Simple Yet Effective: An Information-Theoretic Approach to Multi-LLM Uncertainty Quantification

Maya Kruse[1], Majid Afshar[3], Saksham Khatwani[1,2], Anoop Mayampurath[3], Guanhua Chen[3], Yanjun Gao[1]*

[1]University of Colorado Anschutz Medical Campus
[2]University of Colorado Boulder
[3]University of Wisconsin Madison

Abstract

Large language models (LLMs) often behave inconsistently across inputs, indicating uncertainty and motivating the need for its quantification in high-stakes settings. Prior work on calibration and uncertainty quantification often focuses on individual models, overlooking the potential of model diversity. We hypothesize that LLMs make complementary predictions due to differences in training and the Zipfian nature of language, and that aggregating their outputs leads to more reliable uncertainty estimates. To leverage this, we propose MUSE (Multi-LLM Uncertainty via Subset Ensembles), a simple information-theoretic method that uses Jensen-Shannon Divergence to identify and aggregate well-calibrated subsets of LLMs. Experiments on binary prediction tasks demonstrate improved calibration and predictive performance compared to single-model and naïve ensemble baselines. In addition, we explore using MUSE as guided signals with chain-of-thought distillation to fine-tune LLMs for calibration. MUSE is available at:https://github.com/LARK-NLP-Lab/MUSE.

1 Introduction

Although large language models (LLMs) have shown remarkable performance in a wide range of NLP tasks and domains, their output is not always consistent or reliable (Xiao et al., 2022; Zhao et al., 2024b). The same LLM can generate divergent responses under different decoding settings, even with identical inputs (Wang et al., 2024a; Wei et al., 2022). As LLMs enter high-stakes domains like healthcare, quantifying output variance is essential for trust, safety, and decision-making (Gao et al., 2024b; Savage et al., 2025; Qin et al., 2024).

Quantifying uncertainty is essential to address this challenge: Generating responses with appropriately calibrated confidence helps determine when the answer is trustworthy (Geng et al., 2024). Although prior work has explored uncertainty estimation and calibration through sampling and self-consistency (Rivera et al., 2024; Gao et al., 2024a; Ling et al., 2024), uncertainty-aware training (Liu et al., 2024; Chen and Mueller, 2024; Kapoor et al., 2024), reflection (Zhao et al., 2024a; Zhang et al., 2024b), ranking (Huang et al., 2024), and conformal prediction (Wang et al., 2024b), these methods focus on single LLMs.

This paper introduces a novel approach to uncertainty quantification by aggregating predictions from multiple LLMs. Different LLMs generalize better in distinct regions of the input space, due to the Zipfian nature of language and differences in training corpora, objectives, and architectures (Piantadosi, 2014; Chan et al., 2022). Based on this, we *hypothesize that* combining their outputs offers a principled way to reduce uncertainty, improve robustness, and better approximate ground truth in regions where individual models may falter.

Specifically, we formulate the problem through an information-theoretic lens, using Jensen-Shannon Divergence (JSD) to capture the degree of disagreement among models. JSD offers a symmetric and bounded measure of divergence between probability distributions (Cover, 1999), making it well-suited for comparing predictions across multiple LLMs. We quantify model disagreement to identify reliable consensus and propose MUSE (**M**ulti-LLM **U**ncertainty via **S**ubset **E**nsembles), a simple algorithm that selects and aggregates LLM outputs to balance diversity and reliability.

We evaluate MUSE on three publicly available binary prediction datasets. *TruthfulQA* (TQA) covers general domain questions and answers designed to investigate the truthfulness of the model (Lin et al., 2022); both *EHRShot* (Wornow et al., 2023), and *MIMIC-Extract* (MIMIC) are structured clinical datasets derived from records of hospitalized patients in the real world (Johnson et al., 2016; Wang

*Correspondence: yanjun.gao@cuanschutz.edu

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 30482–30493
November 4-9, 2025

et al., 2020). Focusing on binary prediction enables straightforward evaluation of both discrimination and calibration, and empirical findings provide evidence to support our hypothesis with MUSE improving calibration and robustness.

We also ask whether consensus-driven probabilities can *teach probabilistic reasoning* to individual LLMs. Using MUSE outputs as silver-standard supervision for fine-tuning and CoT distillation, we find that ensemble-derived signals are principled but their effectiveness depends on the underlying model, highlighting an important direction for understanding how LLMs internalize probabilistic reasoning.

2 Related Work

In addition to the work cited in §1, Ling et al. (2024) estimates both aleatoric and epistemic uncertainty using entropy within single-LLM in-context learning. Ensemble method are also common: Lakshminarayanan et al. (2017) introduces ensembles with adversarial training for improved calibration; Chen et al. (2025) focuses on clinical prediction tasks, applying deep ensembles and Monte Carlo dropout to capture uncertainty from a single decoder. In the space of multi-LLM ethods, Zhang et al. (2024a) quantifies uncertainty across LLMs via semantic similarity in long-form generation, and Dey et al. (2025) selects LLMs from a pool to reduce hallucinations based on task accuracy. Our work differs in that multi-LLM subsets are selected by minimizing uncertainty with JSD.

3 Methods

3.1 LLM Uncertainty Quantification

We establish two methods for uncertainty quantification for a single LLM as: (1) *self-consistency-based empirical estimation*, which forms the core of our proposed methods, and (2) *sequence likelihood scoring*, used as a widely adopted baseline in prior work (Geng et al., 2024).

(1) Self-Consistency with Empirical Frequency. Given a binary classification input, we perform stochastic decoding runs k in LLM text generation (GEN), using temperature T sampling ($T = 0.7$ and $k = 10$), resulting in a set of outputs $\{\hat{y}_i\}_{i=1}^{k}$. Each output is mapped to a binary label (yes or no). We define the empirical probability of the label yes as: $\hat{p}^{\texttt{yes}} = \frac{1}{k}\sum_{i=1}^{k}\mathbb{I}(\hat{y}_i = \texttt{yes}), \quad \hat{p}^{\texttt{no}} = 1 - \hat{p}^{\texttt{yes}}$.

To estimate uncertainty, we apply a bootstrapping procedure: we resample 90% of the outputs with replacement and recompute $\hat{p}^{\texttt{yes}}$ over $B = 100$ trials (denoting as GEN^{BS}). From the resulting $\hat{p}^{\texttt{yes}}$ distribution, we compute variance, entropy, and JSD for our proposed algorithms (§ 3.2).

(2) Sequence Likelihood (SLL) Scoring. We adopt the SLL approach used in prior LLM calibration work. For each input x, we compute the total log-likelihood of two candidate completions: `"Answer is Yes"` and `"Answer is No"`, denoted $\text{LL}_{\texttt{yes}}$ and $\text{LL}_{\texttt{no}}$. These are computed using left-to-right autoregressive decoding: $\text{LL}_{\texttt{label}} = \sum_{t=1}^{T} \log P(y_t^{\texttt{label}} \mid x, y_{<t}^{\texttt{label}})$. The final prediction probability is obtained via softmax normalization. We use this predicted distribution to compute AUROC, ECE and Brier Scores (Guo et al., 2017). Although not robust to LLM output variability, SLL provides a deterministic scoring baseline for comparison.

Algorithm 1 MUSE-Greedy version

Require: Prediction set $\mathcal{P} = \{p_i\}_{i=1}^{N}$, confidence $c_i = |p_i^{\text{yes}} - 0.5|$, parameters β, ϵ_{tol}, $m_{\min}$

Sort $\mathcal{P}$ by c_i descending; initialize $\mathcal{S} \leftarrow \{p_1\}$, $u_{\text{epis}}^{\text{prev}} \leftarrow 0$
for each p_j in sorted $\mathcal{P} \setminus \mathcal{S}$ **do**
  $\mathcal{S}' \leftarrow \mathcal{S} \cup \{p_j\}$, $\bar{p} \leftarrow \text{mean}(\mathcal{S}')$
  $u_{\text{epis}} \leftarrow \frac{1}{|\mathcal{S}'|}\sum_{p\in\mathcal{S}'} \text{JS}(p \parallel \bar{p})^2$
  $u_{\text{alea}} \leftarrow \frac{1}{|\mathcal{S}'|}\sum_{p\in\mathcal{S}'} H(p)$
  if $|\mathcal{S}'| \geq m_{\min}$ and $u_{\text{epis}} - u_{\text{epis}}^{\text{prev}} > \epsilon_{\text{tol}}$ **then**
    break
  end if
  $\mathcal{S} \leftarrow \mathcal{S}'$, $u_{\text{epis}}^{\text{prev}} \leftarrow u_{\text{epis}}$
end for
$\hat{p}^{\text{yes}} \leftarrow \text{mean}_{p\in\mathcal{S}}(p^{\text{yes}})$, $u_{\text{total}} \leftarrow u_{\text{epis}}^{\text{prev}} + \beta \cdot u_{\text{alea}}$
return $(\hat{p}^{\text{yes}}, u_{\text{total}}, \mathcal{S})$

3.2 Multi-LLM selective algorithm

The central idea of MUSE is that *disagreement among LLM predictive distributions signals epistemic uncertainty, while consensus indicates more reliable generalization.* This can be measured by JSD (Cover, 1999). Meanwhile, we compute the mean entropy H of individual model predictions to reflect *aleatoric uncertainty*, capturing inherent input ambiguity. Our algorithm identifies model subsets $\mathcal{S}$ with low disagreement and uncertainty, surfacing high-consensus regions while balancing complementary signals ("diversity") against noise that degrades calibration and accuracy.

Problem Setup. Given an input x, let $\mathcal{P}_x = \{\mathbf{p}_i = (p_i, 1-p_i)\}_{i=1}^{N}$ be N predictive distributions from multiple LLMs and/or decoding runs, where p_i denotes the predicted probability of the label yes. Our goal is to select a subset $\mathcal{S}_x \subseteq \mathcal{P}_x$ that yields a

well-calibrated, aggregated prediction $\hat{p}$.

Uncertainty Computation. The two types of uncertainty play an important role in the proposed algorithm. Epistemic uncertainty $\mathcal{U}_{\text{epis}}(\mathcal{S})$ reflects inter-model disagreement and is quantified as the average JSD between each prediction $\mathbf{p}_i$ and the subset mean $\bar{\mathbf{p}}$:

$$\mathcal{U}_{\text{epis}}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} \text{JS}(\mathbf{p}_i \| \bar{\mathbf{p}})$$

Aleatoric uncertainty $\mathcal{U}_{\text{alea}}(\mathcal{S})$ reflects intrinsic noise and is estimated by the average binary entropy:

$$\mathcal{U}_{\text{alea}}(\mathcal{S}) = \frac{1}{|\mathcal{S}|} \sum_{i \in \mathcal{S}} H(p_i),$$

where $H(p) = -p \log p - (1-p) \log(1-p)$.

We focus on *optimizing epistemic uncertainty*, as aleatoric uncertainty stems from inherent data noise and is not reducible via model selection. The total uncertainty of a subset of LLMs, denoted as $\mathcal{S}$, is defined as the sum of its epistemic and aleatoric components: $U(\mathcal{S}) = \mathcal{U}_{\text{epis}}(\mathcal{S}) + \beta \cdot \mathcal{U}_{\text{alea}}(\mathcal{S})$, where β is a weighting factor that controls the trade-off between epistemic disagreement and inherent input ambiguity. Results using total uncertainty $U(\mathcal{S})$ are reported in Appendix A.4.

Multi-LLM Uncertainty via Subset Ensemble. MUSE constructs a well-calibrated set of LLM outputs based on $\mathcal{U}_{\text{epis}}(\mathcal{S})$. It supports two subset selection strategies, *greedy* and *conservative*, which incrementally select a subset of LLMs whose outputs are mutually diverse yet coherent, as determined by pairwise JSD. Two key parameters control the behavior of MUSE: the noise threshold (ϵ_{tol}) and the minimum subset size m_{min} as diversity constraint, controlling the balance between diversity and agreement. Each prediction comes with a confidence score, defined as $c_i = |p_{\text{yes}}^{(i)} - 0.5|$, since 0.5 represents maximum uncertainty in a binary decision. The **greedy** version starts with the most confident LLM prediction and iteratively adds models that increase the overall $\mathcal{U}_{\text{epis}}(\mathcal{S})$ (diversity) of the subset, up to a specified tolerance (as in Algo 1.). The **conservative** version, in contrast, selects models that minimize a joint objective combining epistemic and aleatoric uncertainty. This approach encourages diversity while avoiding instability, resulting in a more calibrated and robust ensemble (see Algo.2).

Once a subset is selected, we compute the final predicted probability by averaging the individual LLM predictions within the subset. Two aggregation strategies are deployed: (1) a simple unweighted mean, and (2) an *aleatoric-aware weighting*, where each LLM's prediction is $\hat{p}^{yes}$ weighted by its $\mathcal{U}_{\text{alea}}(\mathcal{S})$, where each $\hat{p}_i^{yes}$ is weighted by its entropy, i.e., $1 - H(\hat{p}_i^{yes})$. The final prediction is computed as a weighted average, assigning higher weights to more confident (low-entropy) predictions, emphasizing more decisive predictions, particularly when individual models exhibit varying uncertainty levels.

	LLM	Method	AUROC	ECE	Brier
SINGLE	Mistral-7B	SLL	64.99	58.11	55.57
		GENBS	29.48	52.25	54.92
	Qwen2-7B	SLL	58.47	65.54	65.23
		GENBS	60.78	47.71	49.5
	Gemma-7B	SLL	60.65	37.41	36.78
		GENBS	52.14	48.4	55.88
	DS-Qwen-32B	SLL	72.89	57.3	54.16
		GENBS	63.11	18.83	30.32
MUSE Greedy	All LLMs	mean	57.57	38.59	40.68
		weighted	59.28	41.53	43.39
	Excl. Outlier	mean	69.54	40.29	40.45
		weighted	68.93	41.11	41.6
	DS+Qwen2	mean	69.98	40.27	41.3
		weighted	69.86	41.23	42.21
MUSE Conserv.	All LLMs	mean	51.04	40.06	42.1
		weighted	54.45	43.51	45.62
	Excl. Outlier	mean	67.57	39.01	38.48
		weighted	67.3	40.49	39.99
	DS+Qwen2 (Top 2)	mean	†72.33	†38.15	†38.55
		weighted	72.35	39.45	40.25

Table 1: Performance on TruthfulQA. We report SLL and GEN$^{\text{BS}}$ results alongside all MUSE settings (greedy/conservative, with or without aleatoric weighting). † highlight cases where MUSE yields competitive AUROC with lower calibration error despite not being the top performer.

LLMs	AUROC ↑	ECE ↓	Brier Score ↓
Qwen2	53.52	22.22	32.60
Mistral	60.10	13.70	25.50
Gemma	50.20	**9.80**	32.30
DS-Qwen	62.00	10.50	30.00
DS+Qwen2	61.78	17.50	28.89
DS+Mistr	**64.73**	16.57	25.00
DS+Mistr+Qwen2	62.24	11.74	**24.38**
All	60.94	12.76	24.38

Table 2: Performance on the EHRShot Acute myocardial infarction (*acute_mi*) task across single LLMs (GEN$^{\text{BS}}$)and multi-LLM combinations. Greedy and conservative (non-weighted) results are identical. DS = DS-Qwen.

3.3 Uncertainty-Aware Supervised Fine-Tuning

We test whether MUSE-derived probabilities improve model accuracy and calibration through supervised fine-tuning (SFT). We evaluate SFT variants that differ in how probabilities are injected into prompts and whether explicit reasoning is included.

Direct SFT. We inject probabilistic signals via two prompt formats: (i) **Default**: include the MUSE consensus $\hat{p}$ with the patient input and gold label; (ii) **RawProb**: replace $\hat{p}$ with bootstrapped per-model probabilities (e.g., Mistral: [0.62, 0.64,

LLM	Method	LOS3			LOS7			Mort Hosp.		
		AUROC	ECE	Brier	AUROC	ECE	Brier	AUROC	ECE	Brier
DS-Qwen	SLL	41.30	37.70	41.05	40.60	3.80	6.81	46.90	5.20	14.77
	GEN^{BS}	57.80	13.90	31.93	59.10	30.23	26.82	55.52	3.57	11.76
Qwen2-7B	SLL	56.40	36.50	35.96	58.45	55.66	69.60	68.01	47.70	47.38
	GEN^{BS}	56.80	31.10	34.18	58.45	45.80	55.66	59.29	3.30	10.17
Naive	Majo.	54.04	**2.41**	38.64	57.42	**7.15**	**9.18**	58.73	5.76	11.58
	Mean	53.87	6.98	40.99	58.40	3.34	7.76	59.48	2.28	10.15
MUSE	Greedy	**61.47**	12.43	**27.51**	60.47	35.01	26.04	59.55	2.38	10.46
	+Weighted	61.40	18.83	29.71	61.29	34.54	28.22	59.64	2.29	10.44
	Conserv.	60.06	21.43	30.97	61.58	35.13	29.86	59.84	2.61	10.49
	+Weighted	61.04	24.03	32.49	**61.71**	35.42	31.39	59.83	2.76	10.44

Table 3: AUROC, ECE, and Brier Score across clinical prediction tasks. We include DS-Qwen and Qwen-7B because they are the two best-performing single LLMs on this dataset (see more results in Table 6). We compare individual LLMs, simple fusion baselines (majority voting "Majo." and mean), and algorithmic subset selection strategies (greedy and conservative).

0.60]; Qwen: [0.58, 0.55, 0.57]; DS-Qwen: [0.55, 0.53, 0.56]).

Chain-of-thought distillation. We optionally append teacher-generated reasoning using three variants: (i) **Original** (assess whether $\hat{p}$ is reasonable; Table 10), (ii) **Bayesian** (explicit Bayes framing), and (iii) **No** $\hat{p}$ (reason without $\hat{p}$). Each CoT variant is paired with both SFT formats.

4 Experimental Setup

We use *TruthfulQA*(Lin et al., 2022), a benchmark of adversarially designed questions with labeled truthful and untruthful answers. The task is to classify each candidate answer as truthful (`Yes`) or not (`No`), enabling direct evaluation of both discrimination (AUROC) and calibration (ECE, Brier Score). We also test our method using two structured EHR datasets: diagnosis prediction from *EHRShot* (Wornow et al., 2023) and *MIMIC-Extract* (Wang et al., 2020). On MIMIC, the LLMs predict three tasks: hospital length of stay ≥ 3 days (LOS3), ≥ 7 days (LOS7), and in-hospital mortality (Mort Hosp.).

We evaluate the following open-source models: Mistral-7B-Instruct (Jiang et al., 2023), Gemma-7B-it (Team et al., 2023), Qwen2-7B-instruct (Yang et al., 2024), and the latest Deepseek-R1-Distil-Qwen-32B (DS-Qwen) (DeepSeek-AI, 2025). All LLMs are run on a server with 4×A100 40GB GPUs. For DS-Qwen, we apply 8-bit quantization to reduce inference time and memory usage. In addition to single LLM method baseline, we compose two naive multi-LLM baselines: majority voting (counting positive labels), and mean of all LLMs' $\hat{p}^{yes}$ as the final positive probability. On the Uncertainty-aware SFT experiments, we investigate on two models, Mistral-7B-Instruct and Qwen2-7B-instruct, using the MIMIC-Extract dataset for the length-of-stay 3 days (LOS3) prediction task (y=1 if LOS ≥ 3 days, 0 otherwise). For CoT distillation, we leverage a HIPAA-compliant Microsoft Azure GPT-o3-mini as the teacher model, and generate CoT reasonings for the entire LOS-3 training set for each setting. This Azure GPT instances provides high-quality rationales while remaining compliant with the MIMIC-III data use agreement.

MUSE	Task	Total %	GEM %	Qwen %	DS-Qwen %	Mistral %
Conserv	LOS 3	31.6	56.97	35.67	25.01	17.42
Greedy	LOS 3	90.58	27.84	26.76	25.77	27.01
Conserv	TQA	33.33	64.44	19.79	7.89	28.90
Greedy	TQA	76.18	33.46	26.89	17.17	31.95

Method	Total%	GEM%	Qwen%	DS-Qwen%
Greedy	73.13	47.95	39.90	24.71
Conserv.	53.53	69.16	30.69	13.03

Table 4: Breakdown of LLM inclusion in MUSE. Top: frequency of each model's selection under Conservative and Greedy variants on LOS 3 and TruthfulQA (TQA). Bottom: inclusion frequencies when Mistral is removed from the model pool. (GEM: Gemma-7B, Qwen: Qwen2, DS-Qwen: Deepseek-distilled-Qwen32B, Mistral: Mistral-7B-Instruct).

5 Results and Discussion

We organize results around two questions. **Q1: Is MUSE effective for uncertainty estimation?** We evaluate calibration and robustness on general and clinical datasets. **Q2: Can MUSE provide silver-standard supervision for probabilistic reasoning?** We test whether consensus-derived probabilities aid SFT and CoT distillation. Findings show MUSE improves calibration and robustness (Q1), while supervision results are mixed: some models benefit, others do not. Overall, multi-LLM aggregation is promising for uncertainty estimation but presents challenges to improve single-model reasoning.

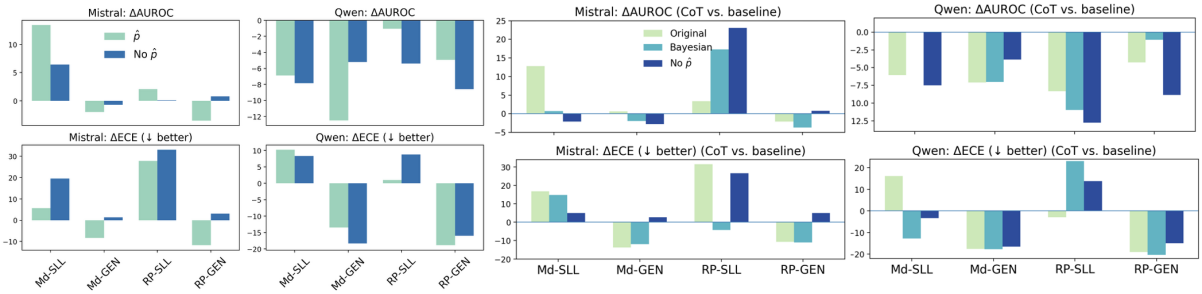

Figure 1: Comparative results of supervised fine-tuning with MUSE-derived probabilities. Common settings at both panels: Md indicates the default consensus probability input, while RP uses the raw bootstrapped probabilities from the model pool. Left (4 panels): Direct SFT performance shown as changes in AUROC and ECE (bottom row) for Mistral and Qwen, when using model SLL and GEN output under settings with and without $\hat{p}$. Improvements are measured relative to no-SFT baselines (positive ΔAUROC, negative ΔECE indicate gains). Right (4 panels): CoT SFT performance under the same models, comparing three prompting strategies (Original, Bayesian, No $\hat{p}$). Results demonstrate that while direct SFT yields modest and model-dependent improvements, CoT-based SFT produces more variable outcomes across prompting strategies.

5.1 MUSE Effectiveness

MUSE improves both AUROC and calibration metrics compared to single LLMs and naive ensembling baselines, demonstrating its effectiveness in producing reliable and well-calibrated predictions through selective multi-model aggregation. Although the SLL method occasionally yields the highest AUROC, such as DS-Qwen achieving 72.89 on *TruthfulQA*, it often suffers from poor calibration (ECE 57.30, Brier 54.16). In contrast, MUSE offers more balanced predictions, with comparable AUROC (72.35) and substantially lower calibration error (ECE 38.15). Similar gains are observed on the EHRShot *acute_mi* task, where DS+Mistral+Qwen achieves the best Brier Score (24.4) and strong AUROC (62.2), improving over all single-LLM baselines.

We address two-subquestions that illustrate the practical significance of the MUSE algorithm.

(1) Does MUSE blindly include weaker models? Table 4 reports selection frequencies under MUSE (Conservative, Greedy). Mistral is chosen less often in Conserv. than Greedy, only when it improves subset consistency. Even strong models (e.g., Qwen) are not always selected, indicating instance-specific, uncertainty-driven choice rather than global accuracy. When a weaker model adds noise, MUSE adapts. Removing Mistral shifts weight to GEM and Qwen, confirming adaptive selection rather than identity-based filtering.

(2) Are gains just from excluding a bad model? Ablations over fixed subsets (Tables 1, 2) and exhaustive pairwise/three-way combinations (Table 8) show that while naive ensembles (e.g., DS+Qwen2) can perform well, MUSE often matches or exceeds them, even when weaker models remain, confirming that it performs input-level rather than fixed global selection. Additional analysis of per-model divergence from the consensus (Table 9) shows small JSD values (< 0.1), supporting our claim that ensemble diversity with selective aggregation underlies MUSE's effectiveness.

5.2 MUSE-Guided Supervised Fine-Tuning

Figure 1 summarizes the impact of incorporating MUSE-derived probabilities into supervised fine-tuning. Direct SFT yields modest, model-dependent effects. For Mistral, $\hat{p}$ sometimes improves discrimination but often worsens calibration, while GEN consistently lowers ECE. Qwen shows the opposite: calibration improves, but AUROC gains are limited or negative. Adding CoT distillation introduces wider variation. Mistral sees large AUROC increases with RawProb, though often at the cost of calibration. Qwen benefits most from Bayesian CoT, with clear calibration gains in GEN. The results highlight that MUSE-derived uncertainty can provide useful supervision, but effects depend on model, supervision style, and whether signals are contextualized through reasoning.

6 Conclusion

We present MUSE, a multi-LLM framework for uncertainty estimation that aggregates predictive distributions into calibrated, uncertainty-aware outputs. Beyond improving accuracy and calibration, we explored using MUSE as a supervisory signal for fine-tuning single LLMs, with mixed but promising results, indicating a direction to explore further.

Acknowledgments

This work is supported by U.S. National Library of Medicine R00 LM014308.

Limitation

Our study evaluates a limited set of open-source LLMs and focuses exclusively on binary prediction tasks, where evaluation of discrimination and calibration is most straightforward. We also assume access to all model outputs during inference, which may not reflect real-time or resource-constrained deployment scenarios. However, our focus is not on maximizing efficiency, but on understanding how model composition and selective aggregation affect uncertainty estimation. Nonetheless, we have provided empirical evidence that MUSE consistently improves both accuracy and calibration, highlighting the value of principled multi-model fusion. Future work will extend to more complex prediction settings and explore efficient selection strategies across broader model ecosystems.

Ethical Consideration

This study uses two publicly available, de-identified clinical datasets (MIMIC-Extract and EHRShot), ensuring no personally identifiable information is accessed or exposed. All models used are open-source LLMs, and no fine-tuning or data logging was performed, eliminating the risk of patient data leakage. While our focus is on evaluating uncertainty and not clinical deployment, we emphasize the need for responsible use of LLMs in sensitive domains. Any generative outputs or predictions from these models should be interpreted with caution, especially in clinical contexts, and subject to domain expert validation prior to real-world application.

References

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. *Advances in neural information processing systems*, 35:18878–18891.

Jiuhai Chen and Jonas Mueller. 2024. Quantifying uncertainty in answers from any language model and enhancing their trustworthiness. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5186–5200.

Zizhang Chen, Peizhao Li, Xiaomeng Dong, and Pengyu Hong. 2025. Uncertainty quantification for clinical outcome predictions with (large) language models. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 7512–7523, Albuquerque, New Mexico. Association for Computational Linguistics.

Thomas M Cover. 1999. *Elements of information theory*. John Wiley & Sons.

DeepSeek-AI. 2025. Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. *Preprint*, arXiv:2501.12948.

Prasenjit Dey, Srujana Merugu, and Sivaramakrishnan Kaveri. 2025. Uncertainty-aware fusion: An ensemble framework for mitigating hallucinations in large language models. *arXiv preprint arXiv:2503.05757*.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024a. Spuq: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346.

Yanjun Gao, Skatje Myers, Shan Chen, Dmitriy Dligach, Timothy A Miller, Danielle Bitterman, Guanhua Chen, Anoop Mayampurath, Matthew Churpek, and Majid Afshar. 2024b. Position paper on diagnostic uncertainty estimation from large language models: Next-word probability is not pre-test probability. In *GenAI for Health: Potential, Trust and Policy Compliance*.

Jiahui Geng, Fengyu Cai, Yuxia Wang, Heinz Koeppl, Preslav Nakov, and Iryna Gurevych. 2024. A survey of confidence estimation and calibration in large language models. In *NAACL-HLT*.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q Weinberger. 2017. On calibration of modern neural networks. In *International conference on machine learning*, pages 1321–1330. PMLR.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 284–312.

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.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Arka Pal, Samuel Dooley, Micah Goldblum, and Andrew Wilson. 2024. Calibration-tuning: Teaching large language models to know what they don't know. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 1–14.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. 2017. Simple and scalable predictive uncertainty estimation using deep ensembles. *Advances in neural information processing systems*, 30.

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252.

Chen Ling, Xujiang Zhao, Xuchao Zhang, Wei Cheng, Yanchi Liu, Yiyou Sun, Mika Oishi, Takao Osaki, Katsushi Matsuda, Jie Ji, and 1 others. 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.

Shudong Liu, Zhaocong Li, Xuebo Liu, Runzhe Zhan, Derek Wong, Lidia Chao, and Min Zhang. 2024. Can llms learn uncertainty on their own? expressing uncertainty effectively in a self-training manner. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 21635–21645.

Steven T Piantadosi. 2014. Zipf's word frequency law in natural language: A critical review and future directions. *Psychonomic bulletin & review*, 21:1112–1130.

Jeremy Qin, Bang Liu, and Quoc Nguyen. 2024. Enhancing healthcare llm trust with atypical presentations recalibration. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2520–2537.

Mauricio Rivera, Jean-François Godbout, Reihaneh Rabbany, and Kellin Pelrine. 2024. Combining confidence elicitation and sample-based methods for uncertainty quantification in misinformation mitigation. In *Proceedings of the 1st Workshop on Uncertainty-Aware NLP (UncertaiNLP 2024)*, pages 114–126.

Thomas Savage, John Wang, Robert Gallo, Abdessalem Boukil, Vishwesh Patel, Seyed Amir Ahmad Safavi-Naini, Ali Soroush, and Jonathan H Chen. 2025. Large language model uncertainty proxies: discrimination and calibration for medical diagnosis and treatment. *Journal of the American Medical Informatics Association*, 32(1):139–149.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, and 1 others. 2023. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*.

Shirly Wang, Matthew BA McDermott, Geeticka Chauhan, Marzyeh Ghassemi, Michael C Hughes, and Tristan Naumann. 2020. Mimic-extract: A data extraction, preprocessing, and representation pipeline for mimic-iii. In *Proceedings of the ACM conference on health, inference, and learning*, pages 222–235.

Xinglin Wang, Yiwei Li, Shaoxiong Feng, Peiwen Yuan, Boyuan Pan, Heda Wang, Yao Hu, and Kan Li. 2024a. Integrate the essence and eliminate the dross: Fine-grained self-consistency for free-form language generation. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11782–11794.

Zhiyuan Wang, Jinhao Duan, Lu Cheng, Yue Zhang, Qingni Wang, Xiaoshuang Shi, Kaidi Xu, Heng Tao Shen, and Xiaofeng Zhu. 2024b. Conu: Conformal uncertainty in large language models with correctness coverage guarantees. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 6886–6898.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, and 1 others. 2022. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Michael Wornow, Rahul Thapa, Ethan Steinberg, Jason Fries, and Nigam Shah. 2023. Ehrshot: An ehr benchmark for few-shot evaluation of foundation models. *Advances in Neural Information Processing Systems*, 36:67125–67137.

Yuxin Xiao, Paul Pu Liang, Umang Bhatt, Willie Neiswanger, Ruslan Salakhutdinov, and Louis-Philippe Morency. 2022. Uncertainty quantification with pre-trained language models: A large-scale empirical analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 7273–7284.

An Yang, Beichen Zhang, Binyuan Hui, Bofei Gao, Bowen Yu, Chengpeng Li, Dayiheng Liu, Jianhong Tu, Jingren Zhou, Junyang Lin, and 1 others. 2024. Qwen2. 5-math technical report: Toward mathematical expert model via self-improvement. *arXiv preprint arXiv:2409.12122*.

Caiqi Zhang, Fangyu Liu, Marco Basaldella, and Nigel Collier. 2024a. Luq: Long-text uncertainty quantification for llms. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 5244–5262.

Wenqi Zhang, Yongliang Shen, Linjuan Wu, Qiuying Peng, Jun Wang, Yueting Zhuang, and Weiming Lu. 2024b. Self-contrast: Better reflection through inconsistent solving perspectives. In *Proceedings*

Algorithm 2 MUSE-Conservative version

Require: Prediction set $\mathcal{P} = \{p_i\}_{i=1}^{N}$, confidence $c_i = |p_i^{\text{yes}} - 0.5|$, parameters $\beta, \tau, m_{\min}$

Sort $\mathcal{P}$ by c_i descending; initialize $\mathcal{S} \leftarrow \{p_1\}$, $u_{\text{total}}^{\text{prev}} \leftarrow \infty$
for each p_j in sorted $\mathcal{P} \setminus \mathcal{S}$ **do**
  $\mathcal{S}' \leftarrow \mathcal{S} \cup \{p_j\}$, $\bar{p} \leftarrow \text{mean}(\mathcal{S}')$
  $u_{\text{epis}} \leftarrow \frac{1}{|\mathcal{S}'|} \sum_{p \in \mathcal{S}'} \text{JS}(p \parallel \bar{p})^2$
  $u_{\text{alea}} \leftarrow \frac{1}{|\mathcal{S}'|} \sum_{p \in \mathcal{S}'} H(p)$
  $u_{\text{total}} \leftarrow u_{\text{epis}} + \beta \cdot u_{\text{alea}}$
  if $|\mathcal{S}'| \geq m_{\min}$ and $u_{\text{total}} > u_{\text{total}}^{\text{prev}} - \tau$ **then**
    break
  end if
  $\mathcal{S} \leftarrow \mathcal{S}'$, $u_{\text{total}}^{\text{prev}} \leftarrow u_{\text{total}}$
end for
$\hat{p}^{\text{yes}} \leftarrow \text{mean}_{p \in \mathcal{S}}(p^{\text{yes}})$
return $(\hat{p}^{\text{yes}}, u_{\text{total}}, \mathcal{S})$

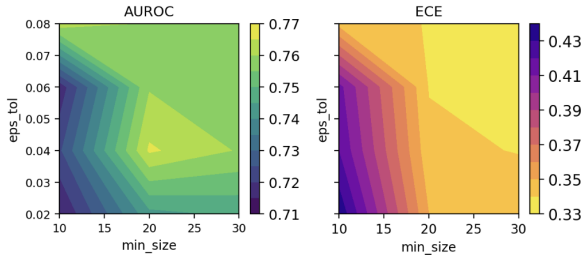

Figure 2: Contour plot of AUROC and ECE as MUSE parameters (m_{size}, ϵ_{tol}) vary, based on a TQA dev set. Main results use m_{size}=20, ϵ_{tol}=0.04. See Appendix for further analysis.

of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 3602–3622, Bangkok, Thailand. Association for Computational Linguistics.

Xinran Zhao, Hongming Zhang, Xiaoman Pan, Wenlin Yao, Dong Yu, Tongshuang Wu, and Jianshu Chen. 2024a. Fact-and-reflection (far) improves confidence calibration of large language models. In *Findings of the Association for Computational Linguistics ACL 2024*, pages 8702–8718.

Yukun Zhao, Lingyong Yan, Weiwei Sun, Guoliang Xing, Shuaiqiang Wang, Chong Meng, Zhicong Cheng, Zhaochun Ren, and Dawei Yin. 2024b. Improving the robustness of large language models via consistency alignment. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 8931–8941.

A More Analysis and Results.

A.1 MUSE Conservative Algorithm

Algorithm 2 presents the MUSE conservative version. We consider total uncertainty as the sum of epistemic and aleatoric components to better balance diversity and reliability in model selection. The conservative version of MUSE adopts a cautious strategy by only adding models when their inclusion leads to a meaningful reduction in total uncertainty. This prevents noisy or unstable predictions from being included, resulting in a more stable and selective ensemble that emphasizes trustworthy aggregation rather than maximizing diversity alone.

Metric	P(Yes)	$U(\mathcal{S})$	Best
LOS 3			
AUROC	**0.5804**	0.5103	P(Yes)
ECE	**0.2015**	0.2208	P(Yes)
Brier Score	0.3321	**0.3168**	$U(\mathcal{S})$
LOS 7			
AUROC	**0.5748**	0.5005	P(Yes)
ECE	**0.1998**	0.2135	P(Yes)
Brier Score	0.3514	**0.1519**	$U(\mathcal{S})$
Mortality			
AUROC	**0.5598**	0.5263	P(Yes)
ECE	**0.0283**	0.1140	P(Yes)
Brier Score	**0.1073**	0.1085	P(Yes)

Table 5: Comparison of predicted probability $p(\texttt{Yes})$ vs. total uncertainty (epistemic + aleatoric) as scoring signals. AUROC favors $p(\texttt{Yes})$, while Brier Score occasionally improves with uncertainty-based scoring.

A.2 Balancing diversity and noise.

A key strength of our approach is its ability to balance diversity with reliability in multi-LLM ensembles. Figure 2 shows that increasing the minimum subset size (m_{size}) and moderately relaxing the epistemic uncertainty threshold (ϵ_{tol}) consistently improves both AUROC and ECE. A larger m_{size} (≥ 20) promotes diversity by including more models, while a moderate ϵ_{tol} ([0.04, 0.08]) allows controlled disagreement without overwhelming the ensemble with noise. The best performance is achieved when both parameters are carefully balanced. This supports our hypothesis that LLMs offer complementary strengths and provides empirical evidence for our subset-based uncertainty aggregation framework.

A.3 Adaptive model selection.

Our method further demonstrates adaptive behavior: performance improves when "stronger" LLMs are present but degrades when weak or noisy models dominate the candidate pool, where strength is defined by each model's single-task performance. This effect is most evident in *acute_mi* and *TQA*. In EHRShot, DS and Mistral form a strong ensemble, but adding a weaker model like Gemma introduces noise that contaminates the pool and harms performance. This supports our hypothesis that selective

LLM	Method	LOS3			LOS7			Mort Hosp.		
		AUROC	ECE	Brier	AUROC	ECE	Brier	AUROC	ECE	Brier
Mistral-7B	SLL	41.41	11.34	26.21	46.08	44.59	27.02	44.15	28.78	18.81
	GEN^{BS}	51.28	15.63	28.37	55.49	38.96	25.38	52.70	14.59	14.07
Gemma-7B	SLL	46.48	24.71	33.85	51.37	16.04	13.44	58.89	10.52	10.48
	GEN^{BS}	55.09	36.91	42.78	53.93	38.29	36.70	55.94	10.40	15.45
MUSE	Greedy	**60.70**	13.34	26.42	60.70	36.43	24.30	62.25	4.36	9.59
	weighted	60.24	18.38	28.34	**61.02**	35.14	26.07	**62.88**	4.02	**9.51**
	Conserv.	58.34	30.87	36.17	58.74	36.29	32.15	61.89	**3.88**	10.90

Table 6: More AUROC, ECE, and Brier Score across clinical prediction tasks (MIMIC-Extract) for Mistral-7B and Gemma-7B. In this table, we also report the MUSE results from Mistral, Gemma, DS-Qwen and Qwen.

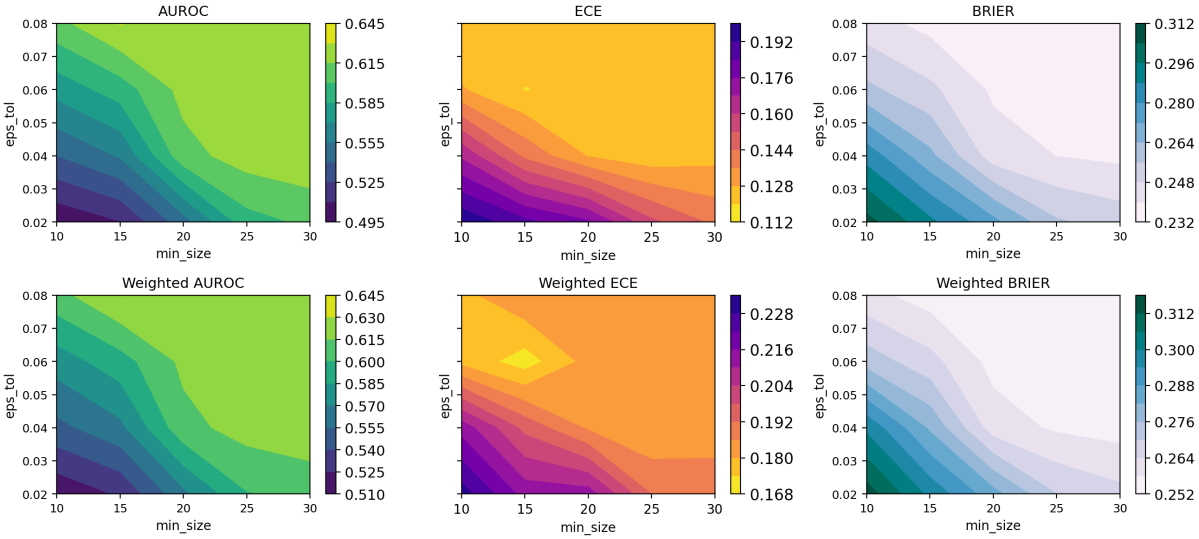

Figure 3: Contour plot for parameter sensitivity analysis using *lupus* prediction task from EHRShot. We report MUSE-Greedy with both weighted and unweighted version, to showcase the differences.

LLMs	AUROC ↑	ECE ↓	Brier Score ↓
Hyperlipidemia (weak models dominate)			
Qwen	46.91	38.92	50.88
Mistral	52.53	14.47	25.69
Gemma	45.18	25.72	35.36
Deepseek-Distill	43.92	22.46	40.97
Greedy (mean)	33.16	24.51	29.13
(weighted)	33.17	27.24	31.16
Lupus (encountered strong, stronger)			
Qwen	45.36	53.09	47.03
Mistral	33.89	19.38	11.39
Gemma	51.60	13.97	11.54
Deepseek-Distill	50.94	3.70	6.84
Greedy (mean)	50.56	10.94	22.48
(weighted)	52.45	11.82	23.09

Table 7: Performance comparison across two EHRShot tasks. In *hyperlipidemia*, where weak models dominate, the multi-LLM algorithm underperforms. In *lupus*, encountering stronger base models allows the algorithm to adapt and perform competitively, reflecting the adaptive behavior pattern.

LLMs Combination	AUROC Naive	AUROC MUSE
Qwen + Mistral	56.78	55.17
Mistral + Gemma	55.14	55.15
Gemma + Qwen	51.87	54.20
Mistral + Gemma + DS	57.41	65.81
Mistral + Gemma + Qwen	54.60	56.47
Qwen + Gemma + DS	55.23	61.05

Table 8: Performance on the EHRShot acute MI task across different multi-LLM combinations. **AUROC Avg. Score** column is a simple average of the bootstrapped score of individual LLMs. **AUROC Algo. Score** signifies the (non-weighted) AUROC score of the combination of LLMs using Greedy approach.

aggregation for trustable consensus is the key to reliable ensemble performance. MUSE makes no assumption that more models lead to better results; rather, it selectively aggregates those that contribute meaningful, calibrated signals to reduce total uncertainty. Table 7 includes a pair of contrasting cases from two prediction tasks in EHRShot for readers who are interested in (*"weak models dominate"* vs. *"encountered strong, stronger"*) behavior.

Model on LOS3	JSD (Conserv)	JSD (Greedy)
GEM	0.0444	0.0709
Qwen2	0.0971	0.0532
DS-Qwen	0.0640	0.06512
Mistral	0.0948	0.05126

Table 9: Comparison of JSD between individual LLM and final MUSE predicted probability.

Base CoT Prompt:
Generate a short chain-of-thought (CoT) reasoning paragraph (maximum 200 words) that explains the significance of the given `p_hat` value in the context of hospital length of stay (LOS) prediction.

Bayesian CoT Prompt:
Generate a short chain-of-thought (CoT) reasoning paragraph (maximum 200 words) that explains the significance of the given `p_hat` value in the context of hospital length of stay (LOS) prediction. Use Bayesian reasoning to justify if `p_hat` is a reasonable estimate. If `p_hat` aligns with the true label, explain why it succeeds. If it does not align, explain why it fails.

Additional Information (for both prompts):
– The `input` field contains clinical data from the MIMIC dataset.
– The `p_hat` value is the predicted probability of the patient staying three or more days.
– The `y_true` is the ground-truth label (0 = LOS $<$ 3 days, 1 = LOS $\geq$ 3 days).

Table 10: Base and Bayesian chain-of-thought prompts used for supervised fine-tuning.

A.4 Comparison of P(Yes) vs. $U(\mathcal{S})$

To compare different scoring strategies, we evaluate the predicted probability $p(\texttt{Yes})$ and the total uncertainty (the sum of epistemic and aleatoric components) as predictors of label correctness. As shown in Table 5, $p(\texttt{Yes})$ consistently achieves higher AUROC and lower ECE across all tasks, indicating better discrimination and calibration. However, total uncertainty yields lower Brier scores in some cases (e.g., LOS7), suggesting it may better reflect the overall confidence–error trade-off in noisier settings. These results indicate that while $p(\texttt{Yes})$ is a strong default for classification, total uncertainty can serve as a complementary signal for soft calibration or abstention.

A.5 More results on MIMIC-Extract

Table 6 presents the performance of two single LLMs and the MUSE multi-LLM approach across three clinical prediction tasks using three uncertainty estimation methods: sequence likelihood (SLL), generation-based prediction with bootstrapped generation (GEN^{BS}). The results show that MUSE, particularly with the Greedy v2 strategy that minimizes total uncertainty, consistently improves AUROC while also reducing calibration error and Brier score compared to individual LLMs. For instance, in the Mortality prediction task, Greedy v2 achieves the highest AUROC (62.88) and the lowest Brier score (9.51), outperforming both Mistral and Gemma models under all methods. Similarly, in LOS3 and LOS7, MUSE achieves competitive or best AUROC while offering substantial improvements in calibration, with ECE as low as 4.02 in Mortality. The Conservative variant further enhances calibration, reaching an ECE of 3.88, though at the cost of slightly lower AUROC. These findings demonstrate the effectiveness of MUSE in producing more reliable and better-calibrated predictions by aggregating complementary strengths from multiple LLMs.

A.6 MUSE JSD from the four LLMs

Table 9 reports the average JSD between each LLM and the MUSE consensus. Both weaker and stronger models show similarly small divergences, with no clear relation between a model's standalone performance and its JSD to MUSE.

A.7 Analyzing the adapting behavior of MUSE method via EHRShot

Table 7 illustrates the adaptive behavior of our multi-LLM calibration algorithm. In the hyperlipidemia task, where all individual LLMs perform modestly, the aggregated model underperforms, indicating that combining weak predictors can degrade performance. In contrast, for the lupus task, where strong base models (e.g., Deepseek-Distill) are available, the algorithm adapts effectively, matching the best AUROC while maintaining good calibration. This contrast demonstrates the algorithm's adaptive strength: it amplifies strong signals when present, but cannot compensate when no reliable model exists.

Table 8 shows the impact of different LLM combinations on AUROC for the EHRShot acute MI task. While naive averaging yields modest performance gains, the MUSE algorithm substantially boosts AUROC by selectively aggregating informative models. Notably, combinations with higher average AUROC do not always lead to better algorithmic performance: e.g., Qwen+Mistral ranks

Model	SFT	Output	With $\hat{p}$ AUROC↑	With $\hat{p}$ Brier↓	With $\hat{p}$ ECE↓	No $\hat{p}$ AUROC↑	No $\hat{p}$ Brier↓	No $\hat{p}$ ECE↓
Mistral	default	SLL	**54.83**	39.51	31.84	**47.82**	61.17	45.69
		GEN^{BS}	49.28	29.39	**20.00**	**50.59**	38.82	29.75
	RP	SLL	**43.46**	53.75	54.04	41.48	59.28	59.28
		GEN^{BS}	47.76	29.19	**16.63**	**52.08**	39.72	31.49
Qwen	default	SLL	49.52	58.89	46.23	48.56	45.25	44.34
		GEN^{BS}	44.31	49.97	**20.70**	51.58	**37.43**	**15.82**
	RP	SLL	55.35	38.94	36.99	51.00	46.54	44.77
		GEN^{BS}	51.86	41.52	**15.29**	48.20	37.36	**18.19**

Table 11: Direct SFT on Mistral-7B and Qwen2-7B under default and RawProb (RP) settings, with and without inclusion of $\hat{p}$. **Bold** indicates improvement over the no-SFT baselines: Mistral-7B: SLL 41.41/11.34/26.21, GEN^{BS} 51.28/15.63/28.37; Qwen2-7B: SLL 56.40/36.50/35.96, GEN^{BS} 56.80/31.10/34.18, reported in Tab 3 and Tab 6.

Model	SFT	Output	Original AUROC	Original Brier	Original ECE	Bayesian AUROC	Bayesian Brier	Bayesian ECE	No $\hat{p}$ AUROC	No $\hat{p}$ Brier	No $\hat{p}$ ECE
Mistral	default	SLL	**54.22**	42.90	42.97	**42.14**	43.04	41.03	39.27	35.09	31.23
		GEN^{BS}	**51.91**	27.97	**14.66**	49.30	28.80	**16.48**	48.52	41.73	31.03
	RawProb	SLL	**44.79**	57.75	57.78	**58.73**	28.90	**21.88**	**64.51**	51.46	52.80
		GEN^{BS}	49.16	28.97	**17.60**	47.56	29.66	**17.38**	52.05	42.01	33.29
Qwen	default	SLL	50.34	52.56	52.08	56.35	**30.76**	**23.23**	48.89	36.79	32.53
		GEN^{BS}	49.72	44.55	**16.63**	49.77	43.65	**16.49**	**52.95**	37.29	**17.59**
	RawProb	SLL	48.07	36.16	**32.97**	45.43	58.95	59.01	43.63	51.21	49.74
		GEN^{BS}	52.55	40.98	**15.19**	55.72	39.83	**13.81**	47.94	40.25	**19.23**

Table 12: SFT on Mistral-7B and Qwen2-7B with CoT prompts (Original, Bayesian, No $\hat{p}$), under Default/RawProb CoT and SLL/GEN^{BS}. **Bold** values indicate improvements over the no-SFT baselines (Qwen2-7B: SLL 56.40/36.50/35.96, GEN^{BS} 56.80/31.10/34.18; Mistral-7B: SLL 41.41/11.34/26.21; GEN^{BS} 51.28/15.63/28.37), reported in Tab 3 and Tab 6.

highest by average but is outperformed by combinations including DeepSeek when selected adaptively. This reinforces that performance is not simply a function of the number of models, but of their individual quality and how well their signals complement one another.

A.8 Parameter sensitivity on EHRShot

Figure 3 shows how MUSE performance varies with the two key hyperparameters: minimum subset size (m_{size}) and epistemic tolerance (ϵ_{tol}), using EHRShot as the evaluation dataset (*lupus* prediction). The top row shows unweighted AUROC, ECE, and Brier scores, while the bottom row shows the same metrics when aleatoric uncertainty is used as weighting in aggregation.

Overall, larger m_{size} and moderate ϵ_{tol} (0.04–0.08) consistently lead to better performance across all metrics. The gains are especially pronounced in calibration (lower ECE and Brier), showing the benefit of including diverse yet coherent model outputs. Aleatoric weighting further improves stability, particularly under looser inclusion criteria. These trends confirm that careful tuning of subset size and disagreement tolerance is key to balancing diversity and reliability in multi-LLM ensembles.

A.9 Prompting Strategy for Chain-of-Thought

Table 10 shows the prompts used for generating the base and Bayesian chain-of-thought reasoning paragraphs.

A.10 Detailed results of MUSE-guided SFT

When comparing against the no-SFT baselines, Direct SFT shows mixed improvements (see Table 11). For Mistral, SLL with $\hat{p}$ markedly improves AUROC (41.41 → 54.83), though calibration deteriorates (ECE 26.21 → 31.84). GEN^{BS} settings yield the strongest calibration benefits, lowering ECE to 16.63 compared to the baseline 28.37, while maintaining competitive AUROC (52.08). For Qwen, SLL with RawProb achieves AUROC

(55.35) close to baseline (56.40) but with worse calibration (ECE 36.99 vs. 34.18). By contrast, GEN^{BS} achieves notable calibration gains, reducing ECE to 15.29–18.19 compared to the baseline 34.18, though AUROC remains lower. Overall, Direct SFT demonstrates that introducing $\hat{p}$ or raw probability signals can help calibration, especially under GEN^{BS}, but often at the cost of discrimination.

CoT distillation SFT (Table 12). Adding teacher-generated reasoning further diversifies the outcomes. For Mistral, Bayesian RawProb SLL produces the highest AUROC (58.73 vs. baseline 41.41), while No $\hat{p}$ RawProb reaches 64.51, albeit with very poor calibration (ECE 52.80). GEN^{BS} again proves more stable, with Original CoT reducing ECE to 14.66, and Bayesian CoT to 16.48, compared to the baseline 28.37. For Qwen, Bayesian Default SLL achieves balanced improvement, raising AUROC to 56.35 while reducing ECE to 23.23 (baseline 34.18). Similarly, GEN^{BS} with RawProb Bayesian CoT yields the lowest calibration error overall (ECE 13.81). These findings suggest that CoT distillation can help models internalize probabilistic reasoning and achieve strong calibration improvements, though sometimes at the expense of discrimination (e.g., Qwen SLL RawProb).

The results highlight that MUSE-derived uncertainty can provide useful supervision, but the benefits are strongly dependent on LLM, supervision style, and whether the probabilistic signals are contextualized through reasoning.