A MULTIDIMENSIONAL VIEW OF LLM UNCERTAINTY QUANTIFICATION VIA SCENARIO INTERVENTIONS

Anonymous authors
Paper under double-blind review

ABSTRACT

Reliable uncertainty quantification for large language models should capture not only whether an answer changes, but whether it changes appropriately when the supporting evidence is altered. We propose Response–Evidence Coupling (REC), a multidimensional framework that estimates answer uncertainty from two coupled dimensions: scenario-conditioned response behavior and scenario-specific evidence support. Starting from an anchor answer, REC constructs evidence-preserving equivalent scenarios and counterfactual replacement scenarios that modify answer-bearing facts. It then evaluates whether the model follows the task-dependent expected response under each intervention and whether the generated answer is supported by the corresponding scenario. These signals are organized into anchor, equivalent-intervention, and counterfactual-intervention risks and aggregated through rank-normalized Fisher fusion. Across four QA benchmarks and four LLMs, REC achieves the best AUROC and AUARC in most settings. Ablation results further show that intervention-aware response behavior and scenario-specific evidence support provide complementary information, supporting a multidimensional view of LLM uncertainty.

1 INTRODUCTION

Large language models (LLMs) can generate fluent and plausible answers that are nevertheless factually unreliable (Manakul et al., 2023; Min et al., 2023; Chen & Mueller, 2024; Sriramanan et al., 2024; Farquhar et al., 2024; Huang et al., 2025). Estimating answer uncertainty is therefore essential for building reliable LLM systems, especially when model outputs are used for selective generation, abstention, adaptive inference, or downstream decision making (Gawlikowski et al., 2023; Ren et al., 2023; Qiu & Miikkulainen, 2024; Chen et al., 2025; Murphy, 2022).

A major line of current uncertainty estimation methods measures reliability from response variation. These methods sample multiple responses, construct semantically diverse generations, or perturb the input, and then estimate uncertainty from answer disagreement, semantic diversity, or instability across variants (Manakul et al., 2023; Kuhn et al., 2023; Farquhar et al., 2024; Nikitin et al., 2024; Kunitomo-Jacquin et al., 2026; Lin et al., 2024b; Aichberger et al., 2025; Da et al., 2025). A common form of input perturbation is to construct equivalent rewrites or semantics-preserving variants and then test whether the answer remains stable across them (Ribeiro et al., 2020). This line of work captures response-side uncertainty, but such a single-signal view is incomplete.

On the one hand, answer invariance under repeated sampling or equivalent rewriting captures only part of intervention consistency. Beyond answer-preserving variants, we can also introduce other scenario interventions, such as replacing key evidence in the context. In this case, the expected answer behavior becomes more nuanced. For factual or open-domain QA, the question itself determines the real-world answer, so the answer should remain unchanged even if key evidence in a pseudo-context is modified. For context-grounded QA, however, the context defines the answer, so replacing answer-bearing facts should lead to a corresponding answer change. In other words, a reliable answer should exhibit expected-consistent behavior under different types of scenario interventions.

On the other hand, answer change itself is only a response-side signal and should be interpreted together with the extent to which the intervened context supports the current answer. If the answer changes in the expected direction but the new answer is not supported by the current scenario evidence,

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

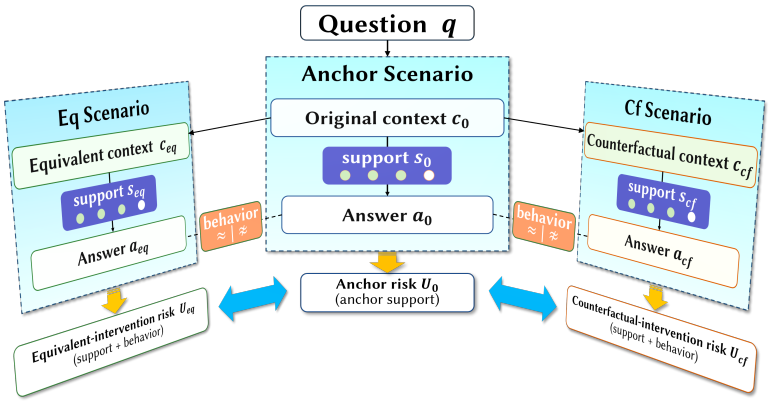


Figure 1: Overview of Response–Evidence Coupling (REC).

such a change may still reflect unsupported drift and thus weaken reliability. Therefore, uncertainty should not be estimated only from whether the answer changes, but from whether the answer behavior follows the expected intervention direction and whether the answer is supported by the evidence in the current scenario.

Based on this insight, we propose *Response–Evidence Coupling* (REC), a scenario-intervention framework for LLM uncertainty estimation. As illustrated in Figure 1, REC first obtains an anchor answer and then constructs two types of intervention scenarios. Equivalent scenarios preserve answer-bearing facts, while counterfactual replacement scenarios modify answer-bearing facts. REC then tests whether the model's answer behavior under different interventions matches the expected behavior. REC evaluates evidence support within each scenario, so that answer behavior is interpreted together with the scenario evidence that supports or fails to support it. The resulting signals are organized into anchor risk, equivalent-intervention risk, and counterfactual-intervention risk, which are fused into the final uncertainty score. Experiments on multiple question-answering benchmarks and mainstream LLMs show that REC outperforms methods based on single-scenario perturbation or single-dimensional uncertainty signals. Our main contributions are as follows:

- We show that answer invariance under repeated sampling or equivalent rewriting captures only part of intervention consistency: reliable answers should exhibit expected-consistent behavior under different types of scenario interventions.
- We propose Response–Evidence Coupling (REC), a scenario-intervention framework that integrates multiple intervention scenarios and multiple uncertainty dimensions by jointly modeling answer behavior and evidence support.
- We conduct extensive experiments on multiple question-answering benchmarks and mainstream LLMs, showing that REC outperforms methods based on single-scenario perturbation or single-dimensional uncertainty signals.

2 Related Work

One line of work estimates uncertainty from signals observed in the original generation scenario, such as model confidence, probability structure, or internal representations. Self-evaluation and verbalized-confidence methods ask the model to judge whether its answer is correct or trustworthy, supporting confidence calibration, knowledge-boundary detection, and selective generation (Lin et al., 2022; Kadavath et al., 2022; Tian et al., 2023; Zhao et al., 2024; Ren et al., 2023; Chen et al., 2025). Likelihood-based approaches use token- or sequence-level probabilities as confidence signals, with recent work refining these scores through contextualized likelihood, semantic token weighting, or semantic-space confidence estimation (Murray & Chiang, 2018; Malinin & Gales, 2020; Lin et al., 2024a; Duan et al., 2024; Bakman et al., 2024; Qiu & Miikkulainen, 2024). White-box and auxiliary-probe methods further exploit logits, hidden states, embeddings, or internal feature gaps to estimate uncertainty or detect hallucination (Chen et al., 2024; Kossen et al., 2024; Vu et al., 2025; Sriramanan et al., 2024; Bakman et al., 2025).

Another important direction estimates uncertainty by examining how model answers vary across repeated generations, perturbed inputs, or semantically related variants. Perturbation-based approaches explicitly probe uncertainty through variations in the input or generation conditions: SPUQ (Gao et al., 2024) aggregates outputs across prompt and sampling perturbations, Inv-Entropy (Song et al., 2025) formulates perturbation-based uncertainty through a probabilistic inverse-modeling framework. Semantic uncertainty methods group generations by meaning and estimate uncertainty from semantic diversity rather than surface-form variation (Kuhn et al., 2023; Farquhar et al., 2024). Subsequent methods refine this idea through pairwise semantic similarity, unobserved answer mass, adaptive sampling budgets, or semantically diverse generation (Nikitin et al., 2024; Kunitomo-Jacquin et al., 2026; Sun et al., 2026; Aichberger et al., 2025). SelfCheckGPT and related black-box methods similarly use agreement, factual consistency, or response dispersion across sampled outputs to detect unreliable generations (Manakul et al., 2023; Chen & Mueller, 2024; Lin et al., 2024b).

REC builds on these two lines of work, but simultaneously evaluates two coupled aspects of uncertainty: whether answer behavior under explicit scenario interventions matches the expected behavior, and whether the answer is supported by the corresponding scenario evidence. By integrating equivalent scenarios, counterfactual replacement scenarios, answer behavior, and evidence support, REC moves beyond single-scenario scoring and single-signal consistency estimation.

3 METHOD

We propose **Response–Evidence Coupling** (**REC**) for estimating the uncertainty of large language model answers. Given a question q, REC first constructs an anchor scenario c_0 and generates an anchor answer a_0. It then constructs two types of intervention scenarios: equivalent scenarios and counterfactual replacement scenarios. Equivalent scenarios preserve answer-relevant evidence. Replacement scenarios modify answer-bearing facts in the context.

REC uses five primitive risk signals: anchor support risk u_{w_0}, equivalent answer drift u_{drift}, equivalent evidence-gap risk $u_{\text{eq-gap}}$, counterfactual behavior risk $u_{\text{cf-behavior}}$, and counterfactual evidence-gap risk $u_{\text{cf-gap}}$. These signals are organized into three risk groups: anchor risk, equivalent-intervention risk, and counterfactual-intervention risk.

3.1 ANCHOR SCENARIO CONSTRUCTION

For each question q, REC defines an anchor scenario c_0 and asks the target language model G_{ans} to generate the anchor answer $a_0 = G_{\mathrm{ans}}(q, c_0)$. For context-grounded QA, c_0 is the original input context. For no-context QA, REC constructs a question-only pseudo-context using the original answer model G_{ans}. Since the purpose of the pseudo-context is to anchor the model's original belief instead of introducing new evidence that may change the answer, using G_{ans} helps keep the pseudo-context aligned with the model's own parametric knowledge. The pseudo-context is generated without conditioning on the candidate answer, thereby avoiding answer-conditioned evidence that trivially supports the model's own answer. The resulting anchor triplet (q, c_0, a_0) serves as the reference point for support estimation, intervention construction, answer comparison, and risk aggregation. The answer-generation prompt and task-specific instructions are provided in Appendix B.1, and the prompt used for pseudo-context generation is provided in Appendix B.2.

3.2 INTERVENTION SCENARIO CONSTRUCTION

Figure 2 illustrates the overall construction process. REC intervenes on the evidence context while keeping the target question fixed. The scenario set consists of the anchor scenario, equivalent scenarios $\mathcal{S}_{\mathrm{eq}}$, and counterfactual replacement scenarios $\mathcal{S}_{\mathrm{cf}}$.

Anchor evidence extraction. After generating the anchor answer, REC uses a lightweight evidence scorer to estimate whether the anchor context supports the answer and to extract supporting evidence spans, denoted as

$$\mathcal{E}_0 = \mathrm{SupportEvidence}(q, c_0, a_0). \tag{1}$$

The prompt used for anchor evidence extraction is provided in Appendix C.1. These spans characterize the main factual path from the question to the anchor answer and provide an interpretable description of the anchor evidence.

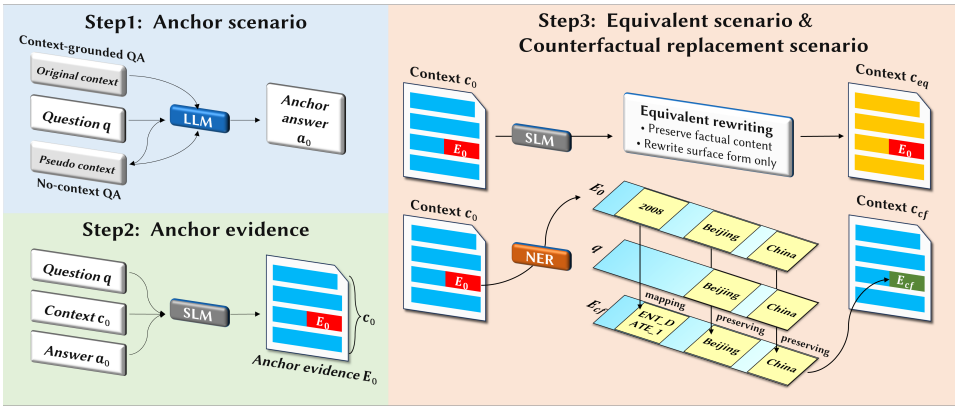


Figure 2: Overview of intervention scenario construction in REC.

Equivalent scenarios. Equivalent scenarios are passage-level semantics-preserving rewrites of the anchor context. REC uses a small language model (SLM) to generate

$$c_i^{\text{eq}} = G_{\text{rew}}(c_0), \tag{2}$$

where G_{rew} rewrites the complete anchor context without conditioning on the question or anchor answer. It is instructed to preserve all factual content, including potentially answer-relevant facts, while varying surface wording or discourse structure. This context-only construction also allows examples sharing the same passage to reuse an identical equivalent rewrite through the passage-level cache. The prompt is provided in Appendix D.1.

Counterfactual replacement scenarios. Counterfactual scenarios are constructed by systematically neutralizing high-value key facts in the context while keeping the question and its explicitly mentioned anchors unchanged. Directly generating such scenarios with a small language model is less reliable than equivalent rewriting, because the model may introduce uncontrolled edits or fail to modify salient factual content. It also incurs additional time and token costs because the generated context must preserve the surrounding textual structure. REC therefore adopts a controlled deterministic replacement strategy that uses a lightweight GLiNER-based span detector to identify high-value entity and key-fact spans throughout the anchor context while excluding candidates protected by question-side anchors. After resolving overlapping or duplicate candidates, all remaining eligible spans are replaced with type-aware placeholders.

Let $\mathcal{M}(q, c_0, a_0) = \{(m_j, \tau_j)\}_{j=1}^{J}$ denote the non-overlapping key-fact spans in c_0 and their inferred types after removing question-protected candidates. The counterfactual context is constructed as

$$c_i^{\text{cf}} = \text{ReplaceAll}\left(c_0, \left\{m_j \to \text{ENT}_{\tau_j}\right\}_{j=1}^{J}\right). \tag{3}$$

Repeated or related mentions are replaced consistently when required, while the question itself remains unchanged. The complete deterministic replacement procedure is provided in Appendix D.2. For each intervention scenario $c_i \in \mathcal{S}_{\text{eq}} \cup \mathcal{S}_{\text{cf}}$, the target model generates a scenario-conditioned answer $a_i = G_{\text{ans}}(q, c_i)$.

Task-aware expected direction. The expected answer-change direction $e_i \in \{0, 1\}$ is task-aware:

$$e_i = \begin{cases} 0, & c_i \in \mathcal{S}_{\text{eq}}, \\ 1, & c_i \in \mathcal{S}_{\text{cf}} \text{ and the task is context-grounded}, \\ 0, & c_i \in \mathcal{S}_{\text{cf}} \text{ and the task is open-domain}. \end{cases} \tag{4}$$

Thus, equivalent evidence should preserve the answer; replacement evidence should change it when context defines the answer, but not when the question fixes a real-world referent.

3.3 Answer Agreement and Evidence Support

REC computes two basic quantities for each intervention scenario: answer agreement and evidence support. Let $g_i = \text{Agree}(a_0, a_i) \in [0, 1]$ be the semantic agreement between the anchor answer and

the intervention answer. The prompt and deterministic preprocessing used for agreement judgment are provided in Appendix E.1. The observed answer change is $d_i = 1 - g_i$, and the response-behavior mismatch is $b_i = |d_i - e_i|$. Thus, b_i is high when the model changes its answer under an answer-preserving intervention or fails to change its answer when a context-grounded counterfactual intervention should induce an answer change.

REC also estimates a support score $s_i \in [0, 1]$ for each scenario-answer pair (c_i, a_i), measuring whether the current scenario supports the answer generated under that scenario. The prompt used for support scoring is provided in Appendix C.2. The anchor support score is s_0, and the anchor support risk is $u_{w_0} = 1 - s_0$.

To measure whether an intervention has sufficient evidential grounding, REC computes an intervention evidential mass:

$$r_i = q_i \eta_i, \qquad q_i = \beta_i s_0 + (1 - \beta_i) s_i, \qquad \eta_i = |2e_i - 1|. \tag{5}$$

Here, q_i is a type-dependent support quality that combines anchor support and intervention support, and η_i is the identifiability of the expected direction. In the equivalent and replacement settings considered here, $e_i \in \{0, 1\}$, so $\eta_i = 1$. The coefficient β_i controls how much the support quality relies on the anchor scenario versus the intervention scenario.

3.4 Risk Construction

REC constructs uncertainty from three complementary risk groups: anchor risk U_{anc}, equivalent-intervention risk U_{eq}, and counterfactual-intervention risk U_{cf}. The anchor group measures whether the original answer is reliable before any intervention is applied. It is directly given by $U_{\text{anc}} = u_{w_0} = 1 - s_0$.

Equivalent and counterfactual replacement scenarios both contribute a behavior risk and an evidence-gap risk. Equivalent scenarios preserve key answer evidence, so their behavior risk corresponds to answer drift. Counterfactual replacement scenarios test whether the model follows the task-aware expected behavior after answer-bearing context facts are replaced. We define

$$\begin{aligned} u_{\text{drift}} &= \frac{1}{|\mathcal{S}_{\text{eq}}|} \sum_{c_i \in \mathcal{S}_{\text{eq}}} b_i, & u_{\text{cf-behavior}} &= \frac{1}{|\mathcal{S}_{\text{cf}}|} \sum_{c_i \in \mathcal{S}_{\text{cf}}} b_i, \\ u_{\text{eq-gap}} &= 1 - \text{Agg}(\{r_i : c_i \in \mathcal{S}_{\text{eq}}\}), & u_{\text{cf-gap}} &= 1 - \text{Agg}(\{r_i : c_i \in \mathcal{S}_{\text{cf}}\}), \\ U_{\text{eq}} &= \text{Fisher}(u_{\text{drift}}, u_{\text{eq-gap}}), & U_{\text{cf}} &= \text{Fisher}(u_{\text{cf-behavior}}, u_{\text{cf-gap}}). \end{aligned} \tag{6}$$

Since $e_i = 0$ for equivalent scenarios, u_{drift} reduces to the average answer-change score $1 - g_i$. For context-grounded tasks, the counterfactual behavior risk is high when the model keeps the original answer after the context has been edited. For factual or open-domain tasks, it is high when the model drifts away from the question-determined answer under a context-only pseudo-counterfactual. The evidence-gap terms are high when the corresponding intervention scenarios have weak evidential grounding. In our experiments, $\text{Agg}(\cdot)$ is mean aggregation.

3.5 Risk Fusion and Final Score

For risks $u_1, \ldots, u_K \in [0, 1]$, REC uses the monotonic Fisher operator

$$\text{Fisher}(u_1, \ldots, u_K) = F_{\chi^2_{2K}}\left(-2 \sum_{k=1}^{K} \log(1 - u_k + \epsilon)\right), \qquad \epsilon = 10^{-6}, \tag{7}$$

which emphasizes jointly occurring risks while retaining evidence from multiple sources. Because group-score scales may differ across tasks, each score is first converted to a task-wise rank percentile,

$$R(z_j) = \frac{\text{rank}(z_j) - 1}{N - 1}, \tag{8}$$

using average ranks for ties. The final score is

$$\begin{aligned} u_{\text{REC}} &= \text{Fisher}\big(R(U_{\text{anc}}), R(U_{\text{eq}}), R(U_{\text{cf}})\big) \\ &= \text{Fisher}\big(R(u_{w_0}), R(\text{Fisher}(u_{\text{drift}}, u_{\text{eq-gap}})), R(\text{Fisher}(u_{\text{cf-behavior}}, u_{\text{cf-gap}}))\big), \end{aligned} \tag{9}$$

where larger values indicate greater answer uncertainty.

4 EXPERIMENTAL SETUP

4.1 DATASETS, BASELINES AND EVALUATION METRICS

Our evaluation covers four QA benchmarks: CoQA (Reddy et al., 2019), HotpotQA (Yang et al., 2018), TriviaQA (Joshi et al., 2017), and Natural Questions (NQ) (Kwiatkowski et al., 2019). CoQA and HotpotQA are treated as context-grounded tasks, whereas TriviaQA and NQ are treated as open-domain factoid QA tasks. For the latter, REC first constructs a question-only pseudo-scenario and then derives equivalent and replacement scenarios from it. The first 500 examples from the corresponding evaluation split of each dataset are used.

The comparison includes P(True) (Kadavath et al., 2022) and SAR (Duan et al., 2024), which estimate uncertainty from self-evaluation or likelihood-based signals in the original generation setting. We also compare with Semantic Entropy (Kuhn et al., 2023; Farquhar et al., 2024), EigenScore (Chen et al., 2024), $\mathrm{SE}_{\mathrm{SDLG}}$ (Aichberger et al., 2025), LexSim (Lin et al., 2024b), Deg (Lin et al., 2024b), and SD (Lin et al., 2024b), which capture response variation, semantic diversity, or representation geometry across multiple generations. Finally, SPUQ (Gao et al., 2024) and Inv-Entropy (Song et al., 2025) are included as input-perturbation baselines that estimate uncertainty from model behavior under perturbed inputs or generation conditions.

Following prior work, uncertainty estimation is evaluated as both an error-detection and a selective-prediction problem. Each method produces an uncertainty score $u \in [0, 1]$, with larger values indicating greater predicted uncertainty. Performance is measured using AUROC (Fawcett, 2006) and AUARC (Nadeem et al., 2009). AUROC measures whether incorrect answers are assigned higher uncertainty than correct ones, whereas AUARC evaluates selective prediction by progressively rejecting high-uncertainty answers and measuring retained-answer accuracy across coverage levels.

4.2 IMPLEMENTATION DETAILS

We evaluate four LLMs: Llama-3.1-8B-Instruct, Llama-3.1-70B-Instruct (Dubey et al., 2024), Qwen3-14B, and Qwen3-32B (Yang et al., 2025). For each evaluated model, the same G_{ans} is used for answer generation and, for no-context QA, question-only pseudo-context generation, ensuring alignment with its parametric knowledge.

REC uses lightweight auxiliary components for scenario construction. Qwen3-1.7B performs anchor evidence extraction and equivalent rewriting, while `gliner-large-v2.1` detects entity spans for type-aware counterfactual substitution. DeepSeek-V4-Flash (DeepSeek-AI, 2026) is used only for support scoring and semantic agreement judgment, answer correctness is independently judged by GPT-5.5. The correctness prompt is provided in Appendix E.2. We additionally calibrate the correctness judge against human annotations and evaluate the validity of the constructed intervention scenarios; the results are reported in Appendices A.1 and A.2, respectively. Each example uses one equivalent and one replacement scenario.

In Eq. 5, we set $\beta_i = 0.5$ for equivalent scenarios to balance semantically consistent anchor and intervened evidence, and $\beta_i = 0.25$ for replacement scenarios to emphasize evidence changes and the model's response to them. Answer generation, support and agreement judgment, and correctness judgment use temperature 0; equivalent rewriting uses temperature 0.4, while sampling-based baselines use temperature 0.8 with $K = 5$ responses. REC caches passage-level equivalent rewrites and GLiNER span candidates. All prompts are provided in the appendices, and repeatable experiments are run three times with mean and standard deviation reported.

5 RESULTS AND ANALYSIS

5.1 MAIN RESULTS

Tables 1 and 2 report the main results under AUROC and AUARC. REC achieves the best AUROC in most dataset–model settings and the best AUARC in nearly all settings, consistently outperforming most non-intervention and perturbation-based baselines. Non-intervention methods use confidence signals from the original generation setting, while perturbation-based methods capture response variation, semantic diversity, or representation geometry. However, neither group explicitly tests

Table 1: Uncertainty estimation results measured by AUROC. Values are reported as percentages, mean ± std. Best results are bolded and second-best results are underlined.

Dataset	Model	P(True)	SAR	SE	Eigen Score	SE_{SDLG}	LexSim	Deg	SD	SPUQ	Inv-Entropy	REC
HotpotQA	Llama3.1-8B	60.6±1.2	71.5±1.3	70.6±1.2	69.8±1.2	72.3±0.8	66.7±1.1	70.1±1.0	71.7±0.7	71.4±1.2	72.8±0.9	**85.5±0.8**
	Llama3.1-70B	65.6±1.1	61.3±1.3	61.6±1.2	61.1±1.2	73.5±0.9	57.9±1.0	60.6±0.8	63.2±0.7	62.0±1.1	63.3±1.0	**74.2±1.1**
	Qwen3-14B	60.1±0.9	65.2±1.6	64.8±1.5	62.8±0.7	**75.5±0.4**	60.3±0.7	64.2±1.1	65.8±0.8	65.6±1.4	66.9±1.3	74.6±0.7
	Qwen3-32B	68.5±0.8	57.9±1.3	58.2±1.2	57.2±0.9	70.1±0.5	54.8±1.1	57.5±1.1	58.6±0.7	58.9±0.9	60.2±1.4	**75.2±1.3**
CoQA	Llama3.1-8B	56.8±1.4	76.5±1.2	75.6±1.0	76.3±0.6	67.2±0.5	64.4±1.2	69.5±1.3	65.0±0.6	66.0±1.0	68.0±1.2	**81.9±0.9**
	Llama3.1-70B	67.0±0.9	60.5±0.4	60.8±0.6	61.4±0.7	62.5±1.0	59.7±0.7	**72.9±1.2**	61.8±0.5	69.6±1.4	71.6±1.1	71.7±0.6
	Qwen3-14B	60.1±1.5	64.5±0.6	64.1±0.9	64.9±1.0	64.8±0.4	62.2±0.8	66.4±1.1	65.7±1.0	63.1±0.8	64.5±0.9	**77.1±1.2**
	Qwen3-32B	64.8±1.2	62.9±0.9	63.2±1.2	64.0±1.2	65.6±0.8	61.7±1.0	65.7±1.4	64.6±0.5	62.6±1.7	63.6±1.5	**78.9±0.8**
TriviaQA	Llama3.1-8B	71.4±0.8	76.7±1.1	76.6±1.4	80.6±1.1	78.6±0.6	78.5±1.1	79.6±1.1	76.6±0.9	78.0±1.2	80.2±1.4	**89.1±0.4**
	Llama3.1-70B	65.4±1.7	66.1±0.8	66.3±1.1	60.3±0.7	78.5±0.6	58.3±1.3	76.4±1.4	67.8±0.6	74.9±1.5	77.1±1.7	**85.2±1.5**
	Qwen3-14B	68.2±0.9	69.9±1.0	70.4±0.9	66.4±1.2	82.1±1.0	62.8±1.2	70.1±1.0	71.6±0.8	68.8±0.7	71.0±0.9	**90.1±0.9**
	Qwen3-32B	68.2±1.2	69.9±1.0	70.1±0.7	65.3±0.7	80.0±0.9	62.1±1.2	70.4±0.6	71.3±0.8	69.0±0.5	71.3±0.4	**88.7±1.1**
NQ	Llama3.1-8B	51.7±1.3	58.4±0.4	58.3±0.6	61.3±1.0	70.4±0.7	56.4±1.1	70.6±1.0	59.8±1.0	71.5±1.3	77.1±0.7	**82.1±1.4**
	Llama3.1-70B	64.3±1.5	56.8±1.2	57.0±1.4	60.7±1.1	59.6±1.0	55.9±0.9	61.4±0.8	59.3±0.6	62.4±1.1	67.4±0.9	**74.0±0.7**
	Qwen3-14B	54.8±0.7	61.6±0.9	62.1±0.6	64.2±0.9	73.2±1.0	60.4±1.0	67.1±1.2	63.6±0.6	68.1±1.0	73.1±1.3	**81.2±1.0**
	Qwen3-32B	60.7±0.8	52.7±1.6	52.9±1.3	56.0±0.9	64.6±0.6	51.8±1.2	56.2±1.0	54.3±1.1	57.2±1.2	62.5±0.8	**82.5±0.6**

Table 2: Uncertainty estimation results measured by AUARC. Values are reported as percentages, mean ± std. Best results are bolded and second-best results are underlined.

Dataset	Model	P(True)	SAR	SE	Eigen Score	SE_{SDLG}	LexSim	Deg	SD	SPUQ	Inv-Entropy	REC
HotpotQA	Llama3.1-8B	68.7±1.1	81.3±0.9	81.9±1.2	82.5±1.2	83.3±0.7	81.1±0.6	82.4±0.6	82.2±0.5	81.3±0.5	81.7±0.7	**87.8±1.2**
	Llama3.1-70B	70.8±0.7	82.4±1.1	84.5±0.9	85.0±0.9	85.8±0.8	83.4±0.5	84.9±0.7	84.2±0.7	82.6±1.0	83.0±0.5	**93.5±0.7**
	Qwen3-14B	66.0±0.7	78.5±0.3	79.2±0.5	81.0±0.5	81.5±0.4	78.4±0.8	79.7±0.7	80.5±0.7	78.6±0.8	79.0±1.0	**90.6±1.0**
	Qwen3-32B	84.6±1.1	75.9±0.5	73.8±0.8	74.0±0.6	85.4±0.7	72.7±0.6	74.0±0.9	73.8±0.4	76.2±0.8	76.6±0.6	**93.4±0.5**
CoQA	Llama3.1-8B	79.5±1.0	**90.7±1.3**	89.3±1.0	90.2±0.6	**90.7±0.4**	89.2±0.7	89.7±0.8	89.1±0.6	84.6±1.0	85.1±1.1	81.3±1.3
	Llama3.1-70B	81.2±1.1	89.6±0.9	91.8±0.7	91.9±0.6	92.5±0.3	91.0±0.3	91.7±0.8	91.2±0.5	88.9±0.6	89.1±1.0	**92.7±0.8**
	Qwen3-14B	77.4±1.4	86.4±0.5	86.6±0.8	88.0±1.1	88.9±0.9	86.2±0.7	87.3±0.7	87.6±0.7	82.5±0.6	83.0±0.4	**91.3±0.6**
	Qwen3-32B	83.0±1.5	80.8±1.4	78.8±1.2	79.8±0.6	90.2±0.3	78.7±0.6	79.3±1.1	79.6±0.7	77.5±1.2	77.8±1.4	**93.3±1.1**
TriviaQA	Llama3.1-8B	82.2±1.5	82.8±1.2	84.0±1.0	83.6±1.1	84.6±0.6	81.1±0.4	82.4±1.1	82.1±1.0	80.3±0.9	80.4±1.2	**87.7±0.7**
	Llama3.1-70B	84.2±1.3	86.6±0.4	88.3±0.5	88.0±0.9	88.9±0.7	85.4±1.0	86.9±0.7	86.3±0.5	86.1±0.6	86.6±0.4	**93.7±1.4**
	Qwen3-14B	69.0±1.3	69.7±1.1	70.9±1.2	70.6±0.6	71.5±0.8	68.2±0.7	69.3±0.8	68.9±0.8	65.6±0.9	66.4±0.5	**80.2±0.9**
	Qwen3-32B	68.1±0.8	69.1±0.9	70.3±1.2	69.8±0.8	70.8±0.7	67.2±0.8	68.6±1.0	68.1±0.7	65.5±0.9	65.9±1.2	**79.1±1.2**
NQ	Llama3.1-8B	44.3±0.7	51.6±0.7	54.3±0.9	53.8±0.6	54.6±0.3	51.7±0.7	53.0±1.0	52.6±0.7	52.0±0.7	53.7±1.1	**57.1±1.0**
	Llama3.1-70B	54.3±0.9	61.2±0.8	62.4±0.9	63.5±0.7	64.0±0.4	61.8±1.0	63.6±0.5	63.0±0.8	59.2±0.8	61.0±0.4	**71.2±0.6**
	Qwen3-14B	42.1±1.1	46.9±1.4	46.4±1.2	45.8±0.6	48.9±0.9	49.0±0.7	49.3±0.8	49.1±0.9	45.7±1.0	47.4±0.6	**50.5±1.5**
	Qwen3-32B	45.8±1.6	45.9±0.9	47.9±1.2	47.4±0.5	48.4±0.7	45.5±1.0	46.7±0.6	46.2±0.9	44.3±0.5	45.8±0.4	**54.1±0.8**

whether answer behavior matches the expected direction under changed evidence or whether the answer remains supported by the corresponding scenario.

REC is especially strong on TriviaQA and NQ, where it obtains the best AUROC for all evaluated models, and remains competitive on HotpotQA and CoQA. The AUARC results further show that REC supports selective prediction: high-uncertainty answers identified by REC are more likely to be unreliable and can be rejected to improve retained-answer accuracy. Overall, these results support the core idea of REC: uncertainty estimation benefits from coupling expected answer behavior with scenario-specific evidence support.

5.2 Ablation Studies

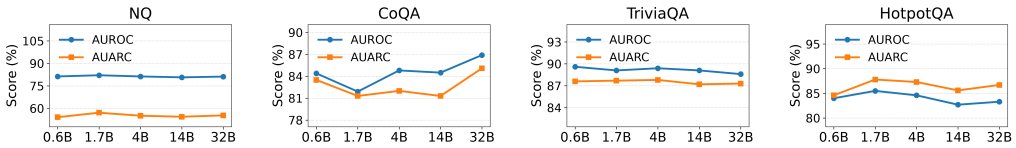
5.2.1 Contribution of Scenario Groups and Diagnostic Signals

Table 3 shows that the full REC formulation consistently outperforms its component-reduced variants. Using only the anchor-support risk u_{w0} decreases AUROC by 2.2–3.8 percentage points and AUARC by 1.3–3.4 points across the four answer models, indicating that the intervention scenario provided additional information that could not be recovered from the anchor scenario alone.

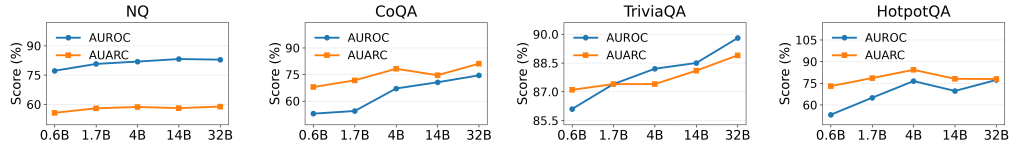
Removing the equivalent group produces a relatively moderate average decrease of 1.15 AUROC points and 0.35 AUARC points. In contrast, removing the counterfactual group causes larger average reductions of 1.93 and 2.03 points, respectively. This indicates that counterfactual replacement provides the stronger individual intervention signal. Nevertheless, the full method remains better than either reduced variant, showing that expected invariance and expected response jointly characterize reliable evidence use.

Table 3: Component ablation averaged across four tasks. Values are percentages; parentheses show percentage-point changes from full REC.

Model	Anchor Only		Scenario-Group Removal				Signal-Component Removal				Full REC	
	u_{w0} only		w/o Eq group		w/o Cf group		w/o Agreement		w/o Support		Full	
	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC
Llama3.1-8B	82.2(−2.5)	77.2(−1.3)	82.5(−2.2)	78.1(−0.4)	82.9(−1.8)	77.5(−1.0)	82.6(−2.1)	78.0(−0.5)	66.0(−18.6)	69.2(−9.3)	**84.7**	**78.5**
Llama3.1-70B	73.6(−2.7)	85.8(−2.0)	75.3(−0.9)	87.3(−0.5)	75.0(−1.3)	86.4(−1.3)	73.6(−2.7)	85.7(−2.1)	65.4(−10.9)	84.6(−3.2)	**76.3**	**87.8**
Qwen3-14B	78.5(−2.2)	75.2(−3.0)	80.2(−0.5)	78.0(−0.2)	79.2(−1.6)	75.6(−2.6)	78.7(−2.0)	76.1(−2.0)	71.5(−9.2)	72.7(−5.5)	**80.8**	**78.2**
Qwen3-32B	77.5(−3.8)	76.6(−3.4)	80.3(−1.0)	79.7(−0.3)	78.4(−3.0)	76.8(−3.2)	78.8(−2.5)	77.7(−2.3)	71.6(−9.8)	75.8(−4.2)	**81.3**	**80.0**



(a) Scenario-construction model scale.



(b) Agreement and support judge model scale.

Figure 3: Sensitivity of REC to auxiliary-model capacity. The answer model and other components are fixed. Vertical ranges differ across datasets.

The signal-component ablations further clarify the roles of support and agreement. Removing agreement, and hence retaining primarily scenario-specific support information, reduces AUROC and AUARC by averages of 2.33 and 1.73 points. Support alone is therefore informative but incomplete. The largest degradation occurs when support is removed. The agreement-only variant loses, on average, 12.13 AUROC points and 5.55 AUARC points. Thus, response agreement cannot be interpreted reliably without verifying whether the generated answers are supported by their corresponding scenarios.

5.2.2 Sensitivity to Auxiliary-Model Capacity

Figure 3 shows no consistent gain from scaling the scenario constructor, suggesting that lightweight models suffice once valid interventions can be produced. Judge capacity matters more: scaling from 0.6B to 32B generally improves NQ, CoQA, and TriviaQA, with CoQA AUROC/AUARC rising from 52.9/67.9 to 74.5/81.1. Hence, under a limited budget, REC should prioritize judge capacity while keeping scenario construction lightweight. We additionally evaluate alternative fusion functions, the complete results and analysis are provided in Appendix A.3, and Appendix A.4 provides further ablation studies on the number of scenes.

5.3 Inference Cost

Table 4 shows that REC introduces an average of 596 additional output tokens per example, of which 62.6% comes from context construction, mainly equivalent rewriting. Counterfactual construction uses deterministic replacement and therefore incurs no additional output-token cost. The overall overhead is largely determined by context length: NQ and TriviaQA require fewer than 300 extra tokens per example, whereas equivalent rewriting accounts for 66.8% and 81.1% of the total cost on CoQA and HotpotQA, respectively.

Although scenario construction dominates the token overhead, it is handled by a lightweight model. In contrast, support and agreement evaluation require stronger reasoning but together account for only 23.3% of the average additional output tokens. REC therefore assigns the token-intensive construction stage to a small model and reserves stronger models for the less token-intensive judging stage, keeping the overall additional inference cost within a controllable range.

Table 4: Additional output-token overhead of REC by task.

Task	Anchor evidence	Context construction			Judge		Answer gen.	Total
		Pseudo-context	Eq	Cf	Support	Agreement		
NQ	50 (17.5%)	30 (10.5%)	30 (10.5%)	0 (0.0%)	100 (35.1%)	60 (21.1%)	15 (5.3%)	**285 (100.0%)**
TriviaQA	80 (28.4%)	27 (9.6%)	27 (9.6%)	0 (0.0%)	70 (24.8%)	60 (21.3%)	18 (6.4%)	**282 (100.0%)**
CoQA	75 (11.4%)	0 (0.0%)	440 (66.8%)	0 (0.0%)	75 (11.4%)	60 (9.1%)	9 (1.4%)	**659 (100.0%)**
HotpotQA	80 (6.9%)	0 (0.0%)	939 (81.1%)	0 (0.0%)	70 (6.0%)	60 (5.2%)	9 (0.8%)	**1158 (100.0%)**
Average	**71.3 (12.0%)**	**14.3 (2.4%)**	**359.0 (60.2%)**	**0.0 (0.0%)**	**78.8 (13.2%)**	**60.0 (10.1%)**	**12.8 (2.1%)**	**596.0 (100.0%)**

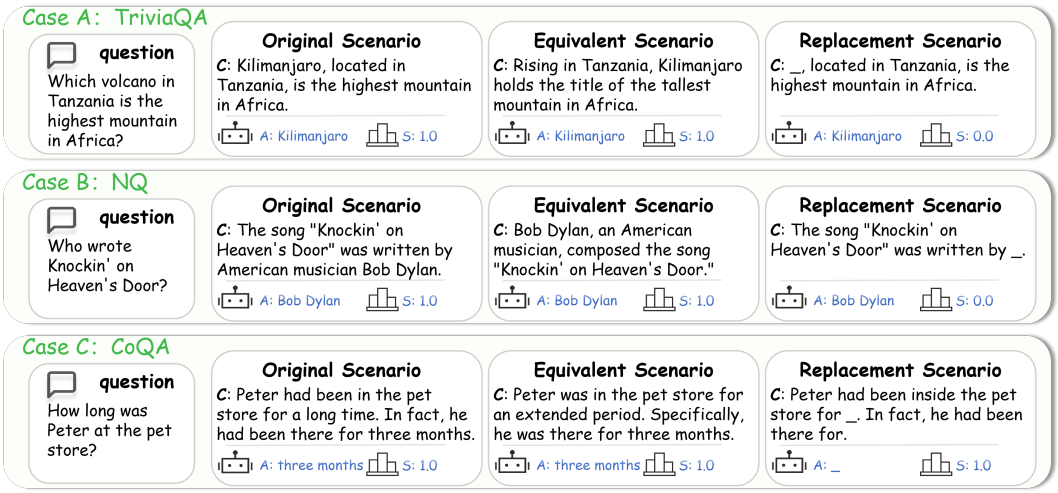


Figure 4: Compressed qualitative cases of REC. Each scenario cell reports the context (**C**), generated answer (**A**), and support score (**S**)

5.4 Case Analysis

Figure 4 illustrates how REC distinguishes uncertainty sources that equivalent rewriting alone cannot reveal. In all three cases, the anchor answer remains unchanged and fully supported under the equivalent scenario. Cases A and B correspond to open-domain QA, where replacement should preserve the question-determined answer. The model consistently retains "Kilimanjaro" and "Bob Dylan," but both answers receive zero support after the answer-bearing evidence is removed. Their behavior is therefore consistent with the task-aware expectation, while the evidence-gap signal reveals their reliance on parametric knowledge rather than the current scenario. In Case C, CoQA is context-grounded, so removing "three months" should induce an answer change. The model appropriately changes to an empty answer with full support, demonstrating evidence-sensitive rather than unstable behavior. Together, these cases show that answer preservation is not necessarily reliable when scenario support is absent, while answer change can be desirable when required by the intervention.

6 Conclusion and Future Work

We presented REC, an uncertainty estimation framework that evaluates whether a language model responds appropriately to controlled changes in its evidence context. Rather than relying on confidence from a single scenario or measuring answer variation alone, REC jointly considers scenario-specific evidence support and the expected response behavior under anchor, equivalent, and counterfactual scenarios. The resulting risks are organized into semantically meaningful groups and aggregated through rank-normalized Fisher fusion. Experiments across four QA datasets and four answer models show that REC consistently provides strong uncertainty estimates and generally outperforms both non-intervention and perturbation-based baselines. Future work may extend REC to more diverse tasks, longer-form generation, and settings where intervention validity and judge reliability must be estimated jointly.

Ethics Statement

This work studies uncertainty estimation for large language models through controlled interventions on evidence contexts. It does not involve human subjects, private user data, sensitive personal information, or proprietary datasets. All evaluated benchmarks, including NQ, TriviaQA, CoQA, and HotpotQA, are publicly available and are used solely for research purposes. The equivalent and counterfactual scenarios constructed by REC are generated for model evaluation rather than for producing factual content for deployment. Nevertheless, because auxiliary language models and judges may inherit biases or errors from their training data, REC scores should be treated as uncertainty indicators rather than definitive guarantees of factual correctness or safety.

Reproducibility Statement

We provide detailed descriptions of REC's anchor, equivalent, and counterfactual scenario construction procedures, together with the definitions of support, agreement, scenario-group risks, and the final Fisher-based aggregation. The datasets, evaluated answer models, auxiliary models, sample sizes, prompting configurations, evaluation metrics, and implementation settings are reported in the main text and appendices. We also report component, fusion-function, and auxiliary-model ablations to clarify the contribution and sensitivity of each design choice. The prompts, pseudocode, preprocessing procedures, and additional experimental details will be included in the supplementary material. We further plan to release anonymized source code and configuration files to support independent replication of the experiments.

AI Use Statement

This manuscript used a large language model only for light editorial support—namely grammar and spelling checks, minor language polishing, and table formatting. The LLM did not generate scientific content, results, analyses, or claims. All edits were reviewed by the authors, and the authors remain fully responsible for the final text.

References

Lukas Aichberger, Kajetan Schweighofer, Mykyta Ielanskyi, and Sepp Hochreiter. Improving uncertainty estimation through semantically diverse language generation. In *International Conference on Learning Representations*, 2025. URL https://proceedings.iclr.cc/paper_files/paper/2025/hash/b94d8b035e2183e47afef9e2f299ba47-Abstract-Conference.html.

Yavuz Bakman, Sungmin Kang, Zhiqi Huang, Duygu Nur Yaldiz, Catarina G. Belém, Chenyang Zhu, Anoop Kumar, Alfy Samuel, Salman Avestimehr, Daben Liu, and Sai Praneeth Karimireddy. Uncertainty as feature gaps: Epistemic uncertainty quantification of LLMs in contextual question-answering, 2025. URL https://arxiv.org/abs/2510.02671.

Yavuz Faruk Bakman, Duygu Nur Yaldiz, Baturalp Buyukates, Chenyang Tao, Dimitrios Dimitriadis, and Salman Avestimehr. MARS: Meaning-aware response scoring for uncertainty estimation in generative LLMs. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7752–7767, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.419. URL https://aclanthology.org/2024.acl-long.419/.

Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. Inside: Llms'internal states retain the power of hallucination detection. In B. Kim, Y. Yue, S. Chaudhuri, K. Fragkiadaki, M. Khan, and Y. Sun (eds.), *International Conference on Learning Representations*, volume 2024, pp. 3056–3076, 2024. URL https://proceedings.iclr.cc/paper_files/paper/2024/file/0d1986a61e30e5fa408c81216a616e20-Paper-Conference.pdf.

Jiuhai Chen and Jonas Mueller. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association*

for Computational Linguistics (Volume 1: Long Papers), pp. 5186–5200, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.283. URL https://aclanthology.org/2024.acl-long.283/.

Lihu Chen, Gerard de Melo, Fabian M. Suchanek, and Gaël Varoquaux. Query-level uncertainty in large language models, 2025. URL https://arxiv.org/abs/2506.09669.

Longchao Da, Xiaoou Liu, Jiaxin Dai, Lu Cheng, Yaqing Wang, and Hua Wei. Understanding the uncertainty of LLM explanations: A perspective based on reasoning topology, 2025. URL https://arxiv.org/abs/2502.17026.

DeepSeek-AI. DeepSeek-V4-Flash. https://huggingface.co/deepseek-ai/DeepSeek-V4-Flash, 2026. Model card. Accessed: 2026-07-12.

Jinhao Duan, Hao Cheng, Shiqi Wang, Alex Zavalny, Chenan Wang, Renjing Xu, Bhavya Kailkhura, and Kaidi Xu. 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)*, pp. 5050–5063, Bangkok, Thailand, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.276. URL https://aclanthology.org/2024.acl-long.276/.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. The Llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, 2024. doi: 10.1038/s41586-024-07421-0. URL https://doi.org/10.1038/s41586-024-07421-0.

Tom Fawcett. An introduction to ROC analysis. *Pattern Recognition Letters*, 27(8):861–874, 2006. doi: 10.1016/j.patrec.2005.10.010.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 2336–2346, St. Julian's, Malta, 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.eacl-long.143. URL https://aclanthology.org/2024.eacl-long.143/.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A survey of uncertainty in deep neural networks. *Artificial Intelligence Review*, 56:1513–1589, 2023. doi: 10.1007/s10462-023-10562-9.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2), 2025. doi: 10.1145/3703155.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 1601–1611, Vancouver, Canada, 2017. Association for Computational Linguistics. doi: 10.18653/v1/P17-1147. URL https://aclanthology.org/P17-1147/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. Language models (mostly) know what they know, 2022. URL https://arxiv.org/abs/2207.05221.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs, 2024. URL https://arxiv.org/abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*, 2023.

Lucie Kunitomo-Jacquin, Edison Marrese-Taylor, Ken Fukuda, and Masahiro Hamasaki. Evidential semantic entropy for LLM uncertainty quantification. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 7107–7122, Rabat, Morocco, 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.eacl-long.334. URL https://aclanthology.org/2026.eacl-long.334/.

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. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466, 2019. doi: 10.1162/tacl_a_00276. URL https://aclanthology.org/Q19-1026/.

Stephanie Lin, Jacob Hilton, and Owain Evans. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*, 2022. URL https://openreview.net/forum?id=8s8K2UZGTZ.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Contextualized sequence likelihood: Enhanced confidence scores for natural language generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pp. 10351–10368, Miami, Florida, USA, 2024a. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.578. URL https://aclanthology.org/2024.emnlp-main.578/.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*, 2024b. ISSN 2835-8856. URL https://openreview.net/forum?id=DWkJCSxKU5.

Andrey Malinin and Mark Gales. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*, 2020.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 9004–9017, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

Sewon Min, Kalpesh Krishna, Xinxi Lyu, Mike Lewis, Wen-tau Yih, Pang Koh, Mohit Iyyer, Luke Zettlemoyer, and Hannaneh Hajishirzi. FActScore: Fine-grained atomic evaluation of factual precision in long-form text generation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 12076–12100, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.741. URL https://aclanthology.org/2023.emnlp-main.741/.

Kevin P. Murphy. *Probabilistic Machine Learning: An Introduction*. Adaptive Computation and Machine Learning. MIT Press, 2022. ISBN 9780262046824. URL https://mitpress.mit.edu/9780262046824/probabilistic-machine-learning/.

Kenton Murray and David Chiang. Correcting length bias in neural machine translation. In *Proceedings of the Third Conference on Machine Translation: Research Papers*, pp. 212–223, Brussels, Belgium, oct 2018. Association for Computational Linguistics. doi: 10.18653/v1/W18-6322. URL https://aclanthology.org/W18-6322/.

Malik Sajjad Ahmed Nadeem, Jean-Daniel Zucker, and Blaise Hanczar. Accuracy-rejection curves (ARCs) for comparing classification methods with a reject option. In *Proceedings of the Third International Workshop on Machine Learning in Systems Biology*, volume 8 of *Proceedings of Machine Learning Research*, pp. 65–81. PMLR, 2009. URL https://proceedings.mlr.press/v8/nadeem10a.html.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for llms from semantic similarities. *Advances in Neural Information Processing Systems*, 37:8901–8929, 2024.

Xin Qiu and Risto Miikkulainen. Semantic density: Uncertainty quantification for large language models through confidence measurement in semantic space. *Advances in neural information processing systems*, 37:134507–134533, 2024.

Siva Reddy, Danqi Chen, and Christopher D. Manning. CoQA: A conversational question answering challenge. *Transactions of the Association for Computational Linguistics*, 7:249–266, 2019. doi: 10.1162/tacl_a_00266. URL https://aclanthology.org/Q19-1016/.

Jie Ren, Yao Zhao, Tu Vu, Peter J. Liu, and Balaji Lakshminarayanan. Self-evaluation improves selective generation in large language models. In *Proceedings on "I Can't Believe It's Not Better: Failure Modes in the Age of Foundation Models" at NeurIPS 2023 Workshops*, volume 239 of *Proceedings of Machine Learning Research*, pp. 49–64. PMLR, 2023. URL https://proceedings.mlr.press/v239/ren23a.html.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pp. 4902–4912, Online, 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.acl-main.442. URL https://aclanthology.org/2020.acl-main.442/.

Haoyi Song, Ruihan Ji, Naichen Shi, Fan Lai, and Raed Al Kontar. Inv-Entropy: A fully probabilistic framework for uncertainty quantification in language models. In *Advances in Neural Information Processing Systems*, volume 38. Curran Associates, Inc., 2025. URL https://proceedings.neurips.cc/paper_files/paper/2025/hash/9afca5107064eb89546fabb4abc379a7-Abstract-Conference.html.

Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. *Advances in Neural Information Processing Systems*, 37:34188–34216, 2024.

Qiyao Sun, Xingming Li, Xixiang He, Ao Cheng, Xuanyu Ji, Hailun Lu, Runke Huang, and Qingyong Hu. Efficient hallucination detection: Adaptive bayesian estimation of semantic entropy with guided semantic exploration. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 40, pp. 33117–33125, 2026. doi: 10.1609/aaai.v40i39.40595. URL https://ojs.aaai.org/index.php/AAAI/article/view/40595.

Katherine Tian, Eric Mitchell, Allan Zhou, Archit Sharma, Rafael Rafailov, Huaxiu Yao, Chelsea Finn, and Christopher Manning. Just ask for calibration: Strategies for eliciting calibrated confidence scores from language models fine-tuned with human feedback. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pp. 5433–5442, Singapore, 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.330. URL https://aclanthology.org/2023.emnlp-main.330/.

Minh N. Vu, Brian Kha Tran, Syed Alamdar Hussain Shah, Geigh Zollicoffer, Nhat Hoang-Xuan, and Manish Bhattarai. HalluField: Detecting LLM hallucinations via field-theoretic modeling, 2025. URL https://arxiv.org/abs/2509.10753.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, Brussels, Belgium, 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Chong Meng, Shuaiqiang Wang, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. Knowing what LLMs DO NOT know: A simple yet effective self-detection method, 2024. URL https://arxiv.org/abs/2310.17918.

A Additional Experiments

A.1 Correctness Judge Calibration

We conduct a human evaluation to calibrate the independent GPT-5.5 correctness judge. For each of CoQA, HotpotQA, TriviaQA, and NQ, we randomly sample 50 generated answers, yielding 200 examples in total. Two human annotators independently receive the question, official reference answers and aliases, and the model prediction, and assign a binary correct/incorrect label. They are not shown the GPT-5.5 judgment, REC signals, or uncertainty score.

Table 5: Agreement between GPT-5.5 correctness judgments and adjudicated human labels.

Dataset	Agreement
CoQA	96.0%
HotpotQA	94.0%
TriviaQA	100.0%
NQ	100.0%
Overall	**97.5%**

As shown in Table 5, GPT-5.5 achieves an overall agreement of 97.5% with the adjudicated human labels. Agreement reaches 100.0% on TriviaQA and NQ, 96.0% on CoQA, and 94.0% on HotpotQA. Overall, the high agreement supports the reliability of GPT-5.5 as an independent, reference-conditioned correctness judge and reduces concerns that the reported results are driven by the preferences of the model used to compute REC's internal signals.

A.2 Intervention Validity Evaluation

REC assumes that equivalent scenarios preserve answer-bearing evidence, whereas counterfactual replacement scenarios systematically alter eligible contextual key facts while preserving the question and its protected anchors. Because violations of these assumptions may distort the expected behavior direction and the resulting uncertainty risks, we directly evaluate these two construction properties.

For each dataset, we randomly sample 100 examples and assess the equivalent and replacement scenarios generated by REC. **Eq. Preserve** measures whether the equivalent rewrite retains the answer-bearing evidence in the anchor scenario and continues to support the same answer. **Rep. Alter** measures whether the replacement successfully alters or removes answer-relevant contextual support through the replacement of eligible key facts, while preserving the question referent and question-side anchors. Rep. Alter does not require the answer to change in open-domain QA, because modifying pseudo-contextual evidence does not redefine the real-world answer determined by the question.

We use GPT-5.5 as an independent validator. For each example, the validator receives the question, anchor scenario, anchor answer, constructed scenario, and intervention type, but not the downstream REC score. The reported validity rate is the proportion of sampled examples judged valid.

As shown in Figure 5, equivalent scenarios achieve an average Eq. Preserve rate of 93.8%, ranging from 91.0% on NQ to 96.1% on CoQA. Replacement scenarios obtain an average Rep. Alter rate of 92.0%, with dataset-level results ranging from 89.1% on HotpotQA to 94.6% on TriviaQA. These results indicate that both intervention types satisfy their intended construction property in most cases. The remaining failures primarily involve ambiguous key-fact spans or long and noisy contexts, where an equivalent rewrite may omit relevant evidence, or the deterministic detector may miss a decisive factual span or leave residual contextual cues.

A.3 Fusion-Function Ablation

Table 6 compares REC with four alternative fusion strategies. NoisyOR, product, and arithmetic-mean fusion are consistently inferior to REC. Among them, product fusion exhibits the largest overall degradation, particularly for Qwen3-32B, where AUROC and AUARC decrease by 3.4 and 3.2 points. A direct product can be dominated by a single small component, whereas an arithmetic mean may

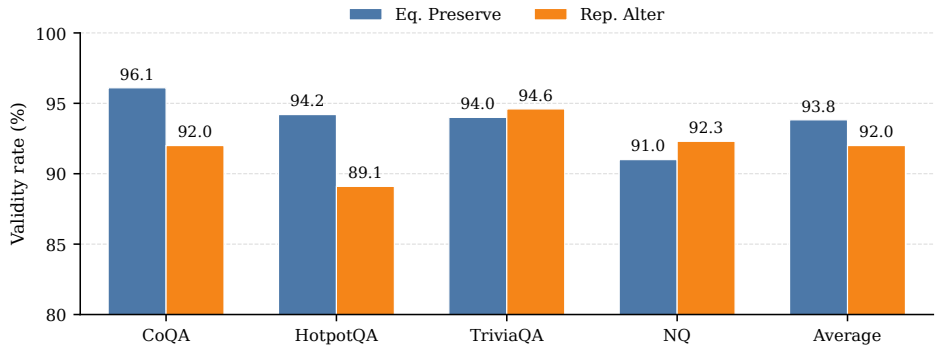


Figure 5: Intervention validity across four datasets. Eq. Preserve measures whether equivalent scenarios retain the answer-bearing evidence, while Rep. Alter measures whether replacement scenarios successfully alter or remove answer-relevant contextual support while preserving question-side anchors.

Table 6: Fusion-function ablation averaged across four tasks. Values are percentages; parentheses show percentage-point changes from REC.

Model	Atomic Fisher		NoisyOR		Product		Mean		REC	
	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC	AUROC	AUARC
Llama3.1-8B	82.5 (−2.2)	78.1 (−0.4)	80.1 (−4.6)	77.0 (−1.5)	81.8 (−2.9)	77.5 (−1.0)	82.3 (−2.4)	77.8 (−0.7)	**84.7**	**78.5**
Llama3.1-70B	76.2 (−0.1)	87.7 (−0.1)	75.1 (−1.2)	87.4 (−0.4)	73.7 (−2.6)	85.8 (−2.0)	75.3 (−0.9)	87.5 (−0.3)	**76.3**	**87.8**
Qwen3-14B	80.8 (+0.0)	**78.3** (+0.1)	79.5 (−1.2)	77.7 (−0.5)	79.0 (−1.8)	75.3 (−2.9)	80.4 (−0.4)	77.9 (−0.3)	**80.8**	78.2
Qwen3-32B	**81.4** (+0.0)	**80.0** (+0.0)	80.0 (−1.4)	79.7 (−0.3)	78.0 (−3.4)	76.8 (−3.2)	81.0 (−0.4)	79.9 (−0.1)	81.3	**80.0**

dilute a strong risk signal with several weak ones. NoisyOR avoids complete cancellation but still imposes a fixed probabilistic interaction among the inputs.

Atomic Fisher is substantially more competitive. It is nearly tied with REC on Qwen3-14B and Qwen3-32B and differs by only 0.1 point on Llama3.1-70B. However, it remains 2.2 AUROC points below REC on Llama3.1-8B. These results indicate that Fisher-style aggregation accounts for an important part of REC's effectiveness, while organizing the risks into semantically meaningful anchor, equivalent, and counterfactual groups improves robustness across answer models.

A.4 Ablation on the Number of Intervention Scenarios

We examine the effect of varying the numbers of equivalent and replacement scenarios, denoted by #Eq/#Rep. Equivalent scenarios test answer preservation under evidence-preserving reformulations, whereas replacement scenarios test task-aware responses to altered evidence.

Figure 6 shows that neither intervention type uniformly dominates the other. The equivalent-only setting $(1/0)$ yields a slightly higher AUROC than the replacement-only setting $(0/1)$, whereas $0/1$ performs better in AUARC. Both remain below the balanced $1/1$ setting, which improves AUROC/AUARC to 84.7/78.5, confirming that evidence-preserving and evidence-altering tests provide complementary information.

Increasing the scenario count to $2/2$ further improves both metrics by reducing sensitivity to a single imperfect intervention. The gain from $2/2$ to $3/3$, however, is marginal, indicating diminishing returns. We therefore use $1/1$ as the default cost-effective setting, while additional scenarios can be used when the extra inference cost is acceptable.

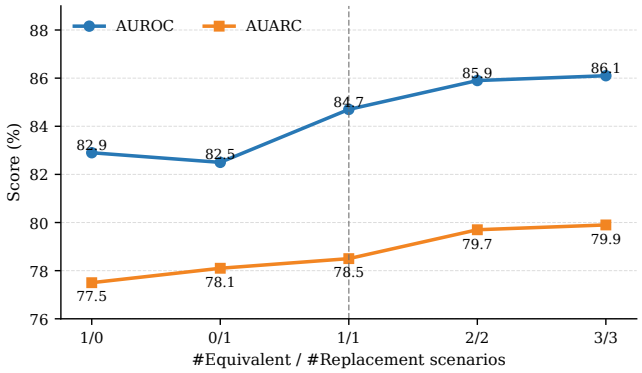


Figure 6: Effect of the number of intervention scenarios on REC using Llama-3.1-8B, averaged across four tasks. The dashed line marks the default $1/1$ setting.

B Answer Generation and Question-Only Pseudo-Context Prompts

B.1 Answer Generation Prompt

REC uses the same answer-generation template for the anchor and intervention worlds. For CoQA and HotpotQA, the anchor answer is generated from the original context. For NQ and TriviaQA, the anchor answer is first generated from the question with no context; the question-only pseudo-context is then constructed independently, and the already generated anchor answer is reused as a_0. Answers for the equivalent and counterfactual worlds are generated from their corresponding constructed contexts.

The exact prompt template used in the implementation is:

```
Answer the question using the context if available.
Return the final answer only inside <answer></answer> tags.
Do not output JSON. Do not use markdown.

[TASK-SPECIFIC ANSWER INSTRUCTION]

Output format:
<answer>your final answer here</answer>

Question:
{question}

Context:
{context}
```

For NQ and TriviaQA, the inserted instruction is:

```
Task-specific instruction for open-domain QA:
- Answer using your knowledge.
- Do not require an external context to answer.
- Do not answer UNKNOWN merely because no context is provided.
- Return the shortest entity, name, title, date, location, number, or phrase that directly
    answers the question.
- If the answer has common aliases, return the most standard concise form.
- If you genuinely do not know the answer, return UNKNOWN.
- Do not include explanations inside the final answer tag.
- Never leave the answer tag empty; if genuinely unknown, write UNKNOWN.
```

For CoQA and HotpotQA, the inserted instruction is:

```
Task-specific instruction for reading QA:
- Answer using the provided context.
- Prefer the shortest answer span or concise paraphrase supported by the context.
- Reasonable inference from the context is allowed, especially for emotions, intentions,
    attitudes, yes/no questions, and causal relations.
```

```
- For yes/no questions, answer YES or NO when the context directly states or reasonably
    implies the polarity.
- Return UNKNOWN only when the context provides neither a direct answer nor enough evidence
    for a reasonable inference.
- Do not return UNKNOWN merely because the exact answer word does not appear in the context.
- Do not include extra events, explanations, or later facts unless the question explicitly
    asks for them.
- Never leave the answer tag empty; if the context is insufficient, write UNKNOWN.
```

B.2 Question-Only Pseudo-Context Prompt

For the no-context datasets NQ and TriviaQA, REC constructs a compact pseudo-context using G_{ans}. Although the implementation function accepts the anchor answer as an argument for compatibility with legacy modes, the reported `question_only` prompt does not insert or otherwise condition on that answer. The output is directly parsed from the `<context>` tags and used as the anchor context c_0.

```
Write a short fact-focused pseudo-context that helps answer the question.
Do not receive, assume, or condition on any candidate answer.
If you know the relevant fact, state it directly and minimally.
If the fact is genuinely uncertain, state only the uncertainty without inventing a fact.
Use declarative evidence-like facts, not step-by-step reasoning.
Only output the context inside <context></context> tags.

Question:
{question}

Output format:
<context>short fact-focused context here</context>
```

C Anchor Evidence Extraction and Support Scoring

C.1 Anchor Evidence Prompt

After obtaining the anchor triplet (q, c_0, a_0), REC invokes Qwen3-1.7B with a dedicated Anchor Evidence Prompt. To preserve the behavior of the original implementation as closely as possible, this prompt retains the same task-specific support rules, input order, UNKNOWN handling, and evidence explanation used by the Support scorer. The only substantive change is that it removes the support-score request and returns only `spans` and `reason`. DeepSeek-V4-Flash independently computes the anchor Support score s_0 using the unchanged Support prompt in Appendix C.2.

```
Identify the minimal context span or spans that support the candidate answer to the question.
Use ONLY the context plus the type of reasoning allowed by the task.
Do not use unrelated outside knowledge.

[TASK-SPECIFIC SUPPORT RULES]

Special rule for UNKNOWN:
- If the candidate answer is UNKNOWN, determine whether UNKNOWN is appropriate.
- Treat UNKNOWN as supported only if the context is genuinely insufficient to answer the
    current question.
- Return an empty span list if the context clearly supports a concrete answer.
- Do not treat UNKNOWN as supported when the answer is explicitly stated or strongly entailed
    by the context.

Important:
- The returned spans must represent evidence of SUPPORT, not your confidence as a judge.
- Return the minimal context span or spans most relevant to whether the candidate answer is
    supported.
- If no supporting evidence exists, return an empty span list.
- Do not output a categorical label.

Return valid JSON only. Do not use markdown.

JSON schema:
{
  "spans": ["minimal span from the context"],
  "reason": "brief reason"
}

Task: {task_name}
```

```
Question:
{question}

Context:
{context}

Candidate answer:
{anchor_answer}
```

For NQ and TriviaQA, the inserted task-specific rules are:

```
Task-specific support rules for open-domain QA with pseudo-contexts:
- The context may be a generated pseudo-context used to externalize knowledge.
- Judge only whether this context supports the candidate answer.
- Do not require the context to be an original retrieved passage.
- Do not use unrelated outside knowledge to override the context.
- A concise entity answer is supported if the context states or clearly entails that entity as
     the answer.
- If the candidate answer is UNKNOWN, treat it as supported only when the context is genuinely
     insufficient.
- If the context states a concrete answer, do not return supporting spans for UNKNOWN.
- Treat common aliases, expanded names, dates in equivalent formats, and concise paraphrases
    as supported when the relation is preserved.
```

For CoQA and HotpotQA, the inserted task-specific rules are:

```
Task-specific support rules for reading QA:
- The answer must be supported by the context.
- Exact wording is not required, but the answer must preserve the same entity, attribute,
    relation, event, or location.
- Accept common aliases, more specific names, or concise paraphrases when they refer to the
    same answer.
- Reasonable inference is allowed when the context clearly supports it.
- If the answer only shares words with the context but the required relation is not supported,
     do not return it as supporting evidence.
```

The returned `spans` constitute $\mathcal{E}_0$. No score from this call enters REC.

C.2 Support Scoring Prompt

The support score represents the degree to which the current context supports the candidate answer. It is a continuous value in $[0, 1]$ and is not treated as a formally calibrated probability.

```
Evaluate how strongly the provided context supports the candidate answer to the question.
Use ONLY the context plus the type of reasoning allowed by the task.
Do not use unrelated outside knowledge.

[TASK-SPECIFIC SUPPORT RULES]

Special rule for UNKNOWN:
- If the candidate answer is UNKNOWN, score whether UNKNOWN is appropriate.
- Give a high score only if the context is genuinely insufficient to answer the current
    question.
- Give a low score if the context clearly supports a concrete answer.
- Do not reward UNKNOWN when the answer is explicitly stated or strongly entailed by the
    context.

Important:
- The score must represent degree of SUPPORT, not your confidence as a judge.
- A higher score means stronger support from the context.
- A lower score means weaker support from the context.
- Do not output a categorical label.

Return valid JSON only. Do not use markdown.

JSON schema:
{
  "score": a support score from 0.0 to 1.0,
  "spans": ["minimal span from the context"],
  "reason": "brief reason"
}

Task: {task_name}
```

```
Question:
{question}

Context:
{context}

Candidate answer:
{answer}
```

For NQ and TriviaQA, the inserted support rules are:

```
Task-specific support rules for open-domain QA with pseudo-contexts:
- The context may be a generated pseudo-context used to externalize knowledge.
- Judge only whether this context supports the candidate answer.
- Do not require the context to be an original retrieved passage.
- Do not use unrelated outside knowledge to override the context.
- A concise entity answer is supported if the context states or clearly entails that entity as
     the answer.
- If the candidate answer is UNKNOWN, score it high only when the context is genuinely
    insufficient.
- If the context states a concrete answer, UNKNOWN should receive low support.
- Treat common aliases, expanded names, dates in equivalent formats, and concise paraphrases
    as supported when the relation is preserved.
```

For CoQA and HotpotQA, the inserted support rules are:

```
Task-specific support rules for reading QA:
- The answer must be supported by the context.
- Exact wording is not required, but the answer must preserve the same entity, attribute,
    relation, event, or location.
- Accept common aliases, more specific names, or concise paraphrases when they refer to the
    same answer.
- Reasonable inference is allowed when the context clearly supports it.
- If the answer only shares words with the context but the required relation is not supported,
     label unsupported.
```

D SCENARIO CONSTRUCTION

REC constructs one equivalent world and one counterfactual world for each example in the main setting. The question is unchanged in both constructions.

D.1 EQUIVALENT-CONTEXT CONSTRUCTION

In the reported setting, equivalent contexts are generated through the context-only rewrite path used by the passage-level cache. The prompt is independent of the question, anchor answer, and extracted evidence spans, which allows examples sharing the same passage to reuse an identical equivalent rewrite.

```
Paraphrase ONLY the context while preserving its complete meaning.
This is a question-independent equivalent-context rewrite.
Preserve all factual content, named entities, numbers, dates, quantities, relations,
    coreference, and negation.
Do not add new facts. Do not remove answer-relevant or potentially answer-relevant facts.
Do not answer any question. Only output the rewritten context inside <context></context> tags.

Original context:
{context}

Output format:
<context>rewritten context here</context>
```

D.2 COUNTERFACTUAL REPLACEMENT PROCEDURE

Counterfactual construction does not use an LLM prompt. In the main setting, the deterministic replacer receives the question, anchor context, and anchor answer. It detects high-value entity and key-fact spans throughout the context with `gliner-large-v2.1` together with deterministic filtering rules, removes candidates protected by question-side anchors, resolves overlapping or duplicate candidates, and replaces all remaining eligible spans with typed placeholders such as

ENT_PERSON_1 or ENT_DATE_1. The same replacement is applied consistently to related context mentions when required. The question remains unchanged, and the surrounding non-replaced textual structure is preserved.

The procedure can be summarized as:

```
Input: question q, anchor context c0, anchor answer a0

1. Detect high-value entity/key-fact candidates throughout c0.
2. Remove candidates protected by the question q.
3. Resolve overlapping or duplicate candidates.
4. Infer a type tau_j for each remaining eligible span m_j.
5. Replace all eligible spans m_j with typed placeholders ENT_tau_j_k.
6. Apply replacements consistently to related mentions when required.
7. Keep q unchanged and return the modified context c_cf.

Output: counterfactual context c_cf
```

E Semantic Agreement and Correctness Evaluation

E.1 Semantic Answer Agreement

Before invoking the LLM judge, the implementation applies conservative deterministic checks for normalized exact matches, yes/no answers, dates, numbers, and high-confidence aliases. Remaining answer pairs are judged by DeepSeek-V4-Flash with the following prompt. The returned score measures answer sameness only; the expected intervention direction is applied later in the REC risk computation.

```
Judge whether the two answers should be considered the same for QA evaluation.
Ignore harmless wording differences.
However, answers must be judged DIFFERENT if they differ in any key factual element, including:

- entity identity
- number or quantity
- color
- location
- relation
- event or action
- yes/no polarity
- owner or agent

Examples:
- 'her mommy and 5 other sisters' vs 'her mommy and 5 other brothers' => different
- 'orange with white stripes' vs 'orange and white' => same
- 'the old farmer' vs 'the old farmer\'s' => same
- 'in a barn' vs 'near a farm house' => different unless both are part of the same complete
    location phrase

Return valid JSON only. Do not use markdown.

JSON schema:
{
  "same": true or false,
  "score": a number from 0.0 to 1.0,
  "reason": "brief reason"
}

Task: {task_name}
Answer A: {answer_a}
Answer B: {answer_b}
```

E.2 Correctness Judge Prompt

The correctness judge is used only to produce evaluation labels for AUROC and AUARC and is not an input to REC. For the four evaluated QA datasets, the implementation uses the following prompt when LLM-based correctness judgment is enabled.

```
Judge whether the model prediction should be considered correct for the QA task.
Use semantic equivalence, not exact string matching.
Accept answers that are more specific or slightly more detailed if they correctly contain or
    entail the gold answer.
```

```
Accept concise paraphrases if they preserve the same entity, relation, attribute, number, date,
      location, or yes/no polarity.
Reject answers that contradict the gold answer or answer a different question.
Reject answers that are only topically related but do not provide the same answer.
Output exactly one tagged field and nothing else:
<judgement>correct</judgement>
or
<judgement>incorrect</judgement>

Task: {task_name}
Question: {question}
Gold answer: {gold_answer}
Prediction: {prediction}

Acceptable aliases:
{aliases}
```

F Output Parsing and Caching Details

Answers are extracted from the `<answer>` tags and normalized conservatively; compact untagged outputs are retained only as a fallback. Support outputs are parsed as JSON, with additional recovery logic for partially generated JSON so that the score, evidence spans, and reason can still be extracted. An unparseable correctness judgment is conservatively marked incorrect.

REC uses two construction caches. The equivalent-context cache is keyed by the normalized context and rewrite configuration, enabling shared passages to reuse a context-only equivalent rewrite. A separate low-level cache stores repeated GLiNER span-candidate extraction. These caches reduce repeated construction calls but do not change the REC score definition.