Black-Box Uncertainty Quantification for Large Language Models via Ensemble-of-Ensembles

Anonymous submission

Abstract

Uncertainty quantification (UQ) is essential for building reliable and trustworthy large language models (LLMs). However, conventional Bayesian or ensemble-based UQ methods are computationally intractable at the scale of modern LLMs and often require white-box access to model parameters or logits. This paper introduces a **two-level ensemble framework** for *black-box* uncertainty estimation that operates entirely at inference time, without retraining or architectural modification. The method is theoretically grounded in the *law of total variance*, decomposing total predictive uncertainty into *aleatoric* and *epistemic* components. The inner ensemble captures stochasticity and ambiguity through repeated stochastic decoding, while the outer ensemble approximates parameter uncertainty via semantically perturbed prompts that serve as proxy samples from the implicit posterior. By measuring variance in a continuous embedding space, our framework yields interpretable and scalable uncertainty estimates across diverse LLMs. Experiments on the TriviaQA and TruthfulQA benchmarks demonstrate that our black-box estimator achieves AUROC performance comparable to or surpassing state-of-the-art white-box baselines, while offering meaningful uncertainty decomposition that distinguishes linguistic ambiguity from knowledge uncertainty.

1 Introduction & Related Work

Uncertainty quantification (UQ) has become a cornerstone of trustworthy AI, providing principled ways to interpret model confidence, handle out-of-distribution data, and ensure reliable decision-making in high-stakes applications. Bayesian neural networks (BNNs) offer a theoretically grounded framework for uncertainty estimation by maintaining a distribution over model parameters instead of point estimates (Neal 1996; Gal and Ghahramani 2016). However, while Bayesian methods have been successfully applied to moderate-scale vision models, their extension to modern large language models (LLMs) with hundreds of billions of parameters remains computationally infeasible.

Advances in LLMs have highlighted their remarkable generative and reasoning capabilities, but also their tendency to produce overconfident or inconsistent predictions. Understanding and quantifying uncertainty in these models is critical for applications such as factuality assessment, hallucination detection, and active learning. Yet, existing UQ techniques in deep learning either require expensive retraining, such as Bayesian inference or deep ensembles (Lakshminarayanan, Pritzel, and Blundell 2017) or specific neural architectures and evidential formulations (Sensoy, Kaplan, and Kandemir 2018; Osband et al. 2023), limiting their scalability to modern LLMs.

More recent work has adapted UQ to the LLM setting through both *white-box* and *black-box* approaches, where the former category exploits internal model signals such as token-level logits or hidden states to estimate semantic uncertainty or detect hallucinations (Kadavath et al. 2022; Fadeeva et al. 2023; Kuhn, Gal, and Farquhar 2023; Farquhar et al. 2024). Black-box methods treat the model as a generator and infer uncertainty from behavioral consistency or learned scoring across sampled outputs (Lin, Trivedi, and Sun 2024; Gao et al. 2024; Xiao et al. 2025; Yaldiz et al. 2025). Recent surveys (Zhou et al. 2024; Choubey et al. 2025) provide systematic overviews of emerging directions.

In this paper, we propose a scalable black-box **two-level ensemble** framework that estimates both *aleatoric* and *epistemic* uncertainty directly from model outputs. Inspired by the *law of total variance*, our method decomposes total uncertainty into within-prompt variability and cross-prompt disagreement. The inner ensemble reflects stochastic decoding noise, while the outer ensemble introduces structured semantic perturbations that emulate parameter posterior sampling. Representing generated texts in a continuous embedding space enables variance-based uncertainty decomposition, providing richer insights than entropy-based token measures and yielding interpretable and scalable uncertainty estimates for black-box LLMs. While prior work has also considered input perturbations or paraphrases for UQ (Gao et al. 2024; Li et al. 2025), here we make a formal connection to uncertainty decomposition through the outer ensemble in Bayesian language. In contrast, (Hou et al. 2023) also modeled uncertainty from a Bayesian perspective but relied on an input clarification and required training an additional classification head. Their method treats the average predictive entropy as epistemic uncertainty (EU) and the mutual information term as aleatoric uncertainty (AU), representing a different modeling paradigm from ours.

Contributions. Our main contributions are threefold:

1. We present a unified formulation of uncertainty decomposition for generative language models, bridging the Bayesian and embedding-based perspectives.

2. We propose a practical two-level ensemble estimator that approximates aleatoric and epistemic uncertainty using only model outputs.
3. Our experimental results show that the method matches and sometimes surpasses some token logit-based white-box methods, and that the decomposed uncertainties are meaningful: aleatoric uncertainty reveals ambiguity or unusual words, and epistemic uncertainty represents the model's knowledge to specific facts or terms.

2 Background

Bayesian Neural Networks

BNNs provide a principled probabilistic framework for modeling uncertainty in deep learning (Goan and Fookes 2020). Instead of learning deterministic parameters, a BNN treats each weight θ as a random variable drawn from a prior distribution $p(\theta)$. Given a dataset $D = \{(x_i, y_i)\}_{i=1}^n$, training a BNN aims to infer the posterior distribution over parameters: $p(\theta \mid D) \propto p(D \mid \theta)\, p(\theta)$, where $p(D \mid \theta)$ denotes the likelihood of the observed data under the model.

The predictive distribution for a new input x is then obtained by marginalizing over the posterior:

$$p(y \mid x, D) = \int p(y \mid x, \theta)\, p(\theta \mid D)\, d\theta = \mathbb{E}_{p(\theta|D)}[p(y \mid x, \theta)].$$

This integral can be interpreted as an ensemble of model predictions weighted by the posterior probability of each parameter configuration (Hüllermeier and Waegeman 2021). From this formulation, one can decompose the total predictive uncertainty into two components: *aleatoric* and *epistemic* (Kendall and Gal 2017; Hüllermeier and Waegeman 2021).

Approximating Parameter Uncertainty via Input Perturbation

Directly sampling from a parameter posterior $p(\theta \mid D)$ is infeasible for LLMs. Following the *local equivalence principle* (Wang and Ji 2024), we approximate small parameter variations $\Delta\theta$ by structured perturbations in the input embedding space: $f(x + \Delta x, \theta) \approx f(x, \theta + \Delta\theta)$.

Under local linearity, perturbing the input embedding x induces output changes statistically similar to parameter uncertainty. Thus, $p(y \mid x, D) \approx \mathbb{E}_{p(\Delta x)}[p(y \mid x + \Delta x, \theta)]$, where $p(\Delta x)$ defines the distribution of input perturbations. This perspective allows us to treat semantically similar variants of a prompt as samples from the implicit posterior, bridging Bayesian theory with black-box model usage.

Types and Quantification of Uncertainty

Aleatoric uncertainty (AU) captures inherent randomness or noise in the data and persists even with perfectly known parameters. **Epistemic uncertainty (EU)** reflects the model's lack of knowledge due to limited or biased training data and decreases as more data are observed (Depeweg et al. 2018; Mucsányi, Kirchhof, and Oh 2024). The uncertainty components take different forms depending on the prediction type:

Classification. For categorical predictions with $p(y \mid x, \theta)$ obtained via softmax,

$$\text{AU} = \mathbb{E}_{p(\theta|D)}[H[p(y \mid x, \theta)]], \quad \text{EU} = I[y, \theta \mid x, D],$$
$$\text{TU} = H[p(y \mid x, D)]$$

where $H[\cdot]$ denotes Shannon entropy and $I[\cdot]$ is the mutual information between model parameters and predictions (Depeweg et al. 2018; Wang 2024).

Regression. For continuous outputs parameterized by $(\mu_\theta(x), \sigma_\theta^2(x))$ (Amini et al. 2020),

$$\text{AU} = \mathbb{E}_{p(\theta|D)}[\sigma_\theta^2(x)], \quad \text{EU} = \text{Var}_{p(\theta|D)}[\mu_\theta(x)],$$
$$\text{TU} = \text{AU} + \text{EU}.$$

This decomposition links *data noise* (AU) and *model uncertainty* (EU), forming the theoretical basis for our proposed black-box uncertainty estimation method for LLMs.

3 Proposed Method: Uncertainty Quantification via Ensemble-of-Ensembles

We propose a scalable and theoretically grounded framework for estimating both *aleatoric* and *epistemic* uncertainty in LLMs for the black-box setting. The core idea is to approximate the Bayesian prediction through two nested ensembles – one over input perturbations and another over stochastic model outputs. This two-level structure captures both intrinsic randomness in decoding and variability in model behavior under small semantic perturbations, enabling principled uncertainty decomposition.

In this work, we adopt the **variance-based formulation** for uncertainty estimation. The inference process is modeled as a stochastic mapping

$$x \xrightarrow{f_\theta} y = f_\theta(x) \xrightarrow{e(\cdot)} e(y) \in \mathbb{R}^d,$$

where f_θ denotes the LLM and $e(\cdot)$ is an embedding function projecting the textual output into a continuous space. This representation bridges the discrete nature of text and the continuous formalism of uncertainty. Additional discussion on the rationale for adopting this formulation is provided in the *Supplementary Material*.

Overview and Intuition

Formally, let the LLM define a conditional generative function $f_\theta : \mathcal{X} \to \mathcal{Y}$, where θ are fixed parameters. Each generated text $y = f_\theta(x)$ can be represented as a continuous embedding $e(y) \in \mathbb{R}^d$ via an encoder $e(\cdot)$. We seek to estimate the predictive variance of the embedding representation:

$$\text{Var}(e(y)) = \mathbb{E}_{p(\Delta x)}[\text{Var}(e(y) \mid x + \Delta x)] + \text{Var}_{p(\Delta x)}(\mathbb{E}[e(y) \mid x + \Delta x]), \quad (1)$$

where Δx denotes perturbations in the input space. The first term of Eq. (1) represents *aleatoric uncertainty* – expected within-input variability – while the second term captures *epistemic uncertainty* arising from systematic differences across semantically perturbed inputs.

Two-Level Ensemble Framework

Generating Semantically Perturbed Inputs. Producing meaningful perturbations for discrete text inputs is nontrivial. We adopt a localized embedding perturbation strategy (Hu et al. 2024). Detailed algorithmic steps and hyperparameter settings are provided in *Supplementary Material*.

Outer Ensemble. We first generate N perturbed versions $\{x_1, \ldots, x_N\}$ of the original input x using above strategies. Each x_i represents a small semantic deviation (e.g., paraphrase or embedding-space perturbation) and serves as a proxy sample from the implicit parameter posterior.

Inner Ensemble. For each perturbed input x_i, we perform M independent stochastic forward passes using the same model f_θ and fixed decoding hyperparameters (e.g., temperature, top-p sampling). This inner loop captures variability due to sampling noise and inherent ambiguity in the model's output distribution $p(y \mid x_i, \theta)$.

Estimation Procedure. Let $e(y_{i,j})$ denote the embedding of the j-th output generated for input x_i. We compute the sample mean and covariance across the M stochastic samples for each perturbed input, and aggregate these statistics across N outer perturbations to estimate the total, aleatoric, and epistemic uncertainty components:

$$\mu_i = \tfrac{1}{M}\sum_{j=1}^{M} e(y_{i,j}), \quad \Sigma_i = \tfrac{1}{M-1}\sum_{j=1}^{M}(e(y_{i,j}) - \mu_i)(e(y_{i,j}) - \mu_i)^\top,$$

$$\widehat{\mathcal{U}}_A = \tfrac{1}{N}\sum_{i=1}^{N}\Sigma_i, \qquad \widehat{\mathcal{U}}_E = \tfrac{1}{N-1}\sum_{i=1}^{N}(\mu_i - \bar{\mu})(\mu_i - \bar{\mu})^\top,$$

$$\bar{\mu} = \tfrac{1}{N}\sum_{i=1}^{N}\mu_i, \qquad \widehat{\mathcal{U}}_T = \widehat{\mathcal{U}}_A + \widehat{\mathcal{U}}_E. \tag{2}$$

Equation (2) compactly summarizes the complete uncertainty estimation process. Here, $\widehat{\mathcal{U}}_A$ measures the expected within-input covariance (aleatoric uncertainty), $\widehat{\mathcal{U}}_E$ captures the variance of mean embeddings across perturbed inputs (epistemic uncertainty), and $\widehat{\mathcal{U}}_T$ is their sum, corresponding to total predictive uncertainty under the law of total variance. This formulation maintains both mathematical rigor and computational efficiency, making it suitable for large-scale LLM inference.

Algorithm 1 summarizes the entire pipeline.

Algorithm 1: Two-Level Ensemble UQ for LLMs

Require: LLM f_θ, embedding $e(\cdot)$, input x, outer size N, inner size M
Ensure: Aleatoric $\mathcal{U}_A$, Epistemic $\mathcal{U}_E$

```
Generate paraphrase pool {x'_k}_{k=1}^K, store embeddings {e(x'_k)}
for i = 1 to N do                                   ▷ Outer ensemble
    x_i ← arg min_{x'_k} D(e(x) + N(0, ε²I), e(x'_k))
    for j = 1 to M do                               ▷ Inner ensemble
        E_i ← E_i ∪ {e(f_θ(x_i))}                   ▷ stochastic decoding
    end for
    μ_i ← mean(E_i),  Σ_i ← cov(E_i)
end for
U_A ← (1/N) Σ_i Σ_i,  μ̄ ← (1/N) Σ_i μ_i
U_E ← (1/(N−1)) Σ_i (μ_i − μ̄)(μ_i − μ̄)^⊤  return U_A, U_E
```

4 Experiments

Experimental Setup

We evaluate our proposed approach on question answering (QA) tasks using the **TriviaQA** (Joshi et al. 2017) and **TruthfulQA** (Lin, Hilton, and Evans 2021) datasets. Our primary results are reported on **TriviaQA**, while additional experiments on **TruthfulQA** using smaller models are presented in *Supplementary Material*. The objective is to examine whether model uncertainty can reliably predict answer correctness, quantified by **Area Under the ROC Curve (AUROC)**, a standard metric for uncertainty estimation. Detailed experimental settings can be found in *Supplementary Material Sec 5*.

Models. We conduct experiments on four representative large language models: **LLaMA3-70B**, **LLaMA3-8B**, **Qwen2.5-72B**, and **Qwen2.5-7B** (Dubey et al. 2024; Bai et al. 2025). All models are evaluated under identical conditions using 20 inner samples ($M = 20$) and 20 outer perturbations ($N = 20$).

Evaluation Metric. We measure answer correctness using two complementary criteria: (1) **Cosine Similarity** between the generated and reference answer embeddings (threshold > 0.8) (Reimers and Gurevych 2019), and (2) **ROUGE-L** overlap score (threshold > 0.5) (Lin 2004). Uncertainty is decomposed into total, aleatoric, and epistemic components using Eq. (2). We report AUROC scores for each uncertainty component, reflecting how effectively uncertainty distinguishes correct from incorrect generations.

Baselines. We compare our approach against six representative uncertainty estimation methods: **G-NLL** (Aichberger, Schweighofer, and Hochreiter 2024), **Predictive Entropy (PE)** and **Length-Normalized Predictive Entropy (LN-PE)** (Malinin and Gales 2020), **Semantic Entropy (SE)**, **Length-Normalized Semantic Entropy (LN-SE)**, and **Discrete Semantic Entropy (D-SE)** (Farquhar et al. 2024; Kuhn, Gal, and Farquhar 2023). Among them, G-NLL, PE, LN-PE, SE, and LN-SE are *white-box*, requiring access to token-level probabilities; D-SE and our method operate in a fully *black-box* setting using only generated text.

Results

Table 1 reports AUROC scores measuring the correlation between predicted uncertainty and answer correctness on the TriviaQA (no-context) dataset. White-box baselines (G-NLL, PE/LN-PE, SE/LN-SE) require token-level logits, whereas D-SE method and our proposed approach are strictly black-box, relying only on generated text.

Overall Comparison. Across all four models, our *Ensemble-of-Ensembles* is better than the black-box baseline and matches or sometimes even surpasses white-box baselines while using only black-box access. For example, on **LLaMA3-70B**, total uncertainty achieves **TU**=0.7967 (cosine), exceeding G-NLL (0.7474) and PE (0.7649); on

Model	Correctness	Ours			Black-box	White-box				
		AU	EU	TU	D-SE	G-NLL	PE	LN-PE	SE	LN-SE
LLaMA3-70B	Cosine (> 0.8)	0.712	0.791	**0.796**	0.622	0.747	0.764	0.781	0.561	0.683
	RougeL (> 0.5)	0.727	0.725	0.774	0.639	**0.788**	0.767	0.761	0.545	0.659
Qwen2.5-72B	Cosine (> 0.8)	**0.760**	0.621	0.749	0.632	0.741	0.672	0.727	0.672	0.673
	RougeL (> 0.5)	**0.798**	0.594	0.766	0.673	0.782	0.710	0.775	0.709	0.710
LLaMA3-8B	Cosine (> 0.8)	0.782	0.556	0.756	0.740	0.817	0.758	**0.829**	0.783	0.786
	RougeL (> 0.5)	0.811	0.531	0.770	0.771	**0.816**	0.737	0.819	0.775	0.775
Qwen2.5-7B	Cosine (> 0.8)	0.781	0.552	0.754	0.695	**0.828**	0.764	0.813	0.756	0.756
	RougeL (> 0.5)	0.792	0.535	0.760	0.738	**0.845**	0.785	0.838	0.784	0.782

Table 1: Comparison of uncertainty-correctness correlation (AUROC) across models and methods on TriviaQA (no-context). Numbers in **bold** represent the overall best performance, underlined numbers indicate the top-performing white-box method, and boxed numbers indicate the top-performing black-box method.

Qwen2.5-72B, **AU**=0.7600 is the strongest single indicator of correctness. These trends underscore the practicality of our method for proprietary/API LLMs without logits.

Large Models. For high-capacity LLMs, the proposed method produces robust and consistent AUROC scores. In **LLaMA3-70B**, epistemic uncertainty (**EU**=0.7914) correlates closely with correctness, reflecting the model's confidence in factual output. On **Qwen2.5-72B**, however, aleatoric uncertainty dominates (**AU**=0.7600), suggesting that variability from perturbed inputs is the primary contributor to overall uncertainty.

This behavior can be partly attributed to differences in training data scale and diversity – Qwen's corpus, though extensive, is smaller and less heterogeneous than LLaMA's, which limits the model's ability to capture comprehensive world knowledge. Consequently, its epistemic uncertainty becomes less informative, reflecting insufficient knowledge coverage rather than genuine model doubt.

Small Models. For lower-capacity models, epistemic uncertainty (**EU**) becomes less reliable ($\approx$ 0.53–0.56 AUROC), often showing weak alignment with correctness. This is expected, as smaller models exhibit high sensitivity to paraphrasing, which can lead to substantially varied outputs. Such instability arises from their limited capacity and insufficient knowledge coverage, which hinders understanding of complex or rare vocabulary (Cox et al. 2025). As a result, their epistemic responses are dominated by superficial variations rather than genuine model doubt (Ahdritz et al. 2024). Conversely, aleatoric uncertainty (**AU**) remains high ($\approx$ 0.78–0.81 AUROC), reflecting the inherent stochasticity of generation and the models' vulnerability to noise and ambiguity. Overall, these observations suggest that small models' uncertainties are shaped more by input sensitivity than well-calibrated knowledge estimation.

These findings demonstrate that our perturbation-based approximation effectively captures both intrinsic (aleatoric) and model-level (epistemic) components without requiring access to internal token distributions.

Key Insights. Three consistent trends emerge: (1) **AU** is the most consistent predictor of correctness across all models; (2) **EU** becomes meaningfully predictive only when factual coverage is sufficient (e.g., LLaMA3-70B), but loses reliability for smaller or less broadly trained models; (3) **TU** serves as a strong overall indicator when both EU and AU are informative (e.g., LLaMA3-70B on TriviaQA and results on TruthfulQA). However, when EU becomes unreliable for various reasons, AU tends to outperform TU.

Overall, our method achieves comparable or superior reliability to white-box approaches, offering a scalable, interpretable, and generalizable framework for UQ in LLMs. Further experimentation on different models and tasks is required to better understand when the proposed method is most suitable.

5 Conclusion

We have presented a scalable, inference-time framework for black-box uncertainty quantification in LLMs that requires no retraining or access to logits, using a *two-level ensemble* that decomposes *aleatoric* and *epistemic* components via nested sampling over stochastic decoding and semantic perturbations, measured in a continuous embedding space. Our estimator matches or outperforms white-box baselines such as G-NLL and PE on QA tasks while providing interpretable decomposition. Empirically, AU is consistently a reliable correlate of correctness across scales which captures the ambiguity in input; EU is informative for models with stronger factual grounding (e.g., LLaMA3-70B) but weakens for smaller models or those trained on less diverse corpora. These findings support black-box UQ as a practical tool for hallucination detection, risk-sensitive generation, and data acquisition. Such work can inform uncertainty mitigation strategies to improve model performance, with uncertainty-guided fine-tuning (e.g. high EU cases highlight the model's lacks of knowledge in specific topic, we can fine-tune the model on related dataset), uncertainty-aware RAG or uncertainty-driven output aggregation strategies.

References

Ahdritz, G.; Qin, T.; Vyas, N.; Barak, B.; and Edelman, B. L. 2024. Distinguishing the knowable from the unknowable with language models. *arXiv preprint arXiv:2402.03563*.

Aichberger, L.; Schweighofer, K.; and Hochreiter, S. 2024. Rethinking uncertainty estimation in natural language generation. *arXiv preprint arXiv:2412.15176*.

Amini, A.; Schwarting, W.; Soleimany, A.; and Rus, D. 2020. Deep evidential regression. *Advances in Neural Information Processing Systems*, 33: 14927–14937.

Bai, S.; Chen, K.; Liu, X.; Wang, J.; Ge, W.; Song, S.; Dang, K.; Wang, P.; Wang, S.; Tang, J.; et al. 2025. Qwen2. 5-vl technical report. *arXiv preprint arXiv:2502.13923*.

Choubey, P.; Liu, X.; Chen, T.; Da, L.; Chen, C.; Lin, Z.; and Wei, H. 2025. Confidence Calibration and Uncertainty Estimation in Large Language Models: A Survey. *ACM Computing Surveys*.

Cox, K.; Xu, J.; Han, Y.; Xu, R.; Li, T.; Hsu, C.-Y.; Chen, T.; Gerych, W.; and Ding, Y. 2025. Mapping from Meaning: Addressing the Miscalibration of Prompt-Sensitive Language Models. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 23696–23703.

Depeweg, S.; Hernandez-Lobato, J.-M.; Doshi-Velez, F.; and Udluft, S. 2018. Decomposition of Uncertainty in Bayesian Deep Learning for Efficient and Risk-sensitive Learning. In Dy, J.; and Krause, A., eds., *Proceedings of the International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, 1184–1193. PMLR.

Dubey, A.; Jauhri, A.; Pandey, A.; Kadian, A.; Al-Dahle, A.; Letman, A.; Mathur, A.; Schelten, A.; Yang, A.; Fan, A.; et al. 2024. The llama 3 herd of models. *arXiv e-prints*, arXiv–2407.

Fadeeva, E.; Vashurin, R.; Tsvigun, A.; Vazhentsev, A.; Petrakov, S.; Fedyanin, K.; Vasilev, D.; Goncharova, E.; Panchenko, A.; Panov, M.; et al. 2023. LM-Polygraph: Uncertainty Estimation for Language Models. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 446–461.

Farquhar, S.; Kossen, J.; Kuhn, L.; and Gal, Y. 2024. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017): 625–630.

Gal, Y.; and Ghahramani, Z. 2016. Dropout as a Bayesian approximation: Representing model uncertainty in deep learning. In *Proceedings of the International Conference on Machine Learning*, 1050–1059. PMLR.

Gao, X.; Zhang, J.; Mouatadid, L.; and Das, K. 2024. SPUQ: Perturbation-Based Uncertainty Quantification for Large Language Models. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics (EACL)*, 2145–2160.

Goan, E.; and Fookes, C. 2020. Bayesian neural networks: An introduction and survey. In *Case Studies in Applied Bayesian Data Science: CIRM Jean-Morlet Chair, Fall 2018*, 45–87. Springer.

Hou, B.; Liu, Y.; Qian, K.; Andreas, J.; Chang, S.; and Zhang, Y. 2023. Decomposing uncertainty for large language models through input clarification ensembling. *arXiv preprint arXiv:2311.08718*.

Hu, W.; Shu, Y.; Yu, Z.; Wu, Z.; Lin, X.; Dai, Z.; Ng, S.-K.; and Low, B. K. H. 2024. Localized zeroth-order prompt optimization. *Advances in Neural Information Processing Systems*, 37: 86309–86345.

Hüllermeier, E.; and Waegeman, W. 2021. Aleatoric and epistemic uncertainty in machine learning: An introduction to concepts and methods. *Machine Learning*, 110(3): 457–506.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. *arXiv preprint arXiv:1705.03551*.

Kadavath, S.; Conerly, T.; Askell, A.; Henighan, T.; Drain, D.; Perez, E.; Schiefer, N.; Hatfield-Dodds, Z.; DasSarma, N.; Tran-Johnson, E.; et al. 2022. Language Models (Mostly) Know What They Know. *arXiv preprint arXiv:2207.05221*.

Kendall, A.; and Gal, Y. 2017. What uncertainties do we need in Bayesian deep learning for computer vision? *Advances in Neural Information Processing Systems*, 30.

Kuhn, L.; Gal, Y.; and Farquhar, S. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Lakshminarayanan, B.; Pritzel, A.; and Blundell, C. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in Neural Information Processing Systems*, 30.

Li, M.; Li, X.; Zhang, W.; and Ma, L. 2025. ESI: Epistemic Uncertainty Quantification via Semantic-preserving Intervention for Large Language Models. *arXiv preprint arXiv:2510.13103*.

Lin, C.-Y. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out (ACL 2004 Workshop)*.

Lin, S.; Hilton, J.; and Evans, O. 2021. Truthfulqa: Measuring how models mimic human falsehoods. *arXiv preprint arXiv:2109.07958*.

Lin, Z.; Trivedi, S.; and Sun, J. 2024. Generating with Confidence: Uncertainty Quantification for Black-Box Large Language Models. *Transactions on Machine Learning Research*.

Malinin, A.; and Gales, M. 2020. Uncertainty estimation in autoregressive structured prediction. *arXiv preprint arXiv:2002.07650*.

Mucsányi, B.; Kirchhof, M.; and Oh, S. J. 2024. Benchmarking Uncertainty Disentanglement: Specialized Uncertainties for Specialized Tasks. In Globerson, A.; Mackey, L.; Belgrave, D.; Fan, A.; Paquet, U.; Tomczak, J.; and Zhang, C., eds., *Advances in Neural Information Processing Systems*, volume 37, 50972–51038. Curran Associates, Inc.

Neal, R. M. 1996. *Bayesian Learning for Neural Networks*, volume 118 of *Lecture Notes in Statistics*. New York: Springer-Verlag.

Osband, I.; Wen, Z.; Asghari, S. M.; Dwaracherla, V.; Ibrahimi, M.; Lu, X.; and Van Roy, B. 2023. Epistemic neural networks. *Advances in Neural Information Processing Systems*, 36: 2795–2823.

Reimers, N.; and Gurevych, I. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.

Sensoy, M.; Kaplan, L.; and Kandemir, M. 2018. Evidential deep learning to quantify classification uncertainty. *Advances in Neural Information Processing Systems*, 31.

Wang, H. 2024. *Towards Explainable and Actionable Bayesian Deep Learning*. Ph.D. thesis, Rensselaer Polytechnic Institute.

Wang, H.; and Ji, Q. 2024. Epistemic uncertainty quantification for pre-trained neural networks. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, 11052–11061.

Xiao, Q.; Bhattacharjya, D.; Ganesan, B.; Marinescu, R.; Mirylenka, K.; Pham, N. H.; Glass, M.; and Lee, J. 2025. The Consistency Hypothesis in Uncertainty Quantification for Large Language Models. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence*.

Yaldiz, D. N.; Bakman, Y. F.; Buyukates, B.; Tao, C.; Ramakrishna, A.; Dimitriadis, D.; Zhao, J.; and Avestimehr, S. 2025. Do Not Design, Learn: A Trainable Scoring Function for Uncertainty Estimation in Generative LLMs. In *Findings of the Association for Computational Linguistics: NAACL*.

Zhou, K.; Mei, Z.; Lidard, J.; Ren, A. Z.; and Majumdar, A. 2024. Uncertainty Quantification for Large Language Models: A Survey. *arXiv preprint arXiv:2404.14224*.

Reproducibility Checklist

Instructions for Authors:

This document outlines key aspects for assessing reproducibility. Please provide your input by editing this `.tex` file directly.

For each question (that applies), replace the "Type your response here" text with your answer.

Example: If a question appears as

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
Type your response here
```

you would change it to:

```
\question{Proofs of all novel claims
are included} {(yes/partial/no)}
yes
```

Please make sure to:

- Replace ONLY the "Type your response here" text and nothing else.
- Use one of the options listed for that question (e.g., **yes**, **no**, **partial**, or **NA**).
- **Not** modify any other part of the `\question` command or any other lines in this document.

You can `\input` this .tex file right before `\end{document}` of your main file or compile it as a stand-alone document. Check the instructions on your conference's website to see if you will be asked to provide this checklist with your paper or separately.

---

1. General Paper Structure

1.1. Includes a conceptual outline and/or pseudocode description of AI methods introduced (yes/partial/no/NA) yes

1.2. Clearly delineates statements that are opinions, hypothesis, and speculation from objective facts and results (yes/no) yes

1.3. Provides well-marked pedagogical references for less-familiar readers to gain background necessary to replicate the paper (yes/no) yes

2. Theoretical Contributions

2.1. Does this paper make theoretical contributions? (yes/no) no

If yes, please address the following points:

2.2. All assumptions and restrictions are stated clearly and formally (yes/partial/no) Type your response here

2.3. All novel claims are stated formally (e.g., in theorem statements) (yes/partial/no) Type your response here

2.4. Proofs of all novel claims are included (yes/partial/no) Type your response here

2.5. Proof sketches or intuitions are given for complex and/or novel results (yes/partial/no) Type your response here

2.6. Appropriate citations to theoretical tools used are given (yes/partial/no) Type your response here

2.7. All theoretical claims are demonstrated empirically to hold (yes/partial/no/NA) Type your response here

2.8. All experimental code used to eliminate or disprove claims is included (yes/no/NA) Type your response here

3. Dataset Usage

3.1. Does this paper rely on one or more datasets? (yes/no) yes

If yes, please address the following points:

3.2. A motivation is given for why the experiments are conducted on the selected datasets (yes/partial/no/NA) yes

3.3. All novel datasets introduced in this paper are included in a data appendix (yes/partial/no/NA) NA

3.4. All novel datasets introduced in this paper will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no/NA) NA

3.5. All datasets drawn from the existing literature (potentially including authors' own previously published work) are accompanied by appropriate citations (yes/no/NA) yes

3.6. All datasets drawn from the existing literature (potentially including authors' own previously published work) are publicly available (yes/partial/no/NA) yes

3.7. All datasets that are not publicly available are described in detail, with explanation why publicly available alternatives are not scientifically satisficing (yes/partial/no/NA) NA

4. Computational Experiments

4.1. Does this paper include computational experiments? (yes/no) yes

If yes, please address the following points:

4.2. This paper states the number and range of values tried per (hyper-) parameter during development of the paper, along with the criterion used for selecting the final parameter setting (yes/partial/no/NA) Type your response here

4.3. Any code required for pre-processing data is included in the appendix (yes/partial/no) yes

4.4. All source code required for conducting and analyzing the experiments is included in a code appendix (yes/partial/no) yes

4.5. All source code required for conducting and analyzing the experiments will be made publicly available upon publication of the paper with a license that allows free usage for research purposes (yes/partial/no) yes

4.6. All source code implementing new methods have comments detailing the implementation, with references to the paper where each step comes from (yes/partial/no) yes

4.7. If an algorithm depends on randomness, then the method used for setting seeds is described in a way sufficient to allow replication of results (yes/partial/no/NA) yes

4.8. This paper specifies the computing infrastructure used for running experiments (hardware and software), including GPU/CPU models; amount of memory; operating system; names and versions of relevant software libraries and frameworks (yes/partial/no) yes

4.9. This paper formally describes evaluation metrics used and explains the motivation for choosing these metrics (yes/partial/no) yes

4.10. This paper states the number of algorithm runs used to compute each reported result (yes/no) yes

4.11. Analysis of experiments goes beyond single-dimensional summaries of performance (e.g., average; median) to include measures of variation, confidence, or other distributional information (yes/no) yes

4.12. The significance of any improvement or decrease in performance is judged using appropriate statistical tests (e.g., Wilcoxon signed-rank) (yes/partial/no) yes

4.13. This paper lists all final (hyper-)parameters used for each model/algorithm in the paper's experiments (yes/partial/no/NA) yes