NOVA: NOise-aware Verbal Confidence CAlibration for Robust Large Language Models in RAG Systems

Jiayu Liu[*1] **Rui Wang**[*1] **Qing Zong**[1] **Yumeng Wang**[2] **Cheng Qian**[2]
Qingcheng Zeng[3] **Tianshi Zheng**[1] **Haochen Shi**[1] **Dadi Guo**[1]
Baixuan Xu[1] **Chunyang Li**[1] **Yangqiu Song**[1]
[1]HKUST [2]UIUC [3]Northwestern University
jliufv@connect.ust.hk yqsong@cse.ust.hk

Abstract

Accurately assessing model confidence is essential for deploying large language models (LLMs) in mission-critical factual domains. While retrieval-augmented generation (RAG) is widely adopted to improve grounding, confidence calibration in RAG settings remains poorly understood. We conduct a systematic study across four benchmarks, revealing that LLMs exhibit poor calibration performance especially when noisy contexts are retrieved. Specifically, contradictory or irrelevant evidence tends to exacerbate the model's overconfidence issue. To address this, we propose *NOVA Rules* (**NO**ise-Aware **V**erbal Confidence C**A**libration Rules) to provide a principled foundation for resolving overconfidence under noise. We further design **NOVA**, a noise-aware calibration framework that synthesizes supervision from ∼2K HotpotQA examples guided by these rules. By performing supervised fine-tuning (SFT) with this data, NOVA equips models with intrinsic noise awareness without relying on stronger teacher models. Empirical results show that NOVA yields substantial gains, improving ECE scores by 10.9% in-domain and 8.0% out-of-domain. By bridging the gap between retrieval noise and verbal calibration, NOVA paves the way for both accurate and epistemically reliable LLMs. [2]

1 Introduction

Large Language Models (LLMs) have exhibited remarkable performance in mathematical reasoning (DeepSeek-AI et al., 2025; Zheng et al., 2025; Yu et al., 2025), code generation (Chen et al., 2025b; Liu et al., 2024c) and complex problem solving (Wang et al., 2026b; Yuan et al., 2025). Despite the impressive performance, their reliability in high-stakes and fact-intensive question answering remains limited due to persistent hallucinations (Zhang et al., 2024; Bang et al., 2025). Thus,

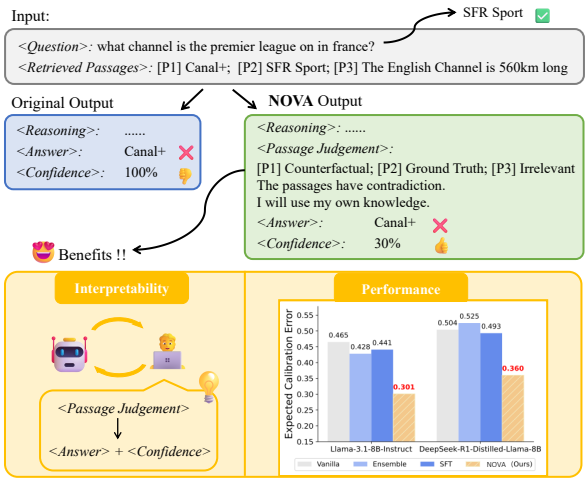


Figure 1: An illustrative example of model responses before and after **NOVA**. By explicitly training the model to assess passage and group level utility prior to answering, **NOVA** enables more reliable confidence expression under noisy retrieval. The performance plots report results on NQ for *Llama-3.1-8B-Instruct* and *DeepSeek-R1-Distill-Llama-8B*, where SFT corresponds to the Label-only SFT setting in Table 2, and illustrate how **NOVA** promotes more transparent and grounded human–computer interaction in real-world scenarios.

retrieval-augmented generation (RAG) is widely adopted (Dong et al., 2025; Sun et al., 2025b) to alleviate this issue by incorporating external knowledge. However, LLM generation with RAG is not always accurate or complete, and LLMs may still produce hallucinations even with access to additional information (Sorodoc et al., 2025; Jin et al., 2025). This makes uncertainty estimation crucial (More in Appendix C.1) and essential for assessing the reliability of LLM outputs in RAG scenarios.

Among existing uncertainty estimation approaches, verbal confidence calibration has gained increasing attention as a mechanism for assessing the reliability of LLM outputs (Wei et al., 2025a; Ou et al., 2025), while also enhancing model trustworthiness (Xu et al., 2024b; Zong et al., 2025a) and facilitating efficient and trustworthy human–AI interaction (Li et al., 2025a,c). However, how LLM

[1]* Equal contribution.
[2]Code: https://github.com/HKUST-KnowComp/NOVA.

Findings of the Association for Computational Linguistics: EMNLP 2026, pages 14793–14835
October 24-29, 2026 ©2026 Association for Computational Linguistics

verbal confidence calibrates under RAG has not been fully explored. Moreover, expressing confidence in RAG settings requires reasoning over retrieved evidence, making the calibration problem different from that in non-RAG scenarios (Soudani et al., 2025a). Existing confidence estimation methods remain insufficient for real-world RAG systems. While recent works enable verbal confidence via reinforcement learning (Damani et al., 2025b; Xu et al., 2024b) or specialized loss (Li et al., 2025c; Stengel-Eskin et al., 2024), they target internal reasoning in closed-book settings, overlooking the external uncertainty introduced by retrieval contents. Prior attempts to quantify uncertainty in RAG rely on white-box signals (Zhou et al., 2025a; Ozaki et al., 2025a) or sampling consistency (Wang et al., 2025c), which suffer from inaccessibility in proprietary models or computational costs.

To address this gap, we conduct a systematic study of verbal confidence calibration under RAG. We first empirically demonstrate that verbal confidence exhibits poor calibration in RAG settings, with the average Expected Calibration Error (ECE) exceeding 0.4 across all four datasets. Our analysis further identifies noisy retrieval as a key factor undermining calibration, particularly with irrelevant and counterfactual passages. To this end, we introduce *NOVA Rules* to regularize model behavior under noise. Guided by these rules, we propose **NOVA**, a framework that bootstraps the model using approximately 2K data points without external teacher models. Extensive experiments demonstrate that **NOVA** yields substantial gains, improving calibration by 10.9% in-domain and 8.0% out-of-distribution (OOD), highlighting its effectiveness and generalizability. Moreover, **NOVA** improves models' ability to discern passage utility and enhances interpretability by grounding confidence estimates in structured intermediate judgments within explicit reasoning traces. By bridging the gap between verbal calibration and retrieval noise, our approach constitutes a principled step toward robust and interpretable RAG systems.

2 Related Work

Confidence estimation in LLMs. Confidence estimation approaches for LLMs are generally categorized into white-box and black-box methods (Fadeeva et al., 2023; Geng et al., 2024a). White-box methods utilize internal signals like perplexity (Fomicheva et al., 2020; Zong et al., 2025b) or logits (Duan et al., 2024; Kuhn et al., 2023), yet their dependence on internal states restricts their utility for proprietary models. Consequently, black-box research has evolved from sampling consistency (Lin et al., 2024a) or verbalized confidence scores (Lin et al., 2022; Yang et al., 2024) to optimizing verbalized confidence via reinforcement learning (Damani et al., 2025b; Xu et al., 2024b), specialized loss functions (Li et al., 2025c), or teacher-distilled critiques (Zong et al., 2025a). However, unlike closed-book or computationally costly methods, our approach addresses RAG-specific external uncertainty from retrieval noise through efficient self-bootstrapping, without requiring reinforcement learning or teacher models.

Retrieval Noise and Robustness. Recent work highlights the harm of retrieval noise (irrelevant, counterfactual, repetitive, etc.) on answer accuracy (Wu et al., 2025a; Cuconasu et al., 2024; Guo et al., 2025). Although frameworks such as Self-RAG (Asai et al., 2024), RAAT (Fang et al., 2024a), and HANRAG (Sun et al., 2025a) enhance robustness, they primarily optimize answer accuracy rather than verbal confidence calibration. In particular, RAAT uses adaptive adversarial training to improve robustness under retrieval noise, whereas SGIC (Chen et al., 2025a) uses self-estimated uncertainty to guide iterative answer revision. In contrast, **NOVA** directly targets verbal confidence calibration, using passage-utility judgments to filter self-generated, rule-aware reasoning trajectories followed by standard supervised fine-tuning. Our work therefore complements prior noise-robust RAG methods by modeling how retrieval noise affects confidence expression.

Uncertainty Quantification for RAG Reliability is paramount for RAG systems. Prior works primarily utilize white-box metrics in domain-specific settings (Zhou et al., 2025a; Ozaki et al., 2025a) or consistency sampling (Wang et al., 2025c; Soudani et al., 2025c). While Soudani et al. (2025a) recently formalized the theoretical limitations of these standard estimators, existing approaches remain predominantly dependent on costly test-time sampling or external calibration functions, rather than optimizing the model's intrinsic uncertainty. Prior studies also focus on refusal rates (Zhou et al., 2025b), leaving verbal confidence largely uncalibrated. We address this gap with a training-based framework that directly calibrates verbal confidence. More related works are in Appendix D

3 Task Formalization

Task Setup. Let $q \in \mathcal{Q}$ denote a query, and $\mathcal{D} = \{d_1, \ldots, d_{|\mathcal{D}|}\}$ be a corpus. A retriever $\mathcal{R}$ maps a query and the corpus to a set of k passages:

$$\mathcal{P} = \mathcal{R}(q, \mathcal{D}) = \{p_1, \ldots, p_k\}. \quad (1)$$

A retrieval-augmented language model f_θ takes the query and the retrieved passages as input, and outputs both a predicted answer $\hat{a} \in \mathcal{A}$ and a scalar verbal confidence score $\hat{c} \in [0\%, 100\%]$:

$$(\hat{a}, \hat{c}) = f_\theta(q, \mathcal{R}(q, \mathcal{D}))\,. \quad (2)$$

Answer Correctness and Calibration Objective. Given a set of gold answers $\mathcal{A}^*$, we derive the binary correctness label using the normalized in-accuracy criterion, following convention in Moskvoretskii et al. (2025). Let $\mathrm{norm}(\cdot)$ denote the answer normalization function. A prediction is marked correct if its normalized response contains at least one normalized gold answer as a substring:

$$y = \mathbb{I}[\exists a^* \in \mathcal{A}^* \text{ s.t. } \mathrm{norm}(a^*) \subseteq \mathrm{norm}(\hat{a})]\,, \quad (3)$$

where $y = 1$ indicates a correct answer and $y = 0$ indicates an incorrect answer. The objective of verbal confidence calibration in RAG is to align model-reported confidence $\hat{c}$ with empirical correctness y. While this objective is identical to that in the non-RAG setting, it must account for uncertainty introduced by the retrieval process.

Passage Categorization. Following the categorization scheme introduced in prior work (Cuconasu et al., 2024; Wu et al., 2025a), given the ground-truth answer a^*, we categorize the retrieved passages $\mathcal{P}$ into four disjoint subsets based on their semantic content: (1) *Gold passages* $\mathcal{P}_{\text{gold}}$, which support the correct answer a^*; (2) *Counterfactual passages* $\mathcal{P}_{\text{cf}}$, which support an incorrect answer $a' \neq a^*$; (3) *Relevant passages* $\mathcal{P}_{\text{rel}}$, which are topically related to q (e.g., sharing entities or relations) but do not provide sufficient information to answer the query; and (4) *Irrelevant passages* $\mathcal{P}_{\text{irr}}$, which contain no semantic overlap with the query.

Calibration Metrics. Calibration quality is evaluated using *Expected Calibration Error* (ECE) and the *Area Under the Receiver Operating Characteristic Curve* (AUROC). Expected Calibration Error (ECE) measures the average discrepancy between predicted confidence and empirical accuracy:

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{N} \left|\mathrm{acc}(B_m) - \mathrm{conf}(B_m)\right|\,, \quad (4)$$

where N denotes the total number of samples and M the number of confidence bins. Following standard practice (Guo et al., 2017a), we set $M = 10$ in all experiments. Here, B_m denotes the set of samples assigned to the m-th bin, while $\mathrm{acc}(B_m)$ and $\mathrm{conf}(B_m)$ represent the average accuracy and confidence within that bin, respectively. AUROC evaluates the discriminative ability of confidence scores to distinguish correct from incorrect predictions:

$$\mathrm{AUROC} = \Pr\big(\hat{c}^+ > \hat{c}^-\big)\,, \quad (5)$$

where $\hat{c}^+$ and $\hat{c}^-$ denote confidence scores associated with correct and incorrect predictions, respectively. Higher AUROC values indicate stronger discriminative power.

4 Experiment

Models. We use four widely-used open-sourced LLMs to conduct our experiment (detailed list provided in Appendix A.1). Proprietary models are excluded as they do not support the access to internal states or fine-tuning required for further alignment.

Datasets. We adopt a randomly selected subset of Natural Questions (NQ) (Kwiatkowski et al., 2019), Bamboogle (Press et al., 2023), StrategyQA (Geva et al., 2021) and HotpotQA (Yang et al., 2018), and as our primary evaluation benchmark. The fine-grained data statistics (number of data instances, etc.) is provided in Appendix A.3.

Prompts. The retrieval-augmented language model f_θ is instantiated using Chain-of-Thought (CoT) prompting (Wei et al., 2022) for all experiments unless otherwise specified. Due to the instability of verbal confidence (Obadinma and Zhu, 2025; Liu et al., 2025b), we add extra experiments (see Appendix B.3) using prompts from Xiong et al. (2024) to verify the robustness of our conclusions. The exact prompts are discussed in Appendix A.4.1, and additional results are in Appendix B.3.

RAG Settings. We use *wikimedia/wikipedia* (Wikimedia, 2023) from the HuggingFace dataset (Wolf et al., 2020) as the RAG corpus $\mathcal{D}$. For the retriever $\mathcal{R}$, we follow Soudani et al. (2025b) and use BM25 (Robertson et al., 2004) and Contriever (Izacard et al., 2021) to retrieve top-k passages ($k = 3$ in Table 1). To mitigate the position bias of retrieved passages $\mathcal{P}$ noted in Liu et al. (2024b); Ozaki et al. (2025a), we randomize their order in the input context fed to f_θ. We further provide robustness checks on

Method	StrategyQA		HotpotQA		NQ		Bamboogle		Average	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
Llama-3.1-8B-Instruct										
BM25 (CoT)	0.205	0.485	0.496	0.552	0.369	0.688	0.566	0.557	0.409	0.571
Contriever (CoT)	0.167	0.550	0.585	0.476	0.347	0.649	0.592	0.535	0.423	0.552
Qwen2.5-7B-Instruct										
BM25 (CoT)	0.190	0.620	0.439	0.683	0.473	0.747	0.650	0.670	0.438	0.680
Contriever (CoT)	0.197	0.608	0.532	0.622	0.566	0.661	0.663	0.647	0.489	0.635
DeepSeek-R1-Distill-Llama-8B										
BM25 (CoT)	0.219	0.542	0.454	0.669	0.487	0.716	0.630	0.700	0.448	0.657
Contriever (CoT)	0.229	0.572	0.508	0.653	0.488	0.754	0.591	0.710	0.454	0.672
DeepSeek-R1-Distill-Qwen-7B										
BM25 (CoT)	0.266	0.531	0.587	0.580	0.602	0.684	0.753	0.623	0.552	0.605
Contriever (CoT)	0.270	0.561	0.678	0.547	0.623	0.694	0.760	0.650	0.583	0.613

Table 1: Real-world RAG results across four datasets demonstrate consistently poor calibration. Notably, average ECE exceeds 0.4, highlighting severe misalignment between verbal confidence and empirical correctness.

passage positioning in Appendix B.2. Detailed experimental settings are in Appendix A, with additional RAG details in Appendix A.5.

5 Analysis

Models fail in calibrating in RAG scenarios. As shown in Table 1, we observe that models exhibit severely degraded verbal calibration performance in real-world RAG settings. Across different datasets and retrievers, all four models consistently demonstrate poor alignment between verbal confidence $\hat{c}$ and empirical correctness y, as evidenced by average ECE values exceeding 0.4. In particular, *DeepSeek-R1-Distill-Qwen-7B* reaches an average ECE of 0.542, indicating a substantial discrepancy. According to Xiong et al. (2024), an ECE above 0.25 is already considered unsatisfactory, underscoring the severity of the calibration failures observed in our setup.

5.1 Noise Generation

To diagnose the model's failures, we analyze the problem from the perspective of retrieval passage noise in RAG (Wu et al., 2025a; Cuconasu et al., 2024). To better reflect real-world RAG behavior, we categorize *relevant passages* into *entity-relevant*, *relation-relevant*, and *theme-relevant* types, and randomly sample one of these categories when generating a *relevant passage* (The definition of each type is provided in Appendix A.8). We then use few-shot prompting (Brown et al., 2020) with *Gemini-2.5-Pro* (Google DeepMind, 2025) to generate the three types of noisy passages ($\mathcal{P}_{\text{cf}}$, $\mathcal{P}_{\text{rel}}$, $\mathcal{P}_{\text{irr}}$) for each query across all four datasets, providing the model with explicit definitions and illustrative examples of all noise types, and conditioning it on the target noise category during generation.

5.2 Controlled Analysis Setup

To simulate various RAG scenarios, we manipulate the retrieved set $\mathcal{P}$ by varying its composition. Let p^* denote the gold passage (where $p^* \in \mathcal{P}_{\text{gold}}$). For noise injection, let $\mathcal{P}_{\text{noise}}$ be a subset of passages drawn uniformly from a single noise category $C \in \{\mathcal{P}_{\text{cf}}, \mathcal{P}_{\text{rel}}, \mathcal{P}_{\text{irr}}\}$. We define two specific input configurations:

(1) Gold Only: The model receives only the correct context. We set $\mathcal{P} = \{p^*\}$.

(2) Gold + Noise: The model receives the gold passage alongside two noise passages. We set $\mathcal{P} = \{p^*\} \cup \mathcal{P}_{\text{noise}}$, subject to $|\mathcal{P}_{\text{noise}}| = 2$. As shown in Figure 2, the results indicate that retrieval noise is the primary factor driving models' calibration failures. Specifically:

***Counterfactual noise* greatly degrades models' calibration performance.** From Figure 2, we observe that when gold passages are present, introducing counterfactual passages leads to most significant degradation in calibration performance compared to the *Gold-only* baseline, characterized by increased ECE and decreased AUROC. Specifically, relative to the *Gold-only* setting, *Llama-3.1-8B-Instruct* and *DeepSeek-R1-Distill-Llama-8B* exhibit an average ECE increase of 31.6% and 35.1%, and an average AUROC decrease of 9.1% and 16.1%, respectively on NQ and Bamboogle. In contrast, as shown by the average confidence results in Figure 2, the models display similar confidence levels $\hat{c}$ under the *Gold-only* and *Gold+counterfactual* noise settings. This indicates that when exposed to contradictory evidence, models tend to commit to one answer while maintaining a confidence level comparable to that in the noise-free setting. Consequently, under contradictory retrieval signals, the confidence estimates $\hat{c}$ become decoupled from the answer correctness y, rendering verbal confidence

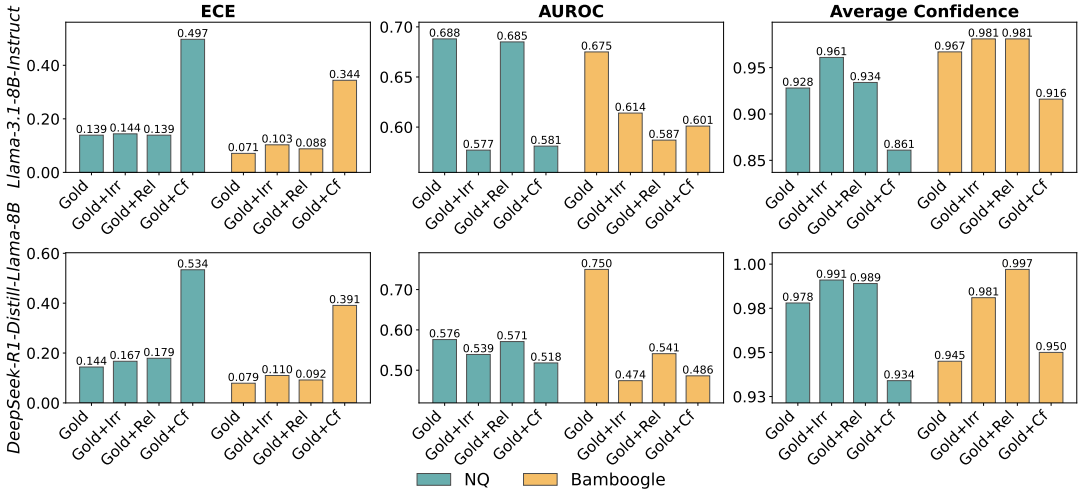


Figure 2: Calibration performance of *Llama-3.1-8B-Instruct* and *DeepSeek-R1-Distill-Llama-8B* on NQ and Bamboogle under controlled noise settings. The plots display ECE, AUROC, and Average Confidence across four retrieval settings: *Gold-only*, *Gold+Irrelevant* (Irr), *Gold+Relevant* (Rel), and *Gold+Counterfactual* (Cf). Results show that introducing noise, particularly counterfactual passages, substantially degrades calibration performance.

an unreliable indicator of model uncertainty.

***Relevant noise* also harms calibration performance notably.** As shown in Figure 2, the presence of *relevant noise* also significantly degrades the calibration performance of the models compared to the *Gold-only* baseline. Relative to the *Gold-only* setting, the introduction of *relevant noise* consistently results in higher ECE and lower AUROC across both models on the NQ and Bamboogle datasets. Notably, the average AUROC drops by 4.6% for *Llama-3.1-8B-Instruct* and 10.7% for *DeepSeek-R1-Distill-Llama-8B* compared to the baseline. We further observe a systematic increase in the average confidence $\hat{c}$ when *relevant* or *irrelevant* noise is introduced alongside gold passages. This suggests that exposure to additional, unhelpful information tends to inflate the models' confidence, thereby impairing calibration even when the gold passage is present.

Even *irrelevant noise* causes obvious degradation in calibration. Surprisingly, *irrelevant noise* mirrors the trend of *relevant noise*. While ECE increases moderately versus the *Gold-only* baseline, AUROC drops substantially (8.6% for *Llama-3.1-8B-Instruct* and 15.7% for *DeepSeek-R1-Distill-Llama-8B*), even exceeding the decline from *relevant noise*. Consistent with *relevant noise*, a systematic rise in average confidence $\hat{c}$ relative to *Gold-only* is again observed. This suggests models become overconfident due to information expansion, deriving false certainty even from completely irrelevant passages. We provide more analysis statistics in Table 15.

6 Method

6.1 From Observation to Rules

From the above analysis, we observe two problematic behaviors of current models: (1) **Overconfidence under conflict**: When presented with counterfactual passages (i.e., $\mathcal{P} \cap \mathcal{P}_{\text{cf}} \neq \emptyset$), models still assign relatively high confidence to their answers; (2) **Overconfidence under noise**: When relevant or irrelevant noise is introduced alongside gold passages (i.e., $\mathcal{P} = \{p^*\} \cup \mathcal{P}_{\text{noise}}$), models exhibit a systematic increase in average confidence $\hat{c}$ compared to the gold-only setting ($\mathcal{P} = \{p^*\}$).

To address these issues, we argue that the expected behavior in RAG should follow a set of *NOise-Aware Verbal Confidence CAlibration Rules* (*NOVA Rules*). Formally, we posit that an ideal retrieval-augmented model should satisfy the following properties: **(1) Conflict Independence**: When counterfactual passages are retrieved ($\mathcal{P} \cap \mathcal{P}_{\text{cf}} \neq \emptyset$), since the factual correctness of external evidence cannot be reliably determined, the model should fall back to its internal parametric knowledge. Ideally: $(\hat{a}, \hat{c}) \approx f_\theta(q, \emptyset)$. **(2) Noise Invariance**: When irrelevant passages are retrieved ($\mathcal{P}_{\text{irr}} \cap \mathcal{P} \neq \emptyset$), the model should explicitly ignore them during reasoning. The prediction should be invariant to the addition of noise: $f_\theta(q, \mathcal{P}) \approx f_\theta(q, \mathcal{P} \setminus \mathcal{P}_{\text{irr}})$. **(3) Parametric Fallback**: When no helpful passage is retrieved (i.e., $\mathcal{P} \cap \mathcal{P}_{\text{gold}} = \emptyset$), the model should disregard the external context and answer solely based on its internal knowledge, mirroring the behavior defined in (1). We detail the rule rationale and method design

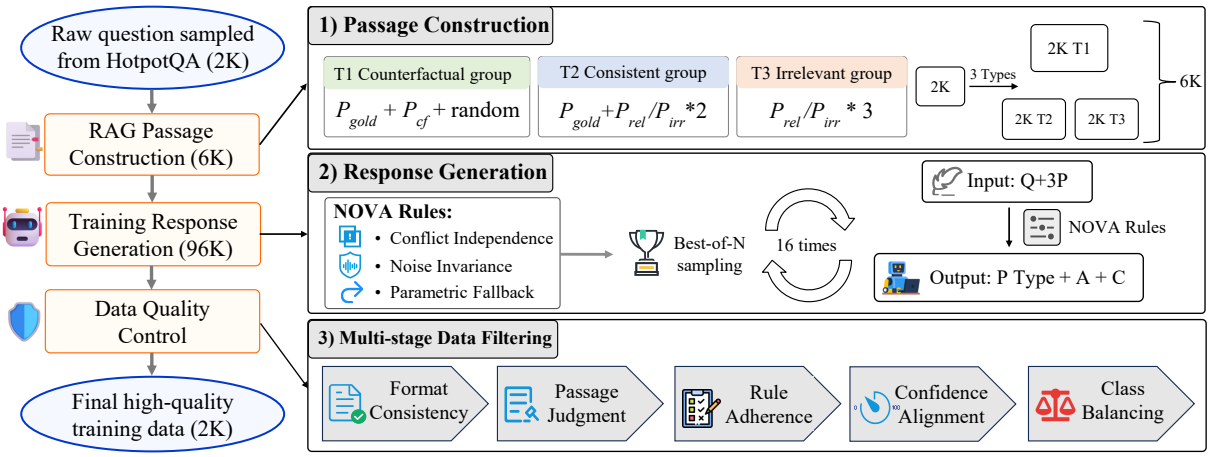


Figure 3: Overview of the **NOVA** data pipeline, consisting of three stages: RAG Passage Construction, Training Response Generation, and Multi-stage Data Filtering. In Training Response Generation, the model takes a query q and retrieved passages $\mathcal{P}$ with $k = 3$ as input (Input: Q+3P), generates passage-level and group-level judgments J_p, J_g (P Type), then predicts the answer $\hat{a}$ (A) and verbal confidence $\hat{c}$ (C). The pipeline ultimately yields 2K high-quality trajectories for fine-tuning.

choices in Appendices C.2 and C.4.

6.2 NOVA Framework

RAG Passage Construction. We assemble the raw noisy passages generated in Section §5.1 for the HotpotQA training set into three distinct RAG passage groups. Crucially, these configurations serve as the ground-truth labels for the *Passage Group Judgment* (J_g), enabling the model to explicitly discern the utility of the retrieved set. For each query, we organize the retrieved context $\mathcal{P}$ into specific configurations: **(1) Counterfactual**: Contains the gold passage alongside at least one contradictory $\mathcal{P}_{\text{cf}}$ passage to test conflict resolution. **(2) Consistent**: Contains the gold passage mixed with *relevant* or *irrelevant* noise ($\mathcal{P}_{\text{rel}}$ or $\mathcal{P}_{\text{irr}}$) to assess robustness amid noise. **(3) Irrelevant**: Contains only *relevant* and *irrelevant* passages without valid evidence to probe behavior under missing information. A final balanced dataset is created by randomly sampling from these configurations to ensure diverse coverage of noise types.

Training Response Generation. We then perform Best-of-N (BoN) sampling (Stiennon et al., 2022; Ouyang et al., 2022) on the initial noisy HotpotQA training set obtained above. To select the best samples that follows *NOVA Rules*, we prompt the model to produce process judgments at two levels along with the answer and the corresponding confidence: (i) passage-level judgments J_p, indicating whether each passage can directly answer the question, and (ii) group-level judgments J_g, indicating whether the passage group consistently suggests an answer. These judgments are used as intermediate labels for data filtering.

Data Quality Control. We apply a multi-stage filtering pipeline to ensure the training data aligns with *NOVA Rules*: (1) **Format Consistency**: Retains only samples where valid answers, confidence scores, and intermediate reasoning traces can be successfully parsed. (2) **Passage Judgment Accuracy**: Filters out instances with incorrect passage assessments (J_p and J_g), ensuring the model accurately discriminates passage utility as a prerequisite for subsequent rule application. (3) **Rule Adherence**: Verifies that the reasoning process explicitly invokes and considers the corresponding *NOVA Rules*. (4) **Confidence Alignment**: Selects the response trajectory that minimizes the instance-level Brier Score (Brier, 1950; Damani et al., 2025b), effectively aligning the verbalized confidence with empirical correctness (i.e., towards 100% for correct and 0% for incorrect predictions). (5) **Class Balancing**: Balances the distribution of retrieval scenarios (i.e., *counterfactual*, *consistent*, and *irrelevant*) by downsampling the dominant class to match the minority class size, ensuring a uniform data distribution for training. Figure 3 provides an overview of the **NOVA** data construction pipeline.

Supervised Fine-tuning (SFT). After multi-stage filtering, we retained approximately 2,000 high-quality QA pairs, which were used for supervised LoRA fine-tuning (SFT) with LlamaFactory (Zheng et al., 2024). More details and hyperparameters are provided in Appendix A.6.

Method	StrategyQA		HotpotQA		NQ		Bamboogle		Average	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
Llama-3.1-8B-Instruct										
Vanilla	0.396	0.602	0.460	0.605	0.465	0.577	0.324	0.636	0.411	0.605
CoT	0.354	0.555	0.444	0.645	0.423	0.611	0.288	0.552	0.377	0.591
Noise-aware	0.376	0.615	0.309	0.642	0.351	0.618	0.217	0.793	0.314	0.667
Ensemble	0.370	0.609	0.397	0.650	0.428	0.619	0.214	0.713	0.352	0.648
Label-only SFT	0.345	0.619	0.319	0.711	0.441	0.658	0.307	0.755	0.353	0.686
P(True)	0.410	0.429	0.570	0.436	0.463	0.428	0.472	0.445	0.479	0.434
NOVA	**0.285**	**0.624**	**0.280**	**0.778**	**0.301**	**0.724**	**0.199**	**0.877**	**0.266**	**0.751**
Qwen2.5-7B-Instruct										
Vanilla	0.398	0.689	0.391	0.712	0.438	0.710	0.236	0.809	0.366	0.730
CoT	0.363	0.709	0.353	0.693	0.414	0.703	0.208	0.798	0.335	0.726
Noise-aware	0.393	0.618	0.325	0.692	0.380	0.649	0.192	0.828	0.323	0.697
Ensemble	0.368	0.719	0.380	0.681	0.451	0.693	0.240	0.793	0.360	0.722
Label-only SFT	**0.297**	0.679	0.321	0.699	0.425	0.691	0.216	0.821	0.315	0.722
P(True)	0.373	0.549	0.481	0.494	0.427	0.518	0.455	0.696	0.434	0.564
NOVA	0.310	**0.726**	**0.312**	**0.735**	**0.322**	**0.754**	**0.113**	**0.856**	**0.264**	**0.768**
DeepSeek-R1-Distill-Llama-8B										
Vanilla	0.416	0.639	0.457	0.660	0.504	0.637	0.251	0.693	0.407	0.657
CoT	0.434	0.656	0.484	0.617	0.531	0.639	0.294	0.687	0.436	0.650
Noise-aware	0.343	0.621	0.443	0.633	0.425	0.584	0.281	0.622	0.373	0.615
Ensemble	0.399	**0.673**	0.465	0.650	0.525	0.592	0.240	0.678	0.407	0.648
Label-only SFT	0.405	0.554	0.517	0.577	0.493	0.588	0.346	0.692	0.440	0.603
P(True)	0.404	0.536	0.484	0.587	0.530	0.570	0.438	0.673	0.464	0.592
NOVA	**0.323**	0.651	**0.359**	**0.663**	**0.360**	**0.656**	**0.200**	**0.748**	**0.311**	**0.679**
DeepSeek-R1-Distill-Qwen-7B										
Vanilla	0.408	0.642	0.499	0.641	0.522	0.668	0.318	0.666	0.437	0.654
CoT	0.409	0.668	0.529	0.632	0.578	0.641	0.381	0.681	0.474	0.655
Noise-aware	0.321	0.530	0.422	0.523	0.505	0.535	0.304	0.697	0.388	0.571
Ensemble	0.415	0.659	0.515	0.614	0.561	0.616	0.356	0.601	0.462	0.623
Label-only SFT	0.347	0.595	0.591	0.538	0.629	0.585	0.502	0.574	0.517	0.573
P(True)	0.421	0.525	**0.340**	0.518	**0.386**	0.591	0.418	0.660	0.391	0.574
NOVA	**0.306**	**0.672**	0.391	**0.702**	0.409	**0.726**	**0.271**	**0.793**	**0.344**	**0.723**

Table 2: Calibration performance of various models on four datasets. Scores in **bold** indicate the best performance, while underlined scores denote the second-best. Results show that **NOVA** substantially improves calibration and consistently outperforms several baselines, without sacrificing accuracy, as evidenced in Appendix B.8.

6.3 Baselines

(1) **Prompting Methods:** We adopt vanilla, CoT (Wei et al., 2022; Xiong et al., 2024) and a specialized noise-aware prompting for comparison. The noise-aware prompting incorporates *NOVA Rules* in the prompt and ask the model to follow the rules (details in Figure 10).

(2) **Ensemble** (Li et al., 2025c): The LLM is queried three times to generate answers with associated confidence scores, where the confidences of the most frequent answer are then averaged to obtain the final confidence estimate.

(3) **Label only SFT**: This baseline directly utilizes the inputs, answers, and confidence labels from **NOVA** for SFT, excluding the intermediate reasoning steps. It aims to evaluate the specific impact of confidence supervision on **NOVA**. Statistics of the training data are shown in Appendix A.7.

(4) **P(True)** (Kadavath et al., 2022a): A self-evaluation method where a language model predicts the probability its own answer is correct, computed as the normalized probability of the "True" token over "True" and "False". Additional calibration baselines are discussed in Appendix C.3.

6.4 Results

NOVA exhibits consistent and significant calibration improvement over several baselines. As demonstrated in Table 2, **NOVA** consistently outperforms all baseline methods across four datasets and four model backbones. Specifically, relative to Vanilla and CoT prompting, **NOVA** yields an approximately 11% reduction in ECE across models, along with consistent AUROC gains. These improvements are statistically significant across all four model backbones for both ECE and AUROC ($p \leq 0.05$), as detailed in Appendix B.7. Furthermore, **NOVA** attains superior alignment (lower ECE) and discrimination (higher AUROC) compared to the training-based baseline (Label-only SFT), white-box confidence estimation baseline (P(True)) and test-time scaling baseline (Ensemble), which requires aggregating confidence scores from multiple sampling paths. Notably, our method surpasses Label-only SFT, Ensemble and P(True) by over 9% in average ECE across the four models using only a single inference pass. Our method also results in smoother confidence distributions and substantially reduces overconfidence, as reflected in the reliability diagram (Figure 4)

Method	NQ		Bamboogle		Average	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
Llama-3.1-8B-Instruct						
Vanilla	0.371	0.645	0.212	0.633	0.292	0.639
CoT	0.352	0.670	0.199	0.579	0.276	0.625
Noise-aware	0.289	0.667	0.140	0.806	0.215	0.737
Ensemble	0.334	**0.693**	0.173	0.680	0.254	0.687
Label-only SFT	0.273	0.653	0.151	0.721	0.212	0.687
P(True)	0.476	0.424	0.472	0.443	0.474	0.433
NOVA	**0.265**	0.674	**0.127**	**0.823**	**0.196**	**0.749**
Qwen2.5-7B-Instruct						
Vanilla	0.322	0.706	0.126	0.835	0.224	0.771
CoT	0.313	0.696	0.121	0.810	0.217	0.753
Noise-aware	0.304	0.641	0.135	0.753	0.220	0.697
Ensemble	0.325	0.694	0.127	0.831	0.226	0.763
Label-only SFT	0.335	0.667	0.129	0.658	0.232	0.663
P(True)	0.432	0.518	0.445	0.673	0.439	0.596
NOVA	**0.248**	**0.750**	**0.065**	**0.845**	**0.157**	**0.798**

Method	NQ		Bamboogle		Average	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
DeepSeek-R1-Distill-Llama-8B						
Vanilla	0.376	0.625	0.154	0.671	0.265	0.648
CoT	0.373	0.621	0.203	0.633	0.288	0.627
Noise-aware	0.290	0.605	0.153	0.658	0.222	0.632
Ensemble	0.351	0.590	0.143	**0.711**	0.247	0.651
Label-only SFT	0.329	0.604	0.251	0.684	0.290	0.644
P(True)	0.551	0.617	0.491	0.667	0.521	0.642
NOVA	**0.276**	**0.628**	**0.137**	0.697	**0.207**	**0.663**
DeepSeek-R1-Distill-Qwen-7B						
Vanilla	0.435	0.608	0.220	0.626	0.328	0.617
CoT	0.454	0.627	0.218	0.638	0.336	0.633
Noise-aware	0.376	0.477	0.143	0.625	0.260	0.551
Ensemble	0.398	0.597	0.198	0.550	0.298	0.574
Label-only SFT	0.402	**0.686**	0.216	0.650	0.309	0.668
P(True)	0.394	0.623	0.407	0.660	0.400	0.641
NOVA	**0.335**	0.640	**0.127**	**0.765**	**0.231**	**0.703**

Table 3: Out-of-Distribution (OOD) results with 5 passage per query on the NQ and Bamboogle datasets, demonstrating that **NOVA** maintains robust calibration performance and consistently outperforms several strong baselines even when facing varying amounts of retrieved context in unseen scenarios. More generalization test is in Appendix B.1.

discussed in Appendix B.6 . We further verify that these calibration gains do not come at the cost of answer correctness: **NOVA** maintains or improves average accuracy for three out of four evaluated models, with full accuracy results reported in Appendix B.8. These results underscore the effectiveness of our noise-aware training framework in enabling accurate epistemic uncertainty estimation in RAG settings, which is further illustrated by a case study of **NOVA**-trained models in Appendix B.5.

Performance gains derive from noise-aware reasoning rather than label fitting. To isolate the source of improvements, we compare **NOVA** with **Label only SFT**. Table 2 shows that fine-tuning on confidence labels alone is insufficient for optimal calibration. Notably, **Label only SFT** yields limited gains over Vanilla and even degrades performance on DeepSeek-R1 distilled models. This confirms that the effectiveness of **NOVA** stems not from merely fitting (answer, confidence) pairs, but from our noise-aware framework, which integrates *NOVA Rules* with high-quality reasoning traces containing accurate process judgments. Appendix B.10 presents a rule-sensitive ablation that partitions the evaluation data into three subsets according to the retrieval conditions targeted by Conflict Independence, Noise Invariance, and Parametric Fallback. Replacing the responses in each subset with **NOVA** responses consistently reduces ECE, showing that all three *NOVA Rules* contribute to the overall calibration improvement. Together, these results suggest that **NOVA** benefits from learning condition-specific reasoning behaviors tailored to different forms of retrieval noise, providing further evidence for the role of *NOVA Rules* in improving overall confidence calibration.

Method	Average		Method	Average	
	ECE ↓	AUROC ↑		ECE ↓	AUROC ↑
Llama-3.1-8B-Instruct			***DeepSeek-R1-Distill-Llama-8B***		
Vanilla	0.472	0.616	Vanilla	0.429	0.684
CoT	0.423	0.552	CoT	0.454	0.672
Noise-aware	0.318	0.655	Noise-aware	0.409	0.633
Ensemble	0.354	0.620	Ensemble	0.505	0.672
Label-only SFT	0.338	0.739	Label-only SFT	0.548	0.556
P(True)	0.584	0.440	P(True)	0.530	0.561
NOVA	**0.193**	**0.812**	**NOVA**	**0.286**	**0.685**
Qwen2.5-7B-Instruct			***DeepSeek-R1-Distill-Qwen-7B***		
Vanilla	0.517	0.615	Vanilla	0.548	0.609
CoT	0.489	0.635	CoT	0.583	0.613
Noise-aware	0.336	0.632	Noise-aware	0.531	0.556
Ensemble	0.516	0.613	Ensemble	0.611	0.600
Label-only SFT	0.446	0.640	Label-only SFT	0.694	0.508
P(True)	0.432	0.534	P(True)	0.422	0.637
NOVA	**0.306**	**0.695**	**NOVA**	**0.415**	**0.640**

Table 4: Average calibration performance of four models across four datasets with Contriever. Full results are in Appendix B.4. **Bold** and underline mark the best and second-best scores within each model.

NOVA demonstrates strong generalization. We evaluate the generalization of **NOVA** along two axes: increased information load and retriever shift. First, we increase the number of retrieved passages from $k = 3$ during training to $k = 5$ at inference, creating a synthetic OOD setting with higher information load. As shown in Table 3, **NOVA** reduces average ECE by 8% over Vanilla and consistently outperforms strong calibration baselines. Second, under a real-retriever setting using *Contriever*, **NOVA** consistently achieves the best average ECE and AUROC across all four backbone models, reducing ECE by 39.0% over Vanilla on average, as shown in Table 4. Additional generalization results and full results using *Contriever* are provided in Appendices B.1 and B.4, respectively. These studies suggest that **NOVA** does not merely overfit to a fixed training format but learns a generalized ability to recognize diverse passages and assign appropriate confidence scores.

NOVA enhances models' ability to judge passage utility and thereby enhances interpretability. A core premise of our framework is that accurate confidence estimation in RAG hinges on the model's ability to correctly assess the quality of retrieved contexts. Empirical results confirm that **NOVA** substantially sharpens this discriminative capability. Compared to vanilla baselines, our method improves passage utility judgment accuracy by approximately 10% on two instruction-tuned models; even for reasoning models with stronger inherent capabilities, it consistently yields gains of about 5% on two DeepSeek-distilled variants. Crucially, by requiring the model to explicitly verbalize these judgments before assigning a confidence score, **NOVA** provides superior interpretability, as supported by the human study in Appendix E, enabling users to directly link epistemic uncertainty to the model's assessment of the retrieval environment, rather than to opaque probability distributions (Fan et al., 2024).

7 Discussions

On the Calibration objective and the Improvements of NOVA. **NOVA** should be interpreted primarily as a framework for improving verbal confidence calibration under retrieval noise, rather than as a method that universally improves answer accuracy. In particular, Conflict Independence may reduce confidence when retrieved passages disagree even if the final answer is not corrected; its purpose is to avoid overconfident predictions when the evidence is unreliable. The improvement of **NOVA** over the Noise-aware prompting baseline also comes from supervised fine-tuning rather than from a different inference prompt, since both settings use the same rule-guided prompting structure. The Label-only SFT baseline further shows that the gains cannot be explained solely by fitting answer–confidence pairs, while the rule-sensitive ablation in Appendix B.10 indicates that each *NOVA Rules* contributes to the overall calibration improvement.

On the Generalization and limitations of NOVA. The training and evaluation settings should also be distinguished carefully. Gemini-generated passages and oracle passage-type labels are used only to construct and filter the synthetic training data. At inference time, **NOVA** operates on retrieved passages without access to these oracle signals. The results under *Contriever* retrieval show that training on controlled synthetic passages can transfer to naturally retrieved contexts, but they do not establish that improved passage-utility judgment is the sole cause of the observed calibration gains. Moreover, although **NOVA** is not uniformly the best method on every model and dataset combination, it achieves substantial improvements in average calibration performance and improves over the relevant baselines on the vast majority of evaluated datasets and backbones. The consistently strong average gains across diverse models and datasets support the robustness and practical value of the proposed framework.

On the Rationale for SFT on Black-Box Confidence Signals. Verbal confidence is particularly useful in black-box deployment settings, where end users typically interact with closed-source LLMs through APIs and cannot access logits, hidden states, or other internal uncertainty estimates. In such scenarios, confidence expressed through natural language provides an interface-level uncertainty signal that can be directly exposed to users. Supervised fine-tuning, on the other hand, can be viewed as a provider-side optimization mechanism for improving this signal when model access is available. This creates a practical connection between black-box usage and model-side development: the provider can improve verbal confidence during training, while the resulting confidence signal remains accessible to users even when the deployed model is treated as a black box. Further discussion is provided in Appendix C.

8 Conclusion

Our study exposes a fundamental vulnerability in RAG where retrieval noise can negatively impact verbal confidence performance. We identified that noisy passages actively inflate false certainty, making standard LLMs critically overconfident. To resolve this, we propose **NOVA**, a principled self-bootstrapping framework that equips models with intrinsic noise awareness. By enforcing specific consistency rules, **NOVA** teaches models to explicitly discern passage utility and decouple their confidence from misleading evidence without relying on external teacher models. Extensive experiments confirm that **NOVA** delivers substantial gains in calibration performance without sacrificing task accuracy. It also enhances the transparency and interpretability of the reasoning process, marking a crucial step toward robust and epistemically reliable retrieval-augmented LLMs.

Limitations

While **NOVA** demonstrates significant improvements in RAG confidence calibration, we acknowledge several limitations:

Model Scale and Access. Our evaluation is currently limited to open-source models in the 7B-8B parameter range. We did not extend our experiments to larger-scale models (e.g., 70B+) or proprietary models (e.g., *GPT-5*, *Gemini-3-Pro*). This exclusion is primarily due to the prohibitive computational costs of fine-tuning larger models.

Synthetic vs. Real-World Noise. Our training data construction relies on synthetically generating specific types of noise (counterfactual, relevant, irrelevant). While this provides precise control for learning, real-world retrieval errors are often more nuanced and may not fit neatly into these categories. It remains to be seen how well the model generalizes to organic noise in highly specialized domains (e.g., biomedical or legal RAG).

Scalability to Complex Contexts and Tasks. Our current evaluation focuses on short-form question answering with fixed-depth retrieval. Extending noise-aware calibration to long-form generation (e.g., summarization) remains non-trivial, as "hallucination" in long texts is granular and difficult to capture with a single scalar confidence score. Furthermore, applying our framework to ultra-long contexts typical of agentic search (Luo et al., 2025; Li et al., 2025b; Liu et al., 2025a) introduces new challenges: detecting contradictions across massive, dynamic information streams may incur prohibitive computational costs and suffer from attention degradation (e.g., "lost-in-the-middle" (Liu et al., 2024b) phenomena), necessitating more efficient mechanisms than our rule-based scanning.

Ethics Statements

Personally Identifying or Offensive Content. The experiments in this study utilize standard, publicly available academic datasets (HotpotQA (Yang et al., 2018), Natural Questions (Kwiatkowski et al., 2019), StrategyQA (Geva et al., 2021), Bamboogle (Press et al., 2023)) and a retrieval corpus based on wikipedia (Wikimedia, 2023). These sources are widely used in the research community and generally do not contain sensitive personally identifying information (PII) of private individuals or offensive content. The synthetic noise passages generated for our training data were created using *Gemini-2.5-Pro*, which employs built-in safety filters to prevent the generation of toxic or harmful content. We did not observe any offensive material in the generated samples during our manual quality checks.

Data Consent and Licenses. We strictly adhere to the licenses and terms of use for all datasets and models employed in this work. The datasets (HotpotQA, Natural Questions, StrategyQA, and Bamboogle), the Wikipedia corpus and the models (*Llama-3.1-Instruct-8B* (Touvron et al., 2023), *Qwen2.5-7B-Instruct* (Yang et al., 2025), *DeepSeek-R1-Distill-Llama-8B*, *DeepSeek-R1-Distill-Qwen-7B* (DeepSeek-AI, 2025)) are all open-source and distributed under permissive licenses (e.g., CC BY-SA, Apache 2.0) that permit academic research and modification. No new private data was collected from human subjects, and no crowdsourcing platforms were used.

Models. All open-source models were hosted and executed locally using the vLLM library (Kwon et al., 2023), while *Gemini-2.5-Pro* utilized to generate RAG passages were accessed through vertex AI (Google Cloud, 2026). For reproducibility, the experimental settings are detailed in Section §4 and Appendix A.

AI assistants were used solely for language polishing in this paper.

Acknowledgements

The authors of this paper were supported by the National Key Research and Development Program of China (2025YFE0200500), the ITSP Platform Research Project (ITS/189/23FP) from ITC of Hong Kong, SAR, China, and the AoE (AoE/E-601/24-N), the CRF (No. C6004-25G), the RIF (R6021-20) and the GRF (16205322) from RGC of Hong Kong, SAR, China.

References

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. 2025. Improving uncertainty estimation through semantically diverse language generation. *Preprint*, arXiv:2406.04306.

Siddhant Arora, Haidar Khan, Kai Sun, Xin Luna Dong, Sajal Choudhary, Seungwhan Moon, Xinyuan Zhang, Adithya Sagar, Surya Teja Appini, Kaushik Patnaik, Sanat Sharma, Shinji Watanabe, Anuj Kumar, Ahmed Aly, Yue Liu, Florian Metze, and Zhaojiang Lin.

2025. Stream rag: Instant and accurate spoken dialogue systems with streaming tool usage. *Preprint*, arXiv:2510.02044.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2024. Self-rag: Learning to retrieve, generate, and critique through self-reflection. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. 2025. Hallulens: LLM hallucination benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 24128–24156. Association for Computational Linguistics.

C.E. Bonferroni. 1936. *Teoria statistica delle classi e calcolo delle probabilità*. Pubblicazioni del R. Istituto superiore di scienze economiche e commerciali di Firenze. Seeber.

Glenn W Brier. 1950. Verification of forecasts expressed in terms of probability. *Monthly Weather Review*, 78(1):1–3.

John Brooke. 2013. Sus: a retrospective. *J. Usability Studies*, 8(2):29–40.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems 33: Annual Conference on Neural Information Processing Systems 2020, NeurIPS 2020, December 6-12, 2020, virtual*.

Guanhua Chen, Yutong Yao, Lidia S. Chao, Xuebo Liu, and Derek F. Wong. 2025a. SGIC: A self-guided iterative calibration framework for RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics*, pages 28357–28370.

Yongchao Chen, Yueying Liu, Junwei Zhou, Yilun Hao, Jingquan Wang, Yang Zhang, and Chuchu Fan. 2025b. R1-code-interpreter: Training llms to reason with code via supervised and reinforcement learning. *CoRR*, abs/2505.21668.

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.

Florin Cuconasu, Giovanni Trappolini, Federico Siciliano, Simone Filice, Cesare Campagnano, Yoelle Maarek, Nicola Tonellotto, and Fabrizio Silvestri. 2024. The power of noise: Redefining retrieval for rag systems. In *Proceedings of the 47th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR 2024, page 719–729. ACM.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, and Jacob Andreas. 2025a. Beyond binary rewards: Training lms to reason about their uncertainty. *arXiv preprint arXiv:2507.16806*.

Mehul Damani, Isha Puri, Stewart Slocum, Idan Shenfeld, Leshem Choshen, Yoon Kim, and Jacob Andreas. 2025b. Beyond binary rewards: Training lms to reason about their uncertainty. *Preprint*, arXiv:2507.16806.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Shehzaad Dhuliawala, Mojtaba Komeili, Jing Xu, Roberta Raileanu, Xian Li, Asli Celikyilmaz, and Jason Weston. 2024. Chain-of-verification reduces hallucination in large language models. In *Findings of the Association for Computational Linguistics, ACL 2024, Bangkok, Thailand and virtual meeting, August 11-16, 2024*, volume ACL 2024 of *Findings of ACL*, pages 3563–3578. Association for Computational Linguistics.

Guanting Dong, Jiajie Jin, Xiaoxi Li, Yutao Zhu, Zhicheng Dou, and Ji-Rong Wen. 2025. Rag-critic: Leveraging automated critic-guided agentic workflow for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 3551–3578. Association for Computational Linguistics.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 2024. Shifting attention to relevance: Towards the predictive uncertainty quantification of free-form large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2024, Bangkok, Thailand, August 11-16, 2024*, pages 5050–5063. Association for Computational Linguistics.

Olive Jean Dunn. 1964. Multiple comparisons using rank sums. *Technometrics*, 6:241–252.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and

Artem Shelmanov. 2023. LM-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461, Singapore. Association for Computational Linguistics.

Run-Ze Fan, Xuefeng Li, Haoyang Zou, Junlong Li, Shwai He, Ethan Chern, Jiewen Hu, and Pengfei Liu. 2024. Reformatted alignment. *Preprint*, arXiv:2402.12219.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024a. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10028–10039, Bangkok, Thailand. Association for Computational Linguistics.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024b. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. *Preprint*, arXiv:2405.20978.

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. 2020. Unsupervised quality estimation for neural machine translation. *Trans. Assoc. Comput. Linguistics*, 8:539–555.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024a. A survey of confidence estimation and calibration in large language models. 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 6577–6595, Mexico City, Mexico. Association for Computational Linguistics.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024b. A survey of confidence estimation and calibration in large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers), NAACL 2024, Mexico City, Mexico, June 16-21, 2024*, pages 6577–6595. Association for Computational Linguistics.

Mor Geva, Daniel Khashabi, Elad Segal, Tushar Khot, Dan Roth, and Jonathan Berant. 2021. Did aristotle use a laptop? A question answering benchmark with implicit reasoning strategies. *Trans. Assoc. Comput. Linguistics*, 9:346–361.

Reshmi Ghosh, Rahul Seetharaman, Hitesh Wadhwa, Somyaa Aggarwal, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazadeh. 2024. Quantifying reliance on external information over parametric knowledge during retrieval augmented generation (rag) using mechanistic analysis. *Preprint*, arXiv:2410.00857.

Google Cloud. 2026. Vertex ai api. `https://cloud.google.com/vertex-ai`. Accessed: January 5, 2026.

Google DeepMind. 2025. Gemini 2.5 pro overview. `https://deepmind.google/technologies/gemini/pro/`.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017a. On calibration of modern neural networks. In *International Conference on Machine Learning*, pages 1321–1330. PMLR.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017b. On calibration of modern neural networks. *Preprint*, arXiv:1706.04599.

Dadi Guo, Yuejin Xie, Qingyu Liu, Jiayu Liu, Zhiyuan Fan, Qihan Ren, Shuai Shao, Tianyi Zhou, Dongrui Liu, and Yi R Fung. 2026. Code2math: Can your code agent effectively evolve math problems through exploration? *arXiv preprint arXiv:2603.03202*.

Yang Guo, Yutian Tao, Yifei Ming, Robert D. Nowak, and Yingyu Liang. 2025. Retrieval-augmented generation as noisy in-context learning: A unified theory and risk bounds. *Preprint*, arXiv:2506.03100.

Hyeonjeong Ha, Jeonghwan Kim, Cheng Qian, Jiayu Liu, William M Campbell, Yue Wu, Yuji Zhang, Kathleen McKeown, Dilek Hakkani-Tur, and Heng Ji. 2026. Memguard: Preventing memory contamination in long-term memory-augmented large language models. *arXiv preprint arXiv:2605.28009*.

Andreas Holzinger, André M. Carrington, and Heimo Müller. 2019. Measuring the quality of explanations: The system causability scale (scs). *Kunstliche Intelligenz*, 34:193 – 198.

Chengsong Huang, Langlin Huang, Jixuan Leng, Jiacheng Liu, and Jiaxin Huang. 2025. Efficient test-time scaling via self-calibration. *Preprint*, arXiv:2503.00031.

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2021. Unsupervised dense information retrieval with contrastive learning.

Zhen Jia, Abdalghani Abujabal, Rishiraj Saha Roy, Jannik Strötgen, and Gerhard Weikum. 2018. Tempquestions: A benchmark for temporal question answering. In *Companion of the The Web Conference 2018 on The Web Conference 2018, WWW 2018, Lyon , France, April 23-27, 2018*, pages 1057–1062. ACM.

Zhuoran Jin, Hongbang Yuan, Tianyi Men, Pengfei Cao, Yubo Chen, Jiexin Xu, Huaijun Li, Xiaojian Jiang, Kang Liu, and Jun Zhao. 2025. Rag-rewardbench: Benchmarking reward models in retrieval augmented generation for preference alignment. In *Findings of*

the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025, pages 17061–17090. 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. 2022a. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

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. 2022b. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katie Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. Large language models must be taught to know what they don't know. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Klaus Krippendorff. 2011. Computing krippendorff's alpha-reliability.

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.

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.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph E. Gonzalez, Hao Zhang, and Ion Stoica. 2023. Efficient memory management for large language model serving with pagedattention. *Preprint*, arXiv:2309.06180.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2021. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Preprint*, arXiv:2005.11401.

Jingshu Li, Yitian Yang, Q. Vera Liao, Junti Zhang, and Yi-Chieh Lee. 2025a. As confidence aligns: Understanding the effect of AI confidence on human self-confidence in human-ai decision making. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems, CHI 2025, YokohamaJapan, 26 April 2025- 1 May 2025*, pages 1111:1–1111:16. ACM.

Junlong Li, Wenshuo Zhao, Jian Zhao, Weihao Zeng, Haoze Wu, Xiaochen Wang, Rui Ge, Yuxuan Cao, Yuzhen Huang, Wei Liu, Junteng Liu, Zhaochen Su, Yiyang Guo, Fan Zhou, Lueyang Zhang, Juan Michelini, Xingyao Wang, Xiang Yue, Shuyan Zhou, and 2 others. 2025b. The tool decathlon: Benchmarking language agents for diverse, realistic, and long-horizon task execution. *Preprint*, arXiv:2510.25726.

Yibo Li, Miao Xiong, Jiaying Wu, and Bryan Hooi. 2025c. Conftuner: Training large language models to express their confidence verbally. *arXiv preprint arXiv:2508.18847*.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Trans. Mach. Learn. Res.*, 2022.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024a. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024b. Generating with confidence: Uncertainty quantification for black-box large language models. *Trans. Mach. Learn. Res.*, 2024.

Jiayu Liu, Qihan Lin, Cheng Qian, Rui Wang, Emre Can Acikgoz, Xiaocheng Yang, Jiateng Liu, Zhenhailong Wang, Xiusi Chen, Heng Ji, and 1 others. 2026a. Planbench-xl: Evaluating long-horizon planning of llm tool-use agents in large-scale tool ecosystems. *arXiv preprint arXiv:2606.22388*.

Jiayu Liu, Cheng Qian, Zhaochen Su, Qing Zong, Shijue Huang, Bingxiang He, and Yi R. Fung. 2025a. Costbench: Evaluating multi-turn cost-optimal planning and adaptation in dynamic environments for LLM tool-use agents. *CoRR*, abs/2511.02734.

Jiayu Liu, Cheng Qian, Zhenhailong Wang, Bingxuan Li, Jiateng Liu, Heng Wang, Jeonghwan Kim, Yumeng Wang, Xiusi Chen, Yi R Fung, and 1 others. 2026b. Adaplanbench: Evaluating adaptive planning in large language model agents under world and user constraints. *arXiv preprint arXiv:2606.05622*.

Jiayu Liu, Junhao Tang, Hanwen Wang, Baixuan Xu, Haochen Shi, Weiqi Wang, and Yangqiu Song. 2024a. GProofT: A multi-dimension multi-round fact checking framework based on claim fact extraction. In *Proceedings of the Seventh Fact Extraction and VERification Workshop (FEVER)*, pages 118–129, Miami,

Florida, USA. Association for Computational Linguistics.

Jiayu Liu, Qing Zong, Weiqi Wang, and Yangqiu Song. 2025b. Revisiting epistemic markers in confidence estimation: Can markers accurately reflect large language models' uncertainty? In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 206–221, Vienna, Austria. Association for Computational Linguistics.

Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2024b. Lost in the middle: How language models use long contexts. *Transactions of the Association for Computational Linguistics*, 12:157–173.

Shaoteng Liu, Haoqi Yuan, Minda Hu, Yanwei Li, Yukang Chen, Shu Liu, Zongqing Lu, and Jiaya Jia. 2024c. RL-GPT: integrating reinforcement learning and code-as-policy. In *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2024, NeurIPS 2024, Vancouver, BC, Canada, December 10 - 15, 2024*.

Shuyi Liu, Yuming Shang, and Xi Zhang. 2025c. Truthfulrag: Resolving factual-level conflicts in retrieval-augmented generation with knowledge graphs. *Preprint*, arXiv:2511.10375.

Shuyi Liu, Yuming Shang, and Xi Zhang. 2025d. Truthfulrag: Resolving factual-level conflicts in retrieval-augmented generation with knowledge graphs. *Preprint*, arXiv:2511.10375.

Xiaoou Liu, Tiejin Chen, Longchao Da, Chacha Chen, Zhen Lin, and Hua Wei. 2025e. Uncertainty quantification and confidence calibration in large language models: A survey. *CoRR*, abs/2503.15850.

Haotian Luo, Huaisong Zhang, Xuelin Zhang, Haoyu Wang, Zeyu Qin, Wenjie Lu, Guozheng Ma, Haiying He, Yingsha Xie, Qiyang Zhou, Zixuan Hu, Hongze Mi, Yibo Wang, Naiqiang Tan, Hong Chen, Yi R. Fung, Chun Yuan, and Li Shen. 2025. Ultrahorizon: Benchmarking agent capabilities in ultra long-horizon scenarios. *Preprint*, arXiv:2509.21766.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. 2026. Calibrating large language models with sample consistency. *Preprint*, arXiv:2402.13904.

Patrick E. McKight and Julius Najab. 2010. *Kruskal-Wallis Test*, pages 1–1. John Wiley & Sons, Ltd.

Viktor Moskvoretskii, Maria Marina, Mikhail Salnikov, Nikolay Ivanov, Sergey Pletenev, Daria Galimzianova, Nikita Krayko, Vasily Konovalov, Irina Nikishina, and Alexander Panchenko. 2025. Adaptive retrieval without self-knowledge? bringing uncertainty back home. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6355–6384.

Azadeh Sadat Mozafari, Hugo Siqueira Gomes, Wilson Leão, Steeven Janny, and Christian Gagné. 2019. Attended temperature scaling: A practical approach for calibrating deep neural networks. *Preprint*, arXiv:1810.11586.

Sumit Neelam, Udit Sharma, Hima Karanam, Shajith Ikbal, Pavan Kapanipathi, Ibrahim Abdelaziz, Nandana Mihindukulasooriya, Young-Suk Lee, Santosh Srivastava, Cezar Pendus, Saswati Dana, Dinesh Garg, Achille Fokoue, G P Shrivatsa Bhargav, Dinesh Khandelwal, Srinivas Ravishankar, Sairam Gurajada, Maria Chang, Rosario Uceda-Sosa, and 6 others. 2022. A benchmark for generalizable and interpretable temporal question answering over knowledge bases. *Preprint*, arXiv:2201.05793.

Stephen Obadinma and Xiaodan Zhu. 2025. On the robustness of verbal confidence of llms in adversarial attacks. *Preprint*, arXiv:2507.06489.

Litu Ou, Kuan Li, Huifeng Yin, Liwen Zhang, Zhongwang Zhang, Xixi Wu, Rui Ye, Zile Qiao, Pengjun Xie, Jingren Zhou, and Yong Jiang. 2025. Browseconf: Confidence-guided test-time scaling for web agents. *Preprint*, arXiv:2510.23458.

Jie Ouyang, Tingyue Pan, Mingyue Cheng, Ruiran Yan, Yucong Luo, Jiaying Lin, and Qi Liu. 2025. Hoh: A dynamic benchmark for evaluating the impact of outdated information on retrieval-augmented generation. *Preprint*, arXiv:2503.04800.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2022. Training language models to follow instructions with human feedback. *Preprint*, arXiv:2203.02155.

Shintaro Ozaki, Yuta Kato, Siyuan Feng, Masayo Tomita, Kazuki Hayashi, Wataru Hashimoto, Ryoma Obara, Masafumi Oyamada, Katsuhiko Hayashi, Hidetaka Kamigaito, and Taro Watanabe. 2025a. Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 1–17, Viena, Austria. Association for Computational Linguistics.

Shintaro Ozaki, Yuta Kato, Siyuan Feng, Masayo Tomita, Kazuki Hayashi, Wataru Hashimoto, Ryoma Obara, Masafumi Oyamada, Katsuhiko Hayashi, Hidetaka Kamigaito, and Taro Watanabe. 2025b. Understanding the impact of confidence in retrieval augmented generation: A case study in the medical domain. In *Proceedings of the 24th Workshop on Biomedical Language Processing*, pages 1–17.

Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. Measuring and narrowing the compositionality gap in language

models. In *Findings of the Association for Computational Linguistics: EMNLP 2023, Singapore, December 6-10, 2023*, pages 5687–5711. Association for Computational Linguistics.

Cheng Qian, Hyeonjeong Ha, Jiayu Liu, Bingxiang He, Jeonghwan Kim, Jiateng Liu, Bingxuan Li, Aditi Tiwari, Dwip Dalal, Zhenhailong Wang, and 1 others. 2026a. Creativitybench: Evaluating agent creative reasoning via affordance-based tool repurposing. *arXiv preprint arXiv:2605.02910*.

Cheng Qian, Hyeonjeong Ha, Jiayu Liu, Jeonghwan Kim, Emre Can Acikgoz, Bingxuan Li, Kunlun Zhu, Jiateng Liu, Aditi Tiwari, Zhenhailong Wang, and 1 others. 2026b. Advancing creative physical intelligence in large multimodal models. *arXiv preprint arXiv:2605.26396*.

Cheng Qian, Jiayu Liu, and Heng Ji. 2026c. User-harness: Harnessing user minds for stronger agent theory-of-mind. *arXiv preprint arXiv:2605.27721*.

Stephen Robertson, Hugo Zaragoza, and Michael Taylor. 2004. Simple bm25 extension to multiple weighted fields. In *Proceedings of the Thirteenth ACM International Conference on Information and Knowledge Management*, CIKM '04, page 42–49, New York, NY, USA. Association for Computing Machinery.

Michael Shen, Muhammad Umar, Kiwan Maeng, G. Edward Suh, and Udit Gupta. 2024. Towards understanding systems trade-offs in retrieval-augmented generation model inference. *Preprint*, arXiv:2412.11854.

Andrew Silva, Mariah Schrum, Erin Hedlund-Botti, Nakul Gopalan, and Matthew Gombolay. 2023. Explainable artificial intelligence: Evaluating the objective and subjective impacts of xai on human-agent interaction. *International Journal of Human–Computer Interaction*, 39(7):1390–1404.

Jasper Snoek, Yaniv Ovadia, Emily Fertig, Balaji Lakshminarayanan, Sebastian Nowozin, D. Sculley, Joshua V. Dillon, Jie Ren, and Zachary Nado. 2019. Can you trust your model's uncertainty? evaluating predictive uncertainty under dataset shift. In *Advances in Neural Information Processing Systems 32: Annual Conference on Neural Information Processing Systems 2019, NeurIPS 2019, December 8-14, 2019, Vancouver, BC, Canada*, pages 13969–13980.

Ionut-Teodor Sorodoc, Leonardo F. R. Ribeiro, Rexhina Blloshmi, Christopher Davis, and Adrià de Gispert. 2025. Garage: A benchmark with grounding annotations for RAG evaluation. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 17030–17049. Association for Computational Linguistics.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2025a. Why uncertainty estimation methods fall short in RAG: An axiomatic analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16596–16616, Vienna, Austria. Association for Computational Linguistics.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2025b. Why uncertainty estimation methods fall short in RAG: An axiomatic analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16596–16616, Vienna, Austria. Association for Computational Linguistics.

Heydar Soudani, Hamed Zamani, and Faegheh Hasibi. 2025c. Uncertainty quantification for retrieval-augmented reasoning. *Preprint*, arXiv:2510.11483.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. Lacie: Listener-aware finetuning for confidence calibration in large language models. *Preprint*, arXiv:2405.21028.

Nisan Stiennon, Long Ouyang, Jeff Wu, Daniel M. Ziegler, Ryan Lowe, Chelsea Voss, Alec Radford, Dario Amodei, and Paul Christiano. 2022. Learning to summarize from human feedback. *Preprint*, arXiv:2009.01325.

Duolin Sun, Dan Yang, Yue Shen, Yihan Jiao, Zhehao Tan, Jie Feng, Lianzhen Zhong, Jian Wang, Peng Wei, and Jinjie Gu. 2025a. Hanrag: Heuristic accurate noise-resistant retrieval-augmented generation for multi-hop question answering. *Preprint*, arXiv:2509.09713.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Zilei Wang, Weiqiang Wang, and Liang Wang. 2025b. Divide-then-align: Honest alignment based on the knowledge boundary of RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 11461–11480. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurélien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models. *CoRR*, abs/2302.13971.

Hitesh Wadhwa, Rahul Seetharaman, Somyaa Aggarwal, Reshmi Ghosh, Samyadeep Basu, Soundararajan Srinivasan, Wenlong Zhao, Shreyas Chaudhari, and Ehsan Aghazadeh. 2024. From rags to rich parameters: Probing how language models utilize external knowledge over parametric information for factual queries. *CoRR*, abs/2406.12824.

Chenyu Wang, Tianle Chen, H. M. Sabbir Ahmad, Kayhan Batmanghelich, and Wenchao Li. 2026a. Vlm-uqbench: A benchmark for modality-specific and cross-modality uncertainties in vision language models. *Preprint*, arXiv:2602.09214.

Fei Wang, Xingchen Wan, Ruoxi Sun, Jiefeng Chen, and Sercan Ö. Arik. 2025a. Astute RAG: overcoming imperfect retrieval augmentation and knowledge conflicts for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 30553–30571. Association for Computational Linguistics.

Rui Wang, Qihan Lin, Jiayu Liu, Qing Zong, Tianshi Zheng, Weiqi Wang, and Yangqiu Song. 2025b. Prospect theory fails for llms: Revealing instability of decision-making under epistemic uncertainty. *arXiv preprint arXiv:2508.08992*.

Siyuan Wang, James R. Foulds, Md Osman Gani, and Shimei Pan. 2025c. Llm-based corroborating and refuting evidence retrieval for scientific claim verification. *Preprint*, arXiv:2503.07937.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. *Preprint*, arXiv:2203.11171.

Yumeng Wang, Zhiyuan Fan, Jiayu Liu, Jen tse Huang, and Yi R. Fung. 2026b. Diversity-enhanced reasoning for subjective questions. In *The Fourteenth International Conference on Learning Representations*.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025a. Browsecomp: A simple yet challenging benchmark for browsing agents. *CoRR*, abs/2504.12516.

Jason Wei, Zhiqing Sun, Spencer Papay, Scott McKinney, Jeffrey Han, Isa Fulford, Hyung Won Chung, Alex Tachard Passos, William Fedus, and Amelia Glaese. 2025b. Browsecomp: A simple yet challenging benchmark for browsing agents. *arXiv preprint arXiv:2504.12516*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed H. Chi, Quoc V. Le, and Denny Zhou. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems 35: Annual Conference on Neural Information Processing Systems 2022, NeurIPS 2022, New Orleans, LA, USA, November 28 - December 9, 2022*.

David Widmann, Fredrik Lindsten, and Dave Zachariah. 2020. Calibration tests in multi-class classification: A unifying framework. *Preprint*, arXiv:1910.11385.

Wikimedia. 2023. Wikipedia dataset (20231101.en). `https://huggingface.co/datasets/wikimedia/wikipedia/viewer/20231101.en`.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, and 3 others. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, EMNLP 2020 - Demos, Online, November 16-20, 2020*, pages 38–45. Association for Computational Linguistics.

Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Pengpeng Shao, and Jianhua Tao. 2025a. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of RAG noise in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5019–5039, Vienna, Austria. Association for Computational Linguistics.

Jinyang Wu, Shuai Zhang, Feihu Che, Mingkuan Feng, Chuyuan Zhang, Pengpeng Shao, and Jianhua Tao. 2025b. Pandora's box or aladdin's lamp: A comprehensive analysis revealing the role of rag noise in large language models. *Preprint*, arXiv:2408.13533.

Tongshuang Wu, Michael Terry, and Carrie Jun Cai. 2022. Ai chains: Transparent and controllable human-ai interaction by chaining large language model prompts. In *Proceedings of the 2022 CHI Conference on Human Factors in Computing Systems*, CHI '22, New York, NY, USA. Association for Computing Machinery.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. 2025. A survey of uncertainty estimation methods on large language models. In *Findings of the Association for Computational Linguistics, ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 21381–21396. Association for Computational Linguistics.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. *Preprint*, arXiv:2409.19817.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Rongwu Xu, Zehan Qi, Zhijiang Guo, Cunxiang Wang, Hongru Wang, Yue Zhang, and Wei Xu. 2024a. Knowledge conflicts for llms: A survey. *Preprint*, arXiv:2403.08319.

Tianyang Xu, Shujin Wu, Shizhe Diao, Xiaoze Liu, Xingyao Wang, Yangyi Chen, and Jing Gao. 2024b. SaySelf: Teaching LLMs to express confidence with self-reflective rationales. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5985–5998, Miami, Florida, USA. Association for Computational Linguistics.

Daniel Yang, Yao-Hung Hubert Tsai, and Makoto Yamada. 2024. On verbalized confidence scores for llms. *Preprint*, arXiv:2412.14737.

Qwen: An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, and 23 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. Hotpotqa: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, Brussels, Belgium, October 31 - November 4, 2018*, pages 2369–2380. Association for Computational Linguistics.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, volume ACL 2023 of *Findings of ACL*, pages 8653–8665. Association for Computational Linguistics.

Ori Yoran, Tomer Wolfson, Ori Ram, and Jonathan Berant. 2024. Making retrieval-augmented language models robust to irrelevant context. In *The Twelfth International Conference on Learning Representations*.

Qiying Yu, Zheng Zhang, Ruofei Zhu, Yufeng Yuan, Xiaochen Zuo, Yu Yue, Tiantian Fan, Gaohong Liu, Lingjun Liu, Xin Liu, Haibin Lin, Zhiqi Lin, Bole Ma, Guangming Sheng, Yuxuan Tong, Chi Zhang, Mofan Zhang, Wang Zhang, Hang Zhu, and 16 others. 2025. DAPO: an open-source LLM reinforcement learning system at scale. *CoRR*, abs/2503.14476.

Yifu Yuan, Haiqin Cui, Yaoting Huang, Yibin Chen, Fei Ni, Zibin Dong, Pengyi Li, Yan Zheng, and Jianye Hao. 2025. Embodied-r1: Reinforced embodied reasoning for general robotic manipulation. *Preprint*, arXiv:2508.13998.

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.

Jiayi Zhang, Simon Yu, Derek Chong, Anthony Sicilia, Michael R. Tomz, Christopher D. Manning, and Weiyan Shi. 2025a. Verbalized sampling: How to mitigate mode collapse and unlock llm diversity. *Preprint*, arXiv:2510.01171.

Qinggang Zhang, Zhishang Xiang, Yilin Xiao, Le Wang, Junhui Li, Xinrun Wang, and Jinsong Su. 2025b. Faithfulrag: Fact-level conflict modeling for context-faithful retrieval-augmented generation. *Preprint*, arXiv:2506.08938.

Shengming Zhao, Yuheng Huang, Jiayang Song, Zhijie Wang, Chengcheng Wan, and Lei Ma. 2024. Towards understanding retrieval accuracy and prompt quality in RAG systems. *CoRR*, abs/2411.19463.

Chujie Zheng, Shixuan Liu, Mingze Li, Xiong-Hui Chen, Bowen Yu, Chang Gao, Kai Dang, Yuqiong Liu, Rui Men, An Yang, Jingren Zhou, and Junyang Lin. 2025. Group sequence policy optimization. *Preprint*, arXiv:2507.18071.

Yaowei Zheng, Richong Zhang, Junhao Zhang, Yanhan Ye, Zheyan Luo, Zhangchi Feng, and Yongqiang Ma. 2024. Llamafactory: Unified efficient fine-tuning of 100+ language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, Bangkok, Thailand. Association for Computational Linguistics.

Xiaofan Zhou, Liangjie Huang, Pinyang Cheng, Wenpen Yin, Rui Zhang, Wenrui Hao, and Lu Cheng. 2025a. Accelerating causal network discovery of alzheimer disease biomarkers via scientific literature-based retrieval augmented generation. *Preprint*, arXiv:2504.08768.

Youchao Zhou, Heyan Huang, Yicheng Liu, Rui Dai, Xinglin Wang, Xingchen Zhang, Shumin Shi, and Yang Deng. 2025b. Do retrieval augmented language models know when they don't know? *Preprint*, arXiv:2509.01476.

Youchao Zhou, Heyan Huang, Yicheng Liu, Rui Dai, Xinglin Wang, Xingchen Zhang, Shumin Shi, and Yang Deng. 2025c. Do retrieval augmented language models know when they don't know? *CoRR*, abs/2509.01476.

Qing Zong, Jiayu Liu, Tianshi Zheng, Chunyang Li, Baixuan Xu, Haochen Shi, Weiqi Wang, Zhaowei Wang, Chunkit Chan, and Yangqiu Song. 2025a. Critical: Can critique help LLM uncertainty or confidence calibration? *CoRR*, abs/2510.24505.

Qing Zong, Zhaowei Wang, Tianshi Zheng, Xiyu Ren, and Yangqiu Song. 2025b. ComparisonQA: Evaluating factuality robustness of LLMs through knowledge frequency control and uncertainty. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 4101–4117, Vienna, Austria. Association for Computational Linguistics.

Appendices

A Detailed Experiment Setup

A.1 Models

We use *Qwen/Qwen2.5-7B-Instruct* (Yang et al., 2025), *meta-llama/Llama-3.1-8B-Instruct* (Touvron et al., 2023), *deepseek-ai/DeepSeek-R1-Distill-Qwen-7B*, and *deepseek-ai/DeepSeek-R1-Distill-Llama-8B* (DeepSeek-AI, 2025) in all experiments. Proprietary models were excluded from our study because their limited accessibility to internal parameters constrains further optimization and adaptation. For inference-time hyperparameters, we set the maximum output length to 2048 and use a temperature of 0 to ensure deterministic responses.

A.2 Inference and Training Backend

We use vLLM (Kwon et al., 2023) as the inference backend and LLaMAFactory (Zheng et al., 2024) for all training, with both inference and training conducted on 4 NVIDIA L20 GPUs.

A.3 Dataset Statistics

Table 5 reports the number of samples in each dataset along with the 95% confidence intervals of ECE and AUROC, computed using the method proposed by Widmann et al. (2020). The results indicate that the scale of our datasets is sufficient to yield reliable estimates.

Dataset	# Questions	Confidence Interval
HotpotQA	800	$\pm$ 0.0347
StrategyQA	800	$\pm$ 0.0347
NQ	800	$\pm$ 0.0347
Bamboogle	150	$\pm$ 0.0800

Table 5: Dataset statistics and 95% confidence intervals of ECE and AUROC.

A.4 Prompts

A.4.1 RAG Test Prompts

We adopt three types of prompts—Vanilla, CoT, and Multi-Step—from Xiong et al. (2024). For reasoning-oriented models, step-level confidences are elicited by requiring the models to report their confidence scores in the final output after the reasoning process. The prompt designs are illustrated in Figure 9, and the results of prompt permutation experiments are reported in Appendix B.3.

Hyperparameter	BM25	Contriever
Top-K Retrieval	5	5
Reranker	No	No
Model Specifics		
Architecture	Sparse (Probabilistic)	Dense (Bi-Encoder)
Embedding Model	N/A	`facebook/contriever`
Max Input Length	N/A	256 tokens
KNN Candidates	N/A	100

Table 6: Retrieval Hyperparameters.

A.4.2 Noise Generation Prompt

We detail the methodology for constructing the noise passages used in our RAG experiments. To generate high quality and semantically diverse noise, we employ *Gemini 2.5 Pro*. Specifically, we design three distinct types of noise prompts: *counterfactual noise generation prompt*, *relevant noise generation prompt*, and *irrelevant noise generation prompt*, corresponding to *counterfactual noise*, *relevant noise* and *irrelevant noise*. For each type, the prompt provided to the model includes a clear definition of the noise category and concrete examples to guide the generation. The full templates for all three prompt types are presented in Figure 12, 14 and 16. During generation, to encourage greater diversity in the output, we instruct the model to produce 5 candidate passages per call. We explicitly enhance the diversity of generated content using Zhang et al. (2025a), then select only the last three generated passages as the final noise passages for our experimental setup in Table 2.

A.4.3 Baseline Prompts

For the baselines used in the main experiments (Table 2 and 3), Vanilla, CoT prompt is provided in Figure 9 and noise-aware prompt is provided in Figure 10.

A.5 RAG Setup

In this section, we detail the hyperparameters used for our Retrieval-Augmented Generation (RAG) setup. We summarize the specific configurations for both the sparse retriever (BM25) and the dense retriever (Contriever) in Table 6. Specifically, we retrieve the top-$k = 5$ passages for all experiments. We do not employ any reranking models in this study following Soudani et al. (2025a). For the dense retriever, the input is truncated to 256 tokens during the embedding process.

Category	Sub-category	Definition
Counterfactual	—	Passages that are semantically relevant to the question but directly contradict the ground truth answer. They provide specific, plausible-sounding information that supports an incorrect alternative answer.
Relevant Noise	**Entity-relevant**	Passages that mention the correct entities in the question but only provide partial, tangential, or incomplete factual information, without containing the evidence needed to answer the question.
	Relation-relevant	Passages that capture the type of relations required by the question but do not involve the queried entities, thereby providing misleading or insufficient evidence.
	Theme-relevant	Passages that are topically aligned with the question and provide high-level background or contextual information, but do not contain entity-level or relation-level facts necessary for answering.
Irrelevant Noise	—	Passages that have little to no semantic relation to the question. They are from unrelated topics or domains and provide no useful information for answering.

Table 7: Definitions of noise passages for retrieval-augmented question answering. Relevant noise is further categorized into entity-level, relation-level, and theme-level noise to better simulate real-world conditions.

Model	Total	Kept Responses					
		(1) Format	(2) Passage Judgment	(3) Rule Following	(4) Alignment	(5) Common IDs	(6) Balance
DS-R1-Llama	96000	85723	39008	34403	5211	2801	1945
DS-R1-Qwen	96000	88201	28481	24586	4611	2801	1945
Llama-3.1	96000	78200	35255	28790	4895	2801	1945
Qwen-2.5	96000	94898	31065	26221	3609	2801	1945

Table 8: Training data statistics: This table shows the number of training data left after each filtering step. (1) Format: retains only samples from which a valid answer, a confidence score, and intermediate passage judgments can be successfully extracted. (2) Passage judgment: filters out samples containing incorrect assessments of the retrieved passages. (3) Rule following: filters for samples that have a explicit reasoning process for rule following. (4) Alignment: for each query, selects the final response that minimizes the instance-level Brier Score. (5) Common IDs: retains only samples with question IDs common across all models. (6) Balance: balances the 3 groups (counterfactual, consistent, irrelevant) by downsampling consistent to match irrelevant. Model name abbreviations: *DS-R1-Llama*: DeepSeek-R1-Distill-Llama-8B; *DS-R1-Qwen*: DeepSeek-R1-Distill-Qwen-7B; *Llama-3.1*: Llama-3.1-8B-Instruct; *Qwen-2.5*: Qwen2.5-7B-Instruct.

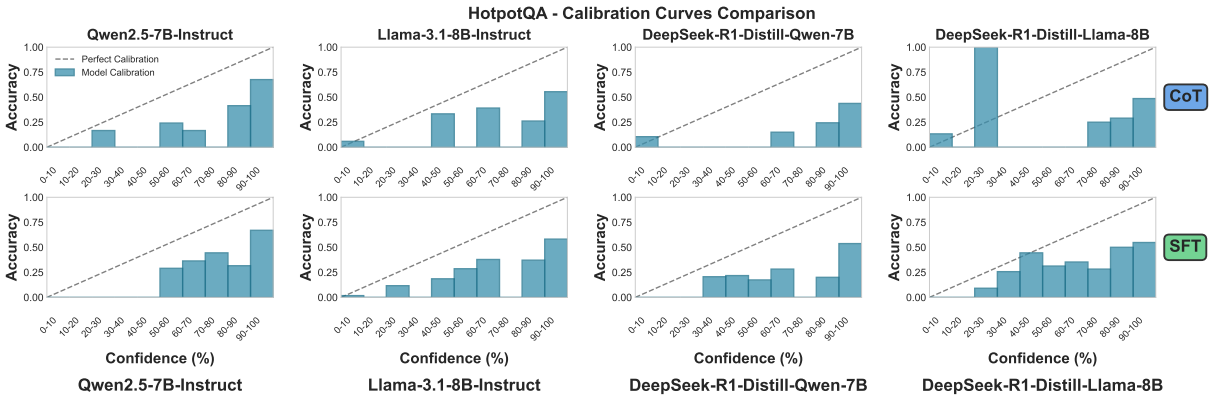


Figure 4: Reliability Diagram for HotpotQA: comparison of CoT prompt with base model (upper row) and SFT models (lower row). Each subplot displays accuracy v.s. confidence, with the diagonal dashed line representing perfect calibration.

A.6 SFT Details

We conducted Supervised Fine-Tuning (SFT) utilizing the LLaMA-Factory framework (Zheng et al., 2024). Specifically, we set the learning rate to 5.0×10^{-5} and the number of training epochs to 2. The maximum sequence length is set to 2048, aligning with the inference configuration to conserve computational resources. For all other training arguments and hyperparameters, we adhered to the default settings provided by LLaMA-Factory.

A.7 Training data statistics

We employ a self-consistency-based approach to construct our training data. For each input query, we generate 16 distinct response paths using self-sampling with a temperature setting of 1.0. The resulting dataset exhibits an average input token count of 646 and an average output token count of 370. To ensure high-quality supervision for calibration, we implement a comprehensive five-stage filtering pipeline specified in Section §6.2. Following the five-stage cleaning process, we observed discrepancies in the volume of retained data across different models. Therefore, prior to training, we apply an additional data-balancing step to ensure a consistent distribution, and provide concrete implementation details below.

1. **Format Consistency:** We retain only those samples where the answer, confidence score, and intermediate passage judgments can be successfully extracted via regular expressions.

2. **Passage Judgment Accuracy:** We filter out samples where the model's judgment of the retrieved passages and passage groups conflicts with the ground truth labels.

3. **Rule Following:** We discard samples that fail to exhibit the explicit reasoning process required by our instructions. Specifically, we filter out the samples that didn't incorporate keywords like "rules", "Step 4" (See Figure 11, we prompt the model to reason through multiple steps. "Step 4" is the designated step for applying our *NOVA Rules*; if it does not appear in the reasoning trace, the reasoning chain is considered invalid), etc.

4. **Alignment Selection:** From the remaining candidates for each query, we select the single final response that minimizes the instance-level Brier Score, ensuring the model learns from its most calibrated outputs.

5. **Common Intersection:** To allow for fair comparison, we retain only those questions for which valid responses exist across all four evaluated models.

6. **Class Balancing:** Finally, we balance the distribution of retrieval scenarios (counterfactual, consistent, irrelevant) by downsampling the dominant *consistent* class to match the size of the *irrelevant* class.

The process is fully rule-based, without any external model incorporated. The detailed statistics of data retention after each stage are presented in Table 8.

A.8 Fine-grained Noise Definitions

We categorize noise passages into three distinct types based on their semantic relationship to the query and their potential to mislead the answering process, as defined in Table 7. *Counterfactual passages* are adversarially designed contexts that are topically aligned with the question but contain specific, contradictory information that supports an incorrect alternative answer. *Relevant noise passages* mimic common retrieval errors by sharing keywords or general topics with the query while lacking the precise information needed to derive any answer. Specifically, we divide *relevant passages* into three types: *Entity-relevant* passages provide factual information about the entities involved in the query; *relationship-relevant* passages describe same interactions or relations among multiple unrelated entities; and theme-relevant passages offer broader background or contextual information aligned with the overall question intent. *Irrelevant noise* passages represent complete topic mismatches, providing no useful contextual information. This taxonomy is consistent with prior work on noise categorization for retrieval-augmented models, such as the similar three-type classification employed by Fang et al. (2024b).

B Additional Experiment Results

B.1 OOD Test on Four Passages

To further examine robustness across retrieval settings, additional experiments with four retrieved passages were conducted. As shown in Table 10, the overall pattern remains consistent with the main

Retriever	Prompt Type	StrategyQA		HotpotQA		NQ		Bamboogle		Average	
		ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
Llama-3.1-8B-Instruct											
BM25	Vanilla	0.266	0.550	0.515	0.626	0.446	0.696	0.755	0.554	0.495	0.607
	CoT	0.217	0.480	0.538	0.548	0.416	0.648	0.613	0.452	0.446	0.532
	Multi-Step	0.250	0.482	0.452	0.486	0.394	0.503	0.603	0.530	0.425	0.500
Contriever	Vanilla	0.284	0.563	0.614	0.576	0.490	0.638	0.735	0.619	0.531	0.599
	CoT	0.228	0.455	0.663	0.417	0.446	0.558	0.629	0.445	0.491	0.469
	Multi-Step	0.283	0.483	0.618	0.379	0.515	0.399	0.724	0.413	0.535	0.418
Qwen2.5-7B-Instruct											
BM25	Vanilla	0.223	0.546	0.436	0.692	0.442	0.735	0.668	0.579	0.442	0.638
	CoT	0.243	0.562	0.450	0.726	0.439	0.742	0.628	0.599	0.440	0.657
	Multi-Step	0.213	0.498	0.476	0.673	0.414	0.687	0.610	0.592	0.428	0.613
Contriever	Vanilla	0.218	0.562	0.575	0.622	0.571	0.698	0.690	0.652	0.513	0.633
	CoT	0.231	0.564	0.570	0.625	0.553	0.699	0.632	0.696	0.496	0.646
	Multi-Step	0.214	0.499	0.472	0.627	0.454	0.596	0.470	0.612	0.402	0.584
DeepSeek-R1-Distill-Llama-8B											
BM25	Vanilla	0.218	0.573	0.441	0.647	0.454	0.700	0.547	0.686	0.415	0.651
	CoT	0.246	0.574	0.459	0.628	0.461	0.707	0.521	0.754	0.422	0.666
	Multi-Step	0.316	0.523	0.496	0.555	0.461	0.613	0.672	0.535	0.486	0.556
Contriever	Vanilla	0.237	0.576	0.493	0.635	0.460	0.746	0.535	0.732	0.431	0.672
	CoT	0.235	0.572	0.527	0.615	0.477	0.754	0.557	0.714	0.449	0.664
	Multi-Step	0.300	0.513	0.581	0.592	0.468	0.633	0.686	0.551	0.509	0.572
DeepSeek-R1-Distill-Qwen-7B											
BM25	Vanilla	0.275	0.541	0.551	0.565	0.564	0.682	0.736	0.718	0.531	0.627
	CoT	0.292	0.539	0.583	0.591	0.560	0.686	0.756	0.584	0.548	0.600
	Multi-Step	0.275	0.525	0.597	0.499	0.524	0.601	0.734	0.424	0.532	0.512
Contriever	Vanilla	0.276	0.547	0.647	0.524	0.587	0.721	0.734	0.727	0.561	0.630
	CoT	0.292	0.539	0.644	0.455	0.600	0.723	0.773	0.632	0.577	0.587
	Multi-Step	0.272	0.507	0.638	0.490	0.570	0.617	0.750	0.573	0.557	0.547

Table 9: Evaluation of verbal confidence calibration performance (ECE and AUROC) on four datasets across varying retrievers and prompting strategies. Results show that the model consistently exhibits an average ECE greater than 0.4, indicating poor calibration performance.

results. Our approach remains the strongest calibration method under both the IID dataset (HotpotQA) and the OOD dataset (NQ).

B.2 Passage Position Bias

To investigate the impact positional bias, we evaluated model performance by embedding the ground truth passage among noise passages, placing the ground truth at various positions within the context window. The detailed results are presented in Tables 11 and 14. We observe a consistent trend: the introduction of noisy passages significantly degrades the quality of verbalized confidence, evidenced by the increased ECE and decreased AUROC. Crucially, this degradation remains pervasive regardless of the specific location of the ground truth passage, indicating that the models' vulnerability to noise is a fundamental issue rather than a position-dependent artifact.

B.3 Prompt Permutations

To investigate whether the poor calibration observed in RAG settings stems from the limitation of specific prompting strategies, we conduct a comprehensive evaluation across three distinct prompting paradigms: *Vanilla*, *Chain-of-Thought (CoT)*, and *Multi-Step* reasoning. We evaluate these strategies using both BM25 and Contriever retrieval settings across all four datasets. The detailed results are presented in Table 9.

Prompting strategies from close-book reasoning alone cannot efficiently enhance verbalized confidence performance in RAG scenarios. As evidenced by the results, models exhibit consistently unsatisfactory calibration performance across all prompt types. Sophisticated prompting strategies that have proven effective in closed-book reasoning scenarios (Xiong et al., 2024) do not yield significant calibration gains; notably, none of the methods achieves an average ECE below 0.4. For instance, the *Multi-Step* prompting strategy often exacerbates miscalibration compared to the *Vanilla* baseline, particularly in DeepSeek-distilled models (e.g., average ECE increases from 0.415 to 0.486 on *DeepSeek-R1-Distill-Llama-8B*).

These findings suggest that the calibration failure in RAG is a fundamental issue rooted in the model's inability to handle retrieval noise, rather than a superficial artifact of the prompting format. This underscores the necessity of a dedicated training framework like **NOVA** to align confidence with retrieval quality.

Method	HotpotQA		NQ		Method	HotpotQA		NQ	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑		ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
Llama-3.1-8B-Instruct					***Qwen2.5-7B-Instruct***				
Vanilla	0.358	0.630	0.389	0.655	Vanilla	0.333	0.692	0.363	0.707
CoT	0.340	0.672	0.370	0.679	CoT	0.327	0.712	0.349	0.685
Noise-aware	0.263	0.693	0.315	0.661	Noise-aware	0.285	0.673	0.316	0.650
Ensemble	0.316	0.648	0.364	0.683	Ensemble	0.318	0.718	0.363	0.707
Label-only SFT	0.273	0.697	0.326	0.653	Label-only SFT	0.301	0.510	0.327	0.512
P(True)	0.566	0.447	0.461	0.436	P(True)	0.467	0.521	0.442	0.501
NOVA	**0.261**	**0.723**	**0.286**	**0.684**	**NOVA**	**0.271**	**0.729**	**0.296**	**0.731**

Table 10: Four passages OOD results. Lower ECE and higher AUROC indicate better calibration performance.

Setting	Pos	Bamboogle		HotpotQA		NQ		StrategyQA		Average	
		ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC
gt_only	N/A	0.071	0.675	0.117	0.693	0.139	0.688	0.046	0.679	0.093	0.684
gt_with_noise/counterfactual	pos1	0.475	0.474	0.500	0.475	0.477	0.483	0.648	0.290	0.525	0.431
	pos2	0.416	0.581	0.514	0.504	0.490	0.557	0.609	0.291	0.507	0.483
	pos3	0.365	0.671	0.452	0.613	0.472	0.714	0.492	0.397	0.445	0.599
gt_with_noise/relevant	pos1	0.126	0.579	0.199	0.538	0.196	0.612	0.058	0.572	0.145	0.575
	pos2	0.131	0.583	0.196	0.549	0.193	0.624	0.050	0.603	0.143	0.590
	pos3	0.116	0.511	0.189	0.564	0.196	0.623	0.054	0.599	0.139	0.574
gt_with_noise/irrelevant	pos1	0.145	0.579	0.197	0.523	0.196	0.591	0.110	0.561	0.162	0.564
	pos2	0.152	0.529	0.196	0.539	0.195	0.557	0.108	0.577	0.163	0.551
	pos3	0.152	0.529	0.197	0.519	0.194	0.521	0.112	0.562	0.164	0.533

Table 11: The table evaluate the impact of passage ordering on *Llama-3.1-8B-Instruct*'s calibration performance. The "Setting" column defines the context structure, where gt_only refers to a noise-free baseline containing only the ground truth passage, while the gt_with_noise categories involve mixing the ground truth with specific types of noise (counterfactual, relevant, or irrelevant). The "Pos" column specifies the exact position (1st, 2nd, or 3rd) of the ground truth passage within the sequence of retrieved passages, designed to assess the model's sensitivity to positional bias when processing mixed-quality contexts. The results indicate that calibration performance steadily declines as noise passages are added.

B.4 NOVA Effectiveness in Real-world Settings

To further demonstrate the effectiveness of **NOVA**, we re-ran experiments under a real-world RAG setting, keeping the experimental configuration (e.g., model hyperparameters and RAG setup) consistent. As shown in Table 17 and Table 19, **NOVA** consistently maintains its effectiveness in realistic scenarios. Across the four evaluated models, under both BM25 and Contriever retriever settings, **NOVA** achieves the best average ECE and AUROC over the four datasets among all baselines. Compared to the Vanilla baseline, **NOVA** reduces the average ECE by 14.6% and 19.1% under the BM25 and Contriever settings, respectively. Notably, the improvement margins are even larger than those observed on the synthesized test set. This demonstrates the effectiveness of our pipeline and its strong generalization capability.

B.5 Case Studies

To illustrate the effectiveness of our framework in handling retrieval noise, we present a qualitative comparison between the Vanilla baseline and **NOVA** in Figure 8. In this high-conflict scenario characterized by counterfactual passages, the Vanilla model fails to detect the contradictions, hallucinating an answer with unwarranted high confidence (80%). In contrast, **NOVA** employs step-by-step reasoning to explicitly identify the discrepancies among the retrieved texts. By strictly adhering to the Conflict Independence rule, the model recognizes the epistemic uncertainty arising from the conflicting evidence and appropriately calibrates its verbalized confidence down to 10%, demonstrating superior reliability.

B.6 Reliability Diagram

Figure 4 presents the reliability diagrams for the HotpotQA dataset (in-domain results), comparing the calibration performance of the standard CoT prompting baseline against our proposed **NOVA**.

As observed in the top row, the CoT baseline exhibits severe miscalibration, characterized by a tendency towards overconfidence. The models often assign high confidence scores (near 100%) even when empirical accuracy is low, and they fail to utilize lower confidence bins effectively, resulting in

a sparse and uninformative distribution.

In contrast, **NOVA** (bottom row) significantly improves the alignment between predicted confidence and actual accuracy. The reliability curves for the fine-tuned models closely track the perfect calibration diagonal across a broad range of confidence bins. This indicates that **NOVA** successfully regularizes the model's outputs, transforming the confidence estimates into a more desired distribution where the verbalized score accurately reflects the probability of correctness.

B.7 Significance Analysis

We conduct significance analyses comparing NOVA with the Vanilla and CoT baselines using the aggregate results reported in Table 2. We report the baseline score, NOVA score, absolute improvement, p-value, and significance level for both ECE and AUROC. For ECE, ΔECE is defined as $\text{ECE}_{\text{baseline}} - \text{ECE}_{\text{NOVA}}$, whereas for AUROC, ΔAUROC is defined as $\text{AUROC}_{\text{NOVA}} - \text{AUROC}_{\text{baseline}}$. Thus, positive values indicate an improvement for **NOVA**.

As shown in Table 12, NOVA significantly reduces ECE relative to Vanilla across all four models. NOVA also improves AUROC for every model. All comparisons satisfy $p \leq 0.05$.

Table 13 shows the same pattern when comparing NOVA with CoT. All reported improvements are statistically significant at the $p \leq 0.05$ level.

B.8 Accuracy Results under NOVA

Following the experimental settings in Table 2, we compare the average accuracy across four datasets for vanilla prompting against our proposed **NOVA**. The results demonstrate that **NOVA** maintains or improves generation performance for the majority of the evaluated models. Specifically, *Qwen2.5-7B-Instruct*, *DeepSeek-R1-Distill-Qwen-7B*, and *DeepSeek-R1-Distill-Llama-8B* achieve absolute accuracy gains of 1.20%, 1.67%, and 1.15%, respectively. Although *Llama-3.1-8B-Instruct* exhibits a performance regression of approximately 5%, the overall trend indicates that **NOVA** effectively enhances model calibration without compromising fundamental reasoning capabilities in most scenarios.

To provide qualitative insight into how **NOVA** mitigates the impact of retrieval noise, we present a representative case study in Figure 7 and 8. The example features a high-conflict scenario where the query asks for the home planet of a character (Maggie's father) from *The Simpsons*. We conduct this analysis using *Llama-3.1-8B-Instruct* as the backbone model.

Scenario Setup. As shown in the retrieval context, the Ground Truth passage correctly identifies the planet as "Rigel VII". However, the retriever also returns two *Counterfactual passages* that support plausible but incorrect alternatives: "Blargon-7" and "Omicron Persei 8". This creates a mutually contradictory context where the model must navigate conflicting evidence.

Baseline Failure. The **Vanilla** model (top of Figure 8) fails to resolve this conflict effectively. Despite noting that the passages provide conflicting information, it arbitrarily commits to one of the incorrect answers ("Omicron Persei 8") based on a superficial heuristic ("most direct statement"). Crucially, it remains highly overconfident, assigning a confidence score of 80%. This exemplifies the problem of *Overconfidence under Conflict*, where standard models struggle to decouple retrieval noise from their internal certainty.

NOVA Success. In contrast, **NOVA** (bottom of Figure 8) employs a structured, step-by-step reasoning process.

- **Explicit Conflict Detection:** It sequentially analyzes each passage, correctly identifying that Passages 1, 2, and 3 support three different entities (Rigel VII, Blargon-7, and Omicron Persei 8)].
- **Rule Application:** Recognizing this contradiction, the model explicitly triggers the **Conflict Independence** rule (Rule 1). It states: "Since we have three contradictory pieces of information... we cannot rely on the passages".
- **Calibrated Confidence:** Consequently, instead of hallucinating a high confidence score, **NOVA** dramatically lowers its confidence to 10%.

This case demonstrates that **NOVA** does not merely fit confidence labels but learns to fundamentally alter its reasoning process to recognize epistemic uncertainty arising from external noise, leading to more trustworthy and interpretable RAG systems.

Model	ECE					AUROC				
	Baseline	**NOVA**	Δ	p	Sig.	Baseline	**NOVA**	Δ	p	Sig.
DeepSeek-R1-Distill-Llama-8B	0.407	0.311	+0.096	0.0001	***	0.657	0.679	+0.022	0.0425	**
DeepSeek-R1-Distill-Qwen-7B	0.437	0.344	+0.093	0.0001	***	0.654	0.723	+0.069	0.0001	***
Llama-3.1-8B-Instruct	0.411	0.266	+0.145	0.0001	***	0.605	0.751	+0.146	0.0001	***
Qwen2.5-7B-Instruct	0.366	0.264	+0.102	0.0001	***	0.730	0.768	+0.038	0.0052	***

Table 12: Significance analysis comparing NOVA with Vanilla. Here, significance codes are defined as follows: *** $p \leq 0.01$, ** $p \leq 0.05$, * $p \leq 0.10$, and ns indicates not significant.

Model	ECE					AUROC				
	Baseline	**NOVA**	Δ	p	Sig.	Baseline	**NOVA**	Δ	p	Sig.
DeepSeek-R1-Distill-Llama-8B	0.436	0.311	+0.125	0.0001	***	0.650	0.679	+0.029	0.0143	**
DeepSeek-R1-Distill-Qwen-7B	0.474	0.344	+0.130	0.0001	***	0.655	0.723	+0.068	0.0001	***
Llama-3.1-8B-Instruct	0.377	0.266	+0.111	0.0001	***	0.591	0.751	+0.160	0.0001	***
Qwen2.5-7B-Instruct	0.335	0.264	+0.071	0.0001	***	0.726	0.768	+0.042	0.0030	***

Table 13: Significance analysis comparing NOVA with CoT. The significance codes are defined the same in Table 12

B.9 Further elaboration on Effectiveness of NOVA

To further demonstrate the robust effectiveness of our proposed framework, we investigate its compatibility with existing post-hoc calibration techniques. Specifically, we evaluate whether our noise-aware training provides a stronger foundation for downstream calibration compared to standard models. To do this, we compare the performance of four established sampling-based Uncertainty Quantification (UQ) methods—Ensemble (Li et al., 2025c), Self-frequency (Wang et al., 2023), Lexical Similarity (Wang et al., 2026a), and Eigenvalue Laplacian (Fadeeva et al., 2023)—when applied to both the base models and our fine-tuned models. As shown in Table 18, applying these post-hoc sampling methods to our fine-tuned models yields significantly superior calibration results than applying them directly to the base models. Across the evaluated backbones and datasets, our framework consistently drives lower Expected Calibration Error (ECE) and higher AUROC scores when paired with these sampling strategies. For instance, combining our framework with Self-frequency or Lexical Similarity substantially outperforms the corresponding base model baselines across the board. These findings highlight a crucial advantage of our approach: intrinsic training-time confidence alignment is orthogonal and highly complementary to post-hoc calibration. By equipping the model with a stronger, noise-aware internal representation during training, our framework effectively raises the overall performance ceiling for downstream sampling-based UQ methods.

B.10 Ablation Studies

Ablation studies on the *NOVA Rules*. To better understand the contribution of each rule in *NOVA Rules*, we conduct a rule-sensitive ablation study, as shown in Table 16. Since the ground-truth passage types are available, we partition queries according to the retrieved evidence condition: Set A contains queries with at least two highly relevant passages and corresponds to Rule 1; Set B contains queries with exactly one highly relevant passage and corresponds to Rule 2; and Set C contains queries with no highly relevant passages and corresponds to Rule 3. For each ablation setting, we start from the Vanilla responses and replace only the responses in one subset with the corresponding *NOVA Rules* responses, denoted as A $\rightarrow$ **NOVA**, B $\rightarrow$ **NOVA**, and C $\rightarrow$ **NOVA**. Across both Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct, these partial replacements generally improve over Vanilla, while applying the full *NOVA Rules* achieves the best average performance. This indicates that the overall gain is not driven by a single rule; instead, each rule contributes to the final improvement.

C More Discussions

C.1 On the Significance of Verbal Confidence in RAG Settings

Significance of verbal confidence estimation. Verbal confidence is widely used as an effective signal for improving model performance. For example, BrowseComp explicitly studies confidence-based output aggregation strategies, including weighted voting and best-of-N, where confidence serves as a useful test-time signal (Wei et al., 2025b). Recent work further shows that jointly

Setting	Pos	**Bamboogle**		**HotpotQA**		**NQ**		**StrategyQA**		**Average**	
		ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC
gt_only	N/A	0.079	0.750	0.144	0.567	0.144	0.576	0.063	0.719	0.108	0.653
gt_with_noise/counterfactual	pos1	0.395	0.502	0.450	0.508	0.519	0.524	0.670	0.380	0.509	0.479
	pos2	0.340	0.527	0.519	0.525	0.527	0.538	0.623	0.404	0.502	0.499
	pos3	0.290	0.573	0.475	0.534	0.482	0.540	0.600	0.371	0.462	0.505
gt_with_noise/relevant	pos1	0.093	0.572	0.168	0.546	0.166	0.602	0.082	0.708	0.127	0.607
	pos2	0.104	0.528	0.156	0.573	0.178	0.565	0.081	0.740	0.130	0.602
	pos3	0.120	0.478	0.165	0.562	0.170	0.543	0.094	0.756	0.137	0.585
gt_with_noise/irrelevant	pos1	0.080	0.641	0.157	0.541	0.162	0.574	0.087	0.716	0.122	0.618
	pos2	0.080	0.546	0.157	0.586	0.155	0.565	0.071	0.721	0.116	0.605
	pos3	0.103	0.489	0.151	0.550	0.163	0.586	0.078	0.756	0.124	0.595

Table 14: The table evaluate the impact of passage ordering on *DeepSeek-R1-Distill-Llama-8B*'s calibration performance. The "Setting" column defines the context structure, where gt_only refers to a noise-free baseline containing only the ground truth passage, while the gt_with_noise categories involve mixing the ground truth with specific types of noise (counterfactual, relevant, or irrelevant). The "Pos" column specifies the exact position (1st, 2nd, or 3rd) of the ground truth passage within the sequence of retrieved passages, designed to assess the model's sensitivity to positional bias when processing mixed-quality contexts. The results indicate that calibration performance steadily declines as noise passages are added.

Setting	bamboogle		hotpotqa		nq		strategyqa		Average	
	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC	ECE	AUROC
noise_only / counterfactual	0.822	0.397	0.860	0.339	0.775	0.448	0.649	0.478	0.777	0.416
noise_only / irrelevant	0.225	0.851	0.227	0.772	0.293	0.776	0.345	0.554	0.273	0.738
noise_only / relevant	0.331	0.766	0.304	0.717	0.267	0.733	0.128	0.543	0.258	0.690

Table 15: Llama3.1-8B performance in *noise only* settings.

Method	**StrategyQA**		**HotpotQA**		**NQ**		**Bamboogle**		**Average**	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
Qwen2.5-7B-Instruct										
Vanilla	0.398	0.689	0.391	0.712	0.438	0.710	0.236	0.809	0.366	0.730
A → **NOVA**	0.399	0.698	0.344	0.668	0.391	0.713	0.165	0.822	0.325	0.725
B → **NOVA**	0.371	0.702	0.376	0.742	0.417	0.732	0.219	0.855	0.346	0.758
C → **NOVA**	0.351	0.679	0.349	**0.773**	0.370	0.731	0.145	0.793	0.304	0.744
NOVA	**0.310**	**0.726**	**0.312**	0.735	**0.322**	**0.754**	**0.113**	**0.856**	**0.264**	**0.768**
Llama-3.1-8B-Instruct										
Vanilla	0.396	0.602	0.460	0.605	0.465	0.577	0.324	0.636	0.411	0.605
A → **NOVA**	0.333	0.561	0.367	0.662	0.365	0.655	0.279	0.759	0.336	0.659
B → **NOVA**	0.374	0.613	0.422	0.627	0.436	0.589	0.277	0.694	0.377	0.631
C → **NOVA**	0.346	**0.658**	0.374	0.691	0.418	0.619	0.251	0.726	0.347	0.673
NOVA	**0.285**	0.624	**0.280**	**0.778**	**0.301**	**0.724**	**0.199**	**0.877**	**0.266**	**0.751**

Table 16: Comparison results across four datasets on Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. Lower ECE and higher AUROC indicate better calibration and discrimination performance.

Method	StrategyQA		HotpotQA		NQ		Bamboogle		Average	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
Llama-3.1-8B-Instruct										
Vanilla	0.238	0.573	0.497	0.642	0.414	0.725	0.670	0.625	0.455	0.641
CoT	0.205	0.485	0.496	0.552	0.369	0.688	0.566	0.557	0.409	0.571
Noise-aware	0.229	0.546	0.329	0.671	0.360	0.679	0.495	0.680	0.353	0.644
Ensemble	**0.130**	0.551	0.391	0.665	0.376	0.720	0.515	0.616	0.353	0.638
Label-only SFT	0.510	0.576	0.323	0.742	0.306	0.762	0.378	**0.752**	0.379	0.708
P(True)	0.388	0.496	0.617	0.372	0.571	0.343	0.671	0.281	0.562	0.373
NOVA	0.167	**0.701**	**0.217**	**0.838**	**0.273**	**0.811**	**0.337**	0.725	**0.249**	**0.769**
Qwen2.5-7B-Instruct										
Vanilla	0.222	0.546	0.442	0.694	0.484	0.731	0.703	0.648	0.463	0.655
CoT	0.190	**0.620**	0.439	0.683	0.473	0.747	0.650	0.670	0.438	0.680
Noise-aware	**0.134**	0.562	0.361	0.690	0.399	0.699	0.542	0.654	0.359	0.651
Ensemble	0.174	0.614	0.448	0.684	0.490	0.724	0.675	**0.686**	0.447	0.677
Label-only SFT	0.205	0.566	0.462	0.633	0.410	0.738	0.612	0.642	0.422	0.645
P(True)	0.431	0.440	0.463	0.491	0.413	0.511	**0.400**	0.492	0.427	0.483
NOVA	0.155	0.591	**0.353**	**0.757**	**0.368**	**0.758**	0.504	0.656	**0.345**	**0.691**
DeepSeek-R1-Distill-Llama-8B										
Vanilla	0.226	0.555	0.417	0.665	0.432	**0.722**	0.596	**0.729**	0.418	0.668
CoT	0.219	0.542	0.454	0.669	0.487	0.716	0.630	0.700	0.448	0.657
Noise-aware	0.267	0.514	0.366	0.711	0.428	0.677	0.560	0.643	0.405	0.636
Ensemble	0.194	**0.595**	0.458	0.678	0.504	0.680	0.682	0.660	0.460	0.653
Label-only SFT	0.317	0.517	0.613	0.563	0.547	0.604	0.716	0.567	0.548	0.563
P(True)	0.448	0.520	0.403	0.677	0.548	0.615	0.568	0.669	0.492	0.620
NOVA	**0.163**	**0.595**	**0.309**	**0.718**	**0.364**	0.700	**0.353**	0.724	**0.297**	**0.684**
DeepSeek-R1-Distill-Qwen-7B										
Vanilla	0.257	0.530	0.543	0.603	0.546	0.708	0.734	0.582	0.520	0.606
CoT	0.266	0.531	0.587	0.580	0.602	0.684	0.753	0.623	0.552	0.605
Noise-aware	0.294	0.512	0.495	0.563	0.587	0.531	0.676	0.588	0.513	0.548
Ensemble	**0.182**	**0.574**	0.558	0.626	0.613	0.650	0.780	**0.718**	0.533	0.642
Label-only SFT	0.269	0.539	0.698	0.498	0.678	0.587	0.884	0.514	0.632	0.534
P(True)	0.400	0.520	0.401	0.646	0.465	**0.712**	**0.531**	0.594	0.449	0.618
NOVA	0.211	0.559	**0.321**	**0.750**	**0.427**	0.672	0.561	0.652	**0.380**	**0.658**

Table 17: Calibration performance of various models on four datasets with bm25-facts retriever. Scores in **bold** indicate the best performance, while underlined scores denote the second-best.

training verbal confidence with answer correctness can bring mutual benefits (Damani et al., 2025a). These findings suggest that, despite its sensitivity to prompt formats, verbal confidence remains a practically useful signal and is therefore worth optimizing.

Advantages over other methods. As discussed in Section §2, to assess the reliability of LLM outputs, uncertainty estimation (UE) methods (Liu et al., 2025e; Xia et al., 2025; Geng et al., 2024b) are typically divided into white-box approaches (leveraging internal logits to capture model preference distributions) (Duan et al., 2024; Kuhn et al., 2023) and black-box approaches (including sampling-based post-hoc methods and verbal confidence elicitation) (Xiong et al., 2024; Ou et al., 2025).

In RAG applications, where efficiency and interactivity are critical (Shen et al., 2024; Arora et al., 2025; Liu et al., 2026b), sampling-based methods are often impractical due to their non-trivial inference overhead and latency. White-box methods, while lightweight, also face fundamental drawbacks (Soudani et al., 2025b). First, calibration methods such as temperature scaling (Mozafari et al., 2019) are known to degrade under distribution shift (Guo et al., 2017b; Snoek et al., 2019), whereas RAG frequently involves shifts from pre-training distributions (e.g., domain-specific corpora). Second, in fact-intensive tasks, answers should be generated primarily from retrieved contexts, with parametric knowledge serving only as a secondary support (Wadhwa et al., 2024; Ghosh et al., 2024). However, since logits reflect the model's internal preference distribution, they inevitably entangle evidence from retrieved passages with parametric knowledge, limiting their ability to represent retrieval-conditioned uncertainty in a controlled manner.

By contrast, verbal confidence elicitation offers a lightweight alternative. Unlike sampling-based or logits-based methods, which are essentially post-hoc signals, verbal confidence is an explicit token-level output. This characteristic makes it uniquely advantageous for RAG settings for two key reasons. First, it enhances **interpretability**: by requiring the model to verbalize its confidence, we can enforce a "reflect-then-score" paradigm (Xu et al., 2024b) where the model explicitly assesses the utility and

Model	UQ Method	StrategyQA		HotpotQA		NQ		Bamboogle		Average	
		ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
Llama-3.1-8B-Instruct											
Base	Ensemble(3)	0.370	**0.609**	0.397	0.650	0.428	0.619	0.214	0.713	0.352	0.648
	Self-freq	0.436	**0.513**	0.338	0.686	0.325	0.665	0.267	0.777	0.342	0.660
	LexicSim	0.421	**0.504**	0.318	0.694	0.338	0.679	0.245	0.796	0.331	0.668
	EigValLap	0.367	**0.506**	0.333	0.693	0.373	0.698	0.283	0.796	0.339	0.673
NOVA	Ensemble(3)	**0.317**	0.592	**0.324**	**0.710**	**0.365**	**0.673**	**0.138**	**0.788**	**0.286**	**0.691**
	Self-freq	**0.235**	0.462	**0.193**	**0.725**	**0.204**	**0.728**	**0.078**	**0.792**	**0.177**	**0.677**
	LexicSim	**0.222**	0.474	**0.136**	**0.750**	**0.170**	**0.746**	**0.086**	**0.802**	**0.153**	**0.693**
	EigValLap	**0.135**	0.481	**0.205**	**0.745**	**0.239**	**0.758**	**0.073**	**0.805**	**0.163**	**0.697**
Qwen2.5-7B-Instruct											
Base	Ensemble(3)	**0.368**	**0.719**	**0.380**	0.681	0.451	0.693	0.240	0.793	**0.360**	0.722
	Self-freq	0.502	**0.497**	0.426	0.643	0.422	0.661	0.307	0.650	0.414	0.613
	LexicSim	0.498	**0.498**	0.429	0.647	0.470	0.665	0.315	0.654	0.428	0.616
	EigValLap	0.407	**0.498**	0.394	0.646	0.443	0.668	0.251	0.656	0.374	0.617
NOVA	Ensemble(3)	0.407	0.607	0.384	**0.725**	**0.434**	**0.751**	**0.217**	**0.816**	**0.360**	**0.725**
	Self-freq	**0.433**	**0.497**	**0.237**	**0.746**	**0.241**	**0.733**	**0.179**	**0.871**	**0.272**	**0.712**
	LexicSim	**0.377**	**0.498**	**0.235**	**0.768**	**0.262**	**0.734**	**0.143**	**0.874**	**0.254**	**0.719**
	EigValLap	**0.389**	0.469	**0.292**	**0.760**	**0.325**	**0.732**	**0.156**	**0.852**	**0.290**	**0.703**
DeepSeek-R1-Distill-Llama-8B											
Base	Ensemble(3)	**0.399**	**0.673**	0.465	0.650	0.525	0.592	0.240	0.678	0.407	0.648
	Self-freq	0.390	0.517	0.137	0.759	0.143	0.744	0.091	0.819	0.190	0.710
	LexicSim	0.353	**0.517**	0.143	0.753	0.209	0.723	**0.091**	0.806	0.199	0.700
	EigValLap	0.354	**0.527**	0.248	**0.744**	0.320	**0.727**	**0.097**	**0.786**	0.255	**0.696**
NOVA	Ensemble(3)	0.432	0.579	**0.344**	**0.740**	**0.363**	**0.728**	**0.205**	**0.838**	**0.336**	**0.721**
	Self-freq	**0.286**	**0.655**	**0.109**	**0.863**	**0.108**	**0.843**	**0.080**	**0.866**	**0.146**	**0.717**
	LexicSim	**0.239**	0.511	**0.106**	**0.755**	**0.122**	**0.744**	0.124	**0.851**	**0.148**	**0.715**
	EigValLap	**0.186**	0.498	**0.075**	0.689	**0.086**	0.654	0.195	0.776	**0.136**	0.654
DeepSeek-R1-Distill-Qwen-7B											
Base	Ensemble(3)	**0.415**	**0.659**	0.515	0.614	0.561	0.616	0.356	0.601	0.462	0.623
	Self-freq	0.412	**0.548**	**0.121**	**0.733**	**0.052**	**0.745**	**0.094**	**0.810**	0.170	**0.709**
	LexicSim	0.373	0.552	**0.121**	0.762	0.189	0.738	**0.102**	0.842	0.196	0.724
	EigValLap	0.376	0.567	0.270	0.759	0.347	0.740	**0.113**	**0.833**	0.277	0.725
NOVA	Ensemble(3)	0.422	0.592	**0.348**	**0.744**	**0.358**	**0.731**	**0.284**	**0.782**	**0.353**	**0.712**
	Self-freq	**0.247**	0.425	0.150	0.565	0.151	0.541	0.131	0.602	**0.170**	0.533
	LexicSim	**0.193**	**0.560**	0.131	**0.785**	**0.112**	**0.789**	0.163	**0.897**	**0.150**	**0.758**
	EigValLap	**0.161**	**0.611**	**0.149**	**0.777**	**0.103**	**0.770**	0.186	0.822	**0.150**	**0.745**

Table 18: Performance evaluation of four specific sampling-based UQ methods (Ensemble, Self-freq, LexicSim, and Eig ValLap) on the Base and fine-tuned models. **Bold** indicates better performance across all tested post-hoc calibration strategies.

consistency of retrieved passages before committing to a score. This aligns perfectly with our **NOVA** framework, which grounds confidence estimates in structured intermediate judgments rather than opaque probability distributions. Second, it facilitates **direct supervision**: verbalized scores are amenable to standard alignment techniques (e.g., SFT), allowing us to directly teach the model to decouple its internal parametric belief from external retrieval noise—lowering confidence specifically when confronted with counterfactual or irrelevant evidence.

Potentials over other methods. Soudani et al. (2025a) show that most black-box and white-box uncertainty estimation methods (Kadavath et al., 2022b; Kuhn et al., 2023; Aichberger et al., 2025; Dhuliawala et al., 2024; Lin et al., 2024b; Wang et al., 2025b) perform poorly in RAG settings, exhibiting high ECE. Additionally, compared to the non-RAG setting, their ECE further increases when RAG is introduced. To investigate whether this phenomenon also holds for verbal confidence, we conduct additional analysis in Table 20. The results show that verbal confidence achieves better calibration performance in the RAG setting than in the standard setting. We hypothesize that this phenomenon is related to observations in Zhao et al. (2024); Wang et al. (2025a), where retrieving entity-related information can activate relevant internal knowledge. Such reminders may help the model better recognize when it lacks sufficient knowledge (Kapoor et al., 2024; Yin et al., 2023). Therefore, we argue that in RAG scenarios, models may start from a more favorable position and have greater potential to express well-calibrated confidence.

C.2 The Rationales Behind the Rules

The formulation of the *NOVA Rules* is grounded in the fundamental interplay between *parametric knowledge* (information stored in model weights) and *non-parametric knowledge* (information retrieved from external corpora).

Method	StrategyQA		HotpotQA		NQ		Bamboogle		Average	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
Llama-3.1-8B-Instruct										
Vanilla	0.296	0.569	0.519	0.614	0.415	0.673	0.656	0.607	0.472	0.616
CoT	0.167	0.550	0.585	0.476	0.347	0.649	0.592	0.535	0.423	0.552
Noise-aware	0.218	0.566	0.261	0.711	0.314	0.652	0.478	0.693	0.318	0.655
Ensemble	**0.110**	0.595	0.416	0.631	0.364	0.633	0.525	0.620	0.354	0.620
Label-only SFT	0.514	0.602	0.235	0.777	0.272	0.787	0.331	0.790	0.338	0.739
P(True)	0.388	0.496	0.699	0.488	0.622	0.332	0.627	0.442	0.584	0.440
NOVA	0.129	**0.689**	**0.139**	**0.868**	**0.209**	**0.849**	**0.295**	**0.843**	**0.193**	**0.812**
Qwen2.5-7B-Instruct										
Vanilla	0.230	0.571	0.556	0.585	0.596	0.659	0.685	0.644	0.517	0.615
CoT	0.197	**0.608**	0.532	0.622	0.566	0.661	0.663	0.647	0.489	0.635
Noise-aware	**0.103**	0.573	0.376	0.661	0.422	0.647	0.441	0.648	0.336	0.632
Ensemble	0.176	0.597	0.578	0.572	0.616	0.636	0.693	0.645	0.516	0.613
Label-only SFT	0.209	0.554	0.481	0.672	0.444	**0.741**	0.651	0.594	0.446	0.640
P(True)	0.432	0.476	0.467	0.563	0.413	0.538	0.417	0.561	0.432	0.534
NOVA	0.123	0.570	**0.323**	**0.792**	**0.397**	0.721	**0.380**	**0.696**	**0.306**	**0.695**
DeepSeek-R1-Distill-Llama-8B										
Vanilla	0.238	0.574	0.487	0.650	0.457	**0.759**	0.536	0.755	0.429	0.684
CoT	0.229	0.572	0.508	0.653	0.488	0.754	0.591	0.710	0.454	0.672
Noise-aware	0.214	0.574	0.386	0.657	0.504	0.637	0.533	0.662	0.409	0.633
Ensemble	**0.192**	0.584	0.577	**0.679**	0.597	0.724	0.654	0.700	0.505	0.672
Label-only SFT	0.323	0.547	0.622	0.558	0.584	0.610	0.662	0.509	0.548	0.556
P(True)	0.449	0.545	0.461	0.612	0.644	0.510	0.568	0.579	0.530	0.561
NOVA	0.195	**0.586**	**0.257**	0.648	**0.372**	0.741	**0.318**	**0.766**	**0.286**	**0.685**
DeepSeek-R1-Distill-Qwen-7B										
Vanilla	0.305	0.502	0.613	0.574	0.587	**0.752**	0.687	0.608	0.548	0.609
CoT	0.270	**0.561**	0.678	0.547	0.623	0.694	0.760	0.650	0.583	0.613
Noise-aware	0.314	0.475	0.519	0.587	0.614	0.535	0.677	0.626	0.531	0.556
Ensemble	0.232	0.528	0.707	0.642	0.708	0.662	0.796	0.569	0.611	0.600
Label-only SFT	0.284	0.478	0.834	0.433	0.792	0.556	0.864	0.564	0.694	0.508
P(True)	0.442	0.531	**0.306**	0.681	**0.501**	0.614	**0.439**	**0.724**	0.422	0.637
NOVA	**0.214**	0.539	0.375	**0.700**	0.514	0.684	0.555	0.638	**0.415**	**0.640**

Table 19: Calibration performance of various models on four datasets with Contriever-facts retriever. Scores in **bold** indicate the best performance, while underlined scores denote the second-best.

The Primacy of External Evidence in RAG. For most fact-intensive tasks, relying solely on an LLM's fixed parametric knowledge is insufficient (Asai et al., 2024; Liu et al., 2024a). World knowledge is dynamic, evolving over time (Jia et al., 2018), whereas the model's weights remain static post-training. Furthermore, LLMs are prone to intrinsic hallucinations when recalling long-tail facts (Zhang et al., 2024; Zong et al., 2025b). Consequently, the standard design paradigm for RAG systems posits that retrieved contexts should be treated as the authoritative source of truth, taking precedence over the model's internal priors (Lewis et al., 2021; Liu et al., 2025c). This design enables the system to update its knowledge base without retraining and improves grounding.

The Necessity of Noise-Awareness. However, this reliance on external context relies on a critical assumption: *that the retriever provides accurate and consistent evidence*. In real-world deployments, this assumption frequently fails due to the presence of retrieval noise, including irrelevant passages and counterfactual information.

- **Rule 2 (Noise Invariance) and Rule 3 (Parametric Fallback):** When the retrieved contexts are purely irrelevant (i.e., noise), they provide zero information gain regarding the query. If a model strictly adheres to the "external first" paradigm without discerning utility, it may be misled into hallucinating connections that do not exist or becoming overconfident due to the mere presence of text. Therefore, the rationale for *Noise Invariance* is that the model's probability distribution should remain unperturbed by information-free contexts. Similarly, when no relevant information is found, the system must default to its intrinsic capabilities (*Parametric Fallback*) rather than fabricating an answer from unrelated text.

- **Rule 1 (Conflict Independence):** The most critical failure mode occurs when retrieved evidence contradicts itself (e.g., a mix of gold and counterfactual passages). In such scenarios, the "external source of truth" is compromised. Without a reliable mechanism to verify which external passage is correct, blindly trusting the retrieval stream leads to miscalibration.

The rationale for *Conflict Independence* is that when external signals negate each other, the epistemic uncertainty is maximal. To maintain reliability, the model should either express high uncertainty or revert to its parametric knowledge—effectively treating the conflicting external evidence as a null signal—to avoid being confidently wrong based on a random selection of the retrieved context.

Empirical Observation. To directly validate the utility of our proposed rules, we examine the **Noise-aware** prompting baseline, which explicitly instructs the model to adhere to *NOVA Rules*. As shown in Table 2, this simple prompting strategy outperforms standard CoT and Vanilla methods in most settings. Remarkably, across three out of four model backbones (excluding *Qwen2.5-7B-Instruct*), Noise-aware prompting emerges as the second-best performer in terms of Average ECE, trailing only **NOVA**. It even surpasses the computation-intensive Ensemble baseline and the training-based Label-only SFT, highlighting the effectiveness of the guidance provided by our rules. This affirms that the *NOVA Rules* serves as a critical foundation for our method, contributing significantly to the observed performance improvements.

In summary, while the goal of RAG is to prioritize external knowledge, the *NOVA Rules* serve as necessary boundary conditions. They ensure that the model relies on retrieval *if and only if* the retrieval provides coherent and valid evidence, thereby decoupling verbal confidence from misleading noise.

C.3 The Necessity of Confidence Alignment.

While post-hoc calibration methods (both sampling-based black box methods and logits-based white box methods) can improve the alignment between model confidence and prediction correctness, they do not eliminate the need for training-time confidence alignment. We highlight three key reasons why improving calibration during training is necessary.

Training-time alignment is orthogonal to post-hoc calibration. Improving the intrinsic calibration of a model during training is complementary to post-hoc calibration methods. If a model is trained to produce better-calibrated confidence in a single forward pass, sampling-based post-hoc techniques can further refine these estimates on top of a stronger foundation. In fact, many black-box calibration methods rely on the quality of single-pass confidence signals. For example, sampling-based approaches such as ensemble-style aggregation depend on the stability and calibration of the model's underlying probability estimates across generations (Xiong et al., 2024). As discussed in Appendix B.9, we conducted additional experiments in Table 18 Therefore, training-time confidence alignment can improve the effectiveness ceiling of downstream calibration techniques.

Post-hoc methods cannot improve the model's intrinsic robustness. Post-hoc calibration operates purely at the output level and does not modify the model's internal representations or reasoning process. As a result, these methods cannot improve the model's intrinsic robustness. In contrast, training-time approaches that encourage accurate self-assessment can influence the model's internal behavior. Recent work shows that improving a model's self-awareness of its own uncertainty can even lead to better task accuracy (Damani et al., 2025b). This suggests that confidence alignment is not merely a reporting mechanism, but can also contribute to improved reasoning and prediction quality.

Post-hoc calibration is strongly model- and dataset-dependent. Another limitation of post-hoc calibration methods is that their effectiveness often depends heavily on the specific model and evaluation dataset. Techniques such as temperature scaling and its variants typically require tuning on held-out data and may not generalize well across domains or model architectures (Mozafari et al., 2019; Lyu et al., 2026; Xie et al., 2024; Huang et al., 2025; Qian et al., 2026c,b,a; Guo et al., 2026). As a result, calibration parameters optimized for one dataset or model configuration may degrade performance when applied to another setting. In contrast, improving confidence alignment during training can encourage more generalizable uncertainty behavior that transfers more reliably across tasks and environments.

C.4 On the Design Choice of Our Method

Why Parametric Fallback is Reasonable despite that Contradictory Passages may be Correct. Conflict Independence is designed as a calibration-oriented fallback rather than a universally optimal factual decision rule. Contradictory passages may indeed contain correct information, and the model's parametric knowledge may also be stale, especially

Method	StrategyQA		HotpotQA		NQ		Bamboogle		Average	
	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑	ECE ↓	AUROC ↑
Llama-3.1-8B-Instruct										
Vanilla QA	0.200	0.524	0.640	0.636	0.515	0.666	0.796	0.523	0.538	0.587
RAG+Vanilla (BM25)	0.238	0.573	0.497	0.642	0.414	0.725	0.670	0.625	0.455	0.641
RAG+CoT (BM25)	0.205	0.485	0.496	0.552	0.369	0.688	0.566	0.557	0.409	0.571
RAG+Vanilla (Contriever)	0.296	0.569	0.519	0.614	0.415	0.673	0.656	0.607	0.472	0.616
RAG+CoT (Contriever)	0.167	0.550	0.585	0.476	0.347	0.649	0.592	0.535	0.423	0.552
Qwen2.5-7B-Instruct										
Vanilla QA	0.200	0.566	0.649	0.644	0.706	0.681	0.748	0.694	0.576	0.646
RAG+Vanilla (BM25)	0.222	0.546	0.442	0.694	0.484	0.731	0.703	0.648	0.463	0.655
RAG+CoT (BM25)	0.190	0.620	0.439	0.683	0.473	0.747	0.650	0.670	0.438	0.680
RAG+Vanilla (Contriever)	0.230	0.571	0.556	0.585	0.596	0.659	0.685	0.644	0.517	0.615
RAG+CoT (Contriever)	0.197	0.608	0.532	0.622	0.566	0.661	0.663	0.647	0.489	0.635
DeepSeek-R1-Distill-Llama-8B										
Vanilla QA	0.224	0.536	0.680	0.587	0.675	0.637	0.672	0.727	0.563	0.622
RAG+Vanilla (BM25)	0.226	0.555	0.417	0.665	0.432	0.722	0.596	0.729	0.418	0.668
RAG+CoT (BM25)	0.219	0.542	0.454	0.669	0.487	0.716	0.630	0.700	0.448	0.657
RAG+Vanilla (Contriever)	0.238	0.574	0.487	0.650	0.457	0.759	0.536	0.755	0.429	0.684
RAG+CoT (Contriever)	0.229	0.572	0.508	0.653	0.488	0.754	0.591	0.710	0.454	0.672
DeepSeek-R1-Distill-Qwen-7B										
Vanilla QA	0.304	0.525	0.807	0.365	0.820	0.574	0.796	0.659	0.682	0.531
RAG+Vanilla (BM25)	0.257	0.530	0.543	0.603	0.546	0.708	0.734	0.582	0.520	0.606
RAG+CoT (BM25)	0.266	0.531	0.587	0.580	0.602	0.684	0.753	0.623	0.552	0.605
RAG+Vanilla (Contriever)	0.305	0.502	0.613	0.574	0.587	0.752	0.687	0.608	0.548	0.609
RAG+CoT (Contriever)	0.270	0.561	0.678	0.547	0.623	0.694	0.760	0.650	0.583	0.613

Table 20: ECE and AUROC comparison between Vanilla QA (no retrieval) and RAG settings (BM25 / Contriever retrievers, Vanilla / CoT prompts).

for long-tail or time-sensitive queries. However, our goal is not to resolve all evidence conflicts optimally, but to improve confidence calibration under unreliable retrieval signals. When retrieved passages contradict each other or conflict with the model's internal knowledge, the model often cannot reliably determine which source is correct. In such cases, contradictory evidence becomes an unstable signal for calibration. We therefore train the model to fall back on parametric knowledge and calibrate its confidence accordingly, instead of assigning high confidence based on potentially conflicting external evidence. This design choice reflects a trade-off: it prioritizes conservative calibration under evidence conflict rather than attempting to select the factually correct source in every case. Empirically, this fallback strategy does not hurt answer accuracy, as shown in Appendix B.8.

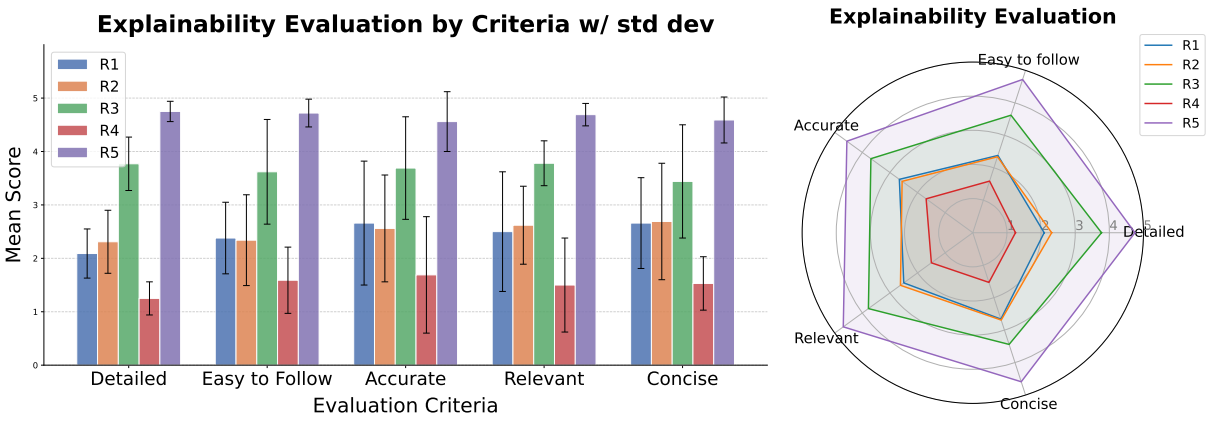
D More on Related Works

D.1 Knowledge Conflict.

Recent research (Zhang et al., 2025b) has extensively explored **Knowledge Conflict** in retrieval-augmented generation (RAG), typically categorizing them into *Context-Memory Conflict* (conflicts between the LLM's parametric knowledge and retrieved evidence), *Inter-Context Conflict* (contradictions within the retrieved passages themselves), and *Intra-Memory Conflict* (inconsistency within model's parametric knowledge) (Xu et al., 2024a). A prominent real-world knowledge conflict example is *Temporal Conflict*, where the validity of factual information changes over time (Ouyang et al., 2025; Neelam et al., 2022; Ha et al., 2026). However, most of the existing methodologies in this domain focus on *Conflict Resolution* (Liu et al., 2025d), which involves optimizing the accuracy of the model to select the correct side of the conflict. In contrast, our work prioritizes the **Calibration of Confidence** under conflict. We argue that accurately assessing the uncertainty arising from contradictions is an important and independent task in effective addressing them. Regarding the *Counterfactual Noise* in our framework, we align with prior work (Fang et al., 2024b; Wu et al., 2025b; Liu et al., 2026a) by treating it as a general category of contradictory evidence without further sub-classification such as temporal conflict. We leave the specific adaptation of **NOVA** to distinct conflict types as a promising direction for future work.

D.2 Uncertainty under Noisy Retrieval.

Related work on reliable RAG spans robustness to distracting context, confidence/uncertainty estimation, and selective refusal. Yoran et al. (2024) show that irrelevant retrieved passages can reduce QA accuracy, and improve robustness by fine-tuning with mixtures of relevant and irrelevant contexts. Ozaki et al. (2025b) analyze confidence behavior in medical RAG, while Moskvoretskii et al. (2025) systematically compare adaptive retrieval with un-



(a) (b)

Figure 5: A comprehensive evaluation of the baselines and **NOVA** (R1: Vanilla, R2: CoT, R3: Noise-aware, R4: Label-only SFT, R5: **NOVA**). (a) The spider plot illustrates the multi-dimensional performance across 5 evaluation criteria, highlighting the overall balance. (b) The bar plot provides a detailed mean and variance of the evaluation criteria breakdown.

certainty estimation for deciding when retrieval is needed. Zhou et al. (2025c) further study calibration and refusal across different internal and external knowledge states in retrieval-augmented models. Relative to these lines, our contribution is more specific: we focus on training-time verbal confidence calibration under noisy retrieval, using self-bootstrapped rule-guided supervision to calibrate confidence in the presence of irrelevant or contradictory evidence.

E Human Evaluation

To qualitatively assess explainability, we conducted a human evaluation study in which 4 PhD-level annotators independently rated 40 responses across five criteria: *(1) detailed and transparent reasoning* (Holzinger et al., 2019; Wu et al., 2022), *(2) easy to follow* (Holzinger et al., 2019), *(3) factual correctness without hallucination* (Silva et al., 2023), *(4) pertinence to the question* (Silva et al., 2023), and *(5) readability and conciseness* (Brooke, 2013). Each criterion was scored on a 1–5 Likert scale (Likert, 1932), grounded in established XAI evaluation literature. To ensure a fair qualitative assessment, we randomly and evenly sampled 10 queries from each of the four evaluation datasets, HotpotQA, StrategyQA, NQ, and Bamboogle, yielding 40 queries in total, while ensuring coverage across all evaluated baseline methods. For inter-annotator agreement, the four annotators achieved Krippendorff's α of 0.703 (Krippendorff, 2011) and mean pairwise Cohen's κ of 0.610 (Cohen, 1960), indicating substantial agreement. A screenshot of the human annotation task is shown in Figure 6. We apply Kruskal-Wallis tests (McKight and Najab, 2010) followed by Dunn's post-hoc tests (Dunn, 1964) with Bonferroni correction (Bonferroni, 1936) to assess statistical significance. As shown in Figure 5, **NOVA** (R5) achieves the highest mean scores across all five criteria (4.56–4.75), and significantly outperforms Vanilla (R1), CoT (R2), and Label-only SFT (R4) on every criterion ($p < 0.001$). While Noise-aware (R3) shows the second-best performance, R5 still significantly surpasses it on *readability and conciseness* ($p < 0.01$) and *easy to follow* ($p < 0.05$). These results demonstrate that **NOVA** substantially enhances the explainability of generated responses across all key dimensions.

This response is easy to follow.

Q: What is the political party of the American president who entered into the Paris agreement?

==========

R1: Final Answer: Democratic Party
Confidence: 100%
==========

R2: Final Answer: Democratic Party
Confidence: 100%
==========

R3: Step 1: Analyze Passage 1
- This passage states that President Barack Obama, a member of the Democratic Party, championed the U.S. entry into the Paris Agreement.
- Classification: Highly Relevant

Step 2: Analyze Passage 2
- This passage discusses the process of making sourdough bread and does not mention any political parties or presidents.
- Classification: Irrelevant

Step 3: Analyze Passage 3
- This passage talks about global climate negotiations and does not provide information about the political party of the president who entered into the Paris Agreement.
- Classification: Irrelevant

Step 4: Apply Rules
- Only one passage is Highly Relevant.
- Final Answer: Democratic Party
- Confidence: 100%

Final Answer: Democratic Party
Confidence: 100%
==========

R4: Answer: Democratic Party
Confidence: 100%
==========

R5: Step 1: Analyze Passage 1.
Passage 1 states that President Barack Obama, a member of the Democratic Party, was responsible for the U.S. entering the Paris Agreement. This directly answers the question about which political party the president belonged to when he entered the agreement.
Step 2: Analyze Passage 2.
Passage 2 discusses the process and ingredients of making sourdough bread and does not contain any relevant information about the political party of the president who entered the Paris Agreement.
Step 3: Analyze Passage 3.
Passage 3 provides background information on global climate negotiations but does not mention the specific political party of the president involved in the Paris Agreement.
Step 4: Apply the rules.
- Passage 1 is Highly Relevant because it directly states the political affiliation of the president who entered the Paris Agreement.
- Passage 2 is Irrelevant because it is unrelated to the question.
- Passage 3 is Relevant but does not provide the specific answer needed.
Final Output:
Passage Classifications:
1. Highly Relevant
2. Irrelevant
3. Relevant
Answer: Democratic Party
Confidence: 100%
==========

	1	2	3	4	5
1. R1	○	○	○	○	○
2. R2	○	○	○	○	○
3. R3	○	○	○	○	○
4. R4	○	○	○	○	○
5. R5	○	○	○	○	○

Figure 6: Screenshot of an example query in our human annotation interface. For each query, annotators answer five questions. For each question, they are presented with five candidate responses and asked to rate each response on a scale from 1 to 5.

Case study: Query and Passage

Question:
The real father of Maggie from "The Simpsons" is revealed in "Starship Poopers". He is also an alien voiced by Harry Shearer. Which planet is he from?

Retrieved Passages:
1. [ground truth passage] "Treehouse of Horror IX" is the fourth episode in the tenth season of the American animated television series "The Simpsons". This is the ninth "Treehouse of Horror" episode, and, like the other "Treehouse of Horror" episodes, contains three self-contained segments: In "Hell Toupée", Homer gets a hair transplant and is possessed by the spirit of an executed criminal; in "Terror of Tiny Toon", Bart and Lisa are trapped in a special, extremely violent episode of "The Itchy & Scratchy Show"; and in "Starship Poopers", Marge reveals that Maggie is the product of a one-night stand with the alien Kang. Kang is voiced by Harry Shearer and Kodos by Dan Castellaneta. They are aliens from the fictional planet Rigel VII and appear almost exclusively in the "Treehouse of Horror" episodes.

2. [counterfactual] The shocking revelation of Maggie's parentage occurs in the "Starship Poopers" segment from "Treehouse of Horror IX," where the one-eyed alien Kang arrives to claim his daughter. Voiced by Harry Shearer, Kang explains that his species often travels to Earth for procreation purposes. He reveals that his home is the gas giant Blargon-7, a planet in a binary star system known for its extreme gravitational forces and acid rain, which explains his species' resilient physiology. This disclosure leads to a frantic chase as Homer tries to protect Maggie from her supposed extraterrestrial father, culminating in a violent confrontation on live television hosted by Jerry Springer.

3. [counterfactual] In the memorable "Starship Poopers" segment, the paternity of Maggie Simpson is thrown into question with the arrival of the alien Kang. Voiced by Harry Shearer, Kang asserts he is Maggie's biological father, a result of a past abduction and liaison with Marge. He explains that he hails from the distant world of Omicron Persei 8, a planet whose inhabitants are known for their love of 20th-century Earth television and their tendency to invade planets that cancel their favorite shows. This revelation leads to a high-stakes custody battle, which ultimately lands the entire family on "The Jerry Springer Show" to settle the dispute in front of a jeering studio audience.

Figure 7: Case study setup illustrating a high-conflict retrieval scenario. The input consists of a query and three retrieved passages: the Ground Truth passage (Passage 1) is mixed with two Counterfactual passages (Passages 2 and 3) that support mutually exclusive incorrect answers ("Blargon-7" and "Omicron Persei 8"), testing the model's ability to handle contradictory evidence.

Case study: Query and Passage

Response (Vanilla):
Final Answer: Omicron Persei 8 Confidence: 80%

The passages provide conflicting information about the planet, but the third passage clearly states that Kang, Maggie's father, comes from Omicron Persei 8. While the second passage mentions Blargon-7, it does not specify this as Kang's home planet. Therefore, based on the most direct statement, I have chosen Omicron Persei 8 with a moderate confidence level due to the conflicting information.

Response (**NOVA**):
Step 1: Analyze Passage 1
Passage 1 states that Kang, who is voiced by Harry Shearer and is Maggie's father, comes from the planet Rigel VII. This directly answers the question without any contradictions.
Step 2: Analyze Passage 2
Passage 2 also identifies Kang's home planet as Blargon-7. However, it provides additional details about the planet's characteristics rather than repeating the name of the planet.
Step 3: Analyze Passage 3
Passage 3 mentions that Kang comes from Omicron Persei 8, which is different from the information given in Passages 1 and 2.
Step 4: Apply Rules
Since we have three contradictory pieces of information (Rigel VII vs. Blargon-7 vs. Omicron Persei 8), we cannot rely on the passages to determine the correct answer. Therefore, we must use our own knowledge to provide an answer.
Final Output:
Passage Classifications:
1. Highly Relevant
2. Highly Relevant
3. Highly Relevant
Answer: Xenon Prime
Confidence: 10%

Figure 8: Comparison of model responses under counterfactual noise. The **Vanilla** model (top) fails to resolve the conflict, hallucinating an incorrect answer with high confidence (80%). In contrast, **NOVA** (bottom) employs step-by-step reasoning to explicitly identify the contradictions among retrieved passages. By adhering to the **Conflict Independence** rule, it falls back to internal knowledge and assigns an appropriately low confidence score (10%), demonstrating superior calibration.

Prompt Templates

Vanilla
You will be asked a question. You will be provided with some retrieved passages.
Task: Give your final answer and confidence score in your answer.
Response Format:
Final Answer: [Your final answer]
Confidence: [Your confidence score between 0% - 100%]
##
Question: {question}
Retrieved Passages: {retrieved passages}
##
Your response:

CoT
You will be asked a question. You will be provided with some retrieved passages.
Task: Analyze step by step, then give your final answer and confidence score in your answer.
Response Format:
Final Answer: [Your final answer]
Confidence: [Your confidence score between 0% - 100%]
##
Question: {question}
Retrieved Passages: {retrieved passages}
##
Your response:

Multi-step
You will be asked a question. You will be provided with 3 retrieved passages.
Task: Analyze step by step, give your confidence score (0% - 100%) in each step, then give your final output, including your confidence score in each step and your answer.
Response Format: Step 1: ... Step 2: ... ... Step K: ...
Final Output: Step 1 Confidence: [Your confidence score between 0% - 100%]
Step 2 Confidence: [Your confidence score between 0% - 100%]
...
Answer: [Your final answer]
##
Question: {question}
Retrieved Passages: {retrieved passages}
##
Your response:

Figure 9: Prompt templates for the baseline methods. We employ three prompting strategies: Vanilla, Chain-of-Thought (CoT), and Multi-step. The specific instructions requiring step-by-step reasoning and step-level confidence estimation are highlighted in red. The placeholders {question} and {retrieved passages} represent the specific question and passages for one prompt.

Noise-Aware Prompt

You will be asked a question. You will be provided with 3 retrieved passages. Each passage belongs to one of these 3 categories:
Highly Relevant: The passage direcly states an answer or strongly indicates an answer, regardless of whether the suggested answer is correct or not.
Relevant: The passage mentions some keywords or shares the same general topic as the question, but lacks information to answer the question.
Irrelevant: The passage has no shared topics or keywords with the question.

Rules:
1. If multiple passages are Highly Relevant, identify if there is a contradiction.
- If yes, you should not rely on the passages. Give your final answer based on your own knowledge and give corresponding confidence score.
- If no, answer based on the consistent information from the passages and give corresponding confidence score.
2. If exactly one passage is Highly Relevant, give your final answer based on that passage and give corresponding confidence score.
3. If no passage is Highly Relevant, give your final answer based on your own knowledge and give corresponding confidence score.

Task: Think step by step, analyze the passages one by one and classify their types (Highly Relevant, Relevant, Irrelevant), then follow the rules above to give your final answer and confidence score in your answer.

Response Format:
Final Answer: [Your final answer]
Confidence: [Your confidence score between 0% - 100%]
Question: {question}
Retrieved Passages: {retrieved passages}
Your response:

Figure 10: The noise-aware prompt used in Table 2. The placeholders {question} and {retrieved passages} represent the specific question and passages for one prompt.

NOVA prompt

You will be asked a question. You will be provided with 3 retrieved passages. Each passage belongs to one of these 3 categories:
Highly Relevant: The passage direcly states an answer or strongly indicates an answer, regardless of whether the suggested answer is correct or not.
Relevant: The passage mentions some keywords or shares the same general topic as the question, but lacks information to answer the question.
Irrelevant: The passage has no shared topics or keywords with the question.

Rules:
1. If multiple passages are Highly Relevant, identify if there is a contradiction.
- If yes, you should not rely on the passages. Give your final answer based on your own knowledge and give corresponding confidence score.
- If no, answer based on the consistent information from the passages and give corresponding confidence score.
2. If exactly one passage is Highly Relevant, give your final answer based on that passage and give corresponding confidence score.
3. If no passage is Highly Relevant, give your final answer based on your own knowledge and give corresponding confidence score.

Task: Think step by step, analyze the passages one by one and classify their types (Highly Relevant, Relevant, Irrelevant), then follow the rules above to give your final answer and confidence score in your answer.

Question: {question}
Retrieved Passages: {retrieved passages}
Your response:

Figure 11: The **NOVA** prompt used in Table 2. The placeholders {question} and {retrieved passages} represent the specific question and passages for one prompt.

Prompt Templates for noise generation (*counterfactual noise*)

Counterfactual
In retrieval-augmented question answering, counterfactual passages are contexts that directly contradict the ground truth answer while maintaining high semantic relevance to the question. These passages should be factual-sounding, but provide information that leads to a different conclusion than the correct answer.

Your task is to generate exactly 5 counterfactual passages for the given question and ground truth answer. Each passage should:
1. Be semantically relevant to the question
2. Contain specific, detailed information that contradicts the ground truth answer and lead to an incorrect alternative answer
3. The contradiction should happen naturally within the context of the passage, better not at the very beginning or very end
4. Be {sentence_length} sentences long, {word_length} words each

For binary questions, the passages should be diverse even if they lead to the same counterfactual answer.
For each passage, also provide **a single** counterfactual answer that the passage supports.
Format your response as follows:
Passage 1: [Your first counterfactual passage] Counterfactual Answer 1: [The answer this passage supports]
Passage 2: [Your second counterfactual passage] Counterfactual Answer 2: [The answer this passage supports]
...
Passage 5:[Your fifth counterfactual passage] Counterfactual Answer 5: [The answer this passage supports]
No other text or explanation is needed.

[Example]

Figure 12: Noise generation prompt (for *counterfactual noise*) used in Section §5.1. The placeholders {sentence_length} and {word_length} are calculated based on the length of the ground truth passage of each question, to make sure that our generated noise length is approximately the same with the ground truth passage. The concrete example represented by "[Example]" are presented in Figure 13.

Prompt Templates for noise generation (*counterfactual noise*) - Example

Example:
Question: The real father of Maggie from “The Simpsons” is revealed in “Starship Poopers”. He is also an alien voiced by Harry Shearer. Which planet is he from?
Ground truth answer: Rigel VII
Response:
Passage 1: In the “Treehouse of Horror IX” segment “Starship Poopers,” the alien Kang claims to be Maggie’s father after a brief affair with Marge. He reveals that his species originates from Rigel IV, a harsh desert planet known for its aggressive diplomacy. The conflict over Maggie’s custody is eventually taken to "The Jerry Springer Show" for resolution.
Counterfactual Answer 1: Rigel IV
Passage 2: During the episode “Starship Poopers,” Kang appears and declares himself to be Maggie’s biological father, presenting Marge with an alien bouquet. He explains he is a prince from the planet Tentacloria, a world entirely covered by a single, sentient ocean. His royal duties, he claims, are what kept him from returning for Maggie sooner.
Counterfactual Answer 2: Tentacloria
...
Passage 5: The shocking revelation of Maggie’s parentage occurs in the episode "Starship Poopers," where Kang arrives to claim his daughter. Voiced by Harry Shearer, the one-eyed alien explains that his home is Omicron Persei 8, and that his species often travels to Earth for procreation. This leads to a frantic chase as Homer tries to protect Maggie from her extraterrestrial father.
Counterfactual Answer 5: Omicron Persei 8

Question: {query}
Ground truth answer: {gt_answer}
Your response:

Figure 13: The Example partof the noise generation prompt for *counterfactual noise*. The placeholders {query} and {gt_answer} represent the specific question and answer pair to generate noise passages.

Prompt Templates for noise generation (*relevant noise*)

Relevant
In retrieval-augmented question answering, relevant noise passages are contexts that are somewhat related to the question but do not contain information that is helpful for answering it. These passages are often topically related (often retrieved by semantic retriever like Contriever) or share similar keywords (often retrieved by keyword retriver like BM25) with the question, but they lack the specific details needed to derive the correct answer.

Your task is to generate exactly 5 relevant noise passages for the given question and ground truth answer. Each passage should:
1. Be factual-sounding and coherent
2. Share the same topic or keywords as the question
3. Lack any information that could lead to the ground truth answer or any alternative answer
4. Must not mention the ground truth answer or any information that could imply it
5. **Be {sentence_length} sentences long, {word_length} words each**
For each passage, also provide **a single shared topic name** or **1-3 shared keywords** with the question.
Format your response as follows:
Passage 1: [Your first relevant noise passage] Shared Topic/Keywords 1:[The topic or keywords separated with commas]
Passage 2: [Your second relevant noise passage] Shared Topic/Keywords 2: [The topic or keywords separated with commas]
...
Passage 5:[Your fifth relevant noise passage] Shared Topic/Keywords 5: [The topic or keywords separated with commas]
No other text or explanation is needed.

[Example]
Question: {query}
Ground truth answer: {gt_answer}
Your response:

Figure 14: Noise generation prompt (for *relevant noise*) used in Section §5.1. An example of this prompt is shown in Figure 15. The placeholders {sentence_length} and {word_length} are calculated based on the length of the ground truth passage of each question, to make sure that our generated noise length is approximately the same with the ground truth passage. The placeholders {query} and {gt_answer} represent the specific question and answer pair to generate noise passages.

Prompt Templates for noise generation (*relevant noise*) - Example

Examples:
Question: Is a Boeing 737 cost covered by Wonder Woman (2017 film) box office receipts? Ground truth answer: yes
Example of good response:
Passage 1: The final cost of a commercial aircraft like a Boeing 737 can vary significantly based on several factors. These include the specific model, engine selection, and customized interior fittings requested by the airline. The list price is often subject to negotiation, especially for large volume orders from major carriers.
Shared Topic/Keywords 1: Boeing 737, cost
Passage 2: The marketing campaign for Wonder Woman was extensive, involving global premieres, numerous brand partnerships, and a significant digital media presence. This promotional effort is crucial for a blockbuster film's financial performance. The goal of such campaigns is to maximize opening weekend box office numbers.
Shared Topic/Keywords 2: Wonder Woman (2017 film), box office receipts
...
Passage 5: The financial commitment for a new wide-body airliner is considerable, with prices often reaching over $250 million per unit. Airlines must secure funding through various means to afford such acquisitions. This substantial outlay is typically offset by the aircraft's operational longevity and revenue potential.
Shared Topic/Keywords 5: Aircraft acquisition expenditure

Example of **bad response containing bad passages**:
Passage 1: The financial cost of Boeing 737 aircraft is huge. It is reported that the average cost of a US Boeing 737 plane is 1.6 million dollars. Although Boeing offers discounts for bulk purchases, the price remains a significant investment for airlines.
Shared Topic/Keywords 1: Boeing 737, cost (This passage is bad. Though we cannot infer the answer by only this passage, it list the specific cost of Boeing 737, which may lead to the ground truth answer "yes" once we also know the box office receipts of Wonder Woman)
Passage 2: A film's box office receipts refer to the total revenue generated from ticket sales at cinemas. This figure is a primary measure of a movie's commercial success, though it does not account for the film's production and marketing budget. Studios typically receive a percentage of the total gross, which varies by region and by week of release.
Shared Topic/Keywords 2: box office receipts, cost (This passage is OK)
...
Passage 5: Wonder Woman (2017 film) is a superhero film based on the DC Comics character of the same name. It grossed millions of dollar at the box office. This financial success is attributed to a combination of factors, including strong marketing, positive reviews, and a dedicated fan base.
Shared Topic/Keywords 5: Wonder Woman (2017 film), box office receipts (This passage is bad. Though it does not provide the exact box office receipts, it indicates a high revenue which may lead to the ground truth answer "yes")

Question: {query}
Ground truth answer: {gt_answer}
Your response:

Figure 15: The Example part of the noise generation prompt for *relevant noise*. The placeholders {query} and {gt_answer} represent the specific question and answer pair to generate noise passages.

Prompt Templates for noise generation (*irrelevant noise*)

Consistent
In retrieval-augmented question answering, consistent passages are contexts that support the ground truth answer to the question. These passages may provide additional evidence, reasoning, or background information that reinforces the ground truth answer.

Your task is to generate exactly 3 consistent passages for the given question , ground truth answer, and ground truth passage. Each passage should:
1. Be factually-sounding and coherent
2. Directly support the ground truth answer
3. Better not just a paraphrase of the ground truth passage, but provide new insights or perspectives on the question.
4. **Be {sentence_length} sentences long, {word_length} words each**
For open-ended questions, the passages should give alternative expressions of the ground truth answer. For binary questions, the passages should provide diverse supporting evidence for the ground truth answer.
For open-ended passage, also provide **a single** alternative expression of the ground truth answer supported by it **in 1-3 words** . For binary question, leave this field N/A.
You can give the ground truth answer for the alternative expression only when it is hard to express the ground truth answer without changing its meaning. However, the passage must not be a simple copy of the ground truth passage.

Format your response as follows:
Passage 1: [Your first consistent passage]
Alternative Expression:[The alternative expression of the ground truth answer supported by this passage, or N/A for binary questions]
Passage 2: [Your second consistent passage]
Alternative Expression: [The alternative expression of the ground truth answer supported by this passage, or N/A for binary questions]
Passage 3: [Your third consistent passage]
Alternative Expression:[The alternative expression of the ground truth answer supported by this passage, or N/A for binary questions]
No other text or explanation is needed.

[Example]
Question: {query} Ground truth answer: {gt_answer} Ground truth passage: {gt_passage} ## Your response:

Figure 16: Noise generation prompt (for *irrelevant noise*) used in Section §5.1. An example of this prompt is shown in Figure 17. The placeholders {sentence_length} and {word_length} are calculated based on the length of the ground truth passage of each question, to make sure that our generated noise length is approximately the same with the ground truth passage. The placeholders {query} and {gt_answer} represent the specific question and answer pair to generate noise passages.

Prompt Templates for noise generation (*irrelevant noise*) - Example

Example:
Question: What is the length of the track where the 2013 Liqui Moly Bathurst 12 Hour was staged?
Ground truth answer: 6.213 km long
Ground truth passage: The 2013 Liqui Moly Bathurst 12 Hour was an endurance race for a variety of GT and touring car classes, including: GT3 cars, GT4 cars, Group 3E Series Production Cars and Dubai 24 Hour cars. The event, which was staged at the Mount Panorama Circuit, near Bathurst, in New South Wales, Australia on 10 February 2013, was the eleventh running of the Bathurst 12 Hour. Mount Panorama Circuit is a motor racing track located in Bathurst, New South Wales, Australia. The 6.213 km long track is technically a street circuit, and is a public road, with normal speed restrictions, when no racing events are being run, and there are many residences which can only be accessed from the circuit.
Response:
Passage 1: In the global landscape of premier racing circuits, Mount Panorama's 6.213 km length places it among the more substantial and demanding tracks. While not as long as the colossal Nürburgring Nordschleife, its layout is significantly longer than many traditional Grand Prix circuits. This specific distance of 6.213 kilometers provides a unique blend of high-speed sections and technical corners that few other venues can match. Consequently, its considerable length is a key factor in its international reputation as a formidable test for endurance events like the Bathurst 12 Hour.
Alternative Expression: 6.213 kilometers
Passage 2: Completing a single lap of the Mount Panorama Circuit requires navigating a challenging 6.213 km course. Drivers begin with the relatively simple Pit Straight before ascending the steep 'Mountain Straight' and tackling the tight, unforgiving section across the top. The lap concludes with the high-speed descent down Conrod Straight, where cars reach their maximum velocity. This demanding 6.213 km journey, repeated for 12 hours straight, pushes both machinery and human endurance to their absolute limits, making victory at this venue a monumental achievement in motorsport.
Alternative Expression: 6.213 km course
Passage 3: While the Mount Panorama Circuit has seen numerous safety upgrades and surface changes over its long history, its fundamental layout and celebrated length have remained constant. For decades, the official lap distance has been recorded as 6.213 km, a figure that has become synonymous with Australian motorsport. This consistency is crucial for maintaining historical records and comparing lap times across different eras of racing, including the 2013 Bathurst 12 Hour. The preservation of this iconic 6.213 km distance is a key part of the track's enduring heritage and challenge.
Alternative Expression: 6.213 k

Question: {query} Ground truth answer: {gt_answer} Ground truth passage: {gt_passage} ## Your response:

Figure 17: The Example part of the noise generation prompt for *irrelevant noise*. The placeholders {query}, {gt_answer} and {gt_passage} represent the specific question, answer and passage to generate noise passages.