# ONE TRAJECTORY IS ENOUGH WITH WISE STEP SELECTION FOR SEMANTIC UNCERTAINTY QUANTIFICATION IN LARGE LANGUAGE DIFFUSION MODELS

**Anonymous authors**
Paper under double-blind review

## ABSTRACT

This paper deals with the high inference cost of sampling-based semantic uncertainty quantification (UQ) for Large Language Diffusion Models (LLDMs). We ask whether multiple independent generations can be replaced by informative states from a *single* denoising trajectory. To this end, we introduce DIMS (*Distribution-Matched Step Selection*), which learns which diffusion steps best reproduce the semantic variability of independent generations. We show that the single-trajectory approximation error decomposes into *distribution mismatch* and *within-trajectory redundancy*, leading to a simple convex MMD-based objective. DIMS requires only a small unlabeled calibration set and no correctness labels. Across three LLDMs, six datasets, and eight semantic UQ estimators, DIMS is best or tied for best among single-trajectory strategies in $34/40$ settings. Overall, our results show that *sampling itself can be optimized for efficient UQ*. Code will be released upon publication.

## 1 INTRODUCTION

Many uncertainty quantification (UQ) methods for large language models rely on *sampling*. Instead of trusting a single answer, they generate several answers and measure how much these answers disagree (Malinin & Gales, 2021; Kuhn et al., 2023; Lin et al., 2024; Nikitin et al., 2024; Vashurin et al., 2025b). This idea is particularly useful for *semantic* UQ, where uncertainty is measured at the level of meaning rather than exact words. For example, Semantic Entropy, graph-based methods, Kernel Language Entropy, and CoCoA compare several generated answers in a semantic space (Kuhn et al., 2023; Farquhar et al., 2024; Lin et al., 2024; Nikitin et al., 2024; Vashurin et al., 2025b). However, this benefit comes at an important cost: obtaining $M$ samples requires $M$ complete generations. As large language models become more expensive to run, this raises a simple question:

> *Can we keep the benefits of sampling-based semantic uncertainty without paying for many independent generations?*

Large Language Diffusion Models (LLDMs) offer a new opportunity to address this question. Unlike autoregressive models, which generate tokens from left to right, masked LLDMs progressively refine a partially masked sequence through a denoising process (Austin et al., 2021; Sahoo et al., 2024; Lou et al., 2024; Nie et al., 2025; Ye et al., 2025). Recent models such as LLaDA and Dream show that this is becoming a competitive paradigm for language generation (Nie et al., 2025; Ye et al., 2025). Importantly, a diffusion generation does not only produce a final answer. At every denoising step, the model also produces an intermediate estimate of the full answer. A single trajectory therefore naturally provides many candidate answers that are already computed during generation.

This suggests an attractive idea: instead of running the model $M$ times, can we reuse the intermediate answers produced along a *single* diffusion trajectory for semantic UQ? Recent work has explored this direction (Vazhentsev et al., 2026), but also shows an important limitation: directly using all intermediate states as a replacement for independent samples can lead to a substantial loss in UQ performance. We argue that this limitation comes from the structure of the diffusion trajectory itself. Intermediate answers are neither independent nor identically distributed. Early states can be far from the distribution of final answers, while late states are often very similar and therefore provide



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.

little additional information. Moreover, because all states belong to the same denoising trajectory, they are strongly dependent. As a result, *not all diffusion steps are equally useful for uncertainty estimation*. This motivates our main idea: rather than using the trajectory as it is, we learn which intermediate states should be selected so that a single trajectory provides the best possible proxy for multiple independent generations.

This leads to the main question studied in this work:

> *Which intermediate states should be selected from a single diffusion trajectory so that they best reproduce the semantic variability of independent generations?*

Our key idea is that *sampling itself can be optimized*. Existing semantic UQ methods mainly focus on how uncertainty should be computed once a set of answers is available. In contrast, we study how this set of answers should be constructed. In LLDMs, the denoising trajectory provides a large set of candidate answers at almost no additional generation cost. The challenge is to select states that are both representative of the final-answer distribution and sufficiently non-redundant.

To this end, we introduce DIMS (*Distribution-Matched Step Selection*). DIMS learns a distribution over diffusion steps such that the selected intermediate answers approximate the *semantic distribution* of independently generated final answers, without paying for the repeated LLDM calls. We formulate this problem using Kernel Mean Embeddings (KME) and Maximum Mean Discrepancy (MMD). Importantly, matching the marginal distributions alone is not sufficient: states selected from the same trajectory can still be highly redundant. We therefore explicitly account for the dependence between intermediate states. Our analysis shows that the error of using a single trajectory can be decomposed into two terms: a *distribution mismatch* term and a *within-trajectory redundancy* term. This gives a principled objective for learning which diffusion steps should be used.

The resulting method is simple and lightweight. DIMS learns a single set of temporal weights offline from a small calibration set, without requiring correctness labels or external reference answers, unlike supervised post-hoc calibration methods. At test time, the LLDM is run only once. DIMS then selects a small set of intermediate answers from this trajectory and feeds them to any standard set-based semantic UQ estimator. Thus, DIMS does not replace existing semantic UQ methods. Instead, it provides them with a more informative set of samples obtained from a single trajectory, at a limited cost. In this sense, our approach separates two questions that are usually treated together: *which samples should be used* and *how uncertainty should be computed from these samples*.

We evaluate DIMS on three instruction-tuned LLDMs, six datasets, different decoding schedules, and eight semantic UQ estimators. Across these settings, DIMS provides the strongest or tied strongest single-trajectory results in most model–estimator configurations and consistently improves over simple choices such as using all, random, or only the last diffusion steps. The results also show an important limit: one trajectory does not fully replace independent sampling for some uncertainty estimators. However, for methods such as CoCoA and AD, a carefully selected single trajectory can approach the performance obtained with multiple independent generations while requiring much less inference time. These results suggest that efficient UQ is not only a question of designing better uncertainty scores: *the sampling strategy itself is also an important part of uncertainty estimation*.

Our contributions are as follows: (1) **We formulate single-trajectory semantic UQ as a sampling problem.** We ask which diffusion states can best replace the multiple independent generations required by standard semantic UQ. (2) **We introduce DIMS and provide theoretical insights based on distribution matching.** We derive an exact decomposition of the approximation error into distribution mismatch and within-trajectory redundancy, and show that minimizing it is equivalent, up to a constant, to approximating $M$ independent generations. This leads to a simple convex MMD-based objective. (3) **We show that optimized sampling is both effective and efficient.** DIMS is best or tied best among single-trajectory strategies in $34/40$ settings. For CoCoA, it comes within $0.021$–$0.044$ AUROC of $20$ independent generations while reducing inference time from about $66.8$ s to $3.5$ s per prompt.

## 2 Related Work

**Semantic uncertainty.** UQ for autoregressive LLMs often relies on token- or sequence-level scores (Malinin & Gales, 2021; Fomicheva et al., 2020; Fadeeva et al., 2023; Aichberger et al., 2026).

Semantic UQ instead measures disagreement between multiple generations, using Semantic Entropy (Kuhn et al., 2023; Farquhar et al., 2024; Hoche et al., 2025; 2026), graph- or kernel-based scores (Lin et al., 2024; Nikitin et al., 2024), or CoCoA (Vashurin et al., 2025b). Most semantic methods, however, still require several complete generations. Probes (Han et al., 2024; Hoche et al., 2025) reduce this cost through supervision; in contrast, our method keeps the downstream UQ estimator unchanged and requires no labels or trained probe.

**Diffusion language models.** Masked diffusion language models (Austin et al., 2021; Lou et al., 2024; Sahoo et al., 2024; Shi et al., 2024), including LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025), generate text through iterative denoising rather than left-to-right decoding. Recent large-scale models (Bie et al., 2025; Fu et al., 2026; DiffusionGemma Team, 2026) make efficient UQ for this new generation process increasingly important.

**UQ for diffusion language models.** UQ for LLDMs is still recent. Existing work adapts likelihood-based scores (Nie et al., 2025; Vazhentsev et al., 2026) or learns detectors from denoising traces (Chang et al., 2026; Qian et al., 2026). Most related to us, Vazhentsev et al. (2026) show that a single diffusion trajectory contains useful uncertainty information and propose several trajectory-based scores. Importantly, their results also show a limitation: directly using all intermediate states as samples for standard semantic UQ can lead to a clear performance drop. Our work starts from this limitation. Rather than designing another uncertainty score, we ask *which diffusion states should be used as samples*. DiMS learns a distribution-matched and dependence-aware step selection so that one trajectory better approximates the semantic variability of independent generations.

## 3 Background

Let $\mathcal{V}$ be the token vocabulary, $\texttt{[MASK]} \notin \mathcal{V}$ a special mask token. A prompt is written $\boldsymbol{x}$ and an answer $\mathbf{y} = [\mathrm{y}^1, \dots, \mathrm{y}^L] \in \mathcal{V}^L$, superscripts indexing token positions; $L$ is fixed for readability and variable length is handled by an end-of-sequence token. We denote $f_{\boldsymbol{\theta}}$ a neural network with parameters $\boldsymbol{\theta}$ and $f^{\ell}_{\boldsymbol{\theta}}(\cdot)$ the estimated token distributions at position $\ell$. For a finite set $\mathcal{A}$, $\Delta(\mathcal{A})$ is the simplex over $\mathcal{A}$, i.e. the set of probability distributions over $\mathcal{A}$. Notation is summarized in App. A.

### 3.1 Large language diffusion models

LLDMs generate text by iteratively unmasking a fully masked sequence (Austin et al., 2021; Sahoo et al., 2024; Lou et al., 2024; Nie et al., 2025). Generation proceeds over a step grid $\{T, \dots, 0\}$ through partially masked *states* $\boldsymbol{z}_t \in (\mathcal{V} \cup \{\texttt{[MASK]}\})^L$, from $\boldsymbol{z}_T = (\texttt{[MASK]}, \dots, \texttt{[MASK]})$ to $\boldsymbol{z}_0 = \mathbf{y}$. Following prior foundational works in discrete-space diffusion (Austin et al., 2021; Sahoo et al., 2024; Shi et al., 2024), we adopt the *mean-parametrization*: the LLDM is made of a neural network $f_{\boldsymbol{\theta}}$ taking a prompt $\boldsymbol{x}$ and a partially masked state $\boldsymbol{z}_t$ and returning logits for all $L$ token positions, trying to predict the clean answer $\mathbf{y}$. At step $t$, we write $\mathcal{M}_t = \{\ell : z_t^{\ell} = \texttt{[MASK]}\}$ the set of masked positions and define the *intermediate answer estimate* at step $t$ as the answer $\tilde{\mathbf{y}}_t = [\tilde{\mathrm{y}}_t^1, \dots, \tilde{\mathrm{y}}_t^L] \in \mathcal{V}^L$ obtained by sampling the masked positions and keeping the committed ones:

$$\forall \ell \in \mathcal{M}_t, \quad \tilde{\mathrm{y}}_t^{\ell} \sim f^{\ell}_{\boldsymbol{\theta}}\big(\cdot \mid \boldsymbol{x}, \boldsymbol{z}_t\big) \; ; \qquad \forall \ell \notin \mathcal{M}_t, \quad \tilde{\mathrm{y}}_t^{\ell} = z_t^{\ell}. \tag{1}$$

LLDMs typically use this intermediate answer estimate $\tilde{\mathbf{y}}_t$ to compute the next state $\boldsymbol{z}_{t-1}$:

$$\boldsymbol{z}_{t-1} \sim \mathcal{G}\big(\cdot \mid \boldsymbol{z}_t, \tilde{\mathbf{y}}_t\big). \tag{2}$$

In the masked diffusion setting we consider here, $\mathcal{G}$ is typically a partial remasking of the intermediate answer estimate $\tilde{\mathbf{y}}_t$, following a masking schedule that gradually reveals more positions. Practical instantiations of LLDMs may use different procedures for $\mathcal{G}$ (Nie et al., 2025; Ye et al., 2025; Wu et al., 2026). Our work is agnostic to the details of $\mathcal{G}$ beyond the fact that the masking schedule is such that every generation has $T$ steps and terminates at a fully unmasked $\boldsymbol{z}_0$. Since $\boldsymbol{z}_0$ is the terminal state, $\tilde{\mathbf{y}}_t$ is indexed in $\mathcal{T} = \{T, \dots, 1\}$. Thus, a single generation produces a whole *trajectory of intermediate answers* $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_T, \dots, \tilde{\mathbf{y}}_1] \in (\mathcal{V}^L)^T$, whose *joint* distribution is denoted $\bar{\pi}_{\boldsymbol{\theta}}(\tilde{\mathbf{Y}} \mid \boldsymbol{x})$. For each denoising step $t$, we denote by $\pi_{\boldsymbol{\theta},t}(\tilde{\mathbf{y}}_t \mid \boldsymbol{x})$ the *marginal* distribution of the intermediate answer at step $t$, obtained by marginalizing over all the other intermediate answers in the trajectory: $\pi_{\boldsymbol{\theta},t}(\tilde{\mathbf{y}}_t \mid \boldsymbol{x}) = \sum_{\{\tilde{\mathbf{y}}_s : s \neq t\}} \bar{\pi}_{\boldsymbol{\theta}}\big(\tilde{\mathbf{y}}_T, \dots, \tilde{\mathbf{y}}_1 \mid \boldsymbol{x}\big)$. We also define the model distribution $p^{\mathrm{Diff}}_{\boldsymbol{\theta}}(\mathbf{y} \mid \boldsymbol{x})$ as the marginal distribution of the final answer produced by the denoising process induced at $t = 0$. By construction, $\pi_{\boldsymbol{\theta},1}(\cdot \mid \boldsymbol{x}) = p^{\mathrm{Diff}}_{\boldsymbol{\theta}}\big(\cdot \mid \boldsymbol{x}\big)$.

### 3.2 SEMANTIC UNCERTAINTY FROM MULTIPLE GENERATIONS

Two answers to the same prompt may differ in wording while carrying the same meaning. Semantic UQ compares answers through a semantic similarity or equivalence function $s$, typically based on a bidirectional natural language inference (NLI) score (Kuhn et al., 2023) or on some similarity function between sequence embeddings. A *sample-based* semantic uncertainty score $U_{\text{sem}}$ maps a set of answers to a scalar that grows with their semantic dispersion. Examples include Semantic Entropy (Kuhn et al., 2023), graph-based scores (Lin et al., 2024), Kernel Language Entropy (Nikitin et al., 2024) and the consistency term of CoCoA (Vashurin et al., 2025b). Standard practice draws

$$\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)} \overset{\text{i.i.d.}}{\sim} p_{\boldsymbol{\theta}}^{\text{Diff}}(\cdot \mid \boldsymbol{x}), \qquad U(\boldsymbol{x}) = U_{\text{sem}}\big(\{\mathbf{y}^{(m)}\}_{m=1}^{M}\big), \tag{3}$$

at a cost of $M$ full generations, i.e. $M\,T$ network calls for an LLDM. Details about these methods are provided in App. C.4. Although $U_{\text{sem}}$ can be evaluated on any set of answers, its interpretation as the uncertainty of $p_{\boldsymbol{\theta}}^{\text{Diff}}(\cdot \mid \boldsymbol{x})$ relies on the samples being representative of this distribution. DIMS keeps $U_{\text{sem}}$ unchanged and replaces independent generations with intermediate answers from denoising trajectories, selected so that their distribution approximates $p_{\boldsymbol{\theta}}^{\text{Diff}}(\cdot \mid \boldsymbol{x})$.

## 4 DIMS: DISTRIBUTION-MATCHED STEP SELECTION

Standard semantic UQ requires $M$ independent generations, leading to roughly $MT$ model evaluations for an LLDM with $T$ denoising steps. We introduce DIMS (***Distribution-Matched Step Selection***) to address this issue. Given a trajectory of intermediate answers $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_T, \dots, \tilde{\mathbf{y}}_1] \sim \bar{\pi}_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})$, DIMS learns to pick $M < T$ intermediate answers among $\tilde{\mathbf{Y}}$, and computes any semantic UQ score on this selection, at a cost of only $T$ model evaluations. The selected answers are chosen so that their semantic distribution is close to the distribution of independently generated final answers with respect to their maximum mean discrepancy (MMD). The selection rule is learned once on a small calibration set. At test time, DIMS requires only one LLDM trajectory.

### 4.1 A MIXTURE OF DIFFUSION STEPS

Let $\mathbf{w} = [w_T, \dots, w_1] \in \Delta(\mathcal{T})$ be a vector in the simplex. Define the corresponding mixture of intermediate answer distributions as

$$\pi_{\mathbf{w}}(\cdot \mid \boldsymbol{x}) = \sum_{t=1}^{T} w_t \pi_{\boldsymbol{\theta},t}(\cdot \mid \boldsymbol{x}). \tag{4}$$

To sample from this distribution, we first sample a denoising step $t \sim \text{Cat}(\cdot; \mathbf{w})$, then sample a trajectory $\tilde{\mathbf{Y}} \sim \bar{\pi}_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})$ and return the intermediate answer $\tilde{\mathbf{y}}_t$ at step $t$. Sampling in an i.i.d. fashion from this mixture distribution requires $M$ independent trajectories, that is approximately $MT$ model evaluations, the same prohibitive cost we want to avoid.[1] Instead, at test time, we sample a single trajectory $\tilde{\mathbf{Y}} \sim \bar{\pi}_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})$ and sample answers according to the following empirical distribution, where $\delta_{\mathbf{y}}(\cdot)$ is the Dirac distribution centered on $\mathbf{y}$:

$$\hat{\pi}_{\mathbf{w}}(\cdot \mid \tilde{\mathbf{Y}}) = \sum_{t=1}^{T} w_t \delta_{\tilde{\mathbf{y}}_t}(\cdot). \tag{5}$$

### 4.2 MATCHING THE DISTRIBUTIONS

To compare the empirical mixture distribution of intermediate answers $\hat{\pi}_{\mathbf{w}}(\cdot \mid \tilde{\mathbf{Y}})$ in (5) and the model distribution $p_{\boldsymbol{\theta}}^{\text{Diff}}\big(\mathbf{y} \mid \boldsymbol{x}\big)$, we use the Maximum Mean Discrepancy (Gretton et al., 2012).

**Kernel mean embedding** We are interested in matching distributions in a semantic way. Let $e : \mathcal{V}^L \to \mathbb{R}^d$ be a semantic embedding function, and let $\psi : \mathbb{R}^d \times \mathbb{R}^d \to \mathbb{R}$. We define the semantic kernel $k : \mathcal{V}^L \times \mathcal{V}^L \to \mathbb{R}$ as

$$k(\mathbf{y}, \mathbf{y}') = \psi\big(e(\mathbf{y}), e(\mathbf{y}')\big). \tag{6}$$

[1] Actually, we could stop each independent trajectory at the step needed, saving a few model evaluations.

The value $k(\mathbf{y}, \mathbf{y}')$ should be large when the answers $\mathbf{y}$ and $\mathbf{y}'$ have similar meanings, and small when their meanings differ. Let $\mathcal{H}_k$ be the Reproducing Kernel Hilbert Space (RKHS) associated with $k$, and let $\phi(\mathbf{y}) = k(\mathbf{y}, \cdot) \in \mathcal{H}_k$ be the feature map. For a fixed prompt $\boldsymbol{x}$, we define the following kernel mean embeddings (KME) (Muandet et al., 2017):

$$\boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x}) = \mathbb{E}_{\mathbf{y}\sim p_{\boldsymbol{\theta}}^{\text{Diff}}(\cdot|\boldsymbol{x})}\left[\phi(\mathbf{y})\right] \quad \boldsymbol{\mu}_t(\boldsymbol{x}) = \mathbb{E}_{\tilde{\mathbf{y}}_t\sim\pi_{\boldsymbol{\theta},t}(\cdot|\boldsymbol{x})}\left[\phi(\tilde{\mathbf{y}}_t)\right] \quad \boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) = \mathbb{E}_{\tilde{\mathbf{y}}\sim\pi_{\mathbf{w}}(\cdot|\boldsymbol{x})}\left[\phi(\tilde{\mathbf{y}})\right], \quad (7)$$

and, given a trajectory $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_T, \ldots, \tilde{\mathbf{y}}_1]$, the KME of the empirical distribution in (5):

$$\hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) = \sum_{t=1}^{T} w_t \phi(\tilde{\mathbf{y}}_t). \quad (8)$$

**Maximum Mean Discrepancy** To measure the distance between two distributions, we use the Maximum Mean Discrepancy (MMD), defined as the distance between the KMEs of two distributions in $\mathcal{H}_k$. We assume throughout that $\psi$ is symmetric, positive definite, and *characteristic* on $\mathbb{R}^d$ (Sriperumbudur et al., 2011), and normalized, $\psi(\boldsymbol{e}, \boldsymbol{e}) = 1$ for every $\boldsymbol{e} \in \mathbb{R}^d$. Because $k = \psi \circ (e \times e)$, the induced discrepancy compares *pushforward* laws in the embedding space,

$$\text{MMD}_k(P, Q) = \text{MMD}_\psi\left(e_{\#}P,\ e_{\#}Q\right), \qquad \text{so} \qquad \text{MMD}_k(P, Q) = 0 \iff e_{\#}P = e_{\#}Q. \quad (9)$$

This is weaker than $P = Q$: two answers with the same semantic embedding are indistinguishable to DiMS, by construction. Yet, for semantic UQ techniques, we do not require this equality at the lexical level, but only that the two distributions induce the same distribution in the semantic embedding space. This is a desirable property for our setting: DiMS is designed to match the *semantic variability* of the model outputs while being insensitive to differences in wording that preserve meaning. The quality of this semantic matching naturally depends on the choice of the embedding $e$ and the kernel $\psi$. We study the robustness of DiMS to these choices in App. E.4.3 and E.4.4.

The quantity DiMS minimizes is the expected squared RKHS error committed at deployment, when a *single* trajectory is used in place of $M$ independent generations:

$$\mathcal{E}_{\boldsymbol{x}}(\mathbf{w}) \ = \ \mathbb{E}_{\tilde{\mathbf{Y}}\sim\bar{\pi}_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})}\Big[\big\|\hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})\big\|^2_{\mathcal{H}_k}\Big]. \quad (10)$$

Note that equation (10) is *not* an MMD between $\pi_{\mathbf{w}}$ and $p_{\boldsymbol{\theta}}^{\text{Diff}}$: the empirical mixture $\hat{\pi}_{\mathbf{w}}(\cdot \mid \tilde{\mathbf{Y}})$ is random, and the expectation is taken outside the squared norm. Section 4.3 shows that the two differ by exactly one term, and that this term is what makes step selection a non-trivial problem.

### 4.3 Deployment error: distribution mismatch and within-trajectory redundancy

In this subsection, we show that $\mathcal{E}_{\boldsymbol{x}}(\mathbf{w})$ is the correct objective for our task, is well suited for optimization, and yields meaningful interpretation.

**Proposition 1** (Exact bias–variance decomposition, proof in App. B.2)**.** *Fix a prompt $\boldsymbol{x}$. For every $\mathbf{w} \in \Delta(\mathcal{T})$, the objective $\mathcal{E}_{\boldsymbol{x}}(\mathbf{w})$ admits the following decomposition:*

$$\mathcal{E}_{\boldsymbol{x}}(\mathbf{w}) = \underbrace{\big\|\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) - \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})\big\|^2_{\mathcal{H}_k}}_{D_{\boldsymbol{x}}(\mathbf{w})} + \underbrace{\mathbf{w}^\top \boldsymbol{C}_{\boldsymbol{x}} \mathbf{w}}_{\textit{within-trajectory variance}} = \mathbf{w}^\top \boldsymbol{B}_{\boldsymbol{x}} \mathbf{w} - 2\boldsymbol{b}_{\boldsymbol{x}}^\top \mathbf{w} + c_{\boldsymbol{x}}, \quad (11)$$

*with $D_{\boldsymbol{x}}(\mathbf{w}) = \text{MMD}_k^2\left(\pi_{\mathbf{w}}(\cdot \mid \boldsymbol{x}),\, p_{\boldsymbol{\theta}}^{\text{Diff}}(\cdot \mid \boldsymbol{x})\right) = \mathbf{w}^\top \boldsymbol{A}_{\boldsymbol{x}} \mathbf{w} - 2\boldsymbol{b}_{\boldsymbol{x}}^\top \mathbf{w} + c_{\boldsymbol{x}}$ and*

$$[\boldsymbol{A}_{\boldsymbol{x}}]_{tu} = \mathbb{E}_{\tilde{\mathbf{y}}_t\sim\pi_{\boldsymbol{\theta},t},\ \tilde{\mathbf{y}}'_u\sim\pi_{\boldsymbol{\theta},u},\text{independently}}\left[k(\tilde{\mathbf{y}}_t, \tilde{\mathbf{y}}'_u)\right] = \left\langle \boldsymbol{\mu}_t(\boldsymbol{x}), \boldsymbol{\mu}_u(\boldsymbol{x})\right\rangle_{\mathcal{H}_k}, \quad (12)$$

$$[\boldsymbol{B}_{\boldsymbol{x}}]_{tu} = \mathbb{E}_{\tilde{\mathbf{Y}}\sim\bar{\pi}_{\boldsymbol{\theta}}}\left[k(\tilde{\mathbf{y}}_t, \tilde{\mathbf{y}}_u)\right], \quad (13)$$

$$[\boldsymbol{b}_{\boldsymbol{x}}]_t = \mathbb{E}_{\tilde{\mathbf{y}}_t\sim\pi_{\boldsymbol{\theta},t},\ \mathbf{y}\sim p_{\boldsymbol{\theta}}^{\text{Diff}}}\left[k(\tilde{\mathbf{y}}_t, \mathbf{y})\right] = \left\langle \boldsymbol{\mu}_t(\boldsymbol{x}), \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})\right\rangle_{\mathcal{H}_k}, \quad (14)$$

$$c_{\boldsymbol{x}} = \mathbb{E}_{\mathbf{y},\mathbf{y}'\overset{i.i.d.}{\sim} p_{\boldsymbol{\theta}}^{\text{Diff}}}\left[k(\mathbf{y}, \mathbf{y}')\right] = \left\|\boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})\right\|^2_{\mathcal{H}_k}, \quad (15)$$

*and $\boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{B}_{\boldsymbol{x}} - \boldsymbol{A}_{\boldsymbol{x}}$. Moreover $\boldsymbol{A}_{\boldsymbol{x}} \succeq 0$, $\boldsymbol{C}_{\boldsymbol{x}} \succeq 0$ and $\boldsymbol{B}_{\boldsymbol{x}} \succeq 0$.*

$\boldsymbol{A}_{\boldsymbol{x}}$ is an expectation over *independent* trajectories and is a Gram matrix of mean embeddings, whereas $\boldsymbol{B}_{\boldsymbol{x}}$ evaluates the kernel on two answers read from the *same* denoising path and therefore sees their coupling. $\boldsymbol{C}_{\boldsymbol{x}}$ is the RKHS covariance matrix of the family $\big(\phi(\tilde{\mathbf{y}}_t)\big)_{t=1}^{T}$ under $\bar{\pi}_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})$.

**Corollary 1** (Equivalence with the $M$-sample target, proof in App. B.3)**.** *Let* $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)} \overset{\text{i.i.d.}}{\sim} p_{\boldsymbol{\theta}}^{\mathrm{Diff}}(\cdot \mid \boldsymbol{x})$ *be independent of* $\tilde{\mathbf{Y}}$*, and* $\hat{\boldsymbol{\mu}}_{\mathrm{iid}} = \frac{1}{M}\sum_m \phi(\mathbf{y}^{(m)})$*. Then, for every* $\mathbf{w} \in \Delta(\mathcal{T})$*,*

$$\mathbb{E}\big\|\hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \hat{\boldsymbol{\mu}}_{\mathrm{iid}}\big\|_{\mathcal{H}_k}^2 = \mathcal{E}_{\boldsymbol{x}}(\mathbf{w}) + \frac{1}{M}\mathrm{Var}_{\mathcal{H}_k}\big(\phi(\mathbf{y})\big), \tag{16}$$

*where the expectation is over the joint distribution of* $\tilde{\mathbf{Y}}$ *and the* $M$ *independent samples from* $p_{\boldsymbol{\theta}}^{\mathrm{Diff}}(\cdot \mid \boldsymbol{x})$*, and where* $\mathrm{Var}_{\mathcal{H}_k}(\phi(\mathbf{y})) = \mathbb{E}\|\phi(\mathbf{y}^{(1)}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\|_{\mathcal{H}_k}^2 = 1 - c_{\boldsymbol{x}}$*.*

Minimizing $\mathcal{E}_{\boldsymbol{x}}$ is thus exactly minimizing the expected squared RKHS distance to what $M$ independent generations would have produced, and the additive constant is independent of $\mathbf{w}$.

### 4.4 Learning and deploying the temporal weights

**Objective.** Following Proposition 1, we optimize $\mathbf{w}$ by solving the constrained quadratic program

$$\min_{\mathbf{w}} \quad J_{\boldsymbol{x}}(\mathbf{w};\lambda) \;=\; D_{\boldsymbol{x}}(\mathbf{w}) \;+\; \lambda\, \mathbf{w}^{\top}\boldsymbol{C}_{\boldsymbol{x}}\mathbf{w}, \qquad \lambda \geq 0, \tag{17}$$

restricted to

$$\mathcal{W}_M = \big\{\mathbf{w} \in \Delta(\mathcal{T}) :\; w_t \leq \tfrac{1}{M}\;\; \forall t \in \mathcal{T}\big\}, \qquad M \leq T. \tag{18}$$

At $\lambda = 1$, the objective (17) equals the deployment error $\mathcal{E}_{\boldsymbol{x}}$, but we tune $\lambda$ in experiments. This tuning and the restriction to $\mathcal{W}_M$ serve complementary purposes. The two terms in (17) pull apart: since $\pi_{\boldsymbol{\theta},1} = p_{\boldsymbol{\theta}}^{\mathrm{Diff}}$, $D_{\boldsymbol{x}}$ vanishes on the last step and is small on late steps, but these are near-copies of one another *within* a trajectory, which is exactly what the redundancy term charges. Without the cap in (18), a small $\lambda$ would collapse $\mathbf{w}$ onto the last step, whose answers are copies of the final one and carry no semantic variability. Thus, $\lambda$ trades marginal fidelity for within-trajectory diversity, and the cap ensures that the weights are spread across at least $M$ steps since its vertices are exactly the uniform weightings of the $M$-step subsets. App. B.5.1 details these choices.

**Calibration.** For a prompt $\boldsymbol{x}$, we draw $R$ independent trajectories $\tilde{\mathbf{Y}}^{(r)} \sim \bar{\pi}_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})$, whose endpoints $\tilde{\mathbf{y}}_1^{(r)}$ then follow the model distribution $p_{\boldsymbol{\theta}}^{\mathrm{Diff}}$. We define the following estimators:

$$\begin{aligned}
[\widehat{\boldsymbol{B}}_{\boldsymbol{x}}]_{tu} &= \frac{1}{R}\sum_{r} k\big(\tilde{\mathbf{y}}_t^{(r)}, \tilde{\mathbf{y}}_u^{(r)}\big), & [\widehat{\boldsymbol{A}}_{\boldsymbol{x}}]_{tu} &= \frac{1}{R^2}\sum_{r,r'} k\big(\tilde{\mathbf{y}}_t^{(r)}, \tilde{\mathbf{y}}_u^{(r')}\big),\\
[\widehat{\boldsymbol{b}}_{\boldsymbol{x}}]_{t} &= \frac{1}{R(R-1)}\sum_{r\neq r'} k\big(\tilde{\mathbf{y}}_t^{(r)}, \tilde{\mathbf{y}}_1^{(r')}\big), & \widehat{c}_{\boldsymbol{x}} &= \frac{1}{R(R-1)}\sum_{r\neq r'} k\big(\tilde{\mathbf{y}}_1^{(r)}, \tilde{\mathbf{y}}_1^{(r')}\big),
\end{aligned} \tag{19}$$

and $\widehat{\boldsymbol{C}}_{\boldsymbol{x}} = \widehat{\boldsymbol{B}}_{\boldsymbol{x}} - \widehat{\boldsymbol{A}}_{\boldsymbol{x}}$. These estimators yield convexity for any $\lambda \geq 0$, at a cost of a small bias:

**Proposition 2** (Proof in App. B.4)**.** *Let* $\lambda \geq 0$ *and* $\widehat{J}_{\boldsymbol{x}}(\mathbf{w};\lambda) = \mathbf{w}^{\top}\big(\widehat{\boldsymbol{A}}_{\boldsymbol{x}} + \lambda\,\widehat{\boldsymbol{C}}_{\boldsymbol{x}}\big)\mathbf{w} - 2\,\widehat{\boldsymbol{b}}_{\boldsymbol{x}}^{\top}\mathbf{w} + \widehat{c}_{\boldsymbol{x}}$*.*
*(a) On every sample* $\tilde{\mathbf{Y}}^{(1)}, \dots, \tilde{\mathbf{Y}}^{(R)}$*,* $\widehat{\boldsymbol{A}}_{\boldsymbol{x}} \succeq 0$ *and* $\widehat{\boldsymbol{C}}_{\boldsymbol{x}} \succeq 0$*, so* $\widehat{J}_{\boldsymbol{x}}(\cdot\,;\lambda)$ *is convex.*
*(b) For every* $\mathbf{w} \in \Delta(\mathcal{T})$*,* $\mathbb{E}\big[\widehat{J}_{\boldsymbol{x}}(\mathbf{w};\lambda)\big] = J_{\boldsymbol{x}}(\mathbf{w};\lambda_{\mathrm{eff}})$*, with* $\lambda_{\mathrm{eff}} = \lambda + \frac{1-\lambda}{R}$*.*

The convex objective $\widehat{J}_{\boldsymbol{x}}(\cdot;\lambda)$ is biased for $J_{\boldsymbol{x}}(\cdot;\lambda)$, but is unbiased for $J_{\boldsymbol{x}}(\cdot;\lambda_{\mathrm{eff}})$. Intuitively, the bias in estimators comes from same-trajectory terms which are captured by $\boldsymbol{C}_{\boldsymbol{x}}$. As $\lambda \mapsto \lambda_{\mathrm{eff}}$ is an increasing bijection from $[0,\infty)$ onto $[1/R,\infty)$, this only reparametrizes the penalty. For $\lambda_{\mathrm{eff}} < 1/R$, convexity breaks (see App. B.5.1). We average the estimators in (19) over a small number of $N_{\mathrm{cal}}$ calibration prompts (denoted by a bar) and optimize $\mathbf{w}$ on this quadratic program:

$$\mathbf{w}^{\star} = \operatorname*{arg\,min}_{\mathbf{w}\in\mathcal{W}_M} \; \mathbf{w}^{\top}\big(\overline{\boldsymbol{A}} + \lambda\,\overline{\boldsymbol{C}}\big)\mathbf{w} - 2\,\overline{\boldsymbol{b}}^{\top}\mathbf{w} + \bar{c}. \tag{20}$$

**Deployment.** At test time, we collect a single trajectory from the LLDM on a prompt and turn $\mathbf{w}^{\star}$ into a set $\mathcal{S}^{\star}$ of steps: for instance the $M$ highest-weight steps, or $M$ distinct steps drawn with inclusion probabilities $M\mathbf{w}^{\star}$, which the cap makes valid. The faithful rule, $M$ i.i.d. draws from $\mathrm{Cat}(\cdot;\mathbf{w}^{\star})$, may instead repeat steps (App. B.5.2). We build $\widehat{\mathcal{Y}}_{\mathrm{DiMS}}(\boldsymbol{x})$ as the intermediate answers at the selected steps in $\mathcal{S}^{\star}$, and pass it through any sample-based semantic uncertainty score:

$$\widehat{\mathcal{Y}}_{\mathrm{DiMS}}(\boldsymbol{x}) = \big\{\tilde{\mathbf{y}}_t\big\}_{t\in\mathcal{S}^{\star}}, \qquad U_{\mathrm{DiMS}}(\boldsymbol{x}) = U_{\mathrm{sem}}\big(\widehat{\mathcal{Y}}_{\mathrm{DiMS}}(\boldsymbol{x})\big). \tag{21}$$

# 5 EXPERIMENTS

## 5.1 EXPERIMENTAL SETUP

**Models, datasets and protocol.** We evaluate DIMS on three instruction-tuned LLDMs: `LLaDA-1.5` (8B) (Zhu et al., 2026), `Dream-v0-Instruct-7B` (Ye et al., 2025) and `Nemotron-Labs-Diffusion-14B` (Fu et al., 2026); and on six datasets spanning general knowledge (**TriviaQA**, Joshi et al., 2017), mathematical reasoning (**GSM8K**, Cobbe et al., 2021; **MATH**, Hendrycks et al., 2021), machine translation (**WMT14 Fr-En**, Bojar et al., 2014), summarization (**SAMSum**, Gliwa et al., 2019) and multi-hop reasoning (**HotpotQA**, Yang et al., 2018). Decoding is zero-shot with $T = L$ denoising steps for $L$ generated tokens (32 on TriviaQA, 128 elsewhere), under the *low-confidence* (`lc`) and *random* (`rd`) remasking schedules; `Nemotron` exposes only `lc`. This gives five model variants $\times$ six datasets $= 30$ runs, each with its own class-stratified $100/200/\text{rest}$ calibration/validation/test split. DIMS is calibrated using $R = 20$ trajectories on the calibration set. Generation settings, prompts and dataset statistics are provided in App. D.1.

**Strategies and estimators.** We compare DIMS with four simple single-trajectory baselines: `Full`[2], `Random`, `Last` and `Centroid`. We also report $M = 20$ *independent* generations as a multi-sample **reference**, not as a competitor. For offline MMD matching, we use `E5-large-v2` (Wang et al., 2022) embeddings with RBF kernel. From the optimal weights $\mathbf{w}^\star$, we select steps with two deterministic rules: DIMS (top-$M$), which selects the $M = 20$ highest-weight steps, and DIMS (cov. 75%), which selects the smallest set of steps carrying 75% of the weight. More details about baselines, similarity measures, DIMS calibration and the $\lambda$ grid are provided in App. D.2. We evaluate each strategy with eight semantic UQ methods (see App. C for details): Semantic Entropy, DegMat, Eccentricity, EigenVal, KLE-Heat, KLE-Matérn, Average Dissimilarity (AD), and CoCoA. Additional results and ablations on experimental settings are provided in App. E.

**Evaluation.** We evaluate uncertainty by its ability to identify incorrect answers (Ielanskyi et al., 2026). We report AUROC in the main text and the Prediction-Rejection Ratio (PRR) (Malinin & Gales, 2021) in App. E.1. Intervals of $\pm 1.96$ standard errors computed from 500 bootstrap resamples are provided in App. E. Following suggestions from Ielanskyi et al. (2026), correctness is evaluated by two large open-weight LLM judges sampled at two different temperatures: 0.5 and 1.0. More details about metrics, judge prompts and agreement and intervals are provided in App. D.3.

## 5.2 RESULTS

**DIMS improves single-trajectory semantic UQ.** Table 1 reports AUROC averaged over the five model variants (per estimator, Table 1a) and over the eight semantic estimators (per variant, Table 1b); full cells are in Table E.1, and per dataset in Tables E.3 to E.7. See Table E.2 for PRR results. Both DIMS rules beat `Full`, `Random`, `Last` and `Centroid` for seven of the eight estimators for four of the five variants. The exception is EigenVal, where `Full` (0.621) stays just ahead of top-$M$ (0.616) and cov. 75% (0.620), and only `LLaDA-1.5` (`lc`) puts `Full` above top-$M$ (0.656 against 0.650), still below cov. 75% (0.659). `Last` is the weakest strategy (0.620 average AUROC, against 0.636 for `Random` and 0.632 for `Centroid`). DIMS is best or tied in 34 of the 40 (estimator, model) blocks of Table E.1, 18 of them under low-confidence remasking.

**How close is one trajectory to twenty generations?** Across the five model variants (Table E.1), the gap with the $M{=}20$ i.i.d. reference stays between 0.021 and 0.044 AUROC (avg. 0.031) above DIMS for CoCoA and between 0.043 and 0.089 (avg. 0.066) for AD (both rules), at 66.7 against 3.4 s per prompt, a $\times 20$ reduction, as expected since generation is the most costly part of the overhead. For the six other estimators that score only the diversity of the sample set, the average gap stays between 0.037 and 0.072 with a similar inference cost reduction. One trajectory thus recovers much of what independent sampling buys, especially with CoCoA. The gap is more pronounced for the pure semantic estimators, but stays competitive thanks to the inference cost reduction. Figure 1 additionally provides the i.i.d. reference for lower values of $M$. Even at low $M$, the i.i.d. reference stays significantly more costly than DIMS, while the AUROC gaps tend to disappear, and even give advantage to DIMS for most UQ estimators. Figure 1 also highlights the cost of `Full` for graph-based estimators and Semantic Entropy because it uses all the steps.

[2]This correspond to Vazhentsev et al. (2026), in particular D-CoCoA-L and D-CoCoA-G (see App. C.4).

Table 1: **Semantic UQ from multiple generations versus a single diffusion trajectory.** We report AUROC (higher is better) averaged in two different ways. Yellow denotes the reference using $M = 20$ independent generations. All other methods use a **single diffusion trajectory**: gray denotes standard single-trajectory baselines (`Full` follows Vazhentsev et al. (2026)), blue denotes our DIMS step-selection strategies, repeated for three sizes $N_{\text{cal}}$ of the unlabeled calibration set (100 by default; the validation and test prompts are identical across the three). Among single-trajectory methods, the best result is in **bold** and the second best is underlined. Smaller calibration sizes ($N_{\text{cal}} = 50$ or 10) do not enter the comparison but highlight stability in low calibration budget. We also report the average computational overhead at inference (See App. D.4 for technical details).

| | Multi-sample | Single trajectory | | | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|
| **UQ estimator** | **Reference** 20 i.i.d. | `Full` | `Random` | `Last` | `Centroid` | **DIMS** ($N_{\text{cal}}$=100) **top-$M$** | **DIMS** ($N_{\text{cal}}$=100) **cov. 75%** | **DIMS** ($N_{\text{cal}}$=50) **top-$M$** | **DIMS** ($N_{\text{cal}}$=50) **cov. 75%** | **DIMS** ($N_{\text{cal}}$=10) **top-$M$** | **DIMS** ($N_{\text{cal}}$=10) **cov. 75%** |
| Semantic Entropy | 0.700 | 0.595 | 0.589 | 0.611 | 0.565 | 0.623 | **0.628** | 0.625 | 0.630 | 0.626 | 0.624 |
| DegMat | 0.701 | 0.638 | 0.629 | 0.595 | 0.627 | **0.643** | **0.643** | 0.641 | 0.644 | 0.640 | 0.642 |
| Eccentricity | 0.679 | 0.623 | 0.615 | 0.591 | 0.616 | 0.627 | **0.628** | 0.626 | 0.628 | 0.620 | 0.623 |
| EigenVal | 0.657 | **0.621** | 0.611 | 0.563 | 0.612 | 0.616 | 0.620 | 0.616 | 0.619 | 0.615 | 0.618 |
| KLE-Heat | 0.705 | 0.634 | 0.629 | 0.596 | 0.633 | **0.646** | 0.644 | 0.642 | 0.644 | 0.641 | 0.644 |
| KLE-Matérn | 0.704 | 0.638 | 0.629 | 0.596 | 0.634 | 0.644 | **0.645** | 0.641 | 0.647 | 0.642 | 0.644 |
| AD | 0.762 | 0.668 | 0.665 | 0.686 | 0.653 | 0.694 | **0.696** | 0.695 | 0.697 | 0.694 | 0.697 |
| CoCoA | 0.763 | 0.724 | 0.722 | 0.723 | 0.720 | **0.732** | **0.732** | 0.732 | 0.731 | 0.732 | 0.733 |
| **Avg. AUROC ↑** | 0.709 | 0.643 | 0.636 | 0.620 | 0.632 | 0.653 | **0.655** | 0.652 | 0.655 | 0.651 | 0.653 |
| **Overhead (s) ↓** | 66.81 | 7.20 | 3.51 | **3.44** | 3.61 | 3.50 | 3.88 | — | — | — | — |

(a) AUROC ↑ for each semantic UQ estimator, averaged over the model variants and datasets.

| **Model variant** | **Reference** 20 i.i.d. | `Full` | `Random` | `Last` | `Centroid` | **DIMS** ($N_{\text{cal}}$=100) **top-$M$** | **DIMS** ($N_{\text{cal}}$=100) **cov. 75%** | **DIMS** ($N_{\text{cal}}$=50) **top-$M$** | **DIMS** ($N_{\text{cal}}$=50) **cov. 75%** | **DIMS** ($N_{\text{cal}}$=10) **top-$M$** | **DIMS** ($N_{\text{cal}}$=10) **cov. 75%** |
|---|---|---|---|---|---|---|---|---|---|---|---|
| LLaDA-1.5 (`lc`) | 0.717 | 0.656 | 0.647 | 0.631 | 0.647 | 0.650 | **0.659** | 0.651 | 0.658 | 0.650 | 0.657 |
| LLaDA-1.5 (`rd`) | 0.728 | 0.669 | 0.659 | 0.605 | 0.662 | 0.674 | **0.677** | 0.675 | 0.678 | 0.675 | 0.675 |
| Dream (`lc`) | 0.747 | 0.669 | 0.664 | 0.680 | 0.650 | **0.697** | 0.694 | 0.695 | 0.695 | 0.689 | 0.695 |
| Dream (`rd`) | 0.704 | 0.634 | 0.626 | 0.594 | 0.629 | 0.637 | **0.641** | 0.638 | 0.641 | 0.640 | 0.639 |
| Nemotron (`lc`) | 0.648 | 0.585 | 0.584 | 0.591 | 0.575 | **0.605** | 0.602 | 0.602 | 0.603 | 0.603 | 0.600 |

(b) AUROC ↑ for each model variant, averaged over the semantic UQ estimators and datasets.

**Ten prompts are enough for calibration** The quadratic program (20) reads no correctness label, so calibration costs only unlabeled generations. The right-hand block of Table 1a shrinks it to 50 and 10 nested prompts, validation and test unchanged. The average AUROC barely moves between $N_{\text{cal}} = 100$ and $N_{\text{cal}} = 50$ or 10 (by at most 0.002). The estimator and model variant averages also stay very stable, shifting by at most 0.008 (intervals in Table. E.14). Reducing instead the number of sampled trajectories per calibration prompt from $R = 20$ to 10 or 5 (Table. E.13) changes no cell by more than 0.005. This suggests that the offline cost of DIMS calibration can be kept small, contributing to the practicality of DIMS in restricted budget scenarios.

**Which steps are selected?** Figure 2 shows that the selection depends mainly on the decoding schedule. Under low-confidence remasking, DIMS keeps both ends of the trajectory: early steps provide diversity, late ones stay close to the final answer. It drops the redundant middle (31% of the kept steps fall in the first third of the trajectory, 64% in the last and 5% in between); random remasking puts most of them early (72% in the first third) and spreads the rest over the middle (20%). Within that pattern, GSM8K, HotpotQA and SAMSum split their steps evenly between the two ends, MATH leans early (69% in the first third) and WMT14 late (50% in the last), and TriviaQA, whose schedule leaves top-$M$ little room ($M$=20 of $T$=32 steps), spreads them over the whole trajectory.

**Ablations, task, and transfer.** A tuned redundancy penalty $\lambda^\star$ gives the best average AUROC for all eight estimators (Table E.10); fixing $\lambda = 1$ instead costs up to 0.031. The other design choices matter less: at most 0.016 across the six step-selection rules, 0.005 across three kernels and 0.003 across three encoders (Tables E.9, E.11 and E.12). Grouped by task, DIMS is the best single-trajectory strategy on all five tasks, 0.006 (QA) to 0.020 (summarization) AUROC above the best baseline, which is `Full` on four tasks and `Last` on summarization (Table E.8). The weights

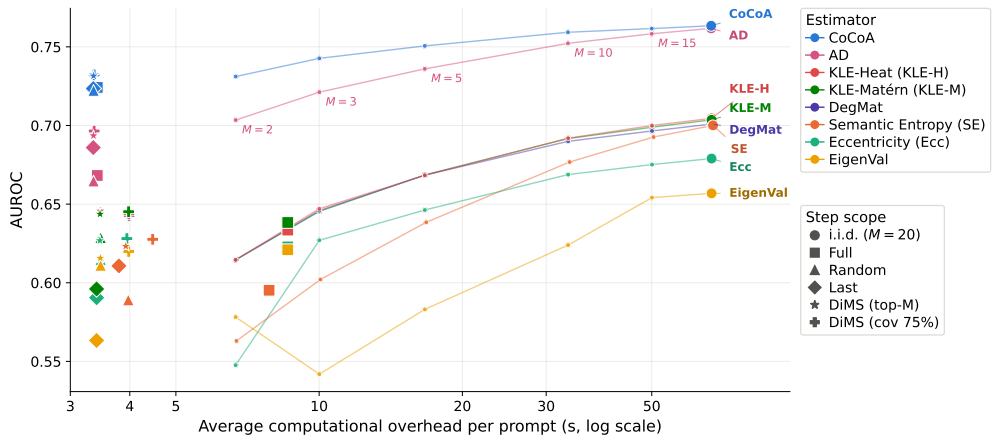



Figure 1: **Accuracy–cost plane**, pooled over the five model variants and six datasets, for all eight semantic UQ estimators (one color per estimator). The $M$ i.i.d. references (circles) are shown for $M \in \{2, 3, 5, 10, 15, 20\}$. Every other marker is a single-trajectory strategy (`Centroid` not shown for clarity). The overhead axis uses a log scale to separate the single-trajectory points.

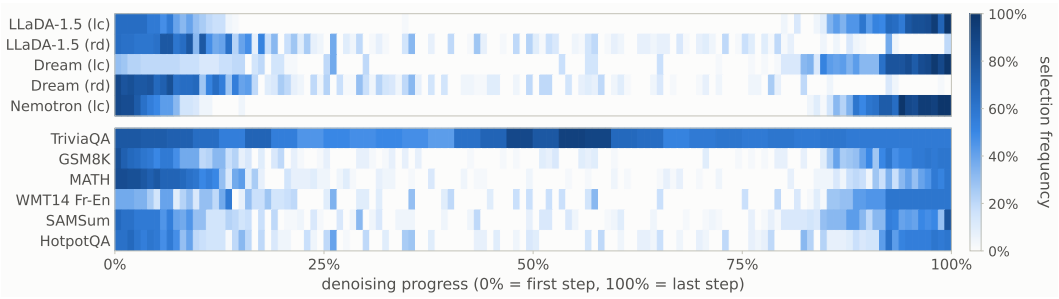



Figure 2: **Which denoising steps** DiMS (top-$M$) keeps for the 8 estimators, as a function of normalized trajectory position (0% = first step, 100% = last). Color is the fraction of (run, estimator) cells for which the step is selected. *Top:* per model variant, averaged over datasets; TriviaQA ($T$=32 steps, 128 elsewhere) is held out of the averages. *Bottom:* per dataset, averaged over models.

also transfer: deployed on another run they lose 0.007 AUROC on average across datasets and 0.017 across models (Tables E.16 and E.15), the remasking schedule being the most harmful axis: which steps carry signal depends more on the decoding dynamics than on the model or task. Yet, since a very restricted calibration budget offers stable performance (see discussion above), it may be valuable to recalibrate DiMS for new models and datasets.

## 6 Conclusion

We introduced DiMS, a simple method to make semantic UQ for LLDMs more efficient by selecting informative states from a single denoising trajectory. Our theory shows that this problem can be written as a trade-off between matching the final-answer distribution and reducing within-trajectory redundancy, leading to a convex MMD-based objective. DiMS is estimator-agnostic, requires no task-specific labels, can be calibrated on very few trajectories, and can be directly plugged into existing semantic UQ methods. Across three LLDMs, six datasets, and eight semantic UQ estimators, DiMS improves over standard single-trajectory baselines and is best or tied best in $34/40$ settings. For some estimators, notably CoCoA, one optimized trajectory recovers much of the performance of multiple independent generations at a much lower inference cost. Overall, our results show that efficient UQ is not only about designing better uncertainty scores, but also about optimizing the sampling strategy itself.

### AI USE STATEMENT

In this work, we used generative AI tools to assist with method implementation, and to improve readability of the manuscript: refining sentence-level phrasing and tables and figures style. We have reviewed all AI-assisted work and LLM-generated pieces of code were verified and tested for correctness. We take responsibility for the final content of this work, including text, claims or artifacts produced with the aid of generative AI.

### REPRODUCIBILITY STATEMENT

We provide the details needed to reproduce both DIMS and our experiments. Section 4 gives the full method, objective, and optimization procedure, while Appendix B contains complete proofs of the theoretical results. The semantic UQ estimators used in our evaluation are described in Appendix C, and Appendix D reports the models, datasets, generation settings, calibration and validation protocols, judge prompts, hyperparameter search, and computational-overhead measurement. Full per-model and per-dataset results, together with ablations and transfer experiments, are provided in Appendix E. An anonymized version of our code is also provided as supplementary material.

## REFERENCES

Lukas Aichberger, Kajetan Schweighofer, and Sepp Hochreiter. Rethinking uncertainty estimation in LLMs: A principled single-sequence measure. In *International Conference on Learning Representations*, 2026.

David Arthur and Sergei Vassilvitskii. k-means++: The advantages of careful seeding. In *Proceedings of the Eighteenth Annual ACM-SIAM Symposium on Discrete Algorithms*, pp. 1027–1035, 2007.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems*, volume 34, pp. 17981–17993, 2021.

Tiwei Bie, Maosong Cao, Kun Chen, Lun Du, Mingliang Gong, Zhuochen Gong, Yanmei Gu, Jiaqi Hu, Zenan Huang, Zhenzhong Lan, Chengxi Li, Chongxuan Li, Jianguo Li, Zehuan Li, Huabin Liu, Lin Liu, Guoshan Lu, Xiaocheng Lu, Yuxin Ma, Jianfeng Tan, Lanning Wei, Ji-Rong Wen, Yipeng Xing, Xiaolu Zhang, Junbo Zhao, Da Zheng, Jun Zhou, Junlin Zhou, Zhanchao Zhou, Liwang Zhu, and Yihong Zhuang. LLaDA2.0: Scaling up diffusion language models to 100B. arXiv preprint arXiv:2512.15745, 2025.

Ondřej Bojar, Christian Buck, Christian Federmann, Barry Haddow, Philipp Koehn, Johannes Leveling, Christof Monz, Pavel Pecina, Matt Post, Hervé Saint-Amand, Radu Soricut, Lucia Specia, and Aleš Tamchyna. Findings of the 2014 workshop on statistical machine translation. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pp. 12–58, 2014.

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

Shenxu Chang, Junchi Yu, Weixing Wang, Yongqiang Chen, Jialin Yu, Philip Torr, and Jindong Gu. TraceDet: Hallucination detection from the decoding trace of diffusion large language models. In *International Conference on Learning Representations*, 2026.

Jianlyu Chen, Shitao Xiao, Peitian Zhang, Kun Luo, Defu Lian, and Zheng Liu. M3-Embedding: Multi-linguality, multi-functionality, multi-granularity text embeddings through self-knowledge distillation. In *Findings of the Association for Computational Linguistics: ACL 2024*, pp. 2318–2335, 2024.

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 preprint arXiv:2110.14168, 2021.

Steven Diamond and Stephen Boyd. CVXPY: A Python-embedded modeling language for convex optimization. *Journal of Machine Learning Research*, 17(83):1–5, 2016.

DiffusionGemma Team. DiffusionGemma technical report. arXiv preprint arXiv:2608.00146, 2026.

Alexander Domahidi, Eric Chu, and Stephen Boyd. ECOS: An SOCP solver for embedded systems. In *European Control Conference*, pp. 3071–3076, 2013.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. LM-Polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pp. 446–461, 2023.

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

Marina Fomicheva, Shuo Sun, Lisa Yankovskaya, Frédéric Blain, Francisco Guzmán, Mark Fishel, Nikolaos Aletras, Vishrav Chaudhary, and Lucia Specia. Unsupervised quality estimation for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8: 539–555, 2020.

Yonggan Fu, Lexington Whalen, Abhinav Garg, Chengyue Wu, Maksim Khadkevich, Nicolai Oswald, Enze Xie, Daniel Egert, Sharath Turuvekere Sreenivas, Shizhe Diao, Chenhan Yu, Ye Yu, Weijia Chen, Sajad Norouzi, Jingyu Liu, Shiyi Lan, Ligeng Zhu, Jin Wang, Jindong Jiang, Morteza Mardani, Mehran Maghoumi, Song Han, Ante Jukić, Nima Tajbakhsh, Jan Kautz, and Pavlo Molchanov. Nemotron-Labs-Diffusion: A tri-mode language model unifying autoregressive, diffusion, and self-speculation decoding. arXiv preprint arXiv:2607.05722, 2026.

Bogdan Gliwa, Iwona Mochol, Maciej Biesek, and Aleksander Wawer. SAMSum corpus: A human-annotated dialogue dataset for abstractive summarization. In *Proceedings of the 2nd Workshop on New Frontiers in Summarization*, pp. 70–79, 2019.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, et al. The Llama 3 herd of models. arXiv preprint arXiv:2407.21783, 2024.

Arthur Gretton, Karsten M. Borgwardt, Malte J. Rasch, Bernhard Schölkopf, and Alexander Smola. A kernel two-sample test. *Journal of Machine Learning Research*, 13(25):723–773, 2012.

Jiatong Han, Jannik Kossen, Muhammed Razzak, Lisa Schut, Shreshth A. Malik, and Yarin Gal. Semantic entropy probes: Robust and cheap hallucination detection in LLMs. In *ICML 2024 Workshop on Foundation Models in the Wild*, 2024.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. DeBERTa: Decoding-enhanced BERT with disentangled attention. In *International Conference on Learning Representations*, 2021.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the MATH dataset. In *Proceedings of the Neural Information Processing Systems Track on Datasets and Benchmarks*, volume 1, 2021.

Joseph Hoche, Andrei Bursuc, David Brellmann, Gilles Louppe, Pavel Izmailov, Angela Yao, and Gianni Franchi. Improving semantic uncertainty quantification in LVLMs with semantic Gaussian processes. arXiv preprint arXiv:2512.14177, 2025.

Joseph Hoche, Maxime Corlay, David Brellmann, Andrei Bursuc, Pavel Izmailov, Angela Yao, and Gianni Franchi. Position: Semantic uncertainty measures disagreement, not reliability. HAL preprint hal-05633506, 2026.

Mykyta Ielanskyi, Kajetan Schweighofer, Lukas Aichberger, and Sepp Hochreiter. Addressing pitfalls in the evaluation of uncertainty estimation methods for natural language generation. In *International Conference on Learning Representations*, 2026.

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

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

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

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. arXiv preprint arXiv:1907.11692, 2019.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. In *Proceedings of the 41st International Conference on Machine Learning*, volume 235 of *Proceedings of Machine Learning Research*, pp. 32819–32848, 2024.

William G. Madow. On the theory of systematic sampling, II. *The Annals of Mathematical Statistics*, 20(3):333–354, 1949.

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

Krikamol Muandet, Kenji Fukumizu, Bharath Sriperumbudur, and Bernhard Schölkopf. Kernel mean embedding of distributions: A review and beyond. *Foundations and Trends in Machine Learning*, 10(1–2):1–141, 2017.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. Large language diffusion models. In *Advances in Neural Information Processing Systems*, volume 38, pp. 50608–50646, 2025.

Alexander Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *Advances in Neural Information Processing Systems*, volume 37, pp. 8901–8929, 2024.

Brendan O'Donoghue, Eric Chu, Neal Parikh, and Stephen Boyd. Operator splitting for conic optimization via homogeneous self-dual embedding. arXiv preprint arXiv:1312.3039, 2013.

Yanyu Qian, Yue Tan, Yixin Liu, Wang Yu, and Shirui Pan. DynHD: Hallucination detection for diffusion large language models via denoising dynamics deviation learning. In *Findings of the Association for Computational Linguistics: EMNLP 2026*, 2026.

Qwen Team. Qwen3.5: Towards native multimodal agents. `https://qwen.ai/blog?id=qwen3.5`, 2026.

Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pp. 3982–3992, 2019.

Subham Sekhar Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T. Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. In *Advances in Neural Information Processing Systems*, volume 37, pp. 130136–130184, 2024.

Jiaxin Shi, Kehang Han, Zhe Wang, Arnaud Doucet, and Michalis Titsias. Simplified and generalized masked diffusion for discrete data. In *Advances in Neural Information Processing Systems*, volume 37, pp. 103131–103167, 2024.

Bharath K. Sriperumbudur, Kenji Fukumizu, and Gert R. G. Lanckriet. Universality, characteristic kernels and RKHS embedding of measures. *Journal of Machine Learning Research*, 12(70): 2389–2410, 2011.

Roman Vashurin, Ekaterina Fadeeva, Artem Vazhentsev, Lyudmila Rvanova, Daniil Vasilev, Akim Tsvigun, Sergey Petrakov, Rui Xing, Abdelrahman Sadallah, Kirill Grishchenkov, Alexander Panchenko, Timothy Baldwin, Preslav Nakov, Maxim Panov, and Artem Shelmanov. Benchmarking uncertainty quantification methods for large language models with LM-Polygraph. *Transactions of the Association for Computational Linguistics*, 13:220–248, 2025a.

Roman Vashurin, Maiya Goloburda, Albina Ilina, Aleksandr Rubashevskii, Preslav Nakov, Artem Shelmanov, and Maxim Panov. CoCoA: A minimum Bayes risk framework bridging confidence and consistency for uncertainty quantification in LLMs. In *Advances in Neural Information Processing Systems*, volume 38, pp. 106236–106281, 2025b.

Artem Vazhentsev, Vladislav Smirnov, David Li, Maxim Panov, Timothy Baldwin, and Artem Shelmanov. Uncertainty quantification for large language diffusion models. arXiv preprint arXiv:2605.14570, 2026.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. Text embeddings by weakly-supervised contrastive pre-training. arXiv preprint arXiv:2212.03533, 2022.

Wenhui Wang, Furu Wei, Li Dong, Hangbo Bao, Nan Yang, and Ming Zhou. MiniLM: Deep self-attention distillation for task-agnostic compression of pre-trained transformers. In *Advances in Neural Information Processing Systems*, volume 33, pp. 5776–5788, 2020.

Adina Williams, Nikita Nangia, and Samuel R. Bowman. A broad-coverage challenge corpus for sentence understanding through inference. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 1112–1122, 2018.

Jingxuan Wu, Zhenglin Wan, Yuzhe Yang, Yiqiao Huang, Chubin Zhang, Xingrui Yu, Ivor Tsang, and Yang You. Time-annealed perturbation sampling: Diverse generation for diffusion language models. In *ICML 2026 Workshop on Structured Probabilistic Inference & Generative Modeling*, 2026.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pp. 2369–2380, 2018.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. Dream 7B: Diffusion large language models. arXiv preprint arXiv:2508.15487, 2025.

Fengqi Zhu, Rongzhen Wang, Shen Nie, Xiaolu Zhang, Chunwei Wu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. LLaDA 1.5: Variance-reduced preference optimization for large language diffusion models. In *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 11425–11460, 2026.

# Supplementary: One Trajectory Is Enough with Wise Step Selection for Semantic Uncertainty Quantification in Large Language Diffusion Models

## TABLE OF CONTENTS


**A Notation** **15**

**B Proofs and method design details** **15**

B.1 Preliminaries . . . . . 15

B.2 Proof of Proposition 1 . . . . . 17

B.3 Proof of Corollary 1 . . . . . 19

B.4 Proof of Proposition 2 . . . . . 20

B.5 Method design details . . . . . 21

**C Uncertainty quantification for text generation** **23**

C.1 Uncertainty scores and their evaluation . . . . . 23

C.2 Information-based estimators . . . . . 24

C.3 Diffusion-specific estimators . . . . . 25

C.4 Semantic-consistency estimators . . . . . 26

C.5 Summary . . . . . 29

**D Additional Details of Experimental Setup** **30**

D.1 Models, datasets and protocol . . . . . 30

D.2 Strategies and estimators . . . . . 32

D.3 Evaluation . . . . . 34

D.4 Overhead measurement . . . . . 36

**E Additional Results** **36**

E.1 Prediction-Rejection Ratio . . . . . 36

E.2 Per-model, per-dataset breakdown . . . . . 37

E.3 Per-task breakdown . . . . . 37

E.4 Ablations . . . . . 37

E.5 Transfer across datasets and models . . . . . 38

# A NOTATION

Table A.1: **Symbols used in the paper.** "Ref." points to the place where the symbol is defined.

| Symbol | Meaning | Ref. |
|---|---|---|
| $T$ | number of denoising steps | §3.1 |
| $M$ | number of i.i.d. answers drawn or selected steps by DIMS | Eq. (3) |
| $N_{\mathrm{cal}}$ | number of calibration prompts for DIMS | Eq. (20) |
| $R$ | trajectories drawn per calibration prompt | §4.4 |
| $\mathcal{V}$ | token vocabulary | §3 |
| $\boldsymbol{x}$ | prompt | §3 |
| $\mathbf{y} = [\mathrm{y}^1, \dots, \mathrm{y}^L]$ | answer; superscripts index token positions | §3 |
| $\boldsymbol{z}_t$ | partially masked state at step $t$; $\boldsymbol{z}_0 = \mathbf{y}$ | §3.1 |
| $\tilde{\mathbf{y}}_t$ | intermediate answer estimate at step $t$ | Eq. (1) |
| $\tilde{\mathbf{Y}}$ | trajectory $(\tilde{\mathbf{y}}_T, \dots, \tilde{\mathbf{y}}_1)$ | §3.1 |
| $f^{\ell}_{\boldsymbol{\theta}}(\cdot \mid \cdot)$ | predicted token distribution at position $\ell$ | §3 |
| $\tau$ | sampling temperature | App. D.1 |
| $p^{\mathrm{Diff}}_{\boldsymbol{\theta}}(\mathbf{y} \mid \boldsymbol{x})$ | LLDM answer distribution (the $t$=0 marginal) | §3.1 |
| $\Delta(\mathcal{T})$ | simplex over $\mathcal{T} = \{T, \dots, 1\}$ | §3 |
| $\mathbf{w} = [w_T, \dots, w_1]$ | step weights in $\Delta(\mathcal{T})$ | Eq. (4) |
| $\pi_{\boldsymbol{\theta},t}$ | marginal distribution of $\tilde{\mathbf{y}}_t$ | §3.1 |
| $\bar{\pi}_{\boldsymbol{\theta}}$ | joint distribution of $\tilde{\mathbf{Y}}$ | §3.1 |
| $\pi_{\mathbf{w}}$ | mixture of step marginals | Eq. (4) |
| $\hat{\pi}_{\mathbf{w}}(\cdot \mid \tilde{\mathbf{Y}})$ | empirical mixture on trajectory $\tilde{\mathbf{Y}}$ | Eq. (5) |
| $k = \psi \circ (e \times e)$ | semantic kernel on answers | Eq. (6) |
| $\phi(\mathbf{y}) = k(\mathbf{y}, \cdot)$ | feature map into $\mathcal{H}_k$ | §4.2 |
| $\boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})$ | KME of $p^{\mathrm{Diff}}_{\boldsymbol{\theta}}$ | Eq. (7) |
| $\boldsymbol{\mu}_t(\boldsymbol{x})$ | KME of $\pi_{\boldsymbol{\theta},t}$ | Eq. (7) |
| $\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x})$ | KME of the mixture $\pi_{\mathbf{w}}$ | Eq. (7) |
| $\hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}})$ | KME of the empirical mixture $\hat{\pi}_{\mathbf{w}}(\tilde{\mathbf{Y}})$ | Eq. (8) |
| $\mathcal{E}_{\boldsymbol{x}}(\mathbf{w})$ | expected squared RKHS deployment error | Eq. (10) |
| $D_{\boldsymbol{x}}(\mathbf{w})$ | distribution-mismatch term (a squared MMD) | Eq. (11) |
| $\boldsymbol{A}_{\boldsymbol{x}}, \boldsymbol{B}_{\boldsymbol{x}}, \boldsymbol{C}_{\boldsymbol{x}}$ | kernel Gram matrices; $\boldsymbol{C}_{\boldsymbol{x}}$=$\boldsymbol{B}_{\boldsymbol{x}}$−$\boldsymbol{A}_{\boldsymbol{x}}$ | Prop. 1 |
| $\boldsymbol{b}_{\boldsymbol{x}}$, $c_{\boldsymbol{x}}$ | cross-term and normalizer of $D_{\boldsymbol{x}}$ | Eq. (14), Eq. (15) |
| $\lambda$ | redundancy penalty weight | Eq. (17) |
| $\mathcal{W}_M$ | capped simplex, feasible set for $\mathbf{w}$ | Eq. (18) |
| $\lambda_{\mathrm{eff}}$ | effective penalty, unbiased plug-in estimator | Prop. 2 |
| $\mathcal{S}^{\star}$ | set of $M$ steps selected at deployment | §4.4 |
| $U_{\mathrm{sem}}$, $U_{\mathrm{DIMS}}$ | semantic UQ score; its DIMS instantiation | Eq. (3), Eq. (21) |

# B PROOFS AND METHOD DESIGN DETAILS

## B.1 PRELIMINARIES

**Reproducing property.** The kernel $k$ of equation (6) is symmetric and positive definite, $\mathcal{H}_k$ is its RKHS and $\phi(\mathbf{y}) = k(\mathbf{y}, \cdot) \in \mathcal{H}_k$. We use repeatedly the reproducing property

$$k(\mathbf{y}, \mathbf{y}') = \big\langle \phi(\mathbf{y}), \phi(\mathbf{y}') \big\rangle_{\mathcal{H}_k}, \qquad \mathbf{y}, \mathbf{y}' \in \mathcal{V}^L. \tag{B.1}$$

Under the normalization assumption of Section 4.2, $\psi(\boldsymbol{e}, \boldsymbol{e}) = 1$ for every $\boldsymbol{e} \in \mathbb{R}^d$, equation (B.1) gives in particular

$$\|\phi(\mathbf{y})\|^2_{\mathcal{H}_k} = k(\mathbf{y}, \mathbf{y}) = 1 \qquad \text{for every } \mathbf{y} \in \mathcal{V}^L. \tag{B.2}$$

**Expectations of $\mathcal{H}_k$-valued random elements.** Because $\mathcal{V}^L$ and $\mathcal{T}$ are finite, every random element appearing below takes finitely many values in $\mathcal{H}_k$. Its expectation is well defined without

further assumptions, and every interchange of expectation with a finite sum or with an inner product performed below is a manipulation of finite sums. We first record the two elementary facts used throughout.

**Lemma 1** (Mean embedding of a step mixture)**.** *Fix a prompt $\boldsymbol{x}$ and $\mathbf{w} \in \Delta(\mathcal{T})$. Then*

$$\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) \;=\; \sum_{t=1}^{T} w_t\, \boldsymbol{\mu}_t(\boldsymbol{x}) \;=\; \mathbb{E}_{\tilde{\mathbf{Y}} \sim \bar{\pi}_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})}\big[\hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}})\big]. \tag{B.3}$$

*Proof.* By definition of the KME (7) and of the mixture $\pi_{\mathbf{w}}$ (4),

$$\begin{aligned}
\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) = \mathbb{E}_{\tilde{\mathbf{y}} \sim \pi_{\mathbf{w}}(\cdot \mid \boldsymbol{x})}\big[\phi(\tilde{\mathbf{y}})\big] &= \sum_{\tilde{\mathbf{y}} \in \mathcal{V}^L} \pi_{\mathbf{w}}(\tilde{\mathbf{y}} \mid \boldsymbol{x})\, \phi(\tilde{\mathbf{y}}) \\
&= \sum_{\tilde{\mathbf{y}} \in \mathcal{V}^L} \sum_{t=1}^{T} w_t\, \pi_{\boldsymbol{\theta},t}(\tilde{\mathbf{y}} \mid \boldsymbol{x})\, \phi(\tilde{\mathbf{y}}) \\
&= \sum_{t=1}^{T} w_t \sum_{\tilde{\mathbf{y}} \in \mathcal{V}^L} \pi_{\boldsymbol{\theta},t}(\tilde{\mathbf{y}} \mid \boldsymbol{x})\, \phi(\tilde{\mathbf{y}}) = \sum_{t=1}^{T} w_t\, \boldsymbol{\mu}_t(\boldsymbol{x}),
\end{aligned}$$

the exchange of the two sums being legitimate since both are finite. For the second equality in (B.3), recall $\hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) = \sum_t w_t \phi(\tilde{\mathbf{y}}_t)$ from equation (8). By linearity of the expectation on the finite-valued $\mathcal{H}_k$-valued elements $\phi(\tilde{\mathbf{y}}_t)$,

$$\mathbb{E}_{\tilde{\mathbf{Y}} \sim \bar{\pi}_{\boldsymbol{\theta}}}\big[\hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}})\big] = \sum_{t=1}^{T} w_t\, \mathbb{E}_{\tilde{\mathbf{Y}} \sim \bar{\pi}_{\boldsymbol{\theta}}}\big[\phi(\tilde{\mathbf{y}}_t)\big] = \sum_{t=1}^{T} w_t\, \mathbb{E}_{\tilde{\mathbf{y}}_t \sim \pi_{\boldsymbol{\theta},t}}\big[\phi(\tilde{\mathbf{y}}_t)\big] = \sum_{t=1}^{T} w_t\, \boldsymbol{\mu}_t(\boldsymbol{x}),$$

where the middle equality holds because $\phi(\tilde{\mathbf{y}}_t)$ is a function of the $t$-th coordinate of $\tilde{\mathbf{Y}}$ alone and $\pi_{\boldsymbol{\theta},t}$ is by definition the $t$-th marginal of $\bar{\pi}_{\boldsymbol{\theta}}$ (§3.1). □

**Lemma 2** (Exchange and factorization in $\mathcal{H}_k$)**.** *Let $\xi, \eta$ be $\mathcal{H}_k$-valued random elements taking finitely many values.*

*a. For every deterministic $f \in \mathcal{H}_k$, $\mathbb{E}\langle \xi, f\rangle_{\mathcal{H}_k} = \langle \mathbb{E}[\xi], f\rangle_{\mathcal{H}_k}$.*

*b. If $\xi$ and $\eta$ are independent, then $\mathbb{E}\langle \xi, \eta\rangle_{\mathcal{H}_k} = \langle \mathbb{E}[\xi], \mathbb{E}[\eta]\rangle_{\mathcal{H}_k}$.*

*In particular, taking $\xi = \phi(\mathbf{u})$ and $\eta = \phi(\mathbf{v})$ with $\mathbf{u}, \mathbf{v}$ independent $\mathcal{V}^L$-valued random variables and using the reproducing property (B.1),*

$$\mathbb{E}\big[k(\mathbf{u}, \mathbf{v})\big] = \big\langle \mathbb{E}[\phi(\mathbf{u})], \mathbb{E}[\phi(\mathbf{v})]\big\rangle_{\mathcal{H}_k}. \tag{B.4}$$

*Proof. (a).* Let $\xi$ take the values $g_1, \ldots, g_J \in \mathcal{H}_k$ with probabilities $p_1, \ldots, p_J$. Since $f$ is deterministic, linearity of the inner product in its first argument applied to a finite sum gives

$$\mathbb{E}\langle \xi, f\rangle_{\mathcal{H}_k} = \sum_{j=1}^{J} p_j\, \langle g_j, f\rangle_{\mathcal{H}_k} = \Big\langle \sum_{j=1}^{J} p_j\, g_j,\; f \Big\rangle_{\mathcal{H}_k} = \langle \mathbb{E}[\xi], f\rangle_{\mathcal{H}_k}.$$

*(b).* By independence the joint law of $(\xi, \eta)$ is the product of the marginals, and both are finitely supported, so Fubini's theorem for finite sums allows to integrate iteratively:

$$\mathbb{E}\langle \xi, \eta\rangle_{\mathcal{H}_k} = \mathbb{E}_{\eta}\Big[\mathbb{E}_{\xi}\langle \xi, \eta\rangle_{\mathcal{H}_k}\Big] = \mathbb{E}_{\eta}\Big[\big\langle \mathbb{E}[\xi], \eta\big\rangle_{\mathcal{H}_k}\Big] = \big\langle \mathbb{E}[\xi], \mathbb{E}[\eta]\big\rangle_{\mathcal{H}_k},$$

where the second equality applies *(a)* for each fixed value of $\eta$ (legitimate, since conditionally on $\{\eta = g\}$ the vector $g$ is deterministic and, by independence, the conditional law of $\xi$ is its unconditional law), and the third applies *(a)* once more with the deterministic vector $f = \mathbb{E}[\xi]$. □

### B.2 Proof of Proposition 1

We prove Proposition 1 in five steps. We first rewrite the quantities $\boldsymbol{A}_{\boldsymbol{x}}$, $\boldsymbol{b}_{\boldsymbol{x}}$, and $c_{\boldsymbol{x}}$ using kernel mean embeddings. We then decompose the deployment error into a distribution-matching term and a within-trajectory variance term. Finally, we identify this variance with $\mathbf{w}^\top \boldsymbol{C}_{\boldsymbol{x}} \mathbf{w}$ and prove the positive-semidefiniteness of the three matrices.

**Step 1: kernel mean expressions.** Recall that the entries of $\boldsymbol{A}_{\boldsymbol{x}}$ compare answers sampled *independently* from two intermediate-step distributions. Therefore, by (B.4) from Lemma 2 we have

$$[\boldsymbol{A}_{\boldsymbol{x}}]_{tu} = \mathbb{E}_{\tilde{\mathbf{y}}_t \sim \pi_{\boldsymbol{\theta},t},\, \tilde{\mathbf{y}}'_u \sim \pi_{\boldsymbol{\theta},u}} \left[k(\tilde{\mathbf{y}}_t, \tilde{\mathbf{y}}'_u)\right] = \langle \boldsymbol{\mu}_t(\boldsymbol{x}), \boldsymbol{\mu}_u(\boldsymbol{x}) \rangle_{\mathcal{H}_k}. \tag{B.5}$$

The same argument applies to $\boldsymbol{b}_{\boldsymbol{x}}$, since the intermediate answer and the final answer are sampled independently:

$$[\boldsymbol{b}_{\boldsymbol{x}}]_t = \mathbb{E}_{\tilde{\mathbf{y}}_t \sim \pi_{\boldsymbol{\theta},t},\, \mathbf{y} \sim p^{\mathrm{Diff}}_{\boldsymbol{\theta}}} \left[k(\tilde{\mathbf{y}}_t, \mathbf{y})\right] = \langle \boldsymbol{\mu}_t(\boldsymbol{x}), \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x}) \rangle_{\mathcal{H}_k}. \tag{B.6}$$

Finally, for two independent final answers,

$$c_{\boldsymbol{x}} = \mathbb{E}_{\mathbf{y},\mathbf{y}' \overset{\text{i.i.d.}}{\sim} p^{\mathrm{Diff}}_{\boldsymbol{\theta}}} \left[k(\mathbf{y}, \mathbf{y}')\right] = \|\boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\|^2_{\mathcal{H}_k}. \tag{B.7}$$

These identities give exactly the second expressions in (12), (14), and (15).

The situation is different for $\boldsymbol{B}_{\boldsymbol{x}}$. Its two answers $\tilde{\mathbf{y}}_t$ and $\tilde{\mathbf{y}}_u$ come from the *same* denoising trajectory. They are therefore generally dependent, so the expectation cannot be written as the inner product of the two marginal mean embeddings. This difference between $\boldsymbol{A}_{\boldsymbol{x}}$ and $\boldsymbol{B}_{\boldsymbol{x}}$ is precisely what will capture the dependence between diffusion steps.

**Step 2: decomposition of the deployment error.** From Lemma 1 we have

$$\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) \;=\; \mathbb{E}_{\tilde{\mathbf{Y}} \sim \bar{\pi}_{\boldsymbol{\theta}}} \big[\hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}})\big] \tag{B.8}$$

Importantly, this equality only uses the marginal distribution at each step; the intermediate answers do not need to be independent.

We can now add and subtract $\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x})$ inside the deployment error:

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{x}}(\mathbf{w}) &= \mathbb{E}_{\tilde{\mathbf{Y}} \sim \bar{\pi}_{\boldsymbol{\theta}}} \left[ \left\| \hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x}) \right\|^2_{\mathcal{H}_k} \right] \\
&= \mathbb{E} \left[ \left\| \hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) + \boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x}) \right\|^2_{\mathcal{H}_k} \right].
\end{aligned} \tag{B.9}$$

Expanding the squared norm gives

$$\begin{aligned}
\mathcal{E}_{\boldsymbol{x}}(\mathbf{w}) = \mathbb{E} &\left[ \left\| \hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) \right\|^2_{\mathcal{H}_k} \right] \\
&+ \|\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\|^2_{\mathcal{H}_k} \\
&+ 2\,\mathbb{E} \left[ \left\langle \hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}), \boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x}) \right\rangle_{\mathcal{H}_k} \right].
\end{aligned} \tag{B.10}$$

The last term is zero from Lemma 2(a). Indeed, $\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})$ is deterministic for a fixed prompt $\boldsymbol{x}$ and by equation (B.8),

$$\mathbb{E} \left[ \hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) \right] = 0.$$

Hence,

$$\mathcal{E}_{\boldsymbol{x}}(\mathbf{w}) = \underbrace{\|\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\|^2_{\mathcal{H}_k}}_{D_{\boldsymbol{x}}(\mathbf{w})} + \mathbb{E} \left[ \left\| \hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) \right\|^2_{\mathcal{H}_k} \right]. \tag{B.11}$$

Equation (B.11) has a direct interpretation. The first term measures whether the *average distribution* induced by the selected diffusion steps matches the final-answer distribution. The second term measures how much the empirical representation obtained from a *single trajectory* varies around this average. The latter is where the dependence between steps appears.

**Step 3: the distribution-mismatch term.** Using $\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) = \sum_{t=1}^{T} w_t \boldsymbol{\mu}_t(\boldsymbol{x})$, we expand the first term of (B.11):

$$\begin{aligned} D_{\boldsymbol{x}}(\mathbf{w}) &= \left\| \sum_{t=1}^{T} w_t \boldsymbol{\mu}_t(\boldsymbol{x}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x}) \right\|_{\mathcal{H}_k}^2 \\ &= \sum_{t=1}^{T} \sum_{u=1}^{T} w_t w_u \left\langle \boldsymbol{\mu}_t(\boldsymbol{x}), \boldsymbol{\mu}_u(\boldsymbol{x}) \right\rangle_{\mathcal{H}_k} \\ &\quad - 2 \sum_{t=1}^{T} w_t \left\langle \boldsymbol{\mu}_t(\boldsymbol{x}), \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x}) \right\rangle_{\mathcal{H}_k} + \left\| \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x}) \right\|_{\mathcal{H}_k}^2 . \end{aligned} \tag{B.12}$$

Using the identities established in Step 1,

$$D_{\boldsymbol{x}}(\mathbf{w}) = \mathbf{w}^{\top} \boldsymbol{A}_{\boldsymbol{x}} \mathbf{w} - 2 \boldsymbol{b}_{\boldsymbol{x}}^{\top} \mathbf{w} + c_{\boldsymbol{x}}. \tag{B.13}$$

Moreover, $\boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x})$ is the kernel mean embedding of $\pi_{\mathbf{w}}(\cdot \mid \boldsymbol{x})$, while $\boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})$ is the kernel mean embedding of $p_{\boldsymbol{\theta}}^{\mathrm{Diff}}(\cdot \mid \boldsymbol{x})$. Therefore,

$$D_{\boldsymbol{x}}(\mathbf{w}) = \mathrm{MMD}_k^2 \left( \pi_{\mathbf{w}}(\cdot \mid \boldsymbol{x}), p_{\boldsymbol{\theta}}^{\mathrm{Diff}}(\cdot \mid \boldsymbol{x}) \right). \tag{B.14}$$

Thus, $D_{\boldsymbol{x}}(\mathbf{w})$ measures the semantic mismatch between the distribution obtained by mixing diffusion steps according to $\mathbf{w}$ and the distribution of independently generated final answers.

**Step 4: the within-trajectory variance term.** We now consider the second term of (B.11). Using the definitions of $\hat{\boldsymbol{\mu}}_{\mathbf{w}}$ and $\boldsymbol{\mu}_{\mathbf{w}}$, we have

$$\begin{aligned} &\mathbb{E}\left[ \left\| \hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) \right\|_{\mathcal{H}_k}^2 \right] \\ &\quad = \mathbb{E}\left[ \left\| \sum_{t=1}^{T} w_t \left( \phi(\tilde{\mathbf{y}}_t) - \boldsymbol{\mu}_t(\boldsymbol{x}) \right) \right\|_{\mathcal{H}_k}^2 \right] \\ &\quad = \sum_{t=1}^{T} \sum_{u=1}^{T} w_t w_u \, \mathbb{E}\left[ \left\langle \phi(\tilde{\mathbf{y}}_t) - \boldsymbol{\mu}_t(\boldsymbol{x}), \phi(\tilde{\mathbf{y}}_u) - \boldsymbol{\mu}_u(\boldsymbol{x}) \right\rangle_{\mathcal{H}_k} \right]. \end{aligned} \tag{B.15}$$

For each pair $(t, u)$, expand the inner product:

$$\begin{aligned} &\mathbb{E}\left[ \left\langle \phi(\tilde{\mathbf{y}}_t) - \boldsymbol{\mu}_t(\boldsymbol{x}), \phi(\tilde{\mathbf{y}}_u) - \boldsymbol{\mu}_u(\boldsymbol{x}) \right\rangle_{\mathcal{H}_k} \right] \\ &= \mathbb{E}\left[ \left\langle \phi(\tilde{\mathbf{y}}_t), \phi(\tilde{\mathbf{y}}_u) \right\rangle_{\mathcal{H}_k} \right] - \left\langle \boldsymbol{\mu}_t(\boldsymbol{x}), \boldsymbol{\mu}_u(\boldsymbol{x}) \right\rangle_{\mathcal{H}_k}. \end{aligned} \tag{B.16}$$

Indeed, the two cross terms reduce to $\langle \boldsymbol{\mu}_t(\boldsymbol{x}), \boldsymbol{\mu}_u(\boldsymbol{x}) \rangle_{\mathcal{H}_k}$ after taking expectations. By the reproducing property,

$$\left\langle \phi(\tilde{\mathbf{y}}_t), \phi(\tilde{\mathbf{y}}_u) \right\rangle_{\mathcal{H}_k} = k(\tilde{\mathbf{y}}_t, \tilde{\mathbf{y}}_u).$$

The two answers in this kernel evaluation belong to the same trajectory. Therefore, from (13),

$$\mathbb{E}\left[ k(\tilde{\mathbf{y}}_t, \tilde{\mathbf{y}}_u) \right] = [\boldsymbol{B}_{\boldsymbol{x}}]_{tu}.$$

Using Step 1 for the second term of (B.16), we obtain

$$\mathbb{E}\left[ \left\langle \phi(\tilde{\mathbf{y}}_t) - \boldsymbol{\mu}_t(\boldsymbol{x}), \phi(\tilde{\mathbf{y}}_u) - \boldsymbol{\mu}_u(\boldsymbol{x}) \right\rangle_{\mathcal{H}_k} \right] = [\boldsymbol{B}_{\boldsymbol{x}}]_{tu} - [\boldsymbol{A}_{\boldsymbol{x}}]_{tu} = [\boldsymbol{C}_{\boldsymbol{x}}]_{tu}. \tag{B.17}$$

Substituting this identity into (B.15) gives

$$\mathbb{E}\left[ \left\| \hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \boldsymbol{\mu}_{\mathbf{w}}(\boldsymbol{x}) \right\|_{\mathcal{H}_k}^2 \right] = \mathbf{w}^{\top} \boldsymbol{C}_{\boldsymbol{x}} \mathbf{w}. \tag{B.18}$$

This also explains the role of $\boldsymbol{C}_{\boldsymbol{x}}$. The matrix $\boldsymbol{A}_{\boldsymbol{x}}$ compares intermediate answers as if they came from independent trajectories, whereas $\boldsymbol{B}_{\boldsymbol{x}}$ compares answers taken from the same trajectory. Their

difference $\boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{B}_{\boldsymbol{x}} - \boldsymbol{A}_{\boldsymbol{x}}$ therefore measures the additional dependence induced by sharing the same denoising path.

Combining (B.11), (B.13), and (B.18), we obtain

$$\begin{aligned}\mathcal{E}_{\boldsymbol{x}}(\mathbf{w}) &= D_{\boldsymbol{x}}(\mathbf{w}) + \mathbf{w}^\top \boldsymbol{C}_{\boldsymbol{x}} \mathbf{w} \\ &= \mathbf{w}^\top \boldsymbol{A}_{\boldsymbol{x}} \mathbf{w} - 2\boldsymbol{b}_{\boldsymbol{x}}^\top \mathbf{w} + c_{\boldsymbol{x}} + \mathbf{w}^\top \boldsymbol{C}_{\boldsymbol{x}} \mathbf{w} \\ &= \mathbf{w}^\top \big(\boldsymbol{A}_{\boldsymbol{x}} + \boldsymbol{C}_{\boldsymbol{x}}\big)\mathbf{w} - 2\boldsymbol{b}_{\boldsymbol{x}}^\top \mathbf{w} + c_{\boldsymbol{x}}. \end{aligned} \tag{B.19}$$

Since $\boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{B}_{\boldsymbol{x}} - \boldsymbol{A}_{\boldsymbol{x}}$, $\boldsymbol{A}_{\boldsymbol{x}} + \boldsymbol{C}_{\boldsymbol{x}} = \boldsymbol{B}_{\boldsymbol{x}}$, and hence

$$\boxed{\mathcal{E}_{\boldsymbol{x}}(\mathbf{w}) = D_{\boldsymbol{x}}(\mathbf{w}) + \mathbf{w}^\top \boldsymbol{C}_{\boldsymbol{x}} \mathbf{w} = \mathbf{w}^\top \boldsymbol{B}_{\boldsymbol{x}} \mathbf{w} - 2\boldsymbol{b}_{\boldsymbol{x}}^\top \mathbf{w} + c_{\boldsymbol{x}}.} \tag{B.20}$$

This proves the decomposition in equation (11).

**Step 5: positive semidefiniteness.** It remains to prove that $\boldsymbol{A}_{\boldsymbol{x}}$, $\boldsymbol{C}_{\boldsymbol{x}}$, and $\boldsymbol{B}_{\boldsymbol{x}}$ are positive semidefinite. The result holds for every $\boldsymbol{v} \in \mathbb{R}^T$; the vector $\boldsymbol{v}$ does not need to belong to the simplex.

First, Step 1 shows that $\boldsymbol{A}_{\boldsymbol{x}}$ is the Gram matrix of the kernel mean embeddings $\boldsymbol{\mu}_1(\boldsymbol{x}), \dots, \boldsymbol{\mu}_T(\boldsymbol{x})$. Therefore,

$$\boldsymbol{v}^\top \boldsymbol{A}_{\boldsymbol{x}} \boldsymbol{v} = \left\| \sum_{t=1}^{T} v_t \boldsymbol{\mu}_t(\boldsymbol{x}) \right\|_{\mathcal{H}_k}^2 \geq 0, \tag{B.21}$$

and hence $\boldsymbol{A}_{\boldsymbol{x}} \succeq 0$.

Second, using (B.17),

$$\boldsymbol{v}^\top \boldsymbol{C}_{\boldsymbol{x}} \boldsymbol{v} = \mathbb{E}_{\tilde{\mathbf{Y}} \sim \bar{\pi}_{\boldsymbol{\theta}}} \left[ \left\| \sum_{t=1}^{T} v_t \left( \phi(\tilde{\mathbf{y}}_t) - \boldsymbol{\mu}_t(\boldsymbol{x}) \right) \right\|_{\mathcal{H}_k}^2 \right] \geq 0. \tag{B.22}$$

Thus, $\boldsymbol{C}_{\boldsymbol{x}} \succeq 0$.

Finally,

$$\boldsymbol{B}_{\boldsymbol{x}} = \boldsymbol{A}_{\boldsymbol{x}} + \boldsymbol{C}_{\boldsymbol{x}}.$$

Since the sum of two positive semidefinite matrices is positive semidefinite, we also have $\boldsymbol{B}_{\boldsymbol{x}} \succeq 0$.

This completes the proof of Proposition 1. □

### B.3 Proof of Corollary 1

Let $\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)} \overset{\text{i.i.d.}}{\sim} p_{\boldsymbol{\theta}}^{\text{Diff}}(\cdot \mid \boldsymbol{x})$ be independent of $\tilde{\mathbf{Y}}$, and set $\hat{\boldsymbol{\mu}}_{\text{iid}} = \frac{1}{M} \sum_{m=1}^{M} \phi(\mathbf{y}^{(m)})$. All expectations below are over the joint law of $\tilde{\mathbf{Y}}$ and of the $M$ independent samples. Note that $\mathbb{E}[\phi(\mathbf{y}^{(m)})] = \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})$ for every $m$, hence $\mathbb{E}[\hat{\boldsymbol{\mu}}_{\text{iid}}] = \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})$.

**Step 1: splitting at $\boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})$.** Write $\hat{\boldsymbol{\mu}}_{\mathbf{w}} - \hat{\boldsymbol{\mu}}_{\text{iid}} = \big(\hat{\boldsymbol{\mu}}_{\mathbf{w}} - \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})\big) - \big(\hat{\boldsymbol{\mu}}_{\text{iid}} - \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})\big)$ and expand:

$$\mathbb{E}\big\|\hat{\boldsymbol{\mu}}_{\mathbf{w}} - \hat{\boldsymbol{\mu}}_{\text{iid}}\big\|_{\mathcal{H}_k}^2 = \mathbb{E}\big\|\hat{\boldsymbol{\mu}}_{\mathbf{w}} - \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})\big\|_{\mathcal{H}_k}^2 + \mathbb{E}\big\|\hat{\boldsymbol{\mu}}_{\text{iid}} - \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})\big\|_{\mathcal{H}_k}^2 - 2\,\Theta',$$

with $\Theta' = \mathbb{E}\big\langle \hat{\boldsymbol{\mu}}_{\mathbf{w}} - \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x}),\ \hat{\boldsymbol{\mu}}_{\text{iid}} - \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x}) \big\rangle_{\mathcal{H}_k}$. The two arguments are $\mathcal{H}_k$-valued, finitely valued, and independent, since the first is a function of $\tilde{\mathbf{Y}}$ and the second a function of $(\mathbf{y}^{(1)}, \dots, \mathbf{y}^{(M)})$. Lemma 2(b) therefore applies and gives

$$\Theta' = \big\langle \mathbb{E}[\hat{\boldsymbol{\mu}}_{\mathbf{w}}] - \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x}),\ \mathbb{E}[\hat{\boldsymbol{\mu}}_{\text{iid}}] - \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x}) \big\rangle_{\mathcal{H}_k} = 0,$$

the second factor being zero. The first term is $\mathcal{E}_{\boldsymbol{x}}(\mathbf{w})$ by definition (10), the inner expectation over the i.i.d. sample being trivial since $\hat{\boldsymbol{\mu}}_{\mathbf{w}}$ depends on $\tilde{\mathbf{Y}}$ only.

**Step 2: the variance of the i.i.d. estimator.** With $\hat{\boldsymbol{\mu}}_{\mathrm{iid}} - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M}\big(\phi(\mathbf{y}^{(m)}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\big)$,

$$\mathbb{E}\big\|\hat{\boldsymbol{\mu}}_{\mathrm{iid}} - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\big\|^2_{\mathcal{H}_k} = \frac{1}{M^2}\sum_{m=1}^{M}\mathbb{E}\big\|\phi(\mathbf{y}^{(m)}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\big\|^2_{\mathcal{H}_k} + \frac{1}{M^2}\sum_{m\neq m'}\mathbb{E}\big\langle\phi(\mathbf{y}^{(m)}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x}),\, \phi(\mathbf{y}^{(m')}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\big\rangle_{\mathcal{H}_k}. \quad \text{(B.23)}$$

For $m \neq m'$ the two centered elements are independent with zero mean, so each term of the second sum vanishes by Lemma 2(b). The $M$ terms of the first sum are equal, since the $\mathbf{y}^{(m)}$ are identically distributed. Denoting

$$\mathrm{Var}_{\mathcal{H}_k}\big(\phi(\mathbf{y})\big) = \mathbb{E}_{\mathbf{y}\sim p^{\mathrm{Diff}}_{\boldsymbol{\theta}}(\cdot|\boldsymbol{x})}\big\|\phi(\mathbf{y}) - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\big\|^2_{\mathcal{H}_k} = \mathbb{E}\big[k(\mathbf{y},\mathbf{y})\big] - \|\boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\|^2_{\mathcal{H}_k} = 1 - c_{\boldsymbol{x}}, \quad \text{(B.24)}$$

the last equality using the normalization (B.2.) Injecting in equation (B.23), we obtain $\mathbb{E}\|\hat{\boldsymbol{\mu}}_{\mathrm{iid}} - \boldsymbol{\mu}_{\mathrm{ref}}(\boldsymbol{x})\|^2_{\mathcal{H}_k} = \frac{1}{M}\mathrm{Var}_{\mathcal{H}_k}(\phi(\mathbf{y}))$.

**Step 3: conclusion.** Combining Steps 1 and 2,

$$\mathbb{E}\big\|\hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}}) - \hat{\boldsymbol{\mu}}_{\mathrm{iid}}\big\|^2_{\mathcal{H}_k} = \mathcal{E}_{\boldsymbol{x}}(\mathbf{w}) + \frac{1}{M}\mathrm{Var}_{\mathcal{H}_k}\big(\phi(\mathbf{y})\big) = \mathcal{E}_{\boldsymbol{x}}(\mathbf{w}) + \frac{1-c_{\boldsymbol{x}}}{M},$$

which is equation (16). The second term does not depend on $\mathbf{w}$, so the two functions of $\mathbf{w}$ differ by an additive constant on $\Delta(\mathcal{T})$, and a fortiori on $\mathcal{W}_M \subseteq \Delta(\mathcal{T})$, and share the same minimizers there. □

### B.4 Proof of Proposition 2

Throughout this subsection the prompt $\boldsymbol{x}$ is fixed. To lighten notation we drop the subscript $\boldsymbol{x}$ from $\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{C}, \boldsymbol{b}, c$ and from their estimators (19), and write $\phi_t^{(r)} = \phi(\tilde{\mathbf{y}}_t^{(r)})$ for $r \in \{1,\dots,R\}$ and $t \in \mathcal{T}$. The $R \geq 2$ trajectories $\tilde{\mathbf{Y}}^{(1)},\dots,\tilde{\mathbf{Y}}^{(R)}$ are i.i.d. draws from $\bar{\pi}_{\boldsymbol{\theta}}$ so answers read from *different* trajectories are independent and answers read from the *same* trajectory are not. Recall the estimators (19) of Section 4.4,

$$\begin{aligned}
[\widehat{\boldsymbol{B}}]_{tu} &= \frac{1}{R}\sum_{r} k\big(\tilde{\mathbf{y}}_t^{(r)}, \tilde{\mathbf{y}}_u^{(r)}\big), & [\widehat{\boldsymbol{A}}]_{tu} &= \frac{1}{R^2}\sum_{r,r'} k\big(\tilde{\mathbf{y}}_t^{(r)}, \tilde{\mathbf{y}}_u^{(r')}\big),\\
[\widehat{\boldsymbol{b}}]_{t} &= \frac{1}{R(R-1)}\sum_{r\neq r'} k\big(\tilde{\mathbf{y}}_t^{(r)}, \tilde{\mathbf{y}}_1^{(r')}\big), & \widehat{c} &= \frac{1}{R(R-1)}\sum_{r\neq r'} k\big(\tilde{\mathbf{y}}_1^{(r)}, \tilde{\mathbf{y}}_1^{(r')}\big),
\end{aligned} \quad \text{(19)}$$

where $\widehat{\boldsymbol{B}}$ uses same-trajectory pairs, $\widehat{\boldsymbol{b}}$ and $\widehat{c}$ use cross-trajectory pairs, and $\widehat{\boldsymbol{A}}$ uses *all* $R^2$ pairs, same-trajectory ones included ($V$-statistic).

**Step 1: moments of the estimators.**

**Lemma 3** (Expectations of the estimators)**.**

$$\mathbb{E}\big[\widehat{\boldsymbol{B}}\big] = \boldsymbol{B}, \qquad \mathbb{E}\big[\widehat{\boldsymbol{A}}\big] = \boldsymbol{A} + \tfrac{1}{R}\,\boldsymbol{C}, \qquad \mathbb{E}\big[\widehat{\boldsymbol{C}}\big] = \big(1 - \tfrac{1}{R}\big)\boldsymbol{C},$$
$$\mathbb{E}\big[\widehat{\boldsymbol{b}}\big] = \boldsymbol{b}, \qquad \mathbb{E}\big[\widehat{c}\big] = c. \quad \text{(B.25)}$$

*Proof.* Each summand of $[\widehat{\boldsymbol{B}}]_{tu}$ is the kernel evaluated on two answers of the same trajectory $\tilde{\mathbf{Y}}^{(r)} \sim \bar{\pi}_{\boldsymbol{\theta}}$, whose expectation is $[\boldsymbol{B}]_{tu}$ by equation (13). The $R^2$ summands of $[\widehat{\boldsymbol{A}}]_{tu}$ split into $R$ same-trajectory pairs ($r = r'$), each of expectation $[\boldsymbol{B}]_{tu}$, and $R(R-1)$ cross-trajectory pairs ($r \neq r'$), in which $\tilde{\mathbf{y}}_t^{(r)} \sim \pi_{\boldsymbol{\theta},t}$ and $\tilde{\mathbf{y}}_u^{(r')} \sim \pi_{\boldsymbol{\theta},u}$ are independent, each of expectation $[\boldsymbol{A}]_{tu}$ by equation (12). Hence $\mathbb{E}[\widehat{\boldsymbol{A}}] = \frac{1}{R}\boldsymbol{B} + \frac{R-1}{R}\boldsymbol{A} = \boldsymbol{A} + \frac{1}{R}(\boldsymbol{B} - \boldsymbol{A}) = \boldsymbol{A} + \frac{1}{R}\boldsymbol{C}$, and $\mathbb{E}[\widehat{\boldsymbol{C}}] = \boldsymbol{B} - \boldsymbol{A} - \frac{1}{R}\boldsymbol{C} = (1 - \frac{1}{R})\boldsymbol{C}$ by linearity. Finally, each summand of $[\widehat{\boldsymbol{b}}]_t$ pairs $\tilde{\mathbf{y}}_t^{(r)} \sim \pi_{\boldsymbol{\theta},t}$ with an independent $\tilde{\mathbf{y}}_1^{(r')} \sim \pi_{\boldsymbol{\theta},1} = p^{\mathrm{Diff}}_{\boldsymbol{\theta}}$, and has expectation $[\boldsymbol{b}]_t$ by equation (14). Likewise, each summand of $\widehat{c}$ pairs two independent draws from $p^{\mathrm{Diff}}_{\boldsymbol{\theta}}$ and has expectation $c$ by equation (15). □

**Step 2: positive semidefiniteness on every sample (Proposition 2(a)).** Fix any realization $\tilde{\mathbf{Y}}^{(1)}, \dots, \tilde{\mathbf{Y}}^{(R)}$ of the trajectories. Let $\bar{\phi}_t = \frac{1}{R}\sum_r \phi_t^{(r)} \in \mathcal{H}_k$ be the empirical mean embedding of step $t$. By the reproducing property (B.1) and bilinearity of the inner product, $[\widehat{\boldsymbol{A}}]_{tu} = \langle \bar{\phi}_t, \bar{\phi}_u \rangle_{\mathcal{H}_k}$ and $[\widehat{\boldsymbol{B}}]_{tu} = \frac{1}{R}\sum_r \langle \phi_t^{(r)}, \phi_u^{(r)} \rangle_{\mathcal{H}_k}$. By bilinearity and definition of $\bar{\phi}$, we get:

$$\begin{aligned}\frac{1}{R}\sum_r \big\langle \phi_t^{(r)} - \bar{\phi}_t,\ \phi_u^{(r)} - \bar{\phi}_u \big\rangle_{\mathcal{H}_k} &= \frac{1}{R}\sum_r \big\langle \phi_t^{(r)}, \phi_u^{(r)} \big\rangle_{\mathcal{H}_k} - \frac{1}{R}\sum_r \big\langle \phi_t^{(r)}, \bar{\phi}_u \big\rangle_{\mathcal{H}_k} \\ &\qquad - \frac{1}{R}\sum_r \big\langle \bar{\phi}_t, \phi_u^{(r)} \big\rangle_{\mathcal{H}_k} + \frac{1}{R}\sum_r \big\langle \bar{\phi}_t, \bar{\phi}_u \big\rangle_{\mathcal{H}_k} \\ &= \frac{1}{R}\sum_r \big\langle \phi_t^{(r)}, \phi_u^{(r)} \big\rangle_{\mathcal{H}_k} - \big\langle \bar{\phi}_t, \bar{\phi}_u \big\rangle_{\mathcal{H}_k} \\ &= [\widehat{\boldsymbol{B}}]_{tu} - [\widehat{\boldsymbol{A}}]_{tu} = [\widehat{\boldsymbol{C}}]_{tu},\end{aligned}$$

so that $\widehat{\boldsymbol{C}}$ is the empirical covariance matrix of the family $(\phi_t^{(r)})_{t\in\mathcal{T}}$ over the $R$ trajectories. Thus, for every $\boldsymbol{v} \in \mathbb{R}^T$,

$$\boldsymbol{v}^\top \widehat{\boldsymbol{A}}\boldsymbol{v} = \Big\| \sum_t v_t\, \bar{\phi}_t \Big\|_{\mathcal{H}_k}^2 \geq 0, \qquad \boldsymbol{v}^\top \widehat{\boldsymbol{C}}\,\boldsymbol{v} = \frac{1}{R}\sum_r \Big\| \sum_t v_t \big(\phi_t^{(r)} - \bar{\phi}_t\big) \Big\|_{\mathcal{H}_k}^2 \geq 0.$$

Hence $\widehat{\boldsymbol{A}} \succeq 0$ and $\widehat{\boldsymbol{C}} \succeq 0$, and $\widehat{\boldsymbol{A}} + \lambda\widehat{\boldsymbol{C}} \succeq 0$ for all $\lambda \geq 0$. Both matrices are symmetric, since $k$ is. As $\widehat{J}_{\boldsymbol{x}}(\cdot\,;\lambda)$ is a quadratic form with Hessian $2(\widehat{\boldsymbol{A}} + \lambda\widehat{\boldsymbol{C}})$ plus an affine term, it is convex on $\mathbb{R}^T$, a fortiori on $\mathcal{W}_M$. Averages of positive semidefinite matrices being positive semidefinite, the averaged matrices $\overline{\boldsymbol{A}}$ and $\overline{\boldsymbol{C}}$ of equation (20) inherit the property, and (20) is a convex quadratic program.

**Step 3: effective penalty (Proposition 2(b)).** By linearity of the expectation and Lemma 3, for every $\mathbf{w} \in \Delta(\mathcal{T})$,

$$\begin{aligned}\mathbb{E}\big[\widehat{J}_{\boldsymbol{x}}(\mathbf{w};\lambda)\big] &= \mathbf{w}^\top \Big(\boldsymbol{A} + \tfrac{1}{R}\boldsymbol{C} + \lambda\big(1 - \tfrac{1}{R}\big)\boldsymbol{C}\Big)\mathbf{w} - 2\,\boldsymbol{b}^\top\mathbf{w} + c \\ &= \big(\mathbf{w}^\top\boldsymbol{A}\mathbf{w} - 2\,\boldsymbol{b}^\top\mathbf{w} + c\big) + \Big(\lambda + \frac{1-\lambda}{R}\Big)\mathbf{w}^\top\boldsymbol{C}\mathbf{w} = J_{\boldsymbol{x}}(\mathbf{w};\lambda_{\text{eff}}),\end{aligned}$$

using $D_{\boldsymbol{x}}(\mathbf{w}) = \mathbf{w}^\top\boldsymbol{A}\mathbf{w} - 2\boldsymbol{b}^\top\mathbf{w} + c$ from Proposition 1. □

### B.5 Method design details

#### B.5.1 Penalized quadratic problem design

**Cap and penalty.** Since $\pi_{\boldsymbol{\theta},1} = p_{\boldsymbol{\theta}}^{\text{Diff}}$, we have $\boldsymbol{\mu}_1(\boldsymbol{x}) = \boldsymbol{\mu}_{\text{ref}}(\boldsymbol{x})$ and $D_{\boldsymbol{x}}(\mathbf{w}_1) = 0$, where $\mathbf{w}_1$ puts all the mass on the last denoising step. So the mismatch term pulls $\mathbf{w}$ onto the last step. The cap and the penalty counter this in complementary ways. Restricting the quadratic program to the capped simplex $\mathcal{W}_M = \frac{1}{M}\{\boldsymbol{v} \in [0,1]^T : \sum_t v_t = M\}$ bounds concentration by construction: every $\mathbf{w} \in \mathcal{W}_M$ has at least $M$ non-zero weights. Moreover, since the vertices of $\mathcal{W}_M$ are the uniform weightings $\frac{1}{M}\mathbf{1}_{\mathcal{S}^\star}$ of the $M$-step subsets, making equation (20) the convex relaxation of selecting $M$ steps. Importantly, $M\mathbf{w}$ is a valid vector of inclusion probabilities, so a $M$ *distinct* steps set $\mathcal{S}^\star$ can be drawn with $\mathbb{P}(t \in \mathcal{S}^\star) = Mw_t$ (see App. B.5.2). In such a case, $\mathbb{E}\big[\frac{1}{M}\sum_{t\in\mathcal{S}^\star}\phi(\tilde{\mathbf{y}}_t) \mid \tilde{\mathbf{Y}}\big] = \hat{\boldsymbol{\mu}}_{\mathbf{w}}(\tilde{\mathbf{Y}})$, i.e. the equal-weight embedding of the scored answers matches, on average, the weighted embedding the objective $\mathcal{E}_{\boldsymbol{x}}$ controls.

The cap, however, only excludes collapse onto *fewer* than $M$ steps, not onto $M$ near-identical late steps. The penalty $\lambda$ addresses this issue. We choose to tune $\lambda$ rather than fix it at $\lambda = 1$ for multiple reasons. First, although $J_{\boldsymbol{x}}(\cdot\,;1) = \mathcal{E}_{\boldsymbol{x}}$ is the expected squared RKHS distance to the empirical mean embedding of $M$ independent generations (Corollary 1), $\mathcal{E}_{\boldsymbol{x}}$ is only a surrogate: $U_{\text{sem}}$ is not a function of the kernel mean embedding, and is even computed using a different semantic similarity score function. Then, the step-selection rules such as $\text{Top}_M$ or coverage do not follow the faithful i.i.d sampling from $\text{Cat}(\cdot;\mathbf{w})$. Finally, we use a single $\mathbf{w}$ on all prompts of a task. Since AUROC and PRR scores are related to rank ordering, the choice of $\lambda$ can significantly impact these metrics. However, $\lambda$ is tuned on a hold-out validation set and use no labeled prompts (see App. D.2).

**Effective penalty and the unbiased estimator of $\boldsymbol{A}$.** $\widehat{\boldsymbol{A}}$ keeps the $R$ same-trajectory pairs ($V$-statistic). Using only cross-trajectory pairs, like $\widehat{\boldsymbol{b}}$ and $\widehat{c}$, gives the unbiased estimator ($U$-statistic) $[\widehat{\boldsymbol{A}}^U]_{tu} = \frac{1}{R(R-1)} \sum_{r \neq r'} k\big(\tilde{\mathbf{y}}_t^{(r)}, \tilde{\mathbf{y}}_u^{(r')}\big)$. Since $\widehat{\boldsymbol{A}} = \frac{1}{R}\widehat{\boldsymbol{B}} + \frac{R-1}{R}\widehat{\boldsymbol{A}}^U$,

$$\widehat{\boldsymbol{A}}^U = \widehat{\boldsymbol{A}} - \frac{1}{R-1}\,\widehat{\boldsymbol{C}}, \qquad \widehat{\boldsymbol{C}}^U = \widehat{\boldsymbol{B}} - \widehat{\boldsymbol{A}}^U = \frac{R}{R-1}\,\widehat{\boldsymbol{C}}, \tag{B.26}$$

the latter being the Bessel-corrected empirical covariance, unbiased for $\boldsymbol{C}$. The unbiased plug-in objective $\widehat{J}_{\boldsymbol{x}}^U(\mathbf{w}; \lambda_{\text{eff}}) = \mathbf{w}^\top\big(\widehat{\boldsymbol{A}}^U + \lambda_{\text{eff}}\,\widehat{\boldsymbol{C}}^U\big)\mathbf{w} - 2\,\widehat{\boldsymbol{b}}^\top\mathbf{w} + \widehat{c}$ thus satisfies $\mathbb{E}\big[\widehat{J}_{\boldsymbol{x}}^U(\mathbf{w}; \lambda_{\text{eff}})\big] = J_{\boldsymbol{x}}(\mathbf{w}; \lambda_{\text{eff}})$, and substituting (B.26) gives, on every sample and for every $\mathbf{w}$,

$$\widehat{J}_{\boldsymbol{x}}(\mathbf{w}; \lambda) = \widehat{J}_{\boldsymbol{x}}^U\big(\mathbf{w}; \lambda_{\text{eff}}\big), \qquad \lambda_{\text{eff}} = \lambda + \frac{1-\lambda}{R} = \frac{1 + \lambda(R-1)}{R}. \tag{B.27}$$

Solving the quadratic program with the $V$-statistic at penalty $\lambda$ is therefore the *same problem*, sample by sample, as solving it with the unbiased estimator at penalty $\lambda_{\text{eff}}$, and Proposition 2(b) follows at once. The map $\lambda \mapsto \lambda_{\text{eff}}$ is an increasing affine bijection from $[0, \infty)$ onto $[1/R, \infty)$, so selecting $\lambda$ amounts to selecting $\lambda_{\text{eff}}$. Its fixed point $\lambda = 1$ is where $\widehat{J}_{\boldsymbol{x}}(\mathbf{w}; 1) = \mathbf{w}^\top\widehat{\boldsymbol{B}}\mathbf{w} - 2\widehat{\boldsymbol{b}}^\top\mathbf{w} + \widehat{c}$ is an unbiased estimate of the deployment error $\mathcal{E}_{\boldsymbol{x}}(\mathbf{w})$ itself, while the grid value $\lambda = 0$ still penalizes redundancy at strength $1/R$ (0.05 at $R = 20$). Reaching $\lambda_{\text{eff}} \in [0, 1/R)$ would require $\lambda < 0$, where $\widehat{\boldsymbol{A}} + \lambda\widehat{\boldsymbol{C}}$ may be indefinite and convexity is no longer guaranteed. This is why we parametrize the problem through the $V$-statistic: the grid $\lambda \geq 0$ automatically stays in the region where Proposition 2(a) guarantees convexity.

### B.5.2 Step selection rules

Deployment turns the learned weights $\mathbf{w}^\star \in \mathcal{W}_M$ into a collection of steps $\mathcal{S}^\star$, whose answers form $\widehat{\mathcal{Y}}_{\text{DIMS}}(\boldsymbol{x})$ of (21). We evaluate six rules. Three spend the fixed budget of $M{=}20$ steps:

- $\text{Top}_M$: the $M$ highest-weight steps. Its uniform weighting $\frac{1}{M}\mathbf{1}_{\mathcal{S}^\star}$ is a vertex of $\mathcal{W}_M$, hence a feasible point of (20), and it recovers $\mathbf{w}^\star$ exactly whenever $\mathbf{w}^\star$ is itself a vertex.
- **Draw with replacement (WR)**: $t_1, \ldots, t_M \overset{\text{i.i.d.}}{\sim} \text{Cat}(\cdot; \mathbf{w}^\star)$, duplicates kept, so that $\mathcal{S}^\star$ may contain the same step several times. This is the faithful rule: the selected answers are $M$ i.i.d. draws from the empirical mixture (5). A repeated step, however, spends part of the budget on a duplicated answer.
- **Draw without replacement (WOR)**: $M$ distinct steps with inclusion probabilities $\mathbb{P}(t \in \mathcal{S}^\star) = M\,w_t^\star$, obtained by Madow's systematic sampling (see below for a short description). Such a design exists because the cap $w_t^\star \leq 1/M$ makes $M\mathbf{w}^\star$ a valid vector of inclusion probabilities. It yields, conditionally on the trajectory, the same mean embedding as WR rule (App. B.5.1), without duplicates.

The other three let the weights dictate the step budget: for a coverage level $c \in \{50\%, 75\%, 90\%\}$, take the smallest set of highest-weight steps whose mass reaches $c$. The two rules the main tables report are therefore

$$\mathcal{S}^\star_{\text{top}} = \text{Top}_M\big(\mathbf{w}^\star\big), \qquad \mathcal{S}^\star_{\text{cov}} = \arg\min\Big\{|\mathcal{S}| : \textstyle\sum_{t \in S} w_t^\star \geq 0.75\Big\}. \tag{B.28}$$

Their sizes contrast: the first is budgeted at $M$, the second is a property of how concentrated $\mathbf{w}^\star$ is rather than a budget we impose, and it varies across runs. Ties are broken by ascending step index, so a coverage rule is deterministic even at a uniform $\mathbf{w}^\star$. App. E.4.1 compares all six rules.

**Madow's systematic sampling.** Madow's systematic sampling is a method for drawing a fixed number of distinct samples from a population, where each unit has a known inclusion probability. Given inclusion probabilities $P_t = M\,w_t^\star \in [0, 1]$ with $\sum_t P_t = M$, the procedure of Madow (1949) draws exactly $M$ distinct steps with $\mathbb{P}(t \in \mathcal{S}^\star) = P_t$, from a single uniform random number. List the steps in a fixed order $t_1, \ldots, t_T$ and lay their probabilities end to end on $[0, M)$: step $t_k$ receives the interval $I_k = [\Pi_{k-1}, \Pi_k)$, where $\Pi_0 = 0$ and $\Pi_k = \sum_{j \leq k} P_{t_j}$, so that $|I_k| = P_{t_k}$ and

$\Pi_T = M$. Draw $u \sim \mathcal{U}([0,1))$ and keep every step whose interval contains one of the $M$ equally spaced points $u, u+1, \ldots, u+M-1$:

$$\mathcal{S}^\star = \{t_k : I_k \cap \{u+m\}_{m=0}^{M-1} \neq \emptyset\}.$$

Three properties follow from the unit spacing of the points. *(i) Distinct steps:* an interval of length $P_{t_k} \leq 1$ contains at most one point. *(ii) Exactly $M$ steps:* the $M$ points $u+m$ lie in $[0, M)$, which the intervals partition, so each point selects one step, and by (i) no step is selected twice. *(iii) Exact inclusion probabilities:* as $u$ is uniform on $[0,1)$, the grid $u+\mathbb{Z}$ meets an interval of length $P_{t_k} \leq 1$ with probability $P_{t_k}$, and its points inside $[0, M)$ are exactly $u, \ldots, u+M-1$. Only the inclusion probabilities of single steps are controlled: joint inclusion depends on the chosen order, and two steps whose intervals fit in a window shorter than one are never selected together. The order therefore does not change the *mean* of the selected embedding given the trajectory, $\mathbb{E}\big[\frac{1}{M}\sum_{t\in\mathcal{S}^\star}\phi(\tilde{\mathbf{y}}_t) \mid \tilde{\mathbf{Y}}\big] = \frac{1}{M}\sum_t \mathbb{P}(t \in \mathcal{S}^\star)\,\phi(\tilde{\mathbf{y}}_t) = \hat{\boldsymbol{\mu}}_{\mathbf{w}^\star}(\tilde{\mathbf{Y}})$ (App. B.5.1), which depends on the single-step probabilities only. It only changes the spread around this mean, whose second moment $\mathbb{E}\big[\|\sum_{t\in\mathcal{S}^\star}\phi(\tilde{\mathbf{y}}_t)\|^2_{\mathcal{H}_k} \mid \tilde{\mathbf{Y}}\big] = \sum_{s,t}\mathbb{P}(s,t\in\mathcal{S}^\star)\,k(\tilde{\mathbf{y}}_s,\tilde{\mathbf{y}}_t)$ involves the joint probabilities.

# C Uncertainty quantification for text generation

This appendix reviews the uncertainty estimators evaluated in Section 5, together with the related scores they build on, in the notation of Section 3. For each estimator we recall the definition given in the original paper, then state how we instantiate it, since several estimators come in variants that the literature does not always distinguish.

Throughout, $\boldsymbol{x}$ is a prompt and $\mathbf{y} = (\mathrm{y}^1, \ldots, \mathrm{y}^L) \in \mathcal{V}^L$ an answer, with superscripts indexing token positions. As in Section 3, variable length is handled by an end-of-sequence token. We write $L_{\text{eff}} \leq L$ for the number of positions before the first end-of-sequence token of the answer being scored. Every length-normalized score below averages over these $L_{\text{eff}}$ positions, except the trajectory scores TrajNLL and TrajEnt (App. C.3), which average over all $L$ positions. We write $\boldsymbol{z}_t$ for the state at denoising step $t$, $\tilde{\mathbf{y}}_t$ for the intermediate answer estimate (1), and $\hat{\mathbf{y}}$ for the *greedy answer*, i.e. the final answer of the trajectory decoded at temperature $\tau = 0$ (App. D.1). Semantic-consistency estimators take a set $\mathcal{Y} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}\}$ and are written $U(\boldsymbol{x}) = U_{\text{sem}}(\mathcal{Y})$, as in (3). In the reference condition, $\mathcal{Y}$ is drawn i.i.d. from $p^{\text{Diff}}_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})$ and $M = 20$. Under the single-trajectory strategies of Section 5.1 and App. D.2, it is read off one trajectory, e.g. $\widehat{\mathcal{Y}}_{\text{DiMS}}(\boldsymbol{x})$ in (21), and $M = |\mathcal{S}|$ is the number of selected steps, which is not always 20 (for instance under the coverage rule of DiMS); if a step is selected twice, its answer is included twice. The estimators below do not depend on how $\mathcal{Y}$ was obtained, with two exceptions that we point out: Semantic Entropy also needs a likelihood for each answer, and AD and CoCoA also need $\hat{\mathbf{y}}$.

## C.1 Uncertainty scores and their evaluation

**Uncertainty and confidence.** Lin et al. (2024) distinguish two kinds of scores. An *uncertainty* score $U(\boldsymbol{x})$ depends on the prompt alone and measures the spread of the answers the model could give to $\boldsymbol{x}$. A *confidence* score $C(\boldsymbol{x}, \mathbf{y})$ refers to one particular generation and is therefore a function of the pair $(\boldsymbol{x}, \mathbf{y})$. To use a single orientation, we write both as $U(\mathbf{y} \mid \boldsymbol{x})$, with larger values meaning a less reliable answer, and drop $\mathbf{y}$ when the score does not depend on it. Scores for which larger values mean *more* reliable are negated.

**Evaluation.** Uncertainty scores are judged by how well they flag unreliable generations. In this work, the answer whose reliability we assess is the *greedy answer* $\hat{\mathbf{y}}$. In the selective-prediction protocol (Ielanskyi et al., 2026; Vashurin et al., 2025a), we are given prompts $\{\boldsymbol{x}_n\}_{n=1}^N$, reference answers $\{\mathbf{y}_n^{\text{ref}}\}_{n=1}^N$, generations $\{\hat{\mathbf{y}}_n\}_{n=1}^N$ and a binary correctness function $c(\boldsymbol{x}_n, \mathbf{y}_n^{\text{ref}}, \hat{\mathbf{y}}_n) \in \{0,1\}$. The quality of an estimator is the association between its scores and the error indicator $1-c$:

$$\text{Assoc}\Big[\big(U(\hat{\mathbf{y}}_n \mid \boldsymbol{x}_n)\big)_{n=1}^N,\ \big(1 - c(\boldsymbol{x}_n, \mathbf{y}_n^{\text{ref}}, \hat{\mathbf{y}}_n)\big)_{n=1}^N\Big], \tag{C.1}$$

where Assoc is the area under the ROC curve (AUROC) or the Prediction-Rejection Ratio (PRR) (Malinin & Gales, 2021); both metrics, the correctness judges and the confidence intervals are detailed in App. D.3. The estimator never sees $\mathbf{y}_n^{\text{ref}}$, which only enters through the correctness label.

*Remark* 1 (Rank equivalence). AUROC and PRR depend on the scores only through the order they induce on the $N$ generations. Two estimators related by a strictly increasing map are therefore indistinguishable under (C.1). We use this below to identify scores that look different but are equivalent, and we point out where one of our variants is *not* rank-equivalent to its published form (MSP, App. C.2; DegMat, App. C.4).

**Three families.** We group the estimators by the signal they use and by the number of generations they need. *Information-based* estimators (App. C.2) aggregate token-level predictive distributions of a single generation. *Diffusion-specific* estimators (App. C.3) are also information-based: they too aggregate token-level predictive distributions of a single generation, but exploit its denoising trajectory, which has no autoregressive counterpart. *Semantic-consistency* estimators (App. C.4) compare a set $\mathcal{Y}$ of $M$ answers to the same prompt. They are among the strongest methods in recent benchmarks (Vashurin et al., 2025a), but are more expensive, since they need multiple generations. Table C.1 lists the inputs of each estimator.

### C.2 Information-based estimators

These estimators turn token-level predictive distributions into a sequence-level score. We first state them for autoregressive (AR) models, for which the left-to-right factorization

$$p_{\boldsymbol{\theta}}^{\mathrm{AR}}(\mathbf{y} \mid \boldsymbol{x}) \;=\; \prod_{\ell=1}^{L_{\mathrm{eff}}} f_{\boldsymbol{\theta}}^{\ell}\big(\mathrm{y}^{\ell} \mid \boldsymbol{x}, \mathbf{y}^{<\ell}\big), \qquad \mathbf{y}^{<\ell} = (\mathrm{y}^{1}, \ldots, \mathrm{y}^{\ell-1}), \tag{C.2}$$

holds exactly. How we compute them for an LLDM is described at the end of the subsection.

**Maximum sequence probability (MSP) and G-NLL.** Aichberger et al. (2026) derive uncertainty measures from proper scoring rules. For the *zero–one* score, the measure is $1 - p_{\boldsymbol{\theta}}^{\mathrm{AR}}(\mathbf{y}^{\star} \mid \boldsymbol{x})$, the probability that an answer sampled from the model differs from the most likely answer $\mathbf{y}^{\star}$. Its rank-equivalent logarithmic form is

$$U_{\mathrm{MSP}}(\boldsymbol{x}) = -\max_{\mathbf{y} \in \mathcal{V}^{L}} \log p_{\boldsymbol{\theta}}^{\mathrm{AR}}(\mathbf{y} \mid \boldsymbol{x}) = -\log p_{\boldsymbol{\theta}}^{\mathrm{AR}}(\mathbf{y}^{\star} \mid \boldsymbol{x}), \quad \mathbf{y}^{\star} = \arg\max_{\mathbf{y} \in \mathcal{V}^{L}} p_{\boldsymbol{\theta}}^{\mathrm{AR}}(\mathbf{y} \mid \boldsymbol{x}). \tag{C.3}$$

This gives a principled justification for maximum sequence probability, a long-standing heuristic baseline implemented, e.g., in LM-Polygraph (Fadeeva et al., 2023). The maximization ranges over $|\mathcal{V}|^{L}$ sequences, so $\mathbf{y}^{\star}$ is intractable. Aichberger et al. (2026) approximate it by greedy decoding, which moves the maximization inside the product, and define G-NLL on the greedy answer $\hat{\mathbf{y}}$, with $\hat{\mathrm{y}}^{\ell} = \arg\max_{v \in \mathcal{V}} f_{\boldsymbol{\theta}}^{\ell}(v \mid \boldsymbol{x}, \hat{\mathbf{y}}^{<\ell})$:

$$U_{\text{G-NLL}}(\boldsymbol{x}) \;=\; -\sum_{\ell=1}^{L_{\mathrm{eff}}} \log\Big(\max_{v \in \mathcal{V}} f_{\boldsymbol{\theta}}^{\ell}\big(v \mid \boldsymbol{x}, \hat{\mathbf{y}}^{<\ell}\big)\Big) \;=\; -\log p_{\boldsymbol{\theta}}^{\mathrm{AR}}(\hat{\mathbf{y}} \mid \boldsymbol{x}) \;\approx\; U_{\mathrm{MSP}}(\boldsymbol{x}). \tag{C.4}$$

The second equality is exact, since G-NLL is the negative log-likelihood of the greedy answer; the only approximation is $\hat{\mathbf{y}} \approx \mathbf{y}^{\star}$. The authors find this single-sequence score competitive with multi-sample methods at a fraction of their cost. An LLDM does not provide the causal conditionals in (C.4), so we do not report G-NLL itself; our MSP feature, described below, is its closest counterpart.

**Perplexity (PPL).** Sequence log-likelihoods favor short answers. Normalizing by length gives the perplexity; Fomicheva et al. (2020) use the length-normalized log-probability of the output as an unsupervised quality estimate:

$$U_{\mathrm{PPL}}(\mathbf{y} \mid \boldsymbol{x}) \;=\; \exp\left(-\frac{1}{L_{\mathrm{eff}}} \sum_{\ell=1}^{L_{\mathrm{eff}}} \log f_{\boldsymbol{\theta}}^{\ell}\big(\mathrm{y}^{\ell} \mid \boldsymbol{x}, \mathbf{y}^{<\ell}\big)\right) \;=\; \exp\Big(-\tfrac{1}{L_{\mathrm{eff}}} \log p_{\boldsymbol{\theta}}^{\mathrm{AR}}(\mathbf{y} \mid \boldsymbol{x})\Big). \tag{C.5}$$

PPL is a strictly increasing function of the length-normalized negative log-likelihood, so the two are rank-equivalent (Remark 1).

**Mean token entropy (MTE).** Instead of scoring the generated tokens, MTE averages the entropy of the predictive distribution over positions ("Softmax-Ent" in Fomicheva et al., 2020):

$$U_{\mathrm{MTE}}(\mathbf{y} \mid \boldsymbol{x}) \;=\; \frac{1}{L_{\mathrm{eff}}} \sum_{\ell=1}^{L_{\mathrm{eff}}} H\big(f_{\boldsymbol{\theta}}^{\ell}(\cdot \mid \boldsymbol{x}, \mathbf{y}^{<\ell})\big), \qquad H(p) = -\sum_{v \in \mathcal{V}} p(v) \log p(v). \tag{C.6}$$

**Computing these scores for an LLDM.** Equations (C.3)–(C.6) assume the causal conditionals $f_{\boldsymbol{\theta}}^{\ell}(\cdot \mid \boldsymbol{x}, \mathbf{y}^{<\ell})$. AR models provide them, but LLDMs do not: generation is parallel and non-causal, and $p_{\boldsymbol{\theta}}^{\mathrm{Diff}}$ is intractable (Section 3.1). We replace the causal conditionals by the token distributions of the *last* denoising step, $f_{\boldsymbol{\theta}}^{\ell}(\cdot \mid \boldsymbol{x}, \boldsymbol{z}_1)$, read off the greedy trajectory and restricted to the first $L_{\mathrm{eff}}$ positions. Three points matter when reading the results of Section 5:

- **Over-confidence.** At the last step, almost every position is already committed. The model attends bidirectionally to these committed tokens and assigns them near-degenerate probabilities, so the resulting scores are expected to be over-confident. This motivates the diffusion-specific scores of App. C.3.
- **Truncated vocabulary.** We store only the top-$k$ logits of each position ($k = 64$), so MTE and MSP are computed from a renormalized top-$k$ distribution. The entropy in (C.6) is therefore slightly underestimated. Perplexity uses the full-vocabulary log-probability of the generated token, and TrajEnt (C.10) is computed during decoding, over the full vocabulary.
- **MSP is a per-position score.** Our MSP feature is the mean, over the retained positions, of the maximum token probability, negated so that larger values mean more uncertainty:

$$U_{\mathrm{MSP}}^{\mathrm{pos}}(\boldsymbol{x}) \;=\; -\frac{1}{L_{\mathrm{eff}}} \sum_{\ell=1}^{L_{\mathrm{eff}}} \max_{v \in \mathcal{V}} f_{\boldsymbol{\theta}}^{\ell}\big(v \mid \boldsymbol{x}, \boldsymbol{z}_1\big). \tag{C.7}$$

  It is not the sequence-level quantity of (C.3). G-NLL (C.4) is $L_{\mathrm{eff}}$ times the negative log of the *geometric* mean of the per-position maxima, whereas our feature uses their *arithmetic* mean. The two are not rank-equivalent across prompts, so (C.3) motivates our feature but does not define it.

### C.3 Diffusion-specific estimators

These estimators use quantities that only a denoising trajectory provides. Throughout, $\boldsymbol{z}_T, \ldots, \boldsymbol{z}_0 = \mathbf{y}$ is the trajectory that produced $\mathbf{y}$, and $t_\ell$ is the step at which position $\ell$ is committed, i.e. the smallest $t$ with $\ell \in \mathcal{M}_t = \{\ell : z_t^{\ell} = \texttt{[MASK]}\}$.

**Monte Carlo negative log-likelihood (MCNLL).** Nie et al. (2025) evaluate conditional likelihoods with a Monte Carlo estimate of the variational *upper* bound on the negative log-likelihood of the answer, obtained by re-noising the answer; Vazhentsev et al. (2026) use it as an uncertainty score under the name MCNLL. We compute it on the greedy answer $\hat{\mathbf{y}}$. For each of $N_{\mathrm{mc}}$ replicates $j$, draw a mask count $n_j \sim \mathcal{U}(\{1, \ldots, L_{\mathrm{eff}}\})$, then mask $n_j$ of the first $L_{\mathrm{eff}}$ positions of $\hat{\mathbf{y}}$, chosen uniformly without replacement, giving a noisy state $\boldsymbol{z}^{(j)}$. The estimate is

$$U_{\mathrm{MCNLL}}(\hat{\mathbf{y}} \mid \boldsymbol{x}) \;=\; -\frac{1}{N_{\mathrm{mc}}} \sum_{j=1}^{N_{\mathrm{mc}}} \frac{L_{\mathrm{eff}}}{n_j} \sum_{\ell=1}^{L_{\mathrm{eff}}} \mathbf{1}\big[z^{(j),\ell} = \texttt{[MASK]}\big] \, \log f_{\boldsymbol{\theta}}^{\ell}\big(\hat{\mathrm{y}}^{\ell} \mid \boldsymbol{x}, \boldsymbol{z}^{(j)}\big). \tag{C.8}$$

Drawing a mask count rather than a masking ratio is the lower-variance form that LLaDA uses for likelihood evaluation. The bound holds for the model distribution defined by a reverse process that unmasks positions in uniformly random order. Under low-confidence remasking, $p_{\boldsymbol{\theta}}^{\mathrm{Diff}}$ differs from that distribution, and MCNLL is a proxy rather than a bound. We report the length-normalized variant $U_{\mathrm{MCNLL}}^{\mathrm{norm}} = U_{\mathrm{MCNLL}}/L_{\mathrm{eff}}$, which makes answers of different lengths comparable, with $N_{\mathrm{mc}} = 100$. MCNLL is a diffusion-aware stand-in for $-\log p_{\boldsymbol{\theta}}^{\mathrm{Diff}}(\hat{\mathbf{y}} \mid \boldsymbol{x})$, but it costs $N_{\mathrm{mc}}$ additional forward passes, a non-negligible overhead. We do not use it inside Semantic Entropy, which relies on the cheaper per-step likelihood $\rho_t$ defined below.

**Trajectory likelihood and entropy (TrajNLL, TrajEnt).** Denoising produces a predictive distribution at every step, so token-level statistics can be averaged along the trajectory instead of read at

its end (Vazhentsev et al., 2026). With $\tilde{\mathbf{y}}_t$ the intermediate estimate from (1),

$$U_{\mathrm{TrajNLL}}(\mathbf{y} \mid \boldsymbol{x}) = -\frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{\ell=1}^{L}\log f_{\boldsymbol{\theta}}^{\ell}\big(\tilde{\mathrm{y}}_t^{\ell} \mid \boldsymbol{x}, \boldsymbol{z}_t\big), \tag{C.9}$$

$$U_{\mathrm{TrajEnt}}(\mathbf{y} \mid \boldsymbol{x}) = \frac{1}{T}\sum_{t=1}^{T}\frac{1}{L}\sum_{\ell=1}^{L} H\big(f_{\boldsymbol{\theta}}^{\ell}(\cdot \mid \boldsymbol{x}, \boldsymbol{z}_t)\big). \tag{C.10}$$

Unlike the other length-normalized scores of this appendix, both average over all $L$ generated positions, including those after the first end-of-sequence token. They are the trajectory-wide counterparts of the length-normalized negative log-likelihood and of MTE: at every step, TrajNLL scores the intermediate estimate $\tilde{\mathbf{y}}_t$ rather than the final answer. Only TrajEnt appears in our tables; TrajNLL is recalled because the per-step likelihood $\rho_t$ defined below builds on it. Vazhentsev et al. (2026) average over the decoding blocks of block-diffusion models. We average over the whole answer instead, which is equivalent in our setting, since every model decodes the answer as a single block (App. D.1).

**Commitment likelihood (CommitNLL).** Equations (C.9)–(C.10) weight all steps equally, including early steps at which the model has almost no committed context. Vazhentsev et al. (2026) instead read each token's probability at the step $t_\ell$ at which it was committed:

$$U_{\mathrm{CommitNLL}}(\mathbf{y} \mid \boldsymbol{x}) \;=\; -\frac{1}{L_{\mathrm{eff}}}\sum_{\ell=1}^{L_{\mathrm{eff}}}\log f_{\boldsymbol{\theta}}^{\ell}\big(\mathrm{y}^{\ell} \mid \boldsymbol{x}, \boldsymbol{z}_{t_\ell}\big). \tag{C.11}$$

CommitNLL is the closest LLDM analogue of the AR sequence negative log-likelihood: each factor is read when the decision is made, before the token enters its own conditioning context. This avoids the over-confidence of the last-step scores of App. C.2. It is not a likelihood, however: the factors are conditioned on different states, and their product is not $p_{\boldsymbol{\theta}}^{\mathrm{Diff}}(\mathbf{y} \mid \boldsymbol{x})$.

**Per-step likelihood proxy ($\rho_t$).** Semantic entropy needs a likelihood for every element of $\mathcal{Y}$, including intermediate answers (App. C.4). We build one, $\rho_t$, from both scores above. As in TrajNLL, it scores the intermediate answer $\tilde{\mathbf{y}}_t$ at step $t$; as in CommitNLL, a committed token is read at its own commitment step. The intermediate answer $\tilde{\mathbf{y}}_t$ is truncated at its first end-of-sequence token, and $\rho_t$ is the mean of the retained token log-probabilities:

$$\rho_t = \frac{1}{L_{\mathrm{eff}}}\sum_{\ell=1}^{L_{\mathrm{eff}}} r_t^{\ell}, \qquad r_t^{\ell} = \begin{cases} \log f_{\boldsymbol{\theta}}^{\ell}\big(\tilde{\mathrm{y}}_t^{\ell} \mid \boldsymbol{x}, \boldsymbol{z}_{t_\ell}\big) & \ell \notin \mathcal{M}_t, \\ \log \max_{v \in \mathcal{V}} f_{\boldsymbol{\theta}}^{\ell}\big(v \mid \boldsymbol{x}, \boldsymbol{z}_t\big) & \ell \in \mathcal{M}_t, \end{cases} \tag{C.12}$$

where $L_{\mathrm{eff}}$ is that of $\tilde{\mathbf{y}}_t$. A position committed before step $t$ thus keeps the log-probability it received at its own commitment step $t_\ell$ instead of being re-scored under $\boldsymbol{z}_t$, and masked positions use the current arg-max log-probability.

### C.4 Semantic-consistency estimators

**Semantic similarity.** Information-based scores ignore *semantic equivalence*: a model that is sure of the meaning of its answer but not of its wording is scored as uncertain (Kuhn et al., 2023). Consistency-based methods instead measure how much the answers in $\mathcal{Y}$ differ in meaning. All of them rely on the semantic similarity $s : \mathcal{V}^L \times \mathcal{V}^L \to [0, 1]$ of Section 3.2, with $s(\mathbf{y}, \mathbf{y}) = 1$, which can be computed in several ways:

- With a natural language inference (NLI) model, such as `microsoft/deberta-v2-xlarge-mnli` (He et al., 2021). It maps an ordered pair to a distribution $\mathrm{NLI}(\mathbf{y}, \mathbf{y}') \in \Delta(\{\mathtt{ent}, \mathtt{neu}, \mathtt{con}\})$ over entailment, neutral and contradiction. Entailment is directional, so both orders are combined. The *hard*, or $0/1$, score sets $s(\mathbf{y}, \mathbf{y}') = 1$ if both directions are predicted as `ent` (e.g., by the arg-max of the NLI output), and 0 otherwise (Kuhn et al., 2023). The *soft* score is a weighted sum of the predicted class probabilities: $s(\mathbf{y}, \mathbf{y}') = \frac{1}{2}\boldsymbol{\alpha}^{\top}\big(\mathrm{NLI}(\mathbf{y}, \mathbf{y}') + \mathrm{NLI}(\mathbf{y}', \mathbf{y})\big)$, with class weights such as $\boldsymbol{\alpha} = (1, \frac{1}{2}, 0)$.

- With a cross-encoder trained for semantic textual similarity (STS), such as `cross-encoder/stsb-roberta-large` (Liu et al., 2019), which scores the ordered pair directly. The score is symmetrized by averaging the two orders and clipped to $[0, 1]$.
- With a similarity, e.g., cosine similarity, between sequence embeddings, $s(\mathbf{y}, \mathbf{y}') = \mathrm{sim}\big(e(\mathbf{y}), e(\mathbf{y}')\big)$, with $e : \mathcal{V}^L \to \mathbb{R}^d$ as in Section 4.2.

Our choices are as follows (the models are listed in App. D.2). Semantic entropy uses the hard NLI score to form clusters. The graph and kernel estimators (DegMat, Eccentricity, EigenVal and KLE) build their graph (C.15) on the STS similarity. AD and CoCoA use the soft NLI score with $\boldsymbol{\alpha} = (1, \frac{1}{2}, 0)$.

**Semantic entropy (SE) and discrete semantic entropy (DSE).** Kuhn et al. (2023) partition $\mathcal{Y}$ into semantic classes $C_1, \dots, C_K$ ($K \le M$) with a greedy procedure. Each answer is compared, in both directions, with one representative of each existing class. It joins the first class for which both directions are entailments, and otherwise starts a new class. Sequence probabilities are summed within each class, and SE is the entropy of the resulting distribution over classes:

$$U_{\mathrm{SE}}(\boldsymbol{x}) \;=\; -\sum_{k=1}^{K} \bar{p}_k \log \bar{p}_k, \qquad \bar{p}_k = \frac{\hat{p}_k}{\sum_{j=1}^{K} \hat{p}_j}, \qquad \hat{p}_k = \sum_{\mathbf{y} \in C_k} p_{\boldsymbol{\theta}}^{\mathrm{AR}}(\mathbf{y} \mid \boldsymbol{x}). \tag{C.13}$$

This requires the sequence likelihood $p_{\boldsymbol{\theta}}^{\mathrm{AR}}$, which is precisely what an LLDM does not provide. In its place we use the exponentiated length-normalized log-likelihood attached to each element of $\mathcal{Y}$: $\rho_1$ of its own trajectory for i.i.d. samples, and $\rho_t$ (C.12) for an intermediate answer $\tilde{\mathbf{y}}_t$. The reference and single-trajectory conditions therefore differ only in the answers being clustered, not in how the entropy is computed.[3] Farquhar et al. (2024) also introduce a *discrete* variant (DSE), which replaces $\bar{p}_k$ by the empirical class frequency and needs no likelihood at all. We do not report it, but we define it because it is the estimator that KLE generalizes:

$$U_{\mathrm{DSE}}(\boldsymbol{x}) \;=\; -\sum_{k=1}^{K} \frac{|C_k|}{M} \log \frac{|C_k|}{M}. \tag{C.14}$$

Both scores rely on a hard partition. The graph scores below use graded similarities instead.

**The semantic graph.** The graph and kernel estimators share one weighted graph on $\mathcal{Y}$, built from the similarity $s$:

$$W_{mm'} = 2\, s\big(\mathbf{y}^{(m)}, \mathbf{y}^{(m')}\big) \in [0, 2] \quad (m \neq m'), \qquad W_{mm} = 0, \tag{C.15}$$

with degree matrix $\boldsymbol{D} = \mathrm{diag}(\boldsymbol{W}\mathbf{1})$. The original papers of these estimators, Lin et al. (2024) and Nikitin et al. (2024), both build $s$ from NLI. Nikitin et al. (2024) set $W_{mm'} = a_{mm'} + a_{m'm}$ with $a_{mm'} = \boldsymbol{\alpha}^{\top}\mathrm{NLI}(\mathbf{y}^{(m)}, \mathbf{y}^{(m')})$, where NLI returns the one-hot vector of the predicted class and $\boldsymbol{\alpha} = (1, \frac{1}{2}, 0)$; this is (C.15) with $s = \frac{1}{2}(a_{mm'} + a_{m'm})$. Lin et al. (2024) instead average the two directional scores, computed from NLI probabilities, so that their weights lie in $[0, 1]$, i.e. $\boldsymbol{W}/2$. This factor of two does not change the symmetric-normalized Laplacian below, but it does rescale the combinatorial Laplacian used by KLE; we keep the $[0, 2]$ scale of Nikitin et al., to which the kernel hyperparameters of (C.20) refer. Two Laplacians appear, and they are not interchangeable. The spectral scores use the symmetric-normalized Laplacian $\boldsymbol{L}_{\mathrm{sym}} = \boldsymbol{I} - \boldsymbol{D}^{-1/2}\boldsymbol{W}\boldsymbol{D}^{-1/2}$, whose eigenvalues $\gamma_1 \le \cdots \le \gamma_M$ lie in $[0, 2]$; a node of zero degree is given a zero row and column, so that it forms its own connected component. The kernel scores use the combinatorial Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$.

**Graph-based scores (DegMat, Eccentricity, EigenVal).** Lin et al. (2024) replace the hard partition by a weighted similarity graph and define three scores, $U_{\mathrm{Deg}}$, $U_{\mathrm{Ecc}}$ and $U_{\mathrm{EigV}}$ in their notation,

[3] Kuhn et al. (2023) use the Monte Carlo approximation $U_{\mathrm{SE}}(\boldsymbol{x}) \approx -K^{-1}\sum_k \log \hat{p}_k$, averaged over classes. The normalized form (C.13) is the one used by Farquhar et al. (2024) and in LM-Polygraph.

which we call DegMat, Eccentricity and EigenVal:

$$U_{\mathrm{DegMat}}(\boldsymbol{x}) = \frac{\mathrm{Tr}(M\boldsymbol{I} - \bar{\boldsymbol{D}})}{M^2} \;=\; 1 - \frac{1}{M^2}\sum_{m,m'} s\big(\mathbf{y}^{(m)}, \mathbf{y}^{(m')}\big), \tag{C.16}$$

$$U_{\mathrm{Eccentricity}}(\boldsymbol{x}) = \Big\| \big[\tilde{\boldsymbol{v}}_1^\top, \ldots, \tilde{\boldsymbol{v}}_M^\top\big] \Big\|_2, \qquad \tilde{\boldsymbol{v}}_m = \boldsymbol{v}_m - \frac{1}{M}\sum_{m'=1}^{M} \boldsymbol{v}_{m'}, \tag{C.17}$$

$$U_{\mathrm{EigenVal}}(\boldsymbol{x}) = \sum_{m=1}^{M} \max\big(0,\, 1 - \gamma_m\big). \tag{C.18}$$

In (C.16), $\bar{\boldsymbol{D}} = \mathrm{diag}(\boldsymbol{S}\mathbf{1})$ is the degree matrix of the similarity matrix $\boldsymbol{S} = \big[s(\mathbf{y}^{(m)}, \mathbf{y}^{(m')})\big]_{m,m'}$, which has a unit diagonal; with (C.15), $\boldsymbol{S} = \boldsymbol{I} + \frac{1}{2}\boldsymbol{W}$. $U_{\mathrm{DegMat}}$ is thus one minus the average pairwise similarity. We average over $m \neq m'$ only, which multiplies (C.16) by $M/(M-1)$: the ranking of prompts is unchanged when $M$ is the same for all of them, but not when it varies (coverage rule). $U_{\mathrm{Eccentricity}}$ measures the spread of the spectral embeddings $\boldsymbol{v}_m$ around their centroid. Here $\boldsymbol{v}_m = [u_{1,m}, \ldots, u_{d,m}]$ collects the $m$-th coordinates of the eigenvectors $\boldsymbol{u}_1, \ldots, \boldsymbol{u}_d$ of $\boldsymbol{L}_{\mathrm{sym}}$ associated with its $d$ smallest eigenvalues. Instead of fixing $d$, we keep every $\boldsymbol{u}_i$ with $\gamma_i < \bar{\gamma}$, and always at least $\boldsymbol{u}_1$. We use $\bar{\gamma} = 0.9$, the upper end of the 0.4–0.9 range explored by Lin et al. (2024). $U_{\mathrm{EigenVal}}$ is a continuous count of the connected components of the graph: the multiplicity of the eigenvalue 0 of $\boldsymbol{L}_{\mathrm{sym}}$ equals the number of connected components, so $U_{\mathrm{EigenVal}}$ approximates the number of distinct meanings and plays the role that $K$ plays in (C.14).

**Kernel Language Entropy (KLE).** Nikitin et al. (2024) generalize semantic entropy. They encode semantic relations in a positive semidefinite kernel matrix $\boldsymbol{K} \in \mathbb{R}^{M\times M}$ with *unit trace*, built from the semantic graph, and take its von Neumann entropy:

$$U_{\mathrm{KLE}}(\boldsymbol{x}) = \mathrm{VNE}(\boldsymbol{K}) = -\,\mathrm{Tr}\,\big[\boldsymbol{K}\log\boldsymbol{K}\big] = -\sum_{m=1}^{M} \eta_m \log \eta_m, \quad \eta_m \in \mathrm{Spec}(\boldsymbol{K}), \tag{C.19}$$

with the convention $0\log 0 = 0$. We use their two graph kernels on the combinatorial Laplacian $\boldsymbol{L} = \boldsymbol{D} - \boldsymbol{W}$,

$$\boldsymbol{K}_\beta = \exp(-\beta\boldsymbol{L}) \quad (\beta = 0.1), \qquad \boldsymbol{K}_{\nu,\kappa} = \Big(\tfrac{2\nu}{\kappa^2}\boldsymbol{I} + \boldsymbol{L}\Big)^{-\nu} \quad (\nu = \kappa = 1), \tag{C.20}$$

which we report as KLE-Heat and KLE-Matérn. The authors' default for the heat kernel is $\beta = 0.3$; the Matérn kernel is at their default. Each kernel is normalized as $\boldsymbol{K} \leftarrow \tilde{\boldsymbol{K}}$, with $\tilde{K}_{mm'} = K_{mm'}\big(K_{mm}K_{m'm'}\big)^{-1/2}/M$: a rescaling to unit diagonal followed by division by $M$, so that $\mathrm{Tr}(\tilde{\boldsymbol{K}}) = 1$. KLE recovers Semantic Entropy for any clustering (Nikitin et al., 2024, Thm. 3.5). In particular, it recovers the discrete form (C.14) with the block kernel $\boldsymbol{K} = M^{-1}\sum_k \mathbf{1}_{C_k}\mathbf{1}_{C_k}^\top$, whose non-zero eigenvalues are the class frequencies $|C_k|/M$, and it extends it by allowing graded similarity between classes. We use the per-answer kernel (KLE) rather than the cluster-level variant (KLE$_c$), so no clustering is involved.

**Average Dissimilarity (AD) and CoCoA.** The scores above describe the whole set and depend on $\boldsymbol{x}$ only. Vashurin et al. (2025b) instead score a *specific* generation $\mathbf{y}$ by its average dissimilarity to the set, their consistency term (there $\widehat{U}_{\mathrm{cons}}$), which we call AD as in Vazhentsev et al. (2026). AD thus has two inputs with distinct roles: an *anchor* $\mathbf{y}$, the generation whose reliability is scored, and the set $\mathcal{Y} = \{\mathbf{y}^{(1)}, \ldots, \mathbf{y}^{(M)}\}$ it is compared with. The anchor need not belong to $\mathcal{Y}$:

$$U_{\mathrm{AD}}(\mathbf{y} \mid \boldsymbol{x}) \;=\; \frac{1}{M}\sum_{m=1}^{M}\Big(1 - s\big(\mathbf{y}, \mathbf{y}^{(m)}\big)\Big). \tag{C.21}$$

Their final score, CoCoA, combines the consistency term *multiplicatively* with an information-based score of the same generation:

$$U_{\mathrm{CoCoA}}(\mathbf{y} \mid \boldsymbol{x}) \;=\; U_{\mathrm{inf}}(\mathbf{y} \mid \boldsymbol{x}) \cdot U_{\mathrm{AD}}(\mathbf{y} \mid \boldsymbol{x}), \tag{C.22}$$

where $U_{\text{inf}}$ is an information-based score such as sequence probability, PPL or MTE.

Vazhentsev et al. (2026) adapt both factors of (C.22) to LLDMs. Their D-CoCoA replaces the information-based factor by a diffusion-aware score, and the i.i.d. set by the trajectory itself. The anchor is then the *final answer of that same trajectory*, and $U_{\text{AD}}$ is (C.21) with this anchor and $\mathcal{Y} = \{\tilde{\mathbf{y}}_t\}_{t=1}^{T}$, the trajectory's intermediate estimates. The local variant is $U_{\text{D-CoCoA-L}} = U_{\text{CommitNLL}} \cdot U_{\text{AD}}$, where the average is taken within each decoding block. The global variant, D-CoCoA-G, takes the same product with the average over the whole sequence instead. With the single-block decoding we use (App. D.1), the two variants coincide.

**Our instantiation of AD and CoCoA.** In every condition, the anchor is the greedy answer $\hat{\mathbf{y}}$, the answer whose reliability is being assessed, and it is produced separately from the set it is compared with. $\hat{\mathbf{y}}$ comes from the greedy pass ($\tau = 0$), whereas $\mathcal{Y}$ comes from the $\tau = 1$ generations (App. D.1). In the reference condition, $\mathcal{Y}$ holds the final answers of the $M$ i.i.d. samples. Under a single-trajectory strategy, it holds intermediate answers of one sampled trajectory. In neither case is $\hat{\mathbf{y}}$ an element of $\mathcal{Y}$, and, unlike in D-CoCoA, the anchor is not the final answer of the trajectory that supplies the set. Each pair (anchor, element of $\mathcal{Y}$) is scored with the symmetrized soft NLI score:

$$U_{\text{AD}}(\hat{\mathbf{y}} \mid \boldsymbol{x}) = \frac{1}{M}\sum_{m=1}^{M}\Big(1 - s\big(\hat{\mathbf{y}}, \mathbf{y}^{(m)}\big)\Big), \qquad s(\hat{\mathbf{y}}, \mathbf{y}) = \tfrac{1}{2}\boldsymbol{\alpha}^{\top}\big(\text{NLI}(\mathbf{y}, \hat{\mathbf{y}}) + \text{NLI}(\hat{\mathbf{y}}, \mathbf{y})\big), \quad \text{(C.23)}$$

with $\boldsymbol{\alpha} = (1, \frac{1}{2}, 0)$, and $s(\hat{\mathbf{y}}, \mathbf{y}) = 1$ when $\mathbf{y}$ is empty or identical to $\hat{\mathbf{y}}$. Under a single-trajectory strategy, the elements of $\mathcal{Y}$ are the intermediate answers $\tilde{\mathbf{y}}_t$ at the selected steps $t \in \mathcal{S}$, so $U_{\text{AD}}$ is the average of $1 - s(\hat{\mathbf{y}}, \tilde{\mathbf{y}}_t)$ over $\mathcal{S}$: the anchor stays the same and only the set changes from one strategy to another. For DiMS, $\mathcal{S} = \mathcal{S}^{\star}$ is obtained from $\mathbf{w}^{\star}$ (B.28). The CoCoA score we report, next to $U_{\text{AD}}$, is $U_{\text{CoCoA}}(\hat{\mathbf{y}} \mid \boldsymbol{x}) = U_{\text{CommitNLL}}(\hat{\mathbf{y}} \mid \boldsymbol{x}) \cdot U_{\text{AD}}(\hat{\mathbf{y}} \mid \boldsymbol{x})$, where the CommitNLL factor is read off the greedy trajectory in every condition.

### C.5 Summary

Table C.1 lists the estimators of this appendix and their inputs. The information-based and diffusion-specific scores form the single-pass block of Table E.1 (its "Info-based" block). They are read off the greedy trajectory at negligible cost, except MCNLL. Every semantic-consistency estimator needs a set of $M$ answers and a pairwise scorer. This is the cost that DiMS reduces, by reading the set off a single trajectory.

Table C.1: **Inputs of each estimator.** "Set" means that the estimator needs $M$ answers, drawn i.i.d. in the reference condition and read off a single trajectory otherwise. "Pairwise model" gives the auxiliary scorer that dominates the marginal cost (App. D.4). Our MSP feature is the per-position variant (C.7) of (C.3) (App. C.2); the graph and kernel estimators use the STS cross-encoder, Semantic Entropy, AD and CoCoA the NLI model.

| Estimator | Needs | Pairwise model | Eq. |
|---|---|---|---|
| Perplexity | last-step logits | – | (C.5) |
| MTE | last-step logits | – | (C.6) |
| MSP | last-step logits | – | (C.7) |
| TrajEnt | trajectory logits | – | (C.10) |
| CommitNLL | commitment log-probs | – | (C.11) |
| MCNLL$^{\text{norm}}$ | $N_{\text{mc}}$ extra passes | – | (C.8) |
| Semantic Entropy | set + $\rho_t$ (C.12) | NLI | (C.13) |
| DegMat | set | STS | (C.16) |
| Eccentricity | set | STS | (C.17) |
| EigenVal | set | STS | (C.18) |
| KLE-Heat, KLE-Matérn | set | STS | (C.19) |
| AD | set + greedy answer | NLI | (C.23) |
| CoCoA | AD inputs + CommitNLL | NLI | (C.22) |

# D Additional Details of Experimental Setup

This appendix details the experimental setup summarized in Section 5.1, following its structure: the models, datasets, prompts and generation protocol (App. D.1), the strategies and estimators compared, with the calibration of DiMS (App. D.2), the evaluation (App. D.3) and the measurement of computational overhead (App. D.4). The estimators themselves are defined in App. C. Code will be released upon publication.

## D.1 Models, datasets and protocol

**Models.** We use three instruction-tuned LLDMs, all loaded from their public Hugging Face checkpoints.

- `LLaDA-1.5` (`GSAI-ML/LLaDA-1.5`, 8B parameters; Zhu et al., 2026): the `LLaDA-8B` masked diffusion model (Nie et al., 2025), further aligned with variance-reduced preference optimization.
- `Dream` (`Dream-org/Dream-v0-Instruct-7B`; Ye et al., 2025): a 7B masked diffusion model initialized from an autoregressive model. It keeps the next-token output convention, i.e. the logits returned at position $\ell - 1$ predict the token at position $\ell$. We shift them by one position so that $f^{\ell}_{\boldsymbol{\theta}}$ always refers to position $\ell$, as for `LLaDA-1.5`.
- `Nemotron` (`nvidia/Nemotron-Labs-Diffusion-14B`; Fu et al., 2026): a 14B model trained jointly with an autoregressive and a block-wise diffusion objective, which can decode autoregressively, by diffusion, or by self-speculation. We use its diffusion mode, which only supports low-confidence remasking; `Nemotron` is therefore evaluated under `lc` only.

All three models decode the whole answer as a single block of $L$ tokens, with no semi-autoregressive or block-wise schedule. Every intermediate answer $\tilde{\mathbf{y}}_t$ is therefore an estimate of the full answer, as assumed in Section 3.1. `Nemotron` has one peculiarity. As in its reference decoder, its first answer token is produced by the autoregressive head, from a causal pass over the prompt, and only the remaining $L - 1$ positions are denoised bidirectionally: a fully masked block is out of distribution for this hybrid model. That first position is therefore already committed at the first denoising step.

**Datasets.** We use six benchmarks covering five task types. Table D.1 gives the subsets, splits and sizes.

- **TriviaQA** (`mandarjoshi/trivia_qa`; Joshi et al., 2017): open-domain factoid question answering, asked without the evidence documents.
- **GSM8K** (`openai/gsm8k`; Cobbe et al., 2021): grade-school math word problems. The model is asked to reason step by step and to end with "`#### [answer]`".
- **MATH** (`qwedsacf/competition_math`; Hendrycks et al., 2021): competition mathematics problems across seven subjects (algebra, counting and probability, geometry, intermediate algebra, number theory, prealgebra, precalculus), with the final answer requested in `\boxed{}`.
- **WMT14 Fr-En** (`wmt/wmt14`; Bojar et al., 2014): machine translation from French into English.
- **SAMSum** (`Samsung/samsum`; Gliwa et al., 2019): abstractive summarization of messenger-style dialogues.
- **HotpotQA** (`hotpotqa/hotpot_qa`; Yang et al., 2018), distractor setting: multi-hop question answering, where each question comes with ten paragraphs, two of them relevant. Including these paragraphs makes its prompts by far the longest.

When a split contains more than 1,000 examples, we keep its first 1,000 examples, in the order of the Hugging Face release.

Table D.1: **Dataset statistics.** *Subset / split*: Hugging Face configuration and split. For MATH, the train split is the only one available on the Hugging Face release we use. *Cal. / Val. / Test*: number of prompts in each part of the split (App. D.2). $T = L$: number of denoising steps and of generated tokens. *Prompt length*: median and maximum number of whitespace-separated words in the filled template of Table D.2, before the chat template is applied.

| Dataset | Subset / split | #Prompts | Cal. / Val. / Test | $T = L$ | Prompt length |
|---|---|---|---|---|---|
| TriviaQA | `rc` / validation | 1000 | 100 / 200 / 700 | 32 | 11 / 36 |
| GSM8K | `main` / test | 1000 | 100 / 200 / 700 | 128 | 73 / 140 |
| MATH | – / train | 1000 | 100 / 200 / 700 | 128 | 35 / 229 |
| WMT14 | `fr-en` / test | 1000 | 100 / 200 / 700 | 128 | 34 / 88 |
| SAMSum | – / test | 819 | 100 / 200 / 519 | 128 | 114 / 556 |
| HotpotQA | `distractor` / validation | 1000 | 100 / 200 / 700 | 128 | 944 / 2274 |

**Prompts.** Table D.2 gives the instruction used for each dataset. The filled instruction is sent as the user turn of the model's chat template, with the system message that template defaults to: "*You are a helpful assistant.*" for `Dream`, none for `LLaDA-1.5`, and an empty one for `Nemotron`, whose assistant turn is opened with an empty `<think></think>` block. For HotpotQA, {documents} is the ten paragraphs, each written as "title: text" and separated by a blank line.

Table D.2: **Instruction templates used for generation.** Fields in braces are filled from the dataset, and `\n` denotes a line break.

| Dataset | Template |
|---|---|
| TriviaQA | `{question}` |
| GSM8K | `Given the following question, reason and give a final answer to the question. Your response should end with "#### [answer]" where [answer] is the response to the question.\n\nQuestion: {question}\nAnswer:` |
| MATH | `Solve the following math problem step by step. Put your final answer in \boxed{}.\n\nProblem: {problem}\nSolution:` |
| WMT14 | `Translate the following French text into English, give only the translation.\n\nOriginal: {source}\nTranslation:` |
| SAMSum | `Provide a summary for the following dialogue. The summary should (1) be rather short, (2) extract all important pieces of information, (3) include names of interlocutors, (4) be written in the third person.\n\nHere's the dialogue:\n{dialogue}\n(End of dialogue).\n\nSummary:` |
| HotpotQA | `The following are the given documents:\n\n{documents}\n\nAnswer the question based strictly on the provided documents.\n\nQuestion: {question}\nAnswer:` |

**Decoding.** All models are prompted zero-shot (no in-context examples; templates in Table D.2) and decoded without classifier-free guidance. The answer length is fixed to $L = 32$ tokens on TriviaQA and $L = 128$ tokens on the five other datasets, and the number of denoising steps equals the answer length, $T = L$, so exactly one position is committed per step (`Nemotron`: after its first token). Answers shorter than $L$ are completed with end-of-sequence tokens. At step $t$, the intermediate answer $\tilde{\mathbf{y}}_t$ is formed as in (1), with the masked positions sampled from $f^{\ell}_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x}, \boldsymbol{z}_t)$ at temperature $\tau$ ($\tau = 0$ meaning arg-max). The decoding configuration $\mathcal{G}$ in (2) then keeps the committed positions of $\tilde{\mathbf{y}}_t$ and commits one new masked position, chosen by one of two remasking schedules (Nie et al., 2025):

- **Low-confidence (`lc`):** the masked position whose proposed token $\tilde{\mathrm{y}}^{\ell}_t$ has the highest probability under $f^{\ell}_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x}, \boldsymbol{z}_t)$;
- **Random (`rd`):** a masked position drawn uniformly at random.

All other positions are masked again. Each (model, remasking schedule) pair defines a *model variant*. With two variants for `LLaDA-1.5` and `Dream` and one for `Nemotron`, the five variants and six datasets give 30 *runs*. When an intermediate answer is compared with other answers, it is detokenized and truncated at its first end-of-sequence token.

**Generations per prompt.** Every prompt is decoded in two generation settings, both under the remasking schedule of the run.

- **One greedy pass** ($\tau = 0$). Its final state is the answer $\hat{\mathbf{y}}$ that is judged for correctness (App. D.3) and whose reliability every estimator scores. The information-based and diffusion-specific scores (Perplexity, MSP, MTE, TrajEnt, CommitNLL, $\mathrm{MCNLL}^{\mathrm{norm}}$) are computed on this trajectory, and AD and CoCoA compare the other answers with $\hat{\mathbf{y}}$ (App. C.4).
- $M = 20$ **independent generations** ($\tau = 1$), i.e. i.i.d. samples from $p_{\boldsymbol{\theta}}^{\mathrm{Diff}}(\cdot \mid \boldsymbol{x})$. We keep the *entire* trajectory of each of them, not only its final answer. These generations supply all three objects the experiments need: their $M$ final answers form the set of the `Reference` in (3), the multi-sample reference of Section 5.1 (*Multi-sample* / 20 *i.i.d.* in the tables); the first of the $M$ trajectories, $\tilde{\mathbf{Y}} \sim \bar{\pi}_{\boldsymbol{\theta}}(\cdot \mid \boldsymbol{x})$, is the sampled trajectory that every single-trajectory strategy reads, so those strategies differ only in the set of steps they keep (see below); and on calibration prompts, the $M = 20$ generations serve as the $R = 20$ trajectories from which the estimators (19) that fit $\mathbf{w}^\star$ are computed (see the DiMS calibration paragraph).

The target distribution $p_{\boldsymbol{\theta}}^{\mathrm{Diff}}$ that DiMS matches is therefore the $\tau = 1$ answer distribution, the same as that of the `Reference`. Reusing the same generations is also what equation ( 19) assumes: on a calibration prompt the reference answers *are* the endpoints $\tilde{\mathbf{y}}_1^{(r)}$ of the $R$ trajectories, so the pair (trajectory $r$, its own endpoint) is coupled rather than a draw from two independent samples. This is why $\widehat{\boldsymbol{b}}_{\boldsymbol{x}}$ and $\widehat{c}_{\boldsymbol{x}}$ sum over $r \neq r'$, and the implementation follows that convention: the $R$ coupled pairs are dropped and the remaining $R(R-1)$ are averaged. Keeping them would bias $\widehat{\boldsymbol{b}}_{\boldsymbol{x}}$ upwards at the late steps, where a trajectory is closest to its own endpoint.

**Hardware.** All generation, calibration and evaluation runs were executed on GPU nodes with NVIDIA GB200 accelerators (`arm64` hosts), inside an Apptainer container.

### D.2 Strategies and estimators

**Single-trajectory strategies.** Given the sampled trajectory $\tilde{\mathbf{Y}} = [\tilde{\mathbf{y}}_T, \ldots, \tilde{\mathbf{y}}_1]$, each strategy keeps a set of steps $\mathcal{S} \subseteq \mathcal{T} = \{T, \ldots, 1\}$ and passes the set $\{\tilde{\mathbf{y}}_t\}_{t \in \mathcal{S}}$ to the estimator, as in (21):

- `Full`: $\mathcal{S} = \mathcal{T}$, i.e. all $T$ intermediate answers;
- `Last`: $\mathcal{S} = \{M, \ldots, 1\}$, the final $M$ steps;
- `Random`: $M$ distinct steps drawn uniformly from $\mathcal{T}$, drawn anew for every prompt from a generator with a fixed seed;
- `Centroid`: an $M$-means clustering of the $T$ intermediate answers in the `E5-large-v2` embedding space (see *Scorers* below); $\mathcal{S}$ keeps, for each cluster, the step nearest its centroid (in case of ties, the one closest to the end of denoising, i.e. the smallest $t$). It is training-free, computed with the $k$-means++ algorithm (Arthur & Vassilvitskii, 2007) with 10 restarts and a fixed seed per prompt. A text repeated over several steps is weighted by its multiplicity, and when there are fewer than $M$ distinct texts, $\mathcal{S}$ keeps one step per distinct text;
- DiMS (top-$M$) and DiMS (cov. 75%): the sets $\mathcal{S}^\star_{\mathrm{top}}$ and $\mathcal{S}^\star_{\mathrm{cov}}$ of (B.28), computed from the learned weights $\mathbf{w}^\star$. Since $w_t^\star \leq 1/M$, reaching 75% of the mass takes at least $\lceil 0.75\, M \rceil = 15$ steps, so the coverage rule keeps at least 15 steps and can keep fewer or more than $M$.

The four other selection rules of App. B.5.2 are used only in App. E.4.1.

**Scorers.** Three pre-trained models provide the similarities used by DiMS and by the estimators.

- **Sentence embedding (DiMS kernel).** The embedding $e$ of (6) is the Sentence-Transformers (Reimers & Gurevych, 2019) model `E5-large-v2` (`intfloat/e5-large-v2`, $d = 1024$; Wang et al., 2022), queried with the instruction prefix `"query: "` it is trained with, and $\ell_2$-normalized. The default $\psi$ is an RBF kernel, $\psi(\boldsymbol{u}, \boldsymbol{v}) = \exp\big(-\|\boldsymbol{u} - \boldsymbol{v}\|_2^2/(2\sigma_{\boldsymbol{x}}^2)\big)$, whose bandwidth is set by the median heuristic (Gretton et al., 2012) *per prompt*: $\sigma_{\boldsymbol{x}}^2$ is the median of the squared pairwise Euclidean distances between all the embeddings that prompt contributes to the quadratic program, i.e. the $RT$ intermediate answers of its $R$ trajectories (zero distances, which duplicate texts produce, are discarded before the median is taken). The same scalar is reused for every entry of $\widehat{\boldsymbol{A}}_{\boldsymbol{x}}$, $\widehat{\boldsymbol{B}}_{\boldsymbol{x}}$, $\widehat{\boldsymbol{b}}_{\boldsymbol{x}}$ and $\widehat{c}_{\boldsymbol{x}}$, so all (intermediate) answers from a prompt are compared under one kernel scale. $\psi$ satisfies $\psi(\boldsymbol{u}, \boldsymbol{u}) = 1$, as assumed in Section 4.2. App. E.4.3 replaces $\psi$ by a Laplacian and a Matérn ($\nu = 3/2$) kernel with the same median rule applied in the metric each kernel uses, and App. E.4.4 replaces $e$ by `all-MiniLM-L6-v2` (`sentence-transformers/all-MiniLM-L6-v2`, $d = 384$), a fine-tuned version of a 6-layer variant of `MiniLM-L12-H384` (`microsoft/MiniLM-L12-H384-uncased`; Wang et al., 2020), and `BGE-M3` (`BAAI/bge-m3`, $d = 1024$; Chen et al., 2024).
- **NLI model.** Semantic Entropy, AD and CoCoA use the NLI classifier `microsoft/deberta-v2-xlarge-mnli` (He et al., 2021), fine-tuned on the Multi-Genre Natural Language Inference corpus (MultiNLI; Williams et al., 2018), for the similarity $s$ of Section 3.2. Following App. C.4, Semantic Entropy clusters answers with the hard (bidirectional entailment) score, while AD and CoCoA use the soft score $P(\text{ent}) + \frac{1}{2}P(\text{neu})$, i.e. $\boldsymbol{\alpha} = (1, \frac{1}{2}, 0)$, averaged over the two directions (C.23).
- **STS model.** DegMat, Eccentricity, EigenVal and the two KLE variants build their graph (C.15) from the cross-encoder `cross-encoder/stsb-roberta-large`, a RoBERTa-large model (Liu et al., 2019) fine-tuned on the Semantic Textual Similarity Benchmark (STS-B; Cer et al., 2017): the scores of the two orderings of a pair are averaged and clipped to $[0, 1]$.

The soft NLI scoring is behind every AD and CoCoA number in the paper, and the STS graph behind every DegMat, Eccentricity, EigenVal and KLE number—Table 1, Figure 1 and every table of App. E, the ablations and the transfer study included—so the captions there do not restate it.

**DiMS calibration.** For each run, the prompts are split into $N_{\text{cal}} = 100$ calibration prompts, $N_{\text{val}} = 200$ validation prompts and a test set with the remaining prompts (Table D.1). The split is drawn once, with a fixed seed, and is stratified by the correctness of the greedy answer, taken as the majority verdict of the four evaluators of App. D.3 (a 2–2 tie counts as incorrect), so that a fold of 100 prompts cannot lose the minority class. Correctness differs from model to model, so the split is drawn *per run* rather than once per dataset: two model variants of the same dataset therefore have different—though overlapping—test sets (pairwise test-set Jaccard index 0.44–0.56). Comparisons *across* model variants are therefore made on different prompt sets, while every comparison *within* a run—which is what all our claims are about—is on identical prompts.

For each calibration prompt, we use the $R = 20$ trajectories at $\tau = 1$, embed all their intermediate answers and compute the estimators in (19). The estimators are averaged over the calibration prompts, and $\mathbf{w}^\star$ solves the quadratic program (20) over the capped simplex $\mathcal{W}_M$ with $M = 20$. We solve it with `CVXPY` (Diamond & Boyd, 2016), using the ECOS solver (Domahidi et al., 2013) with SCS (O'Donoghue et al., 2013) as a fallback. The quadratic program uses no correctness labels.

**Selection of $\lambda$.** We solve (20) for each value of the redundancy penalty in the grid

$$\lambda \in \{0,\ 0.05,\ 0.1,\ 0.2,\ 0.5,\ 0.7,\ 1,\ 2,\ 5,\ 10,\ 20,\ 50,\ 100\}.$$

By Proposition 2, $\lambda = 0$ still corresponds to an effective penalty $\lambda_{\text{eff}} = 1/R = 0.05$, which is why the grid needs no smaller value; and the upper end stops at 100 because the $\lambda \to \infty$ limit is the minimum-variance selection $\arg\min_{\mathbf{w} \in \mathcal{W}_M} \mathbf{w}^\top \overline{\boldsymbol{C}} \mathbf{w}$, a regime the grid already reaches. The weights do not depend on the estimator; only the selected $\lambda$ does. For each run, estimator and

selection rule, $\lambda^\star$ is the grid value that maximizes, on the validation prompts, the Spearman rank correlation between the estimator's DiMS score (one $\tau = 1$ trajectory) and the score of the same estimator computed on the $M = 20$ i.i.d. generations. It reads no correctness label, and one $\lambda^\star$ serves the AUROC and the PRR tables alike. The correlations are rounded to nine decimals, so that saturated scores differing only by floating-point noise tie, and ties are broken towards the smaller $\lambda$; they are point estimates (no bootstrap), the drawn rules average them over their draws, and the prompts whose greedy answer is degenerate are excluded as in the metrics.

**Size of the calibration set.** The quadratic program reads no label, so its input is only unlabeled generations, and Section 5.2 reports what happens when there are fewer of them: $N_{\text{cal}} \in \{100, 50, 10\}$. The two smaller folds are *nested subsets* of the 100, stratified on the same labels, and the validation and test prompts are copied verbatim from the default split. Simply re-running the splitter with a smaller calibration size would not achieve this. Its stratified draw takes the calibration fold first, so shrinking that fold would also re-cut the validation set, which selects $\lambda^\star$, and the test set, on which results are reported. The comparison would then confound three changes at once. With the folds nested, the number of prompts the quadratic program sees is the single varying quantity. Everything downstream is re-run unchanged at each size: the weights are retrained, and $\lambda^\star$ is re-tuned on the same validation prompts.

### D.3 Evaluation

**Correctness labels.** The label $c(\boldsymbol{x}, \mathbf{y}^{\text{ref}}, \hat{\mathbf{y}}) \in \{0, 1\}$ of (C.1) is assigned to the greedy answer $\hat{\mathbf{y}}$ of every prompt by LLM judges. Exact match is not suited to the free-form outputs of most of our tasks, such as translations, summaries or multi-hop answers.

- **Judges.** We use two large open-weight models, `Llama` (`meta-llama/Llama-3.1-405B-Instruct-FP8`; Grattafiori et al., 2024) and `Qwen` (`Qwen/Qwen3.5-397B-A17B-FP8`, a mixture-of-experts model with about 17B active parameters; Qwen Team, 2026), each queried at temperatures 1.0 and 0.5. This gives four evaluators, and hence four label sets per run. An output that cannot be parsed as `yes` or `no` is re-queried with a new seed, for up to three attempts in total per judge; a prompt still unparsed after three attempts carries no label for that evaluator and is dropped from that evaluator's metric alone. This happened twice in total, on two `Nemotron` runs, out of $\sim 8 \times 10^4$ evaluator-prompt pairs on the test splits.
- **Aggregation.** We do not merge the four label sets by majority vote. Every metric is computed separately against each label set, and the four values are averaged (paragraph *Metrics* below). Majority voting is used in one place only, and not for a metric: to give each prompt the single class label that the split is stratified on. Exact-match labels, which our pipeline also computes, are not used in any reported result.
- **Judge prompt.** Each template gives the question, the gold reference(s) and the answer to judge, and asks for a `yes`/`no` verdict (`yes` meaning correct). For TriviaQA and HotpotQA, the template is:

```
We are assessing the quality of answers to the following
question: {question}
The following are correct answers: {reference_answers}
The proposed answer is: {prediction}
Within the context of the question and correct answer, is
the proposed answer correct?
Respond only with yes or no.
Response:
```

  For GSM8K and MATH, the template also gives the full reference solution and the gold final answer extracted from it. For WMT14 and SAMSum, it gives a reference translation or summary and asks whether the proposed translation or summary is correct. The wording and the `yes`/`no` format are otherwise the same for all datasets.

**Accuracy and judge agreement.** Table D.3 gives the accuracy of every model variant under the four evaluators, together with the share of prompts on which all four agree. They agree on 85.6% of prompts overall. Most of the disagreement is between the two judges rather than between the

two temperatures of one judge, and it is concentrated on translation and summarization: SAMSum (63.3% unanimous), where `Qwen` accepts 25 points more summaries than `Llama`, and WMT14 (79.5%), against 91–97% on the four other datasets.

Table D.3: **Accuracy (%) of the greedy answer**, over all prompts of a run. Each cell is the mean, and in brackets the minimum and maximum, over the four evaluators (two judges × two temperatures). *Average*: each evaluator's accuracy is averaged over the runs of the row or column before the mean, minimum and maximum are taken. *Unanimous*: percentage of prompts on which all four evaluators give the same label, averaged over the runs of the row or column; the corner cell covers all 30 runs.

| Model variant | TriviaQA | GSM8K | MATH | WMT14 | SAMSum | HotpotQA | Average | Unanimous |
|---|---|---|---|---|---|---|---|---|
| LLaDA-1.5 (`lc`) | 36.8 (35.9; 38.6) | 57.6 (57.0; 58.3) | 36.5 (36.1; 37.1) | 74.0 (62.7; 79.5) | 64.1 (50.2; 81.6) | 83.5 (79.8; 87.8) | 58.8 (56.4; 63.0) | 87.5 |
| LLaDA-1.5 (`rd`) | 36.4 (35.1; 38.4) | 51.8 (50.6; 53.2) | 29.3 (29.0; 30.1) | 67.2 (55.6; 72.6) | 66.9 (62.1; 74.7) | 82.8 (79.6; 86.7) | 55.7 (54.4; 58.6) | 89.0 |
| Dream (`lc`) | 48.9 (48.3; 49.8) | 43.1 (42.4; 43.9) | 31.8 (31.6; 32.0) | 41.7 (35.5; 44.1) | 51.4 (29.3; 78.9) | 84.9 (83.2; 87.4) | 50.3 (46.5; 55.8) | 87.3 |
| Dream (`rd`) | 47.5 (46.8; 48.9) | 33.9 (32.0; 35.9) | 18.4 (17.9; 19.2) | 53.1 (40.4; 58.6) | 49.0 (28.0; 75.8) | 82.5 (80.9; 85.2) | 47.4 (43.8; 53.5) | 84.4 |
| Nemotron (`lc`) | 51.7 (51.1; 53.0) | 51.5 (37.4; 65.7) | 19.6 (6.1; 33.8) | 67.4 (55.1; 73.9) | 67.7 (64.3; 74.0) | 75.3 (72.9; 76.6) | 55.6 (48.4; 61.1) | 79.9 |
| **Average** | 44.3 (43.5; 45.7) | 47.6 (45.5; 49.6) | 27.1 (24.3; 29.9) | 60.7 (49.9; 65.7) | 59.8 (46.8; 77.0) | 81.8 (79.9; 84.5) | 53.6 (52.3; 56.7) | – |
| **Unanimous** | 96.6 | 91.7 | 91.5 | 79.5 | 63.3 | 91.1 | – | 85.6 |

**Metrics.** Every metric is computed on the test prompts of a run, from the scores $U(\hat{\boldsymbol{y}} \mid \boldsymbol{x})$ and the error indicators $1 - c$ of (C.1).

- **AUROC** is the area under the ROC curve for detecting incorrect answers, with $U$ as the score and incorrect answers as the positive class. It equals the probability that a random incorrect answer receives a higher uncertainty than a random correct one; 0.5 is chance level.
- **PRR** (Malinin & Gales, 2021) compares rejection curves. Answers are rejected in decreasing order of $U$, and the error rate of the retained answers is plotted against the rejected fraction. With $\mathrm{AUC}_U$, $\mathrm{AUC}_{\mathrm{rnd}}$ and $\mathrm{AUC}_{\mathrm{oracle}}$ the areas under this curve for $U$, for a random order and for an order that rejects incorrect answers first, $\mathrm{PRR} = (\mathrm{AUC}_{\mathrm{rnd}} - \mathrm{AUC}_U)/(\mathrm{AUC}_{\mathrm{rnd}} - \mathrm{AUC}_{\mathrm{oracle}})$. As in Vashurin et al. (2025b) and Vazhentsev et al. (2026), the areas are computed only up to a rejection rate of 50% (PRR@50%), so PRR focuses on the most uncertain half of the answers; the curves are evaluated at 25 evenly spaced coverage levels between 100% and 50% and integrated with the trapezoidal rule. PRR is 0 for a random score and 1 for an oracle.

Both metrics depend on $U$ only through the ranking it induces (Remark 1).

**Common support.** A few prompts receive a greedy answer that is empty or unusable. Every score anchored on that answer is then undefined—the information-based scores (Perplexity, MSP, MTE) have no answer span, and every AD and CoCoA variant, at every step scope, measures deviation *from* $\hat{\mathbf{y}}$ and has no anchor—whereas Semantic Entropy and the graph estimators still return a value. Letting each estimator drop its own missing rows would not be neutral: an empty answer is judged incorrect by every judge, so the dropped prompts are almost exclusively the model's failures, and AD/CoCoA would be scored on an easier, higher-accuracy subset than the estimators printed beside them. We therefore exclude those prompts from *every* estimator, on validation and test alike, so that all the numbers in a table are computed on the same prompts (up to the two prompts left unlabeled by a single evaluator, see *Judges* above). The alternative of imputing them at maximal uncertainty is the same asymmetry mirrored—it hands a free true positive only to the estimators that were missing the row—and was measured to reverse the ordering of AD and DegMat on the `Dream` (`lc`) WMT14 run, so it is not used.

The exclusion is a scoring mask, not a re-split: excluded prompts that fall in the calibration fold stay in it, and the quadratic program can still use them, since it matches trajectories against the $\tau = 1$ reference and never reads the greedy answer. In total 494 of the 20,095 test prompts are excluded (2.5%), and 21 of the 30 runs have none. They are concentrated on `Dream` (`lc`), whose WMT14 and MATH runs lose 31% and 17% of their test prompts; no other run loses more than 10%.

Reported numbers on the affected runs are thus conditional on the greedy decode having produced a well-formed answer, and their base error rate differs from the stratified one.

**Intervals.** The uncertainty of a run-level value is estimated with $500$ bootstrap resamples of the test prompts. The four label sets are resampled *jointly*: each iteration draws one resample of prompts, all four evaluators score that same resample, and the iteration's value is their average. The reported standard error is the standard deviation of the bootstrap distribution of that averaged statistic—the quantity the tables actually print—rather than a combination of four separately bootstrapped SEs, which would ignore the correlation between judges scoring the same prompts. Tables report $\pm 1.96$ SE. Averages over $n$ runs (e.g. over the six datasets of a model variant) are unweighted, and their SE is combined in quadrature, $\mathrm{SE} = \left(\sum_{i=1}^{n} \mathrm{SE}_i^2\right)^{1/2}/n$, which treats the runs as independent.

### D.4 Overhead measurement

The *Overhead* values in Tables 1 and E.1 and the horizontal axis of Figure 1 are the mean *marginal* wall-clock time per prompt. *Marginal* means that the greedy decode, which every method shares because it produces the answer being scored, is not counted. The overhead has three components—the generations the strategy needs, the calls to the similarity model (NLI classifier or STS cross-encoder) on the pairs of answers it compares, and the numeric estimator itself—and all three are *measured*, on $100$ test prompts per run (after three warm-up prompts) on the hardware of App. D.1, with a pinned batch shape and thread count. The similarity model is timed with a fixed batch size, without batching across prompts, and with a cache reset for each prompt, so that a pair of identical texts within a prompt is scored once.

`Centroid` adds one term the other strategies do not have, its selection step (embedding and $k$-means++, $85$ ms per prompt), which is measured directly and added to its overhead.

**Single-pass scores.** Perplexity, MSP, MTE, TrajEnt and CommitNLL are read off the greedy trajectory. Their marginal cost is a numeric reduction, negligible next to a forward pass, and is reported as "–". $\mathrm{MCNLL}^{\mathrm{norm}}$ is the exception: it needs $N_{\mathrm{mc}} = 100$ additional forward passes (C.8), which are timed. Their cost grows with the prompt length and is largest on HotpotQA (Table D.1).

**Generation.** The single $\tau = 1$ trajectory read by the single-trajectory strategies is timed at batch size 1. The $M = 20$ independent generations of the `Reference` are charged $M$ times this single-trajectory cost, i.e. as $M$ sequential decodes. Decoding them instead as one batch of $M$ was also measured: it still costs more than $6$ times a single trajectory, so batching narrows the overhead gap but does not close it.

## E Additional Results

This appendix reports the full results behind Section 5.2. Unless stated otherwise, every table uses the default setup of the main experiments: uncertainty signals are read off one $\tau$=1 sampled trajectory; DiMS is calibrated on $N_{\mathrm{cal}}$=100 unlabeled prompts with $R$=20 trajectories each, `E5-large-v2` embeddings and an RBF kernel; and $\lambda^\star$ is selected on the validation split, separately for each (estimator, selection rule), as the grid value that maximizes the Spearman correlation with the estimator's own $M$=20 i.i.d. reference. Since this selection uses no correctness labels, the same $\lambda^\star$ serves AUROC and PRR. Correctness labels are averaged over the four evaluators (App. D.3). Table E.1 gives the full AUROC results per estimator and model variant. Sections E.1–E.3 report the same runs under PRR, per dataset and per task; Section E.4 varies one component of DiMS's calibration at a time, and Section E.5 deploys weights calibrated on another dataset or model variant.

### E.1 Prediction-Rejection Ratio

Table E.2 reports PRR for the configurations of Table E.1, with the same $\lambda^\star$. A DiMS rule is the best or tied-best single-trajectory strategy in 33 of the $40$ (estimator, model) blocks, against $34$ under AUROC: coverage $75\%$ in $18$ blocks and top-$M$ in $16$, with $1$ tie between them (AUROC: $19$ and

18). Among the baselines, `Centroid` is best or tied best in 4 blocks (AUROC: 2), `Full` in 2 (4), `Last` in 2 (2) and `Random` in 1 (0).

### E.2 Per-model, per-dataset breakdown

Tables E.3–E.7 give the per-dataset values behind Tables E.1 and E.2, one table per model variant, with AUROC and PRR side by side.

### E.3 Per-task breakdown

Table E.8 groups the 30 runs by task, pooling GSM8K and MATH into a single Math task, and averages over the eight semantic estimators and the five model variants. A DiMS rule is the best single-trajectory strategy on all five tasks, ahead of the best baseline by 0.006 (QA) to 0.020 (Summarization) AUROC. The best baseline is `Full` on every task except Summarization, where it is `Last`. The gap between the best single-trajectory strategy and the $M{=}20$ i.i.d. reference ranges from 0.029 (Multi-hop QA) to 0.076 (QA).

### E.4 Ablations

We vary one component of the offline calibration of DiMS at a time, holding the others at their default: the rule that turns $\mathbf{w}^\star$ into a step set, the redundancy weight $\lambda$, the MMD kernel, the sentence encoder, the number $R$ of calibration trajectories per prompt, and the number $N_{\text{cal}}$ of calibration prompts. The $\lambda$, kernel and encoder ablations hold the selection rule at top-$M$; the $R$ and $N_{\text{cal}}$ ablations report both top-$M$ and coverage 75%. None of these components affects generation, labeling or the similarity model, so all ablations reuse the same trajectories, splits and labels. The choice of $\lambda$ has the largest effect: a fixed $\lambda{=}1$ lowers an estimator's average AUROC by up to 0.031 relative to $\lambda^\star$, whereas each of the other components changes it by at most 0.016.

#### E.4.1 Step-selection rule

Table E.9 compares the six rules of App. B.5.2: three that spend the budget $M{=}20$ (top-$M$, and draws with (WR) or without (WOR) replacement) and three whose cardinality is set by the weights (coverage 50, 75 and 90%). Each rule has its own $\lambda^\star$.

Per estimator, the spread of average AUROC across the six rules is at most 0.016 (Semantic Entropy), and at most 0.009 for three of the eight estimators. For seven estimators, this spread is smaller than DiMS's margin over `Random` in Table E.1; the exception is EigenVal (spread 0.014, margin 0.005 for top-$M$). A coverage rule has the best or tied-best average for seven estimators: coverage 90% for five of them (tied with coverage 75% for DegMat and KLE-Matérn) and coverage 50% for AD and CoCoA. Top-$M$ is the sole best rule only for KLE-Heat. Coverage 75% keeps more steps than top-$M$, which adds $\sim$14–15% to the overhead of the graph-based estimators and nothing for AD and CoCoA (Table E.1). WR is not the best rule in any of the 40 (estimator, model) blocks.

The main tables report deterministic rules. With the stochastic rules, the uncertainty score $U_{\text{DiMS}}(\boldsymbol{x})$ is itself random: over five replicate draws, the standard deviation of the AUROC is 0.008 for WR and $\sim$0.005 for WOR.

#### E.4.2 Redundancy penalty ($\lambda$)

Table E.10 compares three values of the weight $\lambda$ of the within-trajectory term $\mathbf{w}^\top \boldsymbol{C}_{\boldsymbol{x}}\mathbf{w}$ (Proposition 1): the smallest grid value $\lambda{=}0$ (i.e. $\lambda_{\text{eff}}{=}1/R$, App. B.5.1), a fixed $\lambda{=}1$, and the tuned $\lambda^\star$. $\lambda^\star$ has the best average AUROC for every estimator, ahead of $\lambda{=}0$ by 0.004 (AD) to 0.029 (KLE-Heat) and of $\lambda{=}1$ by 0.003 (CoCoA) to 0.031 (Semantic Entropy). The gain over $\lambda{=}0$ is largest for DegMat, EigenVal and the two KLE variants (0.026–0.029), intermediate for Eccentricity (0.016), and smallest for AD, Semantic Entropy and CoCoA (0.004–0.005). A fixed $\lambda{=}1$ improves on $\lambda{=}0$ for six estimators (by 0.002 to 0.016) but not for Semantic Entropy ($-0.026$) and AD ($-0.003$).

### E.4.3 MMD KERNEL

Table E.11 replaces the RBF kernel with a Laplacian and a Matérn ($\nu$=3/2) kernel, each with its bandwidth set per prompt by the median heuristic (App. D.2). The average AUROC changes by at most 0.005 (Eccentricity and EigenVal). RBF is the best or tied-best kernel for seven of the eight estimators; for EigenVal, the Laplacian kernel is ahead by 0.005.

### E.4.4 SENTENCE ENCODER

Table E.12 replaces `E5-large-v2` (1024-d) with `all-MiniLM-L6-v2` (384-d) and `BGE-M3` (1024-d). The encoder enters only the offline calibration: the deployed estimators use the same NLI or STS model (App. D.2) with all three encoders, so only the selected steps change. The average AUROC changes by at most 0.003. MiniLM is the best or tied-best encoder for seven of the eight estimators, and the sole best for two (Semantic Entropy and EigenVal); for DegMat, `E5-large-v2` is ahead by 0.002.

### E.4.5 CALIBRATION TRAJECTORIES ($R$)

Table E.13 reduces the number of sampled trajectories per calibration prompt from $R$=20 to 10 and 5, with $N_{\text{cal}}$=100 prompts and all other settings unchanged. Across the 32 reduced-$R$ cells, the average AUROC changes by $-0.005$ (Semantic Entropy, coverage 75%, $R$=5) to $+0.002$ (AD, coverage 75%, $R$=10), with a mean change below 0.001 in absolute value. Every change is smaller than the interval half-widths (0.007–0.008).

### E.4.6 CALIBRATION PROMPTS ($N_{\text{cal}}$)

Table E.14 reduces the number of calibration prompts from $N_{\text{cal}}$=100 to 50 and 10 (nested subsets), with $R$=20 trajectories per prompt; it reports the $N_{\text{cal}}$ columns of Table 1 with intervals. Across the 32 reduced-$N_{\text{cal}}$ cells, the average AUROC changes by $-0.007$ (Eccentricity, top-$M$, $N_{\text{cal}}$=10) to $+0.003$ (Semantic Entropy, top-$M$, $N_{\text{cal}}$=10), with a mean change below 0.001 in absolute value. Every change is smaller than the interval half-widths (0.007–0.008).

## E.5 TRANSFER ACROSS DATASETS AND MODELS

Tables E.15 and E.16 deploy the top-$M$ weights $\mathbf{w}^\star$ calibrated on one model variant (resp. dataset) on another, with the other coordinate fixed. Each cell is the change in test AUROC relative to in-domain calibration, averaged over the semantic estimators and every shared run pair; negative values are losses.

**Across datasets.** The mean change is $-0.007$, from 0.000 (HotpotQA $\rightarrow$ WMT14) to $-0.017$ (WMT14 $\rightarrow$ GSM8K). By source, the mean change ranges from $-0.005$ (HotpotQA and SAMSum) to $-0.010$ (GSM8K and WMT14); by target, from $-0.005$ (MATH) to $-0.011$ (HotpotQA).

**Across models.** The mean change is $-0.017$. Weights calibrated under low-confidence remasking (`lc`) lose 0.009 to 0.018 on average; weights calibrated under random remasking (`rd`) lose 0.021 (`LLaDA-1.5`) and 0.024 (`Dream`). The largest losses occur from an `rd` source to the `lc` targets `Dream` and `Nemotron` ($-0.032$ to $-0.047$).

For comparison, DIMS's in-domain margin over `Random` is 0.017–0.019 AUROC, and over `Last` 0.033–0.035 (Table E.8).

Table E.1: **Detailed AUROC comparison across models.** Each model column reports the mean AUROC over the six datasets; intervals correspond to 1.96 standard errors computed from 500 bootstrap resamples. Yellow denotes the **Multi-sample** reference, which uses $M = 20$ independent generations. All rows below it use a **Single trajectory**: gray denotes standard trajectory-selection baselines, while blue denotes our DiMS strategies. **Bold** indicates the best Single-trajectory result within each estimator–model block; the Multi-sample reference does not compete. Information-based scores are reported separately since they do not construct a semantic sample set. We also report the average computational overhead at inference (see App. D.4 for details).

| UQ estimator | Setting | Strategy | AUROC ↑ | | | | | Overhead (s) ↓ |
|---|---|---|---|---|---|---|---|---|
| | | | LLaDA-1.5 (`lc`) | LLaDA-1.5 (`rd`) | Dream (`lc`) | Dream (`rd`) | Nemotron (`lc`) | |
| Info-based | Single pass | Perplexity | 0.625 ± 0.018 | 0.577 ± 0.018 | 0.623 ± 0.019 | 0.562 ± 0.019 | 0.554 ± 0.018 | – |
| | | MTE | 0.659 ± 0.018 | 0.621 ± 0.018 | 0.669 ± 0.017 | 0.607 ± 0.019 | 0.554 ± 0.018 | – |
| | | MSP | 0.644 ± 0.018 | 0.612 ± 0.018 | 0.651 ± 0.018 | 0.600 ± 0.019 | 0.549 ± 0.018 | – |
| | | TrajEnt | 0.671 ± 0.017 | 0.685 ± 0.017 | 0.636 ± 0.018 | 0.654 ± 0.018 | 0.637 ± 0.017 | – |
| | | CommitNLL | 0.722 ± 0.016 | 0.712 ± 0.016 | 0.732 ± 0.016 | 0.719 ± 0.016 | 0.610 ± 0.017 | – |
| | | $\mathrm{MCNLL}^{\mathrm{norm}}$ | 0.612 ± 0.018 | 0.592 ± 0.018 | 0.653 ± 0.018 | 0.648 ± 0.018 | 0.586 ± 0.017 | 0.83 ± 0.38 |
| Sem. Entropy | Multi-sample | 20 i.i.d. | 0.723 ± 0.016 | 0.695 ± 0.017 | 0.760 ± 0.016 | 0.669 ± 0.017 | 0.653 ± 0.017 | 67.26 ± 8.63 |
| | Single trajectory | Full | 0.598 ± 0.018 | 0.619 ± 0.017 | 0.613 ± 0.018 | 0.597 ± 0.019 | 0.550 ± 0.018 | 7.85 ± 1.02 |
| | | Random | 0.581 ± 0.018 | 0.605 ± 0.018 | 0.616 ± 0.018 | 0.594 ± 0.019 | 0.550 ± 0.018 | 3.97 ± 0.43 |
| | | Last | 0.610 ± 0.018 | 0.589 ± 0.018 | 0.695 ± 0.017 | 0.585 ± 0.019 | 0.575 ± 0.018 | 3.80 ± 0.42 |
| | | Centroid | 0.566 ± 0.018 | 0.584 ± 0.018 | 0.585 ± 0.019 | 0.579 ± 0.019 | 0.513 ± 0.018 | 4.10 ± 0.44 |
| | | DiMS (top-$M$) | 0.613 ± 0.018 | 0.624 ± 0.018 | **0.696** ± 0.017 | 0.592 ± 0.020 | **0.591** ± 0.018 | 3.92 ± 0.42 |
| | | DiMS (cov. 75%) | **0.625** ± 0.018 | **0.629** ± 0.017 | 0.695 ± 0.017 | **0.604** ± 0.019 | 0.585 ± 0.018 | 4.47 ± 0.57 |
| DegMat | Multi-sample | 20 i.i.d. | 0.696 ± 0.016 | 0.722 ± 0.016 | 0.738 ± 0.016 | 0.702 ± 0.017 | 0.648 ± 0.017 | 66.78 ± 8.64 |
| | Single trajectory | Full | **0.650** ± 0.017 | 0.653 ± 0.017 | 0.665 ± 0.017 | 0.621 ± 0.018 | 0.599 ± 0.018 | 8.59 ± 1.67 |
| | | Random | 0.639 ± 0.017 | 0.643 ± 0.017 | 0.656 ± 0.017 | 0.609 ± 0.018 | 0.596 ± 0.018 | 3.47 ± 0.45 |
| | | Last | 0.604 ± 0.018 | 0.568 ± 0.016 | 0.653 ± 0.017 | 0.565 ± 0.015 | 0.587 ± 0.017 | 3.41 ± 0.44 |
| | | Centroid | 0.647 ± 0.017 | 0.650 ± 0.017 | 0.638 ± 0.018 | 0.609 ± 0.018 | 0.593 ± 0.018 | 3.57 ± 0.45 |
| | | DiMS (top-$M$) | 0.634 ± 0.018 | **0.665** ± 0.017 | **0.680** ± 0.017 | **0.627** ± 0.018 | **0.610** ± 0.017 | 3.47 ± 0.44 |
| | | DiMS (cov. 75%) | 0.644 ± 0.017 | 0.663 ± 0.017 | 0.678 ± 0.017 | **0.627** ± 0.018 | 0.601 ± 0.018 | 3.99 ± 0.57 |
| Eccentricity | Multi-sample | 20 i.i.d. | 0.689 ± 0.016 | 0.685 ± 0.017 | 0.724 ± 0.016 | 0.670 ± 0.017 | 0.627 ± 0.017 | 66.78 ± 8.64 |
| | Single trajectory | Full | **0.628** ± 0.018 | 0.646 ± 0.017 | 0.644 ± 0.018 | 0.612 ± 0.018 | 0.584 ± 0.018 | 8.59 ± 1.67 |
| | | Random | 0.615 ± 0.018 | 0.638 ± 0.018 | 0.639 ± 0.017 | 0.604 ± 0.018 | 0.579 ± 0.018 | 3.47 ± 0.45 |
| | | Last | 0.595 ± 0.018 | 0.568 ± 0.016 | 0.643 ± 0.018 | 0.564 ± 0.015 | 0.582 ± 0.017 | 3.41 ± 0.44 |
| | | Centroid | 0.621 ± 0.018 | 0.641 ± 0.017 | 0.631 ± 0.018 | 0.607 ± 0.018 | 0.579 ± 0.018 | 3.57 ± 0.45 |
| | | DiMS (top-$M$) | 0.612 ± 0.018 | 0.649 ± 0.017 | **0.665** ± 0.017 | 0.617 ± 0.018 | **0.591** ± 0.017 | 3.46 ± 0.44 |
| | | DiMS (cov. 75%) | 0.615 ± 0.018 | **0.654** ± 0.017 | 0.662 ± 0.017 | **0.621** ± 0.018 | 0.589 ± 0.018 | 3.94 ± 0.55 |
| EigenVal | Multi-sample | 20 i.i.d. | 0.665 ± 0.017 | 0.677 ± 0.017 | 0.700 ± 0.017 | 0.652 ± 0.018 | 0.591 ± 0.018 | 66.78 ± 8.64 |
| | Single trajectory | Full | **0.624** ± 0.018 | 0.646 ± 0.017 | 0.642 ± 0.018 | 0.610 ± 0.018 | 0.583 ± 0.018 | 8.59 ± 1.67 |
| | | Random | 0.615 ± 0.018 | 0.630 ± 0.018 | 0.631 ± 0.018 | 0.601 ± 0.018 | 0.579 ± 0.018 | 3.47 ± 0.45 |
| | | Last | 0.564 ± 0.019 | 0.532 ± 0.016 | 0.631 ± 0.018 | 0.539 ± 0.016 | 0.551 ± 0.017 | 3.41 ± 0.44 |
| | | Centroid | 0.614 ± 0.018 | 0.635 ± 0.018 | 0.629 ± 0.018 | 0.606 ± 0.018 | 0.579 ± 0.018 | 3.57 ± 0.45 |
| | | DiMS (top-$M$) | 0.600 ± 0.018 | 0.642 ± 0.018 | **0.650** ± 0.018 | 0.605 ± 0.018 | 0.580 ± 0.018 | 3.47 ± 0.44 |
| | | DiMS (cov. 75%) | 0.612 ± 0.018 | **0.652** ± 0.017 | 0.640 ± 0.018 | **0.611** ± 0.018 | **0.585** ± 0.018 | 3.98 ± 0.57 |
| KLE-Heat | Multi-sample | 20 i.i.d. | 0.697 ± 0.016 | 0.727 ± 0.016 | 0.739 ± 0.016 | 0.707 ± 0.017 | 0.653 ± 0.017 | 66.78 ± 8.64 |
| | Single trajectory | Full | 0.646 ± 0.018 | 0.651 ± 0.017 | 0.664 ± 0.017 | 0.616 ± 0.018 | 0.591 ± 0.018 | 8.59 ± 1.67 |
| | | Random | 0.638 ± 0.017 | 0.640 ± 0.017 | 0.662 ± 0.017 | 0.607 ± 0.018 | 0.596 ± 0.018 | 3.47 ± 0.45 |
| | | Last | 0.605 ± 0.018 | 0.568 ± 0.016 | 0.657 ± 0.017 | 0.566 ± 0.015 | 0.587 ± 0.017 | 3.41 ± 0.44 |
| | | Centroid | **0.648** ± 0.017 | 0.650 ± 0.017 | 0.648 ± 0.017 | 0.620 ± 0.018 | 0.597 ± 0.018 | 3.57 ± 0.45 |
| | | DiMS (top-$M$) | 0.639 ± 0.018 | **0.664** ± 0.017 | **0.686** ± 0.017 | 0.629 ± 0.018 | **0.611** ± 0.017 | 3.47 ± 0.44 |
| | | DiMS (cov. 75%) | 0.643 ± 0.017 | 0.661 ± 0.017 | 0.682 ± 0.017 | **0.633** ± 0.018 | 0.600 ± 0.018 | 3.98 ± 0.57 |
| KLE-Matérn | Multi-sample | 20 i.i.d. | 0.697 ± 0.016 | 0.724 ± 0.016 | 0.738 ± 0.016 | 0.707 ± 0.017 | 0.653 ± 0.017 | 66.78 ± 8.64 |
| | Single trajectory | Full | **0.650** ± 0.017 | 0.653 ± 0.017 | 0.671 ± 0.017 | 0.620 ± 0.018 | 0.598 ± 0.018 | 8.59 ± 1.67 |
| | | Random | 0.637 ± 0.018 | 0.642 ± 0.017 | 0.661 ± 0.017 | 0.608 ± 0.018 | 0.596 ± 0.018 | 3.47 ± 0.45 |
| | | Last | 0.604 ± 0.018 | 0.568 ± 0.016 | 0.655 ± 0.017 | 0.566 ± 0.015 | 0.587 ± 0.017 | 3.41 ± 0.44 |
| | | Centroid | **0.650** ± 0.017 | 0.654 ± 0.017 | 0.647 ± 0.017 | 0.621 ± 0.018 | 0.596 ± 0.018 | 3.57 ± 0.45 |
| | | DiMS (top-$M$) | 0.627 ± 0.018 | **0.663** ± 0.017 | **0.689** ± 0.017 | 0.627 ± 0.018 | **0.612** ± 0.017 | 3.46 ± 0.44 |
| | | DiMS (cov. 75%) | 0.645 ± 0.017 | **0.663** ± 0.017 | 0.685 ± 0.017 | **0.628** ± 0.018 | 0.606 ± 0.018 | 3.98 ± 0.57 |
| AD | Multi-sample | 20 i.i.d. | 0.784 ± 0.014 | 0.798 ± 0.014 | 0.791 ± 0.015 | 0.751 ± 0.016 | 0.685 ± 0.017 | 66.66 ± 8.63 |
| | Single trajectory | Full | 0.696 ± 0.017 | 0.719 ± 0.017 | 0.703 ± 0.017 | 0.659 ± 0.018 | 0.564 ± 0.019 | 3.42 ± 0.44 |
| | | Random | 0.692 ± 0.017 | 0.715 ± 0.017 | 0.697 ± 0.017 | 0.656 ± 0.018 | 0.563 ± 0.019 | 3.36 ± 0.43 |
| | | Last | 0.721 ± 0.017 | 0.699 ± 0.017 | 0.741 ± 0.016 | 0.650 ± 0.018 | **0.620** ± 0.018 | 3.35 ± 0.43 |
| | | Centroid | 0.673 ± 0.018 | 0.715 ± 0.017 | 0.676 ± 0.017 | 0.654 ± 0.018 | 0.544 ± 0.019 | 3.44 ± 0.44 |
| | | DiMS (top-$M$) | 0.719 ± 0.017 | 0.722 ± 0.017 | **0.748** ± 0.016 | **0.662** ± 0.018 | 0.616 ± 0.018 | 3.36 ± 0.43 |
| | | DiMS (cov. 75%) | **0.727** ± 0.016 | **0.726** ± 0.017 | 0.746 ± 0.016 | **0.662** ± 0.018 | **0.620** ± 0.018 | 3.37 ± 0.43 |
| CoCoA | Multi-sample | 20 i.i.d. | 0.788 ± 0.014 | 0.798 ± 0.014 | 0.785 ± 0.015 | 0.773 ± 0.015 | 0.673 ± 0.017 | 66.66 ± 8.63 |
| | Single trajectory | Full | 0.758 ± 0.015 | 0.764 ± 0.015 | 0.750 ± 0.016 | 0.735 ± 0.016 | 0.614 ± 0.018 | 3.42 ± 0.44 |
| | | Random | 0.756 ± 0.015 | 0.760 ± 0.015 | 0.750 ± 0.016 | 0.732 ± 0.016 | 0.612 ± 0.018 | 3.36 ± 0.43 |
| | | Last | 0.748 ± 0.016 | 0.748 ± 0.015 | 0.762 ± 0.015 | 0.720 ± 0.017 | **0.640** ± 0.017 | 3.35 ± 0.43 |
| | | Centroid | 0.753 ± 0.015 | 0.764 ± 0.015 | 0.744 ± 0.016 | 0.738 ± 0.015 | 0.598 ± 0.018 | 3.44 ± 0.44 |
| | | DiMS (top-$M$) | 0.760 ± 0.015 | 0.764 ± 0.015 | **0.764** ± 0.015 | 0.737 ± 0.016 | 0.633 ± 0.018 | 3.36 ± 0.43 |
| | | DiMS (cov. 75%) | **0.763** ± 0.015 | **0.768** ± 0.015 | 0.762 ± 0.016 | **0.739** ± 0.016 | 0.629 ± 0.018 | 3.37 ± 0.43 |

Table E.2: **Detailed PRR comparison across models.** Each model column reports the mean PRR over the six datasets; intervals correspond to 1.96 standard errors computed from 500 bootstrap resamples. Yellow denotes the **Multi-sample** reference, which uses $M = 20$ independent generations. All rows below it use a **Single trajectory**: gray denotes standard trajectory-selection baselines, while blue denotes our DiMS strategies. **Bold** indicates the best Single-trajectory result within each estimator–model block; the Multi-sample reference does not compete. Information-based scores are reported separately since they do not construct a semantic sample set.

| UQ estimator | Setting | Strategy | PRR ↑ | | | | |
|---|---|---|---|---|---|---|---|
| | | | LLaDA-1.5 (1c) | LLaDA-1.5 (rd) | Dream (1c) | Dream (rd) | Nemotron (1c) |
| Info-based | Single pass | Perplexity | 0.220 ± 0.039 | 0.148 ± 0.041 | 0.241 ± 0.041 | 0.103 ± 0.050 | 0.084 ± 0.043 |
| | | MTE | 0.296 ± 0.039 | 0.236 ± 0.038 | 0.348 ± 0.038 | 0.221 ± 0.045 | 0.081 ± 0.045 |
| | | MSP | 0.269 ± 0.039 | 0.223 ± 0.038 | 0.317 ± 0.039 | 0.211 ± 0.046 | 0.074 ± 0.045 |
| | | TrajEnt | 0.324 ± 0.037 | 0.382 ± 0.035 | 0.244 ± 0.040 | 0.335 ± 0.043 | 0.266 ± 0.043 |
| | | CommitNLL | 0.425 ± 0.035 | 0.412 ± 0.035 | 0.466 ± 0.036 | 0.434 ± 0.041 | 0.217 ± 0.044 |
| | | $\text{MCNLL}^{\text{norm}}$ | 0.222 ± 0.037 | 0.194 ± 0.038 | 0.302 ± 0.040 | 0.291 ± 0.046 | 0.166 ± 0.042 |
| Sem. Entropy | **Multi-sample** | 20 i.i.d. | 0.444 ± 0.034 | 0.384 ± 0.035 | 0.516 ± 0.036 | 0.354 ± 0.040 | 0.290 ± 0.043 |
| | Single trajectory | Full | 0.198 ± 0.036 | 0.244 ± 0.037 | 0.218 ± 0.042 | 0.171 ± 0.048 | 0.083 ± 0.044 |
| | | Random | 0.161 ± 0.037 | 0.224 ± 0.038 | 0.238 ± 0.041 | 0.177 ± 0.047 | 0.076 ± 0.045 |
| | | Last | 0.237 ± 0.036 | 0.215 ± 0.038 | 0.394 ± 0.039 | 0.196 ± 0.047 | 0.146 ± 0.046 |
| | | Centroid | 0.152 ± 0.037 | 0.214 ± 0.037 | 0.168 ± 0.042 | 0.169 ± 0.045 | 0.021 ± 0.047 |
| | | DiMS (top-$M$) | 0.238 ± 0.037 | 0.259 ± 0.038 | **0.397** ± 0.038 | 0.178 ± 0.048 | **0.172** ± 0.044 |
| | | DiMS (cov. 75%) | **0.267** ± 0.036 | **0.261** ± 0.037 | 0.392 ± 0.038 | **0.201** ± 0.046 | 0.158 ± 0.044 |
| DegMat | **Multi-sample** | 20 i.i.d. | 0.357 ± 0.037 | 0.445 ± 0.035 | 0.492 ± 0.036 | 0.417 ± 0.041 | 0.280 ± 0.044 |
| | Single trajectory | Full | 0.267 ± 0.038 | 0.295 ± 0.039 | 0.360 ± 0.038 | 0.237 ± 0.044 | 0.192 ± 0.044 |
| | | Random | 0.248 ± 0.039 | 0.280 ± 0.039 | 0.322 ± 0.039 | 0.217 ± 0.044 | 0.177 ± 0.046 |
| | | Last | 0.182 ± 0.040 | 0.170 ± 0.040 | 0.341 ± 0.039 | 0.164 ± 0.046 | 0.181 ± 0.042 |
| | | Centroid | **0.272** ± 0.038 | 0.285 ± 0.040 | 0.300 ± 0.039 | 0.223 ± 0.043 | 0.164 ± 0.046 |
| | | DiMS (top-$M$) | 0.223 ± 0.039 | **0.317** ± 0.039 | **0.384** ± 0.038 | 0.248 ± 0.044 | **0.220** ± 0.042 |
| | | DiMS (cov. 75%) | 0.257 ± 0.039 | 0.303 ± 0.038 | 0.372 ± 0.038 | **0.262** ± 0.043 | 0.211 ± 0.042 |
| Eccentricity | **Multi-sample** | 20 i.i.d. | 0.340 ± 0.038 | 0.366 ± 0.036 | 0.459 ± 0.037 | 0.361 ± 0.042 | 0.224 ± 0.045 |
| | Single trajectory | Full | **0.233** ± 0.038 | 0.282 ± 0.038 | 0.320 ± 0.039 | 0.228 ± 0.043 | 0.170 ± 0.043 |
| | | Random | 0.212 ± 0.039 | 0.263 ± 0.039 | 0.297 ± 0.039 | 0.214 ± 0.043 | 0.149 ± 0.047 |
| | | Last | 0.171 ± 0.040 | 0.169 ± 0.040 | 0.325 ± 0.039 | 0.156 ± 0.046 | 0.167 ± 0.042 |
| | | Centroid | 0.224 ± 0.038 | 0.281 ± 0.039 | 0.291 ± 0.039 | 0.228 ± 0.043 | 0.157 ± 0.043 |
| | | DiMS (top-$M$) | 0.189 ± 0.039 | 0.288 ± 0.038 | **0.355** ± 0.039 | 0.232 ± 0.044 | **0.201** ± 0.041 |
| | | DiMS (cov. 75%) | 0.200 ± 0.039 | **0.298** ± 0.038 | 0.336 ± 0.039 | **0.252** ± 0.043 | 0.175 ± 0.041 |
| EigenVal | **Multi-sample** | 20 i.i.d. | 0.320 ± 0.038 | 0.365 ± 0.036 | 0.420 ± 0.037 | 0.320 ± 0.043 | 0.154 ± 0.045 |
| | Single trajectory | Full | **0.216** ± 0.038 | 0.280 ± 0.039 | 0.317 ± 0.039 | 0.230 ± 0.043 | 0.169 ± 0.042 |
| | | Random | **0.216** ± 0.040 | 0.255 ± 0.040 | 0.283 ± 0.040 | 0.213 ± 0.043 | 0.152 ± 0.045 |
| | | Last | 0.147 ± 0.040 | 0.131 ± 0.041 | 0.314 ± 0.039 | 0.139 ± 0.047 | 0.134 ± 0.043 |
| | | Centroid | 0.207 ± 0.039 | 0.271 ± 0.039 | 0.290 ± 0.039 | 0.230 ± 0.043 | 0.154 ± 0.044 |
| | | DiMS (top-$M$) | 0.186 ± 0.039 | 0.285 ± 0.039 | **0.334** ± 0.039 | 0.203 ± 0.045 | 0.162 ± 0.042 |
| | | DiMS (cov. 75%) | 0.203 ± 0.039 | **0.286** ± 0.039 | 0.319 ± 0.039 | **0.233** ± 0.043 | **0.178** ± 0.041 |
| KLE-Heat | **Multi-sample** | 20 i.i.d. | 0.364 ± 0.037 | 0.460 ± 0.035 | 0.490 ± 0.037 | 0.425 ± 0.040 | 0.292 ± 0.043 |
| | Single trajectory | Full | 0.271 ± 0.038 | 0.296 ± 0.038 | 0.362 ± 0.038 | 0.236 ± 0.043 | 0.158 ± 0.045 |
| | | Random | 0.246 ± 0.039 | 0.274 ± 0.039 | 0.328 ± 0.039 | 0.212 ± 0.044 | 0.177 ± 0.046 |
| | | Last | 0.184 ± 0.040 | 0.169 ± 0.040 | 0.350 ± 0.039 | 0.164 ± 0.046 | 0.182 ± 0.042 |
| | | Centroid | **0.273** ± 0.038 | 0.295 ± 0.039 | 0.311 ± 0.039 | 0.241 ± 0.043 | 0.177 ± 0.046 |
| | | DiMS (top-$M$) | 0.236 ± 0.039 | **0.315** ± 0.039 | **0.390** ± 0.037 | 0.257 ± 0.044 | **0.216** ± 0.041 |
| | | DiMS (cov. 75%) | 0.254 ± 0.039 | 0.310 ± 0.039 | 0.380 ± 0.037 | **0.276** ± 0.043 | 0.197 ± 0.043 |
| KLE-Matérn | **Multi-sample** | 20 i.i.d. | 0.364 ± 0.037 | 0.453 ± 0.035 | 0.496 ± 0.036 | 0.428 ± 0.040 | 0.290 ± 0.044 |
| | Single trajectory | Full | 0.270 ± 0.038 | 0.295 ± 0.038 | 0.372 ± 0.038 | 0.237 ± 0.044 | 0.189 ± 0.045 |
| | | Random | 0.244 ± 0.039 | 0.278 ± 0.039 | 0.329 ± 0.039 | 0.215 ± 0.044 | 0.175 ± 0.046 |
| | | Last | 0.183 ± 0.040 | 0.170 ± 0.040 | 0.348 ± 0.039 | 0.164 ± 0.046 | 0.182 ± 0.042 |
| | | Centroid | **0.276** ± 0.038 | 0.295 ± 0.039 | 0.306 ± 0.039 | 0.240 ± 0.044 | 0.167 ± 0.047 |
| | | DiMS (top-$M$) | 0.218 ± 0.039 | **0.314** ± 0.039 | **0.390** ± 0.038 | 0.255 ± 0.044 | **0.211** ± 0.042 |
| | | DiMS (cov. 75%) | 0.254 ± 0.039 | 0.308 ± 0.039 | 0.378 ± 0.038 | **0.263** ± 0.044 | 0.198 ± 0.042 |
| AD | **Multi-sample** | 20 i.i.d. | 0.565 ± 0.031 | 0.601 ± 0.030 | 0.592 ± 0.034 | 0.514 ± 0.040 | 0.341 ± 0.046 |
| | Single trajectory | Full | 0.397 ± 0.035 | 0.405 ± 0.037 | 0.397 ± 0.039 | 0.307 ± 0.045 | 0.084 ± 0.049 |
| | | Random | 0.387 ± 0.035 | 0.405 ± 0.037 | 0.374 ± 0.039 | 0.298 ± 0.045 | 0.080 ± 0.049 |
| | | Last | 0.438 ± 0.036 | 0.362 ± 0.039 | 0.475 ± 0.037 | 0.275 ± 0.046 | **0.202** ± 0.049 |
| | | Centroid | 0.358 ± 0.037 | 0.408 ± 0.037 | 0.349 ± 0.039 | 0.307 ± 0.044 | 0.058 ± 0.050 |
| | | DiMS (top-$M$) | 0.433 ± 0.036 | 0.407 ± 0.037 | 0.483 ± 0.037 | 0.320 ± 0.044 | 0.195 ± 0.049 |
| | | DiMS (cov. 75%) | **0.452** ± 0.035 | **0.421** ± 0.037 | **0.484** ± 0.037 | **0.321** ± 0.044 | 0.192 ± 0.048 |
| CoCoA | **Multi-sample** | 20 i.i.d. | 0.568 ± 0.031 | 0.579 ± 0.031 | 0.570 ± 0.034 | 0.561 ± 0.036 | 0.328 ± 0.046 |
| | Single trajectory | Full | 0.507 ± 0.033 | 0.503 ± 0.033 | 0.509 ± 0.036 | 0.479 ± 0.039 | 0.198 ± 0.048 |
| | | Random | 0.503 ± 0.033 | 0.498 ± 0.033 | 0.508 ± 0.036 | 0.475 ± 0.039 | 0.190 ± 0.047 |
| | | Last | 0.498 ± 0.034 | 0.480 ± 0.034 | 0.528 ± 0.035 | 0.449 ± 0.040 | **0.249** ± 0.048 |
| | | Centroid | 0.493 ± 0.034 | 0.499 ± 0.033 | 0.499 ± 0.036 | **0.497** ± 0.037 | 0.175 ± 0.048 |
| | | DiMS (top-$M$) | 0.512 ± 0.033 | 0.500 ± 0.033 | **0.529** ± 0.035 | **0.497** ± 0.038 | 0.229 ± 0.047 |
| | | DiMS (cov. 75%) | **0.516** ± 0.033 | **0.510** ± 0.033 | 0.526 ± 0.035 | **0.497** ± 0.038 | 0.219 ± 0.047 |

Table E.3: **Per-dataset AUROC and PRR for LLaDA-1.5 (`1c`).** Both metrics use the same label-free $\lambda^\star$. The top block holds the information-based scores, which do not construct a semantic sample set; every block below it is one semantic estimator. Yellow denotes the **Multi-sample** reference, which uses $M{=}20$ independent generations; gray denotes standard trajectory-selection baselines, while blue denotes our DiMS strategies. **Bold** indicates the best Single-trajectory result within each estimator block, per dataset and metric; the Multi-sample reference does not compete.

| UQ estimator | Strategy | TriviaQA | | GSM8K | | MATH | | WMT14 Fr-En | | SAMSum | | HotpotQA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR |
| Info-based | Perplexity | 0.643 | 0.294 | 0.608 | 0.191 | 0.632 | 0.291 | 0.645 | 0.223 | 0.556 | 0.092 | 0.667 | 0.231 |
| | MTE | 0.649 | 0.283 | 0.687 | 0.312 | 0.694 | 0.452 | 0.689 | 0.329 | 0.563 | 0.152 | 0.673 | 0.250 |
| | MSP | 0.634 | 0.246 | 0.668 | 0.275 | 0.685 | 0.446 | 0.655 | 0.264 | 0.558 | 0.152 | 0.661 | 0.230 |
| | TrajEnt | 0.706 | 0.450 | 0.635 | 0.255 | 0.714 | 0.502 | 0.626 | 0.219 | 0.592 | 0.099 | 0.750 | 0.420 |
| | CommitNLL | 0.739 | 0.489 | 0.720 | 0.406 | 0.778 | 0.614 | 0.721 | 0.377 | 0.626 | 0.244 | 0.749 | 0.423 |
| | $\text{MCNLL}^{\text{norm}}$ | 0.576 | 0.139 | 0.699 | 0.365 | 0.745 | 0.557 | 0.549 | 0.066 | 0.640 | 0.336 | 0.465 | −0.134 |
| Sem. Entropy | 20 i.i.d. | 0.795 | 0.663 | 0.749 | 0.473 | 0.791 | 0.658 | 0.691 | 0.389 | 0.614 | 0.164 | 0.699 | 0.315 |
| | Full | 0.795 | 0.649 | 0.641 | 0.243 | **0.618** | **0.331** | 0.583 | 0.105 | 0.492 | −0.053 | 0.457 | −0.089 |
| | Random | 0.789 | 0.632 | 0.632 | 0.202 | 0.615 | 0.288 | 0.556 | 0.042 | 0.454 | −0.086 | 0.441 | −0.111 |
| | Last | 0.808 | **0.675** | 0.582 | 0.172 | 0.595 | 0.271 | 0.641 | 0.227 | 0.531 | 0.119 | 0.501 | −0.044 |
| | Centroid | 0.746 | 0.561 | **0.649** | **0.271** | 0.597 | 0.291 | 0.534 | 0.021 | 0.474 | −0.053 | 0.394 | −0.179 |
| | DiMS (top-$M$) | 0.806 | 0.667 | 0.595 | 0.165 | 0.595 | 0.271 | 0.645 | 0.237 | 0.531 | 0.119 | **0.506** | −0.032 |
| | DiMS (cov. 75%) | **0.810** | 0.674 | 0.632 | 0.241 | 0.615 | 0.330 | **0.650** | **0.252** | **0.546** | **0.132** | 0.497 | **−0.026** |
| DegMat | 20 i.i.d. | 0.797 | 0.607 | 0.683 | 0.306 | 0.713 | 0.435 | 0.651 | 0.317 | 0.632 | 0.176 | 0.697 | 0.303 |
| | Full | **0.754** | **0.492** | 0.642 | 0.231 | 0.589 | 0.259 | **0.664** | **0.288** | 0.587 | 0.107 | 0.663 | 0.226 |
| | Random | 0.749 | 0.484 | 0.614 | 0.184 | 0.576 | 0.221 | 0.659 | 0.270 | 0.583 | **0.129** | 0.651 | 0.198 |
| | Last | 0.753 | 0.485 | 0.558 | 0.081 | 0.527 | 0.078 | 0.626 | 0.208 | 0.585 | 0.124 | 0.576 | 0.113 |
| | Centroid | 0.731 | 0.450 | **0.675** | **0.321** | **0.595** | 0.260 | 0.636 | 0.257 | 0.573 | 0.088 | 0.669 | **0.254** |
| | DiMS (top-$M$) | 0.752 | **0.492** | 0.612 | 0.143 | 0.577 | 0.194 | 0.634 | 0.211 | 0.575 | 0.079 | 0.656 | 0.216 |
| | DiMS (cov. 75%) | 0.748 | 0.486 | 0.639 | 0.211 | **0.595** | **0.272** | 0.618 | 0.208 | **0.594** | 0.121 | **0.672** | 0.245 |
| Eccentricity | 20 i.i.d. | 0.770 | 0.540 | 0.673 | 0.273 | 0.712 | 0.420 | 0.651 | 0.317 | 0.648 | 0.224 | 0.680 | 0.265 |
| | Full | **0.702** | **0.424** | **0.648** | **0.234** | **0.607** | **0.270** | 0.654 | 0.275 | 0.566 | 0.094 | 0.588 | 0.103 |
| | Random | 0.685 | 0.394 | 0.622 | 0.218 | 0.584 | 0.220 | **0.660** | 0.278 | 0.554 | 0.084 | 0.584 | 0.076 |
| | Last | 0.700 | 0.386 | 0.562 | 0.091 | 0.530 | 0.091 | 0.623 | 0.200 | 0.584 | **0.156** | 0.571 | 0.101 |
| | Centroid | 0.685 | 0.394 | 0.644 | 0.229 | 0.592 | 0.227 | 0.644 | **0.284** | 0.564 | 0.095 | **0.599** | **0.118** |
| | DiMS (top-$M$) | 0.691 | 0.376 | 0.611 | 0.156 | 0.587 | 0.228 | 0.623 | 0.200 | 0.570 | 0.091 | 0.587 | 0.084 |
| | DiMS (cov. 75%) | 0.697 | 0.367 | 0.609 | 0.188 | 0.589 | 0.210 | 0.616 | 0.201 | **0.590** | 0.127 | 0.591 | 0.104 |
| EigenVal | 20 i.i.d. | 0.747 | 0.465 | 0.656 | 0.292 | 0.699 | 0.425 | 0.591 | 0.273 | 0.634 | 0.226 | 0.665 | 0.241 |
| | Full | **0.701** | **0.417** | 0.639 | 0.209 | **0.604** | 0.252 | **0.664** | **0.279** | 0.561 | 0.080 | 0.576 | 0.055 |
| | Random | 0.685 | 0.407 | 0.621 | 0.216 | 0.596 | 0.243 | 0.653 | 0.266 | 0.556 | 0.097 | 0.577 | 0.067 |
| | Last | 0.687 | 0.351 | 0.531 | 0.078 | 0.500 | 0.059 | 0.590 | 0.207 | 0.568 | **0.146** | 0.508 | 0.043 |
| | Centroid | 0.684 | **0.417** | **0.645** | **0.218** | 0.588 | 0.203 | 0.632 | 0.240 | 0.559 | 0.087 | 0.577 | **0.079** |
| | DiMS (top-$M$) | 0.686 | 0.362 | 0.603 | 0.196 | 0.567 | 0.219 | 0.613 | 0.210 | 0.563 | 0.069 | 0.569 | 0.058 |
| | DiMS (cov. 75%) | 0.691 | 0.392 | 0.639 | 0.216 | **0.604** | **0.261** | 0.576 | 0.182 | **0.577** | 0.102 | **0.582** | 0.062 |
| KLE-Heat | 20 i.i.d. | 0.805 | 0.638 | 0.685 | 0.309 | 0.714 | 0.436 | 0.651 | 0.317 | 0.629 | 0.171 | 0.700 | 0.310 |
| | Full | 0.759 | 0.513 | 0.648 | 0.241 | **0.610** | **0.280** | 0.644 | **0.283** | 0.583 | 0.114 | 0.630 | 0.196 |
| | Random | 0.758 | 0.509 | 0.601 | 0.161 | 0.571 | 0.216 | **0.658** | 0.267 | **0.588** | **0.127** | 0.652 | 0.194 |
| | Last | **0.760** | 0.511 | 0.559 | 0.083 | 0.527 | 0.077 | 0.626 | 0.207 | 0.582 | 0.114 | 0.576 | 0.113 |
| | Centroid | 0.749 | 0.511 | **0.668** | **0.308** | 0.579 | 0.199 | 0.638 | 0.257 | 0.578 | 0.096 | **0.677** | **0.267** |
| | DiMS (top-$M$) | **0.760** | **0.516** | 0.622 | 0.172 | 0.582 | 0.205 | 0.634 | 0.211 | 0.578 | 0.080 | 0.660 | 0.229 |
| | DiMS (cov. 75%) | 0.756 | 0.514 | 0.630 | 0.184 | 0.594 | 0.251 | 0.618 | 0.208 | 0.587 | 0.107 | 0.675 | 0.258 |
| KLE-Matérn | 20 i.i.d. | 0.804 | 0.639 | 0.683 | 0.307 | 0.714 | 0.438 | 0.651 | 0.317 | 0.631 | 0.175 | 0.698 | 0.308 |
| | Full | 0.753 | **0.496** | 0.641 | 0.228 | **0.590** | **0.257** | **0.664** | **0.290** | **0.588** | 0.118 | 0.663 | 0.229 |
| | Random | 0.749 | 0.492 | 0.608 | 0.171 | 0.574 | 0.211 | 0.658 | 0.266 | 0.584 | **0.127** | 0.650 | 0.197 |
| | Last | **0.755** | 0.495 | 0.558 | 0.082 | 0.527 | 0.078 | 0.626 | 0.207 | 0.584 | 0.121 | 0.576 | 0.113 |
| | Centroid | 0.743 | 0.484 | **0.673** | **0.311** | 0.588 | 0.231 | 0.644 | 0.265 | 0.575 | 0.096 | 0.677 | **0.269** |
| | DiMS (top-$M$) | 0.754 | **0.496** | 0.548 | 0.074 | 0.588 | 0.217 | 0.634 | 0.211 | 0.575 | 0.072 | 0.663 | 0.236 |
| | DiMS (cov. 75%) | 0.750 | 0.495 | 0.651 | 0.238 | 0.585 | 0.218 | 0.618 | 0.207 | 0.586 | 0.106 | **0.678** | 0.258 |
| AD | 20 i.i.d. | 0.851 | 0.805 | 0.830 | 0.630 | 0.852 | 0.797 | 0.756 | 0.462 | 0.683 | 0.307 | 0.732 | 0.390 |
| | Full | 0.817 | 0.662 | **0.749** | **0.483** | 0.776 | 0.659 | 0.644 | 0.227 | 0.565 | 0.137 | 0.624 | 0.214 |
| | Random | 0.815 | 0.651 | 0.745 | 0.475 | 0.768 | 0.647 | 0.639 | 0.228 | 0.561 | 0.115 | 0.624 | 0.208 |
| | Last | **0.820** | 0.677 | 0.744 | 0.452 | 0.767 | 0.612 | **0.695** | 0.332 | 0.636 | 0.255 | 0.667 | 0.299 |
| | Centroid | 0.803 | 0.624 | 0.737 | 0.457 | 0.754 | 0.588 | 0.617 | 0.194 | 0.557 | 0.128 | 0.573 | 0.158 |
| | DiMS (top-$M$) | 0.817 | 0.671 | 0.742 | 0.452 | 0.776 | 0.626 | 0.693 | **0.335** | 0.617 | 0.214 | 0.667 | 0.299 |
| | DiMS (cov. 75%) | **0.820** | **0.682** | 0.745 | 0.449 | **0.785** | **0.662** | 0.693 | **0.335** | **0.643** | **0.269** | **0.678** | **0.313** |
| CoCoA | 20 i.i.d. | 0.814 | 0.683 | 0.843 | 0.654 | 0.866 | 0.801 | 0.764 | 0.480 | 0.676 | 0.313 | 0.764 | 0.476 |
| | Full | 0.805 | 0.648 | 0.803 | 0.579 | **0.835** | 0.722 | 0.727 | 0.402 | 0.622 | 0.251 | **0.756** | 0.440 |
| | Random | 0.805 | 0.646 | 0.800 | 0.572 | 0.829 | 0.710 | 0.721 | 0.391 | 0.626 | 0.251 | **0.756** | **0.447** |
| | Last | **0.809** | **0.659** | 0.768 | 0.508 | 0.808 | 0.696 | **0.732** | **0.415** | 0.645 | 0.295 | 0.727 | 0.411 |
| | Centroid | 0.800 | 0.626 | 0.799 | 0.568 | 0.826 | 0.703 | 0.722 | 0.381 | 0.620 | 0.241 | 0.753 | 0.439 |
| | DiMS (top-$M$) | 0.807 | 0.658 | **0.805** | **0.581** | 0.829 | 0.713 | 0.731 | 0.413 | 0.639 | 0.283 | 0.747 | 0.427 |
| | DiMS (cov. 75%) | 0.808 | **0.659** | 0.802 | 0.564 | 0.834 | **0.723** | 0.731 | 0.413 | **0.649** | **0.301** | 0.753 | 0.432 |

Table E.4: **Per-dataset AUROC and PRR for LLaDA-1.5 (`rd`).** Both metrics use the same label-free $\lambda^\star$. The top block holds the information-based scores, which do not construct a semantic sample set; every block below it is one semantic estimator. Yellow denotes the **Multi-sample** reference, which uses $M{=}20$ independent generations; gray denotes standard trajectory-selection baselines, while blue denotes our DIMS strategies. **Bold** indicates the best Single-trajectory result within each estimator block, per dataset and metric; the Multi-sample reference does not compete.

| UQ estimator | Strategy | TriviaQA | | GSM8K | | MATH | | WMT14 Fr-En | | SAMSum | | HotpotQA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR |
| Info-based | Perplexity | 0.612 | 0.249 | 0.550 | 0.092 | 0.605 | 0.246 | 0.574 | 0.113 | 0.518 | 0.019 | 0.607 | 0.168 |
| | MTE | 0.600 | 0.210 | 0.707 | 0.355 | 0.701 | 0.516 | 0.587 | 0.146 | 0.506 | −0.027 | 0.624 | 0.216 |
| | MSP | 0.595 | 0.204 | 0.688 | 0.341 | 0.680 | 0.480 | 0.583 | 0.143 | 0.515 | −0.012 | 0.613 | 0.184 |
| | TrajEnt | 0.726 | 0.541 | 0.704 | 0.401 | 0.749 | 0.655 | 0.624 | 0.215 | 0.622 | 0.196 | 0.685 | 0.287 |
| | CommitNLL | 0.777 | 0.567 | 0.693 | 0.368 | 0.769 | 0.662 | 0.668 | 0.283 | 0.648 | 0.239 | 0.719 | 0.350 |
| | $\mathrm{MCNLL}^{\mathrm{norm}}$ | 0.583 | 0.165 | 0.673 | 0.319 | 0.749 | 0.629 | 0.554 | 0.058 | 0.509 | 0.005 | 0.486 | −0.010 |
| Sem. Entropy | 20 i.i.d. | 0.709 | 0.403 | 0.759 | 0.496 | 0.817 | 0.793 | 0.692 | 0.343 | 0.571 | 0.084 | 0.620 | 0.185 |
| | Full | 0.688 | 0.459 | 0.649 | 0.250 | 0.695 | 0.532 | **0.623** | **0.192** | 0.479 | −0.035 | 0.580 | 0.067 |
| | Random | 0.683 | 0.460 | 0.660 | 0.302 | 0.671 | 0.406 | 0.579 | 0.147 | 0.467 | −0.056 | 0.571 | 0.087 |
| | Last | 0.666 | 0.413 | 0.628 | 0.292 | 0.628 | 0.385 | 0.571 | 0.143 | 0.497 | −0.000 | 0.545 | 0.058 |
| | Centroid | 0.690 | 0.461 | 0.643 | 0.267 | 0.704 | **0.560** | 0.573 | 0.132 | 0.413 | −0.118 | 0.481 | −0.017 |
| | DIMS (top-$M$) | 0.691 | 0.445 | 0.644 | 0.281 | **0.711** | 0.521 | 0.613 | 0.174 | 0.501 | 0.014 | **0.585** | **0.118** |
| | DIMS (cov. 75%) | **0.699** | **0.477** | **0.674** | **0.332** | 0.710 | 0.515 | 0.619 | 0.184 | **0.505** | **0.021** | 0.567 | 0.035 |
| DegMat | 20 i.i.d. | 0.833 | 0.725 | 0.767 | 0.511 | 0.750 | 0.609 | 0.688 | 0.323 | 0.638 | 0.238 | 0.654 | 0.266 |
| | Full | 0.708 | 0.460 | 0.667 | 0.358 | 0.663 | 0.351 | **0.631** | **0.219** | 0.618 | 0.208 | 0.631 | 0.175 |
| | Random | 0.701 | **0.467** | 0.661 | 0.350 | 0.654 | 0.317 | 0.602 | 0.177 | 0.613 | 0.188 | 0.626 | 0.180 |
| | Last | 0.638 | 0.354 | 0.563 | 0.129 | 0.554 | 0.162 | 0.538 | 0.143 | 0.568 | 0.138 | 0.546 | 0.095 |
| | Centroid | 0.690 | 0.416 | 0.681 | 0.356 | 0.670 | 0.363 | 0.606 | 0.167 | 0.621 | 0.209 | 0.630 | 0.198 |
| | DIMS (top-$M$) | 0.712 | 0.461 | **0.710** | **0.411** | **0.686** | **0.420** | 0.622 | 0.192 | 0.617 | 0.208 | **0.643** | **0.212** |
| | DIMS (cov. 75%) | **0.714** | 0.464 | 0.702 | 0.388 | 0.682 | 0.382 | 0.629 | 0.198 | **0.623** | **0.219** | 0.629 | 0.167 |
| Eccentricity | 20 i.i.d. | 0.800 | 0.664 | 0.722 | 0.395 | 0.723 | 0.540 | 0.668 | 0.265 | 0.617 | 0.187 | 0.582 | 0.144 |
| | Full | 0.690 | 0.404 | 0.685 | 0.356 | 0.663 | 0.388 | 0.617 | 0.180 | 0.615 | 0.194 | 0.606 | 0.170 |
| | Random | 0.682 | 0.404 | 0.677 | 0.348 | 0.657 | 0.324 | 0.597 | 0.162 | 0.593 | 0.160 | **0.620** | **0.178** |
| | Last | 0.628 | 0.312 | 0.568 | 0.143 | 0.557 | 0.173 | 0.538 | 0.143 | 0.570 | 0.152 | 0.546 | 0.091 |
| | Centroid | 0.687 | 0.400 | 0.682 | 0.339 | 0.667 | 0.418 | 0.608 | 0.185 | 0.613 | 0.200 | 0.590 | 0.142 |
| | DIMS (top-$M$) | **0.691** | 0.406 | 0.700 | 0.377 | 0.677 | 0.405 | 0.618 | 0.178 | 0.611 | 0.198 | 0.599 | 0.162 |
| | DIMS (cov. 75%) | **0.691** | **0.424** | **0.703** | **0.381** | **0.680** | **0.421** | **0.622** | **0.197** | **0.622** | **0.203** | 0.606 | 0.161 |
| EigenVal | 20 i.i.d. | 0.789 | 0.652 | 0.739 | 0.424 | 0.731 | 0.576 | 0.639 | 0.262 | 0.602 | 0.172 | 0.561 | 0.103 |
| | Full | **0.687** | **0.395** | 0.676 | 0.346 | 0.672 | 0.388 | **0.623** | **0.209** | 0.609 | 0.194 | 0.610 | 0.149 |
| | Random | 0.676 | 0.389 | 0.671 | 0.333 | 0.652 | 0.315 | 0.577 | 0.169 | 0.593 | 0.163 | 0.609 | **0.162** |
| | Last | 0.602 | 0.287 | 0.478 | 0.039 | 0.550 | 0.200 | 0.516 | 0.099 | 0.539 | 0.137 | 0.510 | 0.023 |
| | Centroid | 0.684 | 0.387 | 0.673 | 0.341 | 0.668 | 0.399 | 0.588 | 0.150 | 0.606 | **0.203** | 0.590 | 0.144 |
| | DIMS (top-$M$) | 0.684 | 0.382 | **0.701** | 0.383 | 0.673 | **0.428** | 0.588 | 0.175 | 0.613 | 0.202 | 0.594 | 0.141 |
| | DIMS (cov. 75%) | 0.681 | 0.376 | 0.700 | **0.385** | **0.682** | 0.404 | 0.620 | 0.189 | **0.615** | 0.198 | **0.616** | **0.162** |
| KLE-Heat | 20 i.i.d. | 0.839 | 0.748 | 0.769 | 0.511 | 0.751 | 0.619 | 0.694 | 0.343 | 0.638 | 0.243 | 0.669 | 0.298 |
| | Full | 0.709 | 0.464 | 0.693 | 0.384 | 0.667 | 0.388 | 0.605 | 0.156 | **0.620** | 0.206 | 0.613 | 0.180 |
| | Random | 0.703 | 0.482 | 0.651 | 0.326 | 0.648 | 0.297 | 0.602 | 0.178 | 0.614 | 0.187 | 0.624 | 0.177 |
| | Last | 0.639 | 0.357 | 0.563 | 0.129 | 0.554 | 0.162 | 0.538 | 0.137 | 0.568 | 0.138 | 0.546 | 0.095 |
| | Centroid | 0.711 | **0.503** | 0.667 | 0.338 | 0.663 | 0.344 | 0.602 | 0.175 | **0.620** | 0.208 | 0.637 | 0.205 |
| | DIMS (top-$M$) | 0.712 | 0.473 | **0.701** | **0.390** | 0.684 | 0.408 | 0.624 | 0.196 | 0.619 | 0.207 | **0.645** | **0.214** |
| | DIMS (cov. 75%) | **0.716** | 0.477 | 0.696 | 0.381 | **0.686** | **0.419** | **0.625** | **0.197** | 0.619 | **0.215** | 0.625 | 0.173 |
| KLE-Matérn | 20 i.i.d. | 0.835 | 0.745 | 0.768 | 0.514 | 0.751 | 0.613 | 0.690 | 0.330 | 0.639 | 0.242 | 0.658 | 0.277 |
| | Full | 0.710 | 0.471 | 0.667 | 0.354 | 0.663 | 0.347 | **0.629** | **0.213** | 0.618 | 0.208 | 0.630 | 0.174 |
| | Random | 0.703 | 0.479 | 0.657 | 0.340 | 0.652 | 0.305 | 0.601 | 0.177 | 0.613 | 0.187 | 0.625 | 0.180 |
| | Last | 0.638 | 0.354 | 0.563 | 0.131 | 0.554 | 0.161 | 0.538 | 0.141 | 0.568 | 0.139 | 0.546 | 0.095 |
| | Centroid | **0.722** | **0.500** | 0.674 | 0.341 | 0.664 | 0.343 | 0.610 | 0.178 | 0.620 | 0.210 | 0.634 | 0.199 |
| | DIMS (top-$M$) | 0.714 | 0.478 | **0.702** | **0.391** | 0.683 | **0.408** | 0.623 | 0.191 | 0.617 | 0.208 | **0.641** | **0.206** |
| | DIMS (cov. 75%) | 0.717 | 0.484 | 0.698 | 0.382 | **0.684** | 0.404 | 0.628 | 0.197 | **0.623** | **0.217** | 0.628 | 0.165 |
| AD | 20 i.i.d. | 0.869 | 0.783 | 0.855 | 0.661 | 0.844 | 0.797 | 0.747 | 0.468 | 0.687 | 0.347 | 0.786 | 0.548 |
| | Full | 0.796 | 0.585 | 0.760 | 0.435 | 0.751 | 0.526 | **0.689** | **0.321** | 0.608 | 0.175 | 0.707 | 0.387 |
| | Random | 0.795 | 0.584 | 0.756 | 0.435 | 0.754 | 0.535 | 0.669 | 0.307 | 0.610 | 0.178 | 0.708 | **0.391** |
| | Last | 0.775 | 0.513 | 0.738 | 0.393 | 0.713 | 0.444 | 0.677 | 0.304 | 0.589 | 0.144 | 0.699 | 0.370 |
| | Centroid | 0.802 | **0.599** | 0.772 | 0.453 | **0.777** | **0.605** | 0.661 | 0.294 | 0.599 | 0.161 | 0.680 | 0.337 |
| | DIMS (top-$M$) | 0.801 | 0.589 | 0.762 | 0.437 | 0.772 | 0.556 | 0.680 | 0.310 | 0.609 | 0.170 | **0.709** | 0.380 |
| | DIMS (cov. 75%) | **0.804** | 0.592 | **0.778** | **0.471** | 0.774 | 0.571 | 0.684 | 0.320 | **0.612** | **0.186** | 0.706 | 0.386 |
| CoCoA | 20 i.i.d. | 0.837 | 0.688 | 0.857 | 0.673 | 0.858 | 0.820 | 0.757 | 0.475 | 0.692 | 0.340 | 0.788 | 0.477 |
| | Full | **0.832** | **0.667** | 0.798 | 0.524 | 0.818 | 0.730 | **0.712** | **0.375** | 0.650 | 0.258 | **0.773** | 0.461 |
| | Random | 0.831 | 0.664 | 0.794 | 0.523 | 0.820 | 0.730 | 0.696 | 0.359 | 0.650 | 0.253 | 0.772 | 0.460 |
| | Last | 0.830 | 0.665 | 0.776 | 0.488 | 0.789 | 0.687 | 0.693 | 0.336 | 0.635 | 0.239 | 0.766 | **0.464** |
| | Centroid | 0.828 | 0.662 | 0.809 | **0.554** | 0.832 | 0.758 | 0.702 | 0.344 | 0.650 | 0.248 | 0.763 | 0.429 |
| | DIMS (top-$M$) | 0.830 | 0.660 | 0.805 | 0.541 | 0.828 | **0.763** | 0.704 | 0.351 | 0.650 | 0.250 | 0.763 | 0.434 |
| | DIMS (cov. 75%) | 0.830 | 0.663 | **0.811** | 0.553 | **0.835** | 0.761 | 0.711 | 0.367 | **0.652** | **0.262** | **0.773** | 0.454 |

Table E.5: **Per-dataset AUROC and PRR for Dream (`lc`).** Both metrics use the same label-free $\lambda^\star$. The top block holds the information-based scores, which do not construct a semantic sample set; every block below it is one semantic estimator. Yellow denotes the **Multi-sample** reference, which uses $M{=}20$ independent generations; gray denotes standard trajectory-selection baselines, while blue denotes our DiMS strategies. **Bold** indicates the best Single-trajectory result within each estimator block, per dataset and metric; the Multi-sample reference does not compete.

| UQ estimator | Strategy | TriviaQA | | GSM8K | | MATH | | WMT14 Fr-En | | SAMSum | | HotpotQA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR |
| Info-based | Perplexity | 0.731 | 0.444 | 0.614 | 0.168 | 0.748 | 0.512 | 0.589 | 0.171 | 0.503 | 0.040 | 0.553 | 0.110 |
| | MTE | 0.741 | 0.466 | 0.748 | 0.533 | 0.761 | 0.590 | 0.709 | 0.427 | 0.511 | 0.002 | 0.542 | 0.070 |
| | MSP | 0.725 | 0.426 | 0.739 | 0.514 | 0.759 | 0.577 | 0.652 | 0.334 | 0.495 | −0.017 | 0.538 | 0.065 |
| | TrajEnt | 0.707 | 0.381 | 0.656 | 0.361 | 0.434 | −0.158 | 0.751 | 0.469 | 0.644 | 0.217 | 0.624 | 0.193 |
| | CommitNLL | 0.854 | 0.670 | 0.773 | 0.564 | 0.774 | 0.642 | 0.715 | 0.479 | 0.622 | 0.216 | 0.651 | 0.222 |
| | $\mathrm{MCNLL}^{\mathrm{norm}}$ | 0.788 | 0.586 | 0.648 | 0.275 | 0.583 | 0.252 | 0.729 | 0.471 | 0.624 | 0.218 | 0.548 | 0.010 |
| Sem. Entropy | 20 i.i.d. | 0.855 | 0.685 | 0.782 | 0.584 | 0.806 | 0.668 | 0.805 | 0.599 | 0.665 | 0.305 | 0.649 | 0.254 |
| | Full | 0.836 | 0.614 | 0.532 | 0.103 | 0.543 | 0.129 | 0.637 | 0.265 | 0.494 | 0.002 | 0.638 | 0.197 |
| | Random | 0.822 | 0.590 | 0.550 | 0.200 | 0.524 | 0.109 | 0.635 | 0.247 | 0.529 | 0.075 | 0.634 | 0.204 |
| | Last | 0.838 | 0.614 | **0.690** | **0.410** | **0.641** | 0.416 | 0.703 | 0.410 | **0.607** | **0.215** | **0.692** | **0.302** |
| | Centroid | 0.744 | 0.424 | 0.547 | 0.123 | 0.538 | 0.135 | 0.586 | 0.172 | 0.494 | 0.004 | 0.598 | 0.153 |
| | DiMS (top-$M$) | **0.845** | **0.633** | **0.690** | **0.410** | **0.641** | 0.416 | 0.702 | **0.430** | **0.607** | **0.215** | 0.691 | 0.277 |
| | DiMS (cov. 75%) | 0.842 | 0.628 | **0.690** | 0.407 | 0.639 | **0.417** | **0.709** | 0.412 | 0.604 | 0.196 | 0.687 | 0.292 |
| DegMat | 20 i.i.d. | 0.887 | 0.746 | 0.721 | 0.462 | 0.739 | 0.562 | 0.735 | 0.504 | 0.703 | 0.410 | 0.643 | 0.271 |
| | Full | 0.775 | 0.596 | 0.608 | 0.307 | 0.585 | 0.335 | 0.719 | 0.415 | **0.651** | **0.283** | 0.655 | 0.226 |
| | Random | 0.771 | 0.585 | 0.604 | 0.280 | 0.552 | 0.182 | 0.720 | 0.416 | 0.641 | 0.263 | 0.649 | 0.210 |
| | Last | 0.786 | 0.610 | **0.658** | **0.325** | 0.594 | 0.340 | 0.597 | 0.264 | 0.614 | 0.236 | 0.668 | 0.268 |
| | Centroid | 0.748 | 0.544 | 0.573 | 0.218 | 0.569 | 0.253 | 0.709 | 0.385 | 0.633 | 0.250 | 0.597 | 0.148 |
| | DiMS (top-$M$) | 0.790 | **0.615** | 0.638 | 0.324 | **0.605** | **0.369** | **0.746** | **0.472** | 0.626 | 0.245 | **0.676** | **0.282** |
| | DiMS (cov. 75%) | **0.798** | **0.615** | 0.642 | 0.324 | 0.601 | 0.357 | 0.725 | 0.429 | 0.641 | 0.281 | 0.660 | 0.227 |
| Eccentricity | 20 i.i.d. | 0.874 | 0.707 | 0.708 | 0.436 | 0.704 | 0.492 | 0.730 | 0.488 | 0.682 | 0.342 | 0.649 | 0.290 |
| | Full | 0.752 | 0.563 | 0.579 | 0.236 | 0.563 | 0.253 | 0.688 | 0.375 | **0.643** | **0.269** | 0.642 | 0.227 |
| | Random | 0.751 | 0.559 | 0.581 | 0.246 | 0.549 | 0.200 | 0.677 | 0.337 | 0.630 | 0.234 | 0.643 | 0.206 |
| | Last | 0.762 | 0.589 | 0.655 | 0.316 | **0.580** | **0.309** | 0.580 | 0.218 | 0.617 | 0.253 | 0.665 | 0.265 |
| | Centroid | 0.738 | 0.531 | 0.557 | 0.183 | 0.565 | 0.250 | 0.673 | 0.330 | 0.631 | 0.265 | 0.618 | 0.188 |
| | DiMS (top-$M$) | 0.764 | 0.592 | 0.655 | 0.316 | 0.570 | 0.273 | **0.697** | **0.405** | 0.631 | 0.264 | **0.674** | **0.282** |
| | DiMS (cov. 75%) | **0.773** | **0.594** | **0.656** | **0.317** | 0.557 | 0.226 | 0.695 | 0.396 | 0.638 | 0.268 | 0.652 | 0.212 |
| EigenVal | 20 i.i.d. | 0.852 | 0.684 | 0.681 | 0.363 | 0.643 | 0.365 | 0.707 | 0.467 | 0.696 | 0.386 | 0.620 | 0.255 |
| | Full | 0.745 | 0.559 | 0.586 | 0.249 | 0.564 | 0.254 | **0.688** | 0.369 | **0.637** | 0.262 | 0.633 | 0.211 |
| | Random | 0.742 | 0.554 | 0.577 | 0.224 | 0.547 | 0.195 | 0.660 | 0.312 | 0.631 | 0.235 | 0.627 | 0.178 |
| | Last | **0.754** | **0.590** | **0.621** | **0.310** | **0.579** | **0.301** | 0.572 | 0.200 | 0.605 | 0.233 | 0.653 | **0.250** |
| | Centroid | 0.735 | 0.534 | 0.564 | 0.208 | 0.567 | 0.249 | 0.666 | 0.325 | 0.628 | 0.254 | 0.611 | 0.172 |
| | DiMS (top-$M$) | **0.754** | 0.589 | **0.621** | **0.310** | 0.552 | 0.226 | 0.684 | **0.379** | 0.635 | **0.263** | **0.655** | 0.235 |
| | DiMS (cov. 75%) | **0.754** | 0.589 | 0.594 | 0.297 | 0.550 | 0.201 | 0.662 | 0.351 | 0.624 | 0.259 | 0.653 | 0.220 |
| KLE-Heat | 20 i.i.d. | 0.891 | 0.755 | 0.726 | 0.475 | 0.736 | 0.532 | 0.736 | 0.507 | 0.701 | 0.398 | 0.643 | 0.274 |
| | Full | 0.784 | 0.601 | 0.598 | 0.302 | 0.589 | 0.339 | 0.713 | 0.411 | **0.655** | **0.292** | 0.646 | 0.229 |
| | Random | 0.782 | 0.588 | 0.621 | 0.309 | 0.552 | 0.178 | 0.724 | 0.423 | 0.643 | 0.256 | 0.652 | 0.215 |
| | Last | 0.795 | 0.611 | **0.659** | **0.332** | 0.598 | 0.355 | 0.603 | 0.284 | 0.614 | 0.239 | 0.670 | 0.277 |
| | Centroid | 0.762 | 0.552 | 0.572 | 0.221 | 0.567 | 0.246 | 0.718 | 0.400 | 0.634 | 0.253 | 0.636 | 0.198 |
| | DiMS (top-$M$) | 0.794 | 0.604 | 0.645 | 0.331 | 0.613 | 0.391 | **0.752** | **0.478** | 0.627 | 0.243 | **0.682** | **0.293** |
| | DiMS (cov. 75%) | **0.807** | **0.615** | 0.644 | 0.330 | **0.614** | **0.394** | 0.723 | 0.429 | 0.641 | 0.275 | 0.662 | 0.238 |
| KLE-Matérn | 20 i.i.d. | 0.888 | 0.749 | 0.724 | 0.473 | 0.735 | 0.562 | 0.733 | 0.501 | 0.702 | 0.408 | 0.646 | 0.283 |
| | Full | 0.783 | 0.600 | 0.619 | 0.322 | 0.599 | 0.385 | 0.718 | 0.411 | **0.652** | **0.283** | 0.655 | 0.229 |
| | Random | 0.780 | 0.589 | 0.617 | 0.305 | 0.553 | 0.183 | 0.722 | 0.417 | 0.644 | 0.266 | 0.651 | 0.213 |
| | Last | 0.791 | 0.609 | **0.659** | **0.328** | 0.602 | 0.370 | 0.599 | 0.271 | 0.612 | 0.236 | 0.669 | 0.273 |
| | Centroid | 0.760 | 0.547 | 0.572 | 0.215 | 0.569 | 0.244 | 0.714 | 0.392 | 0.630 | 0.240 | 0.639 | 0.199 |
| | DiMS (top-$M$) | 0.796 | 0.613 | **0.659** | **0.328** | **0.625** | 0.385 | **0.751** | **0.483** | 0.626 | 0.239 | **0.679** | **0.290** |
| | DiMS (cov. 75%) | **0.803** | **0.614** | 0.657 | 0.323 | 0.619 | **0.389** | 0.725 | 0.427 | 0.641 | 0.277 | 0.663 | 0.238 |
| AD | 20 i.i.d. | 0.908 | 0.804 | 0.805 | 0.624 | 0.821 | 0.701 | 0.873 | 0.744 | 0.690 | 0.388 | 0.651 | 0.288 |
| | Full | 0.862 | 0.669 | 0.677 | 0.403 | 0.627 | 0.345 | 0.740 | 0.410 | 0.623 | 0.256 | 0.687 | 0.296 |
| | Random | 0.861 | 0.675 | 0.663 | 0.361 | 0.620 | 0.315 | 0.739 | 0.384 | 0.618 | 0.226 | 0.678 | 0.284 |
| | Last | 0.867 | 0.681 | 0.719 | **0.454** | 0.670 | 0.452 | 0.816 | 0.595 | **0.666** | **0.330** | **0.705** | **0.338** |
| | Centroid | 0.832 | 0.640 | 0.615 | 0.286 | 0.611 | 0.299 | 0.724 | 0.377 | 0.627 | 0.245 | 0.648 | 0.245 |
| | DiMS (top-$M$) | 0.868 | 0.679 | **0.720** | 0.451 | **0.706** | 0.493 | 0.829 | 0.621 | **0.666** | **0.330** | 0.701 | 0.326 |
| | DiMS (cov. 75%) | **0.871** | **0.688** | 0.719 | **0.454** | 0.692 | **0.499** | **0.834** | **0.638** | 0.660 | 0.298 | 0.699 | 0.328 |
| CoCoA | 20 i.i.d. | 0.894 | 0.756 | 0.809 | 0.616 | 0.813 | 0.673 | 0.860 | 0.709 | 0.682 | 0.367 | 0.655 | 0.298 |
| | Full | 0.878 | 0.716 | 0.762 | 0.557 | 0.754 | 0.642 | 0.765 | 0.538 | 0.649 | 0.283 | 0.694 | 0.318 |
| | Random | 0.878 | 0.712 | 0.758 | 0.548 | 0.755 | 0.645 | 0.768 | 0.541 | 0.652 | 0.293 | 0.690 | 0.310 |
| | Last | **0.881** | **0.721** | 0.766 | 0.559 | 0.756 | 0.635 | **0.798** | 0.602 | **0.662** | **0.315** | **0.706** | 0.335 |
| | Centroid | 0.872 | 0.708 | 0.752 | 0.529 | 0.755 | 0.633 | 0.755 | 0.524 | 0.656 | 0.304 | 0.676 | 0.294 |
| | DiMS (top-$M$) | 0.880 | 0.716 | **0.775** | **0.569** | **0.773** | 0.648 | 0.796 | **0.606** | **0.662** | **0.315** | 0.702 | 0.320 |
| | DiMS (cov. 75%) | 0.879 | 0.717 | 0.769 | 0.563 | 0.771 | **0.653** | 0.785 | 0.582 | **0.662** | 0.306 | 0.704 | **0.336** |

Table E.6: **Per-dataset AUROC and PRR for Dream (`rd`).** Both metrics use the same label-free $\lambda^\star$. The top block holds the information-based scores, which do not construct a semantic sample set; every block below it is one semantic estimator. Yellow denotes the **Multi-sample** reference, which uses $M{=}20$ independent generations; gray denotes standard trajectory-selection baselines, while blue denotes our DiMS strategies. **Bold** indicates the best Single-trajectory result within each estimator block, per dataset and metric; the Multi-sample reference does not compete.

| UQ estimator | Strategy | TriviaQA | | GSM8K | | MATH | | WMT14 Fr-En | | SAMSum | | HotpotQA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR |
| Info-based | Perplexity | 0.550 | 0.110 | 0.559 | 0.132 | 0.530 | 0.081 | 0.621 | 0.183 | 0.479 | −0.065 | 0.630 | 0.175 |
| | MTE | 0.557 | 0.078 | 0.644 | 0.371 | 0.604 | 0.298 | 0.675 | 0.322 | 0.499 | −0.032 | 0.663 | 0.286 |
| | MSP | 0.551 | 0.081 | 0.631 | 0.308 | 0.600 | 0.297 | 0.662 | 0.297 | 0.502 | 0.005 | 0.655 | 0.277 |
| | TrajEnt | 0.714 | 0.438 | 0.743 | 0.659 | 0.662 | 0.433 | 0.648 | 0.270 | 0.538 | 0.013 | 0.620 | 0.197 |
| | CommitNLL | 0.807 | 0.574 | 0.754 | 0.544 | 0.729 | 0.567 | 0.715 | 0.399 | 0.587 | 0.168 | 0.719 | 0.349 |
| | MCNLL$^{\text{norm}}$ | 0.723 | 0.415 | 0.657 | 0.271 | 0.659 | 0.372 | 0.678 | 0.319 | 0.598 | 0.261 | 0.570 | 0.108 |
| Sem. Entropy | 20 i.i.d. | 0.714 | 0.373 | 0.718 | 0.558 | 0.726 | 0.618 | 0.597 | 0.143 | 0.570 | 0.116 | 0.686 | 0.317 |
| | Full | 0.650 | 0.233 | 0.601 | 0.233 | 0.600 | 0.244 | **0.559** | 0.041 | 0.552 | 0.102 | 0.621 | 0.175 |
| | Random | 0.647 | **0.242** | 0.614 | 0.231 | 0.604 | 0.302 | 0.553 | 0.047 | 0.531 | 0.072 | 0.614 | 0.171 |
| | Last | 0.628 | 0.201 | 0.614 | 0.254 | 0.593 | 0.325 | **0.559** | **0.131** | **0.564** | 0.110 | 0.556 | 0.153 |
| | Centroid | 0.645 | 0.233 | 0.550 | 0.090 | **0.630** | **0.421** | 0.521 | 0.022 | 0.508 | 0.040 | 0.622 | **0.207** |
| | DiMS (top-$M$) | **0.651** | 0.237 | 0.580 | 0.210 | 0.605 | 0.248 | 0.547 | 0.046 | 0.556 | 0.134 | 0.616 | 0.191 |
| | DiMS (cov. 75%) | 0.649 | 0.232 | **0.617** | **0.269** | 0.614 | 0.317 | 0.558 | 0.064 | 0.560 | **0.138** | **0.626** | 0.186 |
| DegMat | 20 i.i.d. | 0.817 | 0.586 | 0.756 | 0.594 | 0.704 | 0.545 | 0.622 | 0.207 | 0.620 | 0.213 | 0.692 | 0.358 |
| | Full | 0.652 | 0.242 | 0.677 | 0.356 | 0.629 | **0.423** | 0.552 | 0.053 | 0.575 | 0.129 | 0.639 | 0.216 |
| | Random | 0.644 | 0.235 | 0.674 | 0.347 | 0.620 | 0.401 | 0.543 | 0.039 | 0.553 | 0.098 | 0.620 | 0.184 |
| | Last | 0.624 | 0.213 | 0.602 | 0.281 | 0.573 | 0.246 | 0.518 | 0.031 | 0.551 | 0.129 | 0.524 | 0.083 |
| | Centroid | 0.622 | 0.202 | 0.675 | 0.369 | **0.636** | **0.423** | 0.554 | 0.082 | 0.569 | 0.139 | 0.600 | 0.126 |
| | DiMS (top-$M$) | 0.663 | 0.269 | **0.693** | **0.407** | 0.610 | 0.325 | 0.570 | 0.103 | **0.583** | 0.168 | **0.641** | **0.218** |
| | DiMS (cov. 75%) | **0.668** | **0.279** | 0.686 | 0.390 | 0.621 | 0.415 | **0.571** | **0.109** | **0.583** | **0.170** | 0.634 | 0.205 |
| Eccentricity | 20 i.i.d. | 0.782 | 0.518 | 0.653 | 0.429 | 0.679 | 0.513 | 0.613 | 0.193 | 0.615 | 0.213 | 0.677 | 0.301 |
| | Full | 0.648 | 0.232 | 0.650 | 0.327 | 0.631 | 0.414 | 0.550 | 0.060 | 0.576 | 0.163 | 0.618 | 0.171 |
| | Random | 0.638 | 0.228 | 0.650 | 0.325 | 0.629 | 0.416 | 0.541 | 0.047 | 0.561 | 0.105 | 0.607 | 0.161 |
| | Last | 0.618 | 0.186 | 0.605 | 0.292 | 0.571 | 0.237 | 0.518 | 0.030 | 0.551 | 0.110 | 0.524 | 0.083 |
| | Centroid | 0.645 | 0.240 | 0.639 | 0.340 | **0.640** | **0.441** | 0.546 | 0.069 | 0.557 | 0.122 | 0.612 | 0.159 |
| | DiMS (top-$M$) | **0.653** | 0.251 | 0.670 | **0.400** | 0.617 | 0.308 | 0.555 | 0.086 | **0.583** | 0.163 | **0.626** | **0.185** |
| | DiMS (cov. 75%) | 0.651 | **0.260** | **0.671** | 0.390 | 0.637 | 0.413 | **0.562** | **0.098** | 0.581 | **0.171** | 0.622 | 0.178 |
| EigenVal | 20 i.i.d. | 0.767 | 0.498 | 0.641 | 0.352 | 0.687 | 0.485 | 0.602 | 0.172 | 0.597 | 0.196 | 0.619 | 0.216 |
| | Full | 0.643 | 0.233 | 0.650 | 0.333 | 0.631 | 0.417 | 0.552 | 0.073 | 0.582 | **0.189** | **0.599** | 0.137 |
| | Random | 0.627 | 0.205 | 0.646 | 0.295 | 0.629 | 0.443 | 0.548 | 0.067 | 0.563 | 0.135 | 0.592 | 0.130 |
| | Last | 0.594 | 0.154 | 0.535 | 0.197 | 0.549 | 0.258 | 0.523 | 0.048 | 0.523 | 0.114 | 0.508 | 0.063 |
| | Centroid | 0.647 | **0.243** | 0.633 | 0.341 | **0.642** | **0.450** | 0.548 | 0.069 | 0.569 | 0.140 | 0.597 | **0.138** |
| | DiMS (top-$M$) | 0.646 | 0.240 | 0.637 | 0.298 | 0.617 | 0.303 | **0.557** | 0.086 | 0.582 | 0.167 | 0.592 | 0.122 |
| | DiMS (cov. 75%) | **0.648** | 0.238 | **0.655** | **0.356** | 0.628 | 0.408 | **0.557** | **0.097** | **0.583** | 0.186 | 0.593 | 0.115 |
| KLE-Heat | 20 i.i.d. | 0.827 | 0.605 | 0.770 | 0.617 | 0.701 | 0.536 | 0.625 | 0.210 | 0.622 | 0.218 | 0.695 | 0.364 |
| | Full | 0.651 | 0.236 | 0.673 | 0.377 | 0.629 | 0.414 | 0.541 | 0.055 | 0.570 | 0.125 | 0.631 | 0.206 |
| | Random | 0.643 | 0.240 | 0.669 | 0.326 | 0.613 | 0.385 | 0.543 | 0.039 | 0.553 | 0.087 | 0.620 | 0.196 |
| | Last | 0.625 | 0.216 | 0.602 | 0.280 | 0.573 | 0.246 | 0.518 | 0.030 | 0.551 | 0.129 | 0.524 | 0.083 |
| | Centroid | 0.663 | 0.283 | 0.671 | 0.361 | 0.631 | 0.397 | 0.553 | 0.077 | 0.567 | 0.128 | 0.634 | 0.199 |
| | DiMS (top-$M$) | 0.665 | 0.284 | **0.697** | **0.422** | 0.611 | 0.331 | 0.571 | **0.109** | 0.583 | **0.163** | **0.646** | **0.232** |
| | DiMS (cov. 75%) | **0.673** | **0.299** | 0.695 | 0.414 | **0.635** | **0.455** | **0.573** | **0.109** | **0.584** | **0.163** | 0.636 | 0.214 |
| KLE-Matérn | 20 i.i.d. | 0.826 | 0.600 | 0.760 | 0.608 | 0.703 | 0.538 | 0.623 | 0.208 | 0.622 | 0.220 | 0.708 | 0.396 |
| | Full | 0.652 | 0.244 | 0.676 | 0.357 | 0.627 | **0.415** | 0.551 | 0.054 | 0.572 | 0.124 | 0.642 | 0.225 |
| | Random | 0.644 | 0.246 | 0.672 | 0.339 | 0.615 | 0.384 | 0.542 | 0.037 | 0.551 | 0.095 | 0.621 | 0.192 |
| | Last | 0.626 | 0.218 | 0.602 | 0.280 | 0.573 | 0.244 | 0.518 | 0.031 | 0.551 | 0.129 | 0.524 | 0.083 |
| | Centroid | 0.666 | 0.281 | 0.673 | 0.360 | **0.631** | 0.389 | 0.553 | 0.078 | 0.566 | 0.128 | 0.638 | 0.204 |
| | DiMS (top-$M$) | 0.663 | 0.280 | **0.696** | **0.418** | 0.608 | 0.328 | 0.569 | 0.103 | **0.582** | **0.164** | **0.646** | **0.235** |
| | DiMS (cov. 75%) | **0.671** | **0.296** | 0.686 | 0.389 | 0.618 | 0.407 | **0.571** | **0.107** | **0.582** | 0.163 | 0.640 | 0.217 |
| AD | 20 i.i.d. | 0.870 | 0.685 | 0.758 | 0.676 | 0.746 | 0.593 | 0.677 | 0.317 | 0.663 | 0.280 | 0.792 | 0.535 |
| | Full | 0.751 | 0.413 | **0.667** | **0.388** | 0.643 | 0.398 | 0.616 | 0.139 | 0.583 | 0.151 | 0.697 | 0.353 |
| | Random | 0.748 | 0.402 | **0.667** | 0.382 | 0.638 | 0.362 | 0.613 | 0.138 | 0.580 | 0.155 | 0.690 | 0.347 |
| | Last | 0.738 | 0.390 | 0.656 | 0.365 | 0.623 | 0.305 | 0.610 | 0.122 | **0.589** | 0.143 | 0.682 | 0.324 |
| | Centroid | 0.753 | 0.412 | 0.630 | 0.320 | **0.673** | **0.457** | 0.611 | 0.161 | 0.559 | 0.135 | 0.699 | 0.356 |
| | DiMS (top-$M$) | **0.759** | **0.430** | 0.662 | 0.369 | 0.656 | 0.456 | **0.622** | **0.167** | 0.580 | 0.145 | 0.696 | 0.354 |
| | DiMS (cov. 75%) | **0.759** | 0.425 | 0.654 | 0.368 | 0.658 | 0.433 | 0.621 | 0.162 | 0.576 | **0.156** | **0.708** | **0.379** |
| CoCoA | 20 i.i.d. | 0.856 | 0.663 | 0.826 | 0.778 | 0.808 | 0.754 | 0.729 | 0.434 | 0.636 | 0.245 | 0.780 | 0.491 |
| | Full | **0.839** | 0.625 | 0.777 | 0.648 | 0.739 | 0.640 | 0.687 | 0.314 | **0.610** | 0.206 | 0.758 | 0.442 |
| | Random | 0.838 | 0.623 | **0.778** | **0.650** | 0.734 | 0.615 | 0.683 | 0.311 | 0.608 | 0.213 | 0.754 | 0.441 |
| | Last | 0.835 | 0.619 | 0.752 | 0.600 | 0.715 | 0.574 | 0.666 | 0.270 | 0.608 | 0.186 | 0.742 | 0.443 |
| | Centroid | 0.837 | 0.626 | 0.766 | 0.640 | **0.761** | **0.681** | 0.702 | 0.375 | 0.601 | 0.211 | **0.760** | **0.447** |
| | DiMS (top-$M$) | **0.839** | **0.631** | 0.750 | 0.631 | 0.759 | 0.668 | **0.710** | **0.389** | 0.609 | 0.219 | 0.759 | **0.447** |
| | DiMS (cov. 75%) | 0.838 | **0.631** | 0.766 | 0.641 | 0.759 | 0.668 | **0.710** | **0.389** | 0.609 | **0.223** | 0.754 | 0.430 |

Table E.7: **Per-dataset AUROC and PRR for Nemotron (`1c`).** Both metrics use the same label-free $\lambda^\star$. The top block holds the information-based scores, which do not construct a semantic sample set; every block below it is one semantic estimator. Yellow denotes the **Multi-sample** reference, which uses $M{=}20$ independent generations; gray denotes standard trajectory-selection baselines, while blue denotes our DiMS strategies. **Bold** indicates the best Single-trajectory result within each estimator block, per dataset and metric; the Multi-sample reference does not compete.

| UQ estimator | Strategy | TriviaQA | | GSM8K | | MATH | | WMT14 Fr-En | | SAMSum | | HotpotQA | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| | | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR | AUROC | PRR |
| Info-based | Perplexity | 0.563 | 0.069 | 0.619 | 0.207 | 0.493 | −0.016 | 0.573 | 0.140 | 0.544 | 0.119 | 0.533 | −0.016 |
| | MTE | 0.562 | 0.065 | 0.607 | 0.181 | 0.506 | −0.012 | 0.561 | 0.107 | 0.554 | 0.137 | 0.537 | 0.008 |
| | MSP | 0.554 | 0.038 | 0.601 | 0.167 | 0.505 | 0.003 | 0.556 | 0.102 | 0.545 | 0.138 | 0.532 | −0.006 |
| | TrajEnt | 0.579 | 0.233 | 0.695 | 0.427 | 0.539 | 0.132 | 0.647 | 0.264 | 0.671 | 0.257 | 0.688 | 0.282 |
| | CommitNLL | 0.700 | 0.483 | 0.651 | 0.325 | 0.516 | 0.025 | 0.681 | 0.324 | 0.591 | 0.163 | 0.519 | −0.016 |
| | $\mathrm{MCNLL}^{\mathrm{norm}}$ | 0.600 | 0.233 | 0.578 | 0.187 | 0.568 | 0.129 | 0.669 | 0.281 | 0.673 | 0.334 | 0.430 | −0.169 |
| Sem. Entropy | 20 i.i.d. | 0.694 | 0.389 | 0.622 | 0.278 | 0.535 | 0.048 | 0.697 | 0.387 | 0.679 | 0.284 | 0.695 | 0.356 |
| | Full | 0.616 | 0.251 | 0.560 | 0.091 | 0.536 | 0.069 | 0.562 | 0.077 | 0.453 | −0.064 | 0.571 | 0.075 |
| | Random | **0.621** | **0.255** | 0.544 | 0.087 | 0.513 | 0.002 | 0.573 | 0.086 | 0.474 | −0.045 | 0.574 | 0.073 |
| | Last | 0.613 | 0.248 | 0.484 | −0.070 | 0.505 | 0.051 | **0.637** | **0.228** | 0.613 | 0.233 | **0.596** | **0.187** |
| | Centroid | 0.596 | 0.177 | 0.561 | **0.106** | 0.477 | −0.108 | 0.509 | 0.024 | 0.425 | −0.100 | 0.508 | 0.027 |
| | DiMS (top-$M$) | 0.614 | 0.243 | **0.562** | 0.105 | **0.540** | 0.069 | 0.615 | 0.209 | **0.620** | **0.235** | 0.594 | 0.173 |
| | DiMS (cov. 75%) | 0.606 | 0.237 | 0.553 | 0.068 | 0.536 | **0.087** | 0.607 | 0.176 | 0.611 | 0.206 | **0.596** | 0.177 |
| DegMat | 20 i.i.d. | 0.815 | 0.586 | 0.630 | 0.234 | 0.516 | 0.068 | 0.626 | 0.206 | 0.682 | 0.357 | 0.617 | 0.229 |
| | Full | 0.696 | 0.390 | 0.530 | 0.102 | 0.537 | 0.136 | **0.612** | 0.175 | 0.594 | 0.177 | **0.623** | 0.170 |
| | Random | 0.690 | 0.372 | 0.542 | 0.102 | 0.537 | 0.104 | 0.608 | **0.178** | 0.582 | 0.148 | 0.618 | 0.159 |
| | Last | 0.642 | 0.321 | **0.598** | **0.201** | **0.550** | **0.190** | 0.555 | 0.095 | 0.578 | 0.129 | 0.597 | 0.152 |
| | Centroid | 0.691 | 0.382 | 0.562 | 0.125 | 0.514 | −0.003 | 0.588 | 0.150 | 0.590 | 0.158 | 0.613 | 0.173 |
| | DiMS (top-$M$) | **0.715** | **0.431** | 0.550 | 0.148 | 0.535 | 0.153 | 0.602 | 0.177 | **0.639** | **0.235** | 0.617 | **0.175** |
| | DiMS (cov. 75%) | 0.710 | 0.418 | 0.557 | 0.172 | 0.541 | 0.170 | 0.589 | 0.131 | 0.626 | 0.212 | 0.585 | 0.164 |
| Eccentricity | 20 i.i.d. | 0.801 | 0.558 | 0.548 | −0.004 | 0.511 | 0.066 | 0.623 | 0.200 | 0.635 | 0.277 | 0.640 | 0.245 |
| | Full | 0.713 | 0.396 | 0.524 | 0.064 | 0.552 | 0.190 | 0.615 | 0.189 | 0.538 | 0.126 | 0.560 | 0.056 |
| | Random | 0.696 | 0.389 | 0.523 | 0.082 | 0.539 | 0.068 | **0.618** | **0.209** | 0.524 | 0.094 | 0.571 | 0.054 |
| | Last | 0.641 | 0.321 | **0.581** | **0.140** | 0.550 | 0.197 | 0.554 | 0.088 | 0.572 | 0.115 | **0.595** | **0.143** |
| | Centroid | 0.701 | 0.407 | 0.550 | 0.105 | 0.532 | 0.100 | 0.593 | 0.158 | 0.534 | 0.088 | 0.566 | 0.085 |
| | DiMS (top-$M$) | **0.718** | **0.423** | 0.520 | 0.055 | **0.555** | **0.280** | 0.583 | 0.148 | **0.604** | **0.204** | 0.563 | 0.097 |
| | DiMS (cov. 75%) | **0.718** | 0.419 | 0.522 | 0.057 | 0.550 | 0.182 | 0.587 | 0.131 | 0.596 | 0.198 | 0.560 | 0.063 |
| EigenVal | 20 i.i.d. | 0.779 | 0.485 | 0.544 | 0.019 | 0.513 | −0.015 | 0.589 | 0.189 | 0.586 | 0.164 | 0.532 | 0.082 |
| | Full | **0.712** | **0.402** | 0.511 | 0.035 | 0.551 | 0.181 | **0.615** | 0.178 | **0.565** | **0.165** | 0.542 | 0.050 |
| | Random | 0.691 | 0.385 | 0.521 | 0.048 | **0.566** | 0.183 | 0.602 | 0.161 | 0.541 | 0.111 | 0.555 | 0.021 |
| | Last | 0.638 | 0.303 | 0.524 | **0.087** | 0.507 | 0.115 | 0.545 | 0.089 | 0.559 | 0.119 | 0.530 | 0.090 |
| | Centroid | 0.699 | 0.392 | **0.546** | 0.082 | 0.534 | 0.110 | 0.582 | 0.138 | 0.556 | 0.121 | 0.555 | 0.080 |
| | DiMS (top-$M$) | 0.708 | 0.396 | 0.509 | 0.035 | 0.556 | 0.187 | 0.598 | 0.150 | 0.559 | 0.119 | 0.551 | 0.087 |
| | DiMS (cov. 75%) | 0.709 | 0.397 | 0.536 | 0.067 | 0.552 | **0.197** | 0.613 | **0.179** | 0.546 | 0.131 | **0.557** | **0.100** |
| KLE-Heat | 20 i.i.d. | 0.817 | 0.591 | 0.638 | 0.254 | 0.517 | 0.071 | 0.627 | 0.212 | 0.693 | 0.375 | 0.627 | 0.250 |
| | Full | 0.690 | **0.384** | 0.525 | 0.060 | 0.523 | 0.008 | 0.603 | 0.169 | 0.595 | 0.175 | 0.610 | 0.152 |
| | Random | 0.682 | 0.350 | 0.546 | 0.111 | 0.537 | 0.108 | **0.608** | **0.180** | 0.584 | 0.144 | 0.621 | 0.167 |
| | Last | 0.641 | 0.317 | **0.599** | **0.201** | **0.550** | **0.192** | 0.555 | 0.094 | 0.578 | 0.133 | 0.598 | 0.152 |
| | Centroid | 0.683 | 0.371 | 0.564 | 0.131 | 0.523 | 0.035 | 0.586 | 0.148 | 0.593 | 0.168 | **0.632** | **0.213** |
| | DiMS (top-$M$) | **0.699** | **0.384** | 0.562 | 0.182 | 0.538 | 0.172 | 0.603 | 0.173 | **0.639** | **0.221** | 0.622 | 0.162 |
| | DiMS (cov. 75%) | 0.693 | 0.368 | 0.553 | 0.162 | 0.536 | 0.127 | 0.601 | 0.170 | 0.626 | 0.202 | 0.591 | 0.151 |
| KLE-Matérn | 20 i.i.d. | 0.817 | 0.592 | 0.628 | 0.228 | 0.516 | 0.071 | 0.628 | 0.216 | 0.688 | 0.369 | 0.638 | 0.265 |
| | Full | 0.693 | 0.384 | 0.527 | 0.094 | 0.535 | 0.128 | **0.612** | 0.176 | 0.595 | 0.178 | 0.624 | 0.173 |
| | Random | 0.686 | 0.361 | 0.541 | 0.095 | 0.538 | 0.106 | 0.609 | **0.178** | 0.582 | 0.144 | 0.621 | 0.165 |
| | Last | 0.643 | 0.323 | **0.598** | **0.199** | **0.550** | **0.191** | 0.555 | 0.094 | 0.577 | 0.131 | 0.597 | 0.151 |
| | Centroid | 0.690 | 0.375 | 0.560 | 0.115 | 0.510 | −0.017 | 0.592 | 0.155 | 0.592 | 0.165 | **0.631** | **0.207** |
| | DiMS (top-$M$) | **0.708** | **0.412** | 0.548 | 0.133 | 0.529 | 0.106 | 0.602 | 0.177 | **0.653** | **0.255** | 0.630 | 0.182 |
| | DiMS (cov. 75%) | 0.702 | 0.394 | 0.552 | 0.152 | 0.532 | 0.113 | 0.590 | 0.132 | 0.630 | 0.217 | 0.629 | 0.179 |
| AD | 20 i.i.d. | 0.736 | 0.481 | 0.694 | 0.401 | 0.499 | −0.066 | 0.718 | 0.412 | 0.702 | 0.338 | 0.759 | 0.480 |
| | Full | 0.660 | 0.328 | 0.484 | −0.037 | 0.460 | −0.256 | 0.638 | 0.192 | 0.545 | 0.125 | 0.599 | 0.154 |
| | Random | 0.663 | 0.328 | 0.480 | −0.046 | 0.456 | −0.288 | 0.638 | 0.218 | 0.547 | 0.128 | 0.596 | 0.140 |
| | Last | 0.680 | 0.380 | 0.591 | 0.159 | **0.480** | **−0.102** | 0.662 | 0.264 | 0.641 | 0.267 | **0.663** | **0.242** |
| | Centroid | 0.629 | 0.253 | 0.482 | −0.024 | 0.438 | −0.313 | 0.617 | 0.184 | 0.537 | 0.113 | 0.564 | 0.137 |
| | DiMS (top-$M$) | 0.684 | 0.395 | 0.592 | 0.166 | 0.470 | −0.132 | 0.659 | 0.261 | 0.636 | 0.253 | 0.655 | 0.228 |
| | DiMS (cov. 75%) | **0.687** | **0.402** | **0.607** | **0.183** | 0.462 | −0.210 | **0.670** | **0.283** | **0.645** | **0.269** | 0.652 | 0.223 |
| CoCoA | 20 i.i.d. | 0.746 | 0.546 | 0.715 | 0.431 | 0.506 | −0.039 | 0.729 | 0.434 | 0.678 | 0.308 | 0.663 | 0.286 |
| | Full | 0.721 | 0.517 | 0.571 | 0.172 | 0.477 | −0.185 | 0.701 | 0.331 | 0.602 | 0.193 | 0.613 | 0.161 |
| | Random | 0.722 | 0.518 | 0.567 | 0.158 | 0.474 | −0.209 | 0.700 | 0.336 | 0.601 | 0.186 | 0.610 | 0.148 |
| | Last | **0.729** | **0.527** | 0.630 | 0.252 | **0.483** | **−0.099** | 0.694 | 0.324 | **0.652** | **0.264** | **0.651** | **0.224** |
| | Centroid | 0.712 | 0.497 | 0.578 | 0.179 | 0.466 | −0.209 | 0.691 | 0.331 | 0.595 | 0.176 | 0.547 | 0.076 |
| | DiMS (top-$M$) | 0.709 | 0.494 | **0.641** | **0.288** | 0.470 | −0.197 | **0.710** | **0.367** | 0.641 | 0.233 | 0.625 | 0.188 |
| | DiMS (cov. 75%) | 0.711 | 0.496 | 0.629 | 0.256 | 0.475 | −0.187 | 0.694 | 0.320 | 0.641 | 0.239 | 0.622 | 0.187 |

Table E.8: **AUROC per task.** Each cell is the mean over the eight semantic estimators and the five model variants (and, for Math, over GSM8K and MATH); the *Average* row weights every run equally and matches Table 1. Yellow denotes the **Multi-sample** reference, which uses $M=20$ independent generations; gray denotes standard trajectory-selection baselines, while blue denotes our DiMS strategies. **Bold** indicates the best Single-trajectory result per task; the Multi-sample reference does not compete.

| | | | **Multi-sample** | **Baselines** | | | | **DiMS** | |
|---|---|---|---|---|---|---|---|---|---|
| **Task** | **Datasets** | **Runs** | 20 i.i.d. | Full | Random | Last | Centroid | top-$M$ | cov. 75% |
| QA | TriviaQA | 5 | 0.815 | 0.733 | 0.727 | 0.712 | 0.721 | 0.737 | **0.739** |
| Multi-hop QA | HotpotQA | 5 | 0.673 | 0.633 | 0.630 | 0.606 | 0.618 | **0.644** | 0.641 |
| Math | GSM8K, MATH | 10 | 0.713 | 0.631 | 0.624 | 0.605 | 0.628 | 0.641 | **0.647** |
| Translation | WMT14 Fr-En | 5 | 0.687 | 0.640 | 0.632 | 0.606 | 0.623 | **0.648** | 0.645 |
| Summarization | SAMSum | 5 | 0.651 | 0.587 | 0.580 | 0.588 | 0.577 | 0.605 | **0.608** |
| **Average** | – | 30 | 0.709 | 0.643 | 0.636 | 0.620 | 0.632 | 0.653 | **0.655** |

Table E.9: **Step-selection ablation.** Test AUROC of the six rules that turn the learned weights $\mathbf{w}^\star$ into a step set. The top-$M$ rule and the two draws, with (WR) and without (WOR) replacement, spend the budget $M{=}20$; the coverage rules keep the smallest set carrying 50, 75 or 90% of the weight, so their cardinality depends on $\mathbf{w}^\star$. Each rule uses its own label-free $\lambda^\star$. Each model column is the mean over the six datasets and *Average* the mean over the five model variants; intervals correspond to 1.96 standard errors (per-run bootstrap SEs combined in quadrature). Blue denotes the two rules reported in the main tables. **Bold** indicates the best rule within each estimator block.

| UQ estimator | Rule | AUROC ↑ | | | | | |
|---|---|---|---|---|---|---|---|
| | | LLaDA-1.5 (1c) | LLaDA-1.5 (rd) | Dream (1c) | Dream (rd) | Nemotron (1c) | Average |
| Sem. Entropy | top-$M$ | 0.613 ± 0.018 | 0.624 ± 0.018 | 0.696 ± 0.017 | 0.592 ± 0.020 | **0.591** ± 0.018 | 0.623 ± 0.008 |
| | WR | 0.606 ± 0.017 | 0.616 ± 0.018 | 0.675 ± 0.017 | 0.589 ± 0.019 | 0.581 ± 0.018 | 0.613 ± 0.008 |
| | WOR | 0.610 ± 0.018 | 0.621 ± 0.017 | 0.696 ± 0.017 | 0.598 ± 0.019 | 0.584 ± 0.018 | 0.622 ± 0.008 |
| | cov. 50% | 0.607 ± 0.018 | 0.629 ± 0.017 | 0.687 ± 0.017 | 0.599 ± 0.019 | 0.585 ± 0.019 | 0.621 ± 0.008 |
| | cov. 75% | **0.625** ± 0.018 | 0.629 ± 0.017 | 0.695 ± 0.017 | **0.604** ± 0.019 | 0.585 ± 0.018 | 0.628 ± 0.008 |
| | cov. 90% | 0.623 ± 0.017 | **0.633** ± 0.017 | **0.697** ± 0.017 | **0.604** ± 0.019 | 0.588 ± 0.018 | **0.629** ± 0.008 |
| DegMat | top-$M$ | 0.634 ± 0.018 | **0.665** ± 0.017 | 0.680 ± 0.017 | 0.627 ± 0.018 | **0.610** ± 0.017 | **0.643** ± 0.008 |
| | WR | 0.637 ± 0.017 | 0.656 ± 0.017 | 0.663 ± 0.017 | 0.623 ± 0.018 | 0.593 ± 0.018 | 0.634 ± 0.008 |
| | WOR | 0.642 ± 0.017 | 0.664 ± 0.017 | 0.676 ± 0.017 | 0.626 ± 0.018 | 0.603 ± 0.017 | 0.642 ± 0.008 |
| | cov. 50% | 0.635 ± 0.017 | 0.660 ± 0.017 | **0.687** ± 0.017 | 0.626 ± 0.018 | 0.597 ± 0.017 | 0.641 ± 0.008 |
| | cov. 75% | 0.644 ± 0.017 | 0.663 ± 0.017 | 0.678 ± 0.017 | 0.627 ± 0.018 | 0.601 ± 0.018 | **0.643** ± 0.008 |
| | cov. 90% | **0.647** ± 0.017 | 0.662 ± 0.017 | 0.678 ± 0.017 | **0.628** ± 0.018 | 0.601 ± 0.018 | **0.643** ± 0.008 |
| Eccentricity | top-$M$ | 0.612 ± 0.018 | 0.649 ± 0.017 | 0.665 ± 0.017 | 0.617 ± 0.018 | 0.591 ± 0.017 | 0.627 ± 0.008 |
| | WR | 0.604 ± 0.018 | 0.644 ± 0.018 | 0.650 ± 0.018 | 0.616 ± 0.018 | 0.578 ± 0.018 | 0.618 ± 0.008 |
| | WOR | 0.612 ± 0.018 | 0.649 ± 0.017 | 0.654 ± 0.017 | 0.616 ± 0.018 | 0.584 ± 0.018 | 0.623 ± 0.008 |
| | cov. 50% | 0.605 ± 0.018 | 0.647 ± 0.018 | **0.667** ± 0.017 | 0.618 ± 0.018 | 0.592 ± 0.018 | 0.626 ± 0.008 |
| | cov. 75% | 0.615 ± 0.018 | **0.654** ± 0.017 | 0.662 ± 0.017 | **0.621** ± 0.018 | 0.589 ± 0.018 | 0.628 ± 0.008 |
| | cov. 90% | **0.621** ± 0.018 | 0.652 ± 0.017 | 0.659 ± 0.017 | 0.616 ± 0.018 | **0.596** ± 0.018 | **0.629** ± 0.008 |
| EigenVal | top-$M$ | 0.600 ± 0.018 | 0.642 ± 0.018 | 0.650 ± 0.018 | 0.605 ± 0.018 | 0.580 ± 0.018 | 0.616 ± 0.008 |
| | WR | 0.618 ± 0.018 | 0.643 ± 0.018 | 0.637 ± 0.018 | 0.611 ± 0.017 | 0.575 ± 0.018 | 0.617 ± 0.008 |
| | WOR | 0.608 ± 0.018 | 0.643 ± 0.018 | 0.642 ± 0.018 | 0.607 ± 0.018 | 0.585 ± 0.018 | 0.617 ± 0.008 |
| | cov. 50% | 0.579 ± 0.018 | 0.642 ± 0.018 | **0.654** ± 0.018 | 0.611 ± 0.018 | 0.561 ± 0.017 | 0.609 ± 0.008 |
| | cov. 75% | 0.612 ± 0.018 | **0.652** ± 0.017 | 0.640 ± 0.018 | 0.611 ± 0.018 | 0.585 ± 0.018 | 0.620 ± 0.008 |
| | cov. 90% | **0.620** ± 0.018 | 0.650 ± 0.017 | 0.645 ± 0.018 | **0.614** ± 0.018 | **0.587** ± 0.018 | **0.623** ± 0.008 |
| KLE-Heat | top-$M$ | 0.639 ± 0.018 | **0.664** ± 0.017 | 0.686 ± 0.017 | 0.629 ± 0.018 | **0.611** ± 0.017 | **0.646** ± 0.008 |
| | WR | 0.625 ± 0.018 | 0.652 ± 0.017 | 0.663 ± 0.017 | 0.621 ± 0.018 | 0.597 ± 0.018 | 0.632 ± 0.008 |
| | WOR | 0.641 ± 0.017 | 0.662 ± 0.017 | 0.679 ± 0.017 | 0.629 ± 0.018 | 0.603 ± 0.017 | 0.643 ± 0.008 |
| | cov. 50% | 0.640 ± 0.017 | 0.660 ± 0.017 | **0.692** ± 0.017 | 0.628 ± 0.018 | 0.597 ± 0.017 | 0.643 ± 0.008 |
| | cov. 75% | 0.643 ± 0.017 | 0.661 ± 0.017 | 0.682 ± 0.017 | **0.633** ± 0.018 | 0.600 ± 0.018 | 0.644 ± 0.008 |
| | cov. 90% | **0.645** ± 0.017 | 0.658 ± 0.017 | 0.683 ± 0.017 | 0.628 ± 0.018 | 0.600 ± 0.018 | 0.643 ± 0.008 |
| KLE-Matérn | top-$M$ | 0.627 ± 0.018 | **0.663** ± 0.017 | **0.689** ± 0.017 | 0.627 ± 0.018 | **0.612** ± 0.017 | 0.644 ± 0.008 |
| | WR | 0.625 ± 0.018 | 0.651 ± 0.017 | 0.667 ± 0.017 | 0.621 ± 0.018 | 0.595 ± 0.018 | 0.632 ± 0.008 |
| | WOR | 0.641 ± 0.017 | **0.663** ± 0.017 | 0.682 ± 0.017 | 0.627 ± 0.018 | 0.605 ± 0.017 | 0.644 ± 0.008 |
| | cov. 50% | 0.643 ± 0.017 | 0.659 ± 0.017 | **0.689** ± 0.017 | 0.626 ± 0.018 | 0.594 ± 0.017 | 0.642 ± 0.008 |
| | cov. 75% | 0.645 ± 0.017 | **0.663** ± 0.017 | 0.685 ± 0.017 | **0.628** ± 0.018 | 0.606 ± 0.018 | **0.645** ± 0.008 |
| | cov. 90% | **0.650** ± 0.017 | 0.661 ± 0.017 | 0.680 ± 0.017 | 0.627 ± 0.018 | 0.603 ± 0.018 | **0.645** ± 0.008 |
| AD | top-$M$ | 0.719 ± 0.017 | 0.722 ± 0.017 | 0.748 ± 0.016 | **0.662** ± 0.018 | 0.616 ± 0.018 | 0.694 ± 0.008 |
| | WR | 0.721 ± 0.017 | 0.723 ± 0.017 | 0.737 ± 0.016 | 0.661 ± 0.018 | 0.614 ± 0.018 | 0.691 ± 0.008 |
| | WOR | 0.726 ± 0.016 | 0.723 ± 0.017 | 0.746 ± 0.016 | 0.661 ± 0.018 | 0.617 ± 0.018 | 0.695 ± 0.008 |
| | cov. 50% | **0.729** ± 0.016 | **0.726** ± 0.017 | **0.750** ± 0.016 | 0.659 ± 0.018 | **0.627** ± 0.018 | **0.698** ± 0.008 |
| | cov. 75% | 0.727 ± 0.016 | **0.726** ± 0.017 | 0.746 ± 0.016 | **0.662** ± 0.018 | 0.620 ± 0.018 | 0.696 ± 0.008 |
| | cov. 90% | 0.725 ± 0.017 | **0.726** ± 0.017 | 0.739 ± 0.016 | 0.661 ± 0.018 | 0.615 ± 0.018 | 0.693 ± 0.008 |
| CoCoA | top-$M$ | 0.760 ± 0.015 | 0.764 ± 0.015 | 0.764 ± 0.015 | 0.737 ± 0.016 | **0.633** ± 0.018 | 0.732 ± 0.007 |
| | WR | 0.759 ± 0.015 | 0.766 ± 0.015 | 0.762 ± 0.016 | 0.738 ± 0.016 | 0.627 ± 0.018 | 0.730 ± 0.007 |
| | WOR | **0.763** ± 0.015 | 0.765 ± 0.015 | 0.764 ± 0.015 | **0.739** ± 0.016 | 0.626 ± 0.018 | 0.731 ± 0.007 |
| | cov. 50% | **0.763** ± 0.015 | 0.766 ± 0.015 | **0.769** ± 0.015 | **0.739** ± 0.016 | 0.632 ± 0.018 | **0.734** ± 0.007 |
| | cov. 75% | **0.763** ± 0.015 | 0.768 ± 0.015 | 0.762 ± 0.016 | **0.739** ± 0.016 | 0.629 ± 0.018 | 0.732 ± 0.007 |
| | cov. 90% | **0.763** ± 0.015 | **0.769** ± 0.015 | 0.762 ± 0.015 | **0.739** ± 0.016 | 0.629 ± 0.018 | 0.732 ± 0.007 |

Table E.10: **Redundancy-penalty ablation.** Test AUROC at the smallest grid value $\lambda$=0 (i.e. $\lambda_{\text{eff}}$=$1/R$), at a fixed $\lambda$=1, and at the label-free $\lambda^\star$. Step selection is held at DiMS (top-$M$). Each model column is the mean over the six datasets and *Average* the mean over the five model variants; intervals correspond to 1.96 standard errors (per-run bootstrap SEs combined in quadrature). Blue denotes the default configuration, used in every other table: `E5-large-v2` embeddings, an RBF kernel and the label-free $\lambda^\star$. **Bold** indicates the best $\lambda$ within each estimator block.

| UQ estimator | $\lambda$ | AUROC ↑ | | | | | |
|---|---|---|---|---|---|---|---|
| | | LLaDA-1.5 (`lc`) | LLaDA-1.5 (`rd`) | Dream (`lc`) | Dream (`rd`) | Nemotron (`lc`) | Average |
| **Sem. Entropy** | $\lambda$=0 | 0.610 ± 0.018 | 0.617 ± 0.018 | 0.695 ± 0.017 | **0.596** ± 0.019 | 0.573 ± 0.018 | 0.618 ± 0.008 |
| | $\lambda$=1 | 0.599 ± 0.018 | 0.567 ± 0.018 | 0.667 ± 0.018 | 0.560 ± 0.018 | 0.567 ± 0.019 | 0.592 ± 0.008 |
| | $\lambda^\star$ | **0.613** ± 0.018 | **0.624** ± 0.018 | **0.696** ± 0.017 | 0.592 ± 0.020 | **0.591** ± 0.018 | **0.623** ± 0.008 |
| **DegMat** | $\lambda$=0 | 0.604 ± 0.018 | 0.634 ± 0.017 | 0.653 ± 0.017 | 0.607 ± 0.018 | 0.583 ± 0.017 | 0.616 ± 0.008 |
| | $\lambda$=1 | 0.609 ± 0.018 | 0.662 ± 0.017 | 0.649 ± 0.018 | 0.625 ± 0.018 | 0.597 ± 0.017 | 0.628 ± 0.008 |
| | $\lambda^\star$ | **0.634** ± 0.018 | **0.665** ± 0.017 | **0.680** ± 0.017 | **0.627** ± 0.018 | **0.610** ± 0.017 | **0.643** ± 0.008 |
| **Eccentricity** | $\lambda$=0 | 0.595 ± 0.018 | 0.638 ± 0.017 | 0.643 ± 0.018 | 0.599 ± 0.018 | 0.578 ± 0.017 | 0.611 ± 0.008 |
| | $\lambda$=1 | 0.588 ± 0.018 | 0.648 ± 0.017 | 0.633 ± 0.018 | 0.614 ± 0.018 | 0.582 ± 0.017 | 0.613 ± 0.008 |
| | $\lambda^\star$ | **0.612** ± 0.018 | **0.649** ± 0.017 | **0.665** ± 0.017 | **0.617** ± 0.018 | **0.591** ± 0.017 | **0.627** ± 0.008 |
| **EigenVal** | $\lambda$=0 | 0.564 ± 0.019 | 0.614 ± 0.018 | 0.631 ± 0.018 | 0.590 ± 0.018 | 0.552 ± 0.018 | 0.590 ± 0.008 |
| | $\lambda$=1 | 0.568 ± 0.018 | **0.644** ± 0.018 | 0.628 ± 0.018 | **0.609** ± 0.018 | **0.583** ± 0.017 | 0.606 ± 0.008 |
| | $\lambda^\star$ | **0.600** ± 0.018 | 0.642 ± 0.018 | **0.650** ± 0.018 | 0.605 ± 0.018 | 0.580 ± 0.018 | **0.616** ± 0.008 |
| **KLE-Heat** | $\lambda$=0 | 0.605 ± 0.018 | 0.633 ± 0.017 | 0.657 ± 0.017 | 0.608 ± 0.018 | 0.583 ± 0.017 | 0.617 ± 0.008 |
| | $\lambda$=1 | 0.612 ± 0.018 | 0.662 ± 0.017 | 0.655 ± 0.018 | 0.626 ± 0.018 | 0.597 ± 0.017 | 0.630 ± 0.008 |
| | $\lambda^\star$ | **0.639** ± 0.018 | **0.664** ± 0.017 | **0.686** ± 0.017 | **0.629** ± 0.018 | **0.611** ± 0.017 | **0.646** ± 0.008 |
| **KLE-Matérn** | $\lambda$=0 | 0.604 ± 0.018 | 0.634 ± 0.017 | 0.655 ± 0.017 | 0.607 ± 0.018 | 0.583 ± 0.017 | 0.617 ± 0.008 |
| | $\lambda$=1 | 0.611 ± 0.018 | 0.661 ± 0.017 | 0.654 ± 0.018 | 0.625 ± 0.018 | 0.596 ± 0.018 | 0.630 ± 0.008 |
| | $\lambda^\star$ | **0.627** ± 0.018 | **0.663** ± 0.017 | **0.689** ± 0.017 | **0.627** ± 0.018 | **0.612** ± 0.017 | **0.644** ± 0.008 |
| **AD** | $\lambda$=0 | **0.721** ± 0.017 | 0.711 ± 0.017 | 0.741 ± 0.016 | 0.656 ± 0.018 | **0.620** ± 0.018 | 0.690 ± 0.008 |
| | $\lambda$=1 | **0.721** ± 0.017 | 0.707 ± 0.017 | 0.742 ± 0.016 | 0.653 ± 0.018 | 0.612 ± 0.018 | 0.687 ± 0.008 |
| | $\lambda^\star$ | 0.719 ± 0.017 | **0.722** ± 0.017 | **0.748** ± 0.016 | **0.662** ± 0.018 | 0.616 ± 0.018 | **0.694** ± 0.008 |
| **CoCoA** | $\lambda$=0 | 0.748 ± 0.016 | 0.756 ± 0.015 | 0.762 ± 0.015 | 0.728 ± 0.016 | **0.640** ± 0.017 | 0.727 ± 0.007 |
| | $\lambda$=1 | 0.750 ± 0.016 | 0.761 ± 0.015 | 0.763 ± 0.015 | 0.736 ± 0.016 | 0.634 ± 0.017 | 0.729 ± 0.007 |
| | $\lambda^\star$ | **0.760** ± 0.015 | **0.764** ± 0.015 | **0.764** ± 0.015 | **0.737** ± 0.016 | 0.633 ± 0.018 | **0.732** ± 0.007 |

Table E.11: **MMD-kernel ablation.** Test AUROC with an RBF, Laplacian or Matérn ($\nu$=3/2) kernel in the offline calibration. The sentence encoder is fixed to `E5-large-v2` and step selection to DiMS (top-$M$). Each model column is the mean over the six datasets and *Average* the mean over the five model variants; intervals correspond to 1.96 standard errors (per-run bootstrap SEs combined in quadrature). Blue denotes the default configuration, used in every other table: `E5-large-v2` embeddings, an RBF kernel and the label-free $\lambda^\star$. **Bold** indicates the best kernel within each estimator block.

| UQ estimator | Kernel | AUROC ↑ | | | | | |
|---|---|---|---|---|---|---|---|
| | | LLaDA-1.5 (`lc`) | LLaDA-1.5 (`rd`) | Dream (`lc`) | Dream (`rd`) | Nemotron (`lc`) | Average |
| Sem. Entropy | RBF (default) | **0.613** ± 0.018 | **0.624** ± 0.018 | **0.696** ± 0.017 | 0.592 ± 0.020 | **0.591** ± 0.018 | **0.623** ± 0.008 |
| | Laplacian | 0.612 ± 0.018 | 0.609 ± 0.017 | **0.696** ± 0.017 | **0.604** ± 0.019 | 0.586 ± 0.018 | 0.621 ± 0.008 |
| | Matérn ($\nu$=3/2) | 0.598 ± 0.018 | 0.614 ± 0.018 | 0.695 ± 0.017 | 0.599 ± 0.019 | **0.591** ± 0.018 | 0.619 ± 0.008 |
| DegMat | RBF (default) | 0.634 ± 0.018 | **0.665** ± 0.017 | **0.680** ± 0.017 | 0.627 ± 0.018 | **0.610** ± 0.017 | **0.643** ± 0.008 |
| | Laplacian | 0.633 ± 0.018 | 0.663 ± 0.017 | 0.676 ± 0.017 | 0.624 ± 0.018 | 0.605 ± 0.017 | 0.640 ± 0.008 |
| | Matérn ($\nu$=3/2) | **0.641** ± 0.018 | 0.662 ± 0.017 | 0.677 ± 0.017 | **0.628** ± 0.018 | 0.602 ± 0.017 | 0.642 ± 0.008 |
| Eccentricity | RBF (default) | **0.612** ± 0.018 | 0.649 ± 0.017 | **0.665** ± 0.017 | 0.617 ± 0.018 | 0.591 ± 0.017 | **0.627** ± 0.008 |
| | Laplacian | 0.594 ± 0.018 | 0.649 ± 0.017 | 0.661 ± 0.017 | **0.618** ± 0.018 | **0.592** ± 0.018 | 0.623 ± 0.008 |
| | Matérn ($\nu$=3/2) | 0.598 ± 0.018 | **0.650** ± 0.017 | 0.661 ± 0.017 | **0.618** ± 0.018 | 0.585 ± 0.018 | 0.622 ± 0.008 |
| EigenVal | RBF (default) | 0.600 ± 0.018 | 0.642 ± 0.018 | 0.650 ± 0.018 | 0.605 ± 0.018 | 0.580 ± 0.018 | 0.616 ± 0.008 |
| | Laplacian | **0.608** ± 0.018 | **0.647** ± 0.017 | 0.650 ± 0.017 | **0.615** ± 0.018 | 0.585 ± 0.018 | **0.621** ± 0.008 |
| | Matérn ($\nu$=3/2) | 0.598 ± 0.018 | **0.647** ± 0.017 | **0.652** ± 0.017 | 0.611 ± 0.018 | **0.587** ± 0.018 | 0.619 ± 0.008 |
| KLE-Heat | RBF (default) | 0.639 ± 0.018 | **0.664** ± 0.017 | **0.686** ± 0.017 | 0.629 ± 0.018 | **0.611** ± 0.017 | **0.646** ± 0.008 |
| | Laplacian | **0.647** ± 0.017 | 0.663 ± 0.017 | 0.681 ± 0.017 | **0.632** ± 0.018 | 0.608 ± 0.017 | **0.646** ± 0.008 |
| | Matérn ($\nu$=3/2) | 0.638 ± 0.018 | 0.660 ± 0.017 | 0.681 ± 0.017 | 0.630 ± 0.018 | 0.610 ± 0.017 | 0.644 ± 0.008 |
| KLE-Matérn | RBF (default) | **0.627** ± 0.018 | **0.663** ± 0.017 | **0.689** ± 0.017 | 0.627 ± 0.018 | **0.612** ± 0.017 | **0.644** ± 0.008 |
| | Laplacian | 0.626 ± 0.018 | 0.661 ± 0.017 | 0.684 ± 0.017 | **0.630** ± 0.018 | 0.606 ± 0.017 | 0.641 ± 0.008 |
| | Matérn ($\nu$=3/2) | 0.624 ± 0.018 | 0.661 ± 0.017 | 0.683 ± 0.017 | 0.628 ± 0.018 | 0.608 ± 0.017 | 0.641 ± 0.008 |
| AD | RBF (default) | **0.719** ± 0.017 | 0.722 ± 0.017 | **0.748** ± 0.016 | 0.662 ± 0.018 | 0.616 ± 0.018 | **0.694** ± 0.008 |
| | Laplacian | **0.719** ± 0.017 | 0.721 ± 0.017 | 0.746 ± 0.016 | 0.664 ± 0.018 | **0.618** ± 0.018 | **0.694** ± 0.008 |
| | Matérn ($\nu$=3/2) | **0.719** ± 0.017 | **0.723** ± 0.017 | 0.747 ± 0.016 | **0.665** ± 0.018 | 0.617 ± 0.018 | **0.694** ± 0.008 |
| CoCoA | RBF (default) | **0.760** ± 0.015 | 0.764 ± 0.015 | **0.764** ± 0.015 | 0.737 ± 0.016 | **0.633** ± 0.018 | **0.732** ± 0.007 |
| | Laplacian | 0.756 ± 0.015 | **0.765** ± 0.015 | **0.764** ± 0.015 | 0.736 ± 0.016 | 0.629 ± 0.018 | 0.730 ± 0.007 |
| | Matérn ($\nu$=3/2) | 0.757 ± 0.015 | 0.762 ± 0.015 | **0.764** ± 0.015 | **0.740** ± 0.015 | **0.633** ± 0.018 | 0.731 ± 0.007 |

Table E.12: **Sentence-encoder ablation.** Test AUROC with the MMD kernel computed on `MiniLM-L6`, `E5-large-v2` or `BGE-M3` embeddings. The kernel is fixed to RBF and step selection to DiMS (top-$M$). Each model column is the mean over the six datasets and *Average* the mean over the five model variants; intervals correspond to 1.96 standard errors (per-run bootstrap SEs combined in quadrature). Blue denotes the default configuration, used in every other table: `E5-large-v2` embeddings, an RBF kernel and the label-free $\lambda^\star$. **Bold** indicates the best encoder within each estimator block.

| UQ estimator | Encoder | AUROC ↑ | | | | | |
|---|---|---|---|---|---|---|---|
| | | LLaDA-1.5 (`lc`) | LLaDA-1.5 (`rd`) | Dream (`lc`) | Dream (`rd`) | Nemotron (`lc`) | Average |
| Sem. Entropy | MiniLM-L6 | **0.613** ± 0.018 | 0.622 ± 0.017 | **0.699** ± 0.017 | **0.603** ± 0.019 | **0.594** ± 0.019 | **0.626** ± 0.008 |
| | E5-large-v2 (default) | **0.613** ± 0.018 | 0.624 ± 0.018 | 0.696 ± 0.017 | 0.592 ± 0.020 | 0.591 ± 0.018 | 0.623 ± 0.008 |
| | BGE-M3 | 0.605 ± 0.018 | **0.627** ± 0.018 | 0.696 ± 0.017 | 0.594 ± 0.019 | 0.590 ± 0.018 | 0.623 ± 0.008 |
| DegMat | MiniLM-L6 | **0.636** ± 0.018 | 0.663 ± 0.017 | 0.678 ± 0.017 | 0.625 ± 0.018 | 0.606 ± 0.017 | 0.641 ± 0.008 |
| | E5-large-v2 (default) | 0.634 ± 0.018 | **0.665** ± 0.017 | **0.680** ± 0.017 | **0.627** ± 0.018 | **0.610** ± 0.017 | **0.643** ± 0.008 |
| | BGE-M3 | 0.634 ± 0.018 | 0.663 ± 0.017 | **0.680** ± 0.017 | **0.627** ± 0.018 | 0.595 ± 0.017 | 0.640 ± 0.008 |
| Eccentricity | MiniLM-L6 | 0.608 ± 0.018 | **0.651** ± 0.017 | 0.663 ± 0.017 | **0.619** ± 0.018 | **0.592** ± 0.018 | **0.627** ± 0.008 |
| | E5-large-v2 (default) | **0.612** ± 0.018 | 0.649 ± 0.017 | **0.665** ± 0.017 | 0.617 ± 0.018 | 0.591 ± 0.017 | **0.627** ± 0.008 |
| | BGE-M3 | **0.612** ± 0.018 | 0.650 ± 0.017 | 0.663 ± 0.017 | 0.613 ± 0.018 | 0.582 ± 0.017 | 0.624 ± 0.008 |
| EigenVal | MiniLM-L6 | **0.602** ± 0.018 | 0.647 ± 0.017 | **0.653** ± 0.017 | **0.610** ± 0.018 | **0.585** ± 0.018 | **0.619** ± 0.008 |
| | E5-large-v2 (default) | 0.600 ± 0.018 | 0.642 ± 0.018 | 0.650 ± 0.018 | 0.605 ± 0.018 | 0.580 ± 0.018 | 0.616 ± 0.008 |
| | BGE-M3 | 0.596 ± 0.018 | **0.650** ± 0.017 | 0.645 ± 0.018 | 0.605 ± 0.018 | 0.584 ± 0.018 | 0.616 ± 0.008 |
| KLE-Heat | MiniLM-L6 | **0.643** ± 0.018 | **0.665** ± 0.017 | **0.686** ± 0.017 | 0.626 ± 0.018 | **0.612** ± 0.017 | **0.646** ± 0.008 |
| | E5-large-v2 (default) | 0.639 ± 0.018 | 0.664 ± 0.017 | **0.686** ± 0.017 | **0.629** ± 0.018 | 0.611 ± 0.017 | **0.646** ± 0.008 |
| | BGE-M3 | 0.640 ± 0.018 | 0.663 ± 0.017 | **0.686** ± 0.017 | 0.628 ± 0.018 | 0.602 ± 0.017 | 0.644 ± 0.008 |
| KLE-Matérn | MiniLM-L6 | **0.630** ± 0.018 | **0.664** ± 0.017 | 0.688 ± 0.017 | 0.625 ± 0.018 | **0.613** ± 0.017 | **0.644** ± 0.008 |
| | E5-large-v2 (default) | 0.627 ± 0.018 | 0.663 ± 0.017 | **0.689** ± 0.017 | **0.627** ± 0.018 | 0.612 ± 0.017 | **0.644** ± 0.008 |
| | BGE-M3 | **0.630** ± 0.018 | 0.662 ± 0.017 | 0.688 ± 0.017 | **0.627** ± 0.018 | 0.604 ± 0.017 | 0.642 ± 0.008 |
| AD | MiniLM-L6 | **0.723** ± 0.017 | 0.722 ± 0.017 | 0.747 ± 0.016 | 0.665 ± 0.018 | **0.616** ± 0.018 | **0.695** ± 0.008 |
| | E5-large-v2 (default) | 0.719 ± 0.017 | 0.722 ± 0.017 | **0.748** ± 0.016 | 0.662 ± 0.018 | **0.616** ± 0.018 | 0.694 ± 0.008 |
| | BGE-M3 | 0.721 ± 0.017 | **0.726** ± 0.017 | **0.748** ± 0.016 | **0.666** ± 0.018 | 0.615 ± 0.018 | **0.695** ± 0.008 |
| CoCoA | MiniLM-L6 | **0.761** ± 0.015 | 0.763 ± 0.015 | 0.765 ± 0.015 | **0.737** ± 0.016 | **0.633** ± 0.018 | **0.732** ± 0.007 |
| | E5-large-v2 (default) | 0.760 ± 0.015 | **0.764** ± 0.015 | 0.764 ± 0.015 | **0.737** ± 0.016 | **0.633** ± 0.018 | **0.732** ± 0.007 |
| | BGE-M3 | 0.760 ± 0.015 | 0.762 ± 0.015 | **0.766** ± 0.015 | 0.735 ± 0.016 | **0.633** ± 0.018 | 0.731 ± 0.007 |

Table E.13: **Calibration-trajectory ablation.** Test AUROC of DiMS with $R$=20, 10 or 5 sampled trajectories per calibration prompt, with $N_{\text{cal}}$=100 prompts. The budgets are nested (the first $R$ samples of each prompt, used for both the step trajectories and the reference answers), the splits are unchanged, and $\lambda^\star$ is re-selected for each cell. Each value is the mean over the 30 runs; intervals correspond to 1.96 standard errors (per-run bootstrap SEs combined in quadrature). Blue denotes the default configuration ($R$=20).

| UQ estimator | AUROC ↑ | | | | | |
|---|---|---|---|---|---|---|
| | $R$=20 **(default)** | | $R$=10 | | $R$=5 | |
| | top-$M$ | cov. 75% | top-$M$ | cov. 75% | top-$M$ | cov. 75% |
| Sem. Entropy | 0.623 ± 0.008 | 0.628 ± 0.008 | 0.622 ± 0.008 | 0.625 ± 0.008 | 0.623 ± 0.008 | 0.623 ± 0.008 |
| DegMat | 0.643 ± 0.008 | 0.643 ± 0.008 | 0.642 ± 0.008 | 0.643 ± 0.008 | 0.642 ± 0.008 | 0.643 ± 0.008 |
| Eccentricity | 0.627 ± 0.008 | 0.628 ± 0.008 | 0.625 ± 0.008 | 0.627 ± 0.008 | 0.626 ± 0.008 | 0.626 ± 0.008 |
| EigenVal | 0.616 ± 0.008 | 0.620 ± 0.008 | 0.615 ± 0.008 | 0.619 ± 0.008 | 0.614 ± 0.008 | 0.619 ± 0.008 |
| KLE-Heat | 0.646 ± 0.008 | 0.644 ± 0.008 | 0.645 ± 0.008 | 0.644 ± 0.008 | 0.645 ± 0.008 | 0.644 ± 0.008 |
| KLE-Matérn | 0.644 ± 0.008 | 0.645 ± 0.008 | 0.643 ± 0.008 | 0.646 ± 0.008 | 0.643 ± 0.008 | 0.645 ± 0.008 |
| AD | 0.694 ± 0.008 | 0.696 ± 0.008 | 0.695 ± 0.008 | 0.698 ± 0.008 | 0.695 ± 0.008 | 0.697 ± 0.008 |
| CoCoA | 0.732 ± 0.007 | 0.732 ± 0.007 | 0.732 ± 0.007 | 0.732 ± 0.007 | 0.732 ± 0.007 | 0.732 ± 0.007 |

Table E.14: **Calibration-set ablation.** Test AUROC of DIMS with $N_{\text{cal}}$=100, 50 or 10 unlabeled calibration prompts, with $R$=20 trajectories per prompt. The calibration sets are nested ($10 \subset 50 \subset 100$), the validation and test prompts are unchanged, and $\lambda^\star$ is re-selected for each cell; these are the $N_{\text{cal}}$ columns of Table 1, here with intervals. Each value is the mean over the 30 runs; intervals correspond to 1.96 standard errors (per-run bootstrap SEs combined in quadrature). Blue denotes the default configuration ($N_{\text{cal}}$=100).

| UQ estimator | AUROC ↑ | | | | | |
|---|---|---|---|---|---|---|
| | $N_{\text{cal}}$=100 **(default)** | | $N_{\text{cal}}$=50 | | $N_{\text{cal}}$=10 | |
| | top-$M$ | cov. 75% | top-$M$ | cov. 75% | top-$M$ | cov. 75% |
| Sem. Entropy | $0.623 \pm 0.008$ | $0.628 \pm 0.008$ | $0.625 \pm 0.008$ | $0.630 \pm 0.008$ | $0.626 \pm 0.008$ | $0.624 \pm 0.008$ |
| DegMat | $0.643 \pm 0.008$ | $0.643 \pm 0.008$ | $0.641 \pm 0.008$ | $0.644 \pm 0.008$ | $0.640 \pm 0.008$ | $0.642 \pm 0.008$ |
| Eccentricity | $0.627 \pm 0.008$ | $0.628 \pm 0.008$ | $0.626 \pm 0.008$ | $0.628 \pm 0.008$ | $0.620 \pm 0.008$ | $0.623 \pm 0.008$ |
| EigenVal | $0.616 \pm 0.008$ | $0.620 \pm 0.008$ | $0.616 \pm 0.008$ | $0.619 \pm 0.008$ | $0.615 \pm 0.008$ | $0.618 \pm 0.008$ |
| KLE-Heat | $0.646 \pm 0.008$ | $0.644 \pm 0.008$ | $0.642 \pm 0.008$ | $0.644 \pm 0.008$ | $0.641 \pm 0.008$ | $0.644 \pm 0.008$ |
| KLE-Matérn | $0.644 \pm 0.008$ | $0.645 \pm 0.008$ | $0.641 \pm 0.008$ | $0.647 \pm 0.008$ | $0.642 \pm 0.008$ | $0.644 \pm 0.008$ |
| AD | $0.694 \pm 0.008$ | $0.696 \pm 0.008$ | $0.695 \pm 0.008$ | $0.697 \pm 0.008$ | $0.694 \pm 0.008$ | $0.697 \pm 0.008$ |
| CoCoA | $0.732 \pm 0.007$ | $0.732 \pm 0.007$ | $0.732 \pm 0.007$ | $0.731 \pm 0.007$ | $0.732 \pm 0.007$ | $0.733 \pm 0.007$ |

Table E.15: **Weight transfer across models.** Dataset held fixed. Each cell is the change in test AUROC (transferred minus in-domain; negative values are losses) when the DIMS (top-$M$) weights calibrated on the row model variant replace the column model variant's own weights, averaged over the semantic estimators and every shared run pair. *Average* is the off-diagonal mean. A model variant is a (model, remasking) pair, so the `lc` and `rd` variants of one model also transfer to each other.

| Calib. ↓ / Deploy → | LLaDA-1.5 (`lc`) | LLaDA-1.5 (`rd`) | Dream (`lc`) | Dream (`rd`) | Nemotron (`lc`) | Average |
|---|---|---|---|---|---|---|
| **LLaDA-1.5 (`lc`)** | — | −0.027 | −0.022 | −0.015 | −0.007 | −0.018 |
| **LLaDA-1.5 (`rd`)** | −0.005 | — | −0.047 | −0.002 | −0.032 | −0.021 |
| **Dream (`lc`)** | +0.002 | −0.024 | — | −0.011 | −0.005 | −0.009 |
| **Dream (`rd`)** | −0.009 | −0.005 | −0.047 | — | −0.035 | −0.024 |
| **Nemotron (`lc`)** | −0.001 | −0.021 | −0.017 | −0.010 | — | −0.012 |
| **Average** | −0.003 | −0.019 | −0.033 | −0.010 | −0.020 | −0.017 |

Table E.16: **Weight transfer across datasets.** Model variant held fixed. Each cell is the change in test AUROC (transferred minus in-domain; negative values are losses) when the DIMS (top-$M$) weights calibrated on the row dataset replace the column dataset's own weights, averaged over the semantic estimators and every shared run pair. *Average* is the off-diagonal mean. TriviaQA is omitted: its 32-step schedule has no step-aligned counterpart among the 128-step datasets.

| Calib. ↓ / Deploy → | GSM8K | MATH | WMT14 Fr-En | SAMSum | HotpotQA | Average |
|---|---|---|---|---|---|---|
| **GSM8K** | — | −0.003 | −0.012 | −0.008 | −0.016 | −0.010 |
| **MATH** | −0.001 | — | −0.005 | −0.013 | −0.013 | −0.008 |
| **WMT14 Fr-En** | −0.017 | −0.007 | — | −0.003 | −0.011 | −0.010 |
| **SAMSum** | −0.001 | −0.005 | −0.007 | — | −0.006 | −0.005 |
| **HotpotQA** | −0.010 | −0.007 | 0.000 | −0.002 | — | −0.005 |
| **Average** | −0.007 | −0.005 | −0.006 | −0.006 | −0.011 | −0.007 |