EFFICIENT EPISTEMIC UNCERTAINTY ESTIMATION FOR LARGE LANGUAGE MODELS VIA KNOWLEDGE DISTILLATION

Anonymous authors
Paper under double-blind review

ABSTRACT

Quantifying uncertainty in Large Language Models (LLMs) is essential for mitigating hallucinations and enabling risk-aware deployment in safety-critical tasks. However, estimating *Epistemic Uncertainty* (EU) via Deep Ensembles is computationally prohibitive at the scale of modern models. We propose a framework that leverages the small draft models to efficiently estimate token-level EU, bypassing the need for full-scale ensembling. Theoretically grounded in a *Bias-Variance Decomposition*, our approach approximates EU via Jensen-Shannon divergence among drafts (variance proxy) and KL divergence between the draft mixture and the target (bias proxy). We additionally bound finite-reference EU transfer error by teacher-specific predictive mismatch, distinguishing disagreement preservation from predictive-mean approximation. To support accurate estimation, we introduce *Online Stochastic Distillation* (OSD) to efficiently approximate target aggregation and the *Data-Diverse Drafts* (DDD) strategy to enhance draft diversity for better target approximation. Crucially, our approach achieves Hallucination Detection performance competitive with heavy perturbation-based methods like TokUR while replacing repeated target-model uncertainty passes with draft and proxy scoring, offering a practical solution for uncertainty-aware LLM deployment.

1 INTRODUCTION

Large Language Models (LLMs) have demonstrated remarkable performance across a wide range of reasoning, generation, and decision-making tasks Bubeck et al. (2023); Gemini Team (2023). Despite these advances, their tendency to produce fluent yet incorrect responses, which is commonly referred to as *hallucinations*, remains a critical obstacle to reliable real-world deployment Huang et al. (2025b); Alansari & Luqman (2025); Ji et al. (2023). A key challenge underlying hallucinations is the lack of reliable uncertainty estimation, which prevents models from recognizing when they are likely to be wrong.

Uncertainty estimation in neural networks is commonly framed through a Bayesian lens, decomposing predictive uncertainty into *aleatoric uncertainty*, which arises from inherent noise in the data, and *epistemic uncertainty*, which reflects uncertainty over model parameters Kendall & Gal (2017); Hüllermeier & Waegeman (2021). Among these, *epistemic uncertainty* (EU) is particularly valuable for safety-critical applications, as it signals insufficient knowledge and correlates strongly with out-of-distribution (OOD) inputs and hallucination-prone queries Ovadia et al. (2019); Gal & Ghahramani (2016); Farquhar et al. (2024); Zhang et al. (2025). Unlike aleatoric uncertainty which captures inherent data noise, epistemic uncertainty effectively quantifies the model's 'ignorance' stemming from data scarcity. Consequently, high epistemic uncertainty serves as a direct indicator of potential confabulation, making it a vital reliability metric for high-stake domains. Accurately estimating EU enables models to abstain, or trigger verification mechanisms, thereby significantly improving trustworthiness.

However, estimating epistemic uncertainty for modern LLMs remains challenging. Bayesian approaches such as *Deep Ensembles* provide a principled approximation by integrating over multiple independently trained models Lakshminarayanan et al. (2017), but their computational and memory costs are prohibitive for models with billions of parameters. Alternative perturbation-based methods,

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

such as dropout sampling, input paraphrasing, or token-level resampling, avoid maintaining multiple model checkpoints Gal & Ghahramani (2016); Ovadia et al. (2019); Kuhn et al. (2023). However, in autoregressive generation, these approaches often conflate epistemic and aleatoric uncertainty and require repeated forward passes of the same large model Ovadia et al. (2019); Farquhar et al. (2024). These limitations significantly restrict their applicability in latency-sensitive or large-scale settings.

At the same time, recent advances in inference acceleration have introduced an overlooked opportunity. Efficient inference paradigms (Leviathan et al., 2023) employ one or more lightweight draft models to propose candidate tokens, which are subsequently verified by a larger target model. Although originally designed for efficiency, this strategy naturally produces a family of auxiliary models that approximate the target model's predictive behavior while differing in architecture size, training data, and optimization dynamics.

From a Bayesian perspective, epistemic uncertainty arises from variability across plausible models sampled from the posterior over parameters. While directly sampling from the posterior of an LLM is infeasible, we recognize that a single lightweight draft model often lacks the expressivity to fully capture the target's complex distribution. To address this, we leverage *ensembles of draft models*, where each constituent member acts as a lightweight approximation to a different mode of the target model's posterior predictive distribution, collectively capturing the multi-modal nature of it. Under this view, disagreement among draft ensembles provides a natural proxy for epistemic uncertainty: when inputs are well-supported by the training distribution, the ensembles tend to agree, whereas for ambiguous or OOD queries, their predictions diverge substantially. Here, the posterior-proxy view concerns the predictive variation of the chosen teacher reference, not literal posterior sampling by drafts. Matching only the predictive mean can erase this variation, while unrelated draft disagreement can inflate it (Section 2).

Nevertheless, draft models are imperfect approximations of the target and inevitably introduce bias. Thus, effective uncertainty estimation must account for both the diversity among drafts and their mismatch with the target model. Motivated by this observation, we ask the following question:

> *Can we leverage small draft models to estimate the epistemic uncertainty of a large target model both accurately and efficiently?*

In this work, we answer this question affirmatively and propose a principled framework for epistemic uncertainty estimation using draft models. Our approach treats a family of drafts as a predictive proxy for the chosen teacher reference and analyzes estimation error separately from the exact decomposition of the draft score into disagreement and mean-mismatch components. We derive a bias–variance decomposition showing that epistemic uncertainty can be approximated by (i) the Jensen–Shannon divergence among draft predictions as a variance proxy, and (ii) the KL-divergence between the draft mixture and the target model as a bias proxy. Equation (1) decomposes the score itself; Eq. (2) separately identifies its disagreement-transfer, mean-mismatch, and proxy-reference errors.

To approximate the teacher predictive mean while retaining teacher-related draft variation, we introduce *Online Stochastic Distillation (OSD)* combined with a *Data-Diverse Drafts (DDD)* strategy, where drafts are trained on disjoint data partitions. On-policy distillation methods such as GKD Agarwal et al. (2024) or MiniLLM Gu et al. (2024) primarily target student generation quality and teacher alignment, but these objectives do not by themselves guarantee preservation of teacher-family disagreement. Our draft-family design instead explicitly aims to retain complementary teacher-related predictions. In this division of roles, OSD learns the predictive mean, whereas DDD aims to preserve teacher-related variation. The relevant distinction is mean-only versus teacher-specific alignment, not alignment versus diversity.

Empirically, we evaluate our method on GSM8K Cobbe et al. (2021), ASQA Stelmakh et al. (2022) and demonstrate a 37.67% reduction in reduction in RMSE against a perturbation-induced posterior proxy compared to strong baselines. Moreover, our approach achieves hallucination detection performance competitive with heavy perturbation-based methods such as TokUR Zhang et al. (2025), while incurring low marginal overhead under deployment assumptions. In summary, our contributions are:

- We propose a novel perspective that leverages small draft models as a proxy for the target model's posterior predictive distribution.

- We derive a bias–variance decomposition for epistemic uncertainty estimation, grounding draft disagreement in a principled theoretical framework. Theorem 2.4 further bounds finite-reference EU transfer error in terms of teacher-specific predictive fidelity.
- We introduce Online Stochastic Distillation and Data-Diverse Drafts for predictive-mean approximation and teacher-related draft variation, respectively.
- We empirically demonstrate that our method enables accurate and efficient epistemic uncertainty estimation for large language models.

2 Theoretical Analysis

We distinguish an exact decomposition of the draft score from a guarantee that it approximates target epistemic uncertainty (EU). The latter follows from predictive-distribution fidelity, not from treating distilled parameters as posterior samples. Full proofs are given in Appendix A.

2.1 Preliminaries: Token-Level Epistemic Uncertainty

Fix a prefix $h = (x, y_{<t})$. All models are evaluated at this same prefix on a common finite vocabulary $\mathcal{V}$; write $p_\theta(v) = p_\theta(v \mid h)$. Let π_T denote a specified teacher-parameter reference distribution and define

$$p_T = \mathbb{E}_{\theta\sim\pi_T} p_\theta, \qquad U_T = H(p_T) - \mathbb{E}_{\theta\sim\pi_T} H(p_\theta).$$

Under the joint law $\theta \sim \pi_T$, $Y \mid \theta, h \sim p_\theta$, U_T is $I(Y;\theta \mid h)$. An exact Bayesian posterior is one possible reference; the low-rank Gaussian family used here instead defines reference-relative epistemic variability (Appendix A.3). All logarithms are natural, $0 \log 0 = 0$, and KL identities require the displayed divergences to be finite. Entropy and JSD statements allow zero probabilities.

Lemma 2.1 (EU as expected divergence)**.** *At every fixed prefix,* $U_T = \mathbb{E}_\theta \, \mathrm{KL}(p_\theta \| p_T)$.

For a fixed, equally weighted family $p_1, \ldots, p_M$, write $p_T^{(M)} = M^{-1} \sum_i p_i$ and $U_T^{(M)} = \mathrm{JSD}(p_1, \ldots, p_M)$. This finite-reference quantity is distinct from the population U_T.

2.2 Exact Score and Estimation-Error Decompositions

Let $q_{\mathrm{mix}} = K^{-1} \sum_k q_k$, $J_D = K^{-1} \sum_k \mathrm{KL}(q_k \| q_{\mathrm{mix}})$, and $S_r = K^{-1} \sum_k \mathrm{KL}(q_k \| r)$, where $r = p_{\mathrm{mix}}$ is the practical proxy reference. Thus $J_D = \mathrm{JSD}(q_1, \ldots, q_K)$, and S_r is the full draft score, denoted $\widehat{\mathrm{EU}}$. A scored member q_k may itself be an averaged group prediction.

Proposition 2.2 (Divergence decompositions)**.** *For any reference r with the required support,* $\mathbb{E}_\theta \, \mathrm{KL}(p_\theta \| r) = U_T + \mathrm{KL}(p_T \| r)$ *and* $S_r = J_D + \mathrm{KL}(q_{\mathrm{mix}} \| r)$. *In particular,*

$$\frac{1}{K} \sum_{k=1}^{K} \mathrm{KL}(q_k \| p_T) = \underbrace{\mathrm{JSD}(q_1, \ldots, q_K)}_{\textit{variance proxy}} + \underbrace{\mathrm{KL}(q_{\mathrm{mix}} \| p_T)}_{\textit{bias / mean-mismatch proxy}} . \tag{1}$$

These identities do not require predictive alignment. Setting $r = q_{\mathrm{mix}}$ in the teacher-side identity gives an upper bound on U_T, but does not make the oppositely directed draft score an upper bound. Its actual error is

$$S_r - U_T = (J_D - U_T) + \mathrm{KL}(q_{\mathrm{mix}} \| p_T) + \sum_{v\in\mathcal{V}} q_{\mathrm{mix}}(v) \log \frac{p_T(v)}{r(v)}. \tag{2}$$

Only the middle term is necessarily nonnegative. The bias proxy is therefore not an unconditional correction of EU estimation error. Mean alignment alone is insufficient: the two teachers $(0.9, 0.1)$ and $(0.1, 0.9)$ have positive EU, whereas drafts both equal to $(0.5, 0.5)$ have the same mean and zero score with an exact reference.

2.3 Uncertainty Preservation under Distillation

Let μ_T^h be the law of the predictive vector $p_\theta(\cdot \mid h)$ under π_T, and $\mu_D^h = K^{-1} \sum_k \delta_{q_k(\cdot|h)}$.

Assumption 2.3 (Predictive-distribution alignment)**.** At the prefix under consideration, $\mu_D^h = \mu_T^h$.

This is the exact-recovery limit of the posterior-proxy interpretation: it implies $q_{\mathrm{mix}} = p_T$ and $J_D = S_{p_T} = U_T$, without identifying draft parameters with teacher parameters. The following bound replaces an unspecified approximate equality of expectations with a quantitative sufficient condition.

For $M \geq 2$, let $\widetilde{q}_i$ approximate reference member p_i and set $\varepsilon = M^{-1} \sum_i \mathrm{TV}(p_i, \widetilde{q}_i)$, where $\mathrm{TV}(p, q) = \frac{1}{2} \sum_v |p(v) - q(v)|$. Define

$$b_M(e) = \begin{cases} h_2(e) + e \log(M-1), & 0 \leq e \leq 1 - 1/M, \\ \log M, & 1 - 1/M < e \leq 1, \end{cases} \qquad h_2(e) = -e \log e - (1-e) \log(1-e). \tag{3}$$

Theorem 2.4 (Predictive fidelity preserves finite-reference EU)**.** *For any such paired families, with no independence assumption,*

$$\left| \mathrm{JSD}(\widetilde{q}_1, \ldots, \widetilde{q}_M) - U_T^{(M)} \right| \leq b_M(\varepsilon) \leq b_M\left(\min\left\{ 1, \sqrt{\mathcal{L}_{\mathrm{pair}}/2} \right\} \right), \tag{4}$$

where $\mathcal{L}_{\mathrm{pair}} = M^{-1} \sum_i \mathrm{KL}(p_i \| \widetilde{q}_i)$.

The proof represents both JSDs as mutual information between a uniform teacher index and a token, then applies conditional-entropy continuity and Pinsker's inequality (Appendix A.2). The bound depends on M, not $|\mathcal{V}|$, and tends to zero with predictive distillation error. It is a pointwise guarantee: small training loss alone does not assert small $\mathcal{L}_{\mathrm{pair}}$ at every prefix, and matching only the predictive mean does not satisfy the paired condition.

2.4 Structured Draft Families

For the grouping in Section 3.4, write the base drafts as q_{ij}, with M groups and R drafts per group, and let $\bar{q}_i = R^{-1} \sum_j q_{ij}$. Define $J_{\mathrm{between}} = \mathrm{JSD}(\bar{q}_1, \ldots, \bar{q}_M)$ and $J_{\mathrm{within}} = M^{-1} \sum_i \mathrm{JSD}(q_{i1}, \ldots, q_{iR})$.

Proposition 2.5 (Between-group and within-group disagreement)**.** *With* $q_{\mathrm{mix}} = M^{-1} \sum_i \bar{q}_i$*, the flat and grouped full scores satisfy*

$$\underbrace{\frac{1}{MR} \sum_{i,j} \mathrm{KL}(q_{ij} \| r)}_{S_{\mathrm{flat},r}} = \underbrace{J_{\mathrm{between}} + \mathrm{KL}(q_{\mathrm{mix}} \| r)}_{S_{\mathrm{group},r}} + J_{\mathrm{within}}. \tag{5}$$

If $\varepsilon = M^{-1} \sum_i \mathrm{TV}(p_i, \bar{q}_i)$, Theorem 2.4 consequently gives

$$|S_{\mathrm{flat},r} - U_T^{(M)}| \leq b_M(\varepsilon) + J_{\mathrm{within}} + \mathrm{KL}(q_{\mathrm{mix}} \| r). \tag{6}$$

For the grouped score, omit J_{within} from this bound. Between-group disagreement transfers teacher variability only to the extent that group means match their corresponding teachers; within-group disagreement measures additional variation among drafts assigned to the same teacher. Data partitioning alone guarantees neither condition. The two aggregation orders share the same predictive mean but differ by J_{within}, so their transfer guarantees apply to different scores.

2.5 Learning the Proxy Mean via Online Stochastic Distillation

OSD minimizes $\mathcal{L}_{\mathrm{OSD}}(\phi) = \mathbb{E}_{h \sim \mathcal{D}} \mathbb{E}_{\theta \sim \pi_T} \mathrm{KL}(p_\theta \| p_{\mathrm{mix},\phi})$, with teacher realizations sampled independently of the fixed prefix distribution $\mathcal{D}$.

Proposition 2.6 (OSD as predictive-mean approximation)**.** *The population objective satisfies*

$$\mathcal{L}_{\mathrm{OSD}}(\phi) = \mathbb{E}_{h \sim \mathcal{D}} U_T(h) + \mathbb{E}_{h \sim \mathcal{D}} \mathrm{KL}(p_T(\cdot \mid h) \| p_{\mathrm{mix},\phi}(\cdot \mid h)). \tag{7}$$

Thus an attained optimum is the best forward-KL approximation to p_T within the proxy family; exact equality additionally requires realizability and zero population excess risk. This justifies learning

a mean proxy, not a claim of automatic neural-network convergence. OSD on a shared teacher mixture can collapse all drafts to that mean, whereas teacher-specific predictive fidelity preserves disagreement. Appendix A.4 proves these statements and gives sufficient conditions for stability with respect to the proxy reference. Appendix A.5 gives the distinct ideal-sampling regime in which the added KL term is a finite-ensemble bias correction. All scores above are token-level; summation along a generated path is an operational sequence score, not automatically sequence-level mutual information.

3 Experimental Setup

We designed a comprehensive suite of experiments to validate our theoretical framework and determine the optimal configuration for efficient EU estimation and its application to downstream tasks.

3.1 Models and Posterior Simulation

To evaluate robustness across model families and scale ratios, we used four instruction-tuned target–draft pairs spanning Qwen Qwen Team (2024; 2025) and Llama Touvron et al. (2023) families: Qwen3-30B-Instruct → Qwen3-1.7B-Instruct, Llama-3.1-8B-Instruct → Llama-3.2-1B-Instruct, Llama-3.2-3B-Instruct → Llama-3.2-1B-Instruct, and Qwen2.5-3B-Instruct → Qwen2.5-0.5B-Instruct.

To create a ground-truth *Target Family* representing the posterior $\pi_T(\theta)$, we injected Low-rank Gaussian Noise into the parameters of the target model. This process generates a set of perturbed models that act as samples from the posterior. We generated target families of size 3 to evaluate estimation performance under different posterior densities. Here, "ground-truth" denotes a fixed evaluation reference, not a true Bayesian posterior; its finite-family EU is $U_T^{(M)}$ in Section 2.

3.2 Datasets and Data Generation

Experiments were conducted on the GSM8K, ASQA dataset. To train the draft models and the proxy p_{mix}, we generated a synthetic training set where the Target model produced 4 diverse responses for each query in the dataset. This multi-output data captures the aleatoric and epistemic uncertainty of the teacher, providing the necessary signals for distillation. Response diversity supplies training data; the theoretical EU is instead defined across teacher realizations at the same prefix.

3.3 Proxy and Draft Training

We distinguish between proxy training and draft-family training. The proxy p_{mix} is trained with OSD to approximate the target predictive average p_T, which is used as the reference distribution for the bias term. The draft family provides the distributions $\{q_k\}_{k=1}^K$: their disagreement yields the variance term, and their mixture q_{mix} is compared with p_{mix} to compute the bias term.

Training the Proxy p_{mix}. To obtain the proxy model for the Bias term, we apply OSD on the full dataset. By training a single student against stochastic teacher samples, OSD seeks a forward-KL approximation to the teacher predictive mean p_T used in Eq. 1 within the proxy model family, rather than assuming that finite-capacity training recovers p_T exactly. Precisely, Proposition 2.6 characterizes the population optimum: equality with p_T requires realizability and zero population excess risk.

Training the Draft Family. To evaluate how different sources of stochasticity affect posterior approximation, we categorize the draft training into three distinct variants:

- **Data-Diverse Drafts (DDD):** Each draft model is trained on a *disjoint partition* of the target-generated dataset, inducing diversity through variations in the training data distribution.
- **Initialization-Diverse Drafts (IDD):** Each draft model is initialized with diverse parameter noise and trained on the *entire* dataset, isolating the effect of stochastic starting points on posterior coverage.

- **Fully-Diverse Drafts (FDD):** FDD combines both strategies by applying diverse parameter initialization and training each model on disjoint data partitions.

We include GKD and MiniLLM as on-policy distillation baselines for draft training. Unlike DDD, IDD, and FDD, these methods primarily optimize student generation quality and target alignment; such alignment does not by itself guarantee preservation of teacher-family disagreement. Theorem 2.4 applies conditionally on teacher-specific predictive fidelity, not on partitioning alone; $\mathcal{L}_{\text{pair}}$ quantifies that fidelity for analysis rather than introducing an additional training objective.

3.4 Ensemble Configuration

To investigate the interplay between distillation strategies and ensemble construction, we focus on a structured 2×3 configuration designed to capture the diversity of the target posterior. In this setup, we train six independent draft models and ensemble them in pairs to form three *imitated groups*. Each group is specifically optimized to collectively mimic the predictive behavior of a corresponding member within the target family. This hierarchical approach allows the ensemble to effectively preserve the epistemic uncertainty of the target model by leveraging both the diversity induced during independent training and the collaborative refinement provided by group ensembling. For paired mean predictions, $M = 3$, $R = 2$ in Proposition 2.5; flat scoring adds J_{within} to the score of the three means. The transfer guarantee concerns fidelity to each corresponding teacher.

3.5 Evaluation Protocols and Metrics

Uncertainty Estimation. We measure the fidelity of estimated EU against the ground-truth Target EU using *RMSE* (Lower is better) and *Spearman's* ρ (Higher is better). We also report the standard deviation across 5 repeated trials to ensure robustness.

Hallucination Detection. We utilize token-level uncertainty to identify incorrect generations. Since raw scores are not calibrated probabilities, we train a *logistic regression classifier* on a hold-out set to map uncertainty to correctness probabilities. Performance is measured via *AUROC* (discrimination), *ECE*, and *Brier Score* (calibration). We compare our results against *TokUR Zhang et al. (2025)*, a heavy baseline requiring multiple target-level perturbations. Reference fidelity and downstream detection quality are separate criteria; neither alone establishes equality of teacher and draft predictive laws.

4 Experimental Results and Analysis

We evaluate whether lightweight draft scorers can approximate target-perturbation epistemic uncertainty (EU). All main experiments use the 2×3 ensemble configuration and report RMSE (↓) and Spearman correlation (↑). We first evaluate EU estimation fidelity, then discuss scalability, ablations, and downstream hallucination detection with inference cost.

4.1 Uncertainty Estimation Across Model Scales

Table 1 compares DDD with No Train, IDD, FDD, GKD, and MiniLLM across three target–draft model pairs and two datasets. DDD achieves the highest mean Spearman correlation in all six dataset–model settings and the lowest mean RMSE in five of six. The RMSE exception is L8B→L1B on ASQA, where GKD obtains a lower RMSE (0.593 versus 0.632), while DDD achieves a higher Spearman correlation (0.879 versus 0.845).

The gains also extend to the largest target model. On Q30B→Q1.7B, DDD reduces RMSE relative to No Train from 2.031 to 1.067 on GSM8K and from 2.838 to 1.683 on ASQA. Spearman correlation increases from 0.755 to 0.854 and from 0.719 to 0.809, respectively. Taken together, these results indicate that DDD preserves the ordering of target-perturbation EU across the evaluated model scales.

Table 1: **Uncertainty Estimation Performance.** Comparison of RMSE (↓) and Spearman correlation (↑) across target–draft model scales and datasets. Bold indicates the best mean in each setting.

		30B (Target) → 1.7B (Draft)		8B (Target) → 1B (Draft)		3B (Target) → 1B (Draft)	
Dataset	**Method**	**RMSE ↓**	**Spearman ↑**	**RMSE ↓**	**Spearman ↑**	**RMSE ↓**	**Spearman ↑**
GSM8K	No Train	2.031 ± 0.009	0.755 ± 0.001	0.572 ± 0.014	0.819 ± 0.004	0.470 ± 0.006	0.833 ± 0.003
	GKD	1.240 ± 0.003	0.829 ± 0.000	0.624 ± 0.004	0.833 ± 0.001	0.508 ± 0.011	0.818 ± 0.004
	MiniLLM	1.305 ± 0.011	0.843 ± 0.001	0.648 ± 0.003	0.835 ± 0.002	0.520 ± 0.011	0.820 ± 0.004
	IDD	1.478 ± 0.013	0.813 ± 0.001	0.593 ± 0.005	0.884 ± 0.000	0.587 ± 0.009	0.888 ± 0.002
	FDD	1.457 ± 0.010	0.826 ± 0.000	0.564 ± 0.011	0.890 ± 0.003	0.560 ± 0.011	0.890 ± 0.003
	DDD (Ours)	$\mathbf{1.067 \pm 0.016}$	$\mathbf{0.854 \pm 0.000}$	$\mathbf{0.444 \pm 0.010}$	$\mathbf{0.895 \pm 0.001}$	$\mathbf{0.402 \pm 0.005}$	$\mathbf{0.902 \pm 0.002}$
ASQA	No Train	2.838 ± 0.007	0.719 ± 0.002	0.958 ± 0.006	0.793 ± 0.000	0.657 ± 0.012	0.791 ± 0.005
	GKD	1.843 ± 0.007	0.778 ± 0.002	$\mathbf{0.593 \pm 0.005}$	0.845 ± 0.001	0.514 ± 0.015	0.815 ± 0.005
	MiniLLM	1.907 ± 0.004	0.795 ± 0.001	0.828 ± 0.011	0.791 ± 0.002	0.597 ± 0.007	0.805 ± 0.006
	IDD	1.926 ± 0.011	0.808 ± 0.002	0.819 ± 0.003	0.872 ± 0.001	0.593 ± 0.019	0.837 ± 0.005
	FDD	1.933 ± 0.010	0.804 ± 0.002	0.808 ± 0.004	0.870 ± 0.001	0.591 ± 0.017	0.829 ± 0.007
	DDD (Ours)	$\mathbf{1.683 \pm 0.002}$	$\mathbf{0.809 \pm 0.001}$	0.632 ± 0.002	$\mathbf{0.879 \pm 0.001}$	$\mathbf{0.422 \pm 0.012}$	$\mathbf{0.844 \pm 0.003}$

4.2 Ablations and Robustness

Role of the Distilled Proxy. We examine the reference distribution used in the bias term of Eq. 1. Table 5 compares the distilled proxy p_{mix} with the raw unperturbed target distribution under ten perturbations. Using p_{mix} reduces RMSE from 0.0391 to 0.0137, while Spearman correlation increases slightly from 0.9957 to 0.9969. This ablation supports the use of the distilled proxy for more accurate EU estimation in this setting.

Effect of Draft Scorer Capacity. We further evaluate Q3B→Q0.5B as a small-draft setting. As shown in Table 4, DDD achieves the lowest RMSE and highest Spearman correlation among the compared methods on both GSM8K and ASQA. Its absolute estimation performance is lower than in the Q3B→Q1B setting in Table 1, but it retains an advantage over the other methods evaluated with the 0.5B scorer.

4.3 Hallucination Detection and Inference Cost

We evaluate whether draft-based EU helps identify incorrect generations and measure its inference cost. Table 2 reports results for L8B→L1B with the 1×3 DDD configuration. On GSM8K, Full EU achieves the highest mean AUROC among the evaluated methods (0.768) and ties TokUR for the lowest mean Brier score (0.155). On ASQA, its mean AUROC (0.658) exceeds both Draft Disagreement (0.611) and Target Mismatch (0.649), although the gain over Target Mismatch is small. It also outperforms MSP, Total Entropy, Self-Certainty, Semantic Entropy, and Eccentricity in mean ASQA AUROC.

Full EU takes 0.783 s in the measured setting, compared with 0.890 s for TokUR and 20.726–33.196 s for KLE, Eccentricity, and Semantic Entropy. It thus attains TokUR-level detection performance on GSM8K with lower measured runtime, while providing a substantially faster option than the latter three methods on ASQA. The trade-offs are that its ASQA mean AUROC remains below TokUR (0.676) and KLE (0.701), its ECE is higher than TokUR's on both datasets, and its peak VRAM is higher (19.09 versus 13.13 GiB). The large variation in ASQA AUROC also warrants caution when interpreting small differences between methods.

5 Discussion

Validity of the Bias-Variance Decomposition. The reported RMSE and Spearman agreement characterize fidelity to the chosen teacher reference, complementing the analysis in Section 2. More precisely, the decomposition is an algebraic identity, whereas reference agreement and the conditional guarantee in Theorem 2.4 address predictive fidelity. The additional KL term measures mean mismatch; its finite-ensemble bias-correction interpretation uses the ideal-sampling conditions of Appendix A.5.

The Role of Data Diversity. Different initializations alone do not guarantee preservation of teacher-related disagreement. If all drafts exactly fit the same predictive mean, the Variance Proxy vanishes

Table 2: **Hallucination Detection Performance and Computational Cost.** AUROC (↑), ECE (↓), and Brier score (↓) are reported on GSM8K and ASQA. Our method uses the 1×3 DDD configuration. Runtime and peak GPU memory are measured under the same inference setting for all methods. Draft Disagreement is $\mathbb{E}_k[D_{\mathrm{KL}}(q_k \| q_{\mathrm{mix}})]$; Target Mismatch is $D_{\mathrm{KL}}(q_{\mathrm{mix}} \| p_T)$.

	GSM8K			ASQA		
Method	AUROC ↑	ECE ↓	Brier ↓	AUROC ↑	ECE ↓	Brier ↓
MSP	0.746 ± 0.024	0.075 ± 0.019	0.159 ± 0.005	0.513 ± 0.066	0.055 ± 0.033	0.183 ± 0.006
Total Entropy	0.738 ± 0.016	0.066 ± 0.024	0.160 ± 0.005	0.546 ± 0.027	0.067 ± 0.020	0.180 ± 0.005
Self-Certainty	0.752 ± 0.021	0.072 ± 0.021	0.162 ± 0.004	0.599 ± 0.046	0.047 ± 0.010	0.179 ± 0.004
Semantic Entropy	0.746 ± 0.026	0.066 ± 0.011	0.163 ± 0.014	0.597 ± 0.039	0.047 ± 0.004	0.176 ± 0.003
KLE	0.754 ± 0.027	0.055 ± 0.007	0.167 ± 0.010	$\mathbf{0.701 \pm 0.071}$	0.087 ± 0.007	$\mathbf{0.168 \pm 0.011}$
Eccentricity	0.573 ± 0.205	$\mathbf{0.048 \pm 0.069}$	0.201 ± 0.012	0.557 ± 0.101	$\mathbf{0.032 \pm 0.020}$	0.178 ± 0.006
TokUR	0.765 ± 0.011	0.053 ± 0.004	$\mathbf{0.155 \pm 0.001}$	0.676 ± 0.020	0.062 ± 0.029	0.169 ± 0.003
Draft Disagreement	0.758 ± 0.029	0.062 ± 0.017	0.160 ± 0.009	0.611 ± 0.054	0.033 ± 0.007	0.177 ± 0.004
Target Mismatch	0.767 ± 0.025	0.057 ± 0.012	0.156 ± 0.007	0.649 ± 0.103	0.067 ± 0.020	0.173 ± 0.012
Full EU (Ours)	$\mathbf{0.768 \pm 0.024}$	0.062 ± 0.006	$\mathbf{0.155 \pm 0.007}$	0.658 ± 0.102	0.073 ± 0.011	0.172 ± 0.011

Computational cost		
Method	Runtime (s) ↓	Peak VRAM (GiB) ↓
MSP	0.270	7.55
Total Entropy	0.277	7.55
Self-Certainty	**0.262**	**4.38**
Semantic Entropy	33.196	4.45
KLE	20.726	4.45
Eccentricity	20.731	4.45
TokUR	0.890	13.13
Draft Disagreement	0.783	19.09
Target Mismatch	0.783	19.09
Full EU (Ours)	0.783	19.09

($JSD = 0$); Appendix A.4 establishes this shared-objective limit. DDD is intended to preserve teacher-related predictive variation rather than merely increase draft disagreement. This interpretation is conditional on teacher-specific fidelity; data partitioning alone does not prove posterior coverage or remove within-group disagreement (Proposition 2.5).

Lightweight Models as Effective Auditors. Our method's competitive performance in hallucination detection suggests a strong distributional alignment between target and draft families. Notably, 1B models can effectively "audit" 8B teachers; while smaller models lack full generative capacity, they reliably preserve relative uncertainty rankings, serving as efficient proxies for identifying target "confusion."

Efficiency and Practical Integration. By replacing N target passes with $2N$ draft passes and one proxy pass, we significantly reduce costs. Since draft models are typically memory-resident for speculative decoding, this approach enables efficient uncertainty estimation with low overhead compared to standard Deep Ensembles.

Limitations. We note several limitations of our framework. *First*, our ground-truth target family is not an exact Bayesian posterior but a structured low-rank isotropic Gaussian proxy centered at the teacher weights (Appendix A, Section A.3). A more principled alternative would be a curvature-calibrated covariance, e.g., via a projected Hessian, empirical Fisher, or Kronecker-factored approximation; we leave such curvature-aware target construction to future work. *Second*, Assumption 2.3 specifies an ideal predictive-distribution alignment condition, while Theorem 2.4 provides a finite-reference guarantee conditional on teacher-specific predictive fidelity. Distillation and data partitioning do not automatically establish either condition. *Third*, draft capacity matters: we observe a modest performance dip in the 8B→1B regime (Section 4), indicating that very small drafts can lose fidelity when the target–draft scale gap is large. *Fourth*, our empirical evaluation is limited to two datasets (GSM8K and ASQA) and two model families (Llama and Qwen); generalization to other domains (e.g., code, dialogue, multilingual generation) and to larger scale gaps remains to be verified. *Finally*, our framework treats epistemic uncertainty as posterior-induced functional variation in the

teacher; it does not model epistemic gaps that originate from data scarcity outside the support of the draft training distribution, which would require a separate distributional shift analysis. Formally, the approximation concern is departure from the exact predictive-law alignment in Assumption 2.3; Theorem 2.4 bounds finite-reference transfer, separately from the choice of teacher reference.

6 Related Works

Uncertainty estimation in LLMs. Uncertainty estimation (UE) for LLMs aims to assess when model outputs are likely to be unreliable, and has been studied through likelihood- and calibration-based confidence scores, sampling consistency, prompt/input perturbations, and Bayesian approximations Gawlikowski et al. (2023); Xia et al. (2025); Jiang et al. (2021); Ahdritz et al. (2024); Lyu et al. (2025); Manakul et al. (2023); Tonolini et al. (2024). Bayesian approaches provide a principled view of epistemic uncertainty, but direct posterior inference or ensemble-based approximation is costly for modern LLMs Neal (1996); Kendall & Gal (2017); Hüllermeier & Waegeman (2021); Lakshminarayanan et al. (2017). Recent work therefore approximates uncertainty using parameter-efficient Bayesian modules or perturbations, including LoRA-based approaches for token-level uncertainty during autoregressive generation Yang et al. (2024); Wang et al. (2024); Malinin & Gales (2021); Shi et al. (2025); Zhang et al. (2025). Despite their effectiveness, such perturbation-based methods typically require multiple stochastic forward passes of the target LLM, making them expensive in latency-sensitive settings. Our work targets token-level epistemic uncertainty while replacing repeated target-model perturbations with disagreement among lightweight draft models.

Speculative decoding. Speculative decoding accelerates LLM inference by using a lightweight draft model to propose future tokens that are then verified by a larger target model Stern et al. (2018); Xia et al. (2023). Prior work has improved this framework through better drafting, verification, and uncertainty-aware control policies that adapt draft length or verification frequency to increase throughput Xia et al. (2024); Hu et al. (2025); Qin et al. (2025); Huang et al. (2025a); Agrawal et al. (2024). These methods primarily use uncertainty to optimize decoding efficiency. In contrast, we treat the distributional relationship between draft and target models as a source of epistemic information about the target model itself.

Knowledge distillation. Knowledge distillation (KD) trains a compact student to approximate a teacher distribution and is widely used for LLM compression and acceleration Hinton et al. (2015); Yang et al. (2025). Forward-KL distillation tends to be mode-covering, whereas recent on-policy methods such as MiniLLM and GKD use reverse-KL or flexible divergence and sampling choices to better align student generations with the teacher Cao & Kang (2025); Gu et al. (2024); Agarwal et al. (2024). While such alignment is useful for generation quality, fitting every draft only to the same predictive mean can remove inter-draft disagreement even when the teacher family has nonzero epistemic uncertainty. This motivates our OSD/DDD design: OSD approximates the predictive mean, while DDD aims to retain complementary teacher-related predictions. The collapse argument concerns a shared mean objective; teacher-specific alignment instead supports uncertainty transfer under Theorem 2.4.

7 Conclusion

In this paper, we presented a framework to efficiently estimate the Epistemic Uncertainty of LLMs by leveraging lightweight draft models inherent to efficient inference paradigms. Our experimental results confirm that the **Data-Diverse Drafts (DDD)** strategy significantly outperforms existing distillation methods in capturing posterior diversity, while the distilled proxy (p_{mix}) improves approximation of the teacher predictive mean. Furthermore, we demonstrate that this draft-based uncertainty is highly effective for **Hallucination Detection**, achieving performance comparable to computationally expensive baselines like TokUR while replacing repeated target-model uncertainty passes with draft and proxy scoring. By bridging the gap between efficient inference and reliable uncertainty quantification, our framework provides a practical and scalable solution for ensuring the safety and reliability of LLMs in real-world deployment. Our analysis bounds finite-reference uncertainty transfer by predictive fidelity and distinguishes teacher-related disagreement, within-group variation, and predictive-mean mismatch.

AI USE STATEMENT

We used generative AI tools to assist with formulating mathematical claims and developing proofs, implementing the proposed methods, interpreting results, translating text, drafting portions of the manuscript, improving its grammar and readability, writing and revising software code, and searching for and summarizing relevant literature. We did not use generative AI for synthetic data generation, data cleaning or reformatting, figure creation, or brainstorming.

The authors verified the AI-assisted mathematical claims and proofs, code, and analyses underlying the interpretation of results. The authors also reviewed and verified the AI-generated and translated manuscript drafts. We take full responsibility for the final content of this work, including all claims, proofs, results, text, and code produced with the assistance of generative AI.

REFERENCES

Rishabh Agarwal, Nino Vieillard, Yongchao Zhou, Piotr Stanczyk, Sabela Ramos Garea, Matthieu Geist, and Olivier Bachem. On-Policy Distillation of Language Models: Learning from Self-Generated Mistakes. In *ICLR*, 2024.

Sudhanshu Agrawal, Wonseok Jeon, and Mingu Lee. AdaEDL: Early Draft Stopping for Speculative Decoding of Large Language Models via an Entropy-Based Lower Bound on Token Acceptance Probability. In *NeurIPS Workshops*, 2024.

Gustaf Ahdritz, Tian Qin, Nikhil Vyas, Boaz Barak, and Benjamin L. Edelman. Distinguishing the Knowable from the Unknowable with Language Models. In *ICML*, 2024.

Aisha Alansari and Hamzah Luqman. Large Language Models Hallucination: A Comprehensive Survey. *arXiv:2510.06265*, 2025.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, Harsha Nori, Hamid Palangi, Marco Tulio Ribeiro, and Yi Zhang. Sparks of Artificial General Intelligence: Early Experiments with GPT-4. *arXiv:2303.12712*, 2023.

Yihan Cao and Yanbin Kang. On LLM Knowledge Distillation: A Comparison between Forward KL and Reverse KL. In *ICLR Blogposts*, 2025.

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. Training Verifiers to Solve Math Word Problems. *arXiv:2110.14168*, 2021.

Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting Hallucinations in Large Language Models Using Semantic Entropy. *Nature*, 2024.

Yarin Gal and Zoubin Ghahramani. Dropout as a Bayesian Approximation: Representing Model Uncertainty in Deep Learning. In *ICML*, 2016.

Jakob Gawlikowski, Cedrique Rovile Njieutcheu Tassi, Mohsin Ali, Jongseok Lee, Matthias Humt, Jianxiang Feng, Anna Kruspe, Rudolph Triebel, Peter Jung, Ribana Roscher, Muhammad Shahzad, Wen Yang, Richard Bamler, and Xiao Xiang Zhu. A Survey of Uncertainty in Deep Neural Networks. *Artificial Intelligence Review*, 2023.

Gemini Team. Gemini: A Family of Highly Capable Multimodal Models. *arXiv:2312.11805*, 2023.

Yuxian Gu, Li Dong, Furu Wei, and Minlie Huang. MiniLLM: Knowledge Distillation of Large Language Models. In *ICLR*, 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the Knowledge in a Neural Network. *arXiv:1503.02531*, 2015.

Yunhai Hu, Zining Liu, Zhenyuan Dong, Tianfan Peng, Bradley McDanel, and Sai Qian Zhang. Speculative Decoding and Beyond: An In-Depth Survey of Techniques. *arXiv:2502.19732*, 2025.

Kaixuan Huang, Xudong Guo, and Mengdi Wang. SpecDec++: Boosting Speculative Decoding via Adaptive Candidate Lengths. In *COLM*, 2025a.

Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A Survey on Hallucination in Large Language Models: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 2025b.

Eyke Hüllermeier and Willem Waegeman. Aleatoric and Epistemic Uncertainty in Machine Learning: An Introduction to Concepts and Methods. *Machine Learning*, 2021.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. Survey of Hallucination in Natural Language Generation. *ACM Computing Surveys*, 2023.

Zhengbao Jiang, Jun Araki, Haibo Ding, and Graham Neubig. How Can We Know When Language Models Know? On the Calibration of Language Models for Question Answering. *Transactions of the Association for Computational Linguistics*, 2021.

Alex Kendall and Yarin Gal. What Uncertainties Do We Need in Bayesian Deep Learning for Computer Vision? In *NeurIPS*, 2017.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic Uncertainty: Linguistic Invariances for Uncertainty Estimation in Natural Language Generation. In *ICLR*, 2023.

Balaji Lakshminarayanan, Alexander Pritzel, and Charles Blundell. Simple and Scalable Predictive Uncertainty Estimation Using Deep Ensembles. In *NeurIPS*, 2017.

Yaniv Leviathan, Matan Kalman, and Yossi Matias. Fast Inference from Transformers via Speculative Decoding. In *ICML*, 2023.

Qing Lyu, Kumar Shridhar, Chaitanya Malaviya, Li Zhang, Yanai Elazar, Niket Tandon, Marianna Apidianaki, Mrinmaya Sachan, and Chris Callison-Burch. Calibrating Large Language Models with Sample Consistency. In *AAAI*, 2025.

Andrey Malinin and Mark Gales. Uncertainty Estimation in Autoregressive Structured Prediction. In *ICLR*, 2021.

Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative Large Language Models. In *EMNLP*, 2023.

Radford M. Neal. *Bayesian Learning for Neural Networks*. Springer, 1996.

Yaniv Ovadia, Emily Fertig, Jie Ren, Zachary Nado, D. Sculley, Sebastian Nowozin, Joshua V. Dillon, Balaji Lakshminarayanan, and Jasper Snoek. Can You Trust Your Model's Uncertainty? Evaluating Predictive Uncertainty under Dataset Shift. In *NeurIPS*, 2019.

Zongyue Qin, Zifan He, Neha Prakriya, Jason Cong, and Yizhou Sun. Dynamic-Width Speculative Beam Decoding for LLM Inference. In *AAAI*, 2025.

Qwen Team. Qwen2.5 Technical Report. *arXiv:2412.15115*, 2024.

Qwen Team. Qwen3 Technical Report. *arXiv:2505.09388*, 2025.

Haizhou Shi, Yibin Wang, Ligong Han, Huan Zhang, and Hao Wang. Training-Free Bayesianization for Low-Rank Adapters of Large Language Models. In *NeurIPS*, 2025.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. ASQA: Factoid Questions Meet Long-Form Answers. In *EMNLP*, 2022.

Mitchell Stern, Noam Shazeer, and Jakob Uszkoreit. Blockwise Parallel Decoding for Deep Autoregressive Models. In *NeurIPS*, 2018.

Francesco Tonolini, Nikolaos Aletras, Jordan Massiah, and Gabriella Kazai. Bayesian Prompt Ensembles: Model Uncertainty Estimation for Black-Box Large Language Models. In *ACL Findings*, 2024.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. LLaMA: Open and Efficient Foundation Language Models. *arXiv:2302.13971*, 2023.

Yibin Wang, Haizhou Shi, Ligong Han, Dimitris Metaxas, and Hao Wang. BLoB: Bayesian Low-Rank Adaptation by Backpropagation for Large Language Models. In *NeurIPS*, 2024.

Heming Xia, Tao Ge, Peiyi Wang, Si-Qing Chen, Furu Wei, and Zhifang Sui. Speculative Decoding: Exploiting Speculative Execution for Accelerating Seq2seq Generation. In *EMNLP Findings*, 2023.

Heming Xia, Zhe Yang, Qingxiu Dong, Peiyi Wang, Yongqi Li, Tao Ge, Tianyu Liu, Wenjie Li, and Zhifang Sui. Unlocking Efficiency in Large Language Model Inference: A Comprehensive Survey of Speculative Decoding. In *ACL Findings*, 2024.

Zhiqiu Xia, Jinxuan Xu, Yuqian Zhang, and Hang Liu. A Survey of Uncertainty Estimation Methods on Large Language Models. In *ACL Findings*, 2025.

Adam X. Yang, Maxime Robeyns, Xi Wang, and Laurence Aitchison. Bayesian Low-Rank Adaptation for Large Language Models. In *ICLR*, 2024.

Chuanpeng Yang, Yao Zhu, Wang Lu, Yidong Wang, Qian Chen, Chenlong Gao, Bingjie Yan, and Yiqiang Chen. Survey on Knowledge Distillation for Large Language Models: Methods, Evaluation, and Application. *ACM Transactions on Intelligent Systems and Technology*, 2025.

Tunyu Zhang, Haizhou Shi, Yibin Wang, Hengyi Wang, Xiaoxiao He, Zhuowei Li, Haoxian Chen, Ligong Han, Kai Xu, Huan Zhang, Dimitris Metaxas, and Hao Wang. TokUR: Token-Level Uncertainty Estimation for Large Language Model Reasoning. *arXiv:2505.11737*, 2025.

A Theoretical Proofs

All results are stated at a fixed prefix h unless an expectation over prefixes is displayed. Distributions share a finite token vocabulary $\mathcal{V}$; logarithms are natural. We use $0 \log 0 = 0$, omit zero-mass summands, and assume finiteness of the KL quantities in each divergence identity. In particular, the decomposition relative to both p_T and r requires these distributions to be positive on the support of q_{mix}. No uniform probability lower bound is needed for the JSD transfer theorem. The reference distribution π_T and its finite realization are kept distinct throughout.

A.1 Fundamental Identities and Estimation Error

Proof of Lemma 2.1.

Proof. Because $\mathcal{V}$ is finite, the entropy terms are bounded and expectations can be interchanged with the vocabulary sum. Using $\mathbb{E}_\theta p_\theta(v) = p_T(v)$ gives

$$\begin{aligned}\mathbb{E}_\theta \,\mathrm{KL}(p_\theta \| p_T) &= \mathbb{E}_\theta \sum_{v \in \mathcal{V}} p_\theta(v) \log p_\theta(v) - \sum_{v \in \mathcal{V}} \mathbb{E}_\theta p_\theta(v) \log p_T(v) \\ &= -\mathbb{E}_\theta H(p_\theta) + H(p_T) = U_T.\end{aligned}$$

If $p_T(v) = 0$, then $p_\theta(v) = 0$ for π_T-almost every θ, so that coordinate contributes zero. The joint law $\pi_T(d\theta) p_\theta(v)$ has marginals π_T and p_T; the displayed expected KL is consequently its mutual information. For the uniform finite reference, this argument gives

$$U_T^{(M)} = H(p_T^{(M)}) - \frac{1}{M} \sum_i H(p_i) = \frac{1}{M} \sum_i \mathrm{KL}(p_i \| p_T^{(M)}).$$

□

Proof of Proposition 2.2.

Proof. Introduce $p_T(v)$ into the teacher-side logarithm:

$$\mathbb{E}_\theta \,\mathrm{KL}(p_\theta \| r) = \mathbb{E}_\theta \sum_v p_\theta(v) \left(\log \frac{p_\theta(v)}{p_T(v)} + \log \frac{p_T(v)}{r(v)} \right)$$
$$= U_T + \sum_v p_T(v) \log \frac{p_T(v)}{r(v)} = U_T + \mathrm{KL}(p_T \| r).$$

This also proves the general upper bound obtained by taking $r = q_{\mathrm{mix}}$. Independently, introducing $q_{\mathrm{mix}}(v)$ into the draft-side logarithm yields

$$S_r = \frac{1}{K} \sum_k \sum_v q_k(v) \left(\log \frac{q_k(v)}{q_{\mathrm{mix}}(v)} + \log \frac{q_{\mathrm{mix}}(v)}{r(v)} \right)$$
$$= \frac{1}{K} \sum_k \mathrm{KL}(q_k \| q_{\mathrm{mix}}) + \sum_v \left(\frac{1}{K} \sum_k q_k(v) \right) \log \frac{q_{\mathrm{mix}}(v)}{r(v)}$$
$$= J_D + \mathrm{KL}(q_{\mathrm{mix}} \| r).$$

Setting $r = p_T$ proves Eq. (1). Neither computation replaces an expectation over teachers by an expectation over drafts. □

Proof of the estimation-error decomposition.

Proof. Subtract U_T from the draft identity and use

$$\mathrm{KL}(q_{\mathrm{mix}} \| r) = \mathrm{KL}(q_{\mathrm{mix}} \| p_T) + \sum_v q_{\mathrm{mix}}(v) \log \frac{p_T(v)}{r(v)}.$$

This proves Eq. (2). In particular, define

$$\Delta_J = J_D - U_T, \qquad B_T = \mathrm{KL}(q_{\mathrm{mix}} \| p_T), \qquad E_r = \sum_v q_{\mathrm{mix}}(v) \log \frac{p_T(v)}{r(v)}.$$

Then $S_r - U_T = \Delta_J + B_T + E_r$. The first term is a disagreement-transfer error, the second is a nonnegative predictive-mean mismatch, and the third is the signed effect of replacing the exact reference by the proxy. The last term is not generally a KL divergence: its expectation is under q_{mix}, not p_T. □

Exact alignment and its scope. Under Assumption 2.3, equality of the two probability measures on predictive vectors implies equality of their coordinate means and their expected entropies. Hence

$$q_{\mathrm{mix}} = p_T, \qquad \frac{1}{K} \sum_k H(q_k) = \mathbb{E}_\theta H(p_\theta), \qquad J_D = S_{p_T} = U_T.$$

The entropy functional is bounded by $\log |\mathcal{V}|$, so these expectations are well defined. Equality of the complete predictive laws is sufficient, not necessary, for matching a single uncertainty functional. Moreover, a finitely supported draft law need not equal a continuously supported teacher predictive law; the assumption specifies an ideal limit rather than a property guaranteed by distillation. Theorem 2.4 supplies a finite-reference approximation guarantee without requiring this exact law equality.

Why predictive-mean fidelity is insufficient. Consider two equally weighted teachers on a binary vocabulary,

$$p_1 = (0.9, 0.1), \quad p_2 = (0.1, 0.9), \quad p_T^{(2)} = (0.5, 0.5).$$

Their EU is $U_T^{(2)} = \log 2 - h_2(0.1) > 0$. If both drafts equal $(0.5, 0.5)$ and $r = p_T^{(2)}$, then $q_{\mathrm{mix}} = p_T^{(2)}$ but $J_D = S_r = 0$. Conversely, let both teachers equal $(0.5, 0.5)$ and take $q_1 = (0.9, 0.1)$, $q_2 = (0.1, 0.9)$. The means still agree, but $U_T^{(2)} = 0$ while $S_r = J_D = \log 2 - h_2(0.1) > 0$. Thus exact mean matching permits both underestimation and overestimation. In particular, the teacher-side upper-bound identity cannot be transferred to S_r merely by reversing the roles of teachers and drafts.

A.2 Predictive Fidelity and Structured Draft Families

Conditional-entropy continuity. We use the classical conditional-entropy continuity theorem of Alhejji and Smith.[1] For joint distributions P and Q of a variable Z with $M \geq 2$ values and a finite-valued variable Y, it states that

$$\mathrm{TV}(P,Q) \leq e \leq 1 - 1/M \quad \Longrightarrow \quad |H_P(Z \mid Y) - H_Q(Z \mid Y)| \leq h_2(e) + e \log(M-1). \tag{8}$$

For larger e, the difference is at most $\log M$, since both conditional entropies belong to $[0, \log M]$. At $e = 0$, equality of the joint laws gives zero difference. These observations give the modulus b_M in Eq. (3). The first branch is nondecreasing because its derivative on $(0, 1 - 1/M)$ is

$$\log \frac{(M-1)(1-e)}{e} \geq 0,$$

its value at $1 - 1/M$ is $\log M$, and the second branch is constant. The resulting bound is independent of the number of possible values of the conditioning variable Y.

Proof of Theorem 2.4: total-variation form.

Proof. Let Z be uniform on $\{1, \ldots, M\}$ and define two joint laws on $\{1, \ldots, M\} \times \mathcal{V}$:

$$P(Z = i, Y = v) = \frac{1}{M} p_i(v), \qquad Q(Z = i, Y = v) = \frac{1}{M} \widetilde{q}_i(v).$$

Their total variation distance is exactly the paired prediction error:

$$\begin{aligned} \mathrm{TV}(P,Q) &= \frac{1}{2} \sum_{i=1}^{M} \sum_{v \in \mathcal{V}} \left| \frac{p_i(v)}{M} - \frac{\widetilde{q}_i(v)}{M} \right| \\ &= \frac{1}{M} \sum_{i=1}^{M} \mathrm{TV}(p_i, \widetilde{q}_i) = \varepsilon. \end{aligned}$$

Writing $\widetilde{q}_{\mathrm{mix}} = M^{-1} \sum_i \widetilde{q}_i$, the two mutual informations are

$$\begin{aligned} I_P(Z;Y) &= \frac{1}{M} \sum_i \mathrm{KL}(p_i \| p_T^{(M)}) = U_T^{(M)}, \\ I_Q(Z;Y) &= \frac{1}{M} \sum_i \mathrm{KL}(\widetilde{q}_i \| \widetilde{q}_{\mathrm{mix}}) = \mathrm{JSD}(\widetilde{q}_1, \ldots, \widetilde{q}_M). \end{aligned}$$

Both laws have the same uniform Z marginal. Therefore their $H(Z) = \log M$ terms cancel, and

$$|I_Q(Z;Y) - I_P(Z;Y)| = |H_P(Z \mid Y) - H_Q(Z \mid Y)| \leq b_M(\varepsilon),$$

by Eq. (8) and its extension above. The dimension in the bound is the teacher-index dimension M, not the token-vocabulary dimension. The proof concerns deterministic paired distributions and never assumes that the drafts were independently sampled or are parameter-posterior draws. □

Proof of the distillation-loss form.

Proof. For completeness, the needed form of Pinsker's inequality follows by reducing to a binary event. For distributions p, q, let $A = \{v : p(v) \geq q(v)\}$, $a = p(A)$, and $b = q(A)$. Then $a - b = \mathrm{TV}(p, q)$. The log-sum inequality gives

$$\mathrm{KL}(p \| q) \geq d(a \| b) := a \log \frac{a}{b} + (1-a) \log \frac{1-a}{1-b}.$$

[1] Mohammad A. Alhejji and Graeme Smith, *A Tight Uniform Continuity Bound for Equivocation*, arXiv:1909.00787, 2020 version. `https://arxiv.org/abs/1909.00787`. The cited bound is converted from base-two logarithms to natural logarithms here.

For $0 < b < 1$, $d(b\|b) = 0$, its first derivative with respect to a vanishes at $a = b$, and $\partial_a^2 d(a\|b) = 1/[a(1-a)] \geq 4$. Integrating this second-derivative bound gives $d(a\|b) \geq 2(a-b)^2$. Boundary cases follow by limits or are trivial when KL is infinite. Consequently,

$$\mathrm{TV}(p, q) \leq \sqrt{\mathrm{KL}(p\|q)/2}.$$

Apply this inequality to every pair and then use concavity of the square root:

$$\varepsilon \leq \frac{1}{M}\sum_i \sqrt{\mathrm{KL}(p_i\|\widetilde{q}_i)/2} \leq \sqrt{\frac{1}{2M}\sum_i \mathrm{KL}(p_i\|\widetilde{q}_i)} = \sqrt{\mathcal{L}_{\mathrm{pair}}/2}.$$

Together with $\varepsilon \leq 1$ and monotonicity of b_M, this proves the second inequality in Eq. (4). An infinite $\mathcal{L}_{\mathrm{pair}}$ yields only the trivial bound $\log M$; the total-variation statement still applies. For $M = 1$, both teacher and paired-draft JSDs are identically zero and no continuity bound is needed. □

Averaged and grouped predictions. The theorem applies to any specified correspondence between teacher and draft predictive distributions. For $\widetilde{q}_i = \bar{q}_i = R^{-1}\sum_j q_{ij}$, convexity of total variation and of KL in its second argument gives

$$\frac{1}{M}\sum_i \mathrm{TV}(p_i, \bar{q}_i) \leq \frac{1}{MR}\sum_{i,j} \mathrm{TV}(p_i, q_{ij}),$$
$$\mathcal{L}_{\mathrm{pair}} = \frac{1}{M}\sum_i \mathrm{KL}(p_i\|\bar{q}_i) \leq \frac{1}{MR}\sum_{i,j} \mathrm{KL}(p_i\|q_{ij}).$$

The first inequality is the triangle inequality applied coordinatewise; the second follows from convexity of $-\log$ before summing under p_i. Thus fidelity of individual drafts to their assigned teacher is a sufficient, but stronger, condition than fidelity of group means. Neither inequality says that arbitrary data partitions create this teacher correspondence or make its error small.

Proof of Proposition 2.5.

Proof. The entropy form of the JSD of all MR drafts is

$$J_{\mathrm{flat}} = H(q_{\mathrm{mix}}) - \frac{1}{MR}\sum_{i,j} H(q_{ij}).$$

Add and subtract $M^{-1}\sum_i H(\bar{q}_i)$ to obtain

$$\begin{aligned} J_{\mathrm{flat}} &= \left[H(q_{\mathrm{mix}}) - \frac{1}{M}\sum_i H(\bar{q}_i)\right] + \frac{1}{M}\sum_i \left[H(\bar{q}_i) - \frac{1}{R}\sum_j H(q_{ij})\right] \\ &= J_{\mathrm{between}} + J_{\mathrm{within}}. \end{aligned}$$

Applying the draft-side identity in Proposition 2.2 first to all base drafts and then to their group means gives

$$S_{\mathrm{flat},r} = J_{\mathrm{between}} + J_{\mathrm{within}} + \mathrm{KL}(q_{\mathrm{mix}}\|r),$$
$$S_{\mathrm{group},r} = J_{\mathrm{between}} + \mathrm{KL}(q_{\mathrm{mix}}\|r).$$

This proves Eq. (5). The same decomposition has an information-theoretic interpretation. Draw a uniform group index Z and a uniform within-group index W, then draw Y from q_{ZW}. The chain rule $I((Z, W); Y) = I(Z; Y) + I(W; Y \mid Z)$ identifies the flat, between-group, and within-group JSDs, respectively. □

Proof of the hierarchical error bound.

Proof. Let $B_r = \mathrm{KL}(q_{\mathrm{mix}}\|r)$. With $\widetilde{q}_i = \bar{q}_i$, Theorem 2.4 gives $|J_{\mathrm{between}} - U_T^{(M)}| \leq b_M(\varepsilon)$. The exact decomposition therefore implies the two-sided statement

$$J_{\mathrm{within}} + B_r - b_M(\varepsilon) \leq S_{\mathrm{flat},r} - U_T^{(M)} \leq J_{\mathrm{within}} + B_r + b_M(\varepsilon).$$

Since J_{within} and B_r are nonnegative, this proves Eq. (6). For $S_{\text{group},r}$, the corresponding interval is $[B_r - b_M(\varepsilon), B_r + b_M(\varepsilon)]$. In particular, if $\bar{q}_i = p_i$ and $r = p_T^{(M)}$, then

$$S_{\text{group},r} = U_T^{(M)}, \qquad S_{\text{flat},r} = U_T^{(M)} + J_{\text{within}}.$$

The within-group term vanishes exactly when all positive-weight drafts within each group have the same predictive distribution. Nonzero within-group variation is additional to the matched finite-teacher signal; between-group variation can itself contain approximation error when the group means do not match their teachers. □

Unequal groups and weights. For group weights $w_i > 0$ summing to one and within-group weights $a_{ij} > 0$ summing to one for each i, set $\bar{q}_i = \sum_j a_{ij} q_{ij}$ and $q_{\text{mix}} = \sum_i w_i \bar{q}_i$. The same add-and-subtract proof yields

$$\sum_{i,j} w_i a_{ij} \,\mathrm{KL}(q_{ij} \| r) = \sum_i w_i \,\mathrm{KL}(\bar{q}_i \| q_{\text{mix}}) + \sum_{i,j} w_i a_{ij} \,\mathrm{KL}(q_{ij} \| \bar{q}_i) + \mathrm{KL}(q_{\text{mix}} \| r).$$

If teacher and draft group weights agree, the transfer proof also holds for the corresponding weighted JSDs, with $P(i,v) = w_i p_i(v)$, $Q(i,v) = w_i \bar{q}_i(v)$, and $\varepsilon = \sum_i w_i \,\mathrm{TV}(p_i, \bar{q}_i)$: their Z marginals still coincide, and Eq. (8) uses the same cardinality M. Uniform weighting over base drafts with unequal group sizes is not uniform weighting over teachers. The weights must therefore be stated when connecting either score to a teacher reference.

A.3 Justification of the Low-Rank Gaussian Target Family

The perturbed teacher family specifies a structured reference for local predictive variation, rather than exact samples from an inferred Bayesian posterior. The following derivation separates the exact Gaussian law induced by the perturbation rule from the additional assumptions needed for a Laplace interpretation.

Bayesian posterior prediction. For training data D and parameter posterior $p(\theta \mid D)$, Bayesian prediction at a fixed covariate or prefix h averages $p_\theta(\cdot \mid h)$ against the chosen parameter posterior (Neal, 1996). For any specified reference π_T, the analogous predictive mean and disagreement are

$$p_T(\cdot \mid h) = \int p_\theta(\cdot \mid h)\, \pi_T(d\theta), \qquad U_T(h) = H(p_T(\cdot \mid h)) - \int H(p_\theta(\cdot \mid h))\, \pi_T(d\theta).$$

These quantities are mathematically well defined whether or not π_T approximates the true posterior. Their interpretation as Bayesian uncertainty is conditional on the adequacy of that reference.

Full and projected quadratic approximations. Let $\mathcal{L}(\theta) = -\log p(D \mid \theta) - \log p(\theta)$ be twice continuously differentiable near θ_0. Set $g = \nabla \mathcal{L}(\theta_0)$ and $H = \nabla^2 \mathcal{L}(\theta_0)$. For a fixed full-column-rank basis $B \in \mathbb{R}^{d \times m}$ and $\theta = \theta_0 + Bz$, Taylor expansion gives

$$\mathcal{L}(\theta_0 + Bz) = \mathcal{L}(\theta_0) + b^\top z + \tfrac{1}{2} z^\top G z + o(\|z\|^2), \qquad b = B^\top g, \quad G = B^\top H B. \tag{9}$$

The linear term cannot in general be dropped. If $G \succ 0$, completing the square in the quadratic part gives

$$b^\top z + \tfrac{1}{2} z^\top G z = \tfrac{1}{2}(z + G^{-1}b)^\top G (z + G^{-1}b) - \tfrac{1}{2} b^\top G^{-1} b.$$

Neglecting the Taylor remainder therefore gives the local Gaussian approximation on the restricted subspace,

$$z \sim \mathcal{N}(-G^{-1}b, G^{-1}),$$
$$\theta \sim \mathcal{N}\big(\theta_0 - BG^{-1}b,\ BG^{-1}B^\top\big).$$

This is a Gaussian law supported on the affine subspace, not a full-dimensional density when $m < d$. A Gaussian centered at θ_0 additionally requires projected stationarity $B^\top \nabla \mathcal{L}(\theta_0) = 0$. In particular, if $B = I$, θ_0 is a stationary local MAP solution, and $H \succ 0$, the usual full Laplace covariance is H^{-1}. Positive semidefiniteness alone is insufficient to invert the curvature. A damped quadratic model with precision $G + \lambda I \succ 0$ would have mean $-(G + \lambda I)^{-1} b$ in the z coordinates; damping changes the quadratic model rather than proving that the original posterior has this covariance.

Exact law of the isotropic low-rank perturbation. For a fixed teacher matrix $W_0 \in \mathbb{R}^{d_{\text{out}} \times d_{\text{in}}}$, let $U_r \in \mathbb{R}^{d_{\text{out}} \times r}$ contain its leading left singular vectors, with $U_r^\top U_r = I_r$. The perturbation rule is

$$\widetilde{W} = W_0 + \sigma U_r E, \qquad E_{ab} \overset{\text{iid}}{\sim} \mathcal{N}(0,1), \qquad E \in \mathbb{R}^{r \times d_{\text{in}}}.$$

With column-wise vectorization, the identity $\operatorname{vec}(AE) = (I \otimes A)\operatorname{vec}(E)$ implies

$$\operatorname{vec}(\widetilde{W}) = \operatorname{vec}(W_0) + \sigma(I_{d_{\text{in}}} \otimes U_r)\operatorname{vec}(E),$$
$$\operatorname{vec}(\widetilde{W}) \sim \mathcal{N}\big(\operatorname{vec}(W_0),\ \sigma^2(I_{d_{\text{in}}} \otimes U_r U_r^\top)\big).$$

The Gaussian statement is exact conditional on W_0 and the selected basis. Although each matrix perturbation has rank at most r, its vectorized covariance has rank $r\, d_{\text{in}}$ for $\sigma > 0$. Also,

$$\mathbb{E}\|\widetilde{W} - W_0\|_F^2 = \sigma^2 \mathbb{E}\|U_r E\|_F^2 = \sigma^2 \mathbb{E}\|E\|_F^2 = \sigma^2 r\, d_{\text{in}}.$$

Thus the basis restricts the output directions of weight perturbations, while σ controls their scale.

What a Laplace interpretation would require. Taking $B = I_{d_{\text{in}}} \otimes U_r$, the centered isotropic coordinate law corresponds to a projected quadratic model only under the additional approximations $b = 0$ and $G \approx \sigma^{-2} I$. The perturbation construction itself does not establish either condition. It does not estimate a Hessian, Fisher matrix, or generalized Gauss–Newton matrix, and weight singular vectors are not in general posterior-curvature eigenvectors. We therefore use this distribution to define the reference π_T, without identifying its covariance with an estimated Bayesian covariance.

A finite family $\theta_1, \dots, \theta_M$ drawn from this reference has predictive mean $p_T^{(M)} = M^{-1} \sum_i p_{\theta_i}$ and uncertainty $U_T^{(M)}$. Theorem 2.4 controls transfer from that fixed family to the corresponding draft predictions. The choice of reference, finite sampling from it, and compression into drafts are distinct sources of discrepancy; a bound on the last does not by itself bound the first two. The KL decompositions remain exact for this reference regardless of its posterior interpretation.

A.4 OSD, Proxy Approximation, and Mean Collapse

Proof of Proposition 2.6.

Proof. Let $\mathcal{D}$ be a fixed distribution over prefixes. Independently draw $\theta \sim \pi_T$ at each prefix and put $r_\phi = p_{\text{mix},\phi}$. Applying the teacher-side identity at each h and then averaging gives

$$\begin{aligned}\mathcal{L}_{\text{OSD}}(\phi) &= \mathbb{E}_{h\sim\mathcal{D}}\mathbb{E}_{\theta\sim\pi_T} \operatorname{KL}(p_\theta(\cdot \mid h)\|r_\phi(\cdot \mid h)) \\ &= \mathbb{E}_{h\sim\mathcal{D}} U_T(h) + \mathbb{E}_{h\sim\mathcal{D}} \operatorname{KL}(p_T(\cdot \mid h)\|r_\phi(\cdot \mid h)).\end{aligned}$$

The first summand is independent of ϕ. Accordingly, if a minimizer exists in the chosen model family, it minimizes the second summand. Define the representation error and population excess risk by

$$a_{\text{rep}} := \inf_\phi \mathbb{E}_{h\sim\mathcal{D}} \operatorname{KL}(p_T\|r_\phi),$$
$$e_{\text{exc}}(\phi) := \mathcal{L}_{\text{OSD}}(\phi) - \inf_\psi \mathcal{L}_{\text{OSD}}(\psi).$$

For finite risks, the identity implies

$$\mathbb{E}_{h\sim\mathcal{D}} \operatorname{KL}(p_T\|r_\phi) = a_{\text{rep}} + e_{\text{exc}}(\phi).$$

Both terms are nonnegative. Over unrestricted conditional distributions, $r = p_T$ is a minimizer. For a parameterized family, the same conclusion requires realizability and zero population excess risk, and gives equality only $\mathcal{D}$-almost everywhere. The identity does not assert convergence of a particular optimizer, attainment of the infimum, or equality of population and empirical training risks. If teacher sampling instead depends on h, the same proof uses the actual conditional teacher law $\pi_T(d\theta \mid h)$ and its corresponding conditional mean, rather than an unrelated unconditional mean. □

Why the same mean objective can remove draft disagreement. Suppose all K drafts minimize the same unregularized population forward-KL objective against the same stochastic teacher law and the same prefix distribution. If every model family realizes p_T and each attained solution has zero population excess risk, the preceding identity gives $\mathrm{KL}(p_T \| q_k) = 0$ almost everywhere for every k. Intersecting these K full-measure sets yields

$$q_1 = \cdots = q_K = p_T, \qquad J_D = 0 \quad \text{for } \mathcal{D}\text{-almost every prefix.}$$

This does not force $U_T = 0$: the nonidentical teachers in Appendix A.1 provide a counterexample. Mean approximation is therefore the appropriate objective for the OSD reference, but is not by itself a mechanism for preserving a family of teacher-conditioned predictions. Conversely, Theorem 2.4 shows that small teacher-specific prediction error suffices to preserve the finite-reference JSD. These are statements about objectives and predictive fidelity, not about any initialization or partitioning rule automatically satisfying them.

Stability with respect to the proxy reference. For fixed drafts, the effect of replacing p_T by r is exactly

$$S_r - S_{p_T} = \sum_v q_{\mathrm{mix}}(v) \log \frac{p_T(v)}{r(v)}.$$

Consequently, a relative-log error bound on the draft support gives

$$\sup_{v : q_{\mathrm{mix}}(v) > 0} \left| \log \frac{p_T(v)}{r(v)} \right| \le \eta \quad \Longrightarrow \quad |S_r - S_{p_T}| \le \eta. \tag{10}$$

A sufficient additive-probability condition is an explicit floor $p_T(v), r(v) \ge \beta > 0$ on that support. The mean value theorem for $\log$ and Pinsker's inequality then yield

$$\begin{aligned} |S_r - S_{p_T}| &\le \frac{1}{\beta} \sum_v q_{\mathrm{mix}}(v) |p_T(v) - r(v)| \\ &\le \frac{1}{\beta} \|p_T - r\|_1 = \frac{2}{\beta} \, \mathrm{TV}(p_T, r) \le \frac{\sqrt{2 \, \mathrm{KL}(p_T \| r)}}{\beta}. \end{aligned}$$

If the same floor holds $\mathcal{D}$-almost everywhere, averaging and applying Jensen's inequality gives

$$\mathbb{E}_{h \sim \mathcal{D}} |S_{r_\phi}(h) - S_{p_T}(h)| \le \frac{\sqrt{2(a_{\mathrm{rep}} + e_{\mathrm{exc}}(\phi))}}{\beta}.$$

The floor is an additional sufficient condition for this proxy-stability bound, not an assumption of the vocabulary-independent JSD transfer theorem.

Why small mean-fitting KL alone is not enough. To see the need for support control in such a uniform statement, let $a \downarrow 0$ and define strictly positive binary distributions

$$p_T^{[a]} = (1 - a, a), \qquad r^{[a]} = (1 - e^{-1/\sqrt{a}}, e^{-1/\sqrt{a}}), \qquad q_{\mathrm{mix}} = (1/2, 1/2).$$

Their forward KL satisfies

$$\mathrm{KL}(p_T^{[a]} \| r^{[a]}) = (1 - a) \log \frac{1 - a}{1 - e^{-1/\sqrt{a}}} + a \log a + \sqrt{a} \longrightarrow 0.$$

Nevertheless, for any fixed draft family with that mixture,

$$S_{r^{[a]}} - S_{p_T^{[a]}} = \tfrac{1}{2} \log \frac{1 - a}{1 - e^{-1/\sqrt{a}}} + \tfrac{1}{2} \log a + \frac{1}{2\sqrt{a}} \longrightarrow +\infty.$$

The OSD loss controls an expectation weighted by p_T; the full-score reference error is weighted by q_{mix}. This counterexample isolates that difference without contradicting either the OSD objective identity or the paired-JSD bound.

A.5 Finite-Ensemble Bias and the Population Reference

The next result concerns an ideal sampling regime, not arbitrary fixed distilled drafts. In particular, its reference mean is the population p_T, not the mean of the very same finite sample.

Proposition A.1 (Ideal-sampling bias correction)**.** *Fix a prefix. Let $\theta_1, \dots, \theta_K$ be independent draws from π_T, set $q_k = p_{\theta_k}$, and use $r = p_T = \mathbb{E}_\theta p_\theta$. Write $\bar{p}_K = K^{-1} \sum_k p_{\theta_k}$ and $J_K = \mathrm{JSD}(p_{\theta_1}, \dots, p_{\theta_K})$. Then*

$$\mathbb{E} S_{p_T} = U_T, \qquad \mathbb{E} J_K = U_T - \mathbb{E}\,\mathrm{KL}(\bar{p}_K \| p_T). \tag{11}$$

Expectations here are over the teacher sample.

Proof. Linearity of expectation and Lemma 2.1 give

$$\mathbb{E} S_{p_T} = \frac{1}{K} \sum_k \mathbb{E}\,\mathrm{KL}(p_{\theta_k} \| p_T) = U_T.$$

For every realization, Proposition 2.2 gives $S_{p_T} = J_K + \mathrm{KL}(\bar{p}_K \| p_T)$. Take expectations and rearrange to obtain Eq. (11). The expectations exist: on the finite vocabulary, remove coordinates where $p_T(v) = 0$; on the remaining finite support, p_T has a positive minimum and $\mathrm{KL}(p_\theta \| p_T)$ is bounded. Independence is not needed for these expectation identities alone, but is used for the sampling-rate calculation below. □

A finite-sample bound on the expected correction. Let

$$C_T := \sum_{v : p_T(v) > 0} \frac{\mathrm{Var}_\theta(p_\theta(v))}{p_T(v)}.$$

Then independence implies

$$0 \le U_T - \mathbb{E} J_K = \mathbb{E}\,\mathrm{KL}(\bar{p}_K \| p_T) \le \frac{C_T}{K}. \tag{12}$$

Proof. Using $\log u \le u - 1$, for distributions a, b with the required support,

$$\mathrm{KL}(a \| b) \le \sum_v \frac{(a(v) - b(v))^2}{b(v)}.$$

Apply this with $a = \bar{p}_K$, $b = p_T$. Since $\mathbb{E}\bar{p}_K = p_T$ and $\mathrm{Var}(\bar{p}_K(v)) = K^{-1} \mathrm{Var}_\theta(p_\theta(v))$, taking expectations gives the claimed upper bound. Finiteness also follows directly from $\mathrm{Var}(p_\theta(v)) \le p_T(v)(1 - p_T(v))$, so $C_T \le |\mathrm{supp}(p_T)| - 1$. Unlike Theorem 2.4, this sampling bound is not claimed to be vocabulary independent. □

The added KL is thus an unbiased finite-ensemble correction in this exact-sample, exact-population-reference regime. For distilled drafts, predictive approximation and capacity mismatch also contribute to the same algebraic term, so this sampling interpretation is not automatic. For a fixed finite teacher reference with exact paired drafts and $r = p_T^{(M)}$, the correction instead vanishes and its JSD already equals $U_T^{(M)}$.

Separating reference choice, finite sampling, and distillation. If U_{Bayes} denotes EU under a specified true posterior, then any score S satisfies

$$|S - U_{\mathrm{Bayes}}| \le |S - U_T^{(M)}| + |U_T^{(M)} - U_T| + |U_T - U_{\mathrm{Bayes}}|.$$

The transfer and hierarchical bounds control the first term under the stated predictive-fidelity conditions. Proposition A.1 and Eq. (12) characterize the expectation bias associated with the second term in the independent-sampling regime; an expectation-bias result is not a realized absolute-error guarantee. The last term concerns the choice of π_T itself. No posterior-accuracy claim follows solely from the Gaussian perturbation construction or from accurate distillation of that construction.

A.6 Token-Level and Sequence-Level Quantities

At a fixed prefix, each KL or entropy above is a sum over $\mathcal{V}$, not over all possible output sequences. Given a realized response $y_{1:T}$, an operational sequence score can therefore be defined by

$$S_{\mathrm{seq}}(x, y_{1:T}) = \sum_{t=1}^{T} S_{r_t}(x, y_{<t}).$$

The same pathwise aggregation can be applied to the finite-reference uncertainties. For any pointwise bounds $|S_{r_t}(h_t) - U_T^{(M)}(h_t)| \le c_t(h_t)$, the triangle inequality gives

$$\left| S_{\mathrm{seq}}(x, y_{1:T}) - \sum_{t=1}^{T} U_T^{(M)}(h_t) \right| \le \sum_{t=1}^{T} c_t(h_t).$$

This is a guarantee about scores along the same fixed path.

It differs from the sequence mutual information in the generative model that first draws one $\theta \sim \pi_T$ and then draws the entire sequence conditional on that same θ. In that joint model, observing a prefix changes the parameter law to

$$\pi_t(d\theta \mid x, y_{<t}) = \frac{\prod_{s<t} p_\theta(y_s \mid x, y_{<s})\, \pi_T(d\theta)}{\int \prod_{s<t} p_{\theta'}(y_s \mid x, y_{<s})\, \pi_T(d\theta')}.$$

For positive-probability prefixes, the mutual-information chain rule then uses this updated law and an expectation over random prefixes:

$$I(\theta; Y_{1:T} \mid x) = \sum_{t=1}^{T} \mathbb{E}_{Y_{<t}|x} I_{\pi_t}(\theta; Y_t \mid x, Y_{<t}).$$

A realized-path sum with a fixed reference π_T does not in general equal this quantity. Distinguishing the two avoids an implicit exchange of a mixture of autoregressive sequence distributions with a product of token-level mixtures.

B Additional Experimental Details

B.1 Training configurations

Table 3 summarizes the training hyperparameters for all methods. All models used a batch size of 8 and LoRA adapters with rank $r = 16$, scaling factor $\alpha = 32$, and dropout 0.1. For p_{mix}, LoRA was applied to `q_proj` and `v_proj`. For all other methods, LoRA was applied to `q_proj`, `k_proj`, `v_proj`, `o_proj`, `gate_proj`, `up_proj`, and `down_proj`.

Training p_{mix}. Both p_{mix} models used the same training configuration. For each noise-injected teacher model, we generated four answers per question from GSM8K and ASQA. The three resulting datasets were mixed in equal proportions to form the final training set. For each training batch, we sampled $N_{\text{teacher}} = 2$ noise modules, injected them into the teacher model, and computed the KL loss between the student and each sampled teacher. The injected noise followed $\mathcal{N}(0, 0.1)$, matching the target perturbation.

Training IDD, DDD, and FDD. These methods shared the same optimization settings. For IDD and FDD, noise modules were sampled from $\mathcal{N}(0, 0.5)$ and injected into the draft models. Under IDD, all ensemble draft models were trained on the same dataset; under DDD and FDD, each draft model was trained on a different dataset. Each dataset contained four sampled answers per question from GSM8K and ASQA.

Training GKD and MiniLLM. GKD and MiniLLM used the Hugging Face Distillation Trainer and MiniLLM Trainer, respectively. Both methods used on-policy training, with the student generating its own training data. To limit perturbation of the student's distribution during early training, the initial sampling noise followed $\mathcal{N}(0, 0.1)$. Both methods used vLLM. For GKD, we set $\lambda = 0.5$ and $\beta = 0.5$. For MiniLLM, we generated four completions per prompt and enabled conversational prompting and single-step decomposition. The dataset-specific generation settings were identical across GKD and MiniLLM.

Table 3: Training hyperparameters. All methods used a batch size of 8 and LoRA settings $(r, \alpha, \text{dropout}) = (16, 32, 0.1)$. Generation settings in panel (b) apply to both GKD and MiniLLM. A dash indicates a setting not specified in the reported configuration.

(a) Optimization settings

Hyperparameter	p_{mix}	IDD / DDD / FDD	GKD	MiniLLM
Learning rate	3×10^{-4}	7×10^{-5}	7×10^{-5}	7×10^{-5}
Gradient accumulation steps	8	2	2	2
Epochs	12	1	1	1
Weight decay	0.01	0	0	0
Warmup ratio	0.1	—	—	—
Warmup steps	—	150	150	150
Learning-rate scheduler	—	—	Cosine	Cosine

(b) Generation settings for GKD and MiniLLM

Hyperparameter	GSM8K	ASQA
Maximum generated tokens	400	384
Temperature	1.0	0.7
Top-p	1.0	0.9

B.2 Details of Hallucination Detection Test

For the GSM8K test set, we sample one answer per question from the target model and repeat this process three times. During sampling, we instruct the model to place the final answer inside `\boxed{}`, and assign a binary correctness label by comparing the extracted boxed answer with the GSM8K ground-truth solution.

For ASQA, we first prompt the target model to generate an answer for each ambiguous question using the ASQA answer-generation prompt shown in Appendix B.3. We then use a separate judge model to evaluate the target-generated answer. The judge receives the ambiguous question, the gold disambiguations, and the target model's answer, together with the judge prompt shown in Appendix B.4. It assigns one of four labels: *hallucinated*, *supported*, *abstained*, or *incomplete*. We convert these labels into a binary hallucination label by treating *hallucinated* as positive and all other labels as non-hallucinated.

For both datasets, we compute token-level epistemic uncertainty at every generated token and sum these token-level scores into a sequence-level uncertainty score. Since the resulting uncertainty score is not a calibrated probability, we fit a logistic regression classifier on a held-out split to map sequence-level uncertainty to the probability of an incorrect or hallucinated answer. We then evaluate the classifier on the test split using AUROC, ECE, and Brier score.

B.3 Prompt Used for Ambiguous Question Answering

The following prompt was used to instruct the model to answer intentionally ambiguous questions.

```
Ambiguous question: {ambiguous_question}

This dataset contains intentionally ambiguous questions.
Do not refuse to answer just because the question is ambiguous.
Instead, identify the most plausible interpretations and answer each one
    separately.
For each interpretation, give one short answer sentence with the most
    direct factual answer you can provide.
If you do not know the factual answer for a specific interpretation,
    write "unknown" for that interpretation.
Do not write only "ambiguous", "it depends", or placeholders such as
    "...".
Do not leave any interpretation blank.
Do not add background information, explanations, or extra facts beyond
    the direct answer.
```

```
Answer format:
1. If referring to ..., the answer is ...
2. If referring to ..., the answer is ...
```

B.4 Judge Prompt for ASQA Answer Evaluation

We used a separate judge model to evaluate model-generated answers for ASQA. The judge was given the ambiguous question, the gold disambiguations, and the model-generated answer. It was instructed to assign exactly one of four labels: `hallucinated`, `supported`, `abstained`, or `incomplete`. The judge was also instructed to use only the provided inputs and not to rely on outside knowledge. The full prompt template is shown in Listing 1.

Listing 1: Few-shot judge prompt used for ASQA answer evaluation.

```
"""You are a strict judge for ambiguous-question answering.

Your task is to assign exactly one label to the model-generated answer
    :

- "hallucinated"
- "supported"
- "abstained"
- "incomplete"

Judge using only:
1. the ambiguous question,
2. the gold disambiguations,
3. the model-generated answer.

Do not use outside knowledge.

This dataset contains intentionally ambiguous questions.
Therefore, the model should not refuse to answer merely because the
    question is ambiguous.
If the answer only says the question is ambiguous, leaves placeholders
     such as "...", or names interpretations without giving the
    factual answer for them, treat that as abstention rather than a
    valid answer.

Label definitions:

1. "hallucinated"
Assign this label if the answer contains any factual claim that is
    unsupported by, or contradicts, the gold disambiguations.
If even one unsupported or contradicted factual claim appears, choose
    "hallucinated".

2. "abstained"
Assign this label if the answer avoids making substantive factual
    claims for the interpretations it introduces.
This includes cases where the answer:
- says the question is ambiguous without actually answering the
    interpretations,
- uses placeholders such as "...",
- gives only generic setup text without a factual value,
- or says "unknown" for all listed interpretations.
The key condition is: no unsupported factual claim is made, but the
    model does not actually provide a concrete factual answer.

3. "incomplete"
Assign this label if the answer gives at least one factually supported
     concrete answer to at least one gold disambiguation, makes no
    unsupported factual claim, but covers only part of the gold
```

```
    disambiguations or leaves some plausible interpretations
    unanswered or marked as unknown.

4. "supported"
Assign this label if the answer gives factually supported concrete
    answers, makes no unsupported or contradicted factual claims, and
    adequately covers the relevant gold disambiguations.

Important rules:
1. Hallucination takes priority over all other labels.
2. Do not mark an answer as hallucinated just because it is incomplete
    .
3. Do not mark an answer as hallucinated just because it abstains.
4. Because the dataset questions are intentionally ambiguous, merely
    recognizing ambiguity is not enough for "supported" or "incomplete
    "; the answer must also provide factual values for at least one
    interpretation.
5. A vague or cautious answer can be "abstained" or "incomplete",
    depending on whether it gives any supported factual answer.
6. If the answer gives at least one supported factual answer and
    nothing false, it should usually be "incomplete" rather than "
    abstained".
7. Use "supported" only when the answer is factually safe and
    sufficiently complete relative to the gold disambiguations.
8. If the answer gives one supported interpretation and marks another
    as "unknown", that is usually "incomplete", not "abstained".
9. If the answer lists interpretations but leaves them blank, that is
    "abstained".
10. Return JSON only.

Return exactly this JSON schema:
{
  "answer_label": "hallucinated" or "supported" or "abstained" or "
    incomplete",
  "matched_disambiguations": [indices],
  "unsupported_or_contradicted_claims": [strings],
  "abstention_signals": [strings],
  "reason": "brief explanation"
}

Examples:

Example 1
Ambiguous question:
Who was the ruler of France in 1830?

Gold disambiguations:
[0] If referring to the ruler of France in early 1830, the answer is
    Charles X.
[1] If referring to the ruler of France after the July Revolution in
    late 1830, the answer is Louis Philippe I.

Model-generated answer:
Charles X ruled early in 1830, and Louis Philippe I ruled later in
    1830.

Output:
{
  "answer_label": "supported",
  "matched_disambiguations": [0, 1],
  "unsupported_or_contradicted_claims": [],
  "abstention_signals": [],
  "reason": "The answer is factually supported and adequately covers
    both gold disambiguations."
}
```

```
Example 2
Ambiguous question:
Who was the ruler of France in 1830?

Gold disambiguations:
[0] If referring to the ruler of France in early 1830, the answer is
    Charles X.
[1] If referring to the ruler of France after the July Revolution in
    late 1830, the answer is Louis Philippe I.

Model-generated answer:
Napoleon Bonaparte ruled France in 1830.

Output:
{
  "answer_label": "hallucinated",
  "matched_disambiguations": [],
  "unsupported_or_contradicted_claims": [
    "The answer claims Napoleon Bonaparte ruled France in 1830."
  ],
  "abstention_signals": [],
  "reason": "The answer makes a contradicted factual claim."
}

Example 3
Ambiguous question:
Who was the ruler of France in 1830?

Gold disambiguations:
[0] If referring to the ruler of France in early 1830, the answer is
    Charles X.
[1] If referring to the ruler of France after the July Revolution in
    late 1830, the answer is Louis Philippe I.

Model-generated answer:
Charles X ruled France in 1830.

Output:
{
  "answer_label": "incomplete",
  "matched_disambiguations": [0],
  "unsupported_or_contradicted_claims": [],
  "abstention_signals": [],
  "reason": "The answer is factually supported for one gold
    disambiguation, but it is only partial."
}

Example 4
Ambiguous question:
Who was the ruler of France in 1830?

Gold disambiguations:
[0] If referring to the ruler of France in early 1830, the answer is
    Charles X.
[1] If referring to the ruler of France after the July Revolution in
    late 1830, the answer is Louis Philippe I.

Model-generated answer:
The question is ambiguous, and I do not have enough information to
    answer confidently.

Output:
{
  "answer_label": "abstained",
```

```
  "matched_disambiguations": [],
  "unsupported_or_contradicted_claims": [],
  "abstention_signals": [
    "The answer says the question is ambiguous without supplying
    factual answers.",
    "The answer does not commit to a concrete factual answer."
  ],
  "reason": "The answer avoids making a substantive factual claim and
    instead declines to answer."
}

Example 5
Ambiguous question:
Who was the ruler of France in 1830?

Gold disambiguations:
[0] If referring to the ruler of France in early 1830, the answer is
    Charles X.
[1] If referring to the ruler of France after the July Revolution in
    late 1830, the answer is Louis Philippe I.

Model-generated answer:
Charles X ruled early in 1830, and he remained ruler for the entire
    year.

Output:
{
  "answer_label": "hallucinated",
  "matched_disambiguations": [0],
  "unsupported_or_contradicted_claims": [
    "The answer claims Charles X remained ruler for the entire year."
  ],
  "abstention_signals": [],
  "reason": "The answer contains a supported part, but also an
    unsupported or contradicted factual claim, so hallucination takes
    priority."
}

Example 6
Ambiguous question:
Who was the ruler of France in 1830?

Gold disambiguations:
[0] If referring to the ruler of France in early 1830, the answer is
    Charles X.
[1] If referring to the ruler of France after the July Revolution in
    late 1830, the answer is Louis Philippe I.

Model-generated answer:
1. If referring to early 1830, the answer is Charles X.
2. If referring to late 1830, the answer is unknown.

Output:
{
  "answer_label": "incomplete",
  "matched_disambiguations": [0],
  "unsupported_or_contradicted_claims": [],
  "abstention_signals": [
    "The answer marks one interpretation as unknown."
  ],
  "reason": "The answer gives one supported factual answer and safely
    declines another interpretation, so it is incomplete rather than
    abstained."
}
```

```
Now judge the following case.

Ambiguous question:
{question}

Gold disambiguations:
{disambiguations}

Model-generated answer:
{target_answer}"""
```

C Additional Results

C.1 Uncertainty Estimation with smaller draft

Table 4: **Uncertainty Estimation Performance.** Comparison of RMSE (↓) and Spearman correlation (↑) for the 3B target model and 0.5B draft scorer on GSM8K and ASQA.

		3B (Target) → 0.5B (Draft)	
Dataset	**Method**	**RMSE ↓**	**Spearman ↑**
GSM8K	No Train	2.753 ± 0.007	0.288 ± 0.001
	GKD	1.043 ± 0.007	0.795 ± 0.001
	MiniLLM	1.045 ± 0.007	0.813 ± 0.000
	IDD	4.313 ± 0.012	0.202 ± 0.003
	FDD	3.438 ± 0.011	0.175 ± 0.002
	DDD (Ours)	**0.855 ± 0.006**	**0.844 ± 0.000**
ASQA	No Train	4.659 ± 0.006	0.568 ± 0.002
	GKD	2.633 ± 0.004	0.785 ± 0.002
	MiniLLM	2.588 ± 0.004	0.793 ± 0.002
	IDD	3.486 ± 0.011	0.675 ± 0.000
	FDD	3.436 ± 0.005	0.632 ± 0.002
	DDD (Ours)	**2.104 ± 0.007**	**0.821 ± 0.000**

C.2 Effect of training p_{mix}

Table 5: **Robustness of the p_{mix} Proxy.** RMSE (↓) and Spearman correlation (↑) when using the raw target distribution or the distilled p_{mix} proxy under ten perturbations. Each experiment is repeated five times.

Metric	**Target (raw)**	**p_{mix} (Ours)**
RMSE (↓)	0.0391 ± 0.0056	$\mathbf{0.0137 \pm 0.0013}$
Spearman (↑)	0.9957 ± 0.0002	$\mathbf{0.9969 \pm 0.0003}$