highly disputed examples during optimization, allowing the model to focus more on reliable supervision signals.

4.2 GRPO Fine-Tuning

Standard GRPO does not account for ordinal scoring, criterion-level reliability, or annotation disagreement. We therefore introduce three complementary modifications targeting different components of GRPO optimization: reward shaping (Mod1), advantage estimation (Mod2), and policy regularization (Mod3).

Mod1: Ordinal-Aware Reward with Annotator Uncertainty. For criterion c, let $\hat{s}_c$ denote the score predicted by the judge, and let s_{c,e_1} and s_{c,e_2} denote the scores assigned by the two medical experts. All scores are normalized by the criterion-specific maximum score $s_{c,\max}$. For each criterion c, we define a disagreement-aware ordinal reward:

$$r_c = \sum_{i \in \{e_1, e_2\}} \left(1 - \frac{|\hat{s}_c - s_{c,i}|}{s_{c,\max}}\right) w_i$$

where annotator weights decrease as disagreement increases:

$$w_{e_1} = w_{e_2} = \frac{1}{2}\exp\left(-\lambda \cdot \frac{|s_{c,e_1} - s_{c,e_2}|}{s_{c,\max}}\right)$$

where λ controls the sensitivity of the reward to annotator disagreement. Larger values of λ lead to a stronger downweighting of highly disputed annotations. This formulation interprets disagreement as uncertainty rather than annotation noise.

Mod2: IAA-Weighted Criterion-Level Advantage. Rather than computing a single advantage over the aggregated score, we compute criterion-level advantages weighted by IAA:

$$\hat{A}_i = \sum_{c=1}^{C} \alpha_c \cdot \frac{r_{i,c} - \bar{r}_c}{\sigma_{r_c}}$$

where α_c denotes Krippendorff's α computed independently for criterion c. This encourages the policy to rely more strongly on criteria exhibiting higher annotator agreement.

Mod3: Ordinal-Aware Asymmetric KL Penalty. We modify the GRPO Kullback-Leibler (KL) regularization term to account for ordinal distance between scores:

$$D_{KL}^{\text{ord}}[\pi_\theta || \pi_{\text{ref}}] = \sum_s \pi_\theta(s) \log \frac{\pi_\theta(s)}{\pi_{\text{ref}}(s)} \phi(s, s^*)$$

with:

$$\phi(s, s^*) = 1 + \eta \frac{|s - s^*|}{s_{\max}}$$

where η controls the strength of the ordinal distance penalty, s^* denotes the target reference score used for regularization, and s denotes a possible score prediction. This discourages overly confident policy updates on predictions that are ordinarily far from the reference score while remaining tolerant to nearby ordinal deviations.

We evaluate all combinations of the three modifications, yielding eight GRPO variants: *standard*, *Mod1*, *Mod2*, *Mod3*, *Mod1+Mod2*, *Mod1+Mod3*, *Mod2+Mod3*, and *Mod1+Mod2+Mod3*. In addition, we investigate the impact of GRPO initialization by comparing optimization starting from the base Gemma3-4B-it model versus the best SFT checkpoint. Detailed initialization ablations are provided in Appendix E.

4.3 Meta-Evaluator Reinforcement Learning

Following recent work on RL from meta-evaluation in mathematical reasoning (Rentschler and Roberts, 2026b), we investigate whether training without gold annotations can generalize to medical reasoning evaluation. More specifically, we explore whether a frozen LLM meta-evaluator can provide sufficiently informative reward signals for criterion-level medical judgment alignment.

To this end, we use a frozen Qwen3-8B model as a meta-evaluator (see Appendix D for the model selection rationale). For each criterion, the evaluator receives the clinical case, the generated reasoning chain, and the predicted score, and outputs the probability that the score is clinically appropriate. We use the probability $P(\text{Yes})$ as the reward signal. Detailed meta-evaluator prompting templates are provided in Appendix B.2. Criterion-level rewards are aggregated using the same IAA-weighted advantage formulation introduced in Mod2. This setting is trained without reference information.

4.4 Evaluation Protocol

All models are evaluated on the held-out test set using greedy decoding for reproducibility and to minimize generation variability across runs. We report Spearman ρ as the primary evaluation metric, computed separately against each expert's annotations and averaged.

Spearman correlation is particularly suited to our setting as the evaluation criteria correspond to ordinal clinical scores: it captures monotonic agreement between predicted and expert rankings while remaining robust to small ordinal deviations.

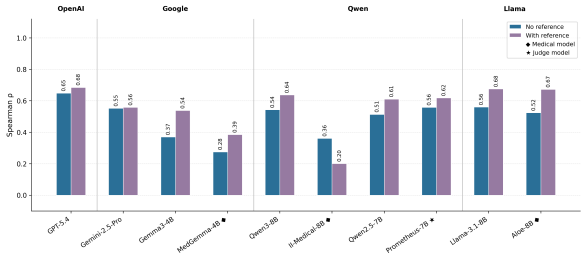

Figure 3: Spearman ρ values of ten LLMs on the test set. Models are grouped based on the base model, with $\diamondsuit$ denoting medically-adapted models, and $\star$ denotes a model aligned to the task of judging.

5 Results

5.1 Baseline results

Figure 3 reports the performance of ten off-the-shelf LLMs as medical judges under both *with-ref* and *no-ref* prompting settings. Overall, providing expert references improves performance across most models, although the gap remains moderate for frontier closed-source systems. For instance, GPT-5.4 improves only from 0.65 to 0.68 overall Spearman correlation, while Gemini-2.5-Pro remains nearly stable (0.55 vs. 0.56), suggesting that strong LLMs already possess intrinsic reasoning evaluation capabilities even without gold answers.

Commercial closed-source models achieve the strongest overall performance, with GPT-5.4 ranking first in both settings. Nevertheless, several open-weight models remain competitive despite substantially smaller parameter counts. Qwen3-8B reaches 0.54 in the *no-ref* setting and 0.64 with references, while Prometheus-7B achieves 0.56 and

0.62 respectively.

Interestingly, medical specialization alone does not systematically improve evaluation quality. Medical-adapted models such as MedGemma-4B and II-Medical-8B generally underperform compared to their general-domain counterparts. For example, II-Medical-8B reaches only 0.20 overall correlation in the *with-ref* setting compared to 0.64 for Qwen3-8B. These results suggest that medical reasoning evaluation requires capabilities beyond medical factual knowledge alone. More detailed analyses are provided in Appendix H.

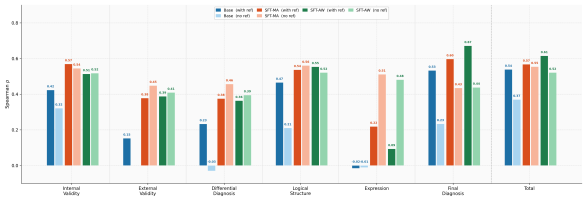

Figure 4: Spearman ρ values of different strategies of disagreement adapted SFT on all criteria, from Gemma-3-4B-it as the Base model. Spearman ρ is not reported where models always output the same score.

5.2 CONCORD Framework results

5.2.1 Disagreement-aware SFT

Figure 4 reports the correlation between model predictions and expert annotations for the different SFT strategies. Overall, both disagreement-aware approaches substantially improve over the base model under both *with-ref* and *no-ref* settings.

SFT-MA generally achieves stronger performance on criteria exhibiting lower IAA, such as *external validity* and *differential diagnosis*. By explicitly preserving multiple expert annotations during training, the model is exposed to a broader range of plausible supervision signals for ambiguous examples. However, treating all annotations equally may also propagate noisy or conflicting supervision signals, which appear to negatively impact more reliable criteria such as *final diagnosis*.

In contrast, *SFT-AW* produces more balanced performance across criteria and achieves the strongest overall agreement with expert annotations, particularly in the *with-ref* setting. These results suggest that explicitly weighting supervision by annotator agreement provides a more stable optimization signal than uniformly training on all annotations. Consequently, we use the *SFT-AW with-ref* checkpoint as initialization for subsequent RL experiments.

5.2.2 Disagreement-Aware GRPO

In this section, we focus on the *no-ref* setting, which better reflects scenarios where expert reference answers are unavailable. Additional analyses, including *with-ref* results and initialization ablations are provided in Appendix I.

Figure 5 reports the performance of the proposed GRPO variants. Overall, RL substantially improves over both the base model and disagreement-aware SFT, confirming the benefit of policy optimization.

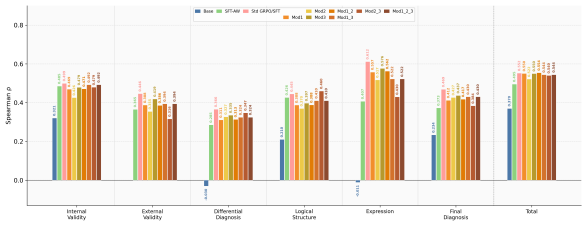

Figure 5: Spearman ρ values of disagreement-aware SFT and GRPO variants of Gemma-3-4B-it (Base), under *no-ref* evaluation setting. Spearman ρ is not reported where models always output the same score.

Standard GRPO initialized from the best disagreement-aware SFT checkpoint (SFT-AW) already yields strong gains over SFT-AW, improving the overall Spearman correlation from 0.495 to 0.552. The largest improvements are observed for *expression*, *logical structure*, and *final diagnosis*, suggesting that RL particularly benefits criteria associated with coherent reasoning and scoring consistency.

The proposed disagreement-aware modifications produce more heterogeneous effects beyond the disagreement-aware initialization already provided by SFT-AW. Mod1 and Mod3 generally provide the strongest individual improvements, while Mod2 alone yields more limited gains. Some combinations, such as Mod1+Mod2, slightly outperform standard GRPO, whereas larger combinations do not consistently improve performance further.

Interpreting these differences is complicated by the fact that evaluation relies on noisy, partially subjective annotations. Since our modifications explicitly model annotator disagreement during optimization, they may align less aggressively to individual noisy labels than standard GRPO. This means lower correlation on the noisy test set may not necessarily reflect inferior evaluation behavior. This is further compounded by the fact that all GRPO variants are initialized from SFT-AW, which already incorporates disagreement-aware su-

pervision; the GRPO-specific modifications should therefore be read as incremental refinements on top of an already uncertainty-aware evaluator, not corrections to a standard baseline.

5.2.3 Uncertainty and Disagreement Sensitivity

While disagreement-aware GRPO variants do not systematically outperform standard GRPO on the full noisy evaluation set, it does not imply that disagreement modeling is ineffective. Stronger alignment with uncertainty-aware supervision may simply not translate into higher correlation on inherently ambiguous examples. To better understand the behavioral impact of our modifications, we analyze prediction entropy under both criterion-level and example-level disagreement conditions. All comparisons are performed under a controlled setup where GRPO variants share the same backbone, training data, decoding strategy, and optimization budget, differing only in the proposed modifications. Standard GRPO initialized from SFT-AW thus serves as the principal baseline, isolating the contribution of disagreement-aware reward shaping, agreement-aware advantage estimation, and ordinal regularization.

Figure 13 reports criterion-level normalized entropy across all models. Disagreement-aware GRPO variants exhibit substantially higher entropy than both the base model and standard GRPO, indicating that the proposed modifications encourage uncertainty-sensitive predictions rather than deterministic scoring. Entropy increases are not uniform: subjective criteria such as *expression*, *differential diagnosis*, and *external validity* consistently show the highest entropy, while more stable criteria such as *internal validity* and *final diagnosis* maintain lower values, suggesting that the modifications preserve uncertainty specifically on ambiguous clinical dimensions.

To further analyze uncertainty calibration, Figure 6 compares entropy between fully agreed and disagreed examples (168 each). Disagreement cases consistently produce higher entropy across all models, confirming greater uncertainty on ambiguous instances. Disagreement-aware variants amplify this separation considerably relative to standard GRPO. Mod2 and Mod3 exhibit the strongest sensitivity: Mod2 increases entropy on disputed examples while preserving low entropy on agreed ones, and Mod3 introduces ordinal-aware regularization for stronger calibration. Their combination produces the largest entropy gap overall, suggesting complementary uncertainty modeling effects.

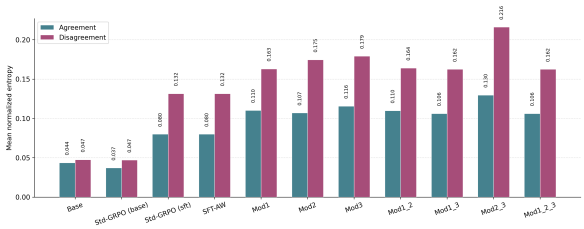

Figure 6: Normalized entropy for fully agreed and disagreed examples across evaluated models.

Notably, standard GRPO initialized from SFT-AW already exhibits higher entropy than its base-model counterpart, confirming that disagreement-aware behavior learned during SFT propagates into subsequent RL optimization and further supporting the importance of incorporating annotator disagreement already at the supervised adaptation stage.

Overall, these analyses suggest that disagreement-aware GRPO primarily improves the calibration and uncertainty behavior of medical LLM judges. Although such effects may not always yield higher correlation on noisy evaluation labels, they indicate that the proposed modifications encourage the model to preserve uncertainty under clinically ambiguous supervision rather than overfitting inconsistent annotations.

5.2.4 Meta-Evaluation Reinforcement Learning

This section presents the results of our preliminary investigation into whether reliable medical LLM judges can be trained without expert annotations by leveraging supervision from a stronger frozen meta-evaluator.

Figure 7 reports the performance of the MetaEval approach on the filtered subset containing only fully agreed examples (168 examples), providing a cleaner evaluation setting with reduced annotation noise. Overall, MetaEval substantially improves

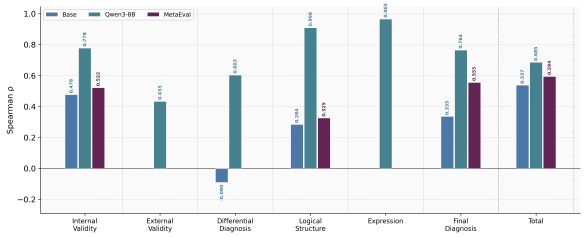

Figure 7: Performance of MetaEval on the filtered subset containing only fully agreed examples.

over the original base model across most evaluation criteria, demonstrating that useful medical evaluation behavior can indeed be transferred from a stronger teacher model. The largest gains are observed for *final diagnosis* (0.335 → 0.555), *internal validity* (0.476 → 0.522), and overall correlation (0.537 → 0.594). These results indicate that meta-evaluation signals capture clinically meaningful judgment patterns even in the absence of direct human supervision.

Nevertheless, a substantial gap remains between MetaEval and the Qwen3-8B meta-evaluator itself, which reaches an overall correlation of 0.685. This suggests that while criterion-level reward distillation successfully transfers part of the evaluation capability, reproducing the full reasoning behavior of a stronger clinical evaluator remains challenging.

Performance improvements are also uneven across evaluation criteria. MetaEval performs particularly well on *final diagnosis*, suggesting that teacher-provided rewards effectively capture high-level clinical correctness. In contrast, criteria where the base model exhibits nearly monotonic or collapsed scoring behavior remain more difficult to improve through meta-evaluation alone. In particular, *expression* and *external validity* show limited adaptation despite strong teacher performance. One possible explanation is that once the student judge collapses toward low-variance predictions, the scalar reward signal provided by the meta-evaluator becomes insufficient to recover fine-grained ordinal distinctions during optimization.

MetaEval training relies exclusively on reward signals from the frozen Qwen3-8B meta-evaluator, with no gold annotations. Despite this challenging setup, the resulting student judge substantially improves over the base model, suggesting that strong frozen evaluators can effectively transfer medical evaluation capabilities to smaller models through RL alone. This is particularly promising for deployment scenarios where expert annotations are scarce: rather than relying on large models at inference time, a stronger teacher can be used during a limited training phase to distill evaluation behavior into a smaller, deployable judge. Overall, these results constitute a promising first step toward scalable medical LLM judges trained entirely without human annotations.

6 Conclusion

In this work, we investigated the problem of aligning medical LLM judges under annotation disagreement and noisy clinical supervision, a low-resource setting that is rarely modeled in existing datasets and particularly prevalent in clinical domains where expert annotations are scarce and costly to obtain. We introduced **CONCORD**, a disagreement-aware framework combining SFT and RL strategies designed to model ordinal uncertainty, criterion reliability, and inter-annotator disagreement in medical reasoning evaluation.

Our analyses first revealed substantial heterogeneity across evaluation criteria, with several dimensions exhibiting relatively low IAA despite expert annotation. These observations motivated the development of disagreement-aware optimization strategies operating at both the supervised and reinforcement learning stages.

Experimental results showed that disagreement-aware SFT consistently improves over the base model, while RL further enhances alignment with expert annotations under realistic no-reference settings. Although disagreement-aware GRPO variants do not systematically dominate standard GRPO from SFT-AW on the full noisy evaluation set, additional analyses on fully agreed examples and uncertainty-sensitive entropy evaluations demonstrate that the proposed modifications encourage more calibrated and uncertainty-aware evaluation behavior under clinically ambiguous supervision.

We additionally explored a meta-evaluation setting in which a judge is trained without gold annotations, using reward signals from a stronger frozen evaluator. Despite the difficulty of this setup, the resulting model substantially improves over the base judge, suggesting that meta-evaluation is a promising direction for scaling medical LLM judges beyond costly annotation pipelines.

Overall, our results highlight that medical LLM-as-a-Judge should not be treated as a standard deterministic classification problem. Instead, clinical evaluation inherently involves ambiguity, uncertainty, and partially subjective reasoning processes that must be explicitly modeled during alignment. We hope this work encourages future research on uncertainty-aware medical evaluation, disagreement-sensitive alignment objectives, and scalable judge training strategies for reliable clinical reasoning assessment.

Limitations

Several limitations should be considered when interpreting our results. First, the annotation setup involves only two medical experts, which limits the robustness of IAA estimates and may not fully capture the diversity of clinical reasoning strategies encountered in real-world medical practice. While disagreement modeling is a central component of our framework, larger multi-expert evaluation settings would provide more reliable estimates of uncertainty and criterion-level variability.

Second, our experiments focus on a single medical reasoning benchmark and a limited set of evaluation criteria. Although the proposed framework is designed to be task-agnostic, additional studies across other clinical domains, reasoning tasks, and annotation protocols are necessary to assess its generalizability.

Third, RL optimization remains challenging under noisy supervision. Although disagreement-aware modifications improve uncertainty sensitivity and calibration behavior, improvements on the full noisy evaluation set remain heterogeneous and sometimes difficult to interpret. This highlights the broader difficulty of evaluating LLM judges when the supervision signal itself is partially inconsistent.

Finally, our meta-evaluation experiments are also preliminary. While the results demonstrate the feasibility of training medical judges entirely without gold annotations, the student evaluator still remains substantially below the stronger frozen teacher model. Future work should explore richer reward formulations, stronger teacher-student alignment strategies, and iterative self-improvement approaches for judge distillation.

References

Adrien Bazoge, Josselin Corvellec, Sofiane Djillali Sid-Ahmed, and Pierre-Antoine Gourraud. 2026. Prompting language influences diagnostic reasoning and accuracy of large language models. *Preprint*, arXiv:2605.19173.

Suhana Bedi, Hejie Cui, Miguel Fuentes, Alyssa Unell, Michael Wornow, Juan M. Banda, Nikesh Kotecha, Timothy Keyes, Yifan Mai, Mert Oez, Hao Qiu, Shrey Jain, Leonardo Schettini, Mehr Kashyap, Jason Alan Fries, Akshay Swaminathan, Philip Chung, Fateme Nateghi Haredasht, Ivan Lopez, and 64 others. 2026. Holistic evaluation of large language models for medical tasks with medhelm. *Nature Medicine*, 32(3):943–951.

Ikram Belmadani, Oumaima El Khettari, Pacôme Constant dit Beaufils, Richard Dufour, and Benoit Favre. 2026. Who judges the judge? evaluating llm-as-a-judge for french medical open-ended qa. In *Proceedings of the 1st Workshop on Linguistic Analysis for Health (HeaLing 2026)*, pages 142–157.

Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, and 1 others. 2024. A survey on evaluation of large language models. *ACM transactions on intelligent systems and technology*, 15(3):1–45.

Gheorghe Comanici, Eric Bieber, Mike Schaekermann, Ice Pasupat, Noveen Sachdeva, Inderjit Dhillon, Marcel Blistein, Ori Ram, Dan Zhang, Evan Rosen, and 1 others. 2025. Gemini 2.5: Pushing the frontier with advanced reasoning, multimodality, long context, and next generation agentic capabilities. *arXiv preprint arXiv:2507.06261*.

Alexander Philip Dawid and Allan M Skene. 1979. Maximum likelihood estimation of observer error-rates using the em algorithm. *Journal of the Royal Statistical Society: Series C (Applied Statistics)*, 28(1):20–28.

Dario Garcia-Gasulla, Jordi Bayarri-Planas, Ashwin Kumar Gururajan, Enrique Lopez-Cuena, Adrian Tormos, Daniel Hinjos, Pablo Bernabeu-Perez, Anna Arias-Duart, Pablo Agustin Martin-Torres, Marta Gonzalez-Mallo, and 1 others. 2026. The aloe family recipe for open and specialized healthcare llms. *npj Digital Medicine*.

Andrea Gingerich, Susan E Ramlo, Cees PM van der Vleuten, Kevin W Eva, and Glenn Regehr. 2017. Inter-rater variability as mutual disagreement: identifying raters' divergent points of view. *Advances in Health Sciences Education*, 22(4):819–838.

Françcois Grolleau, Emily Alsentzer, Timothy Keyes, Philip Chung, Akshay Swaminathan, Asad Aali, Jason Hom, Tridu Huynh, Thomas Lew, April Liang, and 1 others. 2025. Medfacteval and medagentbrief: A framework and workflow for generating and evaluating factual clinical summaries. In *Biocomputing*

2026: Proceedings of the Pacific Symposium, pages 388–399. World Scientific.

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*.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *International Conference on Learning Representations*.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models. *Preprint*, arXiv:2106.09685.

Hui Huang, Xingyuan Bu, Hongli Zhou, Yingqi Qu, Jing Liu, Muyun Yang, Bing Xu, and Tiejun Zhao. 2025. An empirical study of LLM-as-a-judge for LLM evaluation: Fine-tuned judge model is not a general substitute for GPT-4. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 5880–5895, Vienna, Austria. Association for Computational Linguistics.

Intelligent Internet. 2025. Ii-medical-8b: Medical reasoning model.

Emily Jamison and Iryna Gurevych. 2015. Noise or additional information? leveraging crowdsource annotation item agreement for natural language tasks. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 291–297, Lisbon, Portugal. Association for Computational Linguistics.

Jonathan Kim, Anna Podlasek, Kie Shidara, Feng Liu, Ahmed Alaa, and Danilo Bernardo. 2025. Limitations of large language models in clinical problem-solving arising from inflexible reasoning. *Scientific reports*, 15(1):39426.

Dawei Li, Bohan Jiang, Liangjie Huang, Alimohammad Beigi, Chengshuai Zhao, Zhen Tan, Amrita Bhattacharjee, Yuxuan Jiang, Canyu Chen, Tianhao Wu, Kai Shu, Lu Cheng, and Huan Liu. 2025. From generation to judgment: Opportunities and challenges of LLM-as-a-judge. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 2757–2791, Suzhou, China. Association for Computational Linguistics.

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

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

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.

Ellie Pavlick and Tom Kwiatkowski. 2019. Inherent disagreements in human textual inferences. *Transactions of the Association for Computational Linguistics*, 7:677–694.

José Pombal, Dongkeun Yoon, Patrick Fernandes, Ian Wu, Seungone Kim, Ricardo Rei, Graham Neubig, and Andre Martins. 2025. M-prometheus: A suite of open multilingual llm judges. In *Second Conference on Language Modeling*.

Micah Rentschler and Jesse Roberts. 2026a. Reinforcement learning from meta-evaluation: Aligning language models without ground-truth labels. *arXiv preprint arXiv:2601.21268*.

Micah Rentschler and Jesse Roberts. 2026b. Reinforcement learning from meta-evaluation: Aligning language models without ground-truth labels. *Preprint*, arXiv:2601.21268.

Amy Rouillard, Sitwala Mundiab, Linda Camarab, Michael Cameron Gramaniec, Ziyaad Dangorc, Ismail Kallad, Shabir A Madhic, Kajal Morarc, Marlvin T Ncubec, Haroon Saloojeee, and 1 others. 2026. Can llms score medical diagnoses and clinical reasoning as well as expert panels? *arXiv preprint arXiv:2604.14892*.

Ananya B Sai, Akash Kumar Mohankumar, and Mitesh M Khapra. 2022. A survey of evaluation metrics used for nlg systems. *ACM Computing Surveys (CSUR)*, 55(2):1–39.

Andrew Sellergren, Sahar Kazemzadeh, Tiam Jaroensri, Atilla Kiraly, Madeleine Traverse, Timo Kohlberger, Shawn Xu, Fayaz Jamil, Cían Hughes, Charles Lau, and 1 others. 2025. Medgemma technical report. *arXiv preprint arXiv:2507.05201*.

Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, YK Li, Yang Wu, and 1 others. 2024. Deepseekmath: Pushing the limits of mathematical reasoning in open language models. *arXiv preprint arXiv:2402.03300*.

Hwanjun Song, Minseok Kim, Dongmin Park, Yooju Shin, and Jae-Gil Lee. 2022. Learning from noisy labels with deep neural networks: A survey. *IEEE transactions on neural networks and learning systems*, 34(11):8135–8153.

Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. Gemma 3 technical report. *Preprint*, arXiv:2503.19786.

Yidong Wang, Zhuohao Yu, Wenjin Yao, Zhengran Zeng, Linyi Yang, Cunxiang Wang, Hao Chen, Chaoya Jiang, Rui Xie, Jindong Wang, and 1 others. 2024. Pandalm: An automatic evaluation benchmark for llm instruction tuning optimization. In *International Conference on Learning Representations*, volume 2024, pages 43573–43593.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025a. Qwen3 technical report. *Preprint*, arXiv:2505.09388.

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, Junyang Lin, Kai Dang, and 23 others. 2025b. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Wen-wai Yim, Asma Ben Abacha, Zixuan Yu, Robert Doerning, Fei Xia, and Meliha Yetisgen. 2026. Morqa: Benchmarking evaluation metrics for medical open-ended question answering. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026)*, pages 5035–5054, Palma, Mallorca, Spain. European Language Resources Association (ELRA).

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.

Lianghui Zhu, Xinggang Wang, and Xinlong Wang. 2025. Judgelm: Fine-tuned large language models are scalable judges. In *International Conference on Learning Representations*, volume 2025, pages 51257–51296.

A DiagTrace Evaluation Criteria

Table 1 provides the detailed evaluation criteria, score ranges, and annotation guidelines used in DiagTrace in.

B Prompt Templates

All prompts used throughout the experiments follow a unified evaluation framework based on criterion-level medical reasoning assessment. We use two main prompting settings: (i) *with reference*, where the evaluator receives the gold expert diagnosis and reasoning, and (ii) *no-ref*, where evaluation relies solely on the clinical case and the generated prediction.

Prompts are bilingual (English and French) and automatically adapted to the language of the evaluated clinical case.

System prompt

```
Please act as a physician and evaluate the
     quality of the diagnosis and clinical
     reasoning provided by an AI assistant to
     the clinical case displayed below.

You will be given:
- a clinical case
- a predicted diagnosis and reasoning from a
     model

Your job is to evaluate the quality and
     medical soundness of the model-generated
     clinical reasoning according to the
     following criteria:

========================
EVALUATION CRITERIA
========================

[Internal Validity]
This criterion assesses the model's ability
     to extract and interpret the relevant
     elements of the clinical case that are
     useful for diagnosis. It also evaluates
     the model's capacity to translate
     certain semiological descriptions into
     medical terms. For example, a "purlish
     skin lesion that does not fade under
     pressure" should suggest purpura. The
     same applies for clinical measurements,
     for instance mentioning fever in the
     case of temperature of 38.5 C.
Scoring for this criterion:
- 5 points: Exhaustive and accurate
     interpretation of all data relevant to
     the diagnosis. The elements of the case
     are prioritized.
- 4 points: Incomplete analysis of the data,
     but sufficient to make the diagnosis.
     Some irrelevant elements are noted.
- 3 points: Incomplete analysis of the data,
     insufficient to make the diagnosis.
- 2 points: Analysis of the data leads to an
```

Criterion	Score Range	Description
Internal Validity	$[0, 5]$	Ability to extract and interpret all relevant clinical data from the vignette, including appropriate medical terminology and contextualization of numerical findings
External Validity	$[0, 3]$	Scientific accuracy of medical knowledge introduced beyond the case description; defaults to 3 when no external knowledge is invoked
Differential Diagnosis	$[0, 1]$	Generation of relevant diagnostic hypotheses organized by likelihood or severity
Logical Structure	$[0, 4]$	Coherence and organization of reasoning independently of content accuracy
Expression	$[0, 2]$	Absence of syntactic or semantic errors; penalizes responses in the wrong language
Final Diagnosis	$[0, 3]$	Accuracy of the predicted diagnosis relative to the expert reference

Table 1: DiagTrace evaluation criteria and scoring guidelines. All scores are integer-valued.

```
    incorrect diagnosis.
- 1 point: Paraphrases the clinical case
    without analyzing the data.
- 0 points: Hallucination or no response
    generated.

[External Validity]
This criterion assesses the validity of the
    medical knowledge the model draws upon
    during its reasoning, which is not
    initially included in the clinical case.
    Note that if the model does not refer to
    any external knowledge, a default score
    of 3 points will be given, in order to
    not penalize the absence of such
    arguments, which are not always
    necessary.
Scoring for this criterion:
- 3 points: Medical knwoledge is correct and
    useful for making the correct diagnosis.
- 2 points: Partially correct or outdated
    medical knowledge.
- 1 point: Medical knowledge is incorrect.
- 0 points: No response generated.

[Hypotheses/Differential Diagnoses]
This criterion assesses the model's ability
    to formulate hypotheses and suggest
    relevant differential diagnoses.
    Ideally, these should be prioritized in
    terms of likelihood and/or severity. If
    differential diagnoses are mentioned in
    the clinical reasoning of the Expert and
    not in the model response, the score
    will be 0. If the clinical case does not
    lend itself to any differential
    diagnosis and the model does not mention
    any, a default score of 1 point will be
    given to not penalize reasoning that
    remains valid in such a context.
Scoring for this criterion:
- 1 point: The model is able to suggest
    differential diagnoses in relevant
    cases. The diagnoses are explicit and
    prioritized.
- 0 points: No relevant differential
    diagnoses suggested

[Logic]
This criterion evaluates the overall flow of
    reasoning, regardless of the validity of
    the arguments presented. Particular
    attention is paid to the logical order
    of the elements mentioned (for example,
    clinical elements typically precede
    paraclinical ones), the absence of
    contradictions, and the exclusion of
    irrelevant statements that do not
    contribute to the reasoning.
Scoring for this criterion:
- 4 points: Coherent connection between
    mentioned elements, oriented toward the
    correct diagnosis
- 3 points: Coherent connection between
    mentioned elements, but oriented toward
    an incorrect diagnosis.
- 2 points: Incoherent connection of medical
    data, with contradictions present. Also
    applies if not all elements are
    mentioned in the clinical case but the
    reasoning is correct.
- 1 point: No connection between the visible
    medical data.
- 0 points: No response generated.

[Expression]
This criterion penalizes the presence of
    sentences with errors in expression,
    meaning, and/or syntax. It also
    penalizes cases where the model provides
    answers in English when French is
    expected.
Scoring for this criterion:
- 2 points: No syntactic errors or incorrect
    medical phrasing
- 1 points: Language errors or inappropriate
    word choice that do not alter the
    meaning of the sentence.
- 0 point: Error that alters the meaning of
    the text or no response generated.

[Final Diagnosis]
This criterion assesses the consistency
    between the final diagnosis proposed by
    the model and that of the expert. If the
    diagnosis is incomplete or insufficient,
    or if the terminology used is considered
    incorrect (or outdated), it will receive
    1 point. If it is fully accurate, it
    will receive 2 points. If the diagnosis
```

```
    is incorrect or if too many important
    elements are missing, the score will be
    0.
Scoring for this criterion:
- 3 points: Perfect diagnosis.
- 2 points: Correct diagnosis, but
    incomplete (syndromic diagnosis,
    non-lateralized, without specification
    of severity).
- 1 point: Incorrect diagnosis
- 0 points: No response generated.

========================
INSTRUCTIONS
========================

- Evaluate ONLY the predicted reasoning and
    diagnosis.
- Base your judgment solely on the clinical
    case and general medical knowledge.
- Be strict and consistent.
- Do NOT explain your answer.
- Do NOT output anything except a valid JSON.


========================
OUTPUT FORMAT (STRICT)
========================

{
  "scores": {
    "internal_validity": <int>,
    "external_validity": <int>,
    "differential_diagnosis": <int>,
    "logical_structure": <int>,
    "expression": <int>,
    "final_diagnosis": <int>
  }
}

========================
INPUT
========================
- Clinical case: {clinical_case}
- Predicted reasoning: {predicted_reasoning}
- Predicted diagnosis: {predicted_diagnosis}
```

Figure 8: System prompt for no-ref setting for English cases for zero-shot evaluation and training.

French and with-ref prompts can be found in the repository[4].

B.1 Evaluation and Training Prompts

The same prompting structure is reused consistently across zero-shot evaluation, SFT, and GRPO training. Prompts instruct the model to evaluate generated clinical reasoning across the six DiagTrace dataset criteria and output structured JSON scores only. The with-reference setting additionally provides the expert diagnosis and reasoning as calibration guidance for evaluation.

B.2 Meta-Evaluator Prompts

For the meta-evaluator RL setting, Qwen3-8B receives the clinical case, generated reasoning, predicted diagnosis, criterion definition, and predicted criterion-level score, and predicts whether the attributed score is clinically appropriate. More specifically, the evaluator answers a criterion-specific meta-question of the form:

System prompt

```
You are an expert medical education
    evaluator. You are given a clinical
    case, a predicted reasoning, and a
    score attributed to that reasoning for
    a specific evaluation criterion. Your
    task is to judge whether the attributed
    score is reasonable given the clinical
    case, the predicted reasoning, and the
    criterion definition.
Answer with Yes or No only. Do not explain
    your answer.

- Clinical case: {clinical_case}
- Predicted reasoning: {predicted_reasoning}
- Predicted diagnosis: {predicted_diagnosis}
The predicted reasoning received a score of
    {score}/{max_score} for final diagnosis.
- This criterion compares the predicted
    diagnosis with the reference diagnosis
    established by medical experts. It
    assesses whether the final diagnosis is
    appropriate given the clinical evidence
    and reasoning presented.
Is a score of {score}/{max_score} for final
    diagnosis reasonable given the clinical
    case and predicted reasoning above?

Answer with Yes or No only.
```

Figure 9: System prompt for no-ref setting for English cases given to the Meta Evaluator for the final diagnsis criterion.

French, with-ref and other criteria prompts can be found in the repository[5].

> *"Is a score of s/max_c for criterion c reasonable given the clinical case and predicted reasoning above?"*

The model outputs only *Yes* or *No*, and the first-token probability $P(\text{Yes})$ is used as the reward signal during RL.

Full prompting templates are released in the project repository.[6]

[4] https://anonymous.4open.science/r/Concord-F714/README.md

[5] https://anonymous.4open.science/r/Concord-F714/README.md

[6] https://anonymous.4open.science/r/Concord-F714/README.md

C Training Configurations

Table 2 summarizes all training configurations explored in this work.

D Meta-Evaluator Selection

Figure 10 presents the performance of all evaluated baseline judges across both *with-ref* and *no-ref* settings. Since the proposed meta-evaluation framework targets realistic deployment scenarios without access to gold references, the selection of the frozen teacher model was based primarily on the *without-reference* setting.

Although GPT-5.4 achieves the strongest overall performance, we intentionally focus on small open-weight models for practical and reproducibility reasons. Among these candidates, Qwen3-8B exhibits the most stable criterion-level behavior across evaluation dimensions, consistently maintaining strong performance on *internal validity*, *logical structure*, and *expression*, while also achieving competitive overall correlation. In contrast, several other open-weight models display more uneven behavior across criteria despite comparable total scores.

We therefore selected Qwen3-8B as the frozen meta-evaluator for RL-based supervision transfer, balancing evaluation quality, stability across criteria, and realistic open-weight deployment constraints.

E GRPO Initialization Ablation

We investigate the impact of GRPO initialization by comparing optimization starting from: (i) the original Gemma3-4B-it base model, and (ii) the best supervised checkpoint obtained during SFT, namely SFT-AW with reference information.

This ablation evaluates whether disagreement-aware supervised alignment provides a stronger initialization prior for subsequent RL optimization.

F Reward Engineering Details

F.1 Format Reward

In addition to score-based rewards, we include a lightweight formatting reward encouraging valid structured JSON outputs.

Formatting rewards are assigned as follows:

- 0.0: no JSON detected
- 0.25: JSON present but missing criteria keys
- 0.5: all criteria present but invalid values detected
- 1.0: fully valid JSON output

The formatting reward is weighted by 0.2 and combined with the main criterion-level reward.

F.2 Ordinal and Exact-Match Rewards

For criterion c, let $\hat{s}_c$ denote the score predicted by the judge, and let s_{c,e_1} and s_{c,e_2} denote the scores assigned by the two medical experts. All scores are normalized by the criterion-specific maximum score $s_{c,\max}$. The ordinal reward is defined as

$$r_c^{\text{ord}} = 1 - \frac{|\hat{s}_c - s_c|}{s_{c,\max}}$$

To encourage exact score prediction, we additionally include an exact-match bonus:

$$r_c^{\text{exact}} = \begin{cases} 0.5 & \text{if } \hat{s}_c = s_c \\ 0 & \text{otherwise} \end{cases}$$

The final criterion reward becomes:

$$r_c = r_c^{\text{ord}} + r_c^{\text{exact}}$$

The combined reward is normalized to remain within $[0, 1]$ across criteria.

For binary criteria ($s_{c,\max} = 1$), the exact-match bonus is omitted to avoid disproportionate reward scaling.

G Hyperparameters

G.1 Supervised Fine-Tuning Hyperparameters

All supervised fine-tuning experiments use LoRA adapters on top of Gemma3-4B-it. Models are trained using completion-only loss, where prompt tokens are masked and optimization is applied exclusively on generated completions.

Table 3 summarizes the SFT hyperparameters. All models are evaluated every 100 training steps, and the checkpoint achieving the best validation loss is selected for final evaluation.

G.2 GRPO Hyperparameters

All GRPO experiments are initialized either from Gemma3-4B-it or from the best SFT checkpoint depending on the configuration. Table 4 summarizes the reinforcement learning hyperparameters.

H Additional Baseline Analyses

H.1 Criterion-Level Baseline Analysis

Figure 10 reports criterion-level baseline performance across all evaluated models. We observe

Stage	Configuration	Variants
SFT	SFT-MA	with-reference / without-reference
	SFT-AW	with-reference / without-reference
GRPO	Standard (Base Init.)	with-reference / without-reference
	Standard (SFT Init.)	with-reference / without-reference
	Mod1	with-reference / without-reference
	Mod2	with-reference / without-reference
	Mod3	with-reference / without-reference
	Mod1 + Mod2	with-reference / without-reference
	Mod1 + Mod3	with-reference / without-reference
	Mod2 + Mod3	with-reference / without-reference
	Mod1 + Mod2 + Mod3	with-reference / without-reference
Meta-Evaluator RL	Qwen3-8B Meta-Evaluator	without-reference only

Table 2: Summary of all training configurations explored in Concord.

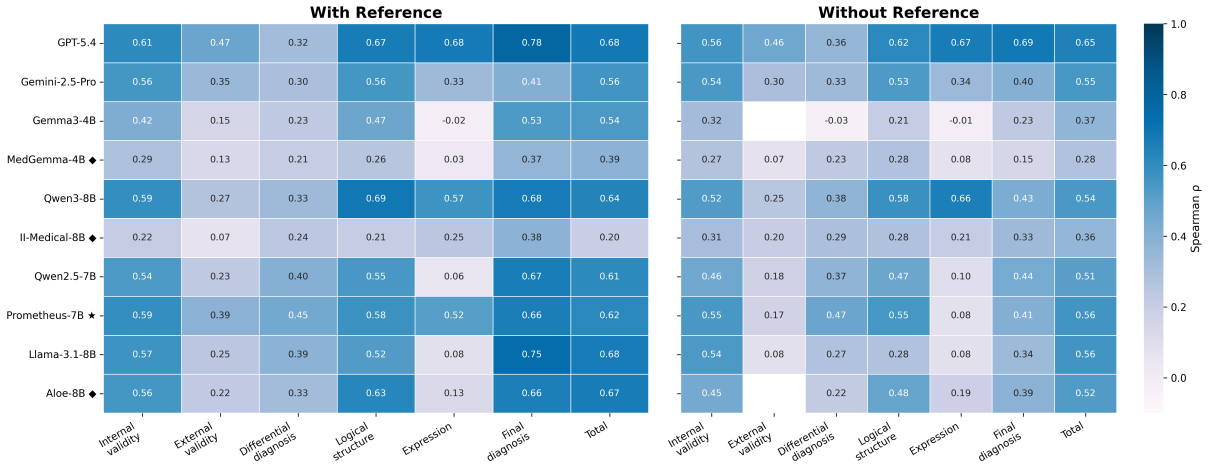

Figure 10: Criterion-level baseline performance across all evaluated LLM judges under *with-reference* and *without-reference* prompting settings. Values correspond to Spearman correlation between LLMs predicted scores and expert annotations for each evaluation criterion. Empty cells indicate cases where the correlation could not be computed because the model predicted a single constant value across all examples.

substantial variability across evaluation dimensions. Criteria associated with structured reasoning and final diagnosis generally achieve higher agreement with expert annotations than external validity or expression quality, reflecting the heterogeneous reliability of different evaluation dimensions.

Across nearly all models, *final diagnosis* and *logical structure* consistently achieve the highest correlations with expert annotations, while *external validity* remains the most difficult criterion. This behavior closely matches the IAA analysis presented in Section 3.3, suggesting that criteria with stronger expert consensus provide more stable supervision and evaluation signals.

We also observe that reference-guided prompting benefits some criteria more strongly than others. In particular, improvements are most visible for *logical structure*, *expression*, and *final diagnosis*, whereas gains remain comparatively limited for *external validity*. This suggests that access to the gold reasoning partially simplifies evaluation of clinically grounded reasoning dimensions but does not fully resolve challenges associated with subjective or knowledge-intensive criteria.

H.2 Language-Specific Baseline Analysis

Figure 11 reports baseline performance separately for French and English evaluation settings. Overall, models exhibit relatively consistent behavior across both languages, suggesting that task difficulty is driven more strongly by medical reasoning evaluation itself than by language-specific effects. Nevertheless, French evaluation settings tend to yield slightly stronger overall performance across

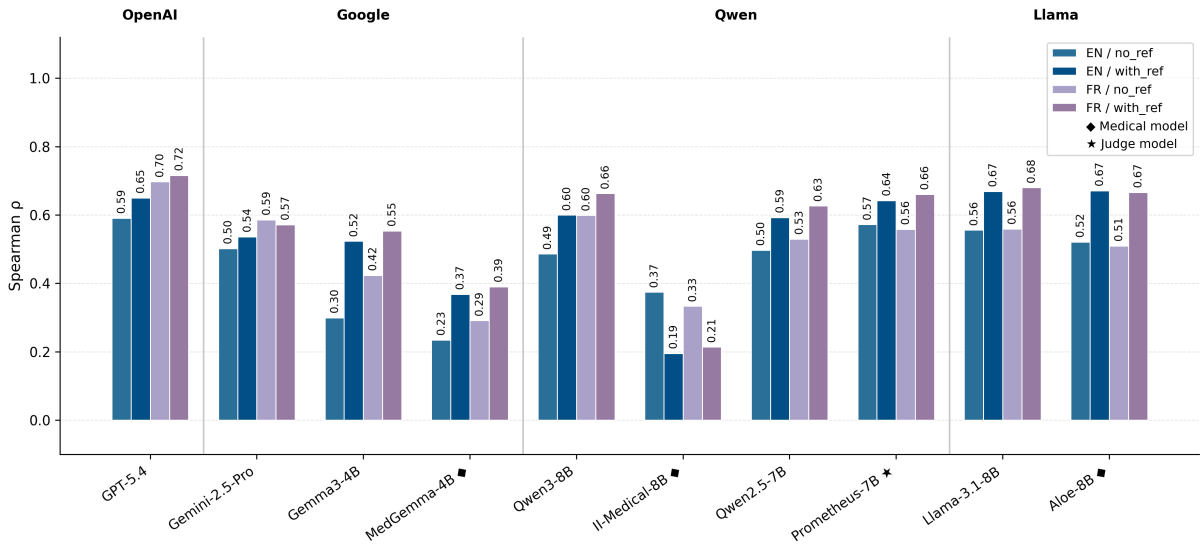

Figure 11: Language-specific baseline performance for all evaluated LLM judges. Results are reported separately for French and English evaluation settings under both *with-reference* and *without-reference* prompting conditions using average Spearman correlation across all criteria. Symbols indicate medical-specialized (♢) and judge-specialized (⋆) models.

Hyperparameter	Value
Epochs	10
Per-device batch size	2
Gradient accumulation	4
Learning rate	5×10^{-6}
Scheduler	Cosine
Warmup ratio	0.05
Weight decay	0.01
Optimizer	AdamW (fused)
bf16	True
Gradient checkpointing	True
Maximum sequence length	2048
LoRA rank (r)	16
LoRA alpha	32
Seed	42
GPU	1 × NVIDIA L40 (48GB)

Table 3: SFT hyperparameters.

several models. This trend may be partially explained by the higher IAA observed for multiple criteria in French annotations, leading to more consistent supervision and evaluation signals. Commercial and multilingual models generally exhibit smaller performance gaps across languages, indicating more stable cross-lingual evaluation capabilities. Interestingly, the relative impact of reference-guided prompting remains consistent across both languages, reinforcing the observation that reference information provides a stable evaluation scaffold independently of linguistic setting.

I Additional Disagreement-Aware GRPO Analysis

I.1 GRPO Initialization Analysis

Before introducing the proposed disagreement-aware GRPO modifications, we first analyze the impact of GRPO initialization. In particular, we compare standard GRPO initialized either from the original base model Gemma3-4B (*Std-GRPO (base)*) or from the disagreement-aware SFT checkpoint (*Std-GRPO (sft)*). Results for both *with-reference* and *without-reference* settings are reported in Figure 12. Overall, initializing GRPO from the disagreement-aware SFT model consistently produces stronger and more stable performance across nearly all evaluation criteria. Under the realistic *with-ref* setting, the overall Spearman correlation improves from 0.557 for *Std-GRPO (base)* to 0.552 for *Std-GRPO (sft)*. While the overall scores remain close, the criterion-level behavior differs substantially.

Most importantly, GRPO initialized from the SFT-AW checkpoint dramatically improves criteria associated with subjective reasoning quality and linguistic evaluation. For example, under the *no-ref* setting, *expression* increases from 0.311 to 0.612, while *logical structure* improves from 0.493 to 0.445 and *final diagnosis* remains competitive (0.505 vs 0.469). Similar trends are observed in the *with-reference* setting, where *Std-GRPO (sft)* achieves the strongest overall score (0.632).

These results suggest that disagreement-aware su-

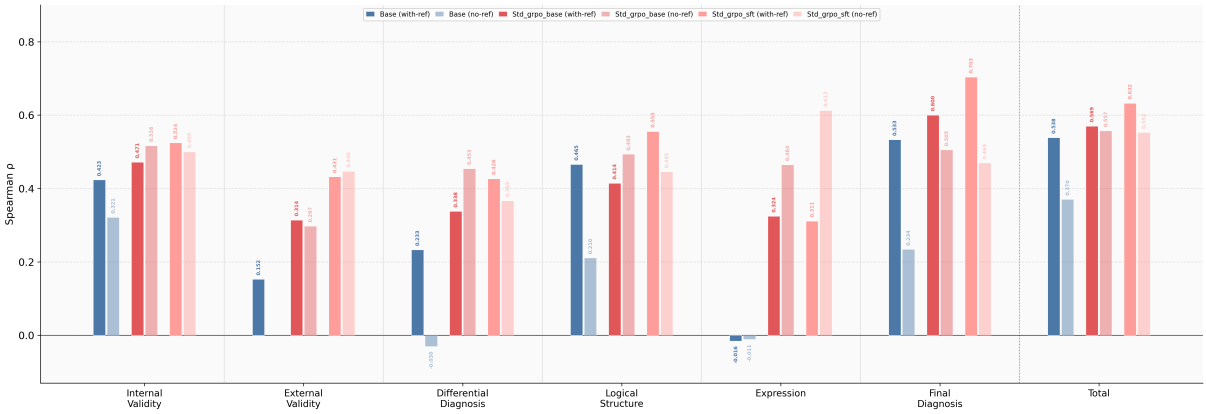

Figure 12: Comparison between standard GRPO initialized from the original base model (*Std-GRPO (base)*) and from the disagreement-aware SFT checkpoint (*Std-GRPO (sft)*).

Hyperparameter	Value
Epochs	5
Per-device batch size	8
Gradient accumulation	8
Learning rate	5×10^{-6}
Scheduler	Cosine
Warmup ratio	0.05
Weight decay	0.01
Optimizer	AdamW (fused)
bf16	True
Gradient checkpointing	True
Number of generations (G)	8
Maximum prompt length	2048
Maximum completion length	512
Temperature	0.9
Top-p	0.95
Format reward weight	0.2
LoRA rank (r)	16
LoRA alpha	32
λ (Mod1)	1.0
η (Mod3)	0.5
β (KL coefficient)	0.04
Seed	42
GPU	3 × NVIDIA H100 (80GB)

Table 4: GRPO hyperparameters.

pervised fine-tuning provides a substantially better initialization space for reinforcement learning optimization. Rather than learning directly from noisy ordinal rewards alone, GRPO benefits from starting from a model already partially aligned with annotator uncertainty and disagreement patterns. Based on these observations, all subsequent disagreement-aware GRPO experiments are initialized from the SFT-AW checkpoint rather than from the original base model.

I.2 Criterion-Level Entropy Analysis

Figure 13 reports criterion-level normalized entropy across all evaluated models on the full test set. As discussed in 5.2.3, disagreement-aware GRPO variants exhibit higher entropy, particularly on lower-agreement and more subjective evaluation criteria, reflecting increased sensitivity to annotation uncertainty.

I.3 With-Reference Evaluation Results

Figure 14 reports the *with-ref* evaluation results for all GRPO variants. In this setting, the judge receives access to the expert reference answer during evaluation, providing a less realistic but more controlled evaluation scenario. Overall, all reinforcement learning approaches substantially outperform both the base model and disagreement-aware SFT, confirming that policy optimization consistently improves alignment with expert annotations even when reference answers are available. Standard GRPO initialized from the SFT-AW checkpoint already increases the overall Spearman correlation from 0.615 to 0.632.

Compared to the *no-ref* setting, performance gains are generally larger across all criteria, particularly for *final diagnosis* and *logical structure*. This suggests that access to the expert reference provides a strong evaluation scaffold that simplifies reward optimization and reduces ambiguity during inference.

Among the proposed disagreement-aware variants, Mod1 and Mod3 provide the strongest individual improvements. Mod1 achieves strong performance on *expression* (0.246) and *final diagnosis* (0.674), while Mod3 improves *external validity* (0.401) and maintains competitive overall performance (0.606).

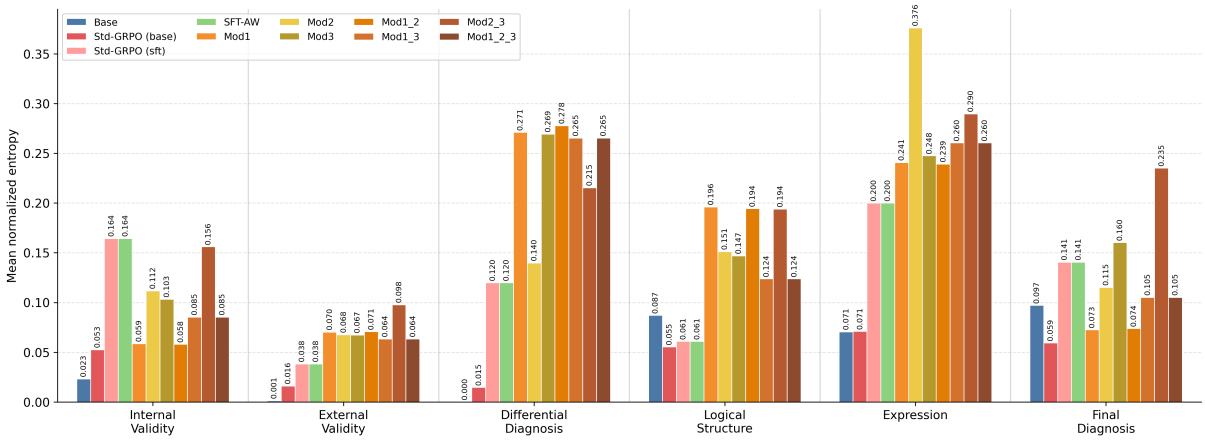

Figure 13: Criterion-level normalized entropy across all evaluated models on the whole test set.

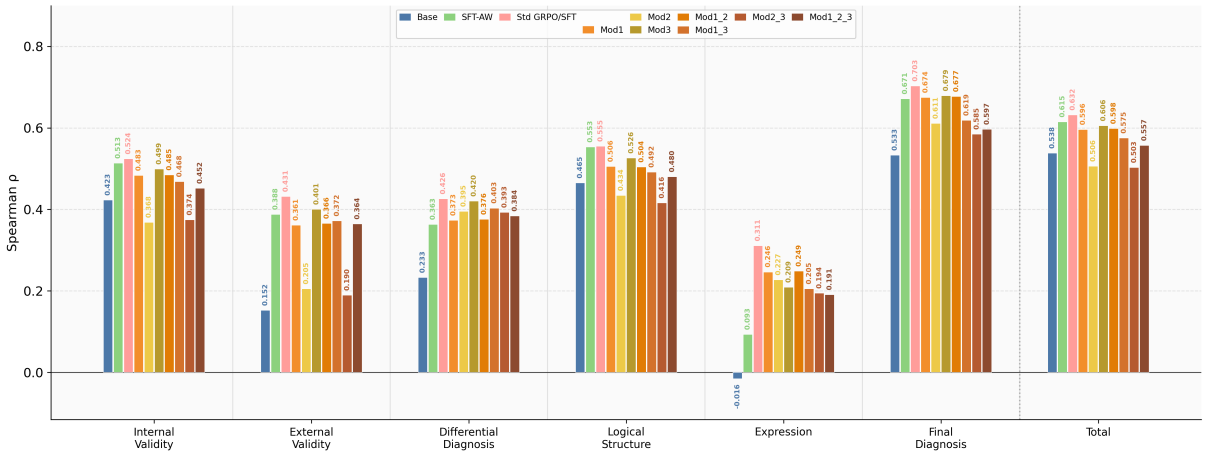

Figure 14: Performance of GRPO variants under the *with-ref* evaluation setting.

Interestingly, Mod1+Mod2 achieves the strongest *final diagnosis* correlation (0.679), whereas the complete Mod1+Mod2+Mod3 combination does not yield further gains, similarly to the observations made in the *no-ref* setting.

Nevertheless, the relative improvements between disagreement-aware variants remain smaller than in the realistic no-reference setting. This likely occurs because access to the reference answer already reduces part of the uncertainty associated with subjective clinical evaluation, partially masking the benefits of disagreement-aware optimization.

Overall, the *with-ref* results remain consistent with the main findings of the paper: reinforcement learning substantially improves medical LLM judges, while disagreement-aware optimization provides additional robustness and more stable criterion-level behavior.