ERA: Evidence-based Reliability Alignment for Honest Retrieval-Augmented Generation

Anonymous ACL submission

Abstract

Retrieval-Augmented Generation (RAG) mitigates LLM hallucinations but introduces knowledge conflicts between parametric memory and retrieved context. Existing alignment methods conflate epistemic and aleatoric uncertainty, forcing models to memorize superficial refusal heuristics rather than learning a principled strategy for conflict resolution. We propose ERA, a framework that disentangles these uncertainties via dual evidential heads capturing parametric and contextual evidence as Dirichlet distributions. These are fused into a conflict score that activates only when both sources are confident yet contradictory, and dynamically modulates the alignment objective to penalize unwarranted confidence. Experiments show that ERA effectively balances answer coverage and safe abstention, improving reliability and generalizing robustly to out-of-distribution settings.

1 Introduction

Recent advances in Large Language Models (LLMs) have profoundly transformed Natural Language Processing, yet they remain prone to hallucinations due to their reliance on static, potentially outdated training data (Kalai et al., 2025; Zhang et al., 2024). While Retrieval-Augmented Generation (RAG) effectively mitigates this by grounding generation in external factual evidence (Lewis et al., 2020), its hybrid nature introduces a new critical challenge: *knowledge conflicts* between a model's internalized parametric memory and the retrieved information (Longpre et al., 2021; Mallen et al., 2023). Such discrepancies create ambiguity during generation, often leading to inconsistent behavior, degraded response quality, or erroneous reasoning when the model fails to appropriately prioritize the correct information source. Consequently, resolving these conflicts is essential for ensuring the trustworthiness of retrieval-augmented systems in real-world applications.

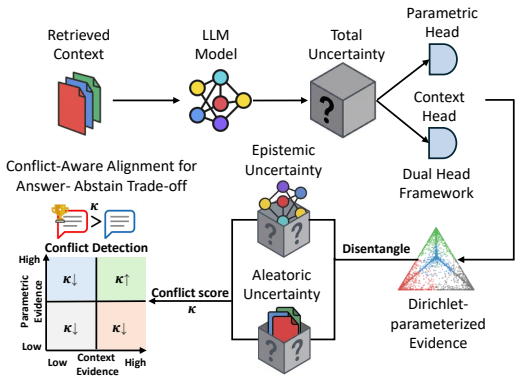


Figure 1: Overview of the ERA framework. By disentangling epistemic from aleatoric uncertainty, it dynamically aligns behavior to optimize the trade-off between answering and safe abstention.

Recent studies have attempted to mitigate this issue by relying on output-level alignment corrections for abstention through the construction of curated boundary datasets. For instance, DTA categorizes training instances into distinct knowledge quadrants—based on whether the model possesses the parametric knowledge and whether the retrieved context is accurate—and uses preference optimization to instill specific abstain patterns (Sun et al., 2025). Similarly, TruthRL constructs reward systems that penalize hallucinations and explicitly incentivize the model to abstain when faced with conflicting information (Wei et al., 2025).

Despite these advances, relying solely on output-level behavioral alignment fails to disentangle *epistemic* uncertainty (a lack of internal knowledge) from *aleatoric* uncertainty (inherent ambiguity in the retrieved context) (Kendall and Gal, 2017). Under this entanglement, the model remains blind to the root cause of its uncertainty, unable to discern whether it arises from its own lack of knowledge or from a flawed retrieval source. This blindness becomes particularly damaging when the parametric guess and the retrieved context present confident yet conflicting signals—for instance, when the

model is parametrically certain of an answer that the retrieved context confidently refutes or when reliable retrieval contradicts a stale parametric belief. Without a principled way to identify which source to trust under such conflicts, the model resorts to memorizing dataset-specific refusal patterns that fail to generalize, causing the answer-abstain trade-off to collapse under distribution shift.

In light of these limitations, we propose ERA (Evidence-based Reliability Alignment) for *honest RAG systems* (Figure 1). The key novelty of our approach lies in *explicitly disentangling epistemic and aleatoric uncertainties* to fundamentally optimize the answer-abstain trade-off. Rather than relying on entangled heuristics, this disentanglement empowers the model to determine which source to trust when confronted with conflicting signals. This framework consists of two key components.

The first component, *Dual Head Evidence Quantification*, models epistemic and aleatoric uncertainty by independently capturing evidential support from parametric memory and retrieved context via separate heads. Specifically, we leverage evidential deep learning to shift confidence estimation from standard scalar probabilities to Dirichlet-parameterized evidence distributions (Sensoy et al., 2018). The dual-head structure, operating on with- and without-context inputs, prevents the entanglement of external aleatoric noise, where Dirichlet parameterization mitigates forced overconfidence by quantifying evidential scarcity. These mechanisms establish a strict epistemic boundary that explicitly exposes what the model does not know.

The second component, *Quantifying Knowledge Conflict*, fuses the two Dirichlet-parameterized evidence distributions into a continuous conflict score κ that serves as a dynamic weighting mechanism for the alignment objective. By design, κ peaks only when both heads are confident yet contradictory, and is naturally suppressed when either is uncertain. This selectivity ensures that the model is not penalized merely for internal ignorance; instead, the loss is amplified precisely when parametric and retrieval signals confidently disagree, forcing the system to learn principled reconciliation rather than surface-level refusal heuristics.

Extensive experiments demonstrate that ERA effectively optimizes the inherent trade-off between answer coverage and safe abstention. Compared to recent baselines on standard RAG benchmarks, our framework significantly improves response reliability: it maintains high accuracy on fully supported queries while selectively withholding answers under unresolvable knowledge conflicts. We further demonstrate this robust generalizability on a specially curated out-of-distribution (OOD) benchmark built from Wikipedia events. Finally, comprehensive empirical analyses confirm that ERA yields exceptionally well-calibrated confidence estimates, ensuring its predicted uncertainty faithfully reflects the true likelihood of correctness.

2 Related Works

Existing works for Honest RAG are bifurcated into inference-time and training-time methods.

2.1 Inference-Time Reliability Methods

Initial efforts focused on estimating reliability without modifying model parameters. LogProb uses average or minimum log-probability across tokens as a confidence metric (Guerreiro et al., 2023), while P_{true} extracts the softmax probability of a "True" token from a verification prompt (Kadavath et al., 2022). Sampling-based approaches exploit generation variability: ICL prompts the model to verbally self-evaluate its outputs (Wei et al., 2022), and Self-Consistency samples multiple reasoning paths and selects the most reliable answer via majority voting (Wang et al., 2023). More recently, decoding-based interventions have emerged. Ada-CAD dynamically adjusts decoding weights based on the Jensen–Shannon divergence between contextual and parametric distributions (Wang et al., 2025a), and RAUQ quantifies uncertainty through attention entropy to modulate generation (Vazhentsev et al., 2025). Fundamentally, inference-time methods share a fundamental limitation: they rely on heuristic thresholds sensitive to a domain and do not reshape the model's underlying policy.

2.2 Training-Time Alignment Methods

A parallel thread frames honest RAG as a training-time alignment problem, embedding abstention directly into the model's policy. TruthRL employs reinforcement learning with a ternary reward system (correct/hallucination/abstention) via GRPO to guide the model toward preferring refusal over fabrication when knowledge is insufficient (Wei et al., 2025). Divide-Then-Align (DTA) partitions data into four knowledge quadrants based on parametric knowledge and retrieval quality, constructing targeted preference pairs for DPO alignment (Sun et al., 2025). Unlike inference-time heuristics,

these approaches eliminate the need for manual threshold calibration.

However, existing alignment methods teach the model *when* to abstain but lack the *epistemic justification*, applying uniform optimization regardless of conflict intensity and failing to distinguish epistemic uncertainty from aleatoric uncertainty. Without this decomposition, the model overfits to dataset-specific refusal patterns that collapse under distribution shift. ERA addresses this by separately quantifying evidential support from parametric and retrieval pathways, fusing them into a conflict score, and dynamically amplifying the alignment penalty when the two sources contradict.

3 Problem Formulation: Honest RAG

We formalize the problem of honest assessment by defining knowledge boundaries and constructing the preference alignment dataset $\mathcal{D}$. In our standard RAG setting, a retriever $\mathcal{R} : \mathcal{Q} \to \mathcal{P}(\mathcal{D}_{\text{corpus}})$ that fetches contexts $\mathcal{C} = \mathcal{R}(q)$, and a generator $\mathcal{G} : \mathcal{Q} \times \mathcal{C} \to \mathcal{A}$ producing an answer via $P(\cdot|q, \mathcal{C})$. To ensure honest behavior, the generator $\mathcal{G}$ is instructed via a system prompt to either provide a substantive answer or explicitly respond with "I don't know" (IDK) when the available information is insufficient, making abstention an integral part of the output space. Let $g : \mathcal{Q} \to \mathcal{A}$ denote the ground truth function, $C(\cdot) : \mathcal{Q} \times \mathcal{A} \to \{0, 1\}$ a correctness indicator, and $CT(\cdot) : \mathcal{C} \times \mathcal{A} \to \{0, 1\}$ a context-containment indicator returning 1 if context c contains $g(q)$.

Knowledge Boundaries. The total knowledge capacity of a RAG system, denoted as $\mathcal{K}_{\text{rag}}$, is defined by the union of its internal parametric boundary ($\mathcal{K}_{\text{param}}$) and its external retrieval boundary ($\mathcal{K}_{\text{retr}}$), such that $\mathcal{K}_{\text{rag}} = \mathcal{K}_{\text{param}} \cup \mathcal{K}_{\text{retr}}$. We formalize these constituent boundaries as follows:

$$\mathcal{K}_{\text{param}} = \{q \in \mathcal{Q} \mid C(\mathcal{G}(q, \emptyset)) = 1\}$$
$$\mathcal{K}_{\text{retr}} = \{q \in \mathcal{Q} \mid \exists c \in \mathcal{R}(q) : CT(c, g(q)) = 1\} \quad \text{(Eq. 1)}$$

To operationalize these definitions, we estimate $\mathcal{K}_{\text{param}}$ by sampling N independent answers from the context-free model $\mathcal{G}(q, \emptyset)$ and requiring the empirical accuracy to exceed a strict threshold δ. For $\mathcal{K}_{\text{retr}}$, we use GPT-4o to judge whether any of the three retrieved passages contains or entails the ground truth $g(q)$ (see the Appendix A for details).

Knowledge Quadrants Dataset. We construct the preference alignment dataset $\mathcal{D} = \{(q,\ y_c,\ y_r,\ y)\}$, where each instance comprises a query q, chosen and rejected responses (y_c, y_r), and a one-hot quadrant label $y \in \{0, 1\}^4$. Each sample is assigned a quadrant label y based on the intersection of $\mathcal{K}_{\text{param}}$ and $\mathcal{K}_{\text{retr}}$ (Figure 2). The preference pairs are then defined per quadrant to enforce the appropriate behavior: ① **Known-Gold** ($q \in \mathcal{K}_{\text{param}} \cap \mathcal{K}_{\text{retr}}$): both sources support the answer, so y_c is the ground truth and y_r is IDK to penalize unnecessary abstention; ② **Known-Noisy** ($q \in \mathcal{K}_{\text{param}} \setminus \mathcal{K}_{\text{retr}}$): the model knows the answer but retrieval is noisy, so y_r includes context-induced hallucinations and IDK to discourage both noise vulnerability and over-conservatism; ③ **Unknown-Gold** ($q \in \mathcal{K}_{\text{retr}} \setminus \mathcal{K}_{\text{param}}$): the model must rely on retrieved evidence, so y_r targets context-ignoring errors, parametric hallucinations, and unjustified IDK; and ④ **Unknown-Noisy** ($q \notin \mathcal{K}_{\text{param}} \cup \mathcal{K}_{\text{retr}}$): neither source is reliable, so y_c is IDK and y_r includes hallucinations and even the ground truth to prevent unsupported lucky guesses.

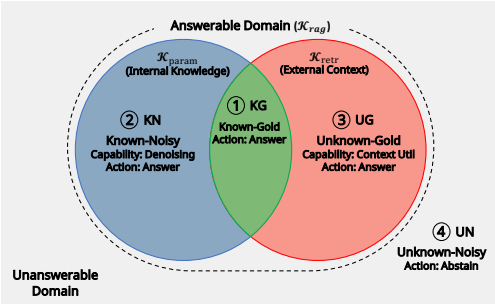


Figure 2: Categorization of Knowledge Domains and Response Strategies in RAG Models.

Evaluation Metrics. We evaluate along three axes. *Answer Quality* (Precision, Recall, F1) measures correctness among answerable queries in $\mathcal{K}_{\text{rag}}$. *Retrieval Handling* isolates two capabilities: Denoise Rate measures accuracy when retrieval is noisy ($\mathcal{K}_{\text{param}} \setminus \mathcal{K}_{\text{retr}}$), and Context Utilization measures accuracy when only retrieved context is reliable ($\mathcal{K}_{\text{retr}} \setminus \mathcal{K}_{\text{param}}$). *Abstain Quality* (Precision, Recall, F1) evaluates refusal behavior on unanswerable queries ($q \notin \mathcal{K}_{\text{param}} \cup \mathcal{K}_{\text{retr}}$). *Overall F1*, the harmonic mean of Answer F1 and Abstain F1, captures the trade-off between the two. Detailed definitions are provided in Table 2 of Appendix A.

4 Proposed Framework: ERA

Given the preference alignment dataset $\mathcal{D}$, our goal is to optimize $\mathcal{G}$ such that it adheres to the preference pairs (y_c, y_r) while jointly learning to predict the knowledge quadrant label y. Figure 3 illustrates the overall workflow. Our framework, ERA, first

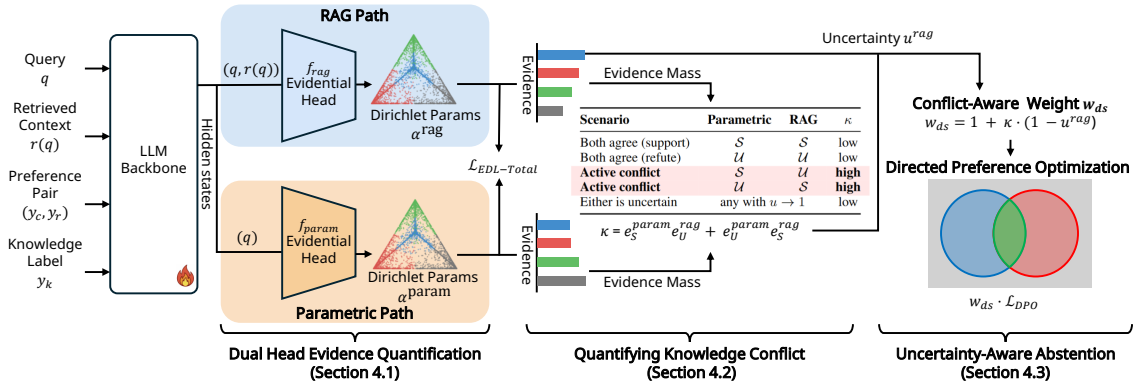


Figure 3: Overview of ERA: (1) Dual-Head Evidence Quantification projects parametric and retrieval representations into Dirichlet distributions; (2) Knowledge Conflict Quantification fuses evidence masses into a conflict score κ; (3) Uncertainty-Aware Abstention modulates the alignment objective based on κ.

integrates dual evidential heads to independently quantify the uncertainty of parametric memory and retrieved context (Section 4.1). It then fuses these estimates to derive an explicit conflict score that measures the discrepancy between internal and external knowledge (Section 4.2). Finally, this conflict signal dynamically modulates the preference alignment objective, amplifying the learning pressure on high-conflict, high-evidence samples while dampening it under noisy retrieval (Section 4.3).

4.1 Dual Head Evidence Quantification

Our goal is to independently quantify evidential support from both parametric memory and retrieved context, since a single uncertainty estimate in standard RAG obscures whether the uncertainty arises from intrinsic ignorance or genuine knowledge conflicts. To address this, we integrate Evidential Deep Learning (EDL) (Sensoy et al., 2018) into a dual-head architecture, achieving uncertatiny disentanglement on two fronts. Structurally, the dual-head framework disentangles the uncertainty sources by physically isolating internal memory from external context. At the optimization level, the EDL objective parameterizes a Dirichlet prior over the predictive distributions, ensuring a strict loss-level disentanglement by penalizing false overconfidence when evidence is lacking.

Dual-Pathway Dirichlet Parameterization. We introduce a dual-pathway design by integrating two auxiliary evidential heads into the LLM $\mathcal{G}$ (f_{rag} and f_{param}), allowing f_{param} to estimate a strictly parametric evidential prior within the same evidential space as f_{rag} for consistent parametric-contextual comparison. These heads operate on two distinct input pathways: the RAG path processes the full retrieval-augmented prompt, whereas the parametric path processes a context-stripped prompt containing only the original query. Finally, the extracted hidden states are projected into logit vectors $z \in \mathbb{R}^K$, where $K = 4$ corresponds to the four distinct knowledge quadrants. [1]

According to the Subjective Logic (Jsang, 2018), predictive outputs are interpreted not as absolute probabilities, but as consisting of an evidence mass e_k for each quadrant and an uncertainty mass u, strictly constrained by $\sum_{k=1}^{K} e_k + u = 1$. This formulation is naturally bridged with the Dirichlet distribution: the evidence and uncertainty masses are analytically derived from the concentration parameters α_k as:

$$e_k = \frac{\alpha_k - 1}{S}, \quad u = \frac{K}{S}, \quad \text{where } S = \sum_{k=1}^{K} \alpha_k \qquad \text{(Eq. 2)}$$

Intuitively, α_k represents the accumulated evidence. When evidence is lacking ($\alpha_k \to 1$ for all k), uncertainty dominates ($u \to 1$), allowing the model to explicitly express ignorance rather than producing false confidence.

Optimization. Having established the evidential decomposition into e_k and u, we now describe how to learn the underlying concentration parameters. Each evidential head projects its hidden state into a logit vector $z \in \mathbb{R}^K$, which is mapped to concentration parameters via $\alpha_k = \text{Softplus}(z_k) + 1$. These parameters define a Dirichlet density over the K-dimensional unit simplex,

[1] Although f_{param} does not explicitly observe the retrieved context, the query itself carries implicit signals regarding likely retrieval quality (e.g., rare or highly specific queries tend to yield noisy retrieval, whereas common factual queries are more likely to be well-supported). Training over all four quadrants thus enables f_{param} to accurately capture these query-conditioned priors.

where $\mathbf{p} = [p_1, \ldots, p_K]^\top$ represents a categorical probability assignment over the K quadrants:

$$\mathrm{Dir}(\mathbf{p} \mid \boldsymbol{\alpha}) = \frac{1}{B(\boldsymbol{\alpha})} \prod_{k=1}^{K} p_k^{\alpha_k - 1} \qquad \text{(Eq. 3)}$$

where $B(\boldsymbol{\alpha})$ is the multinomial beta function.

The evidential heads must learn to produce high α_k for the correct quadrant while reverting to a uniform prior ($\boldsymbol{\alpha} \to \mathbf{1}$) when evidence is lacking. We achieve this with two complementary loss terms. The first term minimizes the average cross-entropy loss over the full Dirichlet distribution, rather than evaluating just one specific probability:

$$\mathcal{L}_{\text{fit}}(\boldsymbol{\alpha}, \boldsymbol{y}) = \int \left(\sum_{k=1}^{K} -y_k \log p_k \right) \mathrm{Dir}(\mathbf{p} \mid \boldsymbol{\alpha})\, d\mathbf{p} \qquad \text{(Eq. 4)}$$

This penalizes the model when α_k for the ground-truth quadrant is low, forcing the network to concentrate evidence on the correct class. The second term suppresses spurious evidence for incorrect classes by penalizing the divergence between the predicted posterior and the uniform prior $\mathrm{Dir}(\mathbf{p} \mid \mathbf{1})$, which represents zero evidence and total uncertainty (u=1):

$$\mathcal{L}_{\text{KL}}(\boldsymbol{\alpha}, \boldsymbol{y}) = \mathrm{KL}\left[\mathrm{Dir}(\mathbf{p} \mid \tilde{\boldsymbol{\alpha}}) \| \mathrm{Dir}(\mathbf{p} \mid \mathbf{1})\right] \qquad \text{(Eq. 5)}$$

where $\tilde{\boldsymbol{\alpha}} = \boldsymbol{y} + (1 - \boldsymbol{y}) \odot \boldsymbol{\alpha}$ retains the target class evidence and regularizes only the non-target classes. The combined per-head loss is $\mathcal{L}_{\text{EDL}} = \mathcal{L}_{\text{fit}} + \lambda_t \mathcal{L}_{\text{KL}}$, where λ_t is an annealing coefficient that increases over training to prevent premature convergence. Applied to both heads f_{rag} and f_{param}, the total evidential objective is:

$$\mathcal{L}_{\text{EDL-Total}} = \mathcal{L}_{\text{EDL}}(\boldsymbol{\alpha}^{\text{rag}}, \boldsymbol{y}) + \mathcal{L}_{\text{EDL}}(\boldsymbol{\alpha}^{\text{param}}, \boldsymbol{y}) \qquad \text{(Eq. 6)}$$

4.2 Quantifying Knowledge Conflict

Given the independently quantified evidential distributions from parametric and retrieval sources, we next assess their conflict which distinguish mutual ignorance (both sources uncertain) from active conflict (both confident yet contradictory). We adopt Dempster-Shafer Theory (DST), whose combination rule naturally separates the two by fusing evidence masses while accounting for uncertainty (Shafer, 2020; Wang et al., 2025b). The conflict score κ is high only when both sources are individually confident yet support mutually exclusive hypotheses—precisely the scenario most likely to produce hallucinations—and is automatically attenuated when either source is uncertain.

Evidence Aggregation. While our evidential heads predict fine-grained knowledge quadrants, the fundamental conflict in RAG scenarios is intrinsically binary: a context either provides sufficient information to support an answer or fails to do so. Known-Gold (KG) and Unknown-Gold (UG) both indicate sufficient grounding (Supported, $\mathcal{S}$), whereas Known-Noisy (KN) and Unknown-Noisy (UN) indicate an absence of valid support (Unsupported, $\mathcal{U}$). We therefore group the quadrants into two mutually exclusive sets:

$$\mathcal{S} = \{\text{KG, UG}\}, \quad \mathcal{U} = \{\text{KN, UN}\} \qquad \text{(Eq. 7)}$$

This reduction reflects the insight that the most severe hallucinations arise not from intra-group misclassifications (e.g., confusing KG with UG), but from a fundamental disagreement about the trustworthiness of the evidence itself. For each source $s \in \{\text{rag, param}\}$, we aggregate the $\mathcal{S}$ and $\mathcal{U}$ evidence by summing the masses of constituent quadrants:

$$e^s_{\mathcal{S}} = \sum_{k \in \mathcal{S}} e^s_k, \quad e^s_{\mathcal{U}} = \sum_{k \in \mathcal{U}} e^s_k \qquad \text{(Eq. 8)}$$

Conflict Score (κ). With the aggregated evidence, we quantify the tension between the model's internal prior (s=param) and the external evidence (s=rag). The conflict score κ captures the mass where the two sources support mutually exclusive hypotheses:

$$\kappa = e^{\text{param}}_{\mathcal{S}} \cdot e^{\text{rag}}_{\mathcal{U}} + e^{\text{param}}_{\mathcal{U}} \cdot e^{\text{rag}}_{\mathcal{S}} \qquad \text{(Eq. 9)}$$

Because uncertainty u^s is strictly isolated from evidence, high ignorance in either source ($u^s \to 1$) shrinks the available evidence ($e^s_{\mathcal{S}}, e^s_{\mathcal{U}} \to 0$), naturally suppressing κ. Conversely, κ activates only when both sources are individually confident yet mutually contradictory—precisely the condition most likely to produce hallucinations.

4.3 Uncertainty-Aware Abstention

The ultimate goal of our framework is to calibrate the model's behavior so that it selectively abstains when intrinsic knowledge is insufficient or contradicts reliable external evidence. Standard preference optimization (e.g., DPO) treats all training samples uniformly, which is suboptimal for learning these delicate decision boundaries (Rafailov

et al., 2023). To address this, we propose a training strategy that prioritizes learning from high-conflict, high-evidence scenarios.

Conflict-Aware Loss Modulation. The most critical signal for learning robustness lies in samples where the model's internal belief conflicts with the retrieved context, but only if that context is trustworthy. We formulate a dynamic weighting factor w_{ds} applied to the alignment loss:

$$w_{\text{ds}} = 1 + \kappa \cdot (1 - u^{\text{rag}}) \qquad \text{(Eq. 10)}$$

Here $(1 - u^{\text{rag}})$ acts as a reliability gate. When conflict κ is high and retrieval uncertainty is low, w_{ds} increases, forcing the model to resolve the dissonance. Conversely, when the retrieval itself is uncertain ($u^{\text{rag}} \approx 1$), the gate closes, keeping w_{ds} near 1.0 and preventing the model from aggressively updating based on noisy retrieval.

Preference Alignment. To align the model with human preferences, we employ Direct Preference Optimization (DPO), which optimizes the policy directly using pairs of chosen (y_c) and rejected (y_r) responses. Given a query q and retrieved context $r(q)$, the DPO loss is:

$$\mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) = -\log \sigma \left(\tau \left[\log \frac{\pi_\theta(y_c|q,r(q))}{\pi_{\text{ref}}(y_c|q,r(q))} - \log \frac{\pi_\theta(y_r|q,r(q))}{\pi_{\text{ref}}(y_r|q,r(q))} \right] \right), \qquad \text{(Eq. 11)}$$

where π_θ is the policy being trained, π_{ref} is the frozen reference model, and τ is the temperature parameter.

Unified Training Objective. We integrate evidential learning and preference alignment into a unified objective:

$$\mathcal{L}_{\text{total}} = w_{\text{ds}} \cdot \mathcal{L}_{\text{DPO}}(\pi_\theta; \pi_{\text{ref}}) + \mathcal{L}_{\text{EDL-Total}} \qquad \text{(Eq. 12)}$$

where w_{ds} is the conflict-aware modulation factor from Eq. Eq. 10. By jointly optimizing this objective, ERA learns to adhere to human preferences while calibrating its evidential distributions to accurately reflect knowledge boundaries.

5 Experiment

5.1 Performance Evaluation

Datasets. Consistent with prior RAG studies (Fang et al., 2024; Sun et al., 2025), we evaluate on three open-domain QA benchmarks: Natural Questions (NQ) (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), and WebQuestions (WebQ) (Berant et al., 2013). We use Dense Passage Retrieval (DPR) (Karpukhin et al., 2020) to fetch the top-3 Wikipedia passages per query. Each query is categorized into one of four knowledge quadrants by cross-referencing the model's parametric knowledge boundary ($\mathcal{K}_{\text{param}}$) with the retrieval quality ($\mathcal{K}_{\text{retr}}$). We adopt the annotation and evaluation protocols from DTA (Sun et al., 2025), using a strict consistency threshold of δ=1.0 with N=10 sampled responses to classify each query as Known or Unknown. We randomly sample 5,000 training and 3,000 evaluation examples from the stratified quadrants across the three datasets.

To further assess generalizability on temporally unseen information, we construct the **Wiki Event** benchmark by parsing Wikipedia annual "Events" pages (2010–2025) and scraping referenced news articles as gold contexts. Each query is paired with a stratified pool of 10 candidate contexts: 1 gold, 4 hard negatives (entity-perturbed counterfactuals), and 5 easy negatives (random corpus samples). Implementation details are provided in Appendix A.

Baselines. We compare against 8 calibration strategies: (1) **Logprob** (Kadavath et al., 2022) calculates the average log-probability of all tokens in the generated answer, serving as a basic confidence metric. (2) **ICL** (Wei et al., 2022) prompts the model to verbally self-evaluate correctness via 3-shot examples; (3) **Self-Consistency** (Wang et al., 2023) measures confidence through pairwise exact-match agreement across 10 sampled outputs, abstaining below a 0.5 threshold; (4) $P_{\textbf{true}}$ (Kadavath et al., 2022) extracts the softmax probability of the "True" token from a binary verification prompt; (5) **RAUQ** (Vazhentsev et al., 2025)leverages attention-based uncertainty metrics; (6) **AdaCAD** (Wang et al., 2025a) dynamically adjusts decoding weights based on the Jensen–Shannon divergence between RAG and context-free outputs; (7) **TruthRL** (Wei et al., 2025) fine-tunes the policy via GRPO with a ternary reward (+1/0/−1) that explicitly incentivizes abstention; and (8) **DTA** (Sun et al., 2025) applies DPO on curated abstention datasets to train selective refusal.

Results. Table 1 demonstrates that ERA achieves the highest *Overall F1* on both Llama3-8B (Dubey et al., 2024) and Qwen3-8B-Base (Yang et al., 2025). This metric directly captures the central goal of honest RAG: *optimally balancing answer coverage against safe abstention*. By explicitly quantifying epistemic evidence, ERA strategically abstains only during genuine knowledge conflicts.

Table 1: Performance comparison across three QA datasets using Qwen3-8B and Llama3.1-8B backbones. For threshold-based methods, we report the threshold achieving the best Overall F1. **Overall F1** represents the harmonic mean balancing Answer Quality F1 and Abstention Quality F1.

Method	Retrieval Handling		Answer Quality			Abstention Quality			Overall F1
	Denoise	Context	Recall	Precision	F1	Recall	Precision	F1	
Qwen3-8B-Base									
Logprob	0.187	0.144	0.197	0.933	0.325	0.689	0.360	0.473	0.385
ICL	0.184	0.142	0.178	0.130	0.150	0.048	0.404	0.085	0.109
Self-Consistency	0.219	0.142	0.203	0.415	0.273	**0.777**	0.350	0.483	0.349
P_{true}	0.245	0.128	0.209	0.931	0.342	0.521	0.399	0.452	0.389
RAUQ	0.450	0.327	0.375	0.309	0.339	0.634	0.835	**0.721**	0.461
AdaCAD	0.578	0.227	0.460	**0.938**	**0.617**	0.367	0.351	0.359	0.454
TruthRL	0.443	0.306	0.357	0.264	0.303	0.511	**0.852**	0.639	0.412
DTA	0.380	0.319	0.436	0.410	0.422	0.476	0.566	0.517	0.465
ERA	**0.606**	**0.454**	**0.622**	0.532	0.573	0.408	0.715	0.520	**0.545**
Llama3.1-8B									
Logprob	**0.568**	**0.392**	0.509	0.910	**0.653**	0.314	0.424	0.361	0.465
ICL	0.548	0.326	0.474	0.336	0.394	0.014	0.800	0.027	0.050
Self-Consistency	0.392	0.300	0.386	0.492	0.432	0.559	0.368	0.444	0.438
P_{true}	0.352	0.207	0.337	0.934	0.495	0.497	0.484	0.490	0.493
RAUQ	0.135	0.064	0.081	0.061	0.070	0.525	**0.846**	**0.648**	0.126
AdaCAD	0.448	0.138	0.348	**0.943**	0.503	0.562	0.319	0.401	0.453
TruthRL	0.353	0.282	0.383	0.359	0.371	0.449	0.835	0.584	0.454
DTA	0.510	0.322	0.490	0.544	0.516	0.596	0.481	0.533	0.524
ERA	0.504	0.367	**0.537**	0.680	0.600	**0.759**	0.506	0.607	**0.604**

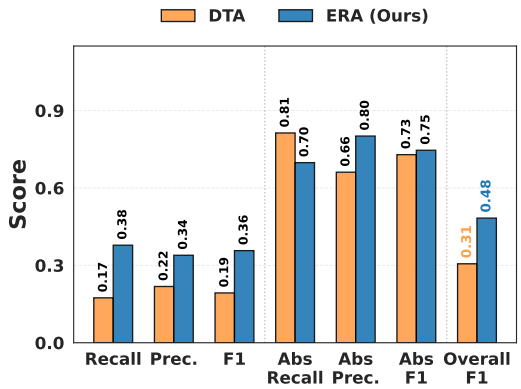


Figure 4: Generalizability evaluation on the OOD Wiki Event dataset using Qwen3-8B-Base, comparing ERA against DTA, the strongest baseline.

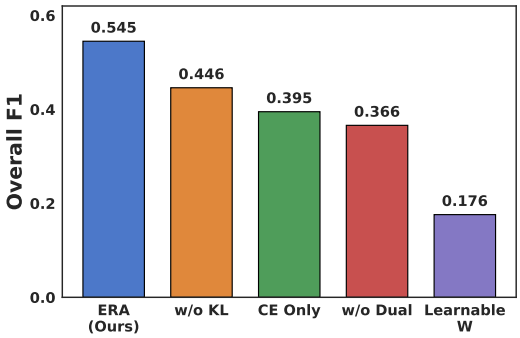


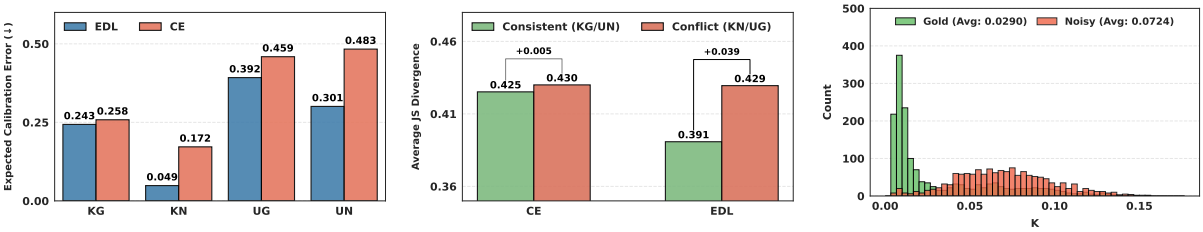
Figure 5: Performance comparison of ablation studies. The consistent degradation in Overall F1 scores observed when removing or altering individual components validates the distinct contribution of each module.

This targeted mechanism yields superior *Abstention Precision* while safely preserving high answer coverage, by ensuring that answerable queries are no longer incorrectly rejected. Consequently, this balanced optimization allows ERA to outperform all baselines in Overall F1.

Figure 4 evaluates models trained on the three QA datasets and directly transferred to the OOD Wiki Event benchmark. ERA significantly outperforms DTA. DTA exhibits severe over-abstention, indiscriminately rejecting valid queries due to overfitted refusal patterns, whereas ERA leverages disentangled uncertainty to establish a precise decision boundary that filters OOD noise without compromising Answer Quality.

5.2 Component Analysis

Ablation Study. Figure 5 presents an ablation study evaluating the contribution of each component in ERA. We compare against four variants targeting the core architecture and learning objectives: (1) **w/o Dual** removes the parametric head, relying solely on RAG context and bypassing internal-external conflict detection; (2) **Learnable W** replaces evidence-based DST fusion with a simple learnable scalar gate; (3) **CE Only** substitutes the evidential objective with standard cross-entropy,



(a) Expected Calibration Error (b) Conflict Sensitivity Analysis (c) Conflict Score Analysis

Figure 6: Quantitative analysis of uncertainty on Wiki Event dataset. This figure compares the proposed Evidential Deep Learning (EDL) model against the Cross-Entropy (CE) baseline across three dimensions: (a) A visualization of prediction probability distributions in conflict situations (KN, UG); (b) A comparison of Expected Calibration Error (ECE) across data types; and (c) An evaluation of model sensitivity via the Jensen-Shannon Divergence (JSD) between consistent and conflicting information.

ignoring explicit uncertainty quantification; and (4) **w/o KL** omits the KL regularization term from the evidential loss. We further analyze (i) evidence fusion structure and (ii) conflict gating strategies; detailed analyses are provided in the Appendix B.

Qualitative Analysis. We first investigate how well our evidential framework calibrates uncertainty compared to standard optimization methods. To visualize this, we compare the uncertainty distributions and Expected Calibration Error (ECE) between the Cross-Entropy (CE) baseline and our Evidential Deep Learning (EDL) approach. This improved calibration is quantitatively supported by Figure 6a, where EDL consistently achieves substantially lower ECE scores across all four knowledge quadrants (KG, KN, UG, UN). Notably, in the high-risk Unknown-Noisy (UN) and Known-Noisy (KN) scenarios, EDL significantly reduces calibration error, demonstrating its effectiveness in mitigating the overconfidence typically observed in hallucinating models.

Next, we examine the model's sensitivity to knowledge conflicts—specifically, determining whether the system can effectively distinguish between scenarios where its internal parametric memory aligns with external retrieved evidence versus when they contradict. To quantify this, we utilize the Jensen-Shannon (JS) divergence to measure the distributional discrepancy between the predictive distributions derived from the parametric and retrieval-augmented paths. Figure 6b reveals that the CE baseline fails to capture this distinction, showing a negligible increase in divergence (+0.005) when a conflict occurs. Conversely, our method demonstrates a marked sensitivity to discordance, with a significant increase in JS divergence (+0.039) for conflicting samples compared to consistent ones. This suggests that our evidential framework effectively disentangles aleatoric uncertainty from epistemic belief, allowing the model to explicitly detect conflicts and thereby filter out contaminated knowledge more effectively than standard probabilistic approaches.

Furthermore, we analyze how this conflict detection mathematically governs the training dynamics via the conflict score (κ). As illustrated in Figure 6c, our framework robustly separates Noisy contexts (KN+UN, mean=0.072) from Gold contexts (KG+UG, mean=0.029) with a substantial effect size (Cohen's d=1.47, $p < 0.001$). This separation directly propagates into the final modulation weight w_{ds} (d=1.17, $p < 0.001$), confirming that high-conflict samples receive a meaningfully stronger gradient signal. This modulation is highly effective because the retrieval path tends to be overconfident on noisy inputs (uncertainty mass $=$ 0.471 for Noisy vs. 0.511 for Gold). Consequently, this overconfidence keeps the reliability gate $(1 - u^{\text{rag}})$ open, amplifying the DPO loss precisely when corrective abstention learning is most needed.

6 Conclusion

This paper proposed ERA, an evidential framework that disentangles epistemic and aleatoric uncertainties within RAG systems. By parameterizing predictive distributions with Dirichlet priors, ERA explicitly measures the evidence and uncertainty of both internal parametric knowledge and external retrieved contexts. Building on this formulation, we introduced a Conflict Score (κ) that isolates genuine structural disagreements from mere ignorance, empowering the model to reliably abstain rather than hallucinate. These findings underscore the importance of moving from point-estimate probabilities to evidential distributions in the pursuit of trustworthy RAG architectures.

Limitations

Our results are best interpreted as demonstrating that ERA enhances the answer-abstain trade-off within a realistic but imperfect approximation of knowledge boundaries. To guarantee a fair comparison, we deliberately employ the same boundary-construction and evaluation protocol established in prior work (Sun et al., 2025), ensuring that ERA and all baselines are benchmarked on identical labels. While the LLM-assisted nature of these procedures might introduce systematic annotation bias, any such bias is distributed uniformly across all methods. The relative performance gains therefore remain entirely valid, even without oracle supervision to define an absolute frontier. The reported metrics should nonetheless be understood within the context of this established evaluation framework.

Ethics Statement

To construct the out-of-distribution Wiki Event benchmark, we scraped the 'Events' sections of Wikipedia annual pages (2010–2025). We acknowledge that Wikipedia text is governed by the CC BY-SA 4.0 license. In strict adherence to this license and general publishing ethics, we ensure proper attribution of the source material. Furthermore, we confirm that our data collection and redistribution process respects the original authors' copyrights and promotes responsible open-source NLP research. Lastly, while our framework is specifically designed to improve the reliability of RAG systems, we caution against over-reliance. Users should remain aware that no system is entirely infallible, and treating an aligned model as fully safe or absolute truth poses inherent risks.

References

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proc. of the Empirical Methods in Natural Language Processing*, pages 1533–1544.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, and 1 others. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Feiteng Fang, Yuelin Bai, Shiwen Ni, Min Yang, Xiaojun Chen, and Ruifeng Xu. 2024. Enhancing noise robustness of retrieval-augmented language models with adaptive adversarial training. In *Proc. of the Association for Computational Linguistics*, pages 10028–10039.

Nuno M. Guerreiro, Elena Voita, and André F. T. Martins. 2023. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proc. of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075.

Mandar Joshi, Eunsol Choi, Daniel S Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proc. of the Meeting of the Association for Computational Linguistics*, pages 1601–1611.

Audun Jsang. 2018. *Subjective Logic: A Formalism for Reasoning Under Uncertainty*. Springer Publishing Company, Incorporated.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2025. Why language models hallucinate. *arXiv preprint arXiv:2509.04664*.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proc. of the Empirical Methods in Natural Language Processing*, pages 6769–6781.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Proc. of the Neural Information Processing Systems*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, and 1 others. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:453–466.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive NLP tasks. In *Proc. of the Neural Information Processing Systems*, pages 9459–9474.

Shayne Longpre, Kartik Perisetla, Anthony Chen, Nikhil Ramesh, Chris DuBois, and Sameer Singh. 2021. Entity-based knowledge conflicts in question answering. In *Proc. of the Empirical Methods in Natural Language Processing*.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proc. of the Association for Computational Linguistics*, pages 9802–9822.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D. Manning, Stefano Ermon, and Chelsea Finn. 2023. Direct preference optimization: Your language model is secretly a reward model. In *Proc. of the Neural Information Processing Systems*, pages 53728–53741.

Murat Sensoy, Lance Kaplan, and Melih Kandemir. 2018. Evidential deep learning to quantify classification uncertainty. In *Proc. of the Neural Information Processing Systems*.

Glenn Shafer. 2020. *A Mathematical Theory of Evidence*. Princeton University Press.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Zilei Wang, Weiqiang Wang, and Liang Wang. 2025. Divide-then-align: Honest alignment based on the knowledge boundary of RAG. In *Proc. of the Association for Computational Linguistics*, pages 11461–11480.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and 1 others. 2025. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for LLMs. *arXiv preprint arXiv:2505.20045*.

Han Wang, Archiki Prasad, Elias Stengel-Eskin, and Mohit Bansal. 2025a. AdaCAD: Adaptively decoding to balance conflicts between contextual and parametric knowledge. In *Proc. of the Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 11636–11652.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *Proc. of the International Conference on Learning Representations*.

Zheng Wang, Xing Xu, Lei Zhu, Yi Bin, Guoqing Wang, Yang Yang, and Heng Tao Shen. 2025b. Evidence-based multi-feature fusion for adversarial robustness. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V. Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *Proc. of the Neural Information Processing Systems*, pages 24824–24837.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, and 1 others. 2025. TruthRL: Incentivizing truthful LLMs via reinforcement learning. *arXiv preprint arXiv:2509.25760*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Muru Zhang, Ofir Press, William Merrill, Alisa Liu, and Noah A. Smith. 2024. How language model hallucinations can snowball. In *Proc. of the International Conference on Machine Learning*, pages 59670–59684.

A Implementation Details

We release the source code and dataset for ERA at https://anonymous.4open.science/r/ERA-318B.

Training Details We employ Qwen3-8B-Base and Llama-3.1-8B as the primary backbone models, optimized via 4-bit Quantized LoRA (QLoRA) with nf4 quantization and bfloat16 precision to ensure computational efficiency. LoRA adapters are applied to all linear layers with a rank $r = 8$, $\alpha = 16$, and a dropout rate of 0.05. We train the model for 1 epoch using the AdamW optimizer with a cosine learning rate scheduler, initialized at $5e-5$ with a warmup ratio of 0.1. For the ERA framework, we introduce two lightweight evidential heads that map hidden states to 4-dimensional Dirichlet parameters using a Softplus activation. We utilize the Digamma loss with KL divergence regularization, applying an annealing schedule over the first $2,000$ steps for stability. The balancing coefficients λ_{rag} and λ_{param} are set to 0.1, The IDK-ratio hyperparameter is set to 0.7 for the training distribution. Additionally, a standard SFT auxiliary loss with a coefficient of 1.0 is incorporated to maintain generation quality. All experiments are performed on 4 NVIDIA A100 GPUs using PyTorch 2.9.0.

Quadrant Annotation To classify each query as Known or Unknown, we employ a highly conservative consistency threshold of δ=1.0 using N=10 sampled responses generated without retrieval. A query is considered Known only if all 10 samples are correct. This strict criterion minimizes false positives (labeling unknown queries as known) and favors a safer failure mode (false negatives), preventing the introduction of hallucination-prone training signals. To verify whether the retrieved context contains the gold answer, we employ GPT-4o as a judge (Figure 7). As validated in DTA (Sun et al., 2025), the GPT-4o-based assessment achieves 93% agreement with human labeling, and the lexical matching protocol yields approximately 90% consistency with human evaluation.

Wiki Event Construction We parse the "Events" sections of Wikipedia annual pages from 2010 to 2025—targeting information that typically post-dates standard LLM pre-training cutoffs—and scrape referenced external news articles to serve as gold contexts. We leverage GPT-4o to generate factual, context-dependent QA pairs directly from these documents. Each query is paired with a stratified pool of 10 candidate contexts: (1) one Golden Context containing the ground truth; (2) four Hard Negatives, synthesized by using GPT-4o to systematically perturb key entities in the gold context, creating counterfactuals with high keyword overlap but incorrect answers; and (3) five Easy Negatives randomly sampled from the global corpus to represent irrelevant noise. This hierarchical structure enables granular assessment of the model's ability to discern correct evidence amidst varying degrees of semantic interference.

> You are an expert at evaluating whether a context contains the correct answer to a question. You should:
>
> 1. Check if the given answer can be found or directly implied by the context
> 2. Return a score of 1 if the context contains or directly implies the answer
> 3. Return a score of 0 if the context does not contain or support the answer
> 4. Provide a brief explanation for your decision
>
> Respond in the following JSON format:
> {
> "score": 0 or 1,
> "explanation": "your explanation here"
> }

Figure 7: Context Evaluation Prompt

Evaluation Metrics. Table 2 summarizes our evaluation framework, which spans three complementary axes. We define $\mathcal{S}_{\text{ans}}$ and $\mathcal{S}_{\text{abs}}$ as the sets of queries where the model decides to answer or abstain, respectively, and intersect these prediction sets with ground-truth partitions to isolate specific behaviors: answerable queries ($\mathcal{I}_{\text{ans}} = \mathcal{K}_{\text{rag}}$), gold-only contexts ($\mathcal{I}_{\text{gold}} = \mathcal{K}_{\text{retr}} \setminus \mathcal{K}_{\text{param}}$), and noisy contexts ($\mathcal{I}_{\text{noise}} = \mathcal{K}_{\text{param}} \setminus \mathcal{K}_{\text{retr}}$). Answer correctness c_i is assessed via lexical matching, following prior work (Sun et al., 2025).

Answer Quality measures the model's ability to answer correctly when it chooses to respond: Recall is the fraction of answerable queries answered correctly, Precision is the fraction of attempted answers that are correct, and their harmonic mean gives the Answer F1. *Retrieval Handling* further decomposes this ability by context type: Denoise Rate measures accuracy on noisy-context queries (where the model must rely on parametric knowledge), while Context Utilization measures accuracy on gold-only-context queries (where the model must extract the answer from retrieval). *Abstain Quality* mirrors Answer Quality for the refusal deci-

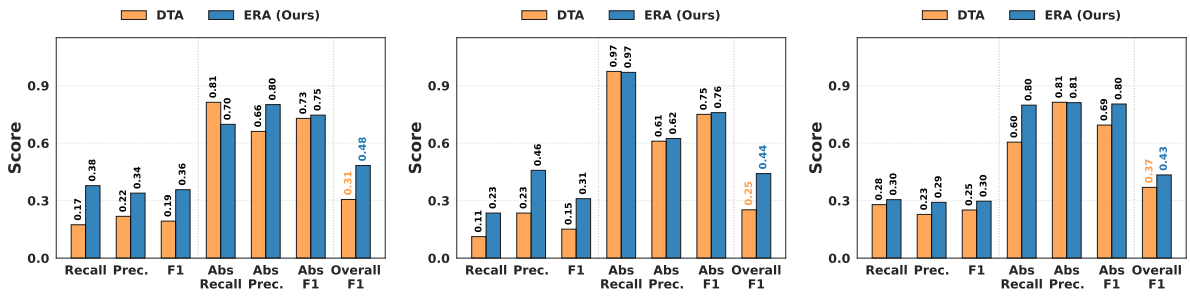
Table 2: Summary of evaluation metrics categorized by Answer Quality, Retrieval Handling, and Abstain Quality, including the Overall F1 for trade-off analysis.

Category	Metric	Formula	Description
Answer Quality	Recall	$\frac{\sum_{i \in \mathcal{S}_{\text{ans}} \cap \mathcal{I}_{\text{ans}}} c_i}{\lvert \mathcal{I}_{\text{ans}} \rvert}$	Ratio of correct answers to all answerable queries
	Precision	$\frac{\sum_{i \in \mathcal{S}_{\text{ans}}} c_i}{\lvert \mathcal{S}_{\text{ans}} \rvert}$	Ratio of correct answers to all attempted answers
	F1	$2 \cdot \frac{\text{Prec} \cdot \text{Rec}}{\text{Prec} + \text{Rec}}$	Harmonic mean of Precision and Recall
Retrieval Handling	Denoise Rate	$\frac{\sum_{i \in \mathcal{S}_{\text{ans}} \cap \mathcal{I}_{\text{noise}}} c_i}{\lvert \mathcal{I}_{\text{noise}} \rvert}$	Accuracy on queries with noisy context
	Context Util.	$\frac{\sum_{i \in \mathcal{S}_{\text{ans}} \cap \mathcal{I}_{\text{gold}}} c_i}{\lvert \mathcal{I}_{\text{gold}} \rvert}$	Accuracy on queries with golden context
Abstain Quality	Abstain Recall	$\frac{\lvert \mathcal{S}_{\text{abs}} \cap \mathcal{I}_{\text{unans}} \rvert}{\lvert \mathcal{I}_{\text{unans}} \rvert}$	Ratio of correct abstentions to all unanswerable queries
	Abstain Precision	$\frac{\lvert \mathcal{S}_{\text{abs}} \cap \mathcal{I}_{\text{unans}} \rvert}{\lvert \mathcal{S}_{\text{abs}} \rvert}$	Ratio of correct abstentions to all abstained queries
	Abstain F1	$2 \cdot \frac{\text{Prec}_{\text{abs}} \cdot \text{Rec}_{\text{abs}}}{\text{Prec}_{\text{abs}} + \text{Rec}_{\text{abs}}}$	Harmonic mean of Abstain Precision and Recall
Overall	Overall F1	$2 \cdot \frac{\text{F1} \cdot \text{F1}_{\text{abs}}}{\text{F1} + \text{F1}_{\text{abs}}}$	Trade-off between Answer Quality and Abstain Quality

sion: Abstain Recall is the fraction of unanswerable queries correctly refused, Abstain Precision is the fraction of refusals that are justified, and their harmonic mean gives the Abstain F1. Finally, the *Overall F1*—the harmonic mean of Answer F1 and Abstain F1—captures the central trade-off between utility (answering) and reliability (abstaining), serving as our primary metric for honest RAG.

Baselines. We implement 8 calibration strategies as baselines. For inference-time baselines, we report the configuration that achieves the highest Overall F1 score in the main table.

- **Logprob:** Calculates the average log-probability of all tokens in the generated answer, serving as a basic confidence metric. We trigger an abstention if this average log-probability falls below a threshold $\tau \in \{-1, -2\}$.
- **ICL:** Prompts the model to verbally self-evaluate the correctness of its answer (Wei et al., 2022). We utilize 3-shot examples generated at a temperature of 0.3, top-k of 50, and top-p of 0.6, determining abstention directly via string matching against predefined refusal phrases without requiring a numerical threshold.
- **Self-Consistency:** Computes confidence based on the agreement rate among multiple sampled outputs for the same query (Wang et al., 2023). We evaluate 10 samples per query at a temperature of 0.7, measuring confidence through a pairwise exact-match agreement score and abstaining when the rate falls below a fixed threshold of 0.5.
- **P_{true}:** Employs a binary verification prompt, providing a continuous confidence score by extracting the specific softmax probability assigned to the "True" token (Kadavath et al., 2022). This is computed using the softmax distribution over A/B tokens with a temperature of 0.0 and a maximum generation length of 1 token, triggering an abstention if the probability falls below a threshold $\tau \in \{0.5, 0.7, 0.9\}$.
- **RAUQ:** Leverages an unsupervised uncertainty metric derived from attention entropy, computing per-sample uncertainty over the top 10% most discriminative attention heads (Vazhentsev et al., 2025). The model abstains when this aggregated uncertainty score exceeds a threshold $\tau \in \{0.1, 0.2, 0.3\}$.
- **AdaCAD:** Dynamically adjusts decoding weights based on the measured conflict between parametric and contextual knowledge to prevent over-correction (Wang et al., 2025a). We detect this conflict by computing the Jensen–Shannon (JS) divergence between probability sequences from a RAG prompt



(a) Qwen3-8B-Base (b) Llama3.1-8B (c) ChatQA

Figure 8: Generalizability evaluation on the Wiki Event dataset.

and a context-free parametric prompt, triggering abstention when the divergence exceeds a specified threshold $\tau \in \{0.5, 0.7, 0.9\}$.

- **TruthRL:** Employs reinforcement learning with a ternary reward system to penalize hallucinations and explicitly incentivize the model to abstain when uncertain (Wei et al., 2025). We fine-tune the policy via GRPO (rewards: $+1$ for correct, 0 for abstained, -1 for hallucinated), allowing the model to intrinsically output refusal phrases on knowledge-gap queries without relying on an external inference-time threshold.

- **DTA:** Employs DPO to train models to abstain from answering when retrieved evidence is insufficient, using curated abstention datasets (Sun et al., 2025). Similar to TruthRL, it learns internal decision boundaries for selective refusal, bypassing the need for manual threshold tuning.

B Further Results on Honest Performance

Due to space limitations in the main manuscript, we present additional results here.

Generalization Analysis. Figure 8 demonstrates that ERA significantly outperforms the strongest baseline, DTA, on the out-of-distribution (OOD) Wiki Event dataset by achieving a superior Overall F1 score. While DTA struggles with severe over-abstention—indiscriminately rejecting valid In-Distribution (ID) queries due to overfitting on refusal patterns—ERA leverages disentangled uncertainty estimates to establish a precise decision boundary. This boundary effectively filters out OOD noise without compromising Answer F1, validating its robust generalization capability across mixed distributions. Furthermore, this advantage seamlessly extends to models explicitly optimized for Retrieval Augmented Fine-Tuning (RAFT), such as ChatQA, where integrating ERA further refines their ability to discriminate between relevant and irrelevant contexts, yielding substantial performance improvements.

Ablation Study. We conduct an extensive ablation study to evaluate the contribution of each component in our full model (Table 3). To systematically analyze our framework, we categorize the ablated variations into three key dimensions: (1) core learning objectives and architecture, (2) belief representation and fusion structure, and (3) conflict gating strategies.

First, to evaluate the core architecture and learning objectives, we compare the following four variations: (1) **w/o Dual** removes the parametric memory head, relying solely on RAG context and thereby bypassing the internal-external conflict detection mechanism; (2) **Learnable W** replaces the evidence-based Dempster-Shafer fusion with a simple learnable scalar gate to weight the loss; (3) **CE Only** substitutes the evidential learning objective with standard cross-entropy loss, ignoring explicit uncertainty quantification; and (4) **w/o KL** omits the KL divergence regularization term from the evidential loss, utilizing only the data fit term.

Second, we investigate the impact of dimensional mapping on evidence fusion. The structural alignment of the two pathways prior to DST fusion is critical, as it determines whether each head can express sufficiently rich evidence to support reliable conflict detection. We compare our **4D→2D** framework against three structural variants.

The **Asymmetric (RAG 4D, Param 2D)** variant assumes only the RAG head requires full 4D granularity while the parametric head directly predicts over a simplified 2D space; the RAG head is then

Table 3: Performance comparison of ablation studies. The consistent degradation in Overall F1 scores observed when removing or altering individual components validates the distinct contribution of each module in the ERA framework.

Qwen3-8B-Base	Recall	Precision	F1	Abstain Recall	Abstain Precision	Abstain F1	Overall F1
w/o KL	0.598	0.471	0.527	0.257	0.778	0.386	0.446
CE Only	0.587	0.455	0.513	0.205	0.746	0.321	0.395
w/o Dual	0.600	0.457	0.519	0.172	0.783	0.282	0.366
Learnable W	0.484	0.353	0.409	0.060	0.754	0.112	0.176
Asymmetric	0.600	0.475	0.530	0.242	0.703	0.360	0.450
2d → 2d	0.371	0.318	0.343	0.659	**0.814**	0.728	0.466
4d → 4d	0.358	0.363	0.360	**0.775**	0.762	**0.768**	0.491
Product	0.348	0.354	0.351	0.762	0.745	0.753	0.479
Max	0.364	0.350	0.357	0.748	0.783	0.765	0.487
Ours	**0.622**	**0.532**	**0.573**	0.408	0.715	0.520	**0.545**

projected down to 2D at the fusion stage to match the parametric head's dimensionality. This design underperforms because the parametric head is deprived of the capacity to express nuanced, query-conditioned priors. Although f_{param} does not explicitly observe the retrieved context, the query itself carries implicit signals regarding likely retrieval quality—rare or highly specific queries tend to yield noisy retrieval, whereas common factual queries are more likely to be well-supported. Training over all four quadrants allows f_{param} to capture these query-conditioned priors, which a 2D formulation cannot represent.

The **2D→2D** variant collapses both pathways too early and loses the fine-grained taxonomy needed to detect conflicts. The **4D→4D** variant preserves all quadrant interactions during fusion but dilutes the conflict score κ with superfluous, semantically noisy cross-products. In contrast, our symmetrical **4D→2D** framework allows both pathways to reason over the full 4D epistemic taxonomy, projecting them to 2D simultaneously prior to fusion. This structural constraint forces the final conflict computation to isolate a single critical dimension—whether the internal and external sources agree on answering the query—without sacrificing the expressive capacity of either head.

Finally, we explore alternative conflict gating formulations. We conducted ablations with two alternative gate formulations incorporating parametric uncertainty: (1) **Product gate** ($g = 1 - u_{\text{rag}} \cdot u_{\text{param}}$), which amplifies the signal only when both sources are confident; and (2) **Max gate** ($g = 1 - \max(u_{\text{rag}}, u_{\text{param}})$), which amplifies when either source is confident. The key insight lies in the UN (Unknown-Noisy) quadrant, where the model must learn to abstain. Here, parametric uncertainty is inherently high. The product gate suppresses the update in this critical scenario, weakening the abstention signal when it is most needed. The max gate partially mitigates this but introduces noise by amplifying the signal when only the parametric source is confident while retrieval is unreliable. Our retrieval-only gate avoids both failure modes: it preserves strong learning pressure in the UN quadrant while suppressing updates only when retrieval itself is genuinely uncertain. This design aligns with the core principle of RAG: the retrieval source is the primary signal to be validated, and the conflict score explicitly encodes parametric disagreement.