Quantifying Aleatoric Uncertainty of In-Context Learning for Robust Measure of LLM Prediction Confidence

Jinseok Chung **Minkyoung Song*** **Hyunji Jung*** **Namhoon Lee**
POSTECH
{jinseok.chung, minkyoung.song, hyunji.jung, namhoon.lee}@postech.ac.kr

Abstract

In-Context Learning (ICL) allows LLMs to adapt to new tasks from a few demonstrations, but its reliability remains a concern: predictions are highly sensitive to both prompt design and the model's ability to understand the context, obscuring whether failures arise from data properties or model limitations. Uncertainty decomposition—separating aleatoric from epistemic sources—is particularly crucial in this setting, yet existing methods, designed for standard generation tasks, fail to capture the unique dynamics of ICL. To address this, we introduce a concept of *self-function vectors*, built upon Bayesian views and the mechanistic interpretability of ICL. These vectors leverage internal model representations to model the latent concept learned during in-context prompting, thereby enabling a direct estimation of aleatoric uncertainty within a Bayesian framework and circumventing the reliance on brittle input or decoding manipulations. Given the lack of established benchmarks and suitable evaluation protocols, we also propose the first and rigorous evaluation protocol, in which data is manipulated in controlled ways so as to quantify aleatoric uncertainty precisely and separately from epistemic uncertainty. With this new evaluation framework, initially grounded in synthetic tasks for conceptual development and subsequently extended to real-world datasets, we show that our proposed methodology can measure uncertainty of LLM predictions made under ICL more reliably than existing alternative methods. Moreover, we show it can be used as a practical tool for trustworthy-related applications, such as hallucination detection. Our findings pave a new direction for connecting the quantitative view of uncertainty with the mechanistic understanding of model behavior.

1 Introduction

In-Context Learning (ICL) enables Large Language Models (LLMs) to adapt to new tasks without parameter updates by conditioning on a handful of input–output demonstrations provided directly in the prompt (Brown et al., 2020). Despite its versatility, ICL is inherently brittle: predictions can fluctuate with small changes to prompt format, the ordering of demonstrations, or data properties (Min et al., 2022; Chan et al., 2022, 2025). On top of that, ICL predictions depend highly on the model's context learnability (Wang et al., 2024), making it difficult to discern whether failures stem from data properties or from model limitations. This highlights the importance of not only measuring the overall level of uncertainty—known as Uncertainty Quantification (Kuhn et al., 2023; Shorinwa et al., 2024)—but also disentangling its underlying sources, a process referred to as Uncertainty Decomposition (Kendall and Gal, 2017; Abbasi-Yadkori et al., 2024; Wang and Holmes, 2025; Jayasekera et al., 2025).

Uncertainty decomposition aims to disentangle uncertainty into its principal components: aleatoric uncertainty (AU), which stems from inherent data ambiguity, and epistemic uncertainty (EU), which reflects limitations in the model's knowledge or representations. This decomposition offers sharper insight into failure modes: AU reflects irreducible noise, while EU highlights uncertainty that could, in principle, be reduced with better modeling or more evidence. This distinction is especially important in ICL, where predictions hinge on both the given context and the model's learnability.

Yet, current research is primarily focused on standard settings such as Question and Answer, and not much has been done for ICL (Shorinwa et al., 2024). Prior approaches typically involve manipulating inputs (Abbasi-Yadkori et al., 2024; Ling et al., 2024; Wang and Holmes, 2025) or altering decoding strategies (Kuhn et al., 2023; Ling et al., 2024), but such methods do not readily transfer to ICL, as seemingly small variations in inputs or decoding may lead to substantial and unpredictable fluctuations in performance (Min et al., 2022).

*Equal contribution.

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 2090–2108

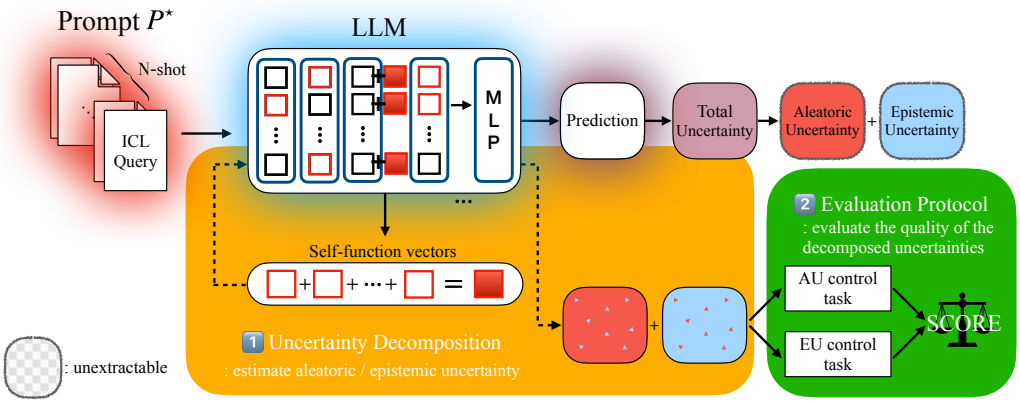


Figure 1: This work comprises two main pillars. (1) Uncertainty Decomposition: We probe the LLM's internal representations to construct self-function vectors from the activations of salient attention heads. These vectors serve as a proxy for the latent concept learned from the prompt, enabling a principled estimation of AU and EU. (2) Evaluation Protocol: We then assess the quality of the decomposition using a novel framework with tasks specifically designed to independently perturb AU and EU, thereby providing a robust score for decomposition fidelity.

Encouragingly, two recent lines of research suggest a path forward. First, interpreting ICL as implicit Bayesian inference provides a natural theoretical foundation for principled decomposition (Xie et al., 2022; Wang et al., 2023; Jiang, 2024). Second, advances in the mechanistic interpretability—examining model internals such as representation dynamics and head activations—offer tools to observe ICL behavior without requiring explicit manipulations of inputs or decoding (Hendel et al., 2023; Todd et al., 2024; Heo et al., 2025). Taken together, these perspectives provide a practical tool for principled and ICL-specific uncertainty decomposition.

At the same time, evaluating decomposition in ICL remains an open challenge: existing studies typically rely on downstream tasks from uncertainty quantification—such as hallucination detection or out-of-distribution identification —since no benchmark directly assesses decomposition quality, leaving it unclear whether AU and EU are genuinely separated or whether apparent gains merely reflect proxy-task performance.

Our contributions are summarized as follows:

1. We introduce an aleatoric uncertainty quantification methodology based on *self-function vectors*—a novel approach that leverages internal model representations for principled uncertainty decomposition in ICL, building upon Bayesian perspectives and mechanistic interpretability.

2. We develop an evaluation protocol tailored to uncertainty decomposition in the ICL setting, grounded in synthetic tasks that can reliably capture ICL dynamics and enable faithful assessment of decomposition quality.

3. We further establish the applicability of our method by showing competitive performance on hallucination detection. This opens a new potential for bridging uncertainty estimation with mechanistic interpretability.

2 Background

LLMs have demonstrated the ability to "learn" new tasks without updating their weights by conditioning on a small number of labeled examples provided on the prompt. This behavior, known as In-Context Learning(ICL), allows the model to condition its predictions on a few labeled examples provided at inference time—without any model updates. Formally, let $\mathcal{D}_{\text{ex}} = \{(x_i, y_i)\}_{i=1}^{k} \subseteq \mathcal{D}_T$ denote the example set sampled from the dataset $\mathcal{D}_T$ of task T, and let $x^\star$ be the query. Conditioned on $\mathcal{D}_{\text{ex}}$, a model defines the predictive distribution:

$$p\big(y \mid x^\star, \mathcal{D}_{\text{ex}}\big). \tag{1}$$

All task adaptation occurs within the context window, with no updates to model weights. This surprising behavior has inspired research to understand its underlying principles. The most prominent perspective is Bayesian inference, which views ICL as posterior updating on the observed data.

Bayesian View of ICL A widely held view is that ICL corresponds to *implicit Bayesian inference* (Xie et al., 2022; Wang et al., 2023; Jiang, 2024). In this perspective, the model infers a latent concept ϕ from the prompt and conditions its predictions on that inferred belief. The resulting predictive distribution approximates

$$p(y \mid x^\star, \mathcal{D}_{\text{ex}}) \approx \int p(y \mid x^\star, \phi) \underbrace{p(\phi \mid \mathcal{D}_{\text{ex}})}_{\text{implicit posterior}} d\phi. \quad (2)$$

Accordingly, the total uncertainty can be decomposed information-theoretically

$$\mathcal{H}[p(y \mid x^\star, \mathcal{D}_{\text{ex}})] = \underbrace{\mathcal{I}(y; \phi \mid x^\star, \mathcal{D}_{\text{ex}})}_{\text{Epistemic}} + \underbrace{\mathbb{E}_{p(\phi \mid \mathcal{D}_{\text{ex}})}[\mathcal{H}(p(y \mid x^\star, \phi))]}_{\text{Aleatoric}}, \quad (3)$$

where $\mathcal{H}$ and $\mathcal{I}$ denote the Shannon entropy and mutual information, respectively. (Wimmer et al., 2023; Falck et al., 2024; Jayasekera et al., 2025).

Mechanistic View of ICL Mechanistic interpretability is a line of research that seeks to understand how models operate by identifying the specific neurons or attention heads that contribute to their behavior and how they work together to produce the model's output. Applied to ICL, this perspective aims to uncover how and which internal representations of the model process understanding context. Early studies highlight *induction heads*—attention heads that copy earlier tokens to later positions—as central to pattern completion (Elhage et al., 2021; Olsson et al., 2022). More recent work identifies a complementary mechanism: *function vectors*, compact task representations encoded in the activations of specific attention heads during ICL (Todd et al., 2024). An function vector v_T is constructed by summing the prompt-wise averaged activations across all attention heads,

$$\bar{h}^T_{(\ell,k)} = \frac{1}{n} \sum_{i=1}^{n} h^{P_i}_{(\ell,k)}, \quad (4)$$

$$v_T = \sum_{(\ell,k) \in \mathcal{S}_T} \bar{h}^T_{(\ell,k)}, \quad (5)$$

where $\mathcal{S}_T$ denotes the set of causal heads (i.e., the attention heads most critical for task T), while $\bar{h}^T_{(\ell,k)}$ is defined as the average activation of the k-th head in layer ℓ computed over the input prompts P_i.

A function vector v_T can be utilized by injecting it into the model's hidden states—that is, by adding the vector to the hidden representations during inference. Injecting v_T shifts the context of the model's predictions—even in zero-shot settings—indicating that function vectors encode task identity and serve as internal task representations. Building on this view, Jiang et al. (2025) show that function vectors serve as latent task embeddings that parameterize task-specific functions, i.e., conditioning on $x^\star$ and v_T induces behavior consistent with $p(y \mid x^\star, \mathcal{D}_{\text{ex}})$. Complementary ablations by Yin and Steinhardt (2025) further reveal that function vector heads, rather than induction heads, are the dominant contributors to ICL performance in larger models. Collectively, these findings establish function vectors as the primary carriers of task structure.

Uncertainty Decomposition before ICL The decomposition of predictive uncertainty into epistemic and aleatoric components has been explored widely before ICL. In computer vision, uncertainty decomposition underpins calibration, segmentation, and out-of-distribution detection, with EU flagging model limitations (Kendall and Gal, 2017). In NLP, it informs machine translation and text generation, separating hallucinations (epistemic) from linguistic ambiguity (aleatoric) (Kuhn et al., 2023). Active learning also exploits this distinction: EU drives sample acquisition, while AU marks uninformative or noisy data (Houlsby et al., 2011; Gal et al., 2017). Despite these advances, evaluating decomposition remains challenging: naive metrics often conflate the two sources, motivating principled benchmarks and task-specific frameworks (Mucsányi et al., 2024; Smith et al., 2025). This evaluation gap is even more pronounced in ICL, where adaptation occurs entirely within the context window, entangling data- and model-driven variability in ways absent from conventional training regimes.

3 Method

We introduce a method for quantifying AU in ICL by leveraging the activations of task-specific attention heads.

Central to this approach is the novel concept of *self-function vectors*, a variant of function vectors, designed to serve as proxies for latent concepts. The method proceeds in four stages:

S1. Causal head selection: Identify salient attention heads $\mathcal{S}_T$ that cause the contextualization of $\mathcal{D}_T$ via causal indirect effect analysis (Section 3.1).

S2. Self-function vector construction: For a given prompt $P^\star = [\mathcal{D}_{\text{ex}}, x^\star]$, extract final-token activations $a_{(\ell,k)}$ from each $(\ell, k) \in \mathcal{S}_T$ and construct a self-function vector $\hat{v}_T$ (Section 3.2).

S3. Self-function vector intervention: Inject $\hat{v}_T$ into the hidden state at the final token during inference, yielding a latent-conditioned prediction $p(y \mid x^\star, \mathcal{D}_{\text{ex}}, \hat{v}_T)$. (Section 3.3).

S4. Uncertainty decomposition: Compute decomposed uncertainties using latent-conditioned predictions (Section 3.4).

Unlike conventional approaches that infer uncertainty only from output variability (e.g., decoding or prompting strategies), our method probes the model's internal representations. This yields a structure-aware and interpretable foundation for uncertainty estimation in ICL (Shorinwa et al., 2024). Our code is available at `https://github.com/LOG-postech/self-fv-icl`.

3.1 Identifying Salient Attention Heads

To identify task-relevant attention heads, we begin with a given task dataset $\mathcal{D}_T$ and estimate the causal contribution of each attention head to the model's predictions. Following Todd et al. (2024), we evaluate this by introducing counterfactual in-context inputs using label-shuffled examples $\tilde{\mathcal{D}}_{\text{ex}}$, which typically degrade performance. Given a corrupted prompt $\tilde{P} = [\tilde{\mathcal{D}}_{\text{ex}}, x_{val}]$, we assess each head (l, k), by replacing its final-token activation during inference with the mean activation vector $\bar{h}^T_{(l,k)}$ computed over correctly labeled inputs. The causal effect (CE) of a head is then defined as the change in prediction probability induced by this replacement:

$$\begin{aligned}\text{CE}(P) =& p_{h^{\tilde{P}}_{(\ell,k)} \to \bar{h}^T_{(\ell,k)}}(y \mid [\tilde{\mathcal{D}}_{\text{ex}}, x_{val}]) \\ &- p(y \mid [\mathcal{D}_{\text{ex}}, x_{val}]).\end{aligned} \tag{6}$$

The notation $\tilde{h}^T_{(\ell,k)} \to \bar{h}^T_{(\ell,k)}$ denotes the activation from a label-shuffled input is substituted with $\bar{h}^T_{(\ell,k)}$. We compute the causal effect multiple times using various prompts, and select the heads with the highest average CE to form the salient set $\mathcal{S}_T$, as they consistently encode task-relevant information (Jiang et al., 2025).

3.2 Self-Function Vectors as Latent Task Representations

Given a test prompt $P^\star$, we extract the final-token activations $h^{P^\star}_{(\ell,k)}$ from each salient head $(\ell, k) \in \mathcal{S}_T$. For each ensemble iteration i, we construct a function vector $\hat{v}^{(i)}_T$ by randomly sampling from $\{h^{P^\star}_{(\ell,k)} \mid (\ell, k) \in \mathcal{S}_T\}$,

$$\hat{v}^{(i)}_T = \sum_{(\ell,k) \subseteq \mathcal{S}_T} h^{P^\star}_{(\ell,k)}. \tag{7}$$

We term these *self-function vectors*, as they encapsulate the prompt's internal representation through task-relevant attention heads. This allows capturing prompt-specific uncertainty, rather than the averaged representation of the function vector. This claim is supported by the experiments in Section 5.

3.3 Self-Function Vectors Intervention

To obtain latent-conditioned predictions, we inject a self-function vector $\hat{v}^{(i)}_T$ into the hidden state at a designated target layer ℓ_t during inference, following Todd et al. (2024). Let h_{ℓ_t} denote the original activation at layer ℓ_t; the intervention modifies it as

$$h'_{\ell_t} = h_{\ell_t} + \hat{v}^{(i)}_T. \tag{8}$$

The modified activation h'_{ℓ_t} is then passed through the language model head to produce the predictive distribution

$$p(y \mid x^\star, \mathcal{D}_{\text{ex}}, \hat{v}^{(i)}_T), \tag{9}$$

as disccused in Jiang et al. (2025). This predictive distribution can be interpreted as a realization of $p(y \mid x^\star, \phi)$ with $\phi \sim p(\phi \mid \mathcal{D}_{\text{ex}})$. This interpretation naturally leads to the Bayesian decomposition of predictive uncertainty.

3.4 Uncertainty Decomposition

We decompose predictive uncertainty using the standard Bayesian decomposition. Specifically, the total uncertainty is given by the entropy of the predictive distribution:

$$\begin{aligned}&\mathcal{H}[p(y \mid x^\star, \mathcal{D}_{\text{ex}})] \\ &= -\sum_{y \in \mathcal{Y}} p(y \mid x^\star, \mathcal{D}_{\text{ex}}) \log p(y \mid x^\star, \mathcal{D}_{\text{ex}}).\end{aligned} \tag{10}$$

AU is approximated by the average entropy of predictions under self-function vector interventions:

$$\mathbb{E}_{p(\phi|\mathcal{D}_{\text{ex}})}[\mathcal{H}(p(y \mid x^\star, \phi))] \approx \frac{1}{N}\sum_{i=1}^{N} \mathcal{H}(p(y \mid x^\star, \mathcal{D}_{\text{ex}}, \hat{v}_T^{(i)})). \quad (11)$$

In practice, we instantiate Equation (11) with $N = 1$, performing a single intervention using a self-function vector $\hat{v}_T$ aggregated from the top causal-effect heads ($top\text{-}k\ \mathcal{S}_T$). Under this setup, AU is a deterministic quantity given a model and an input. We empirically compare this single top-k intervention against ensemble variants that aggregate multiple interventions in Appendix E.2, and find that the top-k single intervention already captures the dominant effect, with ensembling yielding only marginal differences. Finally, EU is obtained as the difference between the total and aleatoric terms.

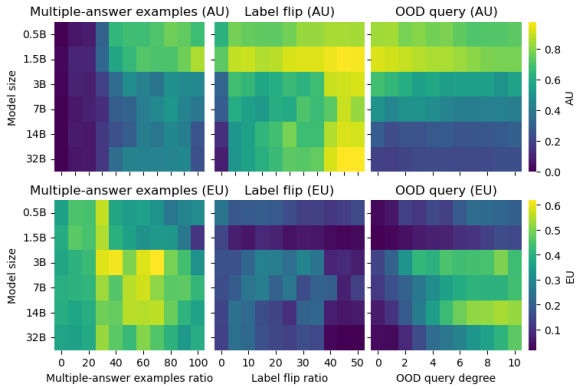


Figure 2: These results show how AU and EU change under various perturbation types and model sizes in synthetic tasks. Each column corresponds to a specific perturbation setting. Within each subplot, the x-axis indicates the degree of perturbation, and the y-axis indicates model size. Color represents the magnitude of each uncertainty. As model size increases, the Multiple-Answer Examples and Label Flip tasks control AU, while the OOD Query task controls EU.

4 Evaluation Protocol

Assessing the quality of uncertainty decomposition is critical for ensuring the reliability and interpretability of model predictions in ICL. However, to the best of our knowledge, no evaluation protocol exists for this purpose. To address this gap, we begin by analyzing how controlled perturbations to examples and queries affect uncertainty in a binary classification. These simplified, well-controlled environments enable the precise attribution of changes in predictive uncertainty to either data- or model-induced factors. Building on this analysis, we propose the first evaluation protocol specifically designed to assess the fidelity of uncertainty decomposition methods on language tasks.

4.1 Toy Experiments

A principled evaluation of uncertainty decomposition requires task settings in which the underlying sources of uncertainty can be manipulated independently. Specifically, it should contain scenarios where AU increases while EU remains constant, and vice versa.

To identify perturbations that induce these distinct uncertainty dynamics, we design synthetic experiments that allow direct control over the data-generating process. This simplified setting offers two primary advantages: it mitigates uncontrolled sources of variation that could affect the uncertainties, and it enables the application of established decomposition techniques (Jayasekera et al., 2025), which are theoretically well-grounded in such simplified settings. The data distributions used in these experiments are visualized in Figure 5.

Figure 2 shows the results of toy experiments, focusing on the tasks in which perturbations produced a clear separation between AU and EU dynamics.

For large models, AU tends to increase as with more perturbation, whereas EU remains relatively unchanged for the Label Flip and Multi-answered Examples tasks as the perturbation degree increases. In contrast, for the OOD Query task, EU increases with perturbation, whereas AU remains mostly fixed. This pattern aligns with prior results in computer vision, where AU is evaluated through human label disagreement and EU through capturing distributional shift (Mucsányi et al., 2024).

Interestingly, smaller models exhibit changes in both AU and EU across all tasks. As model size increases, however, one source of uncertainty tends to stabilize while the other shows a more consistent increase—indicating improved separation of uncertainty sources in larger models.

These results demonstrate that the designed perturbations can isolate distinct sources of uncertainty. Further implementation details and results, including task variants where AU and EU cannot be clearly separated, are provided in Appendix A.

4.2 WordNetMCQ-dataset Construction

From the toy experiments, we identify three tasks controlling AU—Multiple-Answer Examples and

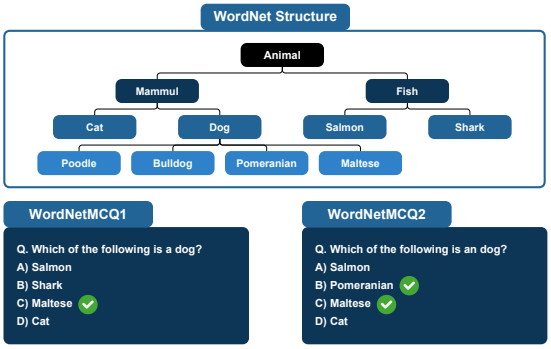


Figure 3: Construction of **WordNetMCQ1** (single-answer) and **WordNetMCQ2** (multiple-answer) question types from WordNet. This design provides the building blocks to carefully control task situations within ICL by varying the proportion of each question type in the demonstration prompts.

Label Flip—and one task controlling EU—OOD Query. We then generalize these synthetic insights into tasks instantiated with WordNet (Miller, 1994), thereby providing a more natural and language-grounded evaluation.

WordNet provides an ideal foundation, as it encodes fine-grained lexical semantics through synsets, hierarchical relations (e.g., hypernymy), and sense-level distinctions. Building on this, we construct multiple-choice questions, which we refer to as WordNetMCQ, following the subjective question setting (Abbasi-Yadkori et al., 2024). Specifically, we distinguish between:

- WordNetMCQ1: single-answer multiple-choice questions, where only one option is correct among four.
- WordNetMCQ2: multiple-answer multiple-choice question dataset in which each question has two correct answers.

This design allows us to carefully control situations for each task in ICL, as illustrated in Figure 3.

Multiple-Answer Examples This task uses WordNetMCQ2 queries, while the examples are intentionally mixed between WordNetMCQ1 and WordNetMCQ2 instances. Consequently, if the examples contain mostly WordNetMCQ1 examples, the model tends to choose only one answer for a WordNetMCQ2 query. This shows how the ICL setup naturally varies the degree of ambiguity.

Label Flip Here, we inject controlled label noise into the examples by flipping their labels according to a predefined ratio. As the flip ratio increases, the supervision becomes more inconsistent, forcing the model to learn from conflicting signals. This systematic perturbation directly raises AU, since the data itself becomes unreliable.

OOD Query To induce EU, we perturb the queries while preserving their original meaning. Specifically, we rephrase queries and insert special characters. By gradually increasing the intensity of these modifications, we adjust the degree of distributional shift for out-of-distribution inputs.

4.3 Evaluation Metric

The degree of perturbation applied to each task is represented as an ordered categorical variable, whereas the resulting uncertainty is measured on a continuous scale. Assessing the relationship between these two variables requires a correlation measure capable of capturing monotonic, and potentially non-linear, dependencies.

The statistical literature offers several such measures, including Pearson's correlation coefficient, Kendall's tau, and Somers' d, each exhibiting different sensitivities to distributional characteristics. Following the comparative framework proposed by Göktaş and Isci Guneri (2011), we empirically evaluated these candidate measures. Among them, Spearman's rank correlation consistently demonstrated the most robust and reliable performance across our experimental conditions.

Based on these findings, we adopt Spearman's rank correlation as our primary evaluation metric. Details on the candidate measures, experimental design, and results are provided in Appendix B.2.

5 Experiments

We evaluate our method through a series of uncertainty-controlled experiments and downstream benchmarks. The goals of our experiments are threefold: (i) to verify that our evaluation protocol can faithfully separate AU and EU in ICL, (ii) to demonstrate that the decomposed signals provide practical benefits in tasks such as hallucination detection, and (iii) to assess the specific contribution of self-function vectors through ablation analyses.

Experimental Setup We use LLaMA2-7B, LLaMA2-13B, and LLaMA2-70B (Touvron et al., 2023), along with Qwen2.5-7B (Qwen et al., 2025) and Mistral-7B (Jiang et al., 2023), as base models to assess robustness across both scale and architecture. We restrict our method to the top-20

causal heads and intervene at one-third of the transformer's depth (e.g., the 10th layer for LLaMA2-7B, the 13th for LLaMA2-13B), following the layer-subsampling strategy (Todd et al., 2024; Liu and Deng, 2025). This design avoids unrealistic full-layer searches. Evaluation primarily relies on WordNetMCQ, which enables controlled manipulation of uncertainty, with additional experiments on AG News (Zhang et al., 2016), Emotion (Saravia et al., 2018), HellaSwag (Zellers et al., 2019), and GSM8K (Cobbe et al., 2021) to assess applicability across classification, reasoning, and math domains (see Appendix D for further details).

Baseline We consider Total entropy, Semantic Entropy (Kuhn et al., 2023), and UQ_ICL (Ling et al., 2024) as representative baselines for uncertainty quantification and decomposition. Semantic Entropy measures uncertainty based on the entropy of semantically clustered model outputs, while UQ_ICL decomposes total uncertainty in in-context learning into AU and EU. We exclude methods that manipulate queries or vary the number of examples for estimation, as it is not straightforward to make direct comparisons with them. Extended baseline comparisons including MaxProb (Hendrycks and Gimpel, 2017) and Lookback Lens (Chuang et al., 2024), as well as a discussion of related work (Vazhentsev et al., 2025), are provided in Appendix F.

5.1 Aleatoric Uncertainty Control

Multiple-Answer Examples Ratio Variation We vary the ratio of WordNetMCQ1 and WordNetMCQ2 examples for WordNetMCQ2 queries. As shown in Table 1, Self-function vector yields the highest correlations across LLaMA2-7B, 13B, and 70B, suggesting that it more clearly reflects the effect of data ambiguity.

Method	LLaMA2-7B	LLaMA2-13B	LLaMA2-70B
Total Entropy	0.514	0.426	0.208
Semantic Entropy	-0.277	-0.248	-0.301
UQ_ICL	-0.093	-0.097	-0.380
Function Vector	0.633	0.429	0.257
Self-FV	**0.640**	**0.435**	**0.292**

Table 1: Spearman correlations (↑) between WordNetMCQ1/2 ratio variation and uncertainty measures.

Label Noise Ratio Variation We vary the flip ratio of example labels to control levels of AU. Results in Table 2 show that Self-function vector often attains higher correlations compared to other methods across LLaMA2 (7B/13B/70B) as well as Qwen2.5-7B and Mistral-7B, indicating that it captures the effect of label noise in a comparatively stable manner and that the mechanism generalizes across modern transformer-based LLMs. This effect is particularly pronounced at larger model scales, consistent with the trend observed in Section 4.1 that uncertainty control becomes more evident as model capacity increases.

Model	Method	WNMCQ1	HellaSwag	GSM8K	AG News	Emotion
LLaMA2-7B	Total Entropy	**0.177**	0.061	0.143	0.739	**0.242**
	Semantic Entropy	0.160	0.028	**0.234**	0.101	0.073
	UQ_ICL	0.172	0.028	0.111	0.166	0.113
	Function Vector	0.164	0.057	0.026	0.668	0.207
	Self-FV	0.167	**0.064**	-0.023	**0.746**	0.238
LLaMA2-13B	Total Entropy	0.385	0.175	0.356	0.730	**0.369**
	Semantic Entropy	0.364	0.105	0.015	0.323	0.159
	UQ_ICL	0.422	**0.200**	0.161	0.443	0.211
	Function Vector	0.316	0.123	0.360	0.586	0.364
	Self-FV	**0.424**	0.183	**0.368**	**0.741**	0.358
LLaMA2-70B	Total Entropy	0.734	0.491	0.389	**0.767**	0.387
	Semantic Entropy	0.712	0.340	0.137	0.554	0.316
	UQ_ICL	0.734	0.430	0.305	**0.807**	0.367
	Function Vector	0.628	0.406	0.355	0.751	0.366
	Self-FV	**0.798**	**0.566**	**0.476**	0.731	**0.416**
Qwen2.5-7B	Total Entropy	0.737	**0.580**	0.499	0.705	**0.517**
	Semantic Entropy	0.605	0.145	0.085	0.262	0.298
	UQ_ICL	0.690	0.456	0.023	0.306	0.181
	Function Vector	0.742	0.573	0.527	0.703	0.500
	Self-FV	**0.751**	0.574	**0.542**	**0.714**	0.487
Mistral-7B	Total Entropy	0.546	0.172	0.392	**0.659**	**0.318**
	Semantic Entropy	0.557	-0.094	0.094	0.539	0.191
	UQ_ICL	0.457	0.018	0.104	0.212	0.027
	Function Vector	0.533	0.170	0.418	0.653	0.314
	Self-FV	**0.657**	**0.173**	**0.436**	0.645	0.311

Table 2: Spearman correlations (↑) between label noise ratio variation and uncertainty.

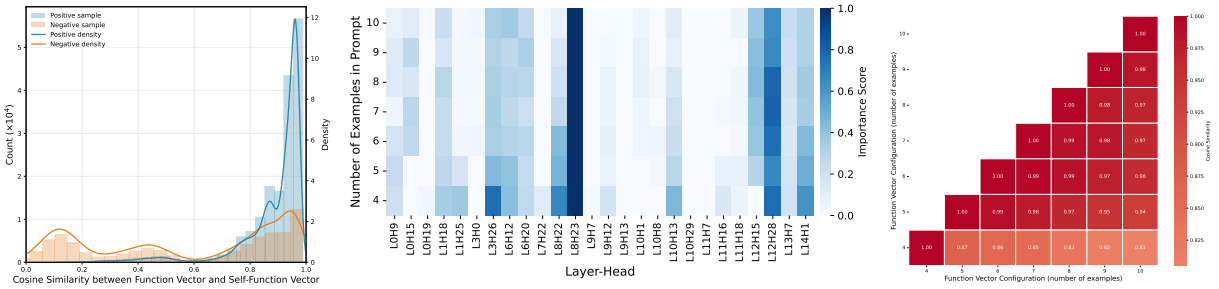
Together, these results show that mechanistic approaches exhibit stronger and more consistent correlations than alternative approaches, indicating that it more reliably reflects controlled sources of ambiguity.

5.2 Epistemic Uncertainty Control

OOD Query Variation This setup targets EU by moving queries progressively out of distribution. As shown in Table 3, Function vector and Self-function vector show lower correlation with OOD variation compared to other entropy-based baselines across LLaMA2 (7B/13B/70B), Qwen2.5-7B, and Mistral-7B, indicating that they better isolate EU.

Method	LLaMA2-7B	LLaMA2-13B	LLaMA2-70B	Qwen2.5-7B	Mistral-7B
Total Entropy	0.213	0.115	0.081	0.295	-0.039
Semantic Entropy	0.288	0.204	0.108	**0.089**	0.557
UQ_ICL	0.310	0.233	0.084	0.110	0.457
Function Vector	0.184	0.089	0.063	0.294	-0.0213
Self-FV	**0.148**	**0.076**	**0.026**	0.289	**-0.0212**

Table 3: Spearman correlations under OOD query variation, evaluated using WNMCQ1. Correlations closer to zero ($|\rho|\downarrow$) indicate better isolation of EU; bold marks the smallest magnitude per column.



(a) Cosine Similarity Distributions by Prediction Outcome

(b) Top 20 Causal Heads

(c) Function Vector Cosine Similarity Matrix

Figure 4: Overall figure caption describing all three subplots. (a) This illustrates the distribution of cosine similarity between the self-function vectors and function vectors, separated by task correctness. Clear distinction between two distribution supports self-function vectors better reflect $P^\star$-specific internal task representation than function vectors. (b) This highlights the top 20 causal heads by their importance score. The x-axis indicates the number of examples used for CIE. The model identifies consistent causal heads across varying numbers of examples. (c) This shows cosine similarity between function vectors derived from various numbers of examples. The similarity of adjacent numbers converges to 1, indicating the consistent formation of the concept.

5.3 Hallucination Detection

We also test on hallucination detection, a common downstream benchmark in uncertainty quantification. As shown in Table 4, mechanistic approaches generally outperform entropy-based baselines and remain competitive with or better than MaxProb; Lookback Lens, which relies on attention maps derived from the generation process, is less compatible with our single-token setting and underperforms accordingly. Self-function vector performs comparably or better across several datasets, indicating practical benefits in trustworthy-related applications, with hyperparameters selected via a grid search (details in Appendix D). PRR results

Model	Method	WNMCQ1	Hellaswag	GSM8K	AG News	Emotion
LLaMA2-7B	Total Entropy	0.872	0.608	0.554	0.843	0.646
	Semantic Entropy	0.859	0.578	0.626	0.532	0.538
	UQ_ICL	0.884	0.575	0.657	0.805	**0.681**
	Function Vector	0.8989	**0.624**	0.719	**0.850**	0.651
	Self-FV	**0.8993**	0.619	**0.749**	0.845	0.647
LLaMA2-13B	Total Entropy	0.873	0.673	0.590	0.832	0.666
	Semantic Entropy	0.904	0.665	0.600	0.598	0.489
	UQ_ICL	**0.912**	0.685	0.616	**0.852**	**0.680**
	Function Vector	0.899	**0.696**	0.627	0.848	0.676
	Self-FV	0.889	0.685	**0.638**	0.845	0.668

Table 4: AUROC (↑) on diverse text classification datasets. Mechanistic methods generally outperform entropy-based baselines.

(Fadeeva et al., 2025) are provided in Appendix G and show a consistent trend. In summary, Self-function vector better reflects controlled AU/EU variations and performs competitively on conventional uncertainty quantification tasks, indicating that a mechanistic perspective offers a promising path for more reliable and interpretable ICL behavior.

5.4 Ablation Study

To validate our approach of AU measurement, we conduct a series of ablation studies, each targeting a component of the method.

Self-function vectors effectively capture prompt-specific concept representations We investigate whether self-function vectors more effectively capture the model's internal representation of the target prompt $P^\star$, compared to function vectors which encode an average representation across prompts.

To test this, we separate prompts that the model generates the correct output with a clear concept from those it does not. We then compare the cosine similarity between self-function vectors and function vectors across these two groups.

As shown in Figure 4a, for the clear prompts that the model predicts the correct answer, cosine similarity is high, that is, the self-function aligns closely with the function vector. In contrast, the prompts that the model couldn't process the task show low cosine similarity, indicating the self-function vector diverges significantly from the function vector.

This trend is consistent across datasets, supporting that self-function vectors reflect the model's internal task representation specific to $P^\star$. The other experiments are reported in Appendix E.1.

Further exploration We support that causal heads and corresponding function vectors indeed capture the internal representation of ICL through two experiments: Causal heads and function vectors under varying numbers of examples.

First, we observe a set of important heads that

consistently emerge across different shot settings. This stable selection implies the existence of certain heads that are crucial for the ICL task itself, rather than being specific to a particular ICL configuration, as shown in Figure 4b. Detailed results are provided in Appendix E.3.

Second, the cosine similarity between the function vectors derived from prompts with adjacent numbers of in-context examples approaches 1 as the number of examples increases, as shown in Figure 4c. This convergence implies that function vectors indeed form a consistent conceptual understanding of the task. Additional results are presented in Appendix E.4.

6 Conclusion

This work addresses the critical challenge of decomposing predictive uncertainty in ICL into its aleatoric and epistemic components. We introduce a novel methodology grounded in Bayesian views and mechanistic interpretability to directly quantify AU from the model's internal representations. Furthermore, we establish the first rigorous evaluation protocol specifically designed for uncertainty decomposition in ICL. Our experiments demonstrate not only a robust capacity to disentangle these two uncertainty sources, the also the practical utility in hallucination detection. Ultimately, this research suggests a new direction that bridges mechanistic understanding with uncertainty estimation.

Limitations

While this work seeks to bridge uncertainty quantification and mechanistic interpretability, several limitations remain.

First, our use of function vectors as an approximation of posterior sampling over latent tasks is indirect and interpretive. The extent to which they capture true posterior structure remains uncertain, and warrants deeper theoretical analysis.

Second, key hyperparameters, such as the intervention layer and the number of causal heads, require per-model tuning, which limits generalization across architectures.

We view these limitations as opportunities for future work, particularly in expanding task coverage, strengthening empirical validation, and refining the theoretical link between internal representations and Bayesian inference.

Acknowledgments

This work was partly supported by the Institute of Information & Communications Technology Planning & Evaluation (IITP) grant funded by the Government of the Republic of Korea (Ministry of Science and ICT) (RS-2019-II191906, Artificial Intelligence Graduate School Program (POSTECH); RS-2022-II220959, (Part 2) Few-Shot Learning of Causal Inference in Vision and Language for Decision Making; RS-2023-00216011, Development of Artificial Complex Intelligence for Conceptually Understanding and Inferring like Humans); and by the High-Performance Computing Support Project, funded by the Government of the Republic of Korea (Ministry of Science and ICT) (RQT-25-070137).

References

Yasin Abbasi-Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvari. 2024. To believe or not to believe your LLM: Iterativeprompting for estimating epistemic uncertainty. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, and 12 others. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Bryan Chan, Xinyi Chen, András György, and Dale Schuurmans. 2025. Toward understanding in-context vs. in-weight learning. In *The Thirteenth International Conference on Learning Representations*.

Stephanie Chan, Adam Santoro, Andrew Lampinen, Jane Wang, Aaditya Singh, Pierre Richemond, James McClelland, and Felix Hill. 2022. Data distributional properties drive emergent in-context learning in transformers. In *Advances in Neural Information Processing Systems*, volume 35, pages 18878–18891. Curran Associates, Inc.

Yung-Sung Chuang, Linlu Qiu, Cheng-Yu Hsieh, Ranjay Krishna, Yoon Kim, and James R. Glass. 2024. Lookback lens: Detecting and mitigating contextual hallucinations in large language models using only attention maps. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias

Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, and 6 others. 2021. A mathematical framework for transformer circuits.

Ekaterina Fadeeva, Aleksandr Rubashevskii, Artem Shelmanov, Sergey Petrakov, Haonan Li, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Pontus Stenetorp. 2025. LM-polygraph: Uncertainty estimation for language models. *Transactions of the Association for Computational Linguistics*, 13.

Fabian Falck, Ziyu Wang, and Christopher C. Holmes. 2024. Is in-context learning in large language models bayesian? a martingale perspective. In *Forty-first International Conference on Machine Learning*.

Yarin Gal, Riashat Islam, and Zoubin Ghahramani. 2017. Deep Bayesian active learning with image data. In *Proceedings of the 34th International Conference on Machine Learning*, volume 70 of *Proceedings of Machine Learning Research*, pages 1183–1192. PMLR.

Atila Göktaş and Oznur Isci Guneri. 2011. A comparison of the most commonly used measures of association for doubly ordered square contingency tables via simulation. *Metodoloski Zvezki*, 8.

Roee Hendel, Mor Geva, and Amir Globerson. 2023. In-context learning creates task vectors. In *The 2023 Conference on Empirical Methods in Natural Language Processing*.

Dan Hendrycks and Kevin Gimpel. 2017. A baseline for detecting misclassified and out-of-distribution examples in neural networks. In *International Conference on Learning Representations*.

Juyeon Heo, Christina Heinze-Deml, Oussama Elachqar, Kwan Ho Ryan Chan, Shirley You Ren, Andrew Miller, Udhyakumar Nallasamy, and Jaya Narain. 2025. Do LLMs "know" internally when they follow instructions? In *The Thirteenth International Conference on Learning Representations*.

Neil Houlsby, Ferenc Huszár, Zoubin Ghahramani, and Máté Lengyel. 2011. Bayesian active learning for classification and preference learning. *Preprint*, arXiv:1112.5745.

I. Shavindra Jayasekera, Jacob Si, Filippo Valdettaro, Wenlong Chen, A. Aldo Faisal, and Yingzhen Li. 2025. Variational uncertainty decomposition for in-context learning. *Preprint*, arXiv:2509.02327.

Albert Q. Jiang, Alexandre Sablayrolles, Arthur Mensch, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Florian Bressand, Gianna Lengyel, Guillaume Lample, Lucile Saulnier, Lélio Renard Lavaud, Marie-Anne Lachaux, Pierre Stock, Teven Le Scao, Thibaut Lavril, Thomas Wang, Timothée Lacroix, and William El Sayed. 2023. Mistral 7b. *Preprint*, arXiv:2310.06825.

Gangwei Jiang, Caigao JIANG, Zhaoyi Li, Siqiao Xue, JUN ZHOU, Linqi Song, Defu Lian, and Ying Wei. 2025. Unlocking the power of function vectors for characterizing and mitigating catastrophic forgetting in continual instruction tuning. In *The Thirteenth International Conference on Learning Representations*.

Hui Jiang. 2024. A latent space theory for emergent abilities in large language models.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

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

Quentin Lhoest, Albert Villanova del Moral, Yacine Jernite, Abhishek Thakur, Patrick von Platen, Suraj Patil, Julien Chaumond, Mariama Drame, Julien Plu, Lewis Tunstall, Joe Davison, Mario Šaško, Gunjan Chhablani, Bhavitvya Malik, Simon Brandeis, Teven Le Scao, Victor Sanh, Canwen Xu, Nicolas Patry, and 13 others. 2021. Datasets: A community library for natural language processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 175–184, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, Guangji Bai, Liang Zhao, and Haifeng Chen. 2024. Uncertainty quantification for in-context learning of large language models. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 3357–3370, Mexico City, Mexico. Association for Computational Linguistics.

Yiting Liu and Zhi-Hong Deng. 2025. Iterative vectors: In-context gradient steering without backpropagation. In *Forty-Second International Conference on Machine Learning*.

George A. Miller. 1994. WordNet: A lexical database for English. In *Human Language Technology: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Sewon Min, Xinxi Lyu, Ari Holtzman, Mikel Artetxe, Mike Lewis, Hannaneh Hajishirzi, and Luke Zettlemoyer. 2022. Rethinking the role of demonstrations: What makes in-context learning work? *Preprint*, arXiv:2202.12837.

Bálint Mucsányi, Michael Kirchhof, and Seong Joon Oh. 2024. Benchmarking uncertainty disentanglement: Specialized uncertainties for specialized tasks. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, and 7 others. 2022. In-context learning and induction heads. *Preprint*, arXiv:2209.11895.

Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jingren Zhou, and 25 others. 2025. Qwen2.5 technical report. *Preprint*, arXiv:2412.15115.

Elvis Saravia, Hsien-Chi Toby Liu, Yen-Hao Huang, Junlin Wu, and Yi-Shin Chen. 2018. CARER: Contextualized affect representations for emotion recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3687–3697, Brussels, Belgium. Association for Computational Linguistics.

Ola Shorinwa, Zhiting Mei, Justin Lidard, Allen Z. Ren, and Anirudha Majumdar. 2024. A survey on uncertainty quantification of large language models: Taxonomy, open research challenges, and future directions. *Preprint*, arXiv:2412.05563.

Freddie Bickford Smith, Jannik Kossen, Eleanor Trollope, Mark van der Wilk, Adam Foster, and Tom Rainforth. 2025. Rethinking aleatoric and epistemic uncertainty. In *Forty-second International Conference on Machine Learning*.

Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. 2024. Function vectors in large language models. In *Proceedings of the 2024 International Conference on Learning Representations*. ArXiv:2310.15213.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, Dan Bikel, Lukas Blecher, Cristian Canton Ferrer, Moya Chen, Guillem Cucurull, David Esiobu, Jude Fernandes, Jeremy Fu, Wenyin Fu, and 49 others. 2023. Llama 2: Open foundation and fine-tuned chat models. *Preprint*, arXiv:2307.09288.

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

Xinyi Wang, Wanrong Zhu, Michael Saxon, Mark Steyvers, and William Yang Wang. 2023. Large language models are latent variable models: Explaining and finding good demonstrations for in-context learning. In *Thirty-seventh Conference on Neural Information Processing Systems*.

Yihan Wang, Si Si, Daliang Li, Michal Lukasik, Felix Yu, Cho-Jui Hsieh, Inderjit S Dhillon, and Sanjiv Kumar. 2024. Two-stage LLM fine-tuning with less specialization and more generalization. In *The Twelfth International Conference on Learning Representations*.

Ziyu Wang and Christopher C. Holmes. 2025. On subjective uncertainty quantification and calibration in natural language generation. In *The 28th International Conference on Artificial Intelligence and Statistics*.

Lisa Wimmer, Yusuf Sale, Paul Hofman, Bernd Bischl, and Eyke Hüllermeier. 2023. Quantifying aleatoric and epistemic uncertainty in machine learning: Are conditional entropy and mutual information appropriate measures? In *Proceedings of the Thirty-Ninth Conference on Uncertainty in Artificial Intelligence*, volume 216 of *Proceedings of Machine Learning Research*, pages 2282–2292. PMLR.

Sang Michael Xie, Aditi Raghunathan, Percy Liang, and Tengyu Ma. 2022. An explanation of in-context learning as implicit bayesian inference. In *International Conference on Learning Representations*.

Fanghua Ye, Mingming Yang, Jianhui Pang, Longyue Wang, Derek F. Wong, Emine Yilmaz, Shuming Shi, and Zhaopeng Tu. 2024. Benchmarking LLMs via uncertainty quantification. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Kayo Yin and Jacob Steinhardt. 2025. Which attention heads matter for in-context learning? *Preprint*, arXiv:2502.14010.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. 2019. Hellaswag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*.

Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. Character-level convolutional networks for text classification. *Preprint*, arXiv:1509.01626.

A Synthetic Data Experiments

Here, we present a detailed description of the construction of the synthetic data in Section 4.1. We also report the corresponding results, which were omitted from the main paper due to space limitations.

A.1 Construction of Synthetic Data

We build binary classification ICL tasks on the Two Moons dataset. Following the synthetic ICL of Jayasekera et al. (2025), numerical values are rendered as string format for use in ICL prompts.

We construct binary classification tasks using the Two Moons dataset, where sampled examples are converted into string representations for ICL.

`make_moons` in scikit-learn generates ICL examples

- $(\cos(x), \sin(x))$ for class 0,
- $(1 - \cos(x), 0, 5 - \sin(x))$ for class 1,

with noise parameter which we set as 0.1. Unless otherwise stated, we draw 30 examples.

To examine the behavior of AU and EU, we perturb both examples and queries under five families of tasks—Multiple-answer examples, Label flip, Ambiguous query, Example number and OOD query. The implementation of each task is summarized in Table 5 and Figure 5.

Multiple-answer Examples From the class posteriors evaluated at each example location, we compute the probability margin between the classes. We then relabel the bottom–$k\%$ of examples by this margin as multi-label, which we denoted as 2. We vary the ratio of multi–labeled examples to single–labeled from 0% to 100% (default). Queries are selected near the decision boundary to simulate the ambiguity that arises when, given single-labeled context, the model must choose one answer from two valid options.

Label Flip A fraction of example labels is flipped uniformly at random. The flip ratio is swept from 0% (default) to 50%; ratios above 50% are symmetric and thus omitted.

Ambiguous Query To move queries toward the class boundary, we contract their radius with a non-negative degree parameter d:

- $\left(\frac{\cos x}{1+d}, \frac{\sin x}{1+d/3}\right)$,
- $\left(1 - \frac{\cos x}{1+d}, 0.5 - \frac{\sin x}{1+d/3}\right)$.

We sweep d from 0 to 100.

Example Number We vary the number of examples from 0 to 50.

OOD Query To place queries Out-Of-Distribution (OOD), we translate small grids away from the moons:

- $\{(0 + i, 1 + j + 0.2 \times \text{degree}) : i, j \in \{0, 0.2\}\}$,
- $\{(1 + i, -0.6 + j - 0.2 \times \text{degree}) : i, j \in \{0, 0.2\}\}$.

We sweep d from 0 to 11.

Among these, we retain only tasks in which either AU or EU consistently increase while the other remains stable, thereby composing the WordNet evaluation protocol.

A.2 Benchmark Validation Results

We apply the uncertainty decomposition method proposed by Jayasekera et al. (2025) to the Qwen2.5 models (Qwen et al., 2025), spanning from 0.5B to 32B parameters. Figure 6 illustrates the behaviors of AU and EU across different model scales under various perturbation settings.

Across all tasks, we observe distinct AU and EU fluctuations depending on model size. We claim that effective control of AU and EU requires a model sufficiently large to perform ICL reliably.

For larger models (typically above 7B parameters), the Multi-Answered Examples task exhibits an increase in AU, while EU fluctuates without a clear trend. The observed drop in AU at the 100% ratio setting likely results from a reduced number of valid answer candidates, as this configuration excludes multi-labeled examples.

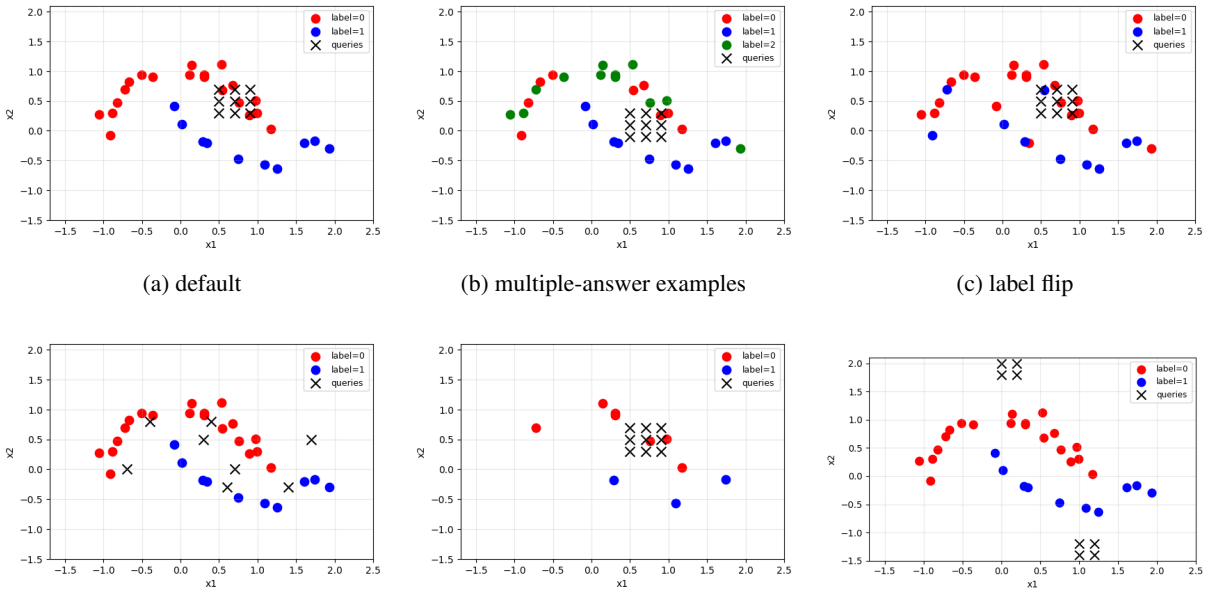
Similarly, in the Label Flip task, AU consistently increases with only minor EU fluctuations. Consequently, we adopt the Multi-Answered Examples and Label Flip tasks as benchmark settings for AU control.

In contrast, the Ambiguous Query task demonstrates simultaneous increases in both AU and EU, while the Example Number task shows decreases in both measures. Therefore, we exclude these tasks from the evaluation protocol.

Finally, the Out-of-Distribution (OOD) Query task yields stable EU and increasing AU as model size grows, making it a suitable protocol for EU control.

Tasks	Control
Multiple-answer examples	For examples with similar class-probability logits, map their labels to 2.
Label flip	Randomly flip example labels.
Ambiguous query	Move the query closer to the decision boundary of the Two Moons dataset.
Example number	Change the number of in-context examples.
OOD query	Move the query further from the decision boundary of the Two Moons dataset.

Table 5: Task Implementation on Synthetic Data



(a) default (b) multiple-answer examples (c) label flip

(d) ambiguous query (e) example number (f) OOD query

Figure 5: Visualization of synthetic data for each task

B Evalution Metric Comparison

In this section, we investigate several candidate metrics to identify the most appropriate one for our task.

B.1 Candidate Metrics

We first provide a brief description of each metric under consideration, outlining their key characteristics and differences.

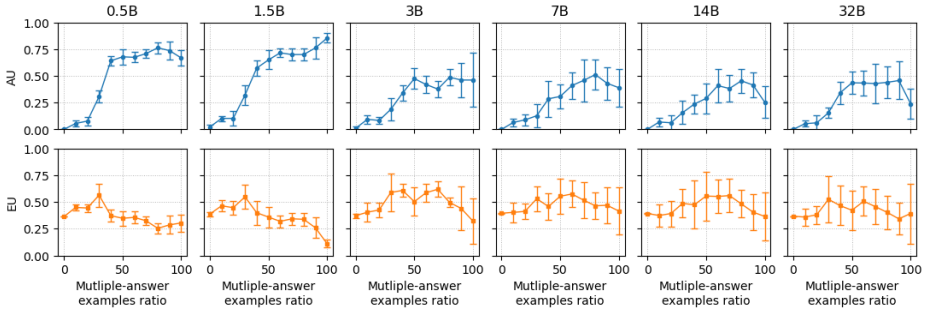
Pearson's correlation. Pearson's correlation quantifies the linear association between two continuous variables. It is defined as the standardized covariance and ranges from -1 to $+1$, with extremal values indicating perfect positive or negative linear dependence. Under joint normality, it fully characterizes dependence, making uncorrelatedness equivalent to independence.

Spearman's rank-correlation. Spearman's correlation is a nonparametric measure of monotonic association. It is obtained by replacing observed values with ranks and then computing the Pearson correlation on these ranks. By focusing on order rather than magnitude, it is robust to non-normality, nonlinear transformations, and outliers.

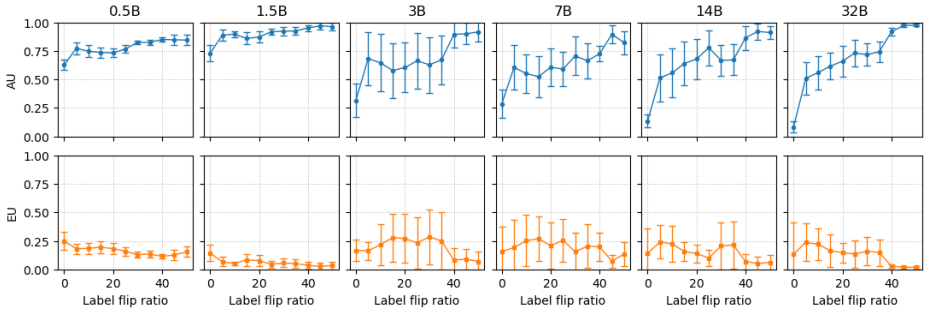
Goodman-Kruskal's gamma. Gamma is a measure of monotonic association between two ordinal variables. It is defined as the difference between concordant and discordant pairs, normalized by their sum. Because ties are excluded, gamma is most suitable when relative ordering is of primary interest rather than their absolute differences.

Kendall's Tau-b. Tau-b also measures ordinal association but explicitly corrects for ties, providing a more balanced statistic when tied observations are common.

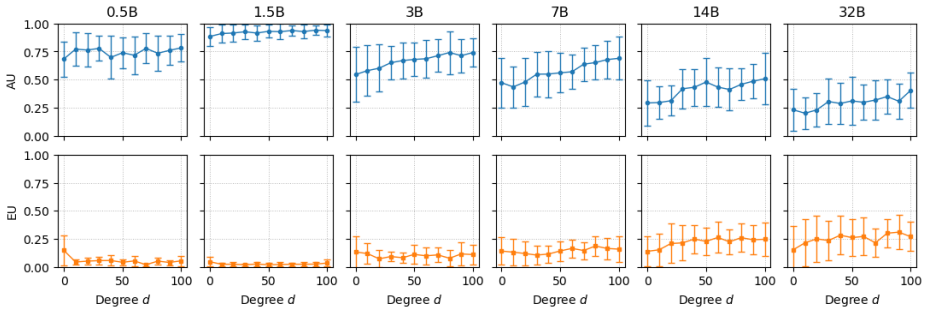
Kendall's Tau-c. While conceptually similar to Kendall's tau-b, Tau-c introduces an adjustment for table dimensionality, ensuring values remain



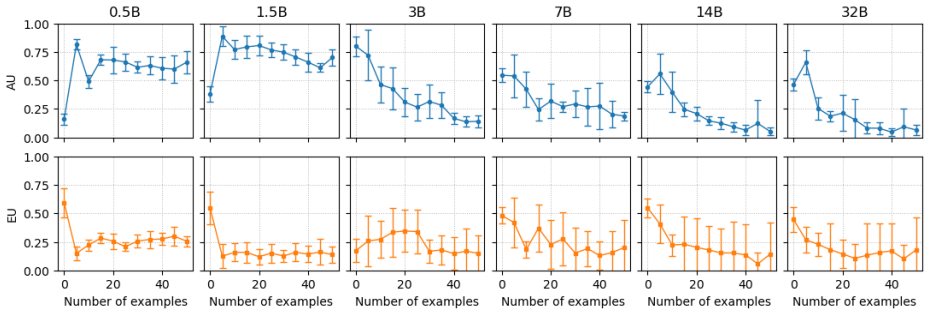
(a) multi-answered examples



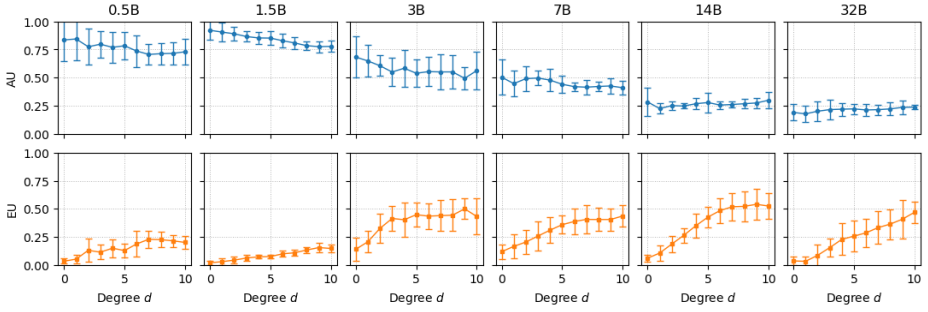
(b) label flip



(c) ambiguous query



(d) example number



(e) OOD query

Figure 6: AU / EU change under each perturbation.

between -1 and $+1$ even when the two variables have different category counts.

Somers' d. Somers' d is an asymmetric extension of Kendall's tau, treating one variable as independent and the other as dependent, and correcting only for ties on the independent variable.

B.2 Experiment Details of Section 4.3

We present a toy example in which we empirically compare the performance of these metrics. Based on this analysis, we determine the most suitable metric to employ in our evaluation.

Recall that each task has a uniform number of data across its ordinal level, while measured uncertainty is on the real line. To find the most suitable metric that captures the correlation between levels and uncertainties, we first generated two random variables with correlation ρ as follows:

- $X = \frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2} X_1 + \frac{\sqrt{1+\rho}-\sqrt{1-\rho}}{2} X_2$,
- $X = \frac{\sqrt{1-\rho}+\sqrt{1-\rho}}{2} X_1 + \frac{\sqrt{1+\rho}+\sqrt{1-\rho}}{2} X_2$,

where X_1 and X_2 are sampled from a normal distribution. The variable X is discretized into ordered categories with approximately uniform frequencies via quantile-based binning, while Y remains continuous and is transformed by monotonic functions such as exponential, arctan, cubic, and cubic-root to induce nonlinear dependence.

This procedure yielded a non-symmetric contingency structure, reflecting an ordered categorical independent variable and a continuous dependent variable with a non-linear dependence. We then measure every candidate metrics for various ρ. The result is in Figure 7. We observe that Spearman's rank correlation most faithfully captures the underlying correlation with the lowest variance, and is therefore adopted as the evaluation metric for our benchmark.

C Prompt Examples

In Section 5.2, we manipulated the OOD (Out-of-Distribution) levels of the query. Below, we present examples of the prompts used at each OOD level. As the OOD level increases, the query text becomes progressively noisier, while the core semantic meaning remains the same.

- **Level 0 Example (Original)**

```
Question: Which of the following is/are
types of debris?
Choices:
A. leg of lamb
B. sexton
C. shopper
D. slack
Answer: D
```

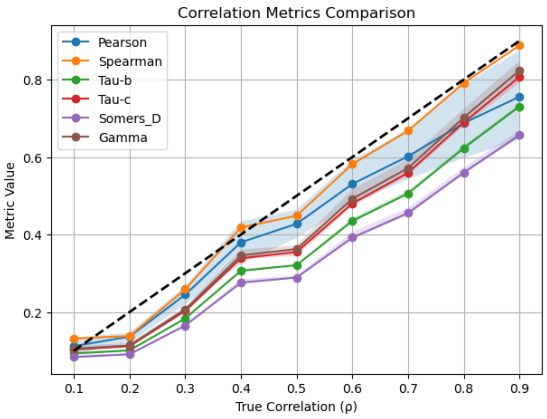


Figure 7: Metric evaluation

- **Level 1 Example**

```
Question: Among the options, which one(s)
classify as types of debris?
Choices:
A. leg of lamb
B. sexton
C. shopper
D. slack
Answer: D
```

- **Level 2 Example**

```
Answer: D

Answer: Question: Among the options,
which !one(s) classify as
types of de@bris?
Choices:
A. leg of lamb
B. sexton
C. shopper
D. slack
Answer: D
```

- **Level 3 Example**

```
Question: Among the options, which
$one(s) classify as types#
of de!br$is?
Choices:
A. leg of lamb
B. sexton
C. shopper
D. slack
Answer: D
```

D Experiment Details

Baseline We consider Total Entropy, Semantic Entropy (Kuhn et al., 2023), and UQ_ICL (Ling et al., 2024) as representative baselines for uncertainty quantification and decomposition. Total Entropy measures predictive uncertainty directly from the token-level probability distribution. Semantic Entropy computes entropy over semantically clustered generations, thereby discounting superficial lexical variation. UQ_ICL decomposes total uncertainty in in-context learning into aleatoric and epistemic components, attributing the former to inherent data ambiguity and the latter to variability arising from model parameters or prompt configurations.

In addition, Wang and Holmes (2025) recently introduced a principled framework for subjective uncertainty quantification and calibration in natural language generation, including discussions relevant to in-context learning, while we respectfully acknowledge the contribution of Jayasekera et al. (2025), who proposed a variational method for decomposing uncertainty in in-context learning. However, these approaches are excluded from our direct comparisons, as they typically require external APIs or well instruction-tuned models to generate auxiliary data, or rely on explicitly varying the number of in-context examples.

Data We use datasets from HuggingFace Datasets (Lhoest et al., 2021): AGNews (Zhang et al., 2016), Emotion (Saravia et al., 2018), HellaSwag (Zellers et al., 2019), and GSM8K (Cobbe et al., 2021), where AGNews and Emotion are evaluated in the standard ICL setting with open-ended responses (Liu and Deng, 2025; Ling et al., 2024), and HellaSwag and GSM8K are formatted as multiple-choice questions (Ye et al., 2024), consistent with our WNMCQ1 setup. For datasets with pre-defined train/test splits, we use them directly; otherwise, we randomly split into train/test with an 8:2 ratio. The validation set used for causal effect estimation is drawn from 10% of the training data. The number of in-context examples is fixed to 10, both for causal effect estimation and for inference. In aleatoric uncertainty control experiments, the proportion of WordNetMCQ1 examples is varied over {0, 30, 60, 90, 100}%, and the label-flipping ratio for noise injection follows the same schedule.

Hyperparameters By default (all tasks except hallucination detection), we intervene at one-third of the transformer: layer 10 for LLaMA2-7B and layer 13 for LLaMA2-13B, using the top-20 causal heads. For hallucination detection, we select hyperparameters via a small grid search with `interv_layer` $\in \{7, 8, 9, 10, 11, 12, 13, 14, 15\}$ and `num_top_heads` $\in \{8, 10, 12, 14, 16, 18, 20\}$, reporting the best validation configuration.

E Additional Ablation Results

Here, we provide experiment settings and results on every dataset that are omitted in Section 5.4.

E.1 Self-function vectors effectively capture prompt-specific concept representations

Here, we present results on every dataset that are omitted in the main section.

We analyzed the relationship between the number of examples in the prompt $[0, 1, \cdots, 10]$ and the similarity between the task-specific head's self-function vector and the original task's function vector. For each prompt configuration, we visualized the cosine similarity distributions using histograms and kernel density estimation (KDE), conditioned on whether the model's prediction (with or without self-function vector intervention) was correct or incorrect. The results indicate that, for the AG News and Emotion datasets, incorrect samples tend to exhibit lower cosine similarity between the self-function vector and the original task function vector. In contrast, datasets such as WNMCQ1, Hellaswag, and GSM8K—which adopt a multiple-choice question format—showed more consistent similarity distributions, likely because their conceptual representations remain stable across examples.

E.2 Single Top-k Intervention vs. Ensemble Variants

We compare our top-k single intervention against ensemble-style variants that aggregate AU estimates over N randomly sampled interventions

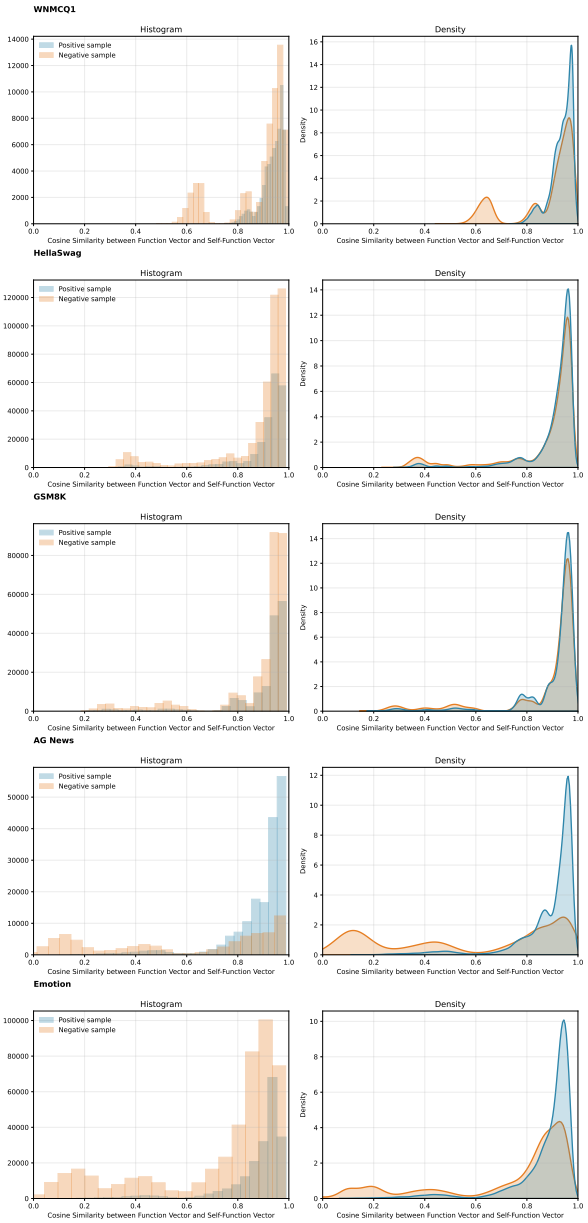


Figure 8: Cosine similarity distributions between self-function vectors and original task function vectors across different numbers of in-context examples $[0, 1, \cdots, 10]$. Incorrect samples in AG News and Emotion exhibit lower similarity, while multiple-choice datasets (WNMCQ1, Hellaswag, GSM8K) show more stable distributions across prompt configurations.

from $\mathcal{S}_T$, in order to empirically justify the $N = 1$ design choice described in Section 3.4.

Concretely, for each prompt we compute the AU estimate using the top-k self-function vector (our default) and compare it against the mean AU estimate obtained from $N \in \{1, 3, 5, 10, 20, 30, 40, 50\}$ randomly sampled interventions. We report the mean absolute difference, its 95% confidence interval, and relative variance across prompts on WNMCQ1 with LLaMA2-7B.

N	Mean Diff	95% CI	Variance	Rel. Variance
1	0.0420	[0.0257, 0.0583]	0.00223	0.00156
3	0.0331	[0.0183, 0.0478]	0.00148	0.00100
5	0.0310	[0.0178, 0.0443]	0.00127	0.00085
10	0.0272	[0.0152, 0.0393]	0.00100	0.00067
20	0.0260	[0.0145, 0.0375]	0.00091	0.00061
30	0.0244	[0.0141, 0.0348]	0.00079	0.00053
40	0.0240	[0.0141, 0.0340]	0.00075	0.00050
50	0.0244	[0.0145, 0.0343]	0.00077	0.00051

Table 6: Mean absolute difference between the top-k single intervention AU estimate and a random ensemble of N members (average AU $= 1.302$). Differences decrease rapidly and plateau after $N \approx 10$, remaining below 2.1% of the average AU.

As shown in Table 6, the mean absolute difference between the top-k estimate and the ensemble converges quickly: it drops from 0.042 at $N = 1$ to roughly 0.024 by $N = 10$ and remains stable thereafter. Relative to the average AU of 1.302, the gap at $N = 1$ is approximately 3.2%, and converges to $\sim$1.8% as $N \to 50$. These marginal differences confirm that the top-k single intervention already captures the dominant effect, with additional ensembling providing limited practical gain relative to the added computational cost.

E.3 Impact of the number of shots on causal head selection

In exploring the identification of causal heads, we varied the number of examples in the prompt to examine how robustly task-specific heads can be discovered and how effectively their activations can be leveraged. Specifically, we varied the number of examples from 4 to 10 and identified causal heads for each task. Our analysis revealed that, regardless of the configuration, there consistently exist task-specific heads that play a crucial role, which further supports the effectiveness of our proposed method.

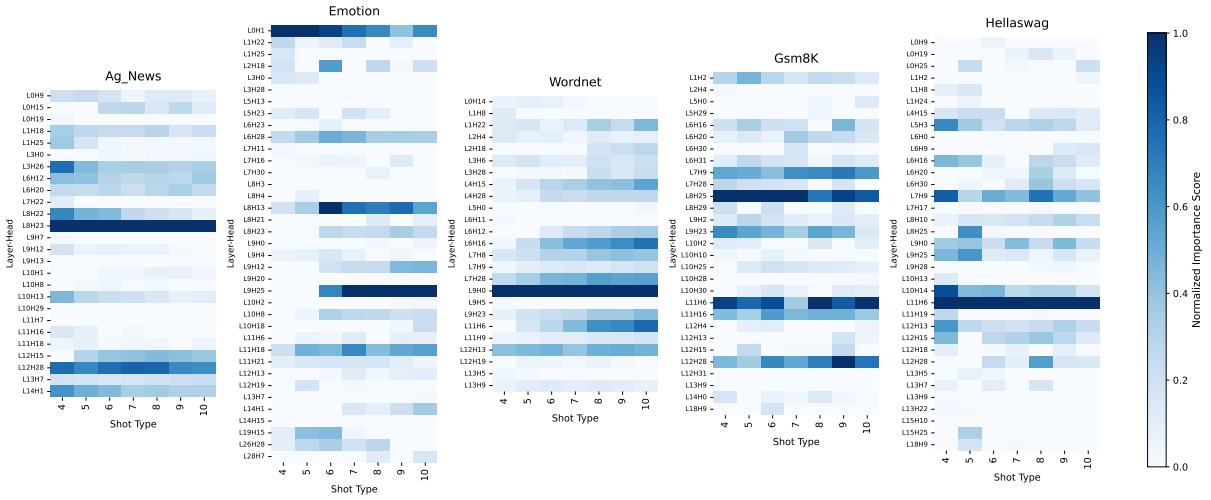


Figure 9: Consistent task-specific causal heads across different shot configurations, suggesting their activations can be effectively leveraged.

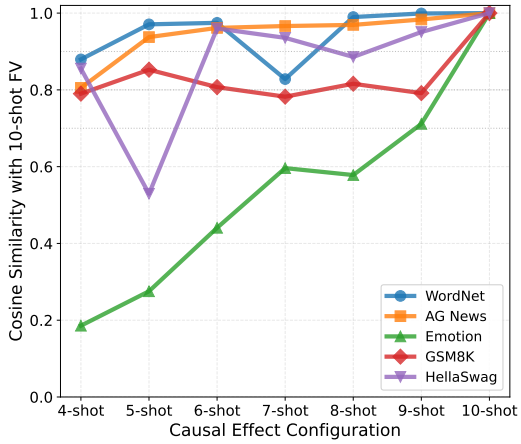


Figure 10: Cosine similarity between function vectors across tasks as the number of in-context examples increases from 4 to 10.

E.4 Function vector converges as number of shots increases

Here, we provide experiment settings and results on every dataset that are omitted in the main section.

To further investigate the stability of function vectors derived under different numbers of in-context examples, we analyzed their behavior in the causal head selection setting described in Section Appendix E.3. For each given task T, we obtained multiple function vectors that represent the task semantics under varying prompt configurations. Specifically, we used the LLaMA-2-7B model and identified causal heads for each task while varying the number of in-context examples from 4 to 10.

We then compared the resulting function vectors across adjacent configurations (e.g., between 4 and 5 examples, 5 and 6 examples, etc.) by computing their cosine similarities. We observed that, as the number of examples increases, the cosine similarity between function vectors derived from neighboring configurations approaches 1, indicating that the function representation of the task converges to a stable direction in representation space. This suggests that, once sufficient contextual information is provided, the model forms a consistent and robust conceptual understanding of the given task, rather than being sensitive to minor variations in the number of examples.

F Additional Baselines and Related Work

F.1 Related Work: RAUQ

Vazhentsev et al. (2025) propose RAUQ, an unsupervised UQ method that leverages attention-head signals within transformer layers to estimate prediction confidence in standard text generation. While RAUQ shares the broad motivation of grounding uncertainty in internal model components, it targets total UQ in text generation rather than aleatoric–epistemic decomposition in ICL, making it complementary to rather than directly comparable with our work.

F.2 MaxProb and Lookback Lens

We extend the baseline comparison to two additional methods: MaxProb (Hendrycks and Gimpel, 2017), a canonical confidence baseline using the maximum softmax probability, and Lookback Lens (Chuang et al., 2024), a recent attention-based hal-

Model	Method	WNMCQ1	Hellaswag	GSM8K	AG News	Emotion
LLaMA2-7B	MaxProb	0.875	**0.624**	0.605	0.837	0.666
	Lookback Lens	0.572	0.530	0.190	0.572	0.562

Table 8: AUROC (↑) for MaxProb and Lookback Lens on hallucination detection (LLaMA2-7B), complementing Table 4. MaxProb is competitive; Lookback Lens underperforms due to incompatibility with single-token generation.

lucination detection method.

Uncertainty Decomposition Protocol Table 7 reports Spearman correlations of MaxProb and Lookback Lens under our AU- and EU-control settings on LLaMA2-7B. Both methods yield small or inconsistent correlations across the AU-control settings and noticeably larger OOD correlations than Self-FV, reflecting that they are designed for overall confidence estimation rather than disentangling AU from EU. These results further motivate specialized decomposition methods for ICL.

Model	Method	Multi-Ans.	WNMCQ1	HellaSwag	GSM8K	AG News	Emotion	OOD
LLaMA2-7B	MaxProb	0.249	0.167	0.022	0.220	0.145	0.097	0.153
	Lookback Lens	-0.970	0.0006	0.0001	0.322	0.188	-0.035	0.406

Table 7: MaxProb and Lookback Lens on the AU- (Multi-Ans., WNMCQ1, HellaSwag, GSM8K, AG News, Emotion) and EU- (OOD) control evaluation protocol for LLaMA2-7B, complementing the main-text results in Tables 1 to 3. AU columns (↑); OOD ($|\rho|\downarrow$).

Hallucination Detection Table 8 reports AUROC for MaxProb and Lookback Lens on LLaMA2-7B, complementing the main hallucination detection results in Table 4. MaxProb is a competitive baseline, while Lookback Lens—which relies on attention maps derived from the generation process—is less compatible with our single-token classification setting and underperforms accordingly.

G PRR Results for Hallucination Detection

Table 9 reports Prediction-Rejection Ratio (PRR) (Fadeeva et al., 2025) for the hallucination detection benchmarks in Section 5, complementing the AUROC results in Table 4. PRR measures the quality of an uncertainty method on the selective generation task: a higher PRR indicates that rejecting predictions with high uncertainty more reliably removes incorrect answers. Mechanistic approaches are generally competitive with or outperform entropy-based baselines across datasets, further supporting the practical utility of self-function vectors for trustworthy generation.

Model	Method	WNMCQ1	Hellaswag	GSM8K	AG News	Emotion
LLaMA2-7B	Total Entropy	0.828	0.242	0.107	0.776	0.364
	Semantic Entropy	0.815	0.191	0.277	0.548	0.283
	UQ_ICL	0.585	0.096	0.159	0.556	0.356
	Function Vector	0.865	**0.291**	0.518	**0.785**	**0.374**
	Self-FV	**0.866**	0.273	**0.571**	0.779	0.368
LLaMA2-13B	Total Entropy	0.841	0.430	0.179	0.758	0.386
	Semantic Entropy	0.859	0.394	0.293	0.536	0.364
	UQ_ICL	0.730	0.229	**0.337**	**0.802**	0.036
	Function Vector	**0.875**	**0.469**	0.261	0.783	**0.410**
	Self-FV	0.862	0.458	0.270	0.777	0.394

Table 9: PRR (↑) on diverse text classification datasets following the LM-Polygraph benchmark (Fadeeva et al., 2025). Mechanistic approaches are generally competitive with or outperform entropy-based baselines.