

Knowledge Boundaries Are Not Defined by Prompts: Improving the Reachability of Knowledge through Question-Level Probing

Wang Bin, Yilu Zhong, Shuping Chen, Hui Li*

Guangdong Provincial Key Laboratory of
Ultra High Definition Immersive Media Technology,
Shenzhen Graduate School, Peking University

* Correspondence: lih64@pku.sz.edu.cn

Abstract

Safe and reliable large language models should abstain when a question falls beyond their knowledge boundary. Existing abstention methods often use prompt-level uncertainty signals, such as semantic entropy or sampling consistency, as proxies for this boundary, treating instability under the current prompt as evidence that the question is unanswerable. We argue that this prompt-level view is overly conservative: the same underlying question can exhibit different stability under different formulations, and high uncertainty under one prompt may indicate that the answer has not been stably reached rather than that the model lacks the relevant knowledge. We propose a test-time question-level probing framework that uses existing prompt-level uncertainty signals rather than introducing a new uncertainty metric. When the original prompt is unstable, the framework probes semantically equivalent reformulations and accepts an answer only when multiple stable views provide consistent low-entropy evidence. This observable behavior serves as an operational proxy for question-level answerability under a fixed model, sampling configuration, and reformulation set. The method requires no target-model training or parameter updates and is compatible with semantic entropy and its variants. Experiments show that question-level probing recovers prompt-sensitive answers incorrectly rejected by prompt-level methods and improves the risk-coverage trade-off with controlled error rates. The code is available at [GitHub repository](#).

1 Introduction

As large language models are increasingly integrated into practical applications, ensuring their reliability has become a central concern (Wang et al., 2026a). In factual question answering, they may produce fluent but factually incorrect answers, a phenomenon known as hallucination (Ji et al.,

2023; Wen et al., 2025; Maynez et al., 2020). Reliable systems must therefore recognize the boundaries of their knowledge and abstain when a question falls beyond them (Yin et al., 2024). Uncertainty quantification provides a central signal for this decision, with common approaches including self-evaluation, sampling consistency, and black-box uncertainty estimation (Kadavath et al., 2022; Manakul et al., 2023; Huang et al., 2025). Semantic entropy has become a prominent tool in this space: it measures the dispersion of sampled responses over semantic equivalence clusters, reducing sensitivity to superficial wording differences when estimating hallucination risk and response reliability (Kuhn et al., 2023; Farquhar et al., 2024; Kossen et al., 2024; Nikitin et al., 2024).

However, existing abstention methods largely make decisions from the model’s behavior under the current prompt. Uncertainty-based selective answering methods typically abstain when a risk score exceeds a threshold (Farquhar et al., 2024; Abbasi-Yadkori et al., 2024; Tjandra et al., 2024; Phillips et al., 2026), while methods based on supervised fine-tuning (Zhang et al., 2024), preference optimization (Cheng et al., 2024), or multi-model collaboration (Feng et al., 2024) require additional annotation, training, or inference cost. Abstention has also been studied from survey and benchmark perspectives (Wen et al., 2025; Madhusudhan et al., 2025). Despite these advances, such methods typically treat instability under the current prompt as evidence that the question should not be answered. We argue that this prompt-level boundary can be overly conservative: high semantic entropy may indicate not that the model lacks the relevant knowledge, but that the current prompt fails to stably elicit it. As a result, prompt-level uncertainty may place some prompt-sensitive but recoverable questions outside the answerable boundary.

A more fundamental issue lies in the granular-

ity of the decision. Standard semantic entropy and Conformal Abstention make abstention decisions at the prompt level, while Semantic Density (Qiu and Miikkulainen, 2024) and verbalized confidence (Tian et al., 2023) rank or calibrate outputs at the response or confidence level. In all cases, the object of observation remains a single prompt or a single response. This is limiting because the same underlying question can be expressed in multiple ways, and model behavior may vary with the formulation or interaction context (Xie et al., 2024; Wang et al., 2026b). For example, the model may answer “Steve Jobs and Steve Wozniak” stably for “Who founded Apple?” but drift among variants including or omitting Ronald Wayne under another equivalent formulation. A prompt-level method would treat the unstable formulation as high-entropy and abstain, even though the underlying question may still be answerable.

To address this issue, we propose a test-time framework for question-level probing and selective answering. Motivated by prior work showing that answer stability can depend on prompt formulation (Kadavath et al., 2022; Li et al., 2025), we shift abstention from the current prompt to the underlying question. The framework answers directly when the original prompt yields stable responses; otherwise, it probes semantically equivalent reformulations and accepts an answer only when they provide consistent low-entropy evidence that the answer can be stably reached. If no equivalent view recovers such evidence, the system abstains. Rather than introducing a new uncertainty metric, the framework computes KLE independently within each prompt and combines conditional triggering, equivalence filtering, and cross-view consistency. Under a fixed model, sampling configuration, and reformulation set, we operationalize reachability as whether multiple stable views consistently support the same answer, using this observable behavior to approximate the question’s latent answerability. The framework is training-free with respect to the target model, requiring no target-model training or parameter updates, and it remains compatible with multiple semantic entropy implementations.

Our main contributions are as follows:

(1) We study questions that are unstable under the original prompt but may still be recoverable, and propose a systematic abstention-recovery approach that reduces their misclassification as unanswerable by prompt-level methods.

(2) We introduce a question-level boundary probing framework that lifts abstention decisions from a single prompt or response to multiple equivalent formulations of the same underlying question, using consistent low-entropy evidence to assess answer reachability.

(3) We design a denoising pre-screening mechanism based on semantic entropy decomposition, which exits early when a stable dominant answer can be identified and thereby reduces the average cost of reformulation-based recovery.

(4) We analyze which paraphrase characteristics support correct-answer recovery, showing that clear, standard, or suitably terminology-aware formulations can reveal stable answer evidence missed under the original prompt.

2 Related Work

2.1 Knowledge Boundaries

Reliable large language models need to recognize the boundaries of their knowledge and abstain when a question falls beyond them. Existing work has studied this capability from the perspectives of training, collaboration, and evaluation. Some methods construct unknown or unanswerable data, apply supervised fine-tuning, or use preference optimization to teach models to answer “I don’t know” when uncertain (Zhang et al., 2024; Cheng et al., 2024). Another line of work identifies model knowledge gaps through multi-model collaboration, thereby triggering abstention in potential hallucination scenarios (Feng et al., 2024). In addition, prior studies have systematically surveyed abstention in large language models or introduced dedicated benchmarks for evaluating abstention on unanswerable questions (Wen et al., 2025; Kirichenko et al., 2025). A recent survey on knowledge boundaries further points out that some knowledge may exist in model parameters but cannot always be reliably used under arbitrary prompt formulations (Li et al., 2025). These works mainly focus on how to identify or handle questions beyond the boundary, whereas we study a more fundamental question: whether the uncertainty boundary induced by the current prompt is equivalent to the reachability boundary of the underlying question.

2.2 Entropy-Based Boundaries

Sampling-based uncertainty estimation assesses whether model responses remain consistent across

multiple generations, and has become an important approach for factuality detection and selective answering. Representative black-box methods judge generation reliability through cross-response consistency checks (Manakul et al., 2023; Huang et al., 2025). Semantic entropy further lifts consistency from the string level to the semantic level, measuring response stability under the current prompt by computing entropy over semantic equivalence clusters (Kuhn et al., 2023; Farquhar et al., 2024). Subsequent work has developed semantic entropy probes and kernel language entropy to improve estimation efficiency or capture finer-grained semantic similarity (Kossen et al., 2024; Nikitin et al., 2024). These methods typically treat stability under the current prompt as evidence of answer reliability. In contrast, we ask whether the knowledge boundary induced by prompt-level stability is overly conservative, and further probe whether the underlying question behind a high-entropy prompt can still be stably reached. Alignment training may also alter the sampling distribution, making the relationship between stability signals and factual knowledge more complex (Liu, 2026).

2.3 Question-Level Probing

Existing selective answering methods typically estimate risk for a given prompt or a given response, and trade off coverage against error rate. Self-evaluation methods prompt the model to judge whether its own answer is correct, thereby obtaining response-level confidence (Kadavath et al., 2022). Verbalized confidence methods study how to elicit calibrated confidence scores from models aligned with human feedback (Tian et al., 2023). Representation-based methods use hidden states or activation features to predict answer truthfulness (Azaria and Mitchell, 2023; Chen et al., 2024; Lavi et al., 2026). In addition, confidence estimation in semantic space and conformal abstention improve selective answering from the perspectives of response-level semantic density and calibration guarantees, respectively (Qiu and Miikkulainen, 2024; Abbasi-Yadkori et al., 2024). Stability across equivalent formulations is also relevant: Elazar et al. (2021) expose inconsistent factual predictions across paraphrased templates; Li and Huai (2025) use perturbations to estimate uncertainty in generated explanations; Kumar et al. (2025) use uncertainty to select chain-of-thought demonstrations; and Cox et al. (2025) sample paraphrases to calibrate prompt-sensitive models and decompose

uncertainty. Our goal is different: we compute KLE independently within each prompt, invoke reformulation only after an uncertainty trigger, filter for semantic equivalence, and accept an answer only from cross-view consistent low-entropy evidence, rather than pooling rewrites into a new uncertainty estimator.

3 Method

Our method uses semantic stability under a single prompt as its basic signal. When the original prompt is unstable, it expands the evaluation to a set of semantically equivalent prompts, so that answer-or-abstain decisions are made at the question level rather than for one prompt alone. We first introduce the semantic stability score used in this work, and then describe the three-stage question-level abstention framework built on top of it.

3.1 Semantic Stability Estimation

Given a prompt p , we independently sample N responses $\{a_1, \dots, a_N\}$ from the model $\mathcal{M}$, and compare whether these responses semantically point to the same answer. Let $s(a_i, a_j)$ denote the semantic similarity between two responses. By default, we compute this similarity using a bidirectional entailment score from an NLI model, which captures answer equivalence across different surface forms. The pairwise similarities form a kernel matrix

$$\mathbf{K} \in \mathbb{R}^{N \times N}, K_{ij} = s(a_i, a_j) \quad (1)$$

On top of this similarity structure, we use Kernel Language Entropy (KLE) (Nikitin et al., 2024) to convert pairwise semantic similarity into a continuous uncertainty score. As illustrated in Figure 1, sampled responses can be viewed as nodes in a graph, and NLI similarities define semantic connections between them. These connections are used to construct and normalize a semantic kernel matrix. After eigendecomposition of the normalized kernel, KLE is defined as the spectral entropy:

$$\mathbf{K}' = \mathbf{K} / \text{tr}(\mathbf{K}), \text{KLE}(p) = - \sum_{j=1}^N \lambda_j \log \lambda_j \quad (2)$$

where $\{\lambda_j\}_{j=1}^N$ are the eigenvalues of $\mathbf{K}'$ and satisfy $\sum_j \lambda_j = 1$. When responses concentrate on a small number of semantic modes, the spectrum is also concentrated, KLE is low, and the model is treated as more certain under the prompt. When

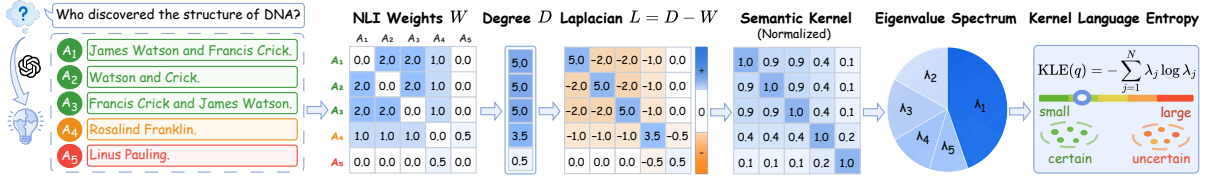


Figure 1: Semantic stability estimation with KLE. Responses sampled from the same prompt are connected by NLI-based semantic similarities, which form a normalized semantic kernel. A concentrated eigenvalue spectrum indicates higher prompt-level certainty, while a dispersed spectrum indicates higher uncertainty.

responses spread across multiple conflicting or unrelated semantic modes, the spectrum is more dispersed, KLE is high, and the model is treated as more uncertain.

The question-level framework below treats the prompt-level uncertainty estimator as a modular component. For any prompt p , the framework only requires an uncertainty score $u(p)$ and a dominant answer $a(p)$ extracted from the sampled responses. In this work, we instantiate $u(p)$ with $\text{KLE}(p)$; lower values indicate that responses concentrate on fewer semantic modes and thus reflect higher prompt-level certainty. We further use $\rho(p)$ to denote the strength of the dominant semantic mode. For the original user prompt p_0 , we write $u_0 = u(p_0)$, $a_0 = a(p_0)$, and $\rho_0 = \rho(p_0)$.

3.2 Question-Level Framework

Before invoking question-level paraphrase probing, the framework first checks whether the original prompt already provides enough evidence for an answer. The first case is direct acceptance: if u_0 is below the threshold τ , the original prompt stably points to its dominant answer, and the framework returns a_0 . The second case is early-exit acceptance: some examples have non-negligible overall uncertainty but still contain a clear dominant semantic mode, with disagreement mainly coming from a small number of scattered responses. For such examples, Variant B accepts a_0 when the dominant-mode strength ρ_0 exceeds a threshold and the remaining responses do not form a competing stable answer, avoiding unnecessary paraphrase probing. Variant A skips this rule and sends all examples that fail direct acceptance to the paraphrase probing stage.

If the original prompt is not accepted by the above rules, the framework enters question-level paraphrase probing. This stage generates a set of prompts $\{p_1, \dots, p_K\}$ that are semantically equivalent to p_0 , and estimates $u_k = u(p_k)$ and $a_k = a(p_k)$ for each prompt. These prompts provide dif-

ferent surface views of the same underlying question; the framework returns an answer only when multiple views provide low-uncertainty and mutually consistent evidence, and abstains otherwise. The next section describes paraphrase generation, equivalence verification, and aggregation.

3.3 Probing and Aggregation

For examples not accepted at the original-prompt stage, the framework probes other surface views of the same underlying question through semantically equivalent paraphrases. A paraphraser generates K candidate prompts $\{p_1, \dots, p_K\}$ from p_0 , preserving the queried subject, factual attribute, and expected answer type while changing only the wording. To cover plausible user formulations, we guide generation with five lightweight roles: *Layperson*, *Student*, *Expert*, *Researcher*, and *Wildcard*. The full prompts are provided in Appendix B.

Paraphrase probing is designed to vary the surface wording while preserving the question-level semantic core. In our factual QA setting, this means preserving the queried subject, factual attribute, and expected answer type. Since the paraphraser may introduce semantic drift, the framework verifies each candidate prompt for equivalence and keeps it only if it points to the same underlying fact as p_0 . Let $\mathcal{P}^+(p_0)$ denote the set of prompts that pass this verification. For each $p_k \in \mathcal{P}^+(p_0)$, the framework resamples responses and computes $u_k = u(p_k)$ and $a_k = a(p_k)$ as in Section 3.1. In this way, Stage 3 collects stability evidence for the same question under different surface views, rather than expanding or replacing the question itself.

After re-estimation, the framework looks for stable evidence that is mutually supported across verified paraphrases. Let $\mathcal{S} = \{k : p_k \in \mathcal{P}^+(p_0), u_k < \tau\}$ denote the set of low-uncertainty paraphrases. The framework then checks whether their dominant answers agree. If some answer

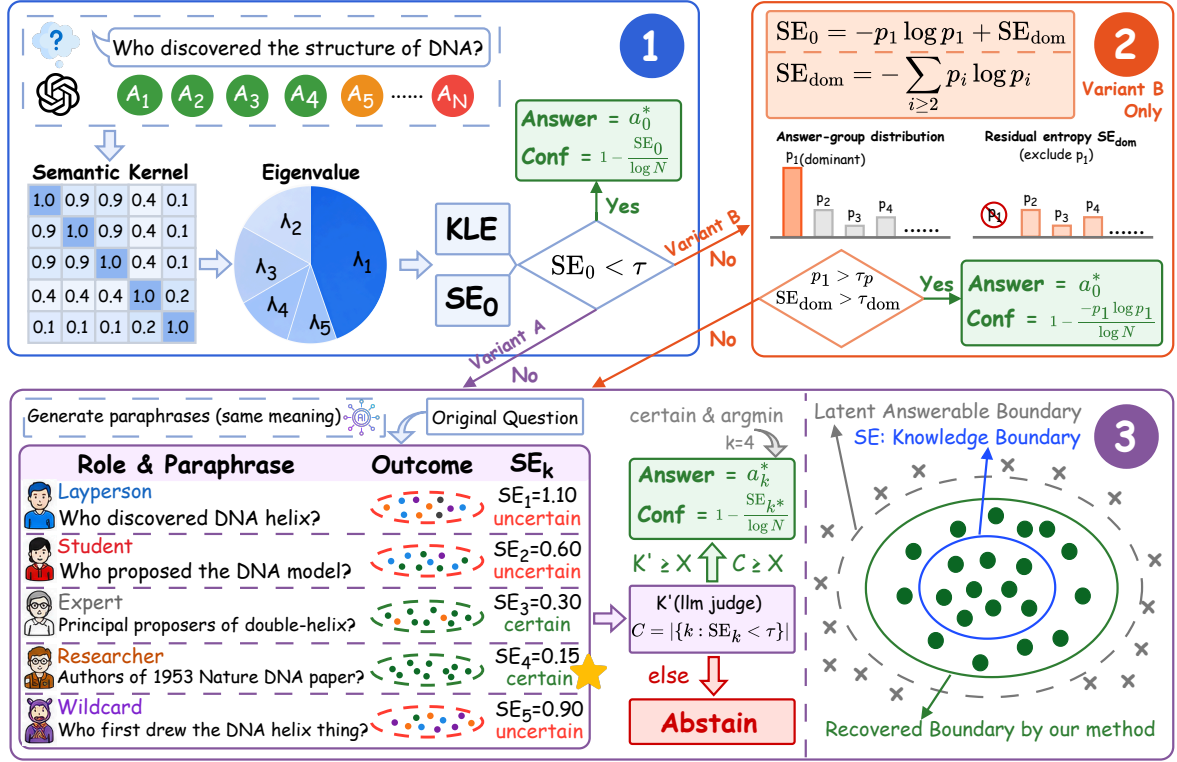


Figure 2: Overview of the question-level abstention framework. The method evaluates the original prompt, optionally applies a dominant-answer early exit in Variant B, and then probes semantically equivalent prompts when needed.

is supported by at least X prompts in $\mathcal{S}$, the framework treats the underlying question as stably answerable across multiple surface views, and returns the dominant answer from the lowest-uncertainty supporting view. If there are too few low-uncertainty views, or if stable views give conflicting answers, the available evidence is insufficient to justify an answer, and the framework abstains.

3.4 Decision Rule and Variants

Putting the stages together, let E_0 denote the early-exit condition for the original prompt, where the dominant mode is strong enough and no competing stable answer is present. The framework returns

$$\text{Output}(p_0) = \begin{cases} a_0, & u_0 < \tau, \\ a_0, & \text{Variant B and } E_0, \\ a_{k^*}, & \exists A : |\mathcal{S}_A| \geq X, \\ \text{Abstain}, & \text{otherwise} \end{cases} \quad (3)$$

where $\mathcal{S}_A = \{k \in \mathcal{S} : a_k = A\}$ is the set of low-uncertainty paraphrases that support answer A , and $k^* = \arg \min_{k \in \mathcal{S}_A} u_k$ selects the most stable supporting view. The first line is direct ac-

ceptance under the original prompt, the second line is the dominant-answer early exit enabled only in Variant B, and the third line is question-level recovery through equivalent prompts. Variant A skips the second line, while Variant B uses the full rule. If none of these evidence paths succeeds, the framework abstains.

4 Experiments

4.1 Experimental Setup

Benchmarks. We evaluate on three complementary factual QA benchmarks. TriviaQA (Joshi et al., 2017) contains human-written trivia-style short-answer questions. Natural Questions (Kwiatkowski et al., 2019) contains real Google search queries with short answers. PopQA (Mallen et al., 2023) is constructed from Wikidata triples and stratifies entities by popularity to cover long-tail knowledge. We use the official validation splits for TriviaQA and Natural Questions, and the 14k-question set for PopQA. For each dataset, we sample 200 questions as a labeled development set for calibrating τ , and evaluate on a disjoint set of 1000 questions under three fixed random seeds. We report means and standard deviations.

Models. We evaluate six instruction-tuned open models from three families and sizes from 7B to 70B: Llama-3-8B-Instruct, Llama-3.1-70B-Instruct, Qwen2.5-7B-Instruct, Qwen2.5-14B-Instruct, Mistral-7B-Instruct-v0.3, and Mistral-Small-3.1-24B-Instruct. This setting tests transfer across model scales and architectures.

Baselines. We compare against six baselines covering five common uncertainty estimation mechanisms: P(True) (Kadavath et al., 2022) for self-evaluation, SelfCheckGPT (Manakul et al., 2023) for sampling consistency, SE (Farquhar et al., 2024) and KLE (Nikitin et al., 2024) for semantic-entropy-based uncertainty estimation, Conformal Abstention (Abbasi-Yadkori et al., 2024) for calibration-based abstention, and Verbalized Confidence (Tian et al., 2023) for verbalized self-confidence. We include both SE and KLE to separate the semantic-entropy signal from its specific implementation.

Metrics. We use AUROC to evaluate how well confidence separates correct from incorrect answers, and AUGRC to measure the overall risk-coverage trade-off of selective answering. Given confidence-correctness pairs $\{(c_i, y_i)\}_{i=1}^M$, where $y_i \in \{0, 1\}$, AUROC is defined as

$$\text{AUROC} = \Pr(c_i > c_j \mid y_i = 1, y_j = 0) \quad (4)$$

After sorting examples by confidence in descending order, let c denote coverage, namely, the fraction of all evaluation examples retained for answering, and let $R(c)$ denote the cumulative risk at that coverage, computed as the number of incorrect accepted answers divided by the total number of evaluation examples. We evaluate this empirical risk-coverage curve at the distinct confidence levels and compute its area using trapezoidal integration:

$$\text{AUGRC} = \int_0^1 R(c) dc \quad (5)$$

Higher AUROC and lower AUGRC indicate better performance. Correctness labels are determined by LLM-as-judge and Exact Match, with human verification on 200 stratified samples.

Implementation Details. For each prompt, we sample $N = 10$ responses at temperature $T = 1.0$ with a maximum length of 100 tokens. Semantic clustering uses bidirectional entailment from DeBERTa-large-MNLI, while prompt rewriting and equivalence judging are performed by Claude Opus 4.7. The threshold τ is calibrated

on the development set of each model-dataset pair by scanning $[0.1, 2.0]$ to maximize AUROC. Variant B uses dominant-mode strength and residual-answer stability for early exit, with $\tau_p = 0.5$ and paraphrase aggregation threshold $X = 2$ by default. Replacing the paraphraser with Qwen2.5-14B-Instruct or Qwen2.5-3B-Instruct yields AUGRC values of 0.110 and 0.108 (versus 0.106 with Claude) at only $0.28\times$ and $0.11\times$ the end-to-end cost, showing that the framework does not require an expensive proprietary paraphraser (Appendix Table 6). We use Claude uniformly in the main experiments to control candidate-generation variation and stably produce diverse role-based reformulations, not because it is required by the method.

4.2 Main Results

Figure 3(a) shows the Risk-Coverage curves on the three datasets. Our framework maintains lower risk in the low-coverage region, and its advantage does not quickly disappear as coverage increases. This suggests that the gain does not only come from selecting the easiest questions more accurately, but also from recovering cases where equivalent formulations yield stable answer evidence despite instability under the original prompt. The trend on PopQA is consistent with TriviaQA and Natural Questions, indicating that the mechanism remains effective for long-tail knowledge questions.

Appendix Table 4 reports the corresponding aggregate AUROC and AUGRC results, and Appendix Tables 10, 11, and 12 further provide per-model results. Overall, our framework outperforms prompt-level baselines across the three datasets and most models. These results show that extending the decision from a single prompt to question-level equivalent-prompt probing improves the ranking quality and risk control of selective answering across the evaluated settings.

At the same mean budget of 29.6 calls, Variant B obtains 0.106 AUGRC, compared with 0.145 for budget-matched original-prompt self-consistency, 0.116 for the Cox et al. (2025) adaptation, and 0.119 for majority voting over rewrites (Appendix Table 5). The reformulation-based methods share the same candidates; Variant B also obtains 0.850 AUROC and 31.8%/4.6% correct/wrong recovery over initially abstained cases, showing that additional sampling or generic reformulation aggregation alone does not explain its gains.

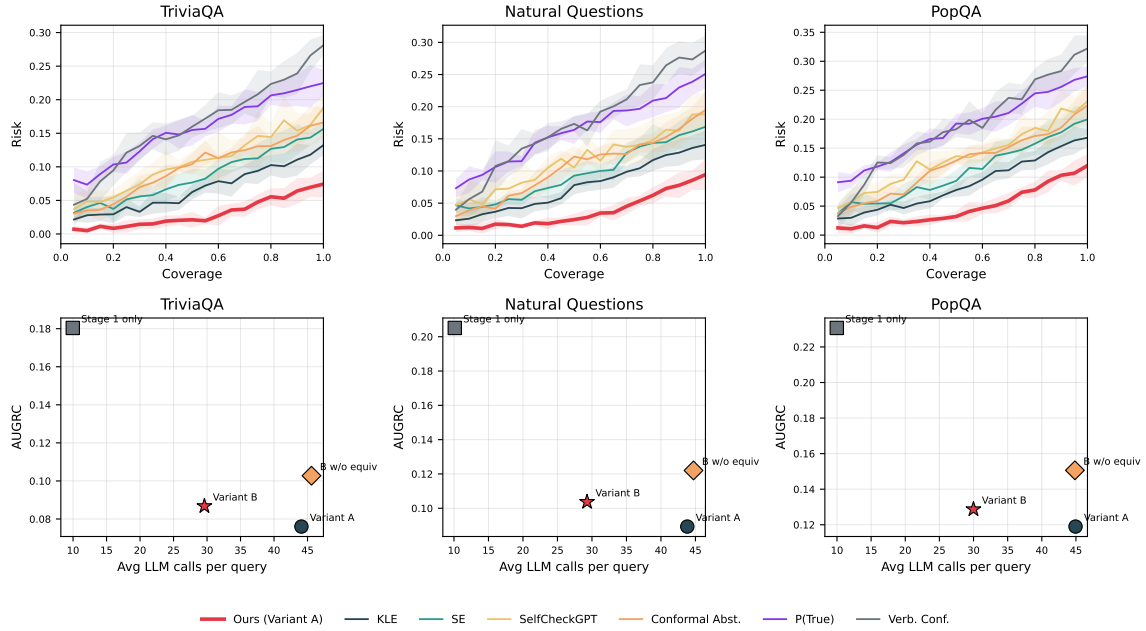


Figure 3: Experimental results for Section 4. (a) Risk-Coverage curves on the three datasets, showing the error risk of each method at different coverage levels. (b) Cost-AUGRC comparison for the structural ablation configurations, where the x-axis is the average number of LLM calls per query and the y-axis is AUGRC.

4.3 Ablation Study

To isolate the contribution of each decision mechanism, we compare four configurations. **Stage1Only** makes the accept-or-abstain decision only from u_0 under the original prompt p_0 , without question-level probing. **Variant A** applies equivalent-prompt probing to questions that are not directly accepted by $u_0 < \tau$. **Variant B** further enables the dominant-answer early-exit condition E_0 . **VariantB-NoEquiv** disables equivalence verification and sends all candidate paraphrases to resampling.

Figure 3(b) shows the cost-AUGRC trade-off among these configurations. Stage1Only has the lowest cost but the highest risk, whereas Variant A has the lowest risk but the highest cost. Relative to Variant A, Variant B reduces mean calls, mean tokens, and P50 latency by 33.2%, 35.3%, and 50.0%, respectively; however, its P95 calls remain 61 and P95 latency remains 18.7 s, so early exit improves average rather than tail cost. VariantB-NoEquiv has higher risk, supporting the role of equivalence verification in filtering semantic drift. Full results are in Appendix Tables 7 and 8.

5 Analysis

5.1 Question-Level Gain

Figure 4(a) shows the outcome distribution. The stable-wrong rate—the fraction of finally accepted answers that have low KLE but are incorrect—is 8.4%, 8.5%, and 8.3% for Stage1Only, Variant A, and Variant B, respectively. Relative to Stage1Only, Variants A and B recover 7.4 and 6.9 correct answers per additional wrong answer, showing no measurable amplification of stable errors (Appendix Table 9). Figure 4(b) shows that our framework retains more answerable examples as the target accuracy increases.

5.2 Original-Prompt Unstable Subset

We define questions with $u_0 > \tau$ as the original-prompt unstable subset. This subset only indicates that sampled answers are unstable under the current formulation; it does not imply that the model lacks the relevant knowledge. Figure 4(c) shows that prompt-level baselines have substantially lower accuracy on this subset. In contrast, our framework still recovers part of the correct answers. This suggests that original-prompt uncertainty alone is insufficient to treat these questions as unanswerable, since equivalent formulations can provide more stable answer evidence.

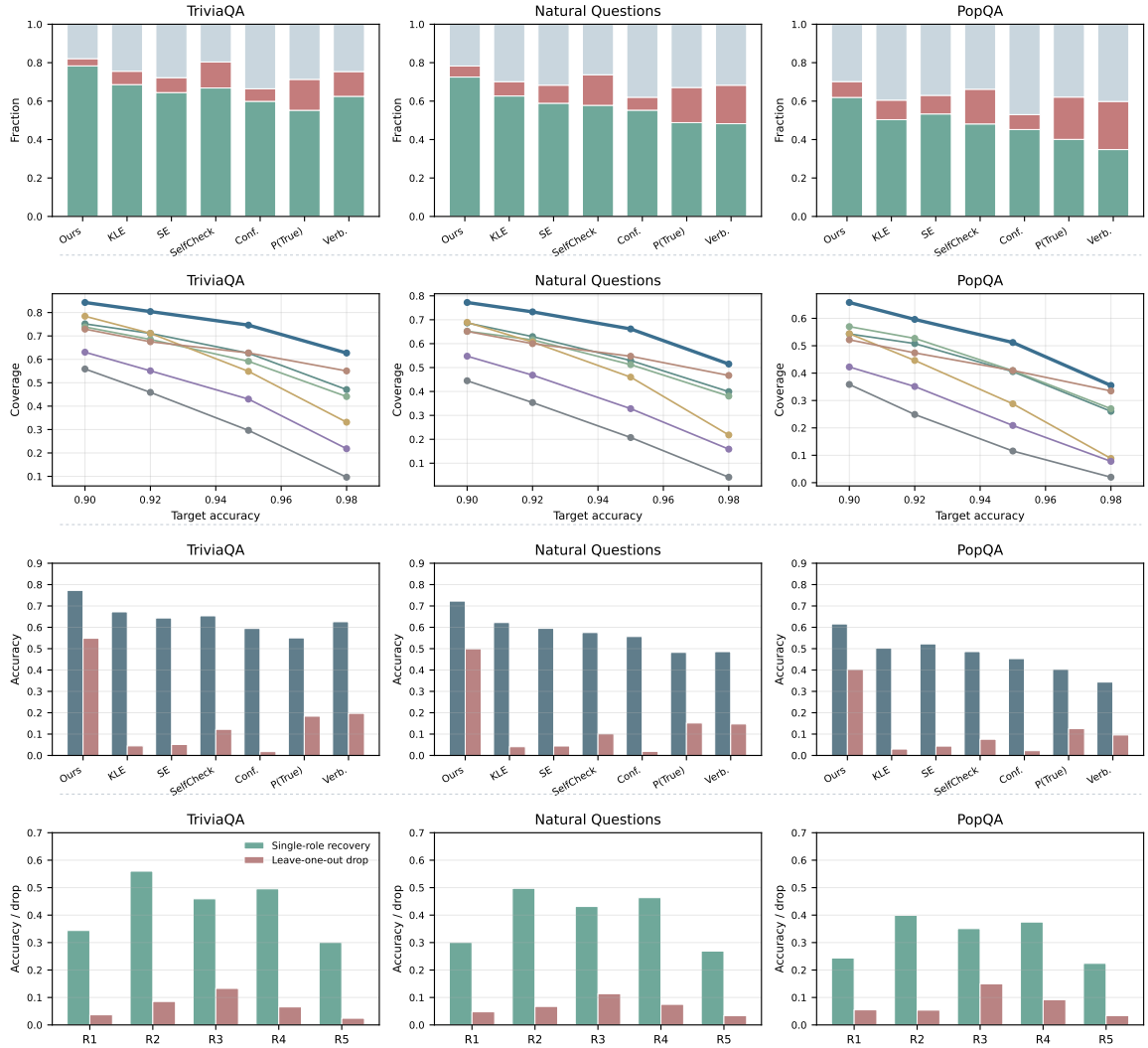


Figure 4: Question-level analysis across TriviaQA, Natural Questions, and PopQA. Columns correspond to datasets, and rows correspond from top to bottom to panels (a)–(d). Panel (a) shows the outcome distribution for each method; panel (b) shows coverage at fixed target accuracy; panel (c) compares accuracy on the full set and the original-prompt unstable subset, with the two bars within each method corresponding to these two subsets in that order; and panel (d) reports single-role recovery and leave-one-out drop for roles R1–R5.

5.3 Paraphrase Role Analysis

Figure 4(d) compares the contribution of the five paraphrase roles. Single-role recovery measures the recovery ability when using only one role, while leave-one-out drop measures the performance loss after removing that role. Student and Researcher have higher single-role recovery, suggesting that clear, standard, or moderately contextualized formulations are more likely to produce stable answers. Expert has a larger leave-one-out impact, suggesting that domain terminology provides complementary information. Layperson and Wildcard contribute less individually, but still help increase expression diversity. Overall, a reformulation must not change the queried fact

or introduce answer-bearing information; its role is to reduce formulation ambiguity and clarify the expected answer type under semantic equivalence. Appendix B.1 gives admissible and rejected cases grounded in questions from the evaluation datasets.

A blinded, stratified audit of 300 pairs across 3 datasets, 5 roles, and 3 paraphraser finds equivalence-judge precision/recall/F1 of 97.2/94.1/95.6%; filtering raises strict equivalence from 94.7% to 99.0% and reduces semantic drift from 5.3% to 1.0% (Appendix D). We observe 0/300 factual injections (an approximate 95% event-rate upper bound of 1%) and no direct answer leakage by automatic matching. Using

only human-verified clean reformulations yields 0.108 AUGRC, still better than KLE’s 0.157.

6 Conclusion

This paper lifts abstention decisions in factual question answering from individual prompts to underlying questions, and proposes a test-time framework for question-level probing and selective answering. The framework first evaluates response stability under the original prompt; when the original prompt is unstable, it probes semantically equivalent reformulations of the same question, outputs an answer if stable and consistent evidence is found, and abstains otherwise. Under a fixed experimental setting, this observable cross-view evidence serves as a proxy for the question’s latent answerability. The early-exit mechanism further avoids unnecessary reformulation calls for some examples, reducing average inference cost. Experiments across three factual QA benchmarks and six open-source models show that our method improves selective answering quality and answer accuracy on the original-prompt unstable subset over prompt-level baselines.

Acknowledgements

This work was supported by the Pkusz-China Mobile Internet Co. Ltd. Joint Laboratory (CMIC-202400488) and financially supported by Guangdong Provincial Key Laboratory of Ultra High Definition Immersive Media Technology (Grant No. 2024B1212010006). We used generative AI tools only for language editing, formatting checks, and limited assistance with experimental code development. All scientific judgments, code review, experimental validation, and final manuscript content are the responsibility of the authors.

Limitations

This framework has several limitations. First, because it relies on semantic entropy, it cannot identify answers that are stable but incorrect; doing so requires external factual verification. Second, recovery depends on the quality of the paraphraser and equivalence verifier; low diversity or semantic drift in reformulations can weaken the evidence. Third, high-uncertainty cases require additional reformulation, verification, and sampling; Variant B’s P95 latency remains 18.7 s, so the method targets reliability-oriented settings that permit additional test-time computation. We evaluate the

framework on factual QA, where answer correctness and selective answering can be assessed systematically; extending the evaluation to broader generation tasks is a natural direction for future work.

Future work could combine question-level probing with external factual verification, develop recovery-oriented reformulation and equivalence-checking mechanisms, and predict recoverability from the original question to reduce unnecessary probing.

Ethical Considerations

This work studies selective answering for factual QA under uncertainty. By distinguishing prompt-level instability from question-level answerability, the proposed framework aims to reduce unsupported factual responses while avoiding unnecessary refusals caused by unstable prompt formulations. This can improve the reliability and usability of factual QA systems, especially in settings where both hallucinated answers and excessive abstention can harm user trust.

In deployment, abstention should be communicated as insufficient evidence under the system’s decision procedure, rather than a definitive judgment about the underlying fact. In high-impact domains such as medicine, law, and finance, the method should be used with appropriate human oversight or external verification. Since question-level probing relies on reformulation and automatic equivalence verification, its behavior may vary across languages, domains, and user expression patterns, and should therefore be evaluated in the intended use context before deployment.

References

- Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. 2024. [Mitigating LLM hallucinations via conformal abstention](#). *CoRR*, abs/2405.01563.
- Amos Azaria and Tom Mitchell. 2023. [The internal state of an LLM knows when it’s lying](#). In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore. Association for Computational Linguistics.
- Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. 2024. Inside: LLMs’ internal states retain the power of hallucination detection. In *International Conference*

- on *Learning Representations*, volume 2024, pages 3056–3076.
- Qinyuan Cheng, Tianxiang Sun, Xiangyang Liu, Wenwei Zhang, Zhangyue Yin, Shimin Li, Linyang Li, Zhengfu He, Kai Chen, and Xipeng Qiu. 2024. Can ai assistants know what they don't know? In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.
- Kyle Cox, Jiawei Xu, Yikun Han, Rong Xu, Tianhao Li, Chi-Yang Hsu, Tianlong Chen, Walter Gerych, and Ying Ding. 2025. [Mapping from meaning: addressing the miscalibration of prompt-sensitive language models](#). In *Proceedings of the Thirty-Ninth AAAI Conference on Artificial Intelligence and Thirty-Seventh Conference on Innovative Applications of Artificial Intelligence and Fifteenth Symposium on Educational Advances in Artificial Intelligence*, AAAI'25/IAAI'25/EAAI'25. AAAI Press.
- Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhishava Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. [Measuring and improving consistency in pretrained language models](#). *Transactions of the Association for Computational Linguistics*, 9:1012–1031.
- Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. [Detecting hallucinations in large language models using semantic entropy](#). *Nat.*, 630(8017):625–630.
- Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024. [Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration](#). In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.
- Yuheng Huang, Jiayang Song, Zhijie Wang, Shengming Zhao, Huaming Chen, Felix Juefei-Xu, and Lei Ma. 2025. [look before you leap: an exploratory study of uncertainty analysis for large language models](#). *IEEE Trans. Softw. Eng.*, 51(2):413–429.
- Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. [Survey of hallucination in natural language generation](#). *ACM Comput. Surv.*, 55(12).
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension](#). In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.
- Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. [Language models \(mostly\) know what they know](#). *CoRR*, abs/2207.05221.
- Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel J. Bell. 2025. [Abstentionbench: Reasoning llms fail on unanswerable questions](#). In *Advances in Neural Information Processing Systems*, volume 38, Main Conference. Curran Associates, Inc.
- Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal. 2024. [Semantic entropy probes: Robust and cheap hallucination detection in llms](#). *CoRR*, abs/2406.15927.
- Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. [Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation](#). In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.
- Shanu Kumar, Saish Mendke, Karody Lubna Abdul Rahman, Santosh Kurasa, Parag Agrawal, and Sandipan Dandapat. 2025. [Enhancing zero-shot chain of thought prompting via uncertainty-guided strategy selection](#). In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 2003–2025, Abu Dhabi, UAE. Association for Computational Linguistics.
- Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. [Natural questions: A benchmark for question answering research](#). *Transactions of the Association for Computational Linguistics*, 7:452–466.
- Maor Juliet Lavi, Tova Milo, and Mor Geva. 2026. [Detecting \(un\)answerability in large language models with linear directions](#). In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 682–699, Rabat, Morocco. Association for Computational Linguistics.
- Moxin Li, Yong Zhao, Wenxuan Zhang, Shuaiyi Li, Wenya Xie, See-Kiong Ng, Tat-Seng Chua, and Yang Deng. 2025. [Knowledge boundary of large language models: A survey](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5131–5157, Vienna, Austria. Association for Computational Linguistics.
- Yangyi Li and Mengdi Huai. 2025. [Quantifying uncertainty in natural language explanations of large](#)

- language models for question answering. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pages 25005–25013, Suzhou, China. Association for Computational Linguistics.
- Mingyi Liu. 2026. The alignment tax: Response homogenization in aligned llms and its implications for uncertainty estimation. *arXiv preprint arXiv:2603.24124*.
- Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do LLMs know when to NOT answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9329–9345, Abu Dhabi, UAE. Association for Computational Linguistics.
- Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.
- Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. 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*, pages 9004–9017, Singapore. Association for Computational Linguistics.
- Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.
- Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. 2024. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, pages 8901–8929. Curran Associates, Inc.
- Edward Phillips, Fredrik K Gustafsson, Sean Wu, Anshul Thakur, and David A Clifton. 2026. Entropy alone is insufficient for safe selective prediction in llms. *arXiv preprint arXiv:2603.21172*.
- Xin Qiu and Risto Miikkulainen. 2024. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. In *Advances in Neural Information Processing Systems*, volume 37, pages 134507–134533. Curran Associates, Inc.
- Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. 2023. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 5433–5442, Singapore. Association for Computational Linguistics.
- Benedict Aaron Tjandra, Muhammed Razzak, Jannik Kossen, Kunal Handa, and Yarin Gal. 2024. Fine-tuning large language models to appropriately abstain with semantic entropy. *CoRR*, abs/2410.17234.
- Bin Wang, Hui Li, Liyang Zhang, Qijia Zhuang, Ao Yang, Dong Zhang, Xijun Luo, and Bing Lin. 2026a. Argus: A multi-agent sensitive information leakage detection framework based on hierarchical reference relationships. In *2026 International Conference on Software Engineering (ICSE 2026)*.
- Bin Wang, YiLu Zhong, MiDi Wan, WenJie Yu, YuanBing Ouyang, YeNan Huang, HuiYu Wu, and Hui Li. 2026b. Is your prompt poisoning code? defect induction rates and security mitigation strategies. *Empirical Softw. Engg.*, 31(6).
- Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. 2025. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556.
- Qiming Xie, Zengzhi Wang, Yi Feng, and Rui Xia. 2024. Ask again, then fail: Large language models’ vacillations in judgment. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10709–10745, Bangkok, Thailand. Association for Computational Linguistics.
- Xunjian Yin, Xu Zhang, Jie Ruan, and Xiaojun Wan. 2024. Benchmarking knowledge boundary for large language models: A different perspective on model evaluation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2270–2286, Bangkok, Thailand. Association for Computational Linguistics.
- Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. R-tuning: Instructing large language models to say ‘I don’t know’. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, Mexico City, Mexico. Association for Computational Linguistics.

A Artifact Licenses and Terms of Use

The public datasets, pretrained models, and baseline implementations used in this paper are used solely for research evaluation. The datasets include TriviaQA, Natural Questions, and PopQA; the evaluated models include the Llama, Qwen,

and Mistral families; and auxiliary components include an NLI model and an LLM-as-judge model. We follow the licenses and terms of use specified by the respective artifact providers, including the Meta Llama Community License for Llama models, the Apache 2.0 license applicable to the Qwen and Mistral open-source models, the relevant terms of use for Gemma-related artifacts, and the licenses or usage guidelines of the QA datasets. PopQA is released under the MIT License.

We do not redistribute the original datasets, model weights, or restricted artifacts. If code is released in the future, the repository will retain references to third-party artifacts, their license information, and instructions for obtaining them. For models or datasets that require login, application, or acceptance of additional terms, users must obtain them directly from the original providers.

B Paraphrase Generation Prompts

This section provides the prompt template used to generate candidate reformulations in Section 3.3. The goal of the paraphraser is not to change the queried fact, introduce answer-bearing information, or alter constraints, but to vary the surface expression while preserving the question-level semantic core. The template requires preserving the queried subject, factual attribute, and expected answer type, and uses five roles to cover different formulations that users may naturally adopt. The generated candidate reformulations are then passed through equivalence verification; only verified prompts enter $\mathcal{P}^+(p_0)$ and participate in subsequent sampling and aggregation.

Paraphrase generation prompt

Instruction. Rewrite the original question into a semantically equivalent question for the assigned role. Preserve the queried subject, factual attribute, and expected answer type. Do not add constraints, remove constraints, or ask for a different fact.

Roles.

Layperson: A casual user asking out of mild curiosity; use a short and informal wording.

Student: A learner who does not know the topic well but wants to understand it; use a clear and standard wording.

Expert: A user with relevant domain knowledge; use appropriate field-specific terminology.

Researcher: A user investigating the topic further; add necessary historical, temporal, or contextual details only when they preserve the same fact.

Wildcard: Use a suitable formulation outside the above four roles to avoid homogeneous paraphrases.

Original prompt: $\{p_0\}$

Output: One rewritten prompt for the assigned role.

B.1 Illustrative Reformulation Cases

Admissible; TriviaQA validation

Original: Who was the man behind The Chipmunks?

Candidate: Which person was responsible for creating the Chipmunks?

Why: Same subject, fact, and person answer type; an idiomatic phrase is replaced by direct wording.

Admissible; PopQA test

Original: What is George Rankin's occupation?

Candidate: Which occupation is associated with George Rankin?

Why: Same subject, occupation attribute, and answer type; only the sentence form changes.

Rejected; Natural Questions validation

Original: who got the first nobel prize in physics

Candidate: Who received the first Nobel Prize in Chemistry?

Why: The queried attribute changes from Physics to Chemistry.

Rejected; PopQA test

Original: What is George Rankin's occupation?

Candidate: Was George Rankin a politician or a journalist?

Why: Candidate answers absent from the original introduce answer-bearing information.

Table 1: Illustrative admissible and rejected reformulations grounded in questions from the evaluation datasets. The assessments apply the semantic-equivalence and answer-bearing-information criteria; they are not additional per-example performance measurements.

Table 1 illustrates the reformulation constraints using questions from the three evaluation datasets. The two admissible cases regularize the wording without changing the subject, queried attribute, or answer type, and without supplying the answer. The rejected cases instead change the queried attribute or add answer-bearing options. Historical, temporal, or contextual information is admissible only when it preserves the same fact and does not carry the answer. These assessments illustrate the equivalence criteria rather than new per-example downstream results.

C Equivalence Verification Judge Prompt

The equivalence verification step in Section 3.3 is implemented with an LLM-as-judge. The prompt template used for this judge is shown below. The judge only determines whether two questions ask about the same underlying fact; it does not assess whether any candidate answer is correct. If a candidate reformulation changes the queried subject, factual attribute, constraint scope, or expected answer type, it is judged as non-equivalent and fil-

tered out. This step reduces the risk of semantic drift introduced by the paraphraser, ensuring that the stability evidence used in subsequent aggregation comes from different formulations of the same question rather than from a silently changed question.

Equivalence verification prompt

Instruction. You are an equivalence judge for question paraphrases.

Given two questions, decide whether they ask about the same underlying fact and would expect the same answer.

Output exactly one of the following labels: EQUIVALENT or NOT_EQUIVALENT.

Reject as NOT_EQUIVALENT if any of the following holds:

Subject shift: The two questions concern different subjects.

Attribute shift: The two questions ask for different attributes of the same subject.

Scope shift: One question is significantly more specific than the other, e.g., asks for a strict subset of the answer.

Answer-type shift: The expected answer types differ, e.g., one expects a person while the other expects a date.

Original question: {q_original}

Paraphrased question: {q_paraphrase}

Decision:

the original that supplies, strongly hints at, or constrains the answer, and factual injection, defined as a new factual claim or entity-level detail that changes the information state relevant to answering. Non-answer-bearing context that does not change the queried fact is not automatically labeled as injection.

The automatic judge treats EQUIVALENT as the positive class. Precision, recall, and F1 are computed over all 300 pairs against adjudicated human labels. Factual injection is reported with an exact numerator and denominator from the human audit; direct answer leakage is checked automatically over all reformulations by matching gold answers and aliases. Raw and post-filter equivalence and drift rates use the same frozen human labels and judge outputs. Table 2 reports the aggregate results.

D Human Audit and Judge Evaluation Protocol

We evaluate the equivalence judge and audit reformulation risks on 300 original–reformulation pairs sampled from the experimental candidate pool, with stratification across TriviaQA, Natural Questions, and PopQA; the five reformulation roles; and Claude Opus 4.7, Qwen2.5-14B-Instruct, and Qwen2.5-3B-Instruct. Sample IDs and presentation order were frozen before annotation, without replacement based on judge or downstream outcomes.

Annotators saw only the original question, candidate reformulation, and annotation guidelines. Paraphraser identity, gold answers and aliases, automatic judge predictions, target-model outputs, and downstream correctness were hidden. Two annotators independently labeled all 300 pairs without discussing individual examples; a third senior annotator adjudicated disagreements on equivalence or either safety label. The adjudicated labels serve as the human reference for judge evaluation.

A pair is labeled EQUIVALENT only when it preserves the same subject, queried attribute, answer type, and scope. Changes to any of these elements, added constraints, temporal or other scope shifts, a different queried fact, or irresolvable ambiguity lead to NOT_EQUIVALENT and a recorded drift reason. Annotators separately label answer-bearing information, defined as information absent from

Setting	Strict Equiv. ($\uparrow$)	Fact Inj. ($\downarrow$)	Drift ($\downarrow$)	Leakage ($\downarrow$)	AUGRC ($\downarrow$)
Raw Candidates	94.7%	0/300	5.3%	0.0%	—
After Equivalence Filter	99.0%	0/300	1.0%	0.0%	0.106
Clean Paraphrases Only	100.0%	0/300	0.0%	0.0%	0.108
KLE	—	—	—	—	0.157

Table 2: Reformulation audit. Strict equivalence, semantic drift, and factual injection are based on the stratified 300-pair human audit; direct answer leakage is checked by matching gold answers and aliases. Clean Paraphrases Only removes every reformulation flagged as questionable by human annotators. Unreported fields are shown as unavailable.

E Hyperparameter Sensitivity

Hyperparam	Value	AUROC ($\uparrow$)	AUGRC ($\downarrow$)
X	1	0.853	0.101
	2	0.859	0.097
	3	0.852	0.105
	4	0.840	0.116
	5	0.820	0.131
N	5	0.850	0.104
	10	0.860	0.097
	20	0.834	0.119
τ_ρ	0.40	0.854	0.100
	0.50	0.860	0.097
	0.60	0.855	0.099
τ_{dom}	0.05	0.857	0.098
	0.10	0.860	0.095
	0.20	0.856	0.100

Table 3: Hyperparameter sensitivity analysis. Each row corresponds to one configuration, with all other hyperparameters fixed to their default values. Gray-shaded rows indicate the default value of the corresponding hyperparameter. Values are averaged across 6 models, 3 datasets, and 3 seeds.

Table 3 reports the sensitivity analysis of four core hyperparameters of the framework (X , N , τ_ρ , and τ_{dom}) around their default values. Each row shows the mean performance of the corresponding configuration across 6 models, 3 datasets, and 3 seeds. Here, X denotes the number of consistent low-uncertainty evidence required for reformulation aggregation, N is the number of samples drawn for each prompt, τ_ρ is the dominant-mode strength threshold in the Variant B early-exit condition E_0 , and τ_{dom} controls whether the residual responses form a competing stable answer.

The results show that the default configuration is close to optimal in both AUROC and AUGRC, and most scanned settings lead to only small performance variations. When X is too large, the framework requires more equivalent reformulations to support the same answer, which reduces coverage and worsens AUGRC. When X is too small, the ev-

idence requirement becomes weaker and may admit less stable recoveries. Increasing the number of samples N from 5 to 10 yields more stable uncertainty estimates, but further increasing it to 20 provides limited gains while incurring higher inference cost. The scans over τ_ρ and τ_{dom} indicate that the early-exit decision in Variant B is stable around the default values, suggesting that the framework can achieve a reliable quality-cost trade-off without fine-grained hyperparameter tuning.

F Additional Experimental Results

This section summarizes the experimental materials discussed in Section 4 that are moved to the appendix due to space constraints. Table 4 reports the aggregated main results across 6 models and 3 random seeds, Table 5 reports the equal-budget baseline comparison, Table 6 reports paraphraser sensitivity, and Tables 7, 8, and 9 report the structural ablation, end-to-end cost, and stable-error results. The main text retains the core trend analysis and Risk-Coverage visualizations, while the tabular results are collected here to facilitate inspection of the exact numbers.

Tables 10, 11, and 12 further break down the main results by dataset, showing AUROC and AUGRC for 8 methods on 6 LLMs. Each number is the mean over 3 seeds. These results show that both variants of our framework maintain stable advantages across model families and model scales, suggesting that the gains from question-level probing are not driven by a single model or dataset. Variant A typically achieves the best AUROC and AUGRC, while Variant B is slightly lower but still clearly outperforms prompt-level baselines, consistent with the quality-cost trade-off discussed in the main text.

Figure 5 presents per-model Risk-Coverage curves. Although the absolute risk differs across models, the relative ordering of the curves is largely stable. Our framework maintains lower

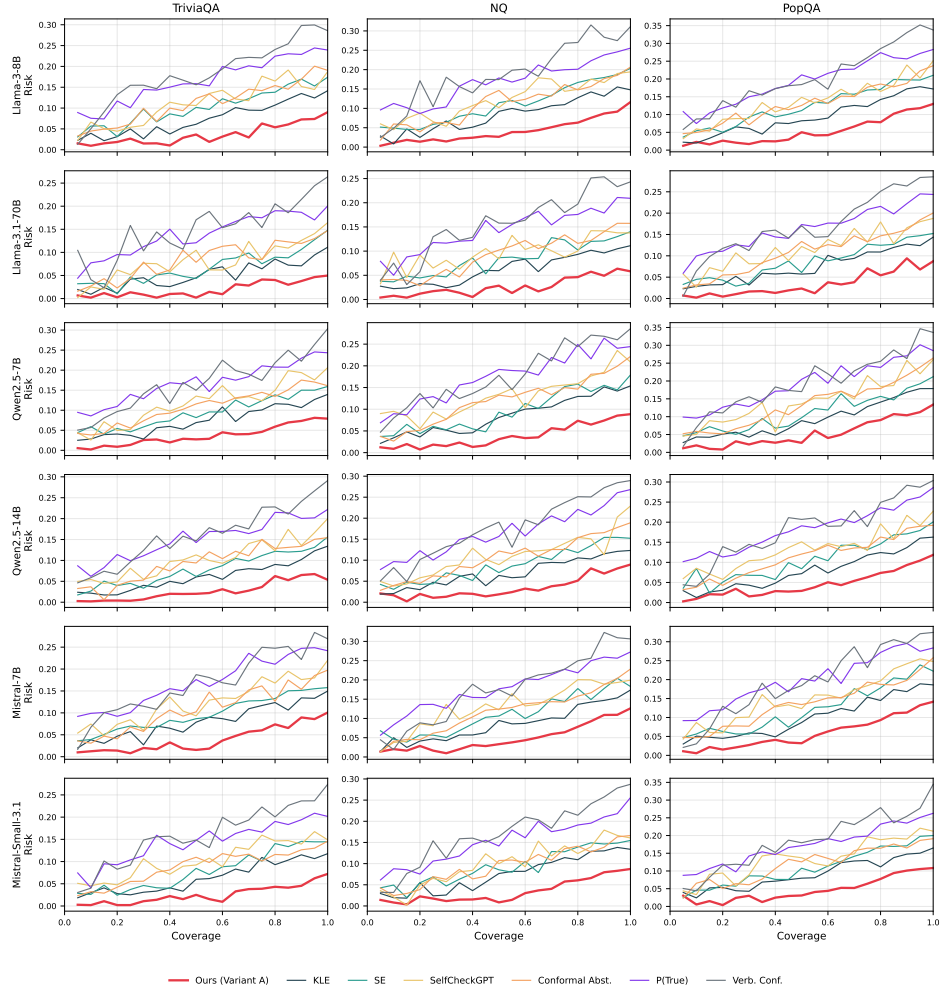


Figure 5: Per-model Risk-Coverage curves. Each row corresponds to one LLM, each column corresponds to one dataset, and each curve represents one method.

risk in both low-coverage and medium-to-high-coverage regions, indicating that its advantage comes not only from ranking high-confidence examples, but also from recovering questions that are unstable under the original prompt as coverage increases.

Method	TriviaQA	NQ	PopQA
<i>AUROC (↑)</i>			
Ours (Variant A)	0.886	0.861	0.832
KLE	0.825	0.809	0.780
SE	0.808	0.781	0.760
SelfCheckGPT	0.765	0.749	0.715
Conformal Abst.	0.777	0.755	0.730
P(True)	0.700	0.680	0.655
Verb. Conf.	0.660	0.634	0.607
<i>AUGRC (↓)</i>			
Ours (Variant A)	0.077	0.093	0.119
KLE	0.129	0.149	0.176
SE	0.150	0.169	0.198
SelfCheckGPT	0.178	0.198	0.224
Conformal Abst.	0.169	0.189	0.216
P(True)	0.236	0.257	0.282
Verb. Conf.	0.266	0.281	0.317

Table 4: Main results. Each cell reports the mean across 6 models and 3 seeds; the standard deviation across configurations is below 0.03. Bold indicates the best value in each dataset column, with higher AUROC and lower AUGRC being better.

Method	AUROC (↑)	AUGRC (↓)	Calls (↓)	Correct Rec. (↑)	Wrong Rec. (↓)
Budget-Matched Self-Consistency	0.811	0.145	29.6	18.2%	6.5%
Cox et al. (2025)	0.839	0.116	29.6	28.7%	5.5%
Majority over Rewrites	0.833	0.119	29.6	27.6%	5.8%
Ours, Variant B	0.850	0.106	29.6	31.8%	4.6%

Table 5: Equal-budget test-time comparison, averaged over 6 models, 3 datasets, and 3 random seeds. All methods use the same mean call budget, and the reformulation-based methods use the same candidate set. Correct and wrong recovery are computed over initially abstained cases.

Paraphraser	AUROC (↑)	AUGRC (↓)	Correct Rec. (↑)	Wrong Rec. (↓)	Rel. Cost (↓)
Claude Opus 4.7	0.850	0.106	31.8%	4.6%	1.00×
Qwen2.5-14B-Instruct	0.846	0.110	30.9%	4.7%	0.28×
Qwen2.5-3B-Instruct	0.848	0.108	31.4%	4.6%	0.11×
KLE	0.796	0.157	—	—	—

Table 6: Paraphraser sensitivity, averaged over 6 models, 3 datasets, and 3 random seeds. Relative cost is end-to-end cost normalized by the Claude Opus 4.7 configuration. Correct and wrong recovery are computed over cases initially abstained by Stage 1. Unreported KLE fields are shown as unavailable.

Configuration	AUROC (↑)	AUGRC (↓)	Mean Calls
Stage 1 only	0.766	0.205	10.0
Variant A	0.859	0.095	44.3
Variant B	0.850	0.106	29.6
Variant B [−]	0.829	0.125	45.1

Table 7: Structural ablation. Each row corresponds to one configuration, and values are averaged across 6 models, 3 datasets, and 3 seeds. Mean Calls is the average number of LLM calls per query. Variant B[−] denotes Variant B with equivalence verification disabled.

Method	Calls, Mean / P95	Mean Tokens	Latency, P50 / P95	Full-Probe Rate	AUGRC
Stage 1 only	10.0 / 10	2.3k	3.0 / 3.8 s	0.0%	0.205
Variant A	44.3 / 61	10.2k	11.8 / 19.4 s	68.6%	0.095
Variant B	29.6 / 61	6.6k	5.9 / 18.7 s	39.2%	0.106

Table 8: End-to-end cost, averaged across 6 models, 3 datasets, and 3 seeds. Calls include target-model generation, reformulation generation, and equivalence verification. Mean Tokens aggregates input and output tokens across all stages. Latency covers the complete pipeline, with stages executed sequentially and operations within each stage parallelized. Full-Probe Rate is the fraction of questions that proceed through complete reformulation, verification, and resampling after original-prompt evaluation.

Method	Stable-Wrong Rate	Correct-to-Wrong Gain
Stage 1 only	8.4%	—
Variant A	8.5%	7.4×
Variant B	8.3%	6.9×

Table 9: Stable-error analysis, averaged across 6 models, 3 datasets, and 3 seeds. Stable-Wrong Rate is the fraction of finally accepted answers that have low KLE but are incorrect. Correct-to-Wrong Gain is the number of newly correct answers per newly wrong answer relative to Stage 1 only.

Method	Llama-8B	Llama-70B	Qwen-7B	Qwen-14B	Mistral-7B	Mistral-24B
<i>AUROC</i> ($\uparrow$)						
Ours (Variant A)	0.873	0.916	0.872	0.904	0.853	0.897
Ours (Variant B)	0.854	0.906	0.854	0.881	0.847	0.880
KLE	0.812	0.858	0.813	0.834	0.800	0.836
SE	0.787	0.836	0.792	0.818	0.785	0.829
SelfCheckGPT	0.753	0.800	0.746	0.780	0.737	0.775
Conformal Abst.	0.756	0.811	0.758	0.791	0.753	0.793
P(True)	0.677	0.735	0.677	0.715	0.681	0.713
Verb. Conf.	0.639	0.686	0.646	0.673	0.635	0.680
<i>AUGRC</i> ($\downarrow$)						
Ours (Variant A)	0.093	0.046	0.093	0.065	0.099	0.063
Ours (Variant B)	0.105	0.054	0.096	0.068	0.104	0.071
KLE	0.144	0.092	0.149	0.118	0.157	0.117
SE	0.169	0.118	0.163	0.134	0.180	0.136
SelfCheckGPT	0.189	0.144	0.196	0.163	0.202	0.174
Conformal Abst.	0.189	0.144	0.180	0.153	0.198	0.152
P(True)	0.255	0.204	0.256	0.222	0.254	0.222
Verb. Conf.	0.284	0.237	0.286	0.250	0.291	0.249

Table 10: AUROC and AUGRC of eight methods on six LLMs for TriviaQA. Each cell reports the mean over three seeds; bold indicates the best value in each LLM column.

Method	Llama-8B	Llama-70B	Qwen-7B	Qwen-14B	Mistral-7B	Mistral-24B
<i>AUROC</i> ($\uparrow$)						
Ours (Variant A)	0.842	0.897	0.845	0.877	0.840	0.865
Ours (Variant B)	0.836	0.884	0.845	0.865	0.826	0.864
KLE	0.789	0.846	0.796	0.822	0.774	0.827
SE	0.761	0.808	0.763	0.798	0.756	0.797
SelfCheckGPT	0.728	0.787	0.735	0.765	0.723	0.758
Conformal Abst.	0.741	0.790	0.733	0.760	0.739	0.768
P(True)	0.666	0.710	0.660	0.691	0.660	0.696
Verb. Conf.	0.618	0.665	0.617	0.650	0.602	0.649
<i>AUGRC</i> ($\downarrow$)						
Ours (Variant A)	0.111	0.057	0.108	0.077	0.123	0.085
Ours (Variant B)	0.113	0.072	0.124	0.086	0.128	0.087
KLE	0.166	0.112	0.165	0.133	0.180	0.138
SE	0.189	0.132	0.185	0.154	0.199	0.152
SelfCheckGPT	0.219	0.168	0.210	0.190	0.223	0.180
Conformal Abst.	0.208	0.162	0.200	0.161	0.216	0.187
P(True)	0.278	0.218	0.282	0.249	0.277	0.239
Verb. Conf.	0.300	0.251	0.296	0.265	0.311	0.261

Table 11: AUROC and AUGRC of eight methods on six LLMs for Natural Questions. Each cell reports the mean over three seeds; bold indicates the best value in each LLM column.

Method	Llama-8B	Llama-70B	Qwen-7B	Qwen-14B	Mistral-7B	Mistral-24B
<i>AUROC</i> ($\uparrow$)						
Ours (Variant A)	0.819	0.854	0.812	0.847	0.811	0.847
Ours (Variant B)	0.811	0.860	0.802	0.823	0.802	0.837
KLE	0.768	0.806	0.771	0.796	0.753	0.787
SE	0.742	0.790	0.743	0.768	0.740	0.775
SelfCheckGPT	0.703	0.742	0.697	0.723	0.698	0.727
Conformal Abst.	0.712	0.765	0.709	0.738	0.708	0.747
P(True)	0.639	0.689	0.638	0.663	0.631	0.673
Verb. Conf.	0.588	0.643	0.590	0.623	0.576	0.625
<i>AUGRC</i> ($\downarrow$)						
Ours (Variant A)	0.133	0.080	0.140	0.108	0.139	0.112
Ours (Variant B)	0.144	0.099	0.147	0.118	0.150	0.124
KLE	0.194	0.141	0.194	0.164	0.198	0.167
SE	0.215	0.168	0.210	0.189	0.218	0.186
SelfCheckGPT	0.244	0.184	0.238	0.221	0.246	0.210
Conformal Abst.	0.231	0.186	0.232	0.201	0.239	0.204
P(True)	0.304	0.242	0.292	0.277	0.307	0.272
Verb. Conf.	0.336	0.286	0.333	0.305	0.339	0.306

Table 12: AUROC and AUGRC of eight methods on six LLMs for PopQA. Each cell reports the mean over three seeds; bold indicates the best value in each LLM column.