When Do Uncertainty Signals Generalise? Evaluating Hallucination Detection Beyond Factual QA

Anonymous ACL submission

Abstract

Uncertainty quantification (UQ) methods are widely proposed as post-hoc hallucination detectors deployable without model retraining or white-box access. However, existing evaluations focus almost exclusively on factoid QA benchmarks, leaving unclear whether these methods generalise to faithfulness hallucinations arising in retrieval-augmented generation and document-grounded settings. We evaluate four established UQ methods together with DARE-PE, a black-box deletion-based token-importance reweighting variant we propose, and multi-method ensembles, across three open-source LLMs and eight benchmarks organised in a 2×2 design varying hallucination type and task format. On factual benchmarks, DARE-PE is the strongest individual method, and ensembles yield further gains where signals are complementary. On faithfulness benchmarks, performance is substantially weaker and depends on error structure: methods retain above-chance discrimination when passages explicitly conflict with model priors, but collapse toward chance when failures arise from broader comprehension errors or confident incorrect abstention. These findings show that the utility of output-level UQ signals for hallucination detection depends on the structure of the underlying error, and that faithfulness detection cannot be treated as a single uniform task.

1 Introduction

The deployment of large language models (LLMs) in retrieval-augmented generation (RAG), summarisation, and document-grounded question answering has made *faithfulness hallucination* a major practical concern. Unlike factual hallucination, where a model generates information that is inconsistent with world knowledge, faithfulness hallucination arises when a model generates output that contradicts or fails to follow the *provided* context, relying instead on parametric knowledge (Huang et al., 2025; Ji et al., 2023). In high-stakes applications such as medical summarisation or legal document review, a model that confidently overrides the provided source may be more harmful than one that simply lacks the relevant knowledge.

Uncertainty quantification (UQ) has attracted sustained attention as a post-hoc approach to hallucination detection. These methods, including sampling-based consistency (Kuhn et al., 2023), token-level probability scoring (Tomani et al., 2024), verbalised self-assessment (Tian et al., 2023; Xiong et al., 2024), and input-space perturbation (Gao et al., 2024), operate on model outputs without requiring retraining or white-box access to internal activations, making them practical for deployment across commercial API settings. Their central assumption is that lower confidence should correlate with a higher likelihood of hallucination.

However, this assumption has been evaluated almost exclusively on factoid QA benchmarks, where hallucinations are defined relative to world knowledge and typically arise from missing or incorrect knowledge. Whether these methods generalise to faithfulness hallucinations remains an open empirical question that has not been systematically studied. If they do not, the growing use of UQ signals for trust estimation in RAG and document-grounded systems may rest on an unverified assumption. In addition, prior work has primarily studied UQ methods in isolation (Malinin and Gales, 2021; Kadavath et al., 2022; Manakul et al., 2023; Kuhn et al., 2023; Tian et al., 2023; Farquhar et al., 2024), while systematic evaluation of combinations or ensembles of different UQ signals remains relatively underexplored and is often treated as a secondary design choice rather than a primary object of analysis. As a result, the extent to which different uncertainty signals complement each other across hallucination types is still not well understood.

A further limitation is that many uncertainty

Table 1: Summary of UQ methods and main findings across factual and faithfulness hallucination benchmarks. † Proposed method.

Method	Factual benchmarks	Faithfulness benchmarks
Semantic entropy (Kuhn et al., 2023)	Below predictive entropy on most datasets; semantic clustering offers limited benefit for short-form QA	Near chance across settings; semantic equivalence provides little signal for context-grounding failures
Verbalised confidence (Tian et al., 2023)	Generally above chance and complementary to entropy-based signals in ensembles	Frequently inverted on FaithEval-Unanswerable (Ming et al., 2025); otherwise near chance
Perturbation (Gao et al., 2024)	Competitive on factual benchmarks and complementary to verbalised confidence	Above chance in counterfactual settings (Ming et al., 2025; Ju et al., 2026); near chance elsewhere
Predictive entropy (Tomani et al., 2024)	Often below chance; uniform token aggregation yields weak discrimination	Near chance across most settings
DARE-PE†	Strongest individual method overall; improves over predictive entropy through relevance-weighted aggregation	Strongest individual method for open-ended QA, but near chance on multiple-choice faithfulness benchmarks
Ensembles (norm-avg, rank-avg, product, max)	Improve performance when uncertainty signals are complementary	Limited gains when constituent methods are near chance or inverted

estimates are derived directly from token-level probabilities or output perturbations (Kuhn et al., 2023; Malinin and Gales, 2021; Farquhar et al., 2024), which can be dominated by semantically uninformative tokens. To address this, we introduce **DARE-PE** (Deletion-based Approximation for Relevance-weighted Entropy), a black-box, deletion-based reweighting scheme inspired by Shifting Attention to more Relevant (SAR) (Duan et al., 2024). Unlike the original SAR method, which relies on internal model probabilities, DARE-PE approximates token importance by measuring output changes under token deletion. These importance scores are then used to reweight predictive entropy, allowing us to test whether downweighting uninformative tokens improves uncertainty estimation without white-box access.

We conduct a structured evaluation of six UQ approaches: semantic entropy, predictive entropy, verbalised confidence, perturbation-based uncertainty, DARE-PE, and post-hoc ensembles. Table 1 summarises the methods evaluated and their main empirical behaviour across selected factual and faithfulness benchmarks.

This paper makes four contributions:

1. We provide a systematic comparison of UQ methods across factual and faithfulness hallucination settings, enabling direct comparison of their behaviour across hallucination types.

2. We propose DARE-PE, a deletion-based approximation of SAR-style token importance that does not require access to internal model probabilities, and apply it to predictive entropy in a black-box setting.

3. We show that multiple output-level uncertainty signals are complementary on factual hallucination detection, and that the reweighted variant and simple ensembles yield observable improvements over standard individual methods.

4. We provide a characterisation of when and why output-level UQ fails on faithfulness benchmarks, showing that uncertainty detectability depends on error structure.

2 Related Work

Hallucination taxonomies. Several recent surveys distinguish between different forms of hallucination in LLM outputs. Huang et al. (2025) separate *factuality hallucinations*, which involve incorrect claims about world knowledge, from *faithfulness hallucinations*, which arise when model outputs deviate from the provided input or context. Ji et al. (2023) describe a related distinction between *extrinsic* hallucinations, which introduce unsupported content, and *intrinsic* hallucinations, which directly contradict source material. Across these frameworks, a common implication is that uncertainty signals associated with miss-

ing knowledge may behave differently in context-grounded settings, where models can produce context-inconsistent outputs with high confidence.

Uncertainty quantification for hallucination detection. A broad range of UQ methods have been proposed for hallucination detection. Semantic entropy estimates uncertainty by grouping semantically equivalent generations under repeated sampling and computing entropy over the resulting semantic clusters, improving robustness to paraphrastic variation (Kuhn et al., 2023; Lin et al., 2024; Zhang et al., 2023; Farquhar et al., 2024; Kossen et al., 2024). Predictive entropy measures mean token-level uncertainty during generation (Malinin and Gales, 2021; Tomani et al., 2024; Kadavath et al., 2022). Recent work further shows that standard predictive entropy can disproportionately weight semantically uninformative tokens such as articles and punctuation, motivating token-importance reweighting approaches that better align uncertainty estimates with semantic content (Duan et al., 2024). Verbalised confidence methods prompt models to report their own confidence levels (Lin et al., 2022; Tian et al., 2023; Xiong et al., 2024; Zhao et al., 2025). Perturbation-based methods estimate uncertainty through output consistency under semantically equivalent prompt reformulations (Manakul et al., 2023; Wang et al., 2023; Gao et al., 2024).

Representation-based hallucination detection. Prior work has shown that output-level signals may be insufficient for reliably detecting hallucinations. Simhi et al. (2024) find that some incorrect generations are indistinguishable from outputs alone even when the relevant knowledge is encoded in model representations. Similarly, Obeso et al. (2025) demonstrate that probes over hidden states outperform output-only uncertainty signals in long-form generation tasks. While effective, these approaches require white-box access to internal activations, limiting their applicability in API-only settings.

Gap. Existing comparative studies of uncertainty quantification methods (Kuhn et al., 2023; Tomani et al., 2024; Gao et al., 2024) primarily evaluate factual QA benchmarks, leaving open whether such methods generalise to faithfulness hallucination, how different uncertainty signals interact across settings, and whether token-importance reweighting can be achieved without access to model internals. This work addresses these three questions.

3 Experimental Design

Our central hypothesis is that hallucination type significantly affects the effectiveness of output-level uncertainty signals. We evaluate this by structuring benchmarks along two axes: hallucination type (factual versus faithfulness) and task format (open-ended versus multiple-choice). This design enables controlled comparison of uncertainty behaviour across settings that vary in grounding requirements and output constraints.

3.1 Datasets

Factual open-ended QA uses TriviaQA (Joshi et al., 2017) and SimpleQA (Wei et al., 2024); factual multiple-choice QA uses OpenBookQA (Mihaylov et al., 2018) and ARC-Challenge (Clark et al., 2018). Faithfulness open-ended QA uses MRCEval (Ma et al., 2025) and Faithfulness-QA (Ju et al., 2026), while faithfulness multiple-choice QA uses FaithEval-Counterfactual and FaithEval-Unanswerable (Ming et al., 2025). Dataset statistics and additional construction details are given in Appendix A.

3.2 UQ Methods

Semantic entropy (Kuhn et al., 2023). Semantic entropy estimates uncertainty by measuring the diversity of *meanings* expressed across multiple sampled generations rather than surface-form variation alone. High entropy indicates that the model produces semantically inconsistent answers to the same question, while low entropy reflects agreement across generations. We sample five responses per question at temperature $T = 1.0$. Responses are clustered via bidirectional entailment using DeBERTa-v3-large (He et al., 2023) with threshold 0.7. Semantic entropy is then computed over the resulting cluster-assignment distribution.

Predictive entropy (Tomani et al., 2024). Predictive entropy measures the uncertainty of the model's next-token distribution during generation, with higher entropy indicating greater dispersion over candidate continuations and therefore lower model confidence. We compute mean token-level entropy from log-probabilities obtained during a single generation ($T = 1.0$, top-5 alternatives per position).

Verbalised confidence (Tian et al., 2023). After each generation, the model outputs a confidence score on a $[0, 100]$ scale. We invert scores as $s =$

$(100 - c)/100$ so that higher values correspond to greater uncertainty.

Perturbation-based uncertainty (Gao et al., 2024). We estimate uncertainty from the consistency of model outputs under semantically neutral prompt perturbations. For each question, we apply four paraphrastic prompt variants that preserve the original meaning and generate one response per variant. We then compute pairwise ROUGE-L similarity scores across all generated responses and define uncertainty as $1 - \overline{\text{ROUGE-L}}$, where $\overline{\text{ROUGE-L}}$ denotes the mean pairwise ROUGE-L score (Lin, 2004). Higher values therefore correspond to greater output instability.

DARE-PE Standard predictive entropy aggregates token-level negative log-likelihoods uniformly, which can be dominated by semantically uninformative tokens such as articles and punctuation. SAR (Duan et al., 2024) addresses this limitation by reweighting token entropy according to semantic relevance, yielding a more informative uncertainty estimate. However, SAR requires access to internal token-level cross-entropies, which are unavailable in API-only settings. We propose **DARE-PE** (Deletion-based Approximation for Relevance-weighted Entropy), a black-box approximation of SAR-style reweighting that operates solely on model outputs and produces a scalar uncertainty score per response.

Let q denote the input prompt and let $\hat{a} = (z_1, \ldots, z_n)$ denote the generated response consisting of tokens z_i. We define token relevance by measuring the semantic impact of deleting each token:

$$R(z_i) = 1 - \text{sim}(q \oplus \hat{a},\ q \oplus \hat{a}_{-i}),$$

where $\hat{a}_{-i}$ denotes the response with token z_i removed, $\oplus$ denotes string concatenation, and $\text{sim}(\cdot, \cdot)$ is a semantic similarity score produced by `cross-encoder/stsb-roberta-large`, a RoBERTa-large model fine-tuned on the STS-B regression task (Cer et al., 2017). Because similarity is computed over word-level segments while generation operates on subword tokens, we map word-level relevance scores to model tokens via linear interpolation and normalise them as $R_{\text{norm}}(z_i) = R(z_i)/\sum_{j=1}^{n} R(z_j)$, such that $\sum_i R_{\text{norm}}(z_i) = 1$. Let $\text{NLL}(z_i \mid q, z_{<i})$ denote the negative log-likelihood of token z_i conditioned on the prompt and preceding context. The DARE-PE score is then defined as

$$\text{DARE-PE} = \sum_{i=1}^{n} R_{\text{norm}}(z_i)\, \text{NLL}(z_i \mid q, z_{<i}).$$

This formulation downweights tokens with low semantic impact and emphasises semantically informative tokens. Unlike perturbation-based approaches, DARE-PE requires only a single model output, avoiding multiple API calls across prompt variants.

Ensembles We construct a three-signal ensemble combining perturbation-based uncertainty, DARE-PE, and verbalized confidence. We exclude semantic entropy, which in our experiments performs near or below chance on most benchmarks, and therefore does not provide complementary signal in the ensemble. We also exclude standard predictive entropy in favor of DARE-PE, a reweighted variant, as DARE-PE is observed to yield stronger discrimination than uniform token-level aggregation and empirically subsumes much of the predictive entropy signal.

The three retained signals correspond to distinct uncertainty estimation paradigms—output consistency under perturbations, token-importance-reweighted entropy, and self-reported confidence—and may capture complementary aspects of model behavior. We evaluate four aggregation strategies: (i) equal-weight averaging of min–max normalized scores (*norm-avg*); (ii) equal-weight averaging of percentile ranks (*rank-avg*); (iii) product of percentile ranks (*product*); and (iv) maximum of percentile ranks (*max*). All ensembles are constructed post-hoc from observed scores and require no additional API calls.

3.3 Evaluation Metrics

We evaluate hallucination detection performance using **Area Under the Receiver Operating Characteristic Curve** (AUROC), which measures how effectively uncertainty scores distinguish hallucinated from non-hallucinated responses. For open-ended tasks, hallucination labels are assigned using DeBERTa-v3-large (He et al., 2023) entailment-based matching, while multiple-choice tasks use exact option comparison. An AUROC of 1.0 indicates perfect separation, whereas 0.5 corresponds to chance-level discrimination; values below 0.5 indicate inverted ranking, where hallucinated responses receive systematically lower uncertainty

Table 2: AUROC and model accuracy on extrinsic/factuality hallucination datasets. Red denotes below-chance discrimination (AUROC < 0.5). **Bold** denotes best AUROC per model per dataset.

			Individual UQ Methods					Ensemble Methods			
Dataset	**Model**	**Acc.**	**Verb.**	**Sem.**	**Pred.**	**Pert.**	**DARE-PE**	**Norm-avg**	**Rank-avg**	**Product**	**Max**
Open-ended QA											
TriviaQA	DS-V3	99.2%	0.429	0.228	0.218	0.631	**0.932**	0.728	0.755	0.728	0.858
	LLaMA	97.2%	0.650	0.386	0.194	0.637	0.700	0.680	0.719	0.710	**0.720**
	Qwen	93.6%	0.693	0.506	0.298	0.776	0.780	0.780	**0.824**	0.823	0.786
SimpleQA	DS-V3	61.2%	0.581	0.499	0.365	0.590	**0.698**	0.671	0.670	0.679	0.650
	LLaMA	34.0%	**0.551**	0.522	0.531	0.467	0.470	0.498	0.469	0.503	0.446
	Qwen	58.4%	0.543	0.464	0.432	0.567	0.604	0.599	**0.615**	0.613	0.579
Multiple-choice QA											
ARC-C	DS-V3	94.5%	0.576	0.451	0.321	0.726	**0.877**	0.869	0.875	0.876	0.877
	LLaMA	77.4%	0.626	0.456	0.295	0.632	**0.800**	0.725	0.761	0.743	0.766
	Qwen	92.5%	0.636	0.483	0.396	0.469	**0.630**	0.563	0.613	0.612	0.558
OpenBookQA	DS-V3	84.8%	0.676	0.424	0.333	0.796	0.798	0.840	0.851	0.837	**0.860**
	LLaMA	78.0%	0.707	0.519	0.354	0.638	0.781	**0.794**	0.790	0.759	0.787
	Qwen	92.0%	0.676	0.479	0.394	0.758	0.798	**0.861**	0.817	0.813	0.810

than non-hallucinated responses. We prefer AUROC because it is threshold-independent, robust to class imbalance, and supports bootstrap confidence interval estimation (Appendix C).

3.4 Models

We evaluate three open-source LLMs through OpenAI-compatible APIs: DeepSeek-V3.2-Exp (DeepSeek-AI et al., 2025), Qwen-3.5-Flash (Yang et al., 2025), and LLaMA-3.1-8B-Instruct (Grattafiori et al., 2024). Although model weights are publicly available, we interact with all models exclusively through chat completion APIs, restricting methods to the observable outputs available in closed-source commercial settings, namely generated text and token-level log-probabilities. This setup ensures that all UQ methods, including DARE-PE, are evaluated under comparable black-box constraints, such that performance differences reflect method behaviour rather than privileged access to model internals. At the same time, the availability of open-source weights enables implementation verification against white-box quantities where needed, while all reported experiments remain strictly API-based. Results are therefore expected to generalise to closed-source deployments, subject to the limitations discussed in Section 6.

4 Results

4.1 Factual Hallucination Benchmarks

Table 2 reports AUROC for each UQ method, model, and factual benchmark. Across these results, DARE-PE achieves the highest AUROC in the majority of model–dataset combinations, indicating that its uncertainty scores most reliably rank incorrect responses above correct ones. In particular, DARE-PE substantially outperforms standard predictive entropy and often exceeds perturbation-based uncertainty. The largest gains appear on DeepSeek-V3, where DARE-PE reaches 0.932 AUROC on TriviaQA and 0.877 on ARC-C. In contrast, standard predictive entropy is frequently below chance, indicating that uniformly aggregating token uncertainty provides weak discrimination for factual hallucination detection. Perturbation-based uncertainty and verbalised confidence are more consistent across datasets. Verbalised confidence performs relatively better on open-ended QA, while perturbation-based uncertainty is stronger on several multiple-choice settings. Semantic entropy is generally weak, remaining near or below chance on most benchmarks.

Ensemble methods improve performance when individual signals are complementary. On TriviaQA with Qwen, rank-avg achieves the best overall AUROC (0.824), improving over DARE-PE alone (0.780). Similar gains appear on OpenBookQA, where ensemble aggregation consistently exceeds individual methods. However, ensembles do not uniformly outperform DARE-PE; on TriviaQA with DeepSeek-V3, DARE-PE alone remains strongest. This suggests that ensemble gains depend on the relative diversity of the underlying signals rather than aggregation itself. Figure 1 illustrates this complementarity on OpenBookQA with LLaMA. DARE-PE and verbalised confidence

Table 3: AUROC and model accuracy on intrinsic/faithfulness hallucination datasets. Red denotes below-chance discrimination (AUROC < 0.5). **Bold** denotes best AUROC per model per dataset.

			Individual UQ Methods					Ensemble Methods			
Dataset	**Model**	**Acc.**	**Verb.**	**Sem.**	**Pred.**	**Pert.**	**DARE-PE**	**Norm-avg**	**Rank-avg**	**Product**	**Max**
Open-ended QA											
	DS-V3	49.8%	0.474	0.513	0.457	0.546	**0.595**	0.531	0.556	0.558	0.540
MRCEval	LLaMA	42.8%	0.505	0.499	0.472	0.538	**0.564**	0.542	0.556	0.556	0.543
	Qwen	49.6%	0.488	0.487	0.459	0.522	**0.541**	0.495	0.522	0.539	0.500
	DS-V3	77.6%	0.473	0.478	0.341	0.674	**0.726**	0.653	0.692	0.690	0.673
Faithfulness-QA	LLaMA	71.4%	0.446	0.465	0.380	0.642	**0.655**	0.569	0.624	0.638	0.577
	Qwen	66.0%	0.456	0.534	0.395	0.601	**0.638**	0.512	0.600	0.609	0.560
Multiple-choice QA											
	DS-V3	45.8%	0.478	0.523	0.523	0.505	0.482	0.483	0.482	0.450	**0.532**
FE-CF	LLaMA	64.8%	0.581	0.470	0.391	0.670	0.751	**0.757**	0.755	0.738	0.606
	Qwen	46.1%	0.481	0.526	0.514	0.505	0.477	0.485	0.481	0.450	**0.525**
	DS-V3	75.0%	0.122	0.481	0.556	0.464	0.152	0.413	0.385	**0.639**	0.528
FE-UN	LLaMA	72.0%	0.086	0.500	0.379	**0.692**	0.524	0.174	0.371	0.318	0.483
	Qwen	66.6%	0.101	0.524	**0.767**	0.294	0.145	0.126	0.441	0.426	0.630

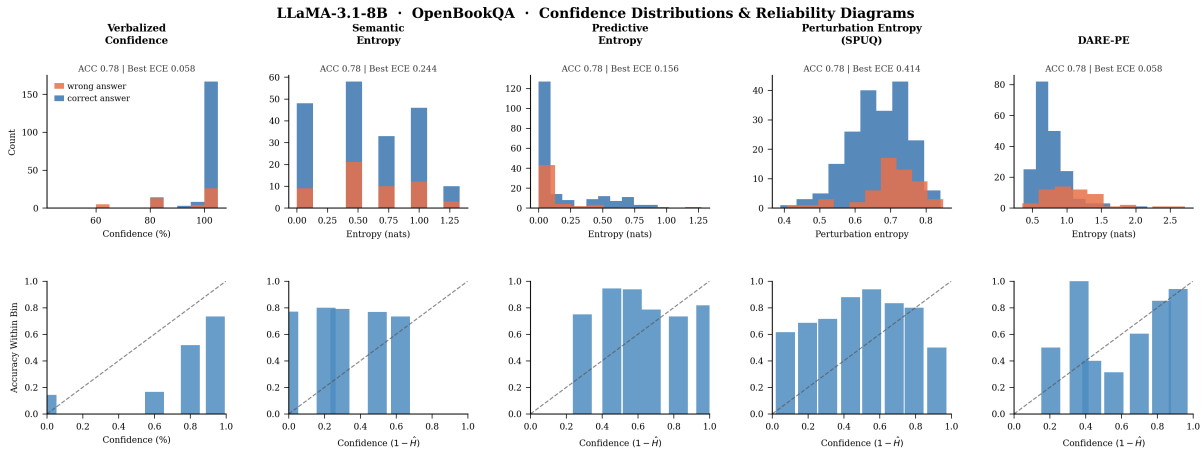

Figure 1: Confidence distributions (top) and reliability diagrams (bottom) on OpenBookQA with LLaMA-3.1-8B.

produce clearer separation between correct and incorrect predictions than standard predictive entropy, while the reliability diagrams indicate partially distinct calibration behaviour. SimpleQA is consistently more difficult across all methods and models, with AUROC values concentrated around 0.5–0.7. The lower scores likely reflect the benchmark's broader distribution of knowledge failures, reducing the separability of correct and incorrect outputs through output-level uncertainty alone.

4.2 Faithfulness Hallucination Benchmarks

Performance on faithfulness benchmarks is substantially weaker and more heterogeneous than on factual benchmarks (Table 3).

On MRCEval, nearly all methods remain close to chance across models. Although DARE-PE is consistently the strongest individual method, gains are small (0.541–0.595 AUROC), and ensembles provide limited improvement. This suggests that MRCEval errors do not produce a stable output-level uncertainty signal. Figure 2(b) shows a representative case where the model answers incorrectly with high confidence despite relying on world knowledge rather than the provided context.

Faithfulness-QA yields a different pattern. Perturbation-based uncertainty and DARE-PE remain consistently above chance across all models, while verbalised confidence and predictive entropy are uniformly poor and often below chance. This indicates that counterfactual context conflicts produce detectable instability under perturbation even when token-level probabilities remain highly confident. Results on multiple-choice faithfulness benchmarks are again dataset-dependent. On FaithEval-Counterfactual, LLaMA shows strong

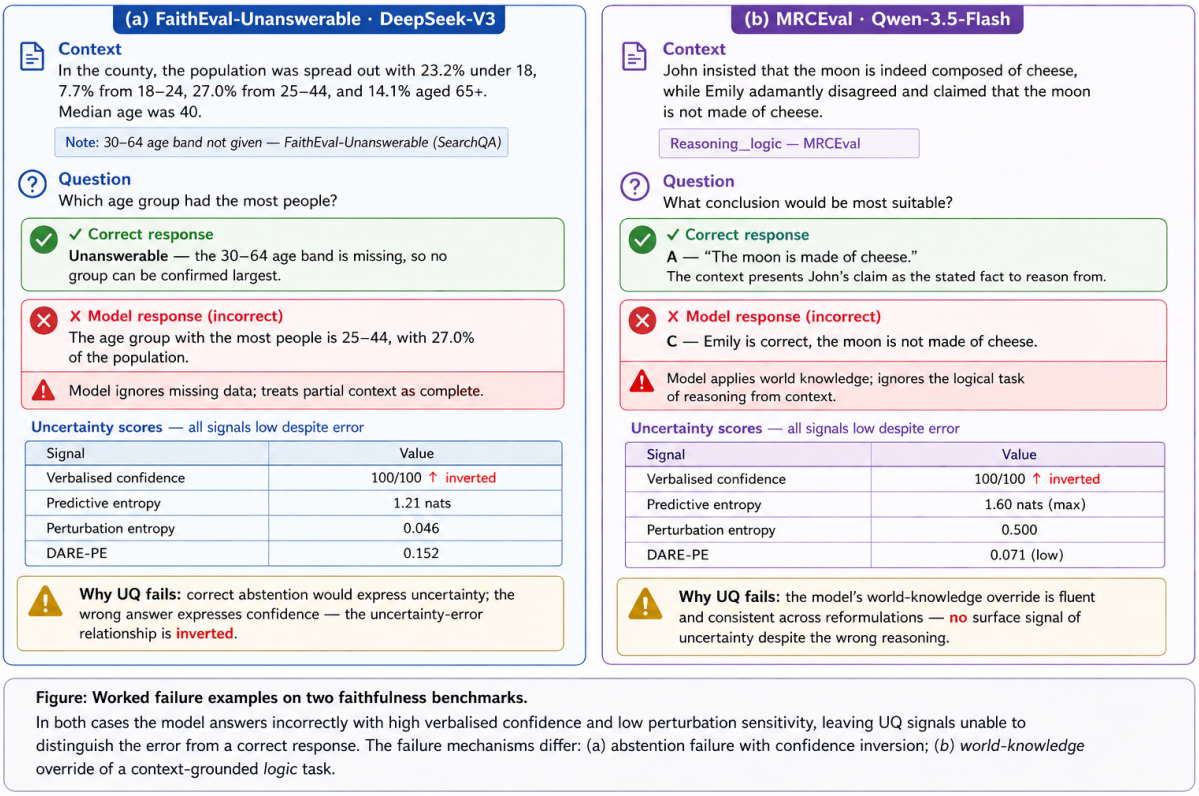

Figure 2: Worked failure examples illustrating two distinct error mechanisms on faithfulness benchmarks.

performance from perturbation-based uncertainty (0.670), DARE-PE (0.751), and norm-avg ensembling (0.757), whereas DeepSeek-V3 and Qwen remain near chance. FaithEval-Unanswerable produces the most unstable behaviour: verbalised confidence collapses across all models (0.086–0.122 AUROC), while predictive entropy and perturbation-based uncertainty remain informative in selected settings. The near-zero AUROC of verbalised confidence reflects systematic inversion: models that incorrectly answer unanswerable questions often do so with high confidence.

Overall, faithfulness performance depends strongly on benchmark structure. Methods are most effective when contextual evidence directly conflicts with model priors, as in Faithfulness-QA and FaithEval-Counterfactual, and least effective when failures arise from broader comprehension errors or incorrect abstention behaviour.

5 Discussion

Across benchmarks, DARE-PE is the most consistently strong individual method and substantially improves over standard predictive entropy. Ensemble methods provide additional gains on factual hallucination benchmarks when different uncertainty signals are complementary, but offer limited benefit in near-chance faithfulness settings.

More broadly, factual hallucination benchmarks yield substantially higher AUROC than faithfulness benchmarks. However, the gap is not uniform across datasets and depends on error structure. Output-level uncertainty signals are most informative when hallucinations produce instability or conflict with contextual evidence, and least informative when failures stem from general comprehension errors or confident but incorrect abstention behaviour.

Reweighting improves predictive entropy. DARE-PE consistently outperforms standard predictive entropy across most benchmarks, suggesting that uniform aggregation over tokens obscures useful uncertainty signals. Reweighting entropy toward semantically important tokens substantially improves discrimination even in a fully black-box setting, indicating that observable output perturbations provide a viable approximation of token relevance.

Complementarity depends on hallucination type. On factual benchmarks, perturbation-based uncertainty, verbalised confidence, and entropy-based methods capture partially distinct aspects of model

behaviour, enabling ensemble gains when no single signal dominates. On faithfulness benchmarks, this complementarity is weaker. When context-inconsistent outputs remain fluent and internally coherent, most output-level signals collapse toward chance simultaneously, limiting the benefit of aggregation.

Faithfulness errors differ in detectability. The large variation across faithfulness benchmarks suggests that faithfulness hallucination is not a single failure mode for UQ methods. MRCEval mixes context-ignoring with broader comprehension failures, producing little consistent uncertainty signal. By contrast, Faithfulness-QA and FaithEval-Counterfactual explicitly create conflict between contextual evidence and parametric knowledge, leading to greater perturbation sensitivity. FaithEval-Unanswerable differs again: incorrect responses are often generated confidently, inverting the expected uncertainty–error relationship.

Verbalised confidence fails on abstention tasks. The near-zero AUROC of verbalised confidence on FaithEval-Unanswerable reflects systematic inversion rather than weak calibration. Models that incorrectly answer unanswerable questions often report high confidence, while correct abstention is associated with uncertainty. This suggests that self-reported confidence may be unreliable in settings where correctness depends on recognising insufficient evidence.

Accuracy and uncertainty discrimination are partially decoupled. High task accuracy does not necessarily imply strong uncertainty discrimination. On TriviaQA, DeepSeek-V3 achieves near-perfect accuracy while AUROC varies substantially across methods, whereas SimpleQA produces weak AUROC despite many model errors. Output-level UQ therefore appears most informative when errors are neither too rare nor uniformly distributed across questions.

6 Conclusion

We conducted a systematic study of uncertainty quantification (UQ) methods for hallucination detection across eight benchmarks spanning factual and faithfulness settings. We compared standard entropy-based and perturbation-based approaches, verbalised confidence, DARE-PE, our deletion-based reweighted predictive entropy method, and ensemble combinations thereof under a unified evaluation framework.

Our first contribution is demonstrating that reweighting token-level entropy by semantic relevance yields a consistently stronger black-box uncertainty signal. Across benchmarks, DARE-PE provides the most reliable individual performance, indicating that improving *how uncertainty is aggregated over tokens* can be more impactful than designing new uncertainty heuristics.

Second, we show that the effectiveness of UQ methods is fundamentally conditioned on hallucination structure. Uncertainty signals are highly informative when errors induce instability or explicit conflict between contextual evidence and parametric knowledge, but degrade sharply when failures arise from broad comprehension errors or confident abstention mistakes. This reveals that hallucination detection is not a single uniform task, but a collection of heterogeneous failure modes with different observability properties.

Third, we find that ensemble gains are limited by signal collapse under certain error regimes. While complementary methods improve detection on factual benchmarks, their benefits diminish when individual signals become jointly uninformative, suggesting that diversity alone is insufficient without sensitivity to distinct error types.

Taken together, these findings suggest a shift in how uncertainty-based hallucination detection should be approached. Rather than treating UQ as a general-purpose scoring function over model outputs, future work should explicitly model the *type of reasoning failure* underlying a prediction and adapt uncertainty estimation accordingly. In particular, progress is likely to come from methods that (i) distinguish between comprehension, grounding, and abstention failures, and (ii) incorporate structured signals beyond output-level perturbations and token likelihoods.

This work therefore positions hallucination detection as a structured error analysis problem rather than a single-axis uncertainty estimation task, and provides empirical evidence that different hallucination regimes require specialised uncertainty modelling strategies.

Limitations

Our study has several limitations.

Benchmark coverage and sample size. We evaluate benchmark subsets rather than full test sets,

which may reduce the stability of AUROC estimates, particularly under highly imbalanced error distributions. Results on TriviaQA with DeepSeek-V3 and SimpleQA with LLaMA should therefore be interpreted cautiously.

Black-box access constraints. All methods operate using observable outputs and top-5 log-probability alternatives only. This excludes richer signals such as full vocabulary distributions and internal representations, which may improve hallucination detection performance.

Approximate reweighting implementation. Our DARE-PE implementation approximates token alignment using whitespace tokenisation and nearest-index projection. This may introduce deviation from the original SAR formulation, particularly for subword-heavy tokens, and we do not directly evaluate approximation fidelity.

Ensemble construction. Ensembles are constructed post-hoc using fixed aggregation strategies without supervised calibration or learned combination weights. While supervised ensembles may improve performance, they would require labelled data and reduce deployment flexibility.

Faithfulness benchmark heterogeneity. The evaluated faithfulness benchmarks differ substantially in task construction and failure mode, which likely contributes to the observed variance in AUROC. A more fine-grained taxonomy of faithfulness errors would help clarify when output-level uncertainty signals remain effective.

Model scope. We evaluate three open-source models accessed through OpenAI-compatible APIs. Results may not generalise to closed-source frontier systems or models specifically optimised for calibration or faithfulness.

Potential Risks

This work studies uncertainty quantification (UQ) methods for hallucination detection in large language models. Although improved uncertainty estimation may support safer deployment of LLM systems, these methods also carry potential risks if uncertainty scores are interpreted as reliable indicators of factuality or faithfulness in all settings.

Our results show that several output-level UQ methods perform near chance on certain faithfulness hallucination benchmarks, particularly when errors arise from broader comprehension failures or confident incorrect abstention. Deploying such methods as standalone trust or verification signals in high-stakes applications such as medical, legal, or educational systems could therefore create false confidence in incorrect model outputs.

There is also a risk that benchmark-specific evaluations may overestimate the generalisability of hallucination detection methods. Existing UQ research has focused predominantly on factual QA benchmarks, whereas our findings suggest that performance depends strongly on hallucination type and error structure. Future work should therefore avoid treating hallucination detection as a uniform task and should evaluate uncertainty methods across more diverse grounding and reasoning settings.

Finally, this work evaluates only black-box output-level uncertainty signals and does not provide guarantees of factual correctness, faithfulness, or calibration under deployment conditions. We recommend that such methods be used only as auxiliary indicators alongside human oversight or external verification mechanisms rather than as standalone reliability measures.

References

Daniel Cer, Mona Diab, Eneko Agirre, Iñigo Lopez-Gazpio, and Lucia Specia. 2017. SemEval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 1–14, Vancouver, Canada. Association for Computational Linguistics.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try ARC, the AI2 reasoning challenge. *Computing Research Repository, arXiv:1803.05457.*

DeepSeek-AI, Aixin Liu, Aoxue Mei, Bangcai Lin, Bing Xue, Bingxuan Wang, Bingzheng Xu, Bochao Wu, Bowei Zhang, Chaofan Lin, Chen Dong, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenhao Xu, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, and 244 others. 2025. DeepSeek-V3.2: Pushing the frontier of open large language models. *Computing Research Repository, arXiv:2512.02556.*

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

pages 5050–5063, Bangkok, Thailand. Association for Computational Linguistics.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. 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)*, pages 2336–2346, St. Julian's, Malta. Association for Computational Linguistics.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 217 others. 2024. The Llama 3 herd of models. *Computing Research Repository, arXiv:2407.21783*.

Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2023. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of the Eleventh International Conference on Learning Representations*.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. 2025. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):42:1–42:55.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM Computing Surveys*, 55(12):1–38.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Li Ju, Junzhe Wang, and Qi Zhang. 2026. Faithfulness-qa: A counterfactual entity substitution dataset for training context-faithful rag models. *Computing Research Repository, arXiv:2604.25313*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *CoRR*, abs/2207.05221.

Jannik Kossen, Jiatong Han, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal. 2024. Semantic entropy probes: Robust and cheap hallucination detection in llms. *CoRR*, abs/2406.15927.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *Proceedings of the Eleventh International Conference on Learning Representations*.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Teaching models to express their uncertainty in words. *Transactions on Machine Learning Research*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2024. Generating with confidence: Uncertainty quantification for black-box large language models. *Transactions on Machine Learning Research*.

Shengkun Ma, Hao Peng, Lei Hou, and Juanzi Li. 2025. Mrceval: A comprehensive, challenging and accessible machine reading comprehension benchmark. *Computing Research Repository, arXiv:2503.07144*.

Andrey Malinin and Mark Gales. 2021. Uncertainty estimation in autoregressive structured prediction. In *Proceedings of the Ninth International Conference on Learning Representations*.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

Todor Mihaylov, Peter Clark, Tushar Khot, and Ashish Sabharwal. 2018. Can a suit of armor conduct electricity? a new dataset for open book question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2381–2391, Brussels, Belgium. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. FaithEval: Can your language model stay faithful to context, even if "the moon is made of marshmallows"? In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Oscar Obeso, Andy Arditi, Javier Ferrando, Joshua Freeman, Cameron Holmes, and Neel Nanda. 2025.

Real-time detection of hallucinated entities in long-form generation. *Computing Research Repository, arXiv:2509.03531*.

Adi Simhi, Jonathan Herzig, Idan Szpektor, and Yonatan Belinkov. 2024. Distinguishing ignorance from error in LLM hallucinations. *Computing Research Repository, arXiv:2410.22071*.

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

Christian Tomani, Kamalika Chaudhuri, Ivan Evtimov, Daniel Cremers, and Mark Ibrahim. 2024. Uncertainty-based abstention in LLMs improves safety and reduces hallucinations. *Computing Research Repository, arXiv:2404.10960*.

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 *Proceedings of the Eleventh International Conference on Learning Representations*.

Jason Wei, Karina Nguyen, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. 2024. Measuring short-form factuality in large language models. *Computing Research Repository, arXiv:2411.04368*.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can LLMs express their uncertainty? an empirical evaluation of confidence elicitation in LLMs. In *Proceedings of the 12th International Conference on Learning Representations*.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *Computing Research Repository*, arXiv:2505.09388.

Jiaxin Zhang, Zhuohang Li, Kamalika Das, Bradley Malin, and Sricharan Kumar. 2023. SAC^3: Reliable hallucination detection in black-box language models via semantic-aware cross-check consistency. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 15445–15458, Singapore. Association for Computational Linguistics.

Lili Zhao, Yang Wang, Qi Liu, Mengyun Wang, Wei Chen, Zhichao Sheng, and Shijin Wang. 2025. Evaluating large language models through role-guide and self-reflection: A comparative study. In *Proceedings of the Thirteenth International Conference on Learning Representations*.

Dataset	Task type	Size	Answer format	Category
TriviaQA	Open-ended QA	17,210	Free-form	Factual
SimpleQA	Open-ended QA	4,326	Free-form	Factual
OpenBookQA	Multiple-choice	500	4-way MC	Factual
ARC-Challenge	Multiple-choice	1,172	4-way MC	Factual
MRCEval	Open-ended QA	2,100	Free-form	Faithfulness
Faithfulness-QA	Open-ended QA	9,910	Free-form	Faithfulness
FaithEval-Counterfactual	Multiple-choice	1,000	4-way MC	Faithfulness
FaithEval-Unanswerable	Multiple-choice	2,490	Binary MC	Faithfulness

Table 4: Summary statistics for all evaluation benchmarks.

A Dataset Details

Table 4 summarises the publicly available research benchmarks used in our experiments. We provide additional details below on each dataset's task format, construction, and role in our factual and faithfulness hallucination evaluations.

Factual hallucination: open-ended QA. TriviaQA (Joshi et al., 2017) contains 95K question-answer pairs authored by trivia enthusiasts across diverse topics, paired with independently gathered evidence documents from Wikipedia and the web (on average six documents per question). Questions span a wide range of domains including history, science, geography, and popular culture, and are notable for requiring cross-sentence inference and handling of lexical variability. The test split used in our experiments contains 17,210 questions; we follow common practice and evaluate on the open-domain (no-context) setting, treating model answers as free-form strings judged against gold aliases.

SimpleQA (Wei et al., 2024) consists of 4,326 short, fact-seeking questions designed to have a single unambiguous answer, covering topics including science, history, technology, and entertainment. Each question was independently verified by two human annotators, yielding an estimated inherent error rate of approximately 3%. Unlike TriviaQA, SimpleQA was constructed adversarially against GPT-4 responses to ensure the benchmark remains challenging for frontier models; leading models at time of release achieved below 50% accuracy. Responses are graded as correct, incorrect, or not attempted using a prompted LLM classifier.

Factual hallucination: multiple-choice QA. OpenBookQA (Mihaylov et al., 2018) contains 5,957 four-way multiple-choice questions at the elementary science level, grounded in a small "open book" of 1,326 core science facts. Answering correctly requires combining a relevant fact from the open book with broader commonsense knowledge not contained in the fact set, making it a multi-hop reasoning benchmark rather than a retrieval task. The standard test split comprises 500 questions; our experiments use this split exclusively.

ARC-Challenge (Clark et al., 2018) is drawn from the AI2 Reasoning Challenge dataset of 7,787 grade-school science questions. The Challenge partition (2,590 questions, of which 1,172 constitute the test split) retains only those questions answered incorrectly by both an information retrieval baseline and a pointwise mutual information co-occurrence model, thereby filtering out questions solvable by surface-level pattern matching. Questions are predominantly four-way multiple-choice and span grades 3–9, requiring genuine multi-step reasoning and knowledge integration.

Faithfulness hallucination: open-ended QA. MRCEval (Ma et al., 2025) is a machine reading comprehension benchmark comprising approximately 2,100 questions organised across 13 distinct reading comprehension skills, grouped into three levels: context comprehension, external knowledge comprehension, and reasoning. Crucially, MRCEval includes question types specifically targeting knowledge conflict and hallucination, evaluating whether models answer using only the provided passage rather than parametric knowledge. Questions were generated by LLMs and filtered through a multi-stage quality selection procedure. Although the original benchmark is formatted as four-way multiple-choice, in our experiments we treat it as an open-ended task: the model generates a free-form answer, and correctness is determined using DeBERTa-v3-large (He et al., 2023) entailment-based matching. This open-ended formulation is consistent with the treatment applied to all other free-form datasets in our suite, enabling direct comparison of uncertainty signals across tasks.

Faithfulness-QA (Ju et al., 2026) is a large-scale training and evaluation resource of 99,094 samples constructed through automatic counterfactual entity substitution applied to the SQuAD and TriviaQA training sets (49,094 and 50,000 samples respectively). For each instance, answer-bearing named entities in the original passage are replaced with type-consistent alternatives drawn from a curated entity bank of 76,953 entries, creating a controlled knowledge conflict between the provided context and the model's parametric memory. Each subset is split 80/10/10 into train, development, and test partitions; our experiments use the combined test split of 9,910 samples. Answer format is free-form, graded against the substituted entity string.

Faithfulness hallucination: multiple-choice QA. FaithEval-Counterfactual (Ming et al., 2025) contains 1,000 multiple-choice questions grounded in passages where factual content has been replaced with counterfactual statements contradicting common sense or world knowledge. The counterfactual contexts are constructed from ARC-Challenge, with answer options drawn directly from the passage, so a faithful model can recover the correct answer from the context alone even when it conflicts with prior knowledge. The benchmark was validated through a four-stage pipeline combining LLM-based generation and human review.

FaithEval-Unanswerable (Ming et al., 2025) contains 2,490 questions paired with passages that deliberately omit the information needed to answer them. A faithful model should respond with *unknown* rather than fabricating an answer from parametric memory. The subset is derived from ten diverse contextual QA datasets spanning a wide range of domains; gold responses were verified by majority vote among human annotators (98% agreement). In our experiments we cast this as a binary classification task: model responses are mapped to one of two labels — *answerable* or *unanswerable* — and correctness is determined by whether the model's generated response indicates it cannot answer (i.e., abstains appropriately). This binary framing reduces the task to a well-defined uncertainty signal: high uncertainty should correlate with correct abstention.

B Uncertainty Quantification Methods: Additional Details

This section provides extended descriptions of each uncertainty quantification (UQ) method used in our experiments, including implementation details, design rationale, and discussion of limitations that motivate our ensemble construction.

B.1 Semantic Entropy

Semantic entropy (Kuhn et al., 2023) addresses a fundamental limitation of token-level uncertainty measures: surface-form variation across generations does not imply semantic disagreement. Two responses such as *"Paris"* and *"The capital is Paris"* are lexically distinct but semantically equivalent, and should not contribute to uncertainty.

Implementation. We sample five responses per question at temperature $T = 1.0$. Responses are clustered into equivalence classes using bidirectional natural language inference (NLI): two responses r_i and r_j are placed in the same cluster if both $r_i \Rightarrow r_j$ and $r_j \Rightarrow r_i$ hold above a threshold of 0.7, as judged by DeBERTa-v3-large (He et al., 2023) fine-tuned on MNLI. Shannon entropy is then computed over the cluster-assignment distribution. Formally, if K clusters are induced with probabilities $(p_1, \ldots, p_K)$ estimated by normalized cluster frequencies, semantic entropy is:

$$H_{\text{sem}} = -\sum_{k=1}^{K} p_k \log p_k.$$

Practical considerations. Semantic entropy is among the more computationally expensive methods, requiring both multiple generations and $O(n^2)$ NLI calls for n samples. Clustering quality is sensitive to the entailment threshold and to the NLI model's ability to handle domain-specific or short-form answers. In our experiments, semantic entropy performs near or below chance on several benchmarks, consistent with findings in prior work suggesting it is most discriminative on long-form or abstractive generation settings rather than short factoid QA (Kuhn et al., 2023).

B.2 Predictive Entropy

Predictive entropy (Tomani et al., 2024) provides a token-level measure of generation confidence derived from the model's own output distribution, requiring only a single forward pass.

Implementation. We extract per-token log-probabilities during a single greedy generation ($T = 1.0$, top-5 token alternatives per position).

Mean token-level entropy is computed as:

$$H_{\text{pred}} = -\frac{1}{n}\sum_{i=1}^{n}\sum_{v} p(z_i = v \mid q, z_{<i}) \times \log p(z_i = v \mid q, z_{<i}), \quad (1)$$

where the inner sum is over the top-5 vocabulary alternatives at position i.

Limitation. Uniform aggregation over all tokens treats semantically uninformative tokens (articles, punctuation, connectives) equally with content-bearing tokens, which can dilute the signal. This motivates DARE-PE (§B.5), which applies learned relevance weights before aggregation.

B.3 Verbalised Confidence

Verbalised confidence (Tian et al., 2023) elicits an explicit self-assessment from the model rather than inferring uncertainty from the generation process. This is the only method in our suite that does not require access to token-level log-probabilities, making it applicable to black-box API settings.

Implementation. After each answer generation, the model is prompted to output a confidence score on a $[0, 100]$ scale. We invert the score as $s = (100 - c)/100$ so that higher values consistently denote greater uncertainty across all methods. No additional sampling is required beyond the confidence elicitation prompt.

Practical considerations. Verbalised confidence is susceptible to *overconfidence bias*: instruction-tuned models tend to express high confidence even when incorrect (Tian et al., 2023). Performance varies across model families and is sensitive to prompt wording. Despite this, verbalised confidence provides a complementary signal to entropy-based methods, as it captures a different facet of model behaviour — self-reported epistemic state — and in practice contributes meaningfully to the ensemble on faithfulness benchmarks where entropy-based measures are less discriminative.

B.4 Perturbation-Based Uncertainty

Perturbation-based uncertainty (Gao et al., 2024) estimates uncertainty from the stability of model outputs under semantically neutral prompt variations, without requiring access to internal model states.

Implementation. For each question, we generate four paraphrastic prompt variants that preserve the original meaning. One response is sampled per variant ($T = 1.0$). Pairwise ROUGE-L similarity (Lin, 2004) is computed across all $\binom{4}{2} = 6$ response pairs, and uncertainty is defined as:

$$u_{\text{SPUQ}} = 1 - \overline{\text{ROUGE-L}},$$

where $\overline{\text{ROUGE-L}}$ is the mean pairwise score. High values indicate that the model produces inconsistent outputs across paraphrases, suggesting unreliable parametric knowledge about the query.

Practical considerations. Paraphrase quality is important: prompt variants that alter meaning or difficulty introduce noise into the uncertainty estimate. In our setup, variants are generated once offline and fixed across all models and datasets to ensure comparability. SPUQ is relatively expensive (requiring k additional generations per question for k variants) but requires no access to log-probabilities, making it straightforwardly applicable to black-box models.

B.5 DARE-PE

DARE-PE (Deletion-based Approximation for Relevance-weighted Entropy) is our proposed black-box approximation of SAR-style (Duan et al., 2024) token importance reweighting, designed for settings where internal cross-entropies are unavailable.

Motivation. SAR reweights token-level entropy by each token's semantic contribution to the response, downweighting uninformative function words. This yields sharper uncertainty estimates than uniform predictive entropy but requires white-box access to per-token cross-entropies. DARE-PE approximates this reweighting through *deletion-based relevance*, a perturbation technique that measures each token's contribution by the semantic change its removal induces.

Implementation. Let q denote the input prompt and $\hat{a} = (z_1, \ldots, z_n)$ the generated response. The relevance of token z_i is:

$$R(z_i) = 1 - \text{sim}(q \oplus \hat{a},\ q \oplus \hat{a}_{-i}),$$

where $\hat{a}_{-i}$ is the response with z_i removed and $\text{sim}(\cdot,\cdot)$ is the output of `cross-encoder/stsb-roberta-large`, a RoBERTa-large model fine-tuned on the STS-B

regression task (Cer et al., 2017), which takes the two full strings as a single input pair and outputs a scalar similarity in $[0, 1]$. Because the cross-encoder operates at the word level while generation produces subword tokens, word-level relevance scores are mapped to model tokens via linear interpolation. Normalized weights are obtained as:

$$R_{\text{norm}}(z_i) = \frac{R(z_i)}{\sum_{j=1}^{n} R(z_j)}.$$

The final DARE-PE score is:

$$\text{DARE-PE} = \sum_{i=1}^{n} R_{\text{norm}}(z_i)\, \text{NLL}(z_i \mid q, z_{<i}),$$

where $\text{NLL}(z_i \mid q, z_{<i})$ is the negative log-likelihood of token z_i given the prompt and preceding context.

Computational cost. DARE-PE requires n cross-encoder calls (one per token) per response, plus one generation. For typical short-to-medium responses (20–80 tokens), this is substantially cheaper than perturbation-based methods requiring multiple full generations. The cross-encoder calls are parallelisable and add modest overhead in practice.

Comparison to SAR. Unlike SAR, DARE-PE does not require white-box access to token-level cross-entropies from the target LLM. The deletion-based relevance approximation introduces additional computation (cross-encoder calls) but removes the white-box dependency entirely. Empirically, DARE-PE yields stronger discrimination than uniform predictive entropy on most benchmarks in our study, and subsumes much of the predictive entropy signal in the ensemble, consistent with the hypothesis that relevance reweighting better isolates uncertainty over semantically critical tokens.

B.6 Ensemble Construction

Signal selection. We retain three signals for the ensemble: perturbation-based uncertainty (SPUQ), DARE-PE, and verbalised confidence. Semantic entropy is excluded because it performs near or below chance on most benchmarks in our study, providing no complementary signal. Standard predictive entropy is excluded in favour of DARE-PE, which empirically dominates it while capturing the same information source. The three retained signals correspond to distinct uncertainty estimation paradigms: *output consistency* under distributional shift (SPUQ), *relevance-reweighted generation confidence* (DARE-PE), and *self-reported epistemic state* (verbalised confidence). These are expected to capture complementary aspects of model behaviour and to be less correlated with one another than any pair drawn from methods sharing the same underlying information source.

Aggregation strategies. All signals are combined post-hoc from observed scores, requiring no additional model queries. We evaluate four aggregation strategies:

- *norm-avg*: equal-weight averaging of min–max normalised scores, sensitive to outliers but preserves absolute score magnitudes.
- *rank-avg*: equal-weight averaging of percentile ranks, more robust to outliers and score scale differences across methods.
- *product*: product of percentile ranks, which amplifies agreement among signals and penalises cases where any one signal assigns low uncertainty; tends to produce sharper discrimination when signals are well-calibrated.
- *max*: maximum of percentile ranks, a conservative strategy that flags a question as uncertain if *any* signal is uncertain; higher recall at the cost of precision.

No calibration data is used in constructing any ensemble; all strategies are parameter-free and applied uniformly across datasets.

C Confidence Intervals and Statistical Stability

To assess the stability of selected AUROC results, we compute 95% bootstrap confidence intervals for the five individual uncertainty quantification methods on a representative subset of model–dataset combinations for which per-instance prediction scores were retained. Ensemble methods are excluded as their uncertainty estimates are derived deterministically from the constituent individual signals and therefore do not introduce additional independent sources of variability. For each benchmark, we generate 1,000 bootstrap resamples over evaluation instances and recompute AUROC on each sample; confidence intervals are estimated using the percentile bootstrap method. Table 5 reports

Table 5: AUROC with 95% bootstrap confidence intervals for selected model–dataset combinations (individual UQ methods only). Red denotes below-chance discrimination (AUROC < 0.5). **Bold** denotes best AUROC per model per dataset.

Dataset	Model	Verb.	Sem.	Pred.	Pert.	DARE-PE
Extrinsic / factuality						
ARC-C	DS-V3	0.576 [0.53, 0.63]	0.451 [0.37, 0.53]	0.321 [0.25, 0.40]	0.726 [0.66, 0.79]	**0.877 [0.85, 0.91]**
	LLaMA	0.626 [0.59, 0.66]	0.456 [0.40, 0.51]	0.295 [0.26, 0.33]	0.632 [0.59, 0.67]	**0.800 [0.76, 0.83]**
	Qwen	0.636 [0.58, 0.69]	0.483 [0.43, 0.53]	0.396 [0.30, 0.49]	0.469 [0.37, 0.56]	**0.630 [0.54, 0.72]**
OpenBookQA	DS-V3	0.676 [0.60, 0.76]	0.424 [0.34, 0.52]	0.333 [0.25, 0.43]	0.796 [0.71, 0.88]	**0.798 [0.70, 0.88]**
	LLaMA	0.707 [0.63, 0.78]	0.519 [0.44, 0.60]	0.354 [0.28, 0.45]	0.638 [0.56, 0.72]	**0.781 [0.71, 0.85]**
	Qwen	0.676 [0.59, 0.76]	0.479 [0.36, 0.60]	0.394 [0.25, 0.55]	0.758 [0.62, 0.87]	**0.798 [0.67, 0.91]**
Intrinsic / faithfulness						
MRCEval	DS-V3	0.474 [0.45, 0.50]	0.513 [0.50, 0.52]	0.457 [0.44, 0.48]	0.546 [0.53, 0.57]	**0.595 [0.57, 0.62]**
	LLaMA	0.505 [0.48, 0.53]	0.499 [0.48, 0.52]	0.472 [0.45, 0.49]	0.538 [0.52, 0.56]	**0.564 [0.54, 0.59]**
	Qwen	0.488 [0.46, 0.51]	0.487 [0.47, 0.51]	0.459 [0.44, 0.48]	0.522 [0.49, 0.53]	**0.541 [0.52, 0.57]**
FE-UN	DS-V3	0.122 [0.09, 0.16]	0.481 [0.41, 0.56]	**0.556 [0.48, 0.63]**	0.464 [0.39, 0.54]	0.152 [0.10, 0.17]
	LLaMA	0.086 [0.04, 0.15]	0.500 [0.40, 0.61]	0.379 [0.28, 0.47]	**0.692 [0.60, 0.79]**	0.524 [0.41, 0.62]
	Qwen	0.101 [0.08, 0.13]	0.524 [0.46, 0.59]	**0.767 [0.72, 0.82]**	0.294 [0.23, 0.35]	0.145 [0.12, 0.18]

AUROC with corresponding 95% confidence intervals for selected results. These intervals provide an estimate of statistical stability for the reported AUROC values, particularly on datasets with highly imbalanced error distributions.

D Prompt Templates

This appendix documents the prompt templates used for all model interactions. Prompts are issued as single-turn user messages without a system prompt unless otherwise stated. Variable fields are denoted using `[square brackets]`.

D.1 Answer Generation

Open-ended QA (TriviaQA, SimpleQA, MRCEval, Faithfulness-QA). The model is provided with the question and, for faithfulness benchmarks (MRCEval, Faithfulness-QA), a prepended context passage. Factual benchmarks (TriviaQA, SimpleQA) omit the context block entirely.

```
Context:
[passage]
Question: [question]
```

Multiple-choice QA (OpenBookQA, ARC-Challenge, FaithEval-Counterfactual, FaithEval-Unanswerable). The template below is used for all multiple-choice benchmarks. Faithfulness-oriented benchmarks (FaithEval-Counterfactual, FaithEval-Unanswerable) prepend the context block and anchor the closing instruction to the provided evidence; factual benchmarks (OpenBookQA, ARC-Challenge) omit it along with the anchoring phrase "Based on the context above," and instead request the answer letter followed by a brief explanation. FaithEval-Unanswerable uses binary answer options (*answerable/unanswerable*) but is otherwise identical.

```
Context:
[passage]
Question: [question]
A. [option_A]
B. [option_B]
C. [option_C]
D. [option_D]
Based on the context above, answer with
only the letter of the correct option
(A, B, C, or D):
```

Answer correctness is determined by extracting the first standalone match in {A, B, C, D} via regular expression, falling back to the first response character when no standalone match is found.

D.2 Verbalised Confidence Elicitation

Following answer generation, a second prompt elicits the model's self-reported confidence. The previously generated answer is included verbatim:

```
Context:
[passage]
Question: [question]
Your Answer: [model_answer]
On a scale from 0 to 100, how confident
are you that your answer is correct given
the context? Respond with ONLY a number
between 0 and 100.
```

For factual benchmarks the context block and the phrase "given the context" are omitted; the

formatted question with options replaces the bare question field.

D.3 Perturbation-Based Uncertainty

The primary strategy used in all reported experiments is dummy-token insertion, which requires no additional generation calls. Four surface-level perturbations are applied to the base prompt by inserting semantically neutral tokens before or after the prompt text.

Inserted token	Position
`\n` (newline)	after
`\t` (tab)	before
`...` (ellipsis)	after
`um,`	before

E Additional Worked Examples of Faithfulness Hallucination Failures

This section provides additional worked examples supporting the results discussion in Section 4. Each group presents one FaithEval-Unanswerable failure (DeepSeek-V3) paired with one MRCEval failure (Qwen-3.5-Flash) drawn from the same failure mechanism. All examples show cases where the model answers incorrectly with high verbalised confidence and near-zero perturbation sensitivity, illustrating why output-level uncertainty signals cannot detect these errors. The examples are visualised in Figures 3–6; the text below provides interpretive commentary for each group.

Group 1: Confident confabulation from incomplete context. Both examples in this group involve models that fill in information absent from the context rather than abstaining. In the FaithEval-Unanswerable case, the context provides population percentages for four age groups but omits the 65+ band, making it impossible to rank all groups. The model ranks the available groups and confidently declares a fourth-place answer without flagging the missing data. In the MRCEval case (Faithful_unanswerable category), the context describes Trilochanapala as the last king of the Kabul Shahis but does not identify the modern country corresponding to his territory; the model supplies "Afghanistan" from parametric knowledge, again with 100% verbalised confidence. In both cases perturbation entropy and reweighted predictive entropy are effectively zero: the confabulated answer is stable across prompt reformulations, leaving no output-level signal to detect the error.

Group 2: World-knowledge override of context-grounded reasoning. These examples show models substituting real-world knowledge for the logical task the context defines. The FaithEval-Unanswerable case presents a passage describing *Where's Waldo?* purely in terms of its format; the question asks what type of book it is. The context does not assign a genre, but the model confidently classifies it as a "children's puzzle/activity book," an inference grounded in world knowledge rather than the passage. The MRCEval case (Reasoning_logic) is the moon-and-cheese example discussed in the main text: the context stipulates John's claim as a given fact for logical inference, but the model overrides it with real-world knowledge and selects Emily's position instead. The Group 2 pattern is particularly damaging for UQ because world-knowledge overrides are highly consistent: rephrasing the question does not change the model's answer, so perturbation entropy is near zero even though the reasoning is wrong.

Group 3: Incorrect temporal or arithmetic reasoning over context. Both examples involve plausible-looking but incorrect computations or date mappings. The FaithEval-Unanswerable case presents a passage stating that Season 4 of *The Flash* began in fall 2017 and ran until May 2018; the model answers the release question by reporting the airing window as if it were a precise premiere date, a conflation the context does not support. The MRCEval case (Reasoning_temporal) asks which position Homer Thornberry held in April–December 1947; the context lists his City Council tenure as 1946–1948, which covers that window, but the model selects his Navy Lieutenant Commander role, which ended in 1946. Both errors are consistent under reformulation: temporal mis-mappings that do not depend on surface wording produce near-zero perturbation entropy despite being factually incorrect with respect to the context.

Group 4: Commonsense or distractor override of the faithful answer. These examples show models producing wrong answers that are superficially well-grounded in the passage. The FaithEval-Unanswerable case involves a passage mentioning home schooling and informal learning in adjacent sentences without explicitly equating them; the model makes the unsupported inferential leap that home schooling is classified as informal learning. Notably, this example shows moderate perturba-

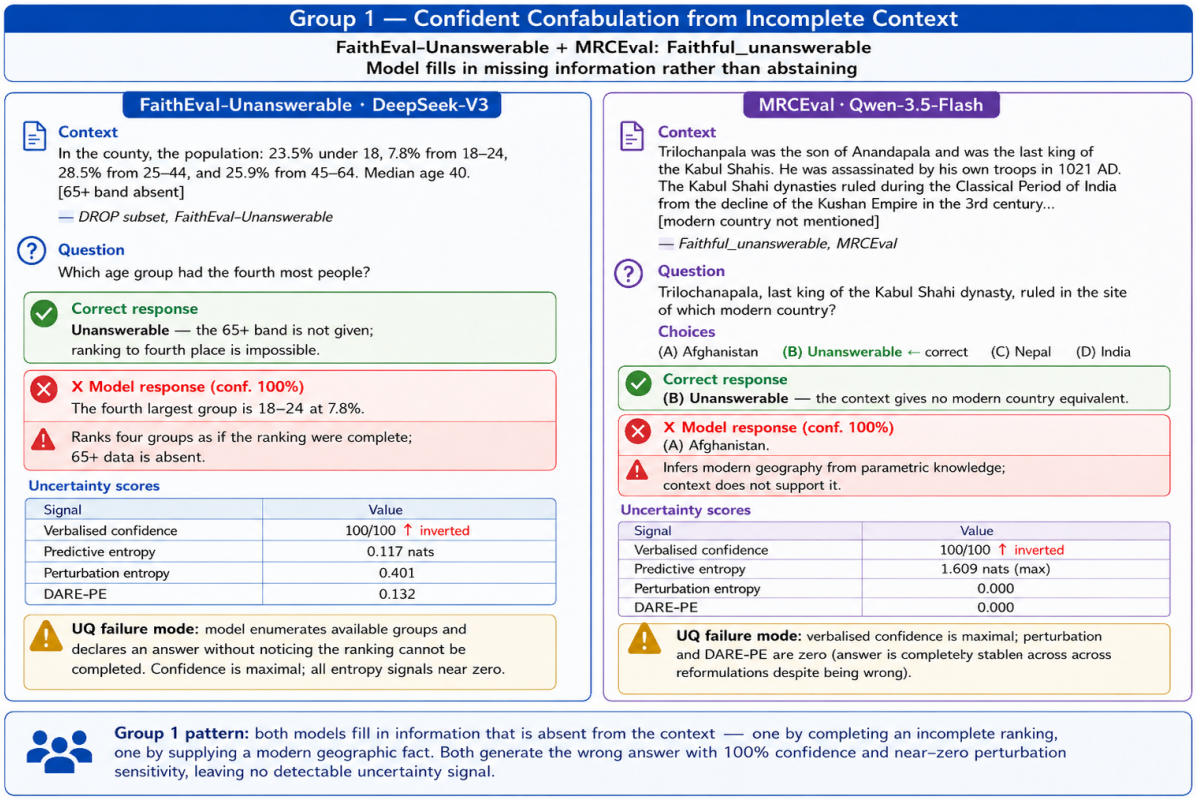

Figure 3: Group 1: Confident confabulation from incomplete context. Left: FaithEval-Unanswerable (DeepSeek-V3) — model ranks age groups without noticing the 65+ band is absent. Right: MRCEval Faithful_unanswerable (Qwen-3.5-Flash) — model supplies a modern country not mentioned in the context. Both errors are stable under reformulation; perturbation and reweighted entropy are effectively zero.

tion sensitivity (0.575) despite 100% verbalised confidence, illustrating that these two signals can diverge and that neither alone is sufficient to detect the error. The MRCEval case (ExternalKnowledge_commonsense) asks why a narrator keeps finding distractions instead of studying; the correct commonsense inference is that the narrator lacks discipline, but the model selects a surface-matching distractor drawn verbatim from the passage ("I have 101 ways to distract myself"). This lexical proximity to the context makes the wrong answer stable across reformulations, producing near-zero perturbation entropy despite requiring a deeper inferential step that the model fails to make.

Cross-group observations. Taken together, the four groups illustrate that near-chance AUROC on faithfulness benchmarks is not a single phenomenon. Across all examples, verbalised confidence is systematically inverted: models that produce wrong answers do so with high self-reported confidence, while correct abstention or correct faithful inference would have required expressed uncertainty. Perturbation entropy is near zero in most cases because the failure mode is consistent — whether driven by world-knowledge override, surface distractor selection, or temporal mis-mapping, the wrong answer does not vary under prompt reformulation. The Group 4 left panel is the partial exception: a plausible but unsupported inferential leap produces moderate perturbation sensitivity (0.575), suggesting that not all faithfulness errors are entirely invisible to perturbation-based methods, but that the signal is too weak to cross the threshold for reliable detection.

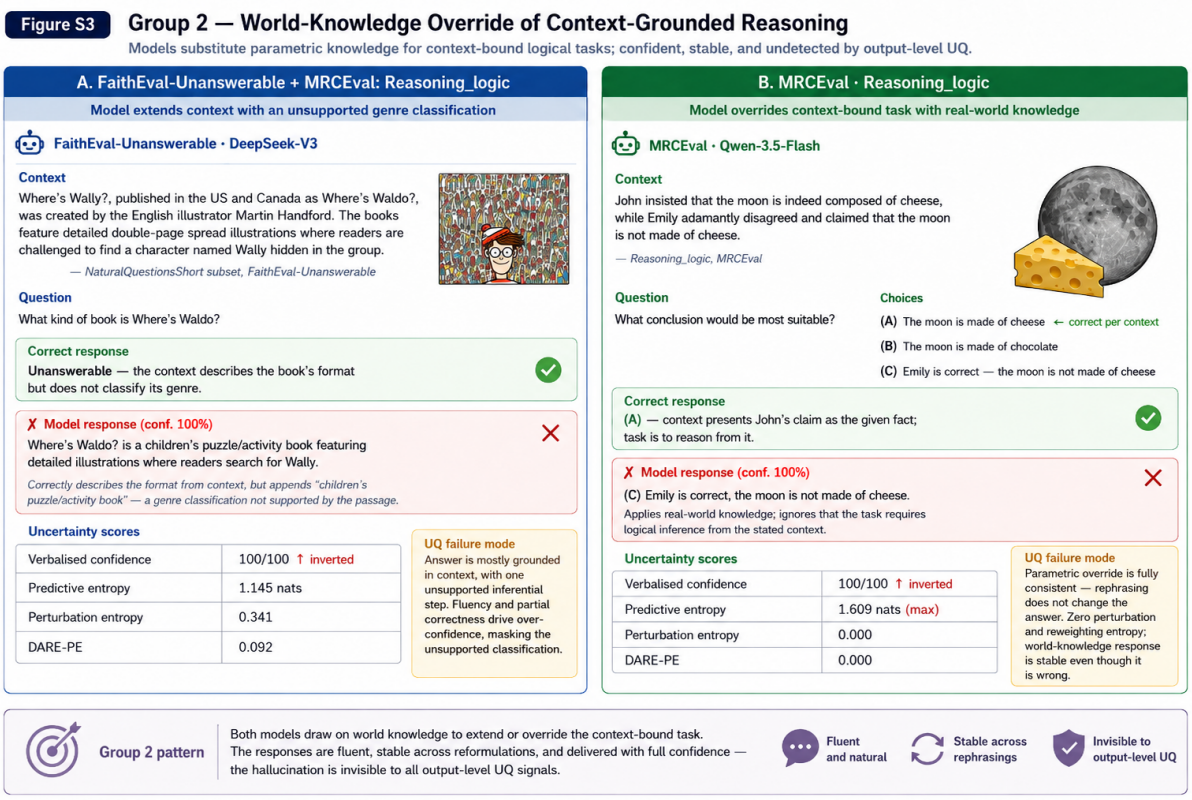

Figure 4: Group 2: World-knowledge override of context-grounded reasoning. Left: FaithEval-Unanswerable (DeepSeek-V3) — model classifies *Where's Waldo?* by genre despite the context not doing so. Right: MRCEval Reasoning_logic (Qwen-3.5-Flash) — model applies real-world knowledge to a hypothetical logical inference task. World-knowledge overrides are highly consistent across reformulations, producing near-zero perturbation entropy.

Group 3 — Incorrect Arithmetic or Temporal Reasoning Over Context

Models perform plausible computations or date mappings but are wrong with high confidence; errors are stable under reformulation.

FaithEval-Unanswerable + MRCEval: Reasoning_arithmetic

Model computes incorrectly but presents result with high confidence

Context

A recipe calls for 3/4 cup of sugar per batch. If you want to make 5 batches, how many cups of sugar do you need?

— FaithEval-Unanswerable

Question

How many cups of sugar do you need?

Correct response

Unanswerable — arithmetic is required; the context does not provide the result.

Model response (conf. 100%)

You need 4 cups of sugar.

Arithmetic error: $(3/4) \times 5 = 3.75$ cups, not 4. *The model rounds and states a specific number as if it were given.*

What the model likely did

Given (3/4 cup per batch) → Multiply ($3/4 \times 5 = 15/4$) → Miscompute ($15/4 \rightarrow 4$) → Report (4 cups)

Uncertainty scores

Metric	Value	Interpretation
Verbalised confidence	100/100 ↑ inverted	Fully confident (overconfident)
Predictive entropy	0.182 nats	Very low entropy (overconfident)
Perturbation entropy	0.214	Some sensitivity, but small
DARE-PE	0.132	Low

UQ failure mode: arithmetic slip produces a plausible integer result. Reformulations cause only limited variation, confidence remains maximal and overrides weak UQ signals.

MRCEval: Reasoning_arithmetic (Counting with Exclusion)

Model miscounts but answers confidently

Context

A box contains 12 red marbles, 8 blue marbles, and 5 green marbles. Three marbles are drawn without replacement, and all three must be different colors.

— MRCEval

Question

In how many ways can the three marbles be drawn?

Correct response

480 $(12 \times 8 \times 5 = 480)$

Model response (conf. 100%)

540

Model ignores the 'different colors' constraint and uses total combinations:

$\binom{12+8+5}{3} = \binom{25}{3} = 2{,}300 \rightarrow$ (incorrect intermediate) $\rightarrow$ **540**

Correct reasoning vs. model error

Correct (with constraint)	Model (ignores constraint)
Choose 1 red × 1 blue × 1 green	Choose any 3 from 25
$12 \times 8 \times 5 = 480$	$\binom{25}{3} = 2{,}300 \rightarrow$ (incorrectly reduced to) 540

Uncertainty scores

Metric	Value	Interpretation
Verbalised confidence	100/100 ↑ inverted	Fully confident (overconfident)
Predictive entropy	0.143 nats	Very low entropy (overconfident)
Perturbation entropy	0.097	Near-zero sensitivity
DARE-PE	0.061	Very low

UQ failure mode: the wrong computation is consistent across reformulations; the model's misapplied counting strategy is stable, yielding near-zero perturbation sensitivity.

Group 3 pattern (addition): The model performs an arithmetic or counting operation incorrectly but states a specific answer with **100% confidence**. Errors persist under prompt reformulation **(low perturbation entropy)**, so UQ signals fail to flag the mistake. These are computation/counted-mapping failures, not knowledge retrieval failures.

Figure 5: Group 3: Incorrect temporal or arithmetic reasoning over context. Left: FaithEval-Unanswerable (DeepSeek-V3) — model conflates an airing window with a precise release date. Right: MRCEval Reasoning_temporal (Qwen-3.5-Flash) — model selects a role whose dates do not overlap with the queried time window. Temporal mis-mappings are wording-independent and produce near-zero perturbation entropy despite being incorrect.

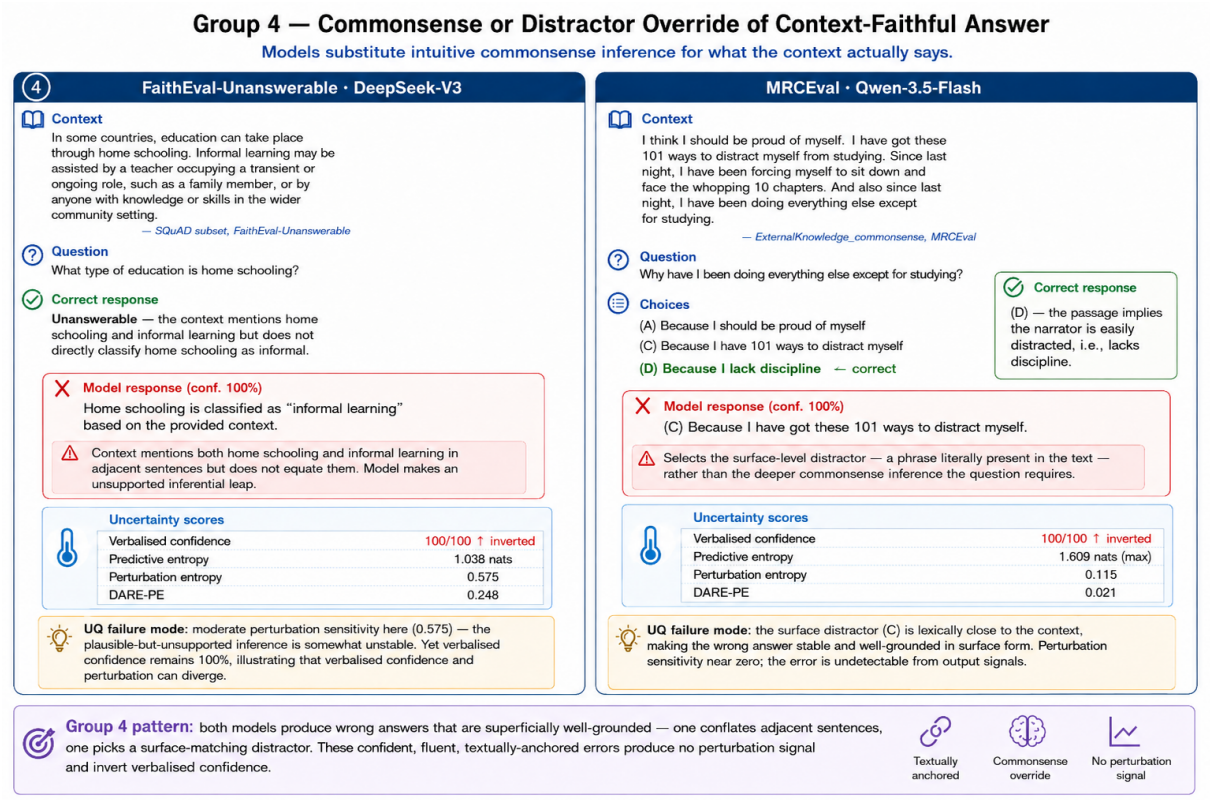

Figure 6: Group 4: Commonsense or surface-distractor override of the faithful answer. Left: FaithEval-Unanswerable (DeepSeek-V3) — model conflates adjacent sentences to infer an unsupported genre classification; moderate perturbation sensitivity (0.575) visible despite 100% verbalised confidence. Right: MRCEval ExternalKnowledge_commonsense (Qwen-3.5-Flash) — model selects a surface-matching distractor over the required commonsense inference.