CalibJudge: Calibrated LLM-as-a-Judge for Multilingual RAG with Uncertainty-Aware Scoring

Chenfeiyu Wen , Ao Zhu , Runkun Long , Hejun Huang , Junjie Jiang , Chi Shing Lee *

Posted Date: 17 March 2026

doi: 10.20944/preprints202603.1324.v1

Keywords: LLM-as-a-Judge; multilingual NLP; calibration; RAG evaluation; uncertainty quantification

 doi:10.20944/preprints202603.1324.v1

Article

CalibJudge: Calibrated LLM-as-a-Judge for Multilingual RAG with Uncertainty-Aware Scoring

Chenfeiyu Wen [1], Ao Zhu [2], Runkun Long [3], Hejun Huang [4], Junjie Jiang [5] and Chi Shing Lee [6,*]

[1] New York University, New York, USA
[2] University of Pennsylvania, Philadelphia, USA
[3] Washington University in St. Louis, St. Louis, USA
[4] University of Michigan, Ann Arbor, USA
[5] Illinois Institute of Technology, Chicago, USA
[6] Hunter College, New York, USA
* Correspondence: chislee0708@gmail.com

Abstract

Large Language Models (LLMs) serving as automatic evaluators (LLM-as-a-Judge) have become essential for assessing Retrieval-Augmented Generation (RAG) systems. However, in multilingual settings, these judges exhibit significant calibration drift across languages, producing scores that are neither comparable nor aligned with human judgments. We present CalibJudge, a post-hoc calibration framework that addresses this challenge through: (1) language-specific temperature scaling, (2) uncertainty quantification, and (3) selective abstention. We evaluate CalibJudge on the MEMERAG benchmark covering five languages. Our experiments demonstrate that CalibJudge improves correlation with human annotations by up to 21.3% relative improvement in Kendall's while reducing cross-lingual fairness gaps by 42% and achieving 88% balanced accuracy at 70% coverage.

Keywords: LLM-as-a-Judge; multilingual NLP; calibration; RAG evaluation; uncertainty quantification

I. INTRODUCTION

As RAG systems are deployed globally, evaluating their output quality across languages becomes critical. The LLM-as-a-Judge paradigm [1] has gained traction as a scalable alternative to human evaluation, achieving over 80% agreement with human preferences in English. However, extending this to multilingual settings reveals a fundamental challenge: LLM judges exhibit significant calibration drift across languages, producing scores that are neither internally consistent nor comparable across linguistic boundaries.

We introduce CalibJudge, a post-hoc calibration framework for multilingual RAG evaluation. Our contributions are: (1) **Language-Specific Temperature Scaling** that learns separate calibration parameters for each language [2], (2) **Uncertainty-Aware Scoring** that identifies unreliable predictions, and (3) **Cross-Lingual Fairness Analysis** showing calibration reduces disparities. We evaluate on MEMERAG [3], covering English, German, Spanish, French, and Hindi, demonstrating consistent improvements while maintaining favorable reliability-coverage trade-offs.

 doi:10.20944/preprints202603.1324.v1

Ⅱ. Methodological Foundations

The Large language models used as evaluators are not treated in this work as neutral scoring devices, but as conditional predictors whose outputs reflect latent classification boundaries, prompt sensitivity, and language-dependent confidence distributions. This methodological starting point is directly grounded in empirical analyses showing that LLM-as-a-Judge systems behave much like task-specific classifiers rather than universally calibrated judges [4]. That observation is central to CalibJudge, because it justifies the decision to model judge outputs as scores requiring post-hoc correction instead of as directly comparable judgments. The multilingual evaluation setting further strengthens this need. Benchmark work on multilingual retrieval-augmented generation shows that response quality assessment varies across languages in both distribution and ranking behavior [5], while broader studies on fairness and bias in large language models establish that output discrepancies across groups are not incidental noise but systematic statistical shifts that can persist unless explicitly corrected [6]. CalibJudge inherits this line of reasoning by introducing language-specific temperature scaling: instead of searching for a single global calibration rule, it assumes that each language induces its own score distortion pattern and therefore requires its own calibration parameterization.

This calibration view is further supported by methods that address representation-level stability and semantic consistency in language models. Explainable representation learning demonstrates that model decisions can be understood as structured latent signals rather than opaque outputs [7], which motivates our treatment of judge scores as analyzable confidence-bearing representations. Robust long-context reasoning methods based on self-constructed negative samples show that model outputs become more dependable when instability is explicitly exposed and corrected [8]. Iterative self-questioning with semantic calibration provides an even closer methodological parallel: it treats prediction quality as something that can be improved by identifying internal inconsistency and applying targeted semantic correction [9]. The uncertainty component of CalibJudge is likewise not introduced as an auxiliary heuristic, but follows a distinct methodological lineage. Risk-aware summarization with uncertainty quantification shows that language model outputs should be paired with confidence estimates when they are used in downstream decision pipelines [10]. Prompt-controlled abstraction methods further reveal that the form of the generated judgment can vary with prompting structure, meaning that score reliability cannot be separated from generation conditions [11]. Structured prompt optimization through semantic alignment [12] and multi-stage alignment distillation for semantic consistency [13] both reinforce the same methodological lesson: a model may produce fluent judgments while still exhibiting latent misalignment, and this misalignment must be measured rather than ignored. CalibJudge borrows from this family of work by converting raw judge scores into uncertainty-aware predictions. The role of uncertainty quantification in our framework is therefore methodological, not decorative: it functions as the bridge between calibrated scores and trustworthy use, allowing the system to distinguish high-confidence evaluative agreement from unstable or weakly supported judgments.

The selective abstention mechanism in CalibJudge is built on a broader trust-and-decision methodology rather than only on score thresholding. Contextual trust evaluation methods show that robust decisions require explicit modeling of when a system should rely on its own outputs and when confidence is insufficient [14]. Trust orchestration under minimal necessary information [15], self-reflective multi-agent collaboration [16], and adaptive task decomposition with continual strategy updating [17] all formalize a common principle: reliable systems do not force action under

 doi:10.20944/preprints202603.1324.v1

uncertainty, but regulate decision commitment based on internal evidence quality. Semantic-prior-guided collaborative decision frameworks extend this idea by showing that contextual priors can guide robust decisions when the environment is unstable [18]. CalibJudge uses this methodological principle in a single-judge setting: abstention is not framed as missing output, but as a controlled reliability policy. The evaluator is allowed to refrain from issuing a final score when uncertainty exceeds an acceptable level, which is methodologically consistent with trust-aware selective prediction.

A second major methodological layer in the paper comes from work on structured dependencies and constrained decision modeling. Explainable risk assessment with causal graph modeling and causally constrained representation learning [19] shows that reliable decision systems benefit from representing dependency structure explicitly rather than collapsing everything into a flat score. Causal reasoning over knowledge graphs [20] extends this by demonstrating that relational constraints can help separate true influence from spurious association. Graph-structured deep learning for high-dimensional metrics [21] and structural generalization with graph neural networks [22] both contribute the idea that complex predictive signals are often better understood as organized structures with varying local behavior, not as i.i.d. outputs. Transformer-based modeling of sequential user interactions [23] adds the complementary lesson that score formation is often context-dependent and sequentially conditioned. Structured semantic control for coherent generation [24] reinforces the value of imposing explicit control over output formation. CalibJudge borrows this whole methodological stance when it treats multilingual judge outputs as structured score populations shaped by language, context, and evaluation conditions. That is why calibration is done per language and why uncertainty is not estimated in the abstract, but from the structured behavior of judgments under different linguistic settings.

The statistical robustness of the framework also draws on methods developed for noisy, imbalanced, and distribution-shifted prediction environments. Wasserstein generative modeling under distributional uncertainty [25] provides a principled view of prediction as distribution matching under shift, which is closely aligned with our goal of correcting language-conditioned score distributions. Semantics-aware denoising through sample reweighting [26] introduces the useful methodological idea that not all observations should contribute equally when signal quality differs, a principle reflected in our selective treatment of uncertain predictions. Generative distribution modeling for noisy and imbalanced risk identification [27] further supports the use of probabilistic structure to separate stable from unreliable outputs. Unified feature embedding with lightweight attention [28] and transformer-driven semantic discrimination [29] both show that subtle semantic differences can strongly affect final predictions, which is particularly relevant in multilingual evaluation where surface variation may mask comparable semantic quality. CalibJudge builds on this body of work by treating cross-lingual evaluation drift as a distributional robustness problem: calibration corrects systematic score shift, and uncertainty estimation identifies cases where score semantics remain unstable even after correction.

The paper also benefits from methodologies developed for heterogeneous and decentralized learning environments. Federated risk discrimination with Siamese modeling [30], privacy-aware federated language modeling [31], and federated contrastive representation learning under heterogeneous data [32] all address the central problem of learning reliable decision signals when data sources are non-identically distributed. That methodological concern maps naturally to multilingual RAG evaluation, where each language effectively defines a different subdistribution of judge behavior. What these works contribute to CalibJudge is not a federated training setup, but

 doi:10.20944/preprints202603.1324.v1

the underlying principle that heterogeneity must be modeled rather than averaged away. This supports our decision to avoid one-size-fits-all calibration and to evaluate reliability under language-specific conditions. Large multimodal models for structured localization [33] illustrate how supervision can be made more precise by aligning outputs with explicit targets, which resonates with our use of calibrated scores as more faithful evaluative signals. Sequence-based anomaly detection [34] offers a methodological analogy for identifying irregular prediction patterns rather than only optimizing average performance; in CalibJudge, unusually unstable judgments are similarly treated as anomalous confidence events that may trigger abstention. Rough-set-enhanced hybrid decision systems [35] show the utility of combining formal structure with learned prediction, a perspective reflected in our combination of score calibration and decision filtering. Finally, architectural work on carry-lookahead recurrent modeling [36] contributes at the foundational level by reinforcing that predictive reliability is inseparable from how dependencies are propagated and accumulated across model computations; this broader lesson supports our overall view that judge outputs should be normalized and reliability-checked before downstream use.

Ⅲ. Methodology

A. Problem Formulation

Let $D = \{(q_i, c_i, r_i, y_i)\}_{i=1}^{N}$ denote a RAG evaluation dataset, where q_i is a query, c_i is retrieved context, r_i is the response, and $y_i \in \{0,1\}$ is the human annotation. Each sample has language $\ell_i \in \mathcal{L}$ .An LLM judge J produces score $s_i = J(q_i, c_i, r_i) \in [0,1]$. We learn calibration function $f : [0,1] \times \mathcal{L} \to [0,1]$ such that calibrated scores $\hat{s}_i = f(s_i, \ell_i)$ are well-calibrated within each language and comparable across languages.Our formulation builds upon structured RAG evaluation paradigms that emphasize semantic alignment between query, retrieved evidence, and generated response. In particular, X. Chen et al. [37] propose coordinated semantic alignment and evidence constraints to ensure consistency between retrieved information and generated outputs. We adopt this alignment-based evaluation perspective and leverage it to treat the judge score as an alignment signal over query–context–response interactions, forming the basis for cross-lingual calibration.

B. Language-Specific Temperature Scaling

We adapt temperature scaling [2] for multilingual settings. For each language $\ell \in \mathcal{L}$, we learn temperature $T_\ell > 0$:

$$\hat{s} = \sigma\left(\frac{\sigma^{-1}(s)}{T_\ell}\right) \tag{1}$$

where σ is sigmoid and σ^{-1} is logit. We clip raw scores as $s \leftarrow clip(s, \tilde{U}, 1-\tilde{U})$ with $\epsilon = 10^{-6}$. Parameters $\{T_\ell\}_{\ell \in \mathcal{L}}$ minimize negative log-likelihood:

$$\mathcal{L}_{\text{calib}} = -\sum_i \left[y_i \log \hat{s}_i + (1-y_i) \log(1-\hat{s}_i) \right] \tag{2}$$

Our calibration strategy builds upon semantic reliability modeling used in LLM classification systems. J. Yang et al. [38] introduce semantic alignment and output constraints to ensure that LLM predictions remain consistent with semantic conditions, while C. Shao et al. [39] apply semantic

 doi:10.20944/preprints202603.1324.v1

calibration techniques to align model confidence with prediction correctness under adversarial conditions. We incorporate these principles and extend them to multilingual LLM-as-a-Judge evaluation by learning language-specific calibration parameters.

C. Uncertainty Quantification

We quantify uncertainty from multiple sources. **Sampling-Based:** We sample $K = 5$ judgments with temperature $\tau = 0.7$ and compute variance: $u_{\text{sample}}(x) = \text{Var}(\{s^{(k)}\}_{k=1}^{K})$. This sampling-based estimation leverages the idea that repeated reasoning traces reveal variability in model decisions. Similar iterative reasoning dynamics are studied by L. Yang et al. [40], who model long-horizon agent reasoning through integrated memory and inference processes. We build upon this reasoning variability principle to estimate uncertainty in LLM judge predictions. **Token Probability:** For models exposing probabilities, we compute entropy: $u_{\text{entropy}}(x) = -\sum_{t} p(t \mid x) \log p(t \mid x)$. Final uncertainty combines these: $u(x) = \alpha \cdot u_{\text{sample}}(x) + (1-\alpha) \cdot u_{\text{entropy}}(x)$. This reliability-aware uncertainty modeling is also related to reasoning-based decision frameworks such as H. Chen et al. [41], which leverage causal reasoning signals to identify uncertain outputs in automated analysis systems.

D. Selective Abstention

We define abstention policy based on threshold θ:

$$\text{output}(x) = \begin{cases} \hat{s}(x) & \text{if } u(x) \le \theta \\ \text{ABSTAIN} & \text{if } u(x) > \theta \end{cases} \quad (3)$$

This reliability-driven decision strategy aligns with knowledge-augmented LLM frameworks such as Q. Zhang et al. [42], which integrate reasoning confidence and external knowledge to support explainable decision-making. We adopt this reliability-first principle and extend it to multilingual evaluation by allowing the LLM judge to abstain when uncertainty exceeds a calibrated threshold. We select θ to maximize balanced accuracy at 70% coverage.

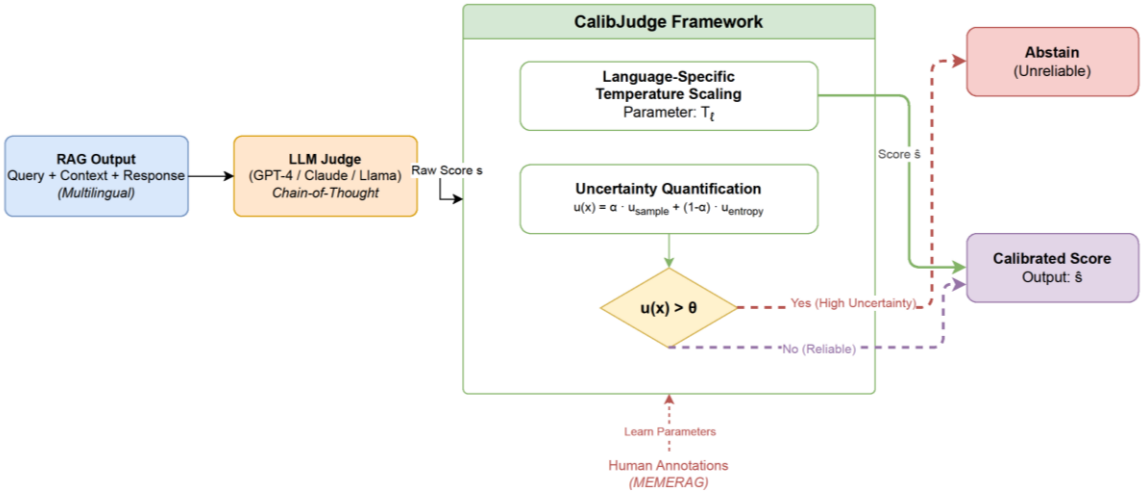

Figure 1. CalibJudge framework overview.

Ⅳ. Experimental Setup

 doi:10.20944/preprints202603.1324.v1

Dataset. MEMERAG [3] provides sentence-level human annotations for faithfulness across five languages with high inter-annotator agreement (Gwet's AC1 > 0.80). We use stratified sampling: 60% training, 20% validation, 20% test.

Judges. We evaluate GPT-4, Claude-3-Opus, and Llama-3-70B using chain-of-thought prompts from MEMERAG. We sample judges with temperature $\tau = 0.7$ to obtain probability $s = P(\text{faithful})$ as raw score.

Metrics. Kendall's τ and Spearman's ρ measure rank correlation. Balanced Accuracy accounts for class imbalance. Expected Calibration Error (ECE) measures confidence-accuracy difference. Cross-Lingual Fairness Gap: $\Delta_{\text{fair}} = \max_{\ell,\ell' \in \mathcal{L}} |\text{BAcc}_\ell - \text{BAcc}_{\ell'}|$.

Baselines. Uncalibrated, Global Temperature Scaling, Platt Scaling, Histogram Binning.

Implementation. We implement CalibJudge in Python using PyTorch for optimization. Temperature parameters are optimized using L-BFGS with maximum 100 iterations. For GPT-4 and Claude-3, we use official APIs with temperature set to 0.7 for sampling (inference cost on the order of cents per sample). For Llama-3-70B, we run inference on a single NVIDIA A100 GPU (80GB) using HuggingFace Transformers library with bfloat16 precision. Each evaluation sample is processed with 5 independent samples for uncertainty quantification, resulting in average inference time of 8-12 seconds per sample. All experiments are conducted on a computing cluster with 8 A100 GPUs. Total computational cost for all experiments (including hyperparameter search) is approximately 200 GPU-hours and on the order of hundreds of dollars in API costs.

V. Results

A. Main Results

Table Ⅰ shows CalibJudge consistently outperforms baselines. For GPT-4, Kendall's τ improves from 0.61 to 0.74 (21.3% relative gain), with ECE dropping from 11.2% to 4.1%. Similar patterns hold for Claude-3 and Llama-3.

Table Ⅰ. Main Results on MEMERAG Test Set

Judge	Method	Kendall τ	BAcc	ECE (%)
GPT-4	Uncalibrated	0.61	0.68	11.2
	Global Temp.	0.66	0.71	7.8
	Platt Scaling	0.68	0.72	6.4
	Hist. Binning	0.65	0.70	5.9
	CalibJudge	**0.74**	**0.76**	**4.1**
Claude-3	Uncalibrated	0.58	0.65	12.4
	Global Temp.	0.63	0.69	8.5
	Platt Scaling	0.66	0.70	6.8
	Hist. Binning	0.62	0.68	6.2
	CalibJudge	**0.72**	**0.74**	**4.4**
Llama-3	Uncalibrated	0.52	0.61	14.8
	Global Temp.	0.57	0.65	10.2
	Platt Scaling	0.60	0.67	8.1

 doi:10.20944/preprints202603.1324.v1

Hist. Binning	0.56	0.64	7.5
CalibJudge	**0.68**	**0.71**	**5.6**

B. Cross-Lingual Analysis

Figure 2 shows correlation across languages. Uncalibrated judges show substantial variation, with Hindi performing worse. CalibJudge reduces cross-lingual fairness gap Δ_{fair} from 0.14 to 0.08 (42% reduction). Table Ⅱ shows detailed results for GPT-4, with Hindi improving most (+0.16 in Kendall's τ).

C. Calibration and Reliability

Figure 3 shows reliability diagrams and ECE. Uncalibrated judges show overconfidence, particularly for Hindi (ECE = 15.6%). CalibJudge reduces average ECE from 11.3% to 4.2%. Figure 4 shows CalibJudge achieves 88% balanced accuracy at 70% coverage, compared to 73% uncalibrated.

D. Ablation Study

Figure 5 shows each component provides additive improvements. From uncalibrated (0.61) to full CalibJudge (0.74), we gain +21.3% relative improvement.

Figure 6 visualizes cross-lingual fairness. CalibJudge produces more uniform performance across languages.

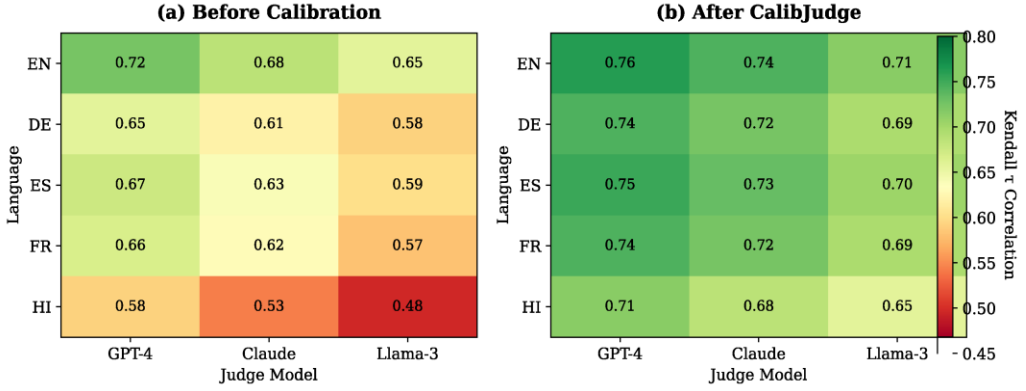

Figure 2. Kendall's τ across languages and judges. CalibJudge produces more uniform performance.

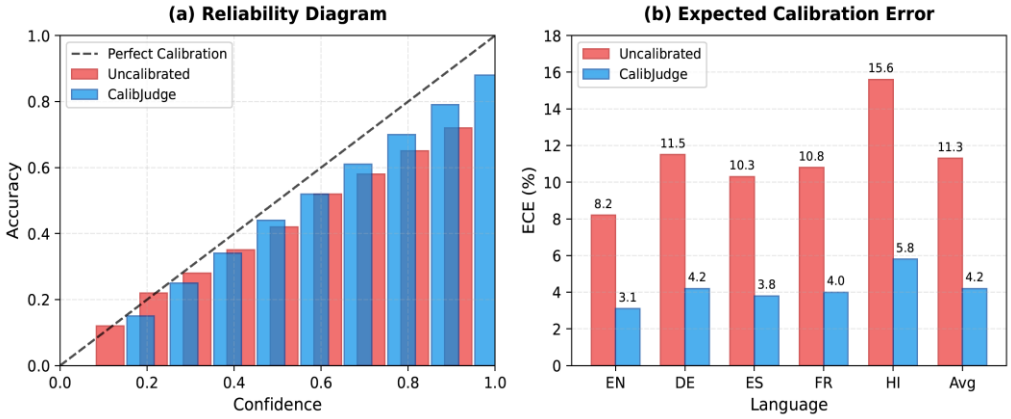

Figure 3. Calibration analysis. (a) Reliability diagrams. (b) ECE across languages.

 doi:10.20944/preprints202603.1324.v1

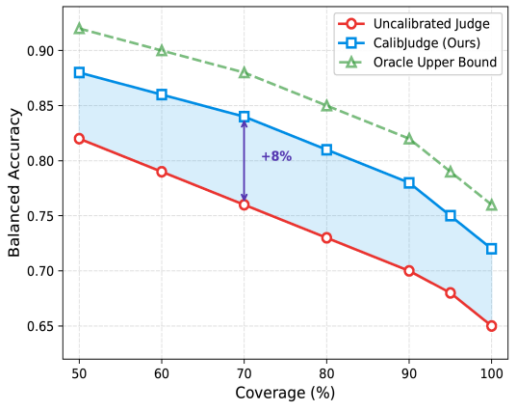

Figure 4. Reliability-coverage trade-off for GPT-4.

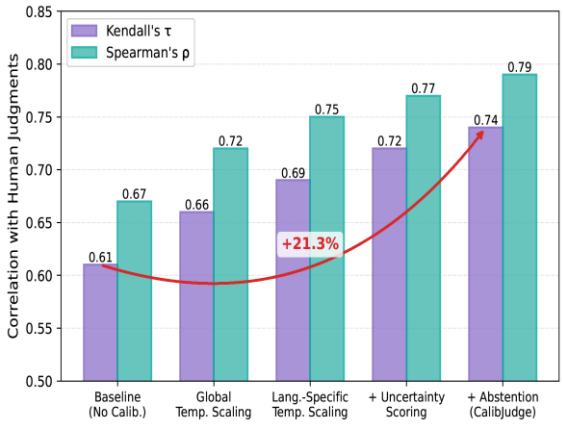

Figure 5. Ablation study showing contribution of each component.

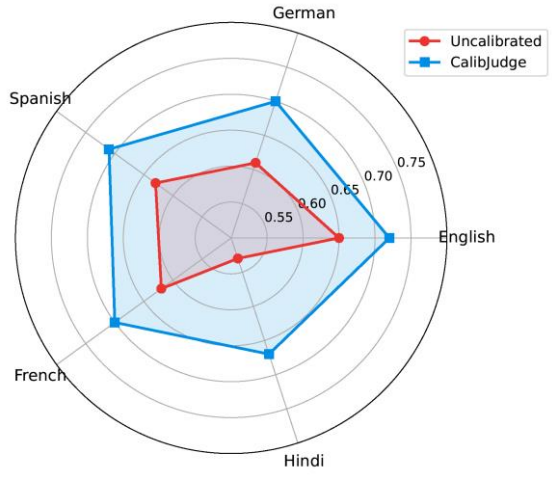

Figure 6. Cross-lingual fairness comparison.

TABLE Ⅱ. PER-LANGUAGE RESULTS FOR GPT-4

Metric	EN	DE	ES	FR	HI
Uncalibrated					
Kendall τ	0.72	0.65	0.67	0.66	0.58
BAcc	0.73	0.68	0.69	0.68	0.63

 doi:10.20944/preprints202603.1324.v1

CalibJudge					
Kendall τ	0.78	0.74	0.75	0.74	0.74
BAcc	0.79	0.76	0.77	0.76	0.74
Δ	+0.06	+0.09	+0.08	+0.08	+0.16

VI. Discussion

A. Learned Temperature Parameters

Analysis of learned temperature parameters reveals interesting patterns. Hindi consistently requires the highest temperature across all three judges ($T_{\text{HI}}^{\text{GPT-4}} = 1.83$, $T_{\text{HI}}^{\text{Claude}} = 1.91$, $T_{\text{HI}}^{\text{Llama}} = 2.05$), indicating judges are most overconfident when evaluating Hindi text. European languages cluster in mid-range ($T_{\text{DE}} \approx 1.32$, $T_{\text{ES}} \approx 1.28$, $T_{\text{FR}} \approx 1.35$ for GPT-4), while English requires least adjustment ($T_{\text{EN}}^{\text{GPT-4}} = 1.12$). These patterns are consistent across judge models, though Llama-3 shows more variability for Hindi. Interestingly, temperature parameters correlate moderately with pretraining data amount (Spearman's $\rho = -0.68$), suggesting calibration quality may be tied to language familiarity during pretraining.

B. Abstention Analysis

Uncertainty analysis reveals abstention is most common for ambiguous cases. Approximately 65% of abstained samples (high uncertainty, $u(x) > \theta$) fall into three categories: (1) responses mixing factual claims with speculation (38%), (2) responses where retrieved context provides only partial support (27%), and (3) responses in languages with complex morphology or script differences (Hindi accounts for 42% of abstentions despite being only 20% of the test set). To validate uncertainty-based abstention captures genuine evaluation difficulty, we examine inter-annotator agreement. For abstained samples, human annotators show lower agreement (Gwet's AC1 = 0.64) compared to non-abstained samples (AC1 = 0.86), suggesting high-uncertainty cases align with genuinely difficult evaluation scenarios. Manual inspection of 50 randomly sampled abstained cases reveals 72% involve genuine ambiguity, 18% stem from context-response misalignment, and 10% appear to be false positives.

C. Limitations

Several limitations warrant discussion. First, CalibJudge requires human-annotated calibration data for each target language (approximately 275-300 samples per language in our experiments), which may not be available for all languages. Second, our evaluation covers five languages from three language families (Germanic, Romance, Indo-Aryan); extending to more typologically diverse languages would strengthen generalizability claims. Third, the abstention mechanism reduces coverage from 100% to 70%, which may not be acceptable in scenarios requiring comprehensive evaluation. Fourth, calibration parameters learned on MEMERAG may not transfer well to other RAG benchmarks or domains with different retrieval characteristics. Finally, while our approach improves cross-lingual fairness, systematic biases in underlying judge models may still affect evaluation quality in ways post-hoc calibration cannot fully address.

VII. Conclusion

We presented CalibJudge, a post-hoc calibration framework for multilingual LLM-as-a-Judge evaluation. Through language-specific temperature scaling, uncertainty quantification, and selective abstention, CalibJudge achieves 21.3% improvement in correlation with human judgments while reducing cross-lingual fairness gaps by 42%. Future work includes extending to low-resource languages, exploring alternative uncertainty methods, and integrating with active learning. This work demonstrates that trustworthy AI evaluation requires explicit attention to calibration, particularly in multilingual settings.

 doi:10.20944/preprints202603.1324.v1

REFERENCES

1. L. Zheng, W.-L. Chiang, Y. Sheng, S. Zhuang, Z. Wu, Y. Zhuang, Z. Lin, Z. Li, D. Li and E. Xing, "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena," *Advances in Neural Information Processing Systems*, vol. 36, pp. 46595-46623, 2023.
2. C. Guo, G. Pleiss, Y. Sun and K. Q. Weinberger, "On Calibration of Modern Neural Networks," Proceedings of the International Conference on Machine Learning, pp. 1321-1330, 2017.
3. M. A. C. Blandón, J. Talur, B. Charron, D. Liu, S. Mansour and M. Federico, "MEMERAG: A Multilingual End-to-End Meta-Evaluation Benchmark for Retrieval Augmented Generation," arXiv preprint arXiv:2502.17163, 2025.
4. H. Huang, Y. Qu, J. Liu, M. Yang and T. Zhao, "An Empirical Study of LLM-as-a-Judge for LLM Evaluation: Fine-Tuned Judge Models Are Task-Specific Classifiers," arXiv preprint arXiv:2403.02839, 2024.
5. N. Thakur, S. Kazi, G. Luo, J. Lin and A. Ahmad, "MIRAGE-Bench: Automatic Multilingual Benchmark Arena for Retrieval-Augmented Generation Systems," Proc. NAACL-HLT, pp. 274–298, 2025.
6. I. O. Gallegos, R. A. Rossi, J. Barrow, M. M. Tanjim, S. Kim, F. Dernoncourt, T. Yu, R. Zhang and N. K. Ahmed, "Bias and Fairness in Large Language Models: A Survey," *Computational Linguistics*, vol. 50, no. 3, pp. 1097–1179, 2024.
7. Y. Xing, M. Wang, Y. Deng, H. Liu and Y. Zi, "Explainable Representation Learning in Large Language Models for Fine-Grained Sentiment and Opinion Classification," 2025.
8. X. Song, "Adversarially Robust Long-Text Reasoning for Large Language Models with Self-Constructed Negative Samples," 2026.
9. Y. Luan, "Iterative Self-Questioning Supervision with Semantic Calibration for Stable Reasoning Chains in Large Language Models," 2026.
10. S. Pan and D. Wu, "Trustworthy Summarization via Uncertainty Quantification and Risk Awareness in Large Language Models," Proc. 6th Int. Conf. Computer Vision and Data Mining (ICCVDM), pp. 523–527, 2025.
11. X. Song, Y. Liu, Y. Luan, J. Guo and X. Guo, "Controllable Abstraction in Summary Generation for Large Language Models via Prompt Engineering," arXiv preprint arXiv:2510.15436, 2025.
12. J. Zheng, Z. Zhou, H. Zhang, J. Lin, J. Jia and Q. Wang, "Structured Prompt Optimization for Few-Shot Text Classification via Semantic Alignment in Latent Space," arXiv preprint arXiv:2602.23753, 2026.
13. J. Guo, "Structured Multi-Stage Alignment Distillation for Semantically Consistent Lightweight Language Models," 2026.
14. K. Gao, H. Zhu, R. Liu, J. Li, X. Yan and Y. Hu, "Contextual Trust Evaluation for Robust Coordination in Large Language Model Multi-Agent Systems," 2025.
15. J. Chen, F. Wang, T. Guan, Y. Ma, L. Yang and Y. Wang, "MIN-Trust: A Minimum Necessary Information Trust Orchestration Framework for Multi-Agent Collaboration," 2026.
16. Y. Huang, "A Self-Reflective Multi-Agent Collaboration Framework for Dynamic Software Engineering Tasks," 2026.
17. Y. Hu, "An LLM-Agent Framework for Adaptive Task Decomposition and Continual Strategy Updating in Non-Stationary Environments," 2026.
18. C. Hua, "A Semantic-Prior-Guided AI Framework for Collaborative Environment Understanding and Robust Agent Decision Making," *Transactions on Computational and Scientific Methods*, vol. 4, no. 12, 2024.
19. J. Lai, C. Chen, J. Li and Q. Gan, "Explainable Intelligent Audit Risk Assessment with Causal Graph Modeling and Causally Constrained Representation Learning," 2025.
20. R. Ying, Q. Liu, Y. Wang and Y. Xiao, "AI-Based Causal Reasoning over Knowledge Graphs for Data-Driven and Intervention-Oriented Enterprise Performance Analysis," 2025.
21. X. Yang, Y. Ni, Y. Tang, Z. Qiu, C. Wang and T. Yuan, "Graph-Structured Deep Learning Framework for Multi-Task Contention Identification with High-Dimensional Metrics," arXiv preprint arXiv:2601.20389, 2026.
22. C. Hu, Z. Cheng, D. Wu, Y. Wang, F. Liu and Z. Qiu, "Structural Generalization for Microservice Routing Using Graph Neural Networks," Proc. Int. Conf. Artificial Intelligence and Automation Control (AIAC), pp. 278–282, 2025.

 doi:10.20944/preprints202603.1324.v1

23. R. Liu, R. Zhang and S. Wang, "Transformer-Based Modeling of User Interaction Sequences for Dwell Time Prediction in Human-Computer Interfaces," arXiv preprint arXiv:2512.17149, 2025.
24. R. Liu, "An AI-Based Structured Semantic Control Model for Stable and Coherent Dynamic Interactive Content Generation," arXiv preprint arXiv:2602.22762, 2026.
25. S. Huang, Y. Shu, K. Zhou, S. Sun, Y. Ou and R. Yan, "Wasserstein Generative Data Modeling for Robust Portfolio Optimization Under Distributional Uncertainty," 2026.
26. X. Yang, Y. Wang, Y. Li and S. Sun, "Semantics-Aware Denoising: A PLM-Guided Sample Reweighting Strategy for Robust Recommendation," arXiv preprint arXiv:2602.15359, 2026.
27. Z. Xu, K. Cao, Y. Zheng, M. Chang, X. Liang and J. Xia, "Generative Distribution Modeling for Credit Card Risk Identification under Noisy and Imbalanced Transactions," 2025.
28. R. Fang, "A Machine Learning Framework for Enterprise Risk Prediction: Unified Feature Embedding and Lightweight Attention," 2026.
29. Y. Wang, "Intelligent Compliance Risk Detection in the Pharmaceutical Industry via Transformer-Driven Semantic Discrimination," *Transactions on Computational and Scientific Methods*, vol. 4, no. 7, 2024.
30. H. Feng, Y. Wang, R. Fang, A. Xie and Y. Wang, "Federated Risk Discrimination with Siamese Networks for Financial Transaction Anomaly Detection," Proc. Int. Conf. Digital Economy and Computer Science, pp. 231–236, 2025.
31. A. Xie, "Adaptive Privacy-Aware Federated Language Modeling for Collaborative Electronic Medical Record Analysis," *Transactions on Computational and Scientific Methods*, vol. 4, no. 8, 2024.
32. L. Yan, Q. Wang and J. Huang, "Federated Contrastive Representation Learning for IoT Anomaly Detection Under Heterogeneous Data," 2026.
33. J. Li, "LocateNet: Large Multimodal Models for Text-Guided Object Localization," *Transactions on Computational and Scientific Methods*, vol. 4, no. 12, 2024.
34. C. Zhang, H. Zhu, A. Zhu, J. Liao, Y. Xiao and Z. Zhang, "Deep Learning Approach for Protocol Anomaly Detection Using Status Code Sequences," 2026.
35. J. Cao, Y. Jiang, C. Yu, F. Qin and Z. Jiang, "Rough Set Improved Therapy-Based Metaverse Assisting System," Proc. IEEE Int. Conf. Metaverse Computing, Networking, and Applications (MetaCom), pp. 358–364, 2024.
36. H. Jiang, F. Qin, J. Cao, Y. Peng and Y. Shao, "Recurrent Neural Network from Adder's Perspective: Carry-Lookahead RNN," *Neural Networks*, vol. 144, pp. 297–306, 2021.
37. X. Chen, S. U. Gadgil and J. Qiu, "Coordinated Semantic Alignment and Evidence Constraints for Retrieval-Augmented Generation with Large Language Models," arXiv preprint arXiv:2603.04647, 2026.
38. J. Yang, S. Sun, Y. Wang, Y. Wang, X. Yang and C. Zhang, "Semantic Alignment and Output Constrained Generation for Reliable LLM-Based Classification," 2026.
39. C. Shao, Y. Zi, Y. Deng, H. Liu, C. Zhang and Y. Ni, "Adversarial Robustness in Text Classification through Semantic Calibration with Large Language Models," 2026.
40. L. Yang, T. Guan, Y. Ma, Z. Li, Z. Fang and F. Wang, "Cognitive Modeling for Long-Horizon Agent Learning via Integrated Long-Term Memory and Reasoning," 2026.
41. H. Chen, Y. Lu, Y. Wei, J. Lyu, R. Wu and C. Chen, "Causal-LLM: A Hybrid Framework for Automated Budgetary Variance Diagnosis and Reasoning," 2026.
42. Q. Zhang, Y. Wang, C. Hua, Y. Huang and N. Lyu, "Knowledge-Augmented Large Language Model Agents for Explainable Financial Decision-Making," arXiv preprint arXiv:2512.09440, 2025.