AlignLLM: Addressing Short Answer Limitations in LLM-as-a-Judge Systems

Ramitha Abeyratne[1], Nirmalie Wiratunga[1], Kyle Martin[1] and Ikechukwu Nkisi-Orji[1]

[1]*Robert Gordon University, Aberdeen, Scotland*

Abstract

Large Language Models are increasingly used as automated evaluators of model-generated outputs, forming the basis of the LLMs-as-judges paradigm. AlignLLM adopts this perspective and reframes evaluation as a semantic alignment problem inspired by Case-Based Reasoning, measuring the consistency between reconstructed problem and solution representations across multiple judges. In this study, we investigate how answer length influences the stability and reliability of alignment-based evaluation within the AlignLLM framework. Using the SQuAD 1.1 dataset, we analyse alignment behaviour across answers of varying lengths using three alignment metrics: ILRAlign, WILRAlign, and Tuned-WILRAlign. Our findings show that very short answers produce comparatively lower and more unstable alignment scores due to insufficient semantic grounding during the reconstruction process. To address this limitation, we introduce a lightweight answer expansion mechanism that slightly enriches short responses while aiming to preserve their original meaning. Experimental results demonstrate that this approach improves alignment score stability within the AlignLLM framework, yielding an average improvement of 32.16% across evaluation settings and increasing sensitivity to correctness by 5.88%. These findings highlight the importance of response richness in alignment-based evaluation and suggest that modest answer expansion improves the robustness of LLM-based automated assessment.

Keywords

LLMs-as-Judges, Case-alignment, Text Expansion

1. Introduction

The growing use of Large Language Models (LLMs) to evaluate the outputs of other models has introduced new challenges for reliable automated assessment. In recent years, LLMs-as-judges frameworks have been proposed as a scalable alternative to human evaluation, where models are prompted to assess the quality of generated responses [1]. Although this paradigm reduces annotation costs and enables large-scale experimentation, individual LLM judges can be unreliable. Their judgments may vary due to prompt sensitivity, internal biases, or instability in reasoning, raising concerns about the robustness of single-model evaluation [2].

To address these limitations, recent research has explored the use of multiple LLM judges whose decisions are aggregated to produce more stable assessments. AlignLLM [3] adopts this multi-judge perspective and proposes a different method for combining judgments. Inspired by the principle of Case Alignment from Case-Based Reasoning (CBR) literature [4], the framework treats evaluation as an alignment problem rather than simple aggregation of scalar scores. Instead of directly comparing numerical ratings, AlignLLM reconstructs semantic representations of the problem and candidate solution from each judge's reasoning. These representations are embedded into a shared latent space, where the consistency of problem–solution relationships across judges can be analysed.

The core assumption is that high-quality answers produce consistent semantic interpretations across independent judges, resulting in stronger alignment among their reconstructed representations [3]. In contrast, weak or incorrect answers tend to produce inconsistent interpretations and weaker alignment, making semantic agreement an implicit indicator of answer quality. Because this mechanism relies on agreement rather than comparison with a reference response, it supports evaluation in settings where gold annotations are unavailable, incomplete, or subjective.

ICCBR Workshop on Memory-Based Reasoning for Responsible GenAI at ICCBR2026, August 15, 2026, Bremen, Germany

r.abeyratne@rgu.ac.uk (R. Abeyratne); n.wiratunga@rgu.ac.uk (N. Wiratunga); k.martin3@rgu.ac.uk (K. Martin); i.nkisi-orji@rgu.ac.uk (I. Nkisi-Orji)

While prior AlignLLM studies demonstrate promising performance on responses with moderate or substantial textual content [3, 5], the framework's behavior under semantically sparse conditions remains underexplored. In particular, very short answers may fail to provide sufficient representational richness for reliable reverse reconstruction, thereby weakening the semantic consistency assumptions that underpin alignment-based evaluation.

In this work, we investigate short-answer sparsity as a case representation problem within the AlignLLM framework, focusing on how limited answer length and insufficient semantic grounding affect alignment quality and robustness. Using the SQuAD 1.1 dataset [6], we analyse the relationship between answer variance, length, and alignment stability, highlighting how sparse responses can hinder the generation of reliable semantic representations during evaluation. To address this issue, we propose a lightweight answer enrichment mechanism that expands short answers prior to evaluation, restoring representational adequacy while preserving the original meaning, and enabling judges to produce richer semantic interpretations that improve the consistency and reliability of alignment-based scoring.

Our main contributions are summarised as follows:

- We identify semantic sparsity in short answers as a case representation limitation within alignment-based LLM-as-a-Judge systems.
- We provide an empirical analysis of how answer length affects reconstruction stability, alignment reliability, and correctness sensitivity.
- We introduce a lightweight answer enrichment mechanism that improves semantic grounding while preserving the original informational content of short answers.

The remainder of the paper is organised as follows. Section 2 reviews prior research related to LLM-based evaluation and alignment methods. Section 3 presents the proposed approach. Section 4 describes the experimental setup and reports the results. Finally, Section 5 concludes the paper and outlines future research directions.

2. Related Work

Case-Based Reasoning [4] provides a conceptual basis for framing evaluation as a structured reasoning task rather than a direct scoring problem. Within this paradigm, the AlignLLM framework models evaluation as an alignment process between reconstructed representations of problems and their corresponding solutions [3]. In traditional CBR systems, new problems are solved by retrieving and adapting solutions from previously observed cases [4], and AlignLLM adopts a similar perspective by treating each question–answer pair as a case. It reconstructs both problem and solution spaces using general-purpose LLMs combined with Retrieval-Augmented Generation (RAG) [3].

The effectiveness of such alignment-based systems depends fundamentally on the adequacy of case representation, as incomplete or weakly specified cases can impair retrieval, adaptation, and comparison quality [4]. Although the reconstruction process in AlignLLM may superficially resemble proportional analogies of the form a:b::c:d, the framework is more appropriately characterised as Case-Based Reasoning rather than Analogy-Based Reasoning. Classical analogy-based approaches focus on transferring relational knowledge between a source and target domain through structure mapping [7], whereas AlignLLM operates on question–answer cases and evaluates the consistency of reconstructed case representations produced by multiple judges [3]. The objective is not to derive new solutions through analogical transfer, but to assess the adequacy and alignment of problem–solution case representations, which is consistent with established CBR notions of case representation and case comparison [8, 4]. In the context of AlignLLM, short answers introduce an analogous challenge: semantically sparse cases may reduce reconstruction fidelity and destabilise inter-judge consistency. This positions answer length not merely as a surface linguistic property, but as a structural factor affecting representational sufficiency within alignment-driven evaluation.

Code available at: https://github.com/RAbeyratne/alignllm-text-expansion

In the AlignLLM framework, evaluation relies on a bidirectional generation mechanism [5]. First, the system reconstructs a representative question conditioned on a candidate answer through reverse generation by an LLM ensemble. This reconstructed problem representation is then used to generate corresponding answers, producing a distribution of plausible solutions. AlignLLM compares these generated representations to measure semantic alignment between reconstructed problem and solution spaces, with the resulting alignment score serving as an indirect indicator of response quality without requiring explicit reference answers.

AlignLLM was initially introduced for domains such as legal question answering, where reliable gold-standard annotations are often limited [3]. By aggregating reasoning outputs from multiple LLM judges, the framework demonstrated greater robustness than single-model evaluators and other unsupervised approaches. Empirical results showed stronger correlations with established ground-truth metrics, suggesting that alignment-based aggregation can serve as an alternative to traditional reference-based evaluation. The compatibility between CBR and reasoning-intensive domains has also been demonstrated in earlier applications, particularly in legal and medical decision-support systems [9].

Subsequent work expanded the empirical analysis of AlignLLM and examined factors affecting its performance [5]. These studies evaluated the framework on general-domain question answering datasets and analysed how contextual information influences the generation process. Incorporating contextual grounding improved evaluation reliability, while assigning greater weight to the reconstructed solution space (9:1 weighting ratio) produced the strongest results. Additionally, the Weighted ILRAlign metric was introduced to aggregate judge responses, with an inverse-distance decay mechanism providing the most effective aggregation strategy.

Beyond alignment-based evaluation, research has examined how properties of generated text influence the reliability of automatic evaluation methods [10]. Both reference-based and reference-free metrics vary depending on the richness of generated text [11]. Short responses often contain limited semantic information, reducing the effectiveness of similarity-based metrics and increasing variability in model judgments [12]. Studies on LLM-based evaluators further show that evaluation quality depends on the reasoning signals available to the judging model, where richer contextual explanations produce more stable and human-aligned outcomes [2]. These findings highlight that the amount of textual information available to evaluators significantly affects the reliability of automated assessment.

The challenge of insufficient textual information has motivated a broader line of research on text and answer expansion. Across question answering, information retrieval, and Natural Language Generation, researchers have investigated methods for augmenting concise responses with additional semantic content in order to improve downstream tasks [13, 14]. These approaches are based on the observation that short texts often fail to expose sufficient semantic signals for reliable processing, motivating the generation of expanded representations that better capture the underlying intent and meaning of the original response. The same principle is relevant to automated evaluation, where semantically sparse answers may provide insufficient evidence for stable judgment.

To address sparse textual inputs, several studies explore methods for enriching generated responses before downstream processing. Techniques such as explanation generation, paraphrasing, and reasoning augmentation introduce additional semantic context that can improve interpretability and evaluation reliability. For example, Chain-of-Thought prompting encourages models to generate intermediate reasoning steps, expanding concise answers into richer explanatory representations [15]. Similarly, Retrieval-Augmented Generation (RAG) incorporates external knowledge sources to augment outputs with contextual information [16]. While these approaches produce more informative semantic structures, RAG-based methods are less suitable in our setting because external knowledge may alter the meaning of the expanded answer. Instead, a simpler expansion function can be used to elaborate answers while preserving their original meaning for subsequent evaluation.

Moreover, our work relates to a broader literature on the reliability of automated evaluation methods. Prior studies have shown that both traditional semantic similarity metrics and LLM-as-a-Judge systems can be sensitive to factors such as response length, reasoning depth, contextual richness, and evaluator bias, which may affect score stability and discriminative power [1, 2, 10, 12]. Our investigation of short-answer sparsity contributes to this broader discussion by examining how limited semantic content

influences evaluation robustness and how lightweight enrichment strategies can improve the quality of evaluation signals.

3. Methodology

3.1. Alignment Metrics Based on Inter-LLM Reconstruction

We adopt the Inter-LLM Reconstruction framework from AlignLLM [3] to quantify semantic alignment between problem and solution spaces. Given a question Q, a generator produces an answer $\bar{A}$ from a simulated RAG context S.

When an answer is semantically sparse (e.g., a single-word response), it may not provide sufficient representational grounding for stable reverse reconstruction. To address this, we introduce a constrained expansion function that enriches the answer by combining task-specific information from the original question with the parametric knowledge of an auxiliary LLM. In this formulation, the question provides immediate contextual anchors, while the model contributes only minimal semantic elaboration required to transform underspecified responses into richer case representations. Importantly, this process preserves the original meaning of the answer while increasing contextual density rather than introducing new information. Formally, when $\bar{A}$ lacks detail, it is expanded via a prompt template $\lambda(\cdot)$ and a model $M(\cdot)$.

$$\bar{A} = \begin{cases} M(\lambda(\bar{A}, Q)) & \text{if insufficient,} \\ \bar{A} & \text{otherwise.} \end{cases}$$

The resulting answer, either original or expanded, is used for reconstruction.

A set of N judge models reconstruct reverse questions Q_i' from $\bar{A}$ and corresponding answers A_i' from Q_i', producing pairs (Q_i', A_i') preserving the semantic relation of $(Q, \bar{A})$.

Contextual embeddings are extracted from the generator $m(\cdot)$ at selected layers $m_l(\cdot)$:

$$\vec{E}_Q = m_{l_Q}(Q), \quad \vec{E}_{\bar{A}} = m_{l_A}(\bar{A}), \quad \vec{E}_{Q_i'} = m_{l_Q}(Q_i'), \quad \vec{E}_{A_i'} = m_{l_A}(A_i'),$$

where l_Q and l_A denote layers for questions and answers. Then these are represented in two spaces, as denoted by Figure 1. Afterwards, similarities are computed via cosine similarity function.

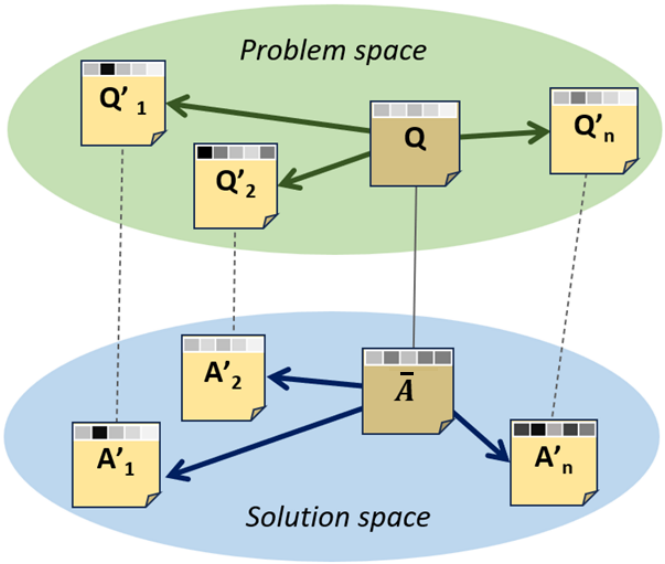

Figure 1: Case representation and alignment for reverse generated question-answer pairs

3.1.1. ILRAlign

The Inter-LLM Reconstruction Alignment score combines problem- and solution-space similarity equally:

$$\text{ILRAlign} = \frac{1}{N}\sum_{i=1}^{N}\big(\text{Sim}(\vec{E}_Q, \vec{E}_{Q_i'}) + \text{Sim}(\vec{E}_{\tilde{A}}, \vec{E}_{A_i'})\big), \tag{1}$$

3.1.2. WILRAlign

To mitigate inconsistent judges, weights are normalised using problem-space similarity:

$$w_i = \frac{\text{Sim}(\vec{E}_Q, \vec{E}_{Q_i'})}{\sum_{j=1}^{N}\text{Sim}(\vec{E}_Q, \vec{E}_{Q_j'})}. \tag{2}$$

The weighted alignment is:

$$\text{WILRAlign} = \sum_{i=1}^{N} w_i\big(\alpha\,\text{Sim}(\vec{E}_Q, \vec{E}_{Q_i'}) + \beta\,\text{Sim}(\vec{E}_{\tilde{A}}, \vec{E}_{A_i'})\big), \tag{3}$$

with $\alpha : \beta = 1 : 1$.

3.1.3. Tuned-WILRAlign

Following [5], we prioritise solution-space consistency with $\alpha : \beta = 1 : 9$, using inverse-distance weighting:

$$w_i = \frac{\frac{1}{1-\text{Sim}(\vec{E}_Q, \vec{E}_{Q_i'})+10^{-6}}}{\sum_{j=1}^{N}\frac{1}{1-\text{Sim}(\vec{E}_Q, \vec{E}_{Q_j'})+10^{-6}}}. \tag{4}$$

4. Effect of Answer Length and Expansion on Alignment

This section evaluates alignment-based evaluation methods using a human-generated dataset, with a particular focus on the role of answer length and the impact of answer expansion.

We use the SQuAD 1.1 training split, which contains 87,599 question–answer pairs across 442 titles. To reduce imbalance across contexts, we first select titles containing no more than 20 questions. From each selected title, we sample 10 instances while explicitly enforcing variation in answer length. Specifically, we group answers into three categories: minimum length (single-word answers), medium length, and maximum length. This results in a total of $20 \times 3 \times 10 = 600$ (titles x answer-lengths x instances-per-answer-length) question–answer pairs, which form the SQuAD[1] test set used in our evaluation.

4.1. Effect of Answer Length

First, we evaluate alignment using ILRAlign, WILRAlign, and Tuned-WILRAlign. Since ground-truth answers are used, higher alignment scores are generally expected because correct answers should preserve stronger semantic consistency across reverse reconstruction cycles. Although perfect alignment (1.0) may not always be achievable due to model variance, lexical ambiguity, and reconstruction noise, relative reductions in alignment remain informative indicators of representational weakness. Following prior work [3], we use four 7 billion-parameter models (Llama, Gemma, Mistral, and Falcon) and adopt a leave-one-out strategy, where three models act as judges and the remaining model serves as the embedding model.

[1] https://huggingface.co/datasets/Ramitha/squad-master-600

Table 1
Alignment scores across answer-length groups. Short answers exhibit lower and more variable alignment scores.

Answer length	Model (m)	ILRAlign	WILRAlign	Tuned-WILRAlign
Minimum	Llama	0.3980	0.4005	0.1573
	Falcon	0.9673	0.9673	0.9467
	Gemma	0.7686	0.7688	0.6272
	Mistral	0.4892	0.4951	0.3312
Medium	Llama	0.5680	0.5706	0.4894
	Falcon	0.9853	0.9853	0.9790
	Gemma	0.8880	0.8882	0.8508
	Mistral	0.5891	0.5942	0.5087
Maximum	Llama	0.6135	0.6170	0.5701
	Falcon	0.9877	0.9877	0.9831
	Gemma	0.9111	0.9113	0.8897
	Mistral	0.6168	0.6232	0.5586

Table 1 shows a clear relationship between answer length and alignment quality. The minimum-length group achieves the lowest average score (0.6098), while medium and maximum lengths reach 0.7414 and 0.7725, respectively. In addition to lower averages, short answers exhibit high variance across models. This instability arises because both forward and reverse generation rely on sufficient contextual grounding. Single-word answers provide limited semantic information, increasing the likelihood of hallucination during reconstruction and reducing alignment reliability.

4.2. Answer Expansion

To address the limitation of semantic information, we perform answer expansion on the 200 instances with minimum-length answers. The goal is to slightly increase answer length while preserving meaning, thereby improving contextual grounding. The following prompt template (λ) is employed for this task.

Listing 1: Prompt for Answer Expansion

```
You are given a question and a short answer.
Your task is to slightly expand the answer by adding a small amount of relevant detail from the
    question.
Keep the response concise. Add approximately 8 to 12 additional words.
Do NOT change the meaning of the original answer.
Do NOT make it long or explanatory.
Do NOT output the original question.
Do NOT add more details which are not in the answer or question.

Question: "{0}"
Short Answer: "{1}"

Expanded Answer:
```

We generate expansions using three models from different families and of varying sizes: SmolLM2-1.7B, Phi-3-mini-3.8B, and Qwen-7B. Alignment evaluation is then repeated on these expanded answers. To avoid evaluation bias, the expansion models are kept distinct from the alignment generator models.

As shown in Table 2, all methods benefit from expansion, with consistent positive gains observed across evaluation settings. ILRAlign and WILRAlign show similar improvements in performance, indicating that both approaches respond comparably to the addition of expanded context, while Tuned-WILRAlign achieves substantially larger improvements, suggesting greater sensitivity to richer input representations. The average improvement across the dataset is reported as 32.16%, highlighting the

Table 2
Percentage improvement in alignment after answer expansion. Tuned-WILRAlign benefits the most from added context.

Model (m)	ILRAlign	WILRAlign	Tuned-WILRAlign
Llama	67.34%	67.37%	298.22%
Falcon	2.43%	2.43%	4.37%
Gemma	22.72%	22.71%	49.68%
Mistral	33.28%	32.84%	82.07%

overall effectiveness of the expansion strategy.

Across expansion models, improvements remain consistent: 32.02% (SmolLM2-1.7B), 35.73% (Phi-3-mini-3.8B), and 28.73% (Qwen-7B), with only moderate variation between models. This consistency indicates that the primary driver of improvement is the addition of contextual information rather than model-specific characteristics, reinforcing the robustness of the approach across different generation backbones.

The observed improvements suggest that answer expansion functions similarly to contextual augmentation in AlignLLM's broader RAG-inspired design [3, 5]. By enriching sparse responses with question-conditioned semantic detail, expansion increases representational adequacy prior to reverse reconstruction. This supports the framework's core assumption that alignment quality depends on sufficient contextual grounding. In effect, answer expansion compensates for semantic under-specification in short answers, enabling judge models to construct more stable problem–solution spaces. The consistency of gains across multiple expansion models further suggests that these improvements derive less from model-specific generation quality than from the structural addition of contextual information itself.

4.3. Sensitivity to Answer Correctness

To further investigate the behavior of minimum-length answers, we construct an additional dataset variant. Specifically, we manually inspect the 200 instances with the shortest answers and replace each original answer with an incorrect answer of the same format for the corresponding question. For example, numerical answers are replaced with different numbers, while named entities are replaced with incorrect names. This process yields a modified test set containing intentionally incorrect answers, which we refer to as SQuAD-min-negative[2].

To examine robustness and sensitivity, we construct six test sets per expansion model by varying the proportion of correct answers (0%, 20%, 40%, 60%, 80%, and 100%), with each set containing 200 samples. Controlled sampling ensures that no instance appears as both correct and incorrect within the same test set. This setup enables a systematic evaluation of how alignment scores respond to changes in answer correctness. We then apply our alignment methods for benchmarking.

Figure 2 and Table 3 together show that alignment scores increase monotonically with correctness across all methods. In contrast, without expansion, the curves remain relatively flat, indicating weak sensitivity to correctness. This behaviour is reflected in the low slope values, confirming the limited discriminative power of the non-expanded setting. Additionally, alignment scores generally begin at higher values when expansion is applied.

With expansion, the alignment curves generally become steeper and more linear, indicating a stronger relationship between correctness and alignment score. Averaged across all settings, the slope increases from 0.0284 without expansion to approximately 0.0301 with expansion, corresponding to a 5.88% improvement. However, ILRAlign and WILRAlign exhibit slight reductions in slope under expansion, suggesting that answer expansion provides little benefit for these methods and may even slightly reduce their sensitivity to correctness. In contrast, Tuned-WILRAlign consistently demonstrates the highest

[2]https://huggingface.co/datasets/Ramitha/squad-master-200-least-negative

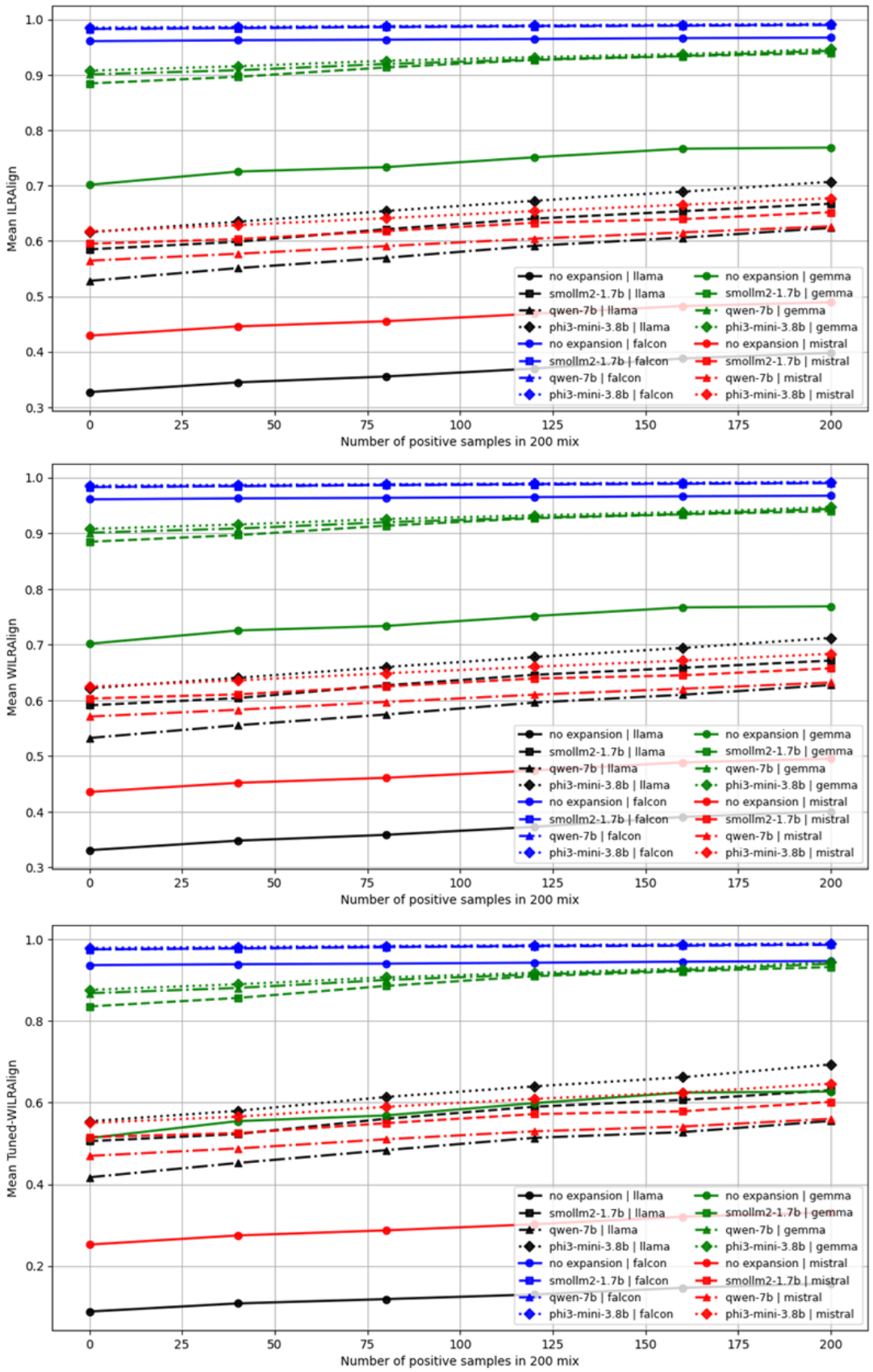

Figure 2: Mean alignment scores as a function of correctness. Answer expansion produces steeper and more linear trends, indicating improved sensitivity.

Table 3
Slope of alignment curves. Higher values indicate stronger sensitivity to correctness.

Model (m)	Base Model	ILRAlign	WILRAlign	Tuned-WILRAlign
No expansion	Falcon	0.0032	0.0032	0.0050
	Gemma	0.0341	0.0341	0.0579
	LLaMA	0.0356	0.0350	0.0336
	Mistral	0.0303	0.0301	0.0391
Phi3-mini-3.8B	Falcon	0.0034	0.0034	0.0056
	Gemma	0.0191	0.0192	0.0333
	LLaMA	0.0455	0.0451	0.0692
	Mistral	0.0302	0.0295	0.0483
Qwen-7B	Falcon	0.0035	0.0035	0.0055
	Gemma	0.0214	0.0214	0.0361
	LLaMA	0.0475	0.0473	0.0677
	Mistral	0.0314	0.0308	0.0454
SmolLM2-1.7B	Falcon	0.0038	0.0038	0.0061
	Gemma	0.0286	0.0286	0.0502
	LLaMA	0.0426	0.0417	0.0645
	Mistral	0.0291	0.0277	0.0438

sensitivity to correctness, with its average slope increasing by approximately 16.93% under expansion. This result indicates that Tuned-WILRAlign is substantially more responsive to the additional contextual information introduced through answer expansion.

Overall, the results demonstrate a clear and consistent progression: short answers produce weak alignment signals, answer expansion improves stability and alignment quality, and expanded representations enhance sensitivity to correctness. The limited variation across different expansion models further suggests that the primary source of improvement arises from the introduction of additional contextual information rather than from the specific expansion model itself.

5. Conclusion

This study examined short-answer sparsity as a representational sufficiency challenge within AlignLLM's alignment-based evaluation framework. Our findings show that semantically sparse answers weaken reconstruction fidelity, reduce alignment stability, and diminish sensitivity to correctness because they provide insufficient grounding for reliable case reconstruction. Viewed through a CBR lens, this suggests that case adequacy is a critical prerequisite for stable alignment-oriented evaluation.

To address this limitation, we introduced a lightweight answer expansion function that slightly enriches short responses while preserving their original meaning. Experiments using three expansion models from different LLM families and parameter sizes produced consistent improvements, increasing alignment scores by an average of 32.16%. The similar gains across models suggest that the effectiveness of the approach is largely independent of the specific expansion model.

To further evaluate robustness, we constructed controlled test sets by systematically varying the proportion of correct and intentionally incorrect answers, where incorrect variants were created by replacing the ground-truth answer with contextually plausible but incorrect alternatives. Using these datasets, we analysed the sensitivity of alignment scores to correctness. Results show that sensitivity improves comparatively with expansion, with a 5.88% increase in the average slope.

Overall, these results indicate that modest answer expansion can reliably enhance short responses, improve the stability of alignment-based evaluation, and strengthen sensitivity to correctness. Future work will investigate adaptive expansion strategies and explore whether similar enrichment techniques

can further improve alignment-based evaluation in other datasets and generation tasks.

6. Declaration on Generative AI

Large Language Models were integrated into this research primarily as components of the experimental pipeline for generating and evaluating textual outputs relevant to the study objectives. Beyond this experimental role, AI-assisted tools were used only for limited language refinement tasks such as spelling correction and formatting support. The core intellectual contributions of this work, including research design, methodological development, analysis, and interpretation, were carried out manually by the authors. All reported results and conclusions were independently verified by the authors, who accept full responsibility for the accuracy and integrity of the final manuscript.

References

[1] H. Li, Q. Dong, J. Chen, H. Su, Y. Zhou, Q. Ai, Z. Ye, Y. Liu, Llms-as-judges: A comprehensive survey on llm-based evaluation methods, 2024. arXiv:2412.05579.

[2] L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li, E. P. Xing, H. Zhang, J. E. Gonzalez, I. Stoica, Judging llm-as-a-judge with mt-bench and chatbot arena, in: 37th NIPS, NIPS '23, Curran Associates Inc., Red Hook, NY, USA, 2023.

[3] R. Abeyratne, N. Wiratunga, K. Martin, I. Nkisi-Orji, L. Jayawardena, Alignllm: Alignment-based evaluation using ensemble of llms-as-judges for q&a, in: Case-Based Reasoning Research and Development: 33rd International Conference, ICCBR 2025, Biarritz, France, June 30–July 3, 2025, Proceedings, Springer-Verlag, Berlin, Heidelberg, 2025, p. 21–36. doi:10.1007/978-3-031-96559-3_2.

[4] A. Aamodt, E. Plaza, Case-based reasoning: foundational issues, methodological variations, and system approaches, AI Commun. 7 (1994) 39–59.

[5] R. Abeyratne, N. Wiratunga, K. Martin, I. Nkisi-Orji, Design principles for alignment-based evaluation of large language models in reference-scarce domains, 2026. doi:10.2139/ssrn.6153652.

[6] P. Rajpurkar, J. Zhang, K. Lopyrev, P. Liang, SQuAD: 100,000+ questions for machine comprehension of text, in: J. Su, K. Duh, X. Carreras (Eds.), Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Austin, Texas, 2016, pp. 2383–2392. URL: https://aclanthology.org/D16-1264/. doi:10.18653/v1/D16-1264.

[7] M. Keane, Analogical mechanisms, Artificial Intelligence Review 2 (1988) 229–251.

[8] D. B. Leake, Case-Based Reasoning: Experiences, Lessons, and Future Directions, AAAI Press / The MIT Press, Menlo Park, CA, 1996.

[9] M. Ahmed, S. Begum, E. Olsson, N. Xiong, P. Funk, Case-Based Reasoning for Medical and Industrial Decision Support Systems, volume 305, 2010, pp. 7–52. doi:10.1007/978-3-642-14078-5_2.

[10] A. Celikyilmaz, E. Clark, J. Gao, Evaluation of text generation: A survey, arXiv preprint arXiv:2006.14799 (2020). URL: https://arxiv.org/abs/2006.14799.

[11] A. R. Fabbri, W. Kryściński, B. McCann, C. Xiong, R. Socher, D. Radev, SummEval: Re-evaluating summarization evaluation, Transactions of the Association for Computational Linguistics 9 (2021) 391–409. URL: https://aclanthology.org/2021.tacl-1.24/. doi:10.1162/tacl_a_00373.

[12] T. Zhang, V. Kishore, F. Wu, K. Q. Weinberger, Y. Artzi, Bertscore: Evaluating text generation with BERT, in: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020, OpenReview.net, 2020. URL: https://openreview.net/forum?id=SkeHuCVFDr.

[13] R. Nogueira, W. Yang, J. J. Lin, K. Cho, Document expansion by query prediction, ArXiv abs/1904.08375 (2019). URL: https://api.semanticscholar.org/CorpusID:119314259.

[14] L. Wang, N. Yang, F. Wei, Query2doc: Query expansion with large language models, in: H. Bouamor, J. Pino, K. Bali (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore, 2023, pp. 9414–9423. URL: https://aclanthology.org/2023.emnlp-main.585/. doi:10.18653/v1/2023.emnlp-main.585.

[15] X. Wang, J. Wei, D. Schuurmans, Q. V. Le, E. H. Chi, S. Narang, A. Chowdhery, D. Zhou, Self-consistency improves chain of thought reasoning in language models, in: The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023, OpenReview.net, 2023. URL: https://openreview.net/forum?id=1PL1NIMMrw.

[16] Y. Gao, Y. Xiong, X. Gao, K. Jia, J. Pan, Y. Bi, Y. Dai, J. Sun, M. Wang, H. Wang, Retrieval-augmented generation for large language models: A survey, 2024. URL: https://arxiv.org/abs/2312.10997. arXiv:2312.10997.