The Confidence Paradox: Unveiling the Latent Discriminative Power of Diffusion Large Language Models in Mathematical Reasoning

Anonymous ACL submission

Abstract

Diffusion large language models (DLLMs) have emerged as a promising alternative to autoregressive (AR) generation, uniquely offering token-level probabilities under bidirectional context. However, the semantics of their native uncertainty estimates remain underexplored. In this work, we uncover a **calibration paradox** inherent to the bidirectional generation mechanism of diverse DLLMs. Concretely, we demonstrate that diffusion confidence is structurally distinct from AR likelihood. Notably, LLaDA-8B is highly miscalibrated (31.2% ECE) on mathematical reasoning benchmarks, yet possesses superior discriminative power (0.826 AUROC), significantly outperforming comparable AR baselines in single-pass settings (0.611 AUROC). We diagnose that this paradox arises because diffusion confidence functions less as a probability of correctness and more as a proxy for structural consistency enabled by the model's bidirectional access to the entire solution path. We further show that lightweight post-hoc calibration can reconcile this gap, reducing ECE by over 60% while preserving the strong ranking signal. Our findings suggest that DLLMs offer a unique, cost-efficient uncertainty signal for reasoning tasks that complements expensive AR approaches.

1 Introduction

Large language models (LLMs) deployed in high-stakes reasoning domains require not only high accuracy but also reliable uncertainty estimates to support downstream decision-making (Guo et al., 2017). Alongside autoregressive (AR) generation, diffusion large language models (DLLMs) have recently emerged as a compelling alternative, offering parallel generation capabilities and bidirectional context awareness. However, unlike the extensively studied probability landscape of AR models, the semantics of uncertainty in this emerging model class remain largely underexplored.

In this work, we investigate semantics of uncertainty in the DLLM model class under a unified, single-pass confidence definition derived from token probabilities, thus allowing a direct comparison to AR baselines. On popular mathematical reasoning benchmarks, we uncover a striking *calibration paradox* phenomenon as shown in Figure 1. We observe that DLLMs are severely miscalibrated as probabilistic estimators, e.g., LLaDA-8B exhibits an Expected Calibration Error (ECE) of 31.2% compared to 11.7% for AR baselines. Yet the same score is a strong discriminator of correctness: LLaDA attains 0.826 AUROC, while the AR baseline attains 0.611 AUROC. This gap suggests that diffusion models possess a latent, high-quality discriminative signal masked by poor scaling.

We diagnose this paradox as a result of the unique bidirectional generation mechanism of DLLMs. Unlike AR models that predict the next token based solely on the prefix, DLLMs refine the entire sequence iteratively, accessing bidirectional context at every step. We hypothesize and provide evidence that their confidence scores function less as probability of correctness and more as a proxy for structural consistency—the degree to which the generated solution path is internally coherent. This makes the signal particularly sensitive to logical and arithmetic contradictions (common in math reasoning) even when the surface form remains fluent, explaining its superior discriminative power in structured domains.

Crucially, we show that this paradox is resolvable. Since the ranking signal is intact, the poor calibration is merely a *structured scaling distortion*. We demonstrate that monotone post-hoc recalibration can translate diffusion confidence into better-calibrated probabilities without destroying discrimination, reducing ECE by over 60% while preserving the strong discriminative signal. This finding positions DLLMs not just as a faster generation alternative, but as a source of cost-efficient,

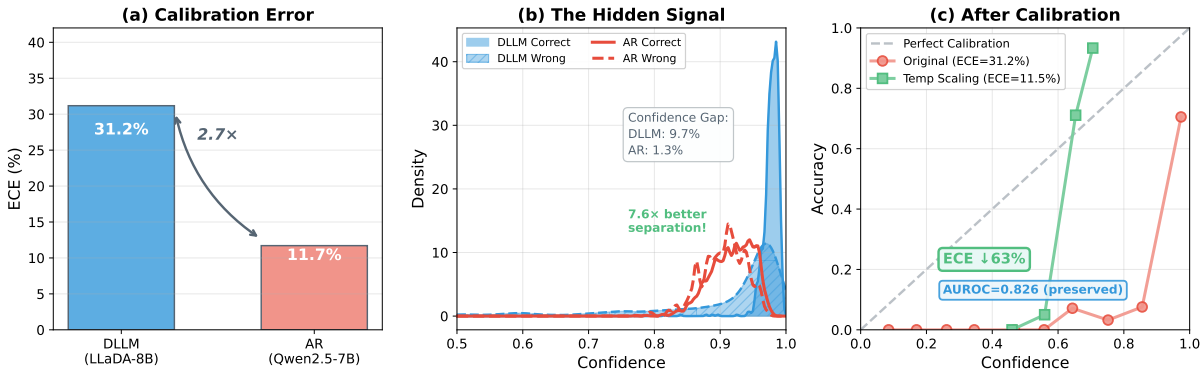

Figure 1: An example of the calibration paradox in LLaDA-8B on GSM8K. Under the same single-pass token-probability confidence definition, LLaDA has higher ECE than an autoregressive baseline, yet higher AUROC. Monotone post-hoc calibration reduces ECE while preserving AUROC, suggesting that diffusion confidence can be strongly discriminative yet mis-scaled.

high-quality uncertainty estimates that can complement expensive AR approaches without the computational overhead of multiple samples.

To summarize, our contributions are three-fold:

(i) We identify and quantify the calibration paradox in DLLMs, where severe probabilistic miscalibration coexists with superior discriminative power, significantly outperforming single-pass autoregressive baselines in reasoning tasks.

(ii) We trace this paradox to the bidirectional generation mechanism of DLLMs, showing that their confidence acts as a proxy for structural consistency rather than pure likelihood, leading to difficulty-dependent scaling distortions.

(iii) We demonstrate that lightweight post-hoc calibration effectively resolves this gap without computational overhead, establishing DLLMs as a source of cost-efficient uncertainty estimates that complement expensive AR approaches.

2 Background

Problem Setting. We study sequence-level confidence estimates produced by reasoning models. Given a prompt, a model generates an answer together with a scalar confidence score. Two properties of this score are central: *calibration*, which measures alignment between confidence and correctness, and *discrimination*, which measures the ability to separate correct from incorrect answers. As we will show, these two properties can be in direct tension.

We first present an overview of baseline accuracy, ECE, and AUROC across benchmarks (Table 1), and then analyze the origin, structure, and persistence of this tension.

2.1 Calibration and Discrimination

A perfectly calibrated model satisfies $\mathbb{P}(y{=}1 \mid p) = p$, where p denotes the model's confidence and $y \in \{0, 1\}$ the correctness label. We quantify deviations from this ideal using the Expected Calibration Error (ECE), which measures the average gap between confidence and accuracy across confidence bins B_b:

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{n} \left|\mathrm{acc}(B_b) - \mathrm{conf}(B_b)\right|. \quad (1)$$

Calibration evaluates probabilistic alignment but does not capture ranking quality. We therefore measure discrimination using the Area Under the ROC Curve (AUROC), defined as the probability that a randomly chosen correct example receives a higher confidence score than a randomly chosen incorrect one (Huang et al., 2024). As a result, a confidence score may exhibit high AUROC even when it is poorly calibrated.

Post-hoc calibration methods such as global temperature scaling (TS) exploit the fact that AUROC is invariant to strictly monotone transformations, allowing ECE to be reduced without affecting discrimination. We focus on TS and a difficulty-aware extension (DATS); while DATS is not globally monotone and may introduce small rank changes, we quantify these effects empirically (Appendix A21, Appendix A12). Additional calibrators are reported in Appendix A8.

2.2 Diffusion Large Language Models

Unlike AR models that generate tokens sequentially, DLLMs such as LLaDA (Nie et al., 2025)

	GSM8K			SVAMP			MATH			GSM-H			TriviaQA		
Model	Acc	ECE↓	AUC↑	Acc	ECE↓	AUC↑	Acc	ECE↓	AUC↑	Acc	ECE↓	AUC↑	Acc	ECE↓	AUC↑
Diffusion Large Language Models (DLLMs)															
LLaDA-8B	62.9	31.2	**.826**	86.2	10.8	**.680**	15.6	73.5	**.836**	24.8	68.2	**.690**	43.4	51.0	.729
Autoregressive LMs (AR)															
Qwen2.5-7B	79.8	11.7	.611	83.1	10.2	.560	46.8	**49.6**	.549	52.4	**41.9**	.671	57.8	**30.3**	**.850**
Llama-3.1-8B	**82.2**	**9.0**	.737	84.3	**7.1**	.790	56.8	33.9	.576	32.1	57.8	.691	73.5	13.1	.839

Table 1: Overview of baseline performance. The calibration paradox is most visible on mathematical reasoning benchmarks (GSM8K, SVAMP, MATH-500, GSM-Hard), where LLaDA has higher discrimination (AUROC) despite poorer calibration (ECE) than AR baselines. On the TriviaQA knowledge task, this pattern does not hold.

iteratively refine a sequence over T steps using bidirectional context. While the corruption schedule and update rule vary across models, the final denoising step yields per-token probabilities for the fully specified sequence. We aggregate these probabilities into a sequence-level confidence score that is directly comparable to token-probability-based confidence in AR baselines.

3 Experimental Setup

3.1 Models and Datasets

Our analysis centers on LLaDA-8B-Instruct (Nie et al., 2025),[1] a state-of-the-art DLLM. We compare it against strong AR counterparts of similar scale: Qwen2.5-7B-Instruct and Llama-3.1-8B-Instruct. We evaluate these models on a suite of benchmarks. For mathematical reasoning, we use GSM8K (Cobbe et al., 2021), SVAMP, the more difficult GSM-Hard (Gao et al., 2023), and the competition-level MATH-500 (Hendrycks et al., 2021). To test domain specificity, we use the open-domain TriviaQA (Joshi et al., 2017) and logical reasoning tasks from BBH (Suzgun et al., 2023).

The baseline performance of all models is summarized in Table 1. These initial results provide the empirical foundation for our entire investigation, motivating the paradox, its diagnosis, and the subsequent analyses. Unless specified otherwise, we report accuracy, ECE, and AUROC on the standard evaluation split for each benchmark (e.g., the full GSM8K test set). For post-hoc calibration experiments, we use a held-out 30/70 calibration/test split as described below.

3.2 Generation and Confidence Extraction

For the main comparisons (LLaDA, Qwen, and Llama), we match task-level prompting and decoding settings within each benchmark. For mathematical reasoning, we use chain-of-thought prompting with a maximum generation length of 512 tokens (256 for the shorter SVAMP problems). For TriviaQA, we use a direct-answer prompt and a 64-token limit. All models generate outputs deterministically (temperature 0 for AR models; greedy decoding for LLaDA with T=128 diffusion steps) and run in BFLOAT16.

We define a unified, single-pass confidence score for both model families. Sequence-level confidence c is the mean probability over generated *output* tokens: $c = \frac{1}{m}\sum_{i=1}^{m} p_i$, where p_i is the softmax probability assigned to the i-th generated output token y_i. For AR models, we decode once and take p_i as the standard next-token probability at generation step i. For LLaDA, after decoding we run one final forward pass at t=T under teacher forcing on the full generated sequence and read the probability of y_i at its position; we do not average probabilities across the denoising trajectory. Unless stated otherwise, we compute ECE with B=15 equal-width bins; for SVAMP, we use B=10 following the dataset-specific evaluation script.

This scoring rule is used to probe how token-probability confidence behaves across model families under a shared definition; we do not assume strict probabilistic equivalence between DLLM final-step token probabilities and AR next-token probabilities. We additionally report robustness checks for alternative aggregations and for path-derived diffusion scores in the appendix (Appendix A8, Appendix A17).

Calibration Protocol. Unless stated otherwise, we evaluate post-hoc calibration with a 30/70 calibration/test split on each dataset (seed 0): temperatures and DATS mappings are fit on the 30% calibration split and evaluated on the held-out 70% test split. When we report full-set calibration curves, we label them as diagnostic upper bounds (Appendix A14).

[1] We will show that the finding generalizes across DLLM variants in Section 6.2.

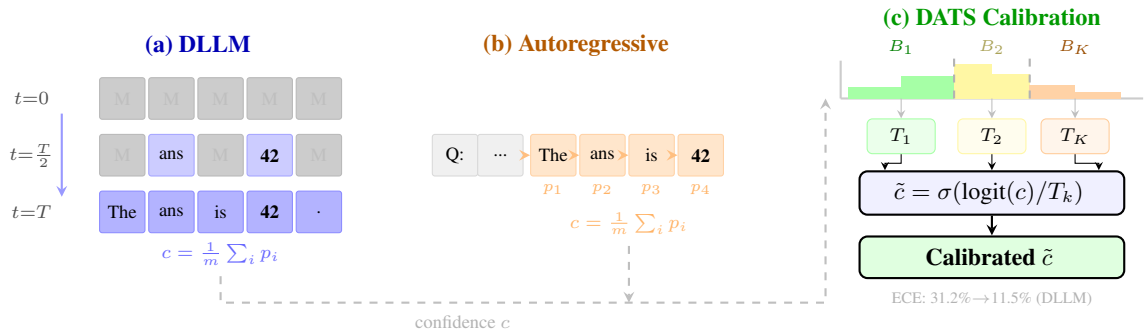

Figure 2: Overview of confidence extraction and calibration. **(a)** For DLLMs, sequence-level confidence c is the mean of per-token probabilities from the final, fully-denoised output. **(b)** For autoregressive models, we use the same aggregation over standard next-token probabilities to ensure a fair comparison. **(c)** Our proposed DATS method partitions examples into K difficulty buckets and learns a separate temperature T_k for each, mapping the raw confidence c to a calibrated score $\tilde{c}$.

Model	Acc.	ECE↓	AUROC↑	Gap
Qwen2.5-7B	79.8%	11.7%	0.611	+1.3%
LLaDA-8B	62.9%	31.2%	**0.826**	+9.7%

Table 2: The calibration paradox on GSM8K. LLaDA is overconfident (high ECE) but more discriminative (high AUROC). The confidence gap (the mean confidence difference between correct and incorrect answers) provides an additional view of this separation.

4 The Calibration Paradox

Model confidence serves two distinct roles that are often conflated: *calibration*, its alignment with the true probability of correctness, and *discrimination*, its ability to separate correct predictions from incorrect ones. On mathematical reasoning, we find that DLLMs can excel at discrimination even when they are poorly calibrated. We refer to this mismatch as the calibration paradox.

4.1 Poor Calibration, Strong Discrimination

Table 2 quantifies this phenomenon on the full GSM8K evaluation set. LLaDA is overconfident under our confidence definition, with an ECE of 31.2% compared to 11.7% for the AR baseline. At the same time, its confidence is a stronger discriminator, achieving an AUROC of 0.826 compared to 0.611 for Qwen. While the models differ in accuracy, AUROC evaluates within-model ranking quality over each model's own correct versus incorrect predictions.

The confidence gap further illustrates this divide: LLaDA assigns 9.7 percentage points more confidence to its correct answers on average, whereas the gap for Qwen is 1.3 points. This indicates that diffusion confidence can be informative for ranking even when it is poorly scaled.

Resolving the Paradox. The tension between poor calibration and strong discrimination motivates the remainder of our analysis. We proceed by diagnosing the structure of this paradox (Section 5), demonstrating its resolution via post-hoc calibration (Section 6), and finally probing the underlying mechanism (Section 7).

5 Difficulty-Conditional Miscalibration

To diagnose the calibration paradox, we ask whether miscalibration varies systematically with problem difficulty. We find that as problems become harder, LLaDA's overconfidence increases while its discriminative signal remains present. This pattern supports a scaling-distortion interpretation: the score continues to separate correct from incorrect answers, but its mapping to probabilities degrades. Figure 2 summarizes our confidence extraction and calibration pipeline.

5.1 Miscalibration Worsens with Difficulty

We first analyze how LLaDA's calibration changes on GSM8K as problems become more complex. Using the number of arithmetic operations in the ground-truth solution as a difficulty proxy,[2] we partition the dataset into four bins. This oracle proxy is used for diagnosis; in Section 6 we also evaluate confidence-binned calibration that does not require ground-truth annotations. Table 3 shows that ECE increases monotonically with difficulty, rising from 21.8% on the easiest problems to 41.7%

[2]We count all occurrences of $\{+, -, \times, \div\}$ in the annotated solution text. See Appendix A15 for binning details.

Difficulty	Acc.	ECE	Corr.
Easy (0-4 ops)	72.4%	21.8%	+0.50
Medium (5-7 ops)	66.7%	28.1%	+0.40
Hard (8-10 ops)	59.4%	34.1%	+0.38
Hardest (11+ ops)	52.0%	41.7%	+0.34

Table 3: Difficulty-conditional analysis on GSM8K. As problem difficulty increases (by operation count), ECE worsens while the confidence–correctness correlation remains positive, consistent with scaling distortion.

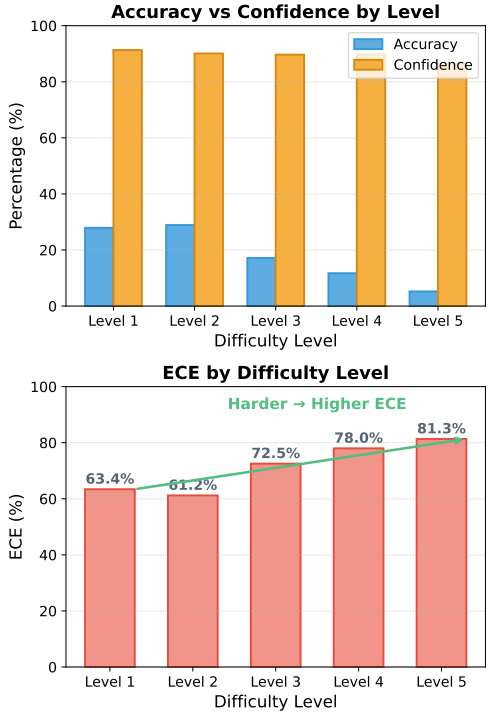

Figure 3: Difficulty-conditional analysis on MATH-500. **Top**: Accuracy decreases with difficulty while average confidence remains high. **Bottom**: ECE rises for harder problems.

on the hardest. At the same time, the confidence–correctness correlation remains positive across all bins. This supports a scaling-distortion diagnosis: the confidence score retains ranking information, but its alignment with true probabilities degrades as problems become more challenging.

5.2 Validation on Harder Benchmarks

To ensure that this pattern is not specific to GSM8K, we validate it on two more challenging benchmarks: MATH-500 and GSM-Hard. Figure 3 shows that on MATH-500, accuracy decreases with difficulty while average confidence remains high, causing ECE to climb from 63.4% to 81.3%.

Table 4 shows that across math benchmarks, LLaDA's ECE increases as tasks become harder,

Dataset	Model	Acc.	ECE↓	AUROC↑	Corr.
GSM8K	LLaDA	62.9%	31.2%	0.826	+0.403
MATH-500	LLaDA	15.6%	73.5%	0.836	+0.221
GSM-Hard	LLaDA	24.8%	68.2%	0.690	+0.197
GSM-Hard	Qwen2.5-7B	52.4%	41.9%	0.671	+0.303

Table 4: Validation across harder math benchmarks. LLaDA maintains strong discrimination (AUROC) even as accuracy drops and ECE rises, while Qwen shows better calibration but only comparable AUROC on GSM-Hard.

while AUROC and correlation remain positive. Notably, on GSM-Hard, the AR model Qwen achieves better calibration but comparable discrimination to LLaDA (AUROC 0.671 vs. 0.690). Overall, these results indicate that the probabilistic interpretation of diffusion confidence degrades with task difficulty, even when its ranking signal remains useful.

6 Difficulty-Aware Temperature Scaling

Our diagnosis in Section 5 suggests that the calibration paradox arises from structured mis-scaling rather than lacking signal. This motivates post-hoc recalibration. We evaluate lightweight calibration methods that map the confidence score to a better-calibrated probability while preserving discrimination.

6.1 Method: Rescaling the Signal

We evaluate two methods. The first is standard **global temperature scaling (TS)** (Guo et al., 2017), which corrects for a consistent over- or under-confidence across all examples. We adapt it to scalar confidence scores $p \in (0, 1)$ by rescaling their log-odds:

$$p_{\text{scaled}} = \sigma(\text{logit}(p)/T). \quad (2)$$

A single temperature T, fit on a calibration set, is applied to all test examples. Since this transformation is strictly monotone, it preserves AUROC by construction.

However, our diagnosis showed that miscalibration worsens with difficulty. This motivates our second method, **Difficulty-Aware Temperature Scaling (DATS)**. Instead of learning one global temperature, DATS first partitions the calibration data into K buckets based on a difficulty proxy and then learns a separate temperature T_k for each bucket. Unless stated otherwise, our split-based results use confidence-bucketed DATS (a deployable proxy); oracle difficulty (operation count) is used only for diagnosis in Section 5 and Appendix A15.

	GSM8K		
	ECE↓	ΔECE	AUROC
LLaDA-8B (DLLM)			
Original	31.0	–	0.826
Global TS	12.1	−61%	0.826
DATS	**10.1**	−67%	0.826
Qwen2.5-7B (AR)			
Original	11.7	–	0.622
Global TS	2.3	−80%	0.622
DATS	**1.6**	−86%	0.626

Table 5: Calibration on GSM8K (30% calibration, 70% test, seed 0). Both TS and DATS substantially reduce ECE. TS leaves AUROC unchanged, while DATS induces a small AUROC change for Qwen ($\Delta = +0.004$) and a minimal change for LLaDA ($\Delta < 0.0001$).

Model	ECE↓ (%)			AUROC↑		
	Orig	TS	DATS	Orig	TS	DATS
LLADA2.0-mini	70.0	38.3	38.3	0.798	0.798	0.798
LLADA2.0-mini-CAP	74.8	43.8	43.8	0.824	0.824	0.824
LLaDA-8B-Instruct-SFT	48.1	28.9	29.2	0.839	0.839	0.809

Table 6: Generalization across different diffusion models on GSM8K-200. The calibration paradox persists across model variants. Global temperature scaling (TS) consistently reduces ECE while preserving discriminative performance (AUROC), while DATS further improves calibration with minor AUROC degradation in some cases.

Protocol. For each benchmark, calibration mappings are fit on a 30% split and evaluated using ECE and AUROC on the remaining 70%.

6.2 Results

We first report results on the held-out 70% test split of GSM8K, with all calibration mappings fit on the 30% calibration split. Table 5 shows that post-hoc calibration substantially reduces ECE. Global TS reduces LLaDA's ECE from 31.0% to 12.1%, while preserving AUROC at 0.826 by construction. DATS further reduces ECE to 10.1%. For Qwen, DATS increases AUROC from 0.622 to 0.626, whereas the change for LLaDA is minimal ($\Delta < 0.0001$). These results indicate that much of the initial miscalibration can be corrected by a monotone (or near-monotone) transformation of confidence.

Rank Preservation. TS preserves AUROC exactly due to monotonicity, while DATS may induce small rank changes across buckets; we report AUROC variations and rank inversion statistics across random seeds in Appendix A21.

Generalization Across DLLM Variants. To test whether the calibration paradox generalizes beyond a single model, we evaluate several DLLM variants that share a common backbone but differ in scale and training objectives, including LLADA2.0-mini, its CAP-trained variant (Bie et al., 2025), and LLaDA-8B-Instruct-SFT. All models are evaluated on GSM8K-200 under the same 30/70 calibration/test protocol (seed 0). As shown in Table 6, all variants exhibit strong discrimination but poor raw calibration. Post-hoc calibration consistently reduces ECE across models, with only minor AUROC degradation in some cases, indicating that the calibration behavior generalizes across DLLM variants.

To visualize the effect of calibration, Figure 4 presents a diagnostic analysis on the *full* GSM8K evaluation set, where calibration maps are both fit and evaluated on the same data. While these full-set numbers serve as an optimistic upper bound and may differ slightly from the strict split results in Table 5, they illustrate the mechanism. The reliability diagram (middle panel) shows calibrated scores are much closer to the ideal diagonal, and this improvement holds for both easy and hard problems (right panel).

Beyond GSM8K, we observe similar trends under the same 30/70 split protocol: global TS reduces ECE while leaving AUROC unchanged on MATH-500, GSM-Hard, and TriviaQA (Appendix A13). In summary, these results support the view that a large component of diffusion overconfidence is correctable mis-scaling, and that a simple post-hoc map can translate the ranking signal into better-calibrated probabilities.

7 Domain Specificity of the Signal

Having established the calibration paradox and shown that it can be mitigated by post-hoc scaling, we next ask what the diffusion confidence score is sensitive to. We test the hypothesis that the discriminative signal is domain-specific and related to *structural consistency* rather than a generic probability of correctness. This section evaluates this hypothesis via cross-domain comparisons and targeted probes.

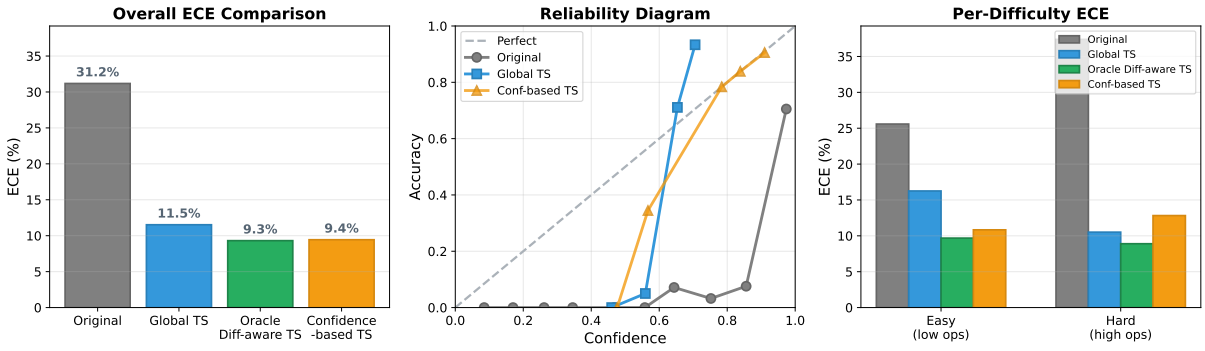

Figure 4: The effect of calibration on LLaDA-8B (full GSM8K set). **Left**: Overall ECE drops from 31.2% to below 12% with TS and DATS. **Middle**: The reliability diagram shows that calibrated predictions (blue/green) are much closer to the ideal diagonal than the original overconfident scores (red). **Right**: Recalibration improves ECE for both easy and hard difficulty bins, demonstrating broad effectiveness.

Task Type	Dataset	Corr.	AUROC
Math	GSM8K	+0.403	0.826
	SVAMP	+0.181	0.680
	MATH-500	+0.221	0.836
	GSM-Hard	+0.197	0.690
Knowledge	TriviaQA	+0.290	0.729

Table 7: Domain specificity of the discriminative signal. The strong discrimination (high AUROC and correlation) is most prominent on mathematical reasoning tasks and weaker on open-domain knowledge QA, suggesting the signal is not a universal correctness detector.

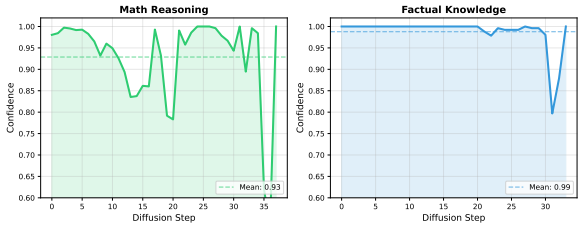

Figure 5: Step-wise confidence traces for representative examples. **Left**: On a math problem, confidence can dip around arithmetic steps, suggesting sensitivity to computational consistency. **Right**: On a factual recall task, confidence remains high and relatively flat.

7.1 The Signal is Strongest in Structured Domains

We begin by comparing structured mathematical reasoning to open-domain knowledge retrieval. Table 7 shows that the high AUROC observed on GSM8K persists across other math benchmarks. On TriviaQA, AUROC is 0.729, which is lower than on the math benchmarks and lower than the AR baseline on this task (0.850). This pattern suggests that the diffusion confidence signal is most informative in domains with strong internal constraints.

7.2 Probing the Mechanism

To probe what drives this domain dependence, we test whether token-probability confidence is sensitive to violations of *structural consistency*. In a math problem, an intermediate arithmetic error introduces a contradiction with other quantities in the chain, whereas a factual error can remain locally fluent. We evaluate this hypothesis with two probes.

Step-wise Denoising Analysis. We examine how the discriminative signal evolves during generation. Quantitatively, per-step AUROC increases from early to late denoising on GSM8K (Appendix A16). Figure 5 provides qualitative illustrations of step-wise confidence traces for representative examples.

Intervention Probe. We use a controlled intervention to isolate sensitivity to arithmetic versus factual errors. We construct minimal pairs that differ by a single substitution, making the statement arithmetically or factually incorrect, and evaluate n=500 pairs for each probe (Appendix A20). Figure 6 shows a representative example for LLaDA: forcing an arithmetic error (e.g., $23 + 45 = 72$) causes a 28pp drop in confidence, whereas forcing a factual swap (e.g., "Yale is in Boston") causes a 3.4pp drop. In contrast, the same factual-swap probe yields a larger confidence gap for the AR baseline (Appendix A20), indicating that this asymmetry is probe- and model-dependent. Overall, these results suggest that arithmetic inconsistencies are strongly penalized, while factual substitutions exhibit model- and scoring-dependent behavior.

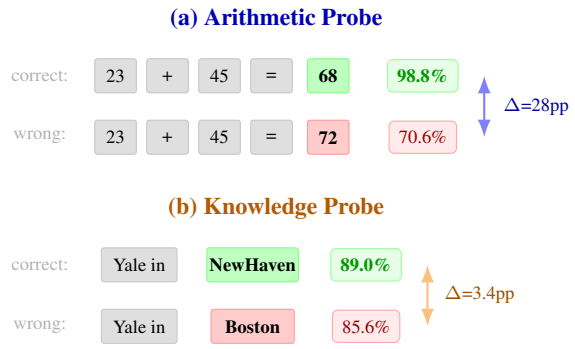

Figure 6: Intervention probe comparison. **(a) Arithmetic Probe:** Swapping a correct result for an incorrect one causes a 28pp drop in confidence (AUROC=1.0). **(b) Knowledge Probe:** Swapping a correct answer for an incorrect one causes a 3.4pp drop (AUROC=0.6).

Summary of Evidence. The results are consistent with a structural-consistency interpretation of diffusion confidence, without establishing a causal mechanism, and show that the calibration paradox is strongest on mathematical reasoning and weaker on less structured tasks such as BBH (Appendix A9).

8 Related Work

Calibration of Neural Networks and LLMs. Post-hoc calibration, particularly temperature scaling (Guo et al., 2017), is a standard approach for aligning a model's confidence with empirical correctness probabilities (Kull et al., 2019; Minderer et al., 2021). Recent work extends these ideas to LLMs, studying self-knowledge, adaptive calibration procedures, and harmonized uncertainty estimation (Kadavath et al., 2022; Yin et al., 2023; Xie et al., 2024; Kapoor et al., 2024; Li et al., 2025). We contribute not by proposing a new calibrator, but by using established calibration tools to study the *semantics* of a DLLM's confidence signal and to contrast it with AR baselines under a unified confidence definition.

Uncertainty Signals for Reasoning. Many approaches to LLM uncertainty aim to construct stronger confidence signals, for example via multi-sample agreement (self-consistency) (Wang et al., 2023), verbalized confidence (Xiong et al., 2024), or semantic entropy (Kuhn et al., 2023). These methods often improve discrimination but incur additional computation or supervision. In contrast, we focus on the native, single-pass token-probability confidence of DLLMs and study when this signal is informative for discrimination, when it fails as a probability, and how lightweight post-hoc calibration can translate it into better-calibrated probabilities.

Selective Prediction and Ranking-Based Use. Confidence is frequently used for abstention and selective prediction (Geifman and El-Yaniv, 2017). Our emphasis on discrimination (AUROC) complements calibration-focused analyses by explicitly separating ranking utility from probabilistic accuracy. Relatedly, rank-calibration evaluates uncertainty through ranking-based criteria and highlights that ranking quality and probability calibration can diverge (Huang et al., 2024).

Diffusion Language Models. Diffusion models were adapted to discrete sequences by Austin et al. (2021) and Hoogeboom et al. (2021). Recent diffusion LMs, including LLaDA (Nie et al., 2025) and Dream (Ye et al., 2025), have begun to show competitive performance on reasoning tasks. The calibration and uncertainty behavior of these models is less studied than that of AR LMs. Our work provides an empirical analysis of calibration versus discrimination for DLLM confidence under a unified confidence definition, and relates the observed paradox to the bidirectional, iterative generation process.

9 Conclusion

In this work, we identify, diagnose, and mitigate a calibration paradox in DLLMs. On mathematical reasoning tasks, LLaDA is poorly calibrated yet highly discriminative under a unified single-pass confidence definition. We show that miscalibration increases with difficulty while discrimination remains high, and provide evidence consistent with a structural-consistency interpretation of the confidence signal.Simple post-hoc calibration improves ECE while largely preserving discrimination; Appendix A18 provides context via self-consistency analysis. Practically, our findings suggest that a diffusion model's single-pass confidence can be useful for ranking and selective prediction on structured reasoning tasks, even before it is well calibrated as a probability. When probabilistic outputs are required, monotone calibration offers a lightweight way to translate this signal while maintaining its ranking utility. Future work includes isolating which architectural components drive this behavior and improving calibration robustness across domains.

Limitations

Our analysis focuses on the emerging class of DLLMs using LLADA-8B as a representative variant. Observations on less capable models suggest the identified paradox may be an emergent property tied to model scale and architectural maturity. Besides, we prioritize structured reasoning domains where the proposed mechanism of structural consistency is most salient, and restrict our evaluation to native single-pass confidence measures to highlight the potential for cost-efficient uncertainty estimation. We leave the exploration of unstructured tasks and computationally expensive sampling-based variants to future work.

References

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. 2021. Structured denoising diffusion models in discrete state-spaces. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 17981–17993.

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, and 12 others. 2025. Llada2.0: Scaling up diffusion language models to 100b. *Preprint*, arXiv:2512.15745.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. *CoRR*, abs/2110.14168.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2023. PAL: program-aided language models. In *International Conference on Machine Learning, ICML 2023, 23-29 July 2023, Honolulu, Hawaii, USA*, volume 202 of *Proceedings of Machine Learning Research*, pages 10764–10799. PMLR.

Yonatan Geifman and Ran El-Yaniv. 2017. Selective classification for deep neural networks. In *Advances in Neural Information Processing Systems 30: Annual Conference on Neural Information Processing Systems 2017, December 4-9, 2017, Long Beach, CA, USA*, pages 4878–4887.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning, ICML 2017, Sydney, NSW, Australia, 6-11 August 2017*, volume 70 of *Proceedings of Machine Learning Research*, pages 1321–1330. PMLR.

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

Emiel Hoogeboom, Didrik Nielsen, Priyank Jaini, Patrick Forré, and Max Welling. 2021. Argmax flows and multinomial diffusion: Learning categorical distributions. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 12454–12465.

Xinmeng Huang, Shuo Li, Mengxin Yu, Matteo Sesia, Hamed Hassani, Insup Lee, Osbert Bastani, and Edgar Dobriban. 2024. Uncertainty in language models: Assessment through rank-calibration. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 284–312. Association for Computational Linguistics.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics, ACL 2017, Vancouver, Canada, July 30 - August 4, Volume 1: Long Papers*, pages 1601–1611. Association for Computational Linguistics.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, Scott Johnston, Sheer El-Showk, Andy Jones, Nelson Elhage, Tristan Hume, Anna Chen, Yuntao Bai, Sam Bowman, Stanislav Fort, and 17 others. 2022. Language models (mostly) know what they know. *Preprint*, arXiv:2207.05221.

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

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Meelis Kull, Miquel Perelló-Nieto, Markus Kängsepp, Telmo de Menezes e Silva Filho, Hao Song, and Peter A. Flach. 2019. Beyond temperature scaling: Obtaining well-calibrated multiclass probabilities with dirichlet calibration. *CoRR*, abs/1910.12656.

Rui Li, Jing Long, Muge Qi, Heming Xia, Lei Sha, Peiyi Wang, and Zhifang Sui. 2025. Towards harmonized uncertainty estimation for large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2025, Vienna, Austria, July 27 - August 1, 2025*, pages 22938–22953. Association for Computational Linguistics.

Matthias Minderer, Josip Djolonga, Rob Romijnders, Frances Hubis, Xiaohua Zhai, Neil Houlsby, Dustin Tran, and Mario Lucic. 2021. Revisiting the calibration of modern neural networks. In *Advances in Neural Information Processing Systems 34: Annual Conference on Neural Information Processing Systems 2021, NeurIPS 2021, December 6-14, 2021, virtual*, pages 15682–15694.

Alexandru Niculescu-Mizil and Rich Caruana. 2005. Predicting good probabilities with supervised learning. In *Machine Learning, Proceedings of the Twenty-Second International Conference (ICML 2005), Bonn, Germany, August 7-11, 2005*, volume 119 of *ACM International Conference Proceeding Series*, pages 625–632. ACM.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Ji-Rong Wen, and Chongxuan Li. 2025. Large language diffusion models. *Preprint*, arXiv:2502.09992.

Mahdi Pakdaman Naeini, Gregory Cooper, and Milos Hauskrecht. 2015. Obtaining well calibrated probabilities using bayesian binning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 29(1).

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc V. Le, Ed H. Chi, Denny Zhou, and Jason Wei. 2023. Challenging big-bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 13003–13051. Association for Computational Linguistics.

Juozas Vaicenavicius, David Widmann, Carl R. Andersson, Fredrik Lindsten, Jacob Roll, and Thomas B. Schön. 2019. Evaluating model calibration in classification. In *The 22nd International Conference on Artificial Intelligence and Statistics, AISTATS 2019, 16-18 April 2019, Naha, Okinawa, Japan*, volume 89 of *Proceedings of Machine Learning Research*, pages 3459–3467. PMLR.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V. Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations, ICLR 2023, Kigali, Rwanda, May 1-5, 2023*. OpenReview.net.

Johnathan Xie, Annie S. Chen, Yoonho Lee, Eric Mitchell, and Chelsea Finn. 2024. Calibrating language models with adaptive temperature scaling. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing, EMNLP 2024, Miami, FL, USA, November 12-16, 2024*, pages 18128–18138. Association for Computational Linguistics.

Miao Xiong, Zhiyuan Hu, Xinyang Lu, Yifei Li, Jie Fu, Junxian He, and Bryan Hooi. 2024. Can llms express their uncertainty? an empirical evaluation of confidence elicitation in llms. In *The Twelfth International Conference on Learning Representations, ICLR 2024, Vienna, Austria, May 7-11, 2024*. OpenReview.net.

Jiacheng Ye, Zhihui Xie, Lin Zheng, Jiahui Gao, Zirui Wu, Xin Jiang, Zhenguo Li, and Lingpeng Kong. 2025. Dream 7b: Diffusion large language models. *CoRR*, abs/2508.15487.

Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. 2023. Do large language models know what they don't know? In *Findings of the Association for Computational Linguistics: ACL 2023, Toronto, Canada, July 9-14, 2023*, pages 8653–8665. Association for Computational Linguistics.

Bianca Zadrozny and Charles Elkan. 2002. Transforming classifier scores into accurate multiclass probability estimates. In *Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, July 23-26, 2002, Edmonton, Alberta, Canada*, pages 694–699. ACM.

A Appendix

A1 DATS Hyperparameter Stability

Figure A1.1 studies the stability of Difficulty-Aware Temperature Scaling (DATS) as the number of confidence buckets varies on GSM8K (full-set diagnostic). For LLaDA-8B, ECE remains stable at approximately 9.5–9.8% for $K \in \{3, 5, 10\}$. For the AR baseline (Qwen2.5-7B), DATS achieves near-perfect calibration (ECE $< 1\%$) for sufficiently large K. Since DATS is only monotone within buckets, AUROC can change slightly as K varies; Appendix A21 and Appendix A12 report the observed rank effects under our split protocol.

A2 ROC Curves

Figure A2.1 shows ROC curves comparing LLaDA-8B and Qwen2.5-7B on GSM8K. Despite worse calibration, LLaDA attains higher AUROC (0.826 vs 0.611), indicating stronger discrimination under the shared single-pass confidence definition.

A3 Additional Reliability Diagram

Figure A3.1 shows a reliability diagram comparing LLaDA-8B and Qwen2.5-7B on the full

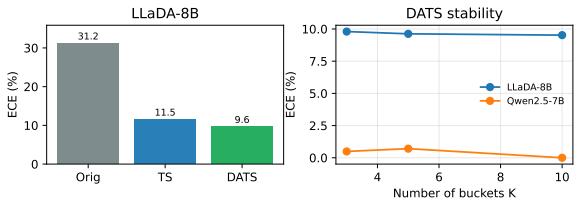

Figure A1.1: DATS stability across different numbers of buckets K (full-set diagnostic). Left: ECE comparison across calibration methods. Right: ECE as a function of K for DATS, showing stable calibration for LLaDA and rapidly improving calibration for Qwen as K increases.

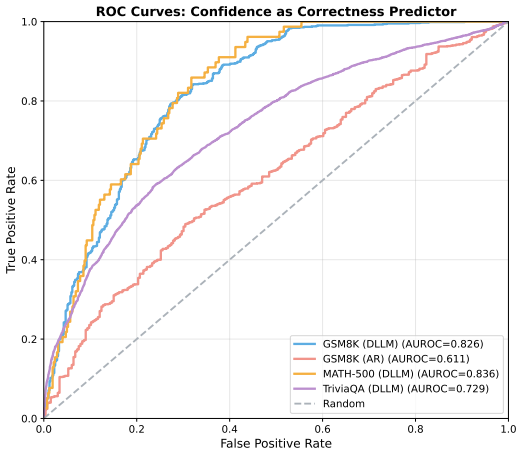

Figure A2.1: ROC curves on GSM8K. LLaDA achieves higher AUROC (0.826) than Qwen2.5 (0.611) despite worse calibration.

GSM8K evaluation set. The autoregressive baseline lies closer to the diagonal (lower ECE 11.7% vs 31.2%).

A4 Additional Arithmetic Consistency Results

We report arithmetic consistency results for the autoregressive baseline on the same synthetic probe used in Section 7. For each randomly sampled equation $a+b$, we construct two prompts differing only in the final result c vs. $d \neq c$ and measure the mean probability assigned to the result tokens under Qwen2.5-7B-Instruct. Across 100 such pairs, Qwen assigns 97.1% average probability to correct results and 56.8% to incorrect ones, yielding a gap of 40.3 percentage points and AUROC $= 1.0$ when using these scores to distinguish correct from wrong chains. This mirrors the DLLM probe (LLaDA: 98.8% vs 70.6%, gap $\approx$28pp, AUROC $= 1.0$) and indicates that local arithmetic consistency is captured by both model families; the main text focuses on how this signal manifests in

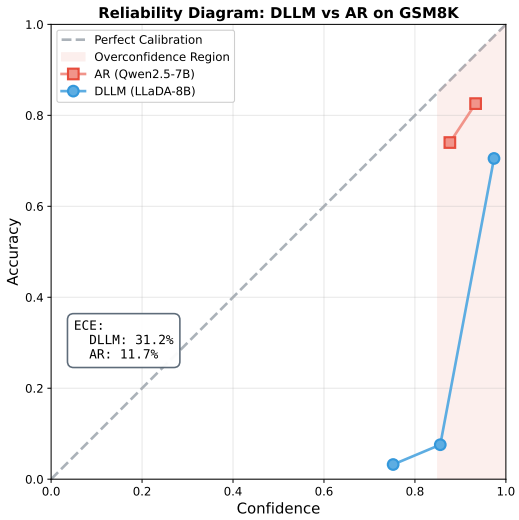

Figure A3.1: Reliability diagram on GSM8K. Qwen2.5 is better calibrated while LLaDA exhibits overconfidence.

sequence-level discrimination on real benchmarks.

A5 Additional Metric Definitions

We summarize the definitions of auxiliary metrics used in the paper. Let $s \in [0, 1]$ denote a confidence score and $y \in \{0, 1\}$ the correctness label.

Expected Calibration Error (ECE). Given confidence bins B_b, ECE is

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|B_b|}{n} \left|\mathrm{acc}(B_b) - \mathrm{conf}(B_b)\right|. \quad (3)$$

Discrimination (AUROC). The Area Under the ROC Curve is

$$\mathrm{AUROC} = \mathbb{P}\big(s^+ > s^-\big), \quad (4)$$

the probability that a randomly chosen correct prediction (s^+) receives higher confidence than a randomly chosen incorrect one (s^-).

Correlation and Confidence Gap. The confidence–correctness correlation is the Pearson correlation

$$\mathrm{Corr.} = \mathrm{corr}(s, y), \quad (5)$$

and the confidence gap is the difference between mean confidence on correct and incorrect predictions

$$\mathrm{Gap} = \mathbb{E}[s \mid y = 1] - \mathbb{E}[s \mid y = 0]. \quad (6)$$

Brier Score. The Brier score is a proper scoring rule that combines calibration and discrimination:

$$\text{Brier} = \mathbb{E}\big[(s - y)^2\big]. \tag{7}$$

These quantities are used as supporting diagnostics; the main conclusions in the paper are based on ECE and AUROC.

A6 Selective Prediction and Risk–Coverage

We report selective prediction performance on GSM8K using risk–coverage curves (Geifman and El-Yaniv, 2017). Following standard practice, we sort examples by confidence and vary a threshold to obtain coverage (fraction of examples we answer) and risk (error rate among answered examples). The Area Under the Risk–Coverage curve (AURC; lower is better) summarizes this trade-off.

Figure A6.1 shows risk–coverage curves for LLaDA-8B and Qwen2.5-7B using original and calibrated confidences. Qwen achieves a lower AURC (0.144 vs. 0.176), reflecting its higher overall accuracy, while temperature scaling (and DATS where evaluated) leaves AURC essentially unchanged. In this setting, selective prediction is therefore dominated by accuracy, whereas the diffusion advantage primarily appears in AUROC-based discrimination at full coverage.

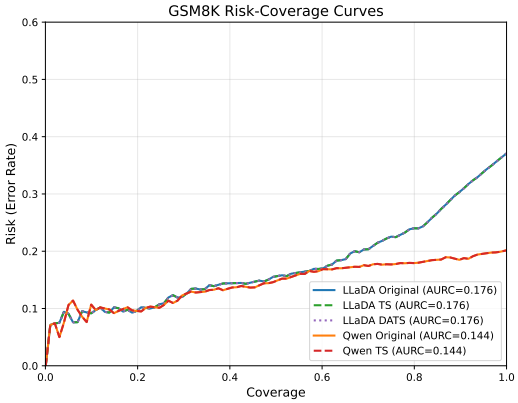

Figure A6.1: Risk–coverage curves on GSM8K. Qwen2.5-7B attains a lower AURC than LLaDA, consistent with its higher accuracy, while calibration has little effect on AURC. In contrast, LLaDA maintains higher AUROC at full coverage in the main results.

A7 Cross-Task Calibration for AR Baselines

Table A7.1 reports the effect of transferring GSM8K calibration parameters to other math benchmarks for the autoregressive baseline Qwen2.5-7B. We use the same protocol as for diffusion: a single temperature and confidence-based DATS mapping learned on GSM8K are applied without retraining.

Target	ECE_{orig}	ECE_{TS}	ECE_{DATS}	AUROC
GSM-Hard	41.9	29.9	32.7	0.671
SVAMP	10.2	5.1	3.8	0.560

Table A7.1: Cross-task calibration transfer for Qwen2.5-7B. Temperatures and DATS buckets are learned on GSM8K and applied to other math benchmarks without re-fitting. Global temperature scaling reduces ECE (15-bin ECE) on GSM-Hard and SVAMP while preserving AUROC, suggesting a transferable scaling component within the math domain.

A8 Additional Calibration Baselines and Aggregators

Non-parametric post-hoc calibration. We fit histogram binning and isotonic regression calibrators (Kull et al., 2019; Vaicenavicius et al., 2019; Zadrozny and Elkan, 2002; Pakdaman Naeini et al., 2015; Niculescu-Mizil and Caruana, 2005) on GSM8K confidences. Both methods drive ECE to (near) zero for LLaDA and Qwen (0.31→0.00 and 0.12→0.00, respectively) while leaving AUROC essentially unchanged (0.826→0.825/0.833 for LLaDA, 0.611→0.623/0.626 for Qwen). This supports the interpretation that the dominant error is global mis-scaling, which can be corrected post-hoc without materially changing the underlying ranking signal.

Aggregation ablations. We probe the robustness of sequence-level confidence to the choice of aggregation over token probabilities. For the autoregressive baseline on GSM8K, using the same per-token probabilities, we compare mean aggregation (default), minimum token probability, geometric mean over all tokens, and an answer-region proxy given by the geometric mean over the last $K{=}8$ tokens. Mean and full-sequence geometric mean yield similar ECE and AUROC (0.12 vs 0.10 ECE, 0.611 vs 0.614 AUROC). In contrast, the minimum aggregator substantially worsens ECE (0.52) despite slightly higher AUROC (0.663), and the tail-based answer proxy performs worse (ECE≈0.14, AUROC≈0.49). For LLaDA-8B, a GSM8K subsample with token-level confidences shows a similar ordering: mean aggregation attains the highest AUROC (≈0.855), geometric mean is slightly worse (≈0.84), and minimum aggregation degrades AUROC substantially (≈0.72). These ablations

indicate that the main qualitative conclusions do not hinge on a particular reasonable aggregation choice; we therefore adopt mean pooling for both model families.

A9 BBH Validation on Logical Reasoning

To probe the generality of the diffusion confidence signal beyond math and factual QA, we evaluate LLaDA-8B-Instruct on three Big-Bench Hard (BBH) subtasks: logical_deduction_three_objects, boolean_expressions, and date_understanding, using the same confidence extraction protocol as in the main text. Table A9.1 reports accuracy, mean confidence, correlation, and confidence gap on the held-out test sets.

Subtask	Acc	Conf.	Corr.	Gap
logical_deduction_3obj	86.8	98.7	+0.04	+0.08
boolean_expressions	0.0	97.8	+0.00	−97.79
date_understanding	68.4	97.1	+0.02	+0.07

Table A9.1: BBH validation results for LLaDA-8B-Instruct. The model is extremely overconfident across all subtasks (confidence ≈97–99%), with weak or even degenerate discrimination on boolean_expressions. We therefore use BBH as a stress test and focus our main analysis on math word problems and factual QA, where the discriminative signal is more stable.

A10 Brier Score Summary on GSM8K

Table A10.1 summarizes Brier scores on GSM8K before and after global temperature scaling. For LLaDA-8B, Brier improves from 0.299 to 0.212 (a reduction of 0.087); for Qwen2.5-7B, Brier improves from 0.172 to 0.158. These trends mirror the ECE reductions reported in the main text under a proper scoring rule.

Model	$\text{Brier}_{\text{orig}}$	Brier_{TS}	AUROC
LLaDA-8B (DLLM)	0.299	0.212	0.826
Qwen2.5-7B (AR)	0.172	0.158	0.611

Table A10.1: Brier scores on GSM8K before and after temperature scaling. Lower is better; both models benefit from TS while preserving AUROC.

A11 ECE Estimator Robustness on GSM8K

ECE depends on the choice of binning scheme. To test whether our conclusions are artifacts of a particular estimator, we recompute GSM8K ECE using equal-width bins with $B \in \{10, 15, 30\}$ as well as equal-mass (quantile) bins with B=15. Table A11.1 shows that the reported ECE values are unchanged under these variants for both LLaDA and Qwen, consistent with the fact that miscalibration is dominated by a large, global mis-scaling.

Model	EW-B=10	EW-B=15	EW-B=30	EM-B=15
LLaDA-8B (DLLM)	31.2	31.2	31.2	31.2
Qwen2.5-7B (AR)	11.7	11.7	11.7	11.7

Table A11.1: ECE robustness on GSM8K (full set). EW denotes equal-width bins and EM denotes equal-mass (quantile) bins. ECE is reported in %.

A12 AUROC Under Calibration

Global temperature scaling is strictly monotone and therefore preserves AUROC exactly. DATS is monotone within each bucket but can violate global monotonicity across bucket boundaries; AUROC invariance is therefore not guaranteed. Table A12.1 reports AUROC before and after calibration on GSM8K under the 30/70 split protocol (seed 0). We observe no AUROC change for global temperature scaling, and only a small AUROC change for DATS.

Dataset	Model	$\text{AUROC}_{\text{orig}}$	AUROC_{TS}	$\text{AUROC}_{\text{DATS}}$	ΔAUROC
GSM8K	LLaDA-8B	0.8260	0.8260	0.8260	+0.0000
GSM8K	Qwen2.5-7B	0.6224	0.6224	0.6261	+0.0037

Table A12.1: AUROC before and after calibration on GSM8K (30% calibration, 70% test, seed 0). Global temperature scaling preserves AUROC exactly, while DATS induces a small AUROC change due to cross-bucket effects.

A13 Additional Split Calibration Results

For non-GSM8K benchmarks we use the same 30/70 calibration protocol as on GSM8K: temperatures are fit on a 30% calibration split and evaluated on the remaining 70% test split (seed 0). Table A13.1 reports results for LLaDA-8B on MATH-500, GSM-Hard, and TriviaQA. In all three cases, TS substantially reduces ECE while leaving AUROC unchanged, suggesting that the calibration gains observed on GSM8K carry over to other tasks.

A14 Full-Set Calibration Diagnostics

Table A14.1 reports calibration results when temperatures are fit and evaluated on the full evaluation sets. These numbers provide optimistic upper bounds on achievable ECE reductions and are useful for comparison with prior work that does not

Dataset	ECE_{orig}	ECE_{TS}	$\text{AUROC}_{\text{orig}}$	AUROC_{TS}
MATH-500	73.7	42.5	0.848	0.848
GSM-Hard	67.6	33.2	0.684	0.684
TriviaQA	51.0	17.6	0.727	0.727

Table A13.1: Additional split calibration results for LLaDA-8B (DLLM). Temperatures are fit on 30% of the data and ECE/AUROC are reported on the held-out 70% test split. TS reduces ECE substantially while leaving AUROC unchanged, confirming that the calibration gains observed on GSM8K generalize to other tasks.

use a held-out calibration split. Our main claims in the paper are based on the 30/70 split protocol (Tables 5 and A13.1).

Dataset	ECE_{orig}	ECE_{TS}	ECE_{DATS}	AUROC
GSM8K	31.2	11.5	9.6	0.826
MATH-500	73.5	41.7	41.7	0.836
GSM-Hard	68.2	33.6	33.6	0.690
TriviaQA	51.0	17.2	17.7	0.729

Table A14.1: Full-set calibration diagnostics for LLaDA-8B. Temperatures and DATS buckets are fit on the full evaluation sets and evaluated on the same data. These are diagnostic upper bounds; our main claims use the 30/70 split protocol.

A15 Difficulty Bin Operationalization

For the difficulty-conditional analysis in Section 5, we use a simple operation-count proxy computed from the GSM8K ground-truth solution annotation. For each example, we count occurrences of arithmetic symbols ($+$, $-$, $\times$, $\div$) appearing in the annotated solution text. This proxy is deterministic, requires no model-generated reasoning, and is reproducible from the public GSM8K annotations.

Using this proxy, we partition the 1,319 GSM8K examples into four bins: Easy (0–4 ops, n=279), Medium (5–7 ops, n=414), Hard (8–10 ops, n=355), and Hardest (11+ ops, n=271). We release the exact implementation and the derived per-example counts with the code.

A16 Step-Wise Discrimination Protocol

For the step-wise diffusion analysis in Section 7, we compute per-step AUROC as follows. At each diffusion step $t \in \{0, 1, \ldots, T-1\}$, we record the mean token probability over currently unmasked positions. We then compute AUROC by treating this per-step confidence as the score and the final correctness label as the binary target. On GSM8K, early steps (0–41) yield mean AUROC 0.519 (near random), middle steps (42–84) yield 0.566, and late steps (85–127) yield 0.593, despite average confidence remaining saturated (> 0.99) throughout. This indicates that discriminative signal emerges gradually and concentrates in the final denoising phase when the model must commit to globally consistent token sequences.

A17 DLLM Scoring-Choice Sensitivity

Our main DLLM confidence is computed from a final teacher-forced forward pass on the fully denoised sequence (Section 3). Since diffusion sampling also records a per-step confidence trace, we evaluate simple path-derived alternatives that aggregate this trace. Table A17.1 shows that step-aggregated path scores are highly saturated (mean ≈ 0.999) and yield substantially lower AUROC than the final forward score. This motivates reporting step-wise AUROC as a diagnostic of when discrimination emerges, while using the final forward score as the primary single-pass confidence for downstream ranking and calibration.

DLLM score (GSM8K)	AUROC	ECE (%)
Final forward (paper default)	0.826	31.2
Path: last step	0.547	37.1
Path: mean over all steps	0.724	37.0
Path: mean over last 8 steps	0.731	37.0

Table A17.1: Sensitivity to DLLM scoring choice on GSM8K (full set). "Path" scores are computed from the per-step confidence trace recorded during diffusion sampling.

A18 Self-Consistency Baseline for AR Models

To address the concern that logit-based AR confidence may underestimate the discriminative potential of autoregressive models, we evaluate a self-consistency baseline (Wang et al., 2023) on GSM8K using Qwen2.5-7B-Instruct. For each question, we sample K=5 chains of thought with temperature 0.7, extract the final numeric answer from each sample, and define self-consistency confidence as the fraction of samples agreeing with the majority-vote answer. Correctness is determined by whether the majority answer matches the ground truth.

Table A18.1 shows that self-consistency strengthens AR discrimination (AUROC 0.863) but requires 5$\times$ sampling. In contrast, LLaDA attains AUROC 0.826 in a single pass, highlighting a cost–discrimination trade-off between multi-sample AR

Method	AUROC	Accuracy	Cost
AR logit-based (Qwen)	0.611	79.8%	1×
DLLM (LLaDA)	0.826	62.9%	1×
AR self-consistency (K=5)	0.863	89.4%	5×

Table A18.1: Self-consistency baseline on GSM8K. With K=5 samples, AR self-consistency achieves AUROC 0.863 at 5× decoding cost. For reference, single-pass token-probability confidence yields AUROC 0.611 for Qwen and 0.826 for LLaDA at 1× cost.

confidence and single-pass diffusion confidence.

A19 Alternative Self-Consistency Scores

Self-consistency produces a distribution over K sampled answers. Beyond the majority-agreement ratio $p_{\max}$, one can derive confidence from predictive entropy or from the margin between the most frequent and second-most frequent answers. Table A19.1 reports these variants on GSM8K for Qwen2.5-7B-Instruct using the same K=5 samples and majority-vote correctness labels. All variants yield similar AUROC (0.863–0.865), suggesting that the gain is driven primarily by multi-sample agreement rather than a specific functional form.

Score	AUROC
$p_{\max}$ (agreement ratio)	0.863
$1 - H/\log K$ (entropy confidence)	0.864
$p_{\max} - p_2$ (margin)	0.865

Table A19.1: Alternative self-consistency confidence scores on GSM8K (Qwen2.5-7B, K=5). All agreement-derived scores yield similar AUROC.

A20 Probe Scaling: n=100 vs. n=500

To address concerns that our intervention probes use small sample sizes, we repeat the arithmetic and TriviaQA swap probes with n=500 pairs. Table A20.1 shows that the qualitative contrast persists: arithmetic swaps induce a large confidence drop, while TriviaQA answer swaps induce a much smaller drop and weaker discrimination.

A21 DATS Monotonicity Across Seeds

DATS is monotone within each bucket but not necessarily globally monotone across bucket boundaries. We therefore quantify the induced rank inversions and AUROC changes under the 30/70 split protocol across five random seeds (0–4). Table A21.1 shows that for LLaDA with K=5, DATS is nearly globally monotone on the test split

Probe	Model	n	p_{correct}	p_{wrong}	Gap	AUROC
Arithmetic	LLaDA	100	98.8	70.6	28.2	1.000
Arithmetic	LLaDA	500	98.7	73.1	25.6	0.996
TriviaQA swap	LLaDA	100	89.0	85.6	3.4	0.603
TriviaQA swap	LLaDA	500	88.8	86.6	2.2	0.571
TriviaQA swap	Qwen2.5-7B	100	30.3	14.1	16.2	0.702
TriviaQA swap	Qwen2.5-7B	500	31.3	13.5	17.9	0.734
Arithmetic	Qwen2.5-7B	100	97.1	56.8	40.3	1.000
Arithmetic	Qwen2.5-7B	500	97.8	57.3	40.5	1.000

Table A20.1: Scaled intervention probes. Probabilities are means over answer/result tokens (in %).

(inversion fraction $\approx 10^{-6}$) and AUROC changes are negligible. For larger K, and for Qwen, global non-monotonicity becomes more pronounced and AUROC can change accordingly. These results motivate reporting AUROC deltas explicitly and using global temperature scaling when strict rank preservation is required.

Model	K	ΔAUROC (mean±std)	Inversion frac. (mean)
LLaDA	5	$+0.0000 \pm 0.0000$	2.8×10^{-6}
LLaDA	10	-0.0080 ± 0.0054	6.3×10^{-2}
Qwen2.5-7B	5	-0.0160 ± 0.0160	1.2×10^{-1}

Table A21.1: DATS rank effects on GSM8K under the 30/70 split protocol, aggregated over seeds 0–4. ΔAUROC is measured relative to the uncalibrated confidence on the same test split.

A22 Length-Stratified GSM8K Analysis

To test whether the observed discrimination differences are driven by output length, we bin GSM8K examples by the generated output length (word-count proxy) and recompute AUROC and ECE within each bin. Table A22.1 shows that LLaDA retains strong discrimination across bins, while Qwen's AUROC decreases on the longest traces as accuracy drops and calibration degrades.

Model	Length bin	n	Mean len.	Acc (%)	ECE (%)	AUROC
LLaDA	9–155	440	116.5	68.9	24.1	0.852
LLaDA	155–224	440	187.4	69.8	26.9	0.762
LLaDA	224–510	439	293.4	50.1	42.6	0.811
Qwen2.5-7B	61–155	440	124.9	91.4	0.8	0.665
Qwen2.5-7B	155–216	440	182.1	85.2	6.9	0.647
Qwen2.5-7B	216–410	439	286.6	62.9	28.4	0.585

Table A22.1: Length-stratified GSM8K results using a word-count proxy for generated output length.