When the Model Shouldn't Cite: Calibrated Abstention for Islamic Citation Hallucination

Mansur Ali Khan [1] **Mehmet Efe Akengin** [2] **Ahmad A. Rushdi** [3]

Abstract

Large language models sometimes misquote the Qur'anic verses and Hadiths they cite --- a high-stakes failure where confidently fabricating sacred text is worse than ordinary hallucination. The 2025 IslamicEval shared task benchmarks this, but every system must commit a label on every span; none abstains. We add a plug-in abstention layer: combine publicly released LLM outputs with cross-family runs of our own ($K{=}7$ total), score each span by majority-disagreement, and abstain when disagreement is high. The threshold is chosen on a $60/40$ calibration split via a union-bound-adjusted Clopper-Pearson upper bound, giving a calibration-sample high-probability bound on selective risk. On the Subtask 1C dev set, the selective-accuracy curve exceeds the leaderboard reference at 14% abstention and reaches 94% conditional accuracy at 44%; risk-controlled operating points certify the 44% region. Public test labels are unavailable, so leaderboard scores serve as reference lines, not head-to-head. The layer is plug-and-play, requires no retraining, and is reusable across tasks.

Task

Subtask 1C span (`C-Q35`, hallucinated hadith): system must return the canonical correction or "wrong".
Gold: فيقول: من يدعوني فأستجيب له، من يسألني فأعطيه، من يستغفرني فأغفر له.

#	variant	output
1	gpt-5-nano	× **wrong**
2	gpt-5 (low)	✓ **gold**
3	gpt-5 (low+codex)	× **other**
4	gpt-5 (high+codex)	× **other**
5	claude-opus-4.7	× **other**
6	qwen3-235b	× **wrong**
7	llama-3.3-70b	× **wrong**

Accuracy only (leaderboard)
majority wrong $(3/7)$
↓
commit ×

Accuracy + abstention (ours)
disagreement $(4/7) > \hat{\tau}$
↓
abstain ✓

Figure 1. An illustrative example on Subtask 1C (C-Q35) with $K{=}7$ variants: 3 emit the refusal-token **wrong**, 3 emit a different hadith (**other**), 1 returns the canonical **gold**. Without abstention, the $K{=}7$ majority commits the largest cluster (**wrong**); our calibrated layer abstains when disagreement exceeds a calibrated threshold $\hat{\tau}$. Here, the largest cluster covers only $3/7$, leaving $4/7$ in disagreement (above threshold, so we abstain ✓).

1. Introduction

Religious text holds a singular position in Arabic NLP: the Qur'an and the canonical Hadith collections are, for Muslims, both linguistic objects and objects of sacred and legal authority. When an LLM quotes scripture, the standard is not "fluent" but "the exact canonical wording, attributed correctly, and not fabricated." A confident misquotation is a categorical failure, not a near miss.

The 2025 IslamicEval shared task (Mubarak et al., 2025) is the first benchmark targeting this. Thirteen teams compete on flagging and correcting hallucinated Qur'anic and Hadith citations; the winning Subtask 1C (Correction) system (Omayrah et al., 2025) reaches 68.18% exact-match on the held-out test set, barely above the 67.52% majority baseline. Every submission must commit to a label on every span (a canonical correction or the refusal label "wrong", خطأ); none of the five system papers reports a per-span confidence, calibration curve, or system-level abstention rule.

Our contribution is a *selective prediction* layer for Islamic citation correction --- not a new correction model: given K outputs per span, commit when they agree, abstain when they don't. We score each span by cross-variant disagreement $s = (K - a)/K$ and apply a split-calibrated threshold (Vovk et al., 2005; Angelopoulos & Bates, 2023; Geifman & El-Yaniv, 2017; Abbasi-Yadkori et al., 2024). Figure 1 walks through one representative span. The whole curve is the contribution, not a single point: a deployer picks the abstention budget matching its tolerance for confident miscitation.

[1]University of Washington, Bothell, WA, USA [2]Fatima Institute for Global AI Research [3]Stanford University, Stanford, CA, USA. Correspondence to: Mansur Ali Khan <mkhan30@uw.edu>, Ahmad A. Rushdi <rushdi@stanford.edu>.

Proceedings of the 43rd International Conference on Machine Learning, Seoul, South Korea. PMLR 306, 2026. Copyright 2026 by the author(s).

2. Related Work

A growing line of work benchmarks LLM competence on Arabic and Islamic content: IslamicMMLU (Abdelaal et al., 2026), IslamTrust (Lahmar et al., 2025) (Sunni ethical alignment), PalmX 2025 (Alwajih et al., 2025) (cultural and Islamic knowledge), and AraHalluEval (Alansari & Luqman, 2025) (fine-grained hallucination indicators across twelve Arabic LLMs). Accuracy and F1 dominate the reporting; none of these benchmarks publishes a calibration curve, conformal coverage, or abstention rule.

The benchmark we target, IslamicEval 2025 (Mubarak et al., 2025), hand-annotates 1,506 responses from six Arabic LLMs (ALLaM-7B, Jais-13B, Qwen3-8B, Llama-3.1-8B, and two Gemma-2-9B variants) to the 251 questions of *Qur'an QA 2023* (Malhas et al., 2023) across three sub-subtasks; Subtask 1C is scored by exact-match after default-diacritic normalization, with a public dev set (50 spans per sub-subtask) and a CodaBench-scored test set (104). Thirteen teams competed; five published system descriptions (BurhanAI (Al Adel et al., 2025), HUMAIN (Omayrah et al., 2025), TCE (ElKoumy et al., 2025), ThinkDrill (Elrefai et al., 2025), Isnad AI (Mohamed Emad Elden, 2025)) spanning RAG, agentic tool-use, multi-stage seq-to-seq correction, and rule-based patterns. The headline result is informative about the difficulty: HUMAIN's 68.18% on 1C only nudges past the 67.52% majority baseline (which predicts the refusal token for every span), and *none* of the five system papers reports a per-span confidence, calibration curve, or abstention rule.

Standard NLP uncertainty machinery elsewhere: split-conformal prediction (Vovk et al., 2005; Angelopoulos & Bates, 2023) for finite-sample coverage, applied to LLM hallucination by Abbasi-Yadkori et al. (2024); semantic entropy via NLI clustering (Farquhar et al., 2024); SelfCheckGPT's black-box consistency (Manakul et al., 2023); selective classification (Geifman & El-Yaniv, 2017); and conditional-coverage limits (Barber et al., 2021). None has been applied to Islamic-content evaluation. Our cross-variant disagreement is the natural ensemble analogue: we construct K variant outputs from released BurhanAI runs plus three cross-family runs, with no resampling required.

The closest prior work, Atif et al. (2025), thresholds single-model token log-probabilities for abstention on FiqhQA. We differ in three ways: a split-calibrated threshold selection with empirical holdout validation (not a heuristic threshold); a cross-variant ensemble disagreement score (not single-model per-token logprob, brittle on Arabic generation); and a comparison to published IslamicEval 2025 leaderboard reference scores (with dev/test caveats), not a stand-alone analysis.

3. Method

Our approach converts K variant outputs (made available by the shared-task setting) into a calibrated abstention layer. We formalize Subtask 1C (§3.1), define cross-variant disagreement as an uncertainty score (§3.2), apply risk-controlled split calibration via a union-bound-adjusted Clopper-Pearson UCB (§3.3), and list the seven variants (§3.4).

3.1. Setting

IslamicEval 2025 Subtask 1 is a three-stage pipeline over LLM free-text answers to religious questions: 1A identifies the spans that claim a verse or Hadith, 1B classifies each span as correct or incorrect against canonical sources, and 1C provides a corrected version for any span flagged as incorrect (or the refusal label "wrong" if no canonical correction exists). *This paper targets only 1C* (Correction), where the hallucination is concrete: a confidently-cited verse or hadith that does not match the canonical text. Appendix A gives the official three-stage breakdown and an example.

For each span $i \in \{1, \ldots, n\}$ in the 1C dev or test set, we are given an erroneous quotation x_i (the "Original Span" in the gold TSV) and a gold canonical correction $y_i^\star$ that is either the verse or hadith text the original was attempting to cite, or the refusal label "wrong" (خطأ) when no valid correction exists. The public 1C files expose each item as the erroneous *Original Span* together with a span-type label (`WrongAyah` or `WrongHadith`), without the upstream full response, question, or gold correction (those live in 1A artifacts not in the 1C release). We run the three added cross-family models using only these public 1C fields, returning either the canonical correction or خطأ.[1] The K=7 comparison is therefore aligned at the 1C item level; the leaderboard comparison remains dev-set selective accuracy vs. published test-set references.

The task page describes 1C scoring via word error rate, but the released leaderboard (Mubarak et al., 2025) and the public scoring script use exact match between $\hat{y}_i$ and $y_i^\star$ after default-diacritic normalization (`remove_default_diac` from the QCRI evaluation script); we follow that script. The official metric is the marginal accuracy $\text{acc} = \frac{1}{n}\sum_i \mathbf{1}[\hat{y}_i = y_i^\star]$. Under this rule, a single missing word or non-default diacritic counts as a full error, and there is no partial credit for a paraphrased gloss or for matching only the opening hemistich --- the same hardness that motivates an abstention rule. Every system must commit to a label for every span; "wrong" is therefore an *answer*, not a system-level abstention --- which is what the rest of this paper supplies.

[1] The exact prompt template is reproduced in Appendix B.

3.2. Cross-variant disagreement: uncertainty score

We treat the K variant outputs as an ensemble. For each variant V_k ($k = 1, \ldots, K$), let $\hat{y}_i^{(k)}$ denote its prediction on span i. Two ensemble statistics determine the score: the *plurality* answer (the largest output cluster, which need not exceed $K/2$) $m_i = \arg\max_y \#\{k : \hat{y}_i^{(k)} = y\}$, with ties broken by the listing order in §3.4 (the score below is tie-break-invariant), and the agreement count $a_i = \#\{k : \hat{y}_i^{(k)} = m_i\} \in \{1, \ldots, K\}$. The *disagreement fraction*

$$s(i) = \frac{K - a_i}{K} \in \left\{0, \tfrac{1}{K}, \ldots, \tfrac{K-1}{K}\right\} \tag{1}$$

is our uncertainty score: $s(i) = 0$ when all K variants agree (highest confidence), $s(i) = (K-1)/K$ when every variant gives a different answer. When the ensemble does not abstain, its committed prediction is m_i.

3.3. Risk-controlled split calibration

We turn $s(\cdot)$ into an abstention rule with a finite-sample high-probability calibration bound on selective risk, in the spirit of selective classification and risk-controlling prediction sets (Vovk et al., 2005; Angelopoulos & Bates, 2023; Geifman & El-Yaniv, 2017; Abbasi-Yadkori et al., 2024). Fix a target selective error β and a failure budget δ. On the discrete grid $T = \{0, 1/K, \ldots, (K-1)/K\}$, for each $\tau \in T$ let $\mathcal{A}_\tau = \{i \in \mathcal{C} : s(i) \leq \tau\}$, $m_\tau = |\mathcal{A}_\tau|$, e_τ = errors in $\mathcal{A}_\tau$, and U_τ = one-sided Clopper-Pearson upper bound on e_τ/m_τ at level $1 - \delta/|T|$ (with $U_\tau = 1$ if $m_\tau = 0$).[2] We select

$$\hat{\tau} = \max\left\{\tau \in T : m_\tau > 0 \wedge U_\tau \leq \beta\right\}, \tag{2}$$

abstaining on every holdout span if no valid τ exists. By a union bound over T, with probability at least $1 - \delta$ over the calibration sample every grid point's population selective risk is bounded by its UCB; under exchangeability of $\mathcal{C}$ and $\mathcal{H}$, the selected $\hat{\tau}$ therefore has population selective risk at most β. We use $\delta = 0.05$, $|T| = 7$, and report holdout performance over 200 random $60/40$ splits.

3.4. Variants

Of the five IslamicEval 2025 system-paper teams, only BurhanAI (Al Adel et al., 2025) publicly released per-span 1C predictions; we use all four of their released GPT-5 configurations (`nano`, `low`, `low+codex`, `high+codex`). We add three cross-family models that we run ourselves via OpenRouter at $T = 0$: `claude-opus-4.7`, `qwen3-235b-a22b`, and `llama-3.3-70b`. We report $K = 4$ (single-family) and $K = 7$ (cross-family) to ablate family diversity.

[2] Concretely, $U_\tau = \text{Beta}^{-1}(1 - \delta/|T|;\ e_\tau + 1,\ m_\tau - e_\tau)$, with $U_\tau = 1$ when $e_\tau = m_\tau$.

4. Experiments

We evaluate the layer on the IslamicEval Subtask 1C dev set in four steps: setup (§4.1); no-abstention marginal accuracies (§4.2); cross-variant agreement structure (§4.3); and the headline selective-accuracy curve with a K=4 vs K=7 ablation and a calibration-to-holdout check (§4.4).

4.1. Setup

The dev set comprises $n = 179$ Subtask 1C spans across 50 questions, released in `TaskC_GT.tsv`; we score under the official exact-match-after-diacritic-normalization rule from Mubarak et al. (2025). We report two ensemble sizes: $K = 4$, using only BurhanAI's (Al Adel et al., 2025) released GPT-5 variants, and $K = 7$, which adds our three runs (Claude Opus 4.7, Qwen3-235B, Llama-3.3-70B). The split calibration of $\hat{\tau}$ is averaged over 200 random $60/40$ calibration/holdout splits. All experiments use Python's `random.seed(42)` and `numpy.random.seed(42)`; the prompts, scripts, and per-variant predictions will be released publicly upon publication.

4.2. No variant matches the leaderboard

Table 1: the strongest single variant is `gpt-5 (high+codex)` at 62.6%, consistent with BurhanAI's test-set 66.56% (Al Adel et al., 2025); the K=7 majority reaches 64.8%, within 3.4 points of HUMAIN's leaderboard-leading 68.18% (Omayrah et al., 2025) but not crossing it. The abstention layer closes and overtakes that gap (green rows).

4.3. Disagreement tracks correctness

We first verify that cross-variant agreement tracks correctness. At $K = 4$, spans with $4/4$ agreement reach 75.0% conditional accuracy, dropping to 62.6% at $3/4$, 55.6% at $2/4$, and 5.9% at $1/4$. At $K = 7$ the pattern sharpens into a bimodal split: $76/179 = 42\%$ of spans with ≥ 5 agreeing variants reach 96.1%, while $78/179 = 44\%$ with ≤ 3 agreeing drop to 26.9% --- the structural property the abstention layer exploits.

4.4. Calibrated abstention beats the reference

Figure 2 shows the full-dev sweep (green) and certified operating points (red diamonds, from §3.3). The K=7 curve exceeds HUMAIN's 68.18% test reference (Omayrah et al., 2025) (reference line, not head-to-head) at 14% abstention and reaches 94.1% at 44%, a $\sim$ 26-point gap. The 14% crossing is descriptive; the UCB rule certifies the 44% region.

K=4 (blue) trails until very high abstention; K=7 adds $\sim$ 21 points at 30%. Random abstention on K=7 (orange)

Setting	Accuracy
gpt-5-nano (BurhanAI)	0.134
gpt-5 (low)	0.514
gpt-5 (low+codex)	0.358
gpt-5 (high+codex)	**0.626**
claude-opus-4.7 (ours)	0.419
qwen3-235b-a22b (ours)	0.531
llama-3.3-70b (ours)	0.531
K=4 ensemble majority (no abst.)	0.553
K=7 ensemble majority (no abst.)	0.648
published test-set references (not head-to-head)	
majority baseline	0.6752
BurhanAI (Al Adel et al., 2025)	0.6656
HUMAIN (Omayrah et al., 2025) (1st place)	**0.6818**
ours (dev): K=7, calibrated abst., 14%	**0.708**
ours (dev): K=7, calibrated abst., 44%	**0.941**

Table 1. IslamicEval Subtask 1C accuracy across settings. Red/orange/green rows are dev-set measurements ($n = 179$); gray rows reproduce published test-set no-abstention scores from (Mubarak et al., 2025) as reference lines, not as official head-to-head comparisons with our dev-set selective accuracy.

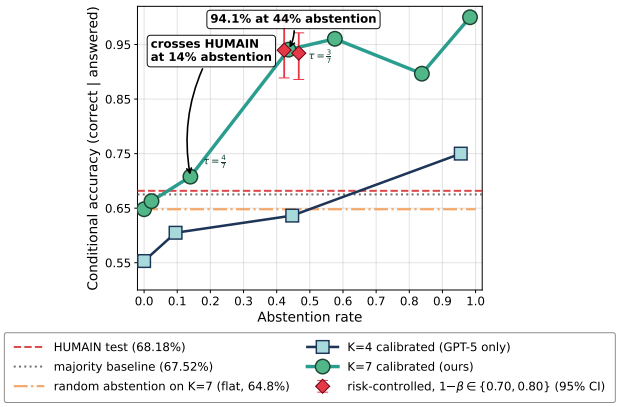

Figure 2. Conditional accuracy vs. abstention rate on IslamicEval Subtask 1C (n=179). Horizontal lines are published test-set references (Mubarak et al., 2025), not head-to-head comparisons. Red diamonds: risk-controlled operating points at $1-\beta \in \{0.70, 0.80\}$ over 200 random 60/40 splits.

stays flat at 64.8%: the lift comes from calibrated selection on s, since abstention only helps when the abstained set is error-enriched. The risk-controlled rule selects $\hat{\tau} = 3/7$ at $1-\beta \in \{0.70, 0.80\}$ ($\approx$ 44--45% abstention, $\approx$ 93--94% accuracy; red diamonds), and every valid split meets the target.

5. Discussion, Ethics, and Limitations

Several caveats temper the results. The dev set is small (n=179), so the UCB rule becomes uncertifiable at aggressive targets (e.g. $1-\beta = 0.90$ is certified in only 3/200 splits) --- a feature of the conservative finite-sample bound, not a failure of the method. Our ensemble uses the one IslamicEval team (BurhanAI) that released per-span predictions plus three cross-family runs; a fuller sample of the thirteen teams could shift it. Those three added models saw only the truncated *Original Span* and its label --- less context than the BurhanAI runs --- which may inflate their disagreement, and hence abstention. And $s(\cdot)$ is blind to *correlated* error: if all variants share a blind spot, disagreement stays low and the layer commits a confident mistake --- a failure mode we flag but leave for future analysis.

Are the gains in Figure 2 genuine uncertainty detection, or just skipping hard spans no model could answer? Two diagnostics probe this. An *oracle any-correct* aggregator (gold if any of the K=7 variants returns it) reaches $127/179 = 70.9\%$ on the full dev set --- an upper bound on what any aggregator over these seven outputs could extract, $\sim$6 points above the K=7 majority's 64.8%.

At the 44% operating point, only 30/78 abstained spans have any variant correct (48 have no exact-gold output among the seven), and $61/78 = 78.2\%$ carry a canonical-text gold --- well above the full-dev base rate of $68/179 = 38.0\%$: the layer disproportionately abstains where a valid correction *does* exist, exactly the high-stakes cases. Per-label, Wrong-Hadith conditional accuracy is 100% (69/84 answered) vs. 81.2% on Wrong-Ayah (32/95 answered).

Ayah hallucinations more often involve diacritic-level mismatches that split the ensemble, driving higher abstention (66% vs. 18%); Qur'an-citation deployments will see higher abstention budgets than Hadith-only ones. Appendix C walks through three representative dev-set spans that exemplify the decision logic.

The IslamicEval 2025 organizers (Mubarak et al., 2025) release the data to improve LLM safety on religious material and caution against misuse; a calibrated refusal layer reduces confident hallucinations on exactly that task. We make no exegetical or jurisprudential claim: the layer operates only over text strings already accepted as gold by the organizers' specialists. The analogy to *tawaqquf* --- abstaining from authoritative speech under uncertainty --- is suggestive: calibrated refusal as a design virtue, not a failure.

6. Conclusion

Our contribution is a selective prediction layer, not a new correction model: given K=7 outputs per span, it commits or abstains on cross-variant disagreement, exceeding HUMAIN's 68.18% at 14% abstention and reaching a risk-controlled 94.1% at 44%. Muslim-facing AI evaluation should report calibrated abstention --- when a model should answer, and when it should not cite.

References

Abbasi-Yadkori, Y., Kuzborskij, I., Stutz, D., György, A., Fisch, A., Doucet, A., Beloshapka, I., Weng, W.-H., Yang, Y.-Y., Szepesvári, C., Cemgil, A. T., and Tomasev, N. Mitigating LLM hallucinations via conformal abstention. *arXiv preprint arXiv:2405.01563*, 2024.

Abdelaal, A., Al Haffar, M. N., Fawzi, M., and Magdy, W. IslamicMMLU: A benchmark for evaluating LLMs on Islamic knowledge. *arXiv preprint arXiv:2603.23750*, 2026.

Al Adel, A., Soliman, A. B., Sawan, M. S., Al-Najjar, R., and Amin, S. BurhanAI at IslamicEval 2025 shared task: Combating hallucinations in LLMs for Islamic content; evaluation, correction, and retrieval-based solution. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, pp. 503--508, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.arabicnlp-sharedtasks.69. URL https://aclanthology.org/2025.arabicnlp-sharedtasks.69/.

Alansari, A. and Luqman, H. AraHalluEval: A fine-grained hallucination evaluation framework for Arabic LLMs. In *Proceedings of the Third Arabic Natural Language Processing Conference*, pp. 148--161, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.arabicnlp-main.12. URL https://aclanthology.org/2025.arabicnlp-main.12/. arXiv:2509.04656.

Alwajih, F., El Mekki, A., Mubarak, H., Hawasly, M., Mohamed, A., and Abdul-Mageed, M. PalmX 2025: The first shared task on benchmarking LLMs on Arabic and Islamic culture. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, pp. 774--789, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.arabicnlp-sharedtasks.107. URL https://aclanthology.org/2025.arabicnlp-sharedtasks.107/. arXiv:2509.02550.

Angelopoulos, A. N. and Bates, S. Conformal prediction: A gentle introduction. *Foundations and Trends in Machine Learning*, 16(4):494--591, 2023. doi: 10.1561/2200000101.

Atif, F., Askarbekuly, N., Darwish, K., and Choudhury, M. Sacred or synthetic? Evaluating LLM reliability and abstention for religious questions. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society (AIES)*, pp. 217--226, 2025. doi: 10.1609/aies.v8i1.36543. arXiv:2508.08287.

Barber, R. F., Candès, E. J., Ramdas, A., and Tibshirani, R. J. The limits of distribution-free conditional predictive inference. *Information and Inference: A Journal of the IMA*, 10(2):455--482, 2021. doi: 10.1093/imaiai/iaaa017.

ElKoumy, M., Allam, K., Tamer, A., and Alqablawi, M. TCE at IslamicEval 2025: Retrieval-augmented LLMs for Quranic and Hadith content identification and verification. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, pp. 515--527, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.arabicnlp-sharedtasks.71. URL https://aclanthology.org/2025.arabicnlp-sharedtasks.71/.

Elrefai, E., Khaled, T., and Soliman, A. ThinkDrill at IslamicEval 2025 shared task: LLM hybrid approach for Qur'an and Hadith question answering. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, pp. 528--533, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.arabicnlp-sharedtasks.72. URL https://aclanthology.org/2025.arabicnlp-sharedtasks.72/.

Farquhar, S., Kossen, J., Kuhn, L., and Gal, Y. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625--630, 2024. doi: 10.1038/s41586-024-07421-0.

Geifman, Y. and El-Yaniv, R. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems*, volume 30, 2017.

Lahmar, A., Arafat, M. E., Farou, Z., and Mahmud, M. IslamTrust: A benchmark for LLMs alignment with Islamic values. In *5th Muslims in ML Workshop, NeurIPS*, 2025. URL https://openreview.net/forum?id=PBcv90iKFB.

Malhas, R., Mansour, W., and Elsayed, T. Qur'an QA 2023 shared task: Overview of passage retrieval and reading comprehension tasks over the holy Qur'an. In *Proceedings of ArabicNLP 2023*, pp. 690--701. Association for Computational Linguistics, December 2023. doi: 10.18653/v1/2023.arabicnlp-1.76. URL https://aclanthology.org/2023.arabicnlp-1.76/.

Manakul, P., Liusie, A., and Gales, M. J. F. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural*

Language Processing, pp. 9004--9017, Singapore, December 2023. Association for Computational Linguistics. URL https://aclanthology.org/2023.emnlp-main.557/.

Mohamed Emad Elden, F. Isnad AI at IslamicEval 2025: A rule-based system for identifying religious texts in LLM outputs. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, pp. 540--559, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.arabicnlp-sharedtasks.74. URL https://aclanthology.org/2025.arabicnlp-sharedtasks.74/.

Mubarak, H., Malhas, R., Mansour, W., Mohamed, A., Fawzi, M., Hawasly, M., Elsayed, T., Darwish, K. M., and Magdy, W. IslamicEval 2025: The first shared task of capturing LLMs hallucination in Islamic content. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, pp. 480--493, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.arabicnlp-sharedtasks.67. URL https://aclanthology.org/2025.arabicnlp-sharedtasks.67/.

Omayrah, A., Alkhereyf, S., Abdelali, A., Al-Thubaity, A., Kuriakose, J., and AbdulMajeed, I. HUMAIN at IslamicEval 2025 shared task 1: A three-stage LLM-based pipeline for detecting and correcting hallucinations in Quran and Hadith. In *Proceedings of the Third Arabic Natural Language Processing Conference: Shared Tasks*, pp. 509--514, Suzhou, China, November 2025. Association for Computational Linguistics. doi: 10.18653/v1/2025.arabicnlp-sharedtasks.70. URL https://aclanthology.org/2025.arabicnlp-sharedtasks.70/.

Vovk, V., Gammerman, A., and Shafer, G. *Algorithmic Learning in a Random World*. Springer, New York, 2005.

A. IslamicEval Subtask 1: task definition and worked example

This appendix reproduces the official Subtask 1 definition from the IslamicEval 2025 task page (Mubarak et al., 2025) and walks through one worked example, both to anchor the terminology used in the main body and to make the paper self-contained for readers unfamiliar with the shared task.

A.1. Three-subtask pipeline

Subtask 1 decomposes a free-text LLM response to a religious question into three sequential sub-subtasks; some quotations of Qur'anic verses (Ayahs) or Prophetic Hadiths in the response may be hallucinated. **Subtask 1A**: Identification of intended Ayahs and Hadiths. Identify the character-level (start, end) spans of each "intended" Ayah or Hadith citation, excluding Surah/Ayah numbers, narrators, and punctuation. Evaluation: span-level precision/recall. **Subtask 1B**: Validation of content accuracy. Classify each 1A span as *correct* or *incorrect* relative to canonical Islamic sources (QCRI-provided Qur'an text and Hadith collections); incorrect diacritics count as errors. Evaluation: classification precision/recall. **Subtask 1C**: Correction of erroneous content. For each 1B-incorrect span, return the canonical wording (and diacritics), or the refusal label *"wrong"* (خطأ) when no canonical source matches. The task page specifies WER, but the released leaderboard and public scoring script use exact-match after default-diacritic normalization; we follow that.

A.2. Worked example

Question: ما الدلالة على أن الأنبياء والرسل لا يعلمون الغيب؟ (*"What is the evidence that the prophets and messengers do not know the unseen?"*)

LLM-generated answer (excerpts; Arabic with inline citations). The model cites two Qur'anic verses and one Hadith; each is a Subtask 1A candidate span, labeled ①--③ below.

① (claimed Ayah, *correct*): قُل لا يَعلَمُ مَن فِي السَّماواتِ وَالأَرضِ الغَيبَ إِلَّا اللَّهُ (Surah 27:65).

② (claimed Hadith, *incorrect* as cited): مفاتيح الغيب خمس لا يعلمهن إلا الله (narrated by al-Bukhari).

③ (claimed Ayah, *incorrect* as cited): وَما كانَ لِسُلَيمانَ نَفَقَةٌ وَلا هُوَ يَعلَمُ ما في غَيبِ السَّماواتِ وَالأَرضِ وَأَنَّهُ عَلِمٌ بِما تَعمَلونَ (claimed Surah 27:16).

The three subtask outputs on this example:

1A (Identification). Locate ①--③ in the character stream.

1B (Validation). ① correct (canonical 27:65); ② incorrect (canonical wording differs on one pronoun); ③ incorrect (no canonical verse matches).

1C (Correction). Return a canonical correction or *"wrong"* per incorrect span:

- ②: canonical Hadith مفاتيح الغيب خمس لا يعلمها إلا الله --- a one-pronoun fix (يعلمها for يعلمهن).
- ③: no canonical verse matches; gold is *"wrong"* (خطأ).

Subtask 1C in the dev set repeats this last step over n=179 flagged spans; our abstention layer withholds when ensemble disagreement is high.

B. Prompt template for the three cross-family runs

The three added LLMs (Claude Opus 4.7, Qwen3-235B-a22b, Llama-3.3-70B) were called via OpenRouter at temperature T=0 with the following prompt template, populated with the truncated `Original Span` and `Label` fields from the dev TSV:

You are given an Arabic text that claims to be a Qur'anic verse (Ayah) or a Hadith but contains errors. Provide the CORRECT canonical text (in Othmani script for Qur'an, exact wording for Hadith).

If the text cannot be matched to any real verse or hadith (i.e. it is wholly fabricated), respond with exactly: خطأ

Original (erroneous) text: `{original}`
Type: `{label}` (`WrongAyah` = Qur'anic verse, `WrongHadith` = Hadith from one of the six canonical collections)

Respond with ONLY a JSON object (no markdown fences, no preamble):
{ `"rationale"`: "<brief explanation>",
`"correction"`: "<canonical text in Arabic, OR خطأ if not correctable>" }

We then extract the `correction` field and apply the same `remove_default_diac` normalization the QCRI scoring script uses, ensuring our three additional outputs are processed identically to BurhanAI's released outputs.

C. Sample IslamicEval Subtask 1C spans

To illustrate the diversity of failure modes the abstention layer handles, we walk through three spans. Each box names the span, shows the erroneous quotation and the gold correction, lists the seven variant outputs, and reports our layer's decision (commit the majority m, or abstain) at the calibrated threshold $\hat{\tau}$ used for the headline operating point in Figure 2.

Case A: confident refusal (gold = "wrong"; $s = 0/7 \rightarrow$ commit)

Span: `C-Q20`, hallucinated hadith.
Erroneous quotation: لَوْ كُنتُ مُكَلِّمُكَ (*a fabricated hadith with no canonical source*).
Gold correction: *"wrong"* (no canonical hadith matches the original).
K=7 outputs: all seven variants emit *"wrong"* (خطأ). Agreement $a = 7$, so $s = (K - a)/K = 0$.
Decision: commit majority (*"wrong"*) --- **correct refusal**.

Case B: subtle diacritic edge case ($s = 3/7 \rightarrow$ abstain)

Span: `C-Q19`, wrong Qur'anic ayah.
Erroneous quotation (truncated): ولقد أرسلنا ...نتم مؤمنين.
Gold correction (Sūrah 7:85 fragment): وَإِلَىٰ مَدْيَنَ أَخَاهُمْ شُعَيْبًا ۗ ...إِن كُنتُم مُّؤْمِنِينَ.
K=7 outputs: four BurhanAI GPT-5 variants return the same canonical text but with differing shadda diacritics; the three cross-family variants (Claude, Qwen, Llama) all return *"wrong"*. Majority agrees on the (near-)canonical text with count $a = 4$, giving $s = 3/7$.
Decision: $s = 3/7$ exceeds the $\hat{\tau}$ at the 14% operating point, so the layer **abstains** rather than commit a diacritically incorrect answer that would fail the exact-match metric.

Case C: full hallucination split ($s = 4/7 \rightarrow$ abstain)

Span: `C-Q35` (also Figure 1), hallucinated hadith with a canonical correction.
Gold correction: فيقول: من يدعوني فأستجيب له، من يسألني فأعطيه، من يستغفرني فأغفر له.
K=7 outputs: three variants (gpt-5-nano, Qwen, Llama) emit *"wrong"*; three (gpt-5 low+codex, gpt-5 high+codex, Claude-Opus-4.7) fabricate three *different* hadiths from the nightly-descent tradition; one (gpt-5 low) returns the canonical gold. Majority $m =$*"wrong"*, $a = 3$, $s = 4/7$.
Decision: the leaderboard's no-abstention rule commits *"wrong"* and misses the canonical fix (which a single variant found). Our layer **abstains** instead, surfacing the high-uncertainty span to a downstream human reviewer rather than emitting a confidently incorrect refusal.

These three patterns --- confident agreement on refusal (Case A), partial agreement with diacritic-level disagreement (Case B), and ensemble split across refusal, fabrication, and gold (Case C) --- are typical of the Subtask 1C dev set. The calibrated threshold $\hat{\tau}$ converts the disagreement signal s into a deployer-tunable commit/abstain decision; the headline Pareto in Figure 2 aggregates this case-by-case behavior across all $n = 179$ dev spans.