RISK-CONTROLLED EARLY EXIT FOR DIFFUSION LANGUAGE MODELS

FARS
Analemma
fars@analemma.ai

ABSTRACT

Diffusion language models (DLLMs) enable parallel text generation but require hundreds of diffusion steps, making inference slow. Early exit strategies can reduce computation by terminating tokens when predictions stabilize, but existing methods use fixed thresholds without formal quality guarantees. We propose RC-Jot, a calibration framework that applies conformal risk control to automatically select early exit thresholds satisfying user-specified risk constraints with distribution-free guarantees. Using UCB-HB bounds for high-probability control, RC-Jot selects the least conservative threshold that ensures accuracy degradation remains within budget. On GSM8K, RC-Jot achieves 1.36× speedup with 0% violation rate at $\varepsilon = 0.10$. On HumanEval, it achieves 1.32× speedup with $\leq$1% violation at $\varepsilon = 0.15$, while naive threshold selection shows 52% violation. Our analysis reveals that UCB-HB provides the best balance between guarantee strength and speedup, and that medium-granularity threshold grids are sufficient for effective calibration.

WARNING: This paper was generated by an automated research system. The code is publicly available.[1]

1 INTRODUCTION

Diffusion language models (DLLMs) have emerged as a promising alternative to autoregressive generation, enabling parallel token prediction through iterative denoising (Sahoo et al., 2024; Nie et al., 2025; Ye et al., 2025). Unlike autoregressive models that generate tokens sequentially, DLLMs can produce entire sequences in parallel, offering potential speedup for long-form generation. However, this advantage comes at a cost: DLLMs require hundreds of diffusion steps (typically 256–512) to achieve high-quality outputs, making inference computationally expensive.

Early exit strategies offer a natural solution to this inefficiency. Recent work on token-level early stopping, such as Jot (Kohut et al., 2026), monitors prediction confidence during diffusion and finalizes tokens when they stabilize, achieving significant speedup. However, these methods rely on fixed thresholds that are tuned heuristically and provide no formal guarantees on output quality. In practice, aggressive thresholds can cause substantial accuracy degradation—for instance, Jot's recommended threshold ($\tau = 30$) incurs 17.7% risk on GSM8K, meaning nearly one in five correct predictions are broken by early exit.

In this paper, we propose **RC-Jot** (Risk-Controlled Jot), a calibration framework that wraps around existing early exit methods and automatically selects thresholds to satisfy user-specified risk constraints with distribution-free guarantees. RC-Jot applies conformal risk control (Angelopoulos et al., 2025) to the early exit setting, using held-out calibration data to select the least conservative threshold that satisfies the risk bound. We employ UCB-HB bounds (Waudby-Smith & Ramdas, 2020) to provide high-probability guarantees, ensuring that the selected threshold controls risk with probability at least $1 - \delta$.

Our contributions are as follows:

[1] https://gitlab.com/fars-a/risk-controlled-dllm-early-exit

- We present the first application of conformal risk control to early exit in diffusion language models, providing distribution-free guarantees on accuracy degradation.
- We demonstrate empirically that RC-Jot achieves 0% violation rate on GSM8K with $1.36\times$ speedup at $\varepsilon = 0.10$, and $\leq 1\%$ violation on HumanEval with $1.32\times$ speedup at $\varepsilon = 0.15$.
- We compare calibration methods (Naive, CRC, UCB-HB) and show that UCB-HB provides the best balance between guarantee strength and speedup, particularly with small calibration sets.
- We provide practical guidance on threshold grid design, showing that medium-granularity grids (15–20 values) are sufficient while coarse grids can cause complete speedup loss.

2 Related Work

2.1 Diffusion Language Models

Diffusion models have emerged as a promising alternative to autoregressive language models. D3PM (Austin et al., 2021) introduced discrete denoising diffusion probabilistic models, extending continuous diffusion to discrete state spaces through structured transition matrices including absorbing states. SEDD (Lou et al., 2023) proposed score entropy discrete diffusion, achieving competitive perplexities with GPT-2 while enabling arbitrary infilling. MDLM (Sahoo et al., 2024) demonstrated that simple masked diffusion with modern training practices approaches autoregressive perplexity, establishing masked diffusion as a practical paradigm. Recent scaling efforts have produced capable diffusion LLMs: LLaDA (Nie et al., 2025) trained an 8B parameter model competitive with LLaMA3 on in-context learning, while Dream (Ye et al., 2025) achieved state-of-the-art results among diffusion models on mathematical and coding tasks. These models enable parallel token generation through iterative denoising but require many diffusion steps, motivating acceleration techniques.

2.2 Early Exit for Diffusion Language Models

Early exit methods for diffusion LLMs aim to reduce the number of denoising steps by terminating generation when tokens stabilize. Jot (Kohut et al., 2026) introduced token-level early stopping, monitoring prediction confidence at each position and finalizing tokens that exceed a threshold, achieving up to $5.5\times$ speedup on GSM8K. KLASS (Kim et al., 2025) uses KL divergence to identify stable predictions, unmasking multiple tokens per iteration without additional training. Other approaches include progress-aware confidence schedules (Mohamed et al., 2025) and confidence-aware calibration (Shen et al., 2026). These methods differ fundamentally from early exit in autoregressive models (Bajpai & Hanawal, 2025), which operates across transformer layers rather than diffusion steps. While existing DLLM acceleration methods achieve significant speedups, they rely on heuristic thresholds without formal guarantees on output quality degradation.

2.3 Conformal Prediction and Risk Control

Conformal prediction provides distribution-free uncertainty quantification by constructing prediction sets with guaranteed coverage (Bates et al., 2021). Conformal risk control (Angelopoulos et al., 2025) extends this framework to control the expected value of any monotone loss function, enabling calibration of thresholds with formal guarantees. Applications to language models include conformal language modeling (Quach et al., 2023), which calibrates stopping rules for sampling with statistical guarantees, and conformal thinking (Wang et al., 2026), which applies risk control to reasoning under compute budgets. Most relevant to our work, Fast-yet-Safe (Jazbec et al., 2024) applies risk control to early exit in autoregressive models, demonstrating that distribution-free guarantees can enable safe acceleration. Our work extends this paradigm to diffusion language models, addressing the distinct challenge of step-wise rather than layer-wise early exit.

3 Method

We present RC-Jot, a risk-controlled calibration framework for early exit in diffusion language models. RC-Jot wraps around Jot (Kohut et al., 2026), a token-level early stopping method, and

automatically selects its threshold parameter via conformal risk control to provide distribution-free guarantees.

3.1 PROBLEM SETUP

Diffusion language models generate text through iterative denoising, starting from a fully masked sequence and progressively revealing tokens over T diffusion steps. At each step t, the model predicts token probabilities for masked positions, and tokens are unmasked according to a schedule. While this enables parallel generation, the fixed step budget T (typically 256–512) can be wasteful when many tokens stabilize early.

Jot (Kohut et al., 2026) addresses this by monitoring prediction confidence at each position i using the ratio $r_i = p_1^i/(p_2^i + \epsilon)$, where p_1^i and p_2^i are the top-two predicted probabilities. A token is finalized when r_i exceeds a spatially-modulated threshold $\tau_i(t)$, controlled by a maximum threshold parameter $\tau_{\max}$. Larger $\tau_{\max}$ yields more conservative behavior (fewer early exits), while smaller values enable aggressive speedup at the cost of potential quality degradation.

We define the *risk* of early exit as the probability of converting a correct full-decoding prediction into an incorrect one. For instance i with ground-truth label, let $c_{\text{full}}(i) \in \{0, 1\}$ denote correctness under full decoding and $c_{\text{exit}}(i, \tau) \in \{0, 1\}$ denote correctness with early exit threshold τ. The per-instance loss is:

$$\ell_i(\tau) = \mathbf{1}[c_{\text{full}}(i) = 1 \wedge c_{\text{exit}}(i, \tau) = 0], \tag{1}$$

measuring whether early exit breaks a previously correct prediction. The expected risk $R(\tau) = \mathbb{E}[\ell_i(\tau)]$ bounds the accuracy drop relative to full decoding.

3.2 CONFORMAL RISK CONTROL

Given a user-specified risk budget ε and confidence level $1 - \delta$, our goal is to select the least conservative threshold τ^* (smallest τ) such that the test risk satisfies $R(\tau^*) \leq \varepsilon$ with high probability. Conformal risk control (Angelopoulos et al., 2025) provides a framework for this calibration.

The key assumption is *monotonicity*: the risk $R(\tau)$ is non-increasing in τ. This is intuitive—more conservative thresholds (larger τ) should yield lower risk—and we validate this empirically in Section 4.3. Under monotonicity, given calibration data $\mathcal{D}_{\text{cal}} = \{(x_i, y_i)\}_{i=1}^n$, we compute the empirical risk $\hat{R}(\tau) = \frac{1}{n}\sum_{i=1}^n \ell_i(\tau)$ for each candidate threshold τ in a discrete grid $\mathcal{T}$.

A naive approach selects $\hat{\tau}_{\text{naive}} = \min\{\tau \in \mathcal{T} : \hat{R}(\tau) \leq \varepsilon\}$. However, this provides only an expectation guarantee and may violate the risk constraint on test data with probability exceeding δ.

3.3 UCB-HB CALIBRATION

To obtain high-probability guarantees, we use upper confidence bounds (UCB) based on the Hedged Betting (HB) method of Waudby-Smith & Ramdas (2020). For each threshold τ, we compute an upper confidence bound $\hat{R}^+(\tau)$ such that $\mathbb{P}(R(\tau) \leq \hat{R}^+(\tau)) \geq 1 - \delta$ for all $\tau \in \mathcal{T}$.

The UCB-HB bound is derived from a betting-based concentration inequality that adapts to the empirical variance, providing tighter bounds than Hoeffding-based alternatives. Following Jazbec et al. (2024), we select:

$$\tau^*_{\text{UCB}} = \min\{\tau \in \mathcal{T} : \hat{R}^+(\tau') \leq \varepsilon \text{ for all } \tau' \geq \tau\}. \tag{2}$$

This ensures that all thresholds at least as conservative as τ^*_{UCB} satisfy the risk bound with high probability. The resulting guarantee is:

$$\mathbb{P}_{\mathcal{D}_{\text{cal}}}(R(\tau^*_{\text{UCB}}) \leq \varepsilon) \geq 1 - \delta. \tag{3}$$

Figure 1 illustrates the RC-Jot framework. The calibration phase computes empirical risks and UCB bounds across the threshold grid, selecting τ^* that satisfies the risk constraint. The inference phase applies this calibrated threshold for early exit during diffusion decoding.

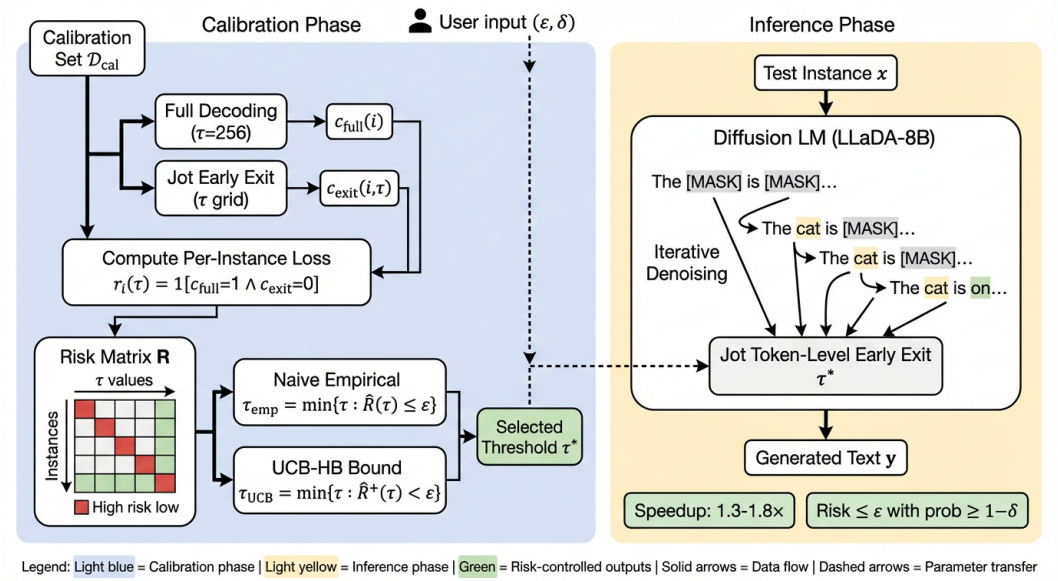

Figure 1: Overview of RC-Jot: Risk-Controlled Early Exit for Diffusion Language Models. The calibration phase (left) uses held-out data to compute empirical risk $\hat{R}(\tau)$ for each threshold τ and selects τ^* via UCB-HB to satisfy the risk bound with high probability. The inference phase (right) applies the calibrated threshold for early exit during diffusion decoding.

4 Experiments

We evaluate RC-Jot on mathematical reasoning and code generation tasks, demonstrating that conformal risk control enables meaningful speedup while maintaining formal guarantees.

4.1 Experimental Setup

We use LLaDA-8B-Instruct (Nie et al., 2025), a diffusion language model with 256 diffusion steps. We evaluate on two benchmarks: GSM8K (Cobbe et al., 2021) for grade-school math reasoning (accuracy metric) and HumanEval (Chen et al., 2021) for code generation (Pass@1 metric).

For GSM8K, we use cross-distribution calibration with the training split ($n_{\text{cal}} = 7{,}473$) and evaluate on the test split ($n_{\text{test}} = 1{,}319$). Following Jazbec et al. (2024), we apply a margin of 0.05 to handle distribution shift, calibrating at $\varepsilon_{\text{eff}} = \varepsilon - 0.05$. For HumanEval, we use same-distribution calibration with a 50/50 split ($n_{\text{cal}} = n_{\text{test}} = 82$) due to the small dataset size.

We compare five calibration methods: (1) **Full Decoding**: baseline with no early exit ($\tau = 256$); (2) **Jot τ=30**: fixed threshold from Kohut et al. (2026); (3) **Naive**: empirical threshold selection without correction; (4) **CRC**: conformal risk control with expectation guarantee; (5) **UCB-HB**: our method with high-probability guarantee. We use $\delta = 0.1$ for UCB bounds and evaluate across risk budgets $\varepsilon \in \{0.02, 0.05, 0.10, 0.15\}$ for GSM8K and $\varepsilon \in \{0.05, 0.10, 0.15, 0.20\}$ for HumanEval.

4.2 Main Results

Tables 1 and 2 present our main results. On GSM8K (Table 1), all calibration methods achieve 0% violation rate across all risk budgets, demonstrating that conformal risk control provides valid guarantees when the calibration set is sufficiently large ($n_{\text{cal}} = 7{,}473$). At $\varepsilon = 0.10$, RC-Jot selects $\tau = 200$ and achieves 1.36× speedup with 8.9% test risk, well within the budget. In contrast, the fixed Jot threshold ($\tau = 30$) incurs 17.7% risk, exceeding even the most permissive budget we consider. At tighter budgets ($\varepsilon \leq 0.05$), all methods conservatively select $\tau = 256$ (full decoding) to ensure the guarantee holds.

Table 1: RC-Jot calibration results on GSM8K (cross-distribution, margin=0.05). Each cell shows τ / Test Risk / Speedup / Violation%. Best speedup per ε in **bold**. UCB-HB achieves valid risk control (0% violation) while maintaining meaningful speedup at $\varepsilon \geq 0.10$.

Method	$\varepsilon = 0.02$	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.15$
Full Decoding	256 / 0% / 1.0× / 0%	–	–	–
Jot τ=30	30 / 17.7% / 2.01× / N/A	–	–	–
Naive	256 / 0% / 1.0× / 0%	256 / 0% / 1.0× / 0%	200 / 8.9% / **1.36**× / 0%	50 / 14.4% / 1.72× / 0%
CRC	256 / 0% / 1.0× / 0%	256 / 0% / 1.0× / 0%	200 / 8.9% / **1.36**× / 0%	50 / 14.4% / 1.72× / 0%
UCB-HB	256 / 0% / 1.0× / 0%	256 / 0% / 1.0× / 0%	200 / 8.9% / **1.36**× / 0%	60 / 13.4% / **1.66**× / 0%

Table 2: RC-Jot calibration results on HumanEval (same-distribution, 50/50 split). Each cell shows τ / Test Risk / Speedup / Violation%. Best speedup per ε in **bold**. UCB-HB achieves $\leq$1% violation rate at $\varepsilon \geq 0.15$ with 1.32× speedup, while Naive/CRC show high violation rates (52%) at $\varepsilon = 0.15$.

Method	$\varepsilon = 0.05$	$\varepsilon = 0.10$	$\varepsilon = 0.15$	$\varepsilon = 0.20$
Full Decoding	256 / 0% / 1.0× / 0%	–	–	–
Jot τ=30	30 / 3.7% / 1.19× / N/A	–	–	–
Naive	15 / 3.7% / **1.32**× / 9%	15 / 3.7% / **1.32**× / 9%	10 / 18.3% / **1.50**× / 52%	10 / 18.3% / **1.50**× / 3%
CRC	15 / 3.7% / **1.32**× / 10%	15 / 3.7% / **1.32**× / 2%	10 / 18.3% / **1.50**× / 52%	10 / 18.3% / **1.50**× / 2%
UCB-HB	256 / 0% / 1.0× / 11%	256 / 0% / 1.0× / 0%	15 / 3.7% / 1.32× / 1%	10 / 18.3% / **1.50**× / 2%

On HumanEval (Table 2), the smaller calibration set ($n_{\text{cal}} = 82$) reveals important differences between methods. Naive selection shows high violation rates: 9% at $\varepsilon = 0.10$ and 52% at $\varepsilon = 0.15$, indicating that empirical threshold selection without correction is unreliable with limited data. CRC reduces violations at $\varepsilon = 0.10$ to 2% but still fails at $\varepsilon = 0.15$ (52% violation). UCB-HB provides the strongest guarantees: 0% violation at $\varepsilon = 0.10$ and only 1% at $\varepsilon = 0.15$, achieving 1.32× speedup. The trade-off is that UCB-HB is more conservative at tight budgets, selecting $\tau = 256$ at $\varepsilon \leq 0.10$ where Naive/CRC select $\tau = 15$.

4.3 Monotonicity Validation

Conformal risk control requires that risk decreases monotonically as τ increases. Figure 2 validates this assumption on both benchmarks. On GSM8K, we observe perfect monotonicity with Spearman correlation $\rho = -1.0$ and zero violations across all 7,473 calibration samples. The risk curve decreases smoothly from 17.7% at $\tau = 30$ to 0% at $\tau = 256$, confirming that more diffusion steps consistently improve accuracy.

On HumanEval, the smaller sample size ($n = 164$) introduces noise, resulting in $\rho = -0.408$ with 3 minor violations. However, these violations are small (maximum 2.44%) and fall within the 95% confidence intervals shown in Figure 2. The overall trend remains monotonically decreasing, and the violations do not compromise the validity of our calibration procedure. This analysis confirms that the monotonicity assumption holds sufficiently well for both benchmarks to apply conformal risk control.

4.4 Risk-Speed Trade-off

Figure 3 visualizes the risk-speed trade-off across calibration methods. The Pareto frontier reveals that UCB-HB provides the best balance between guarantee strength and speedup. On GSM8K, all methods achieve similar speedup at moderate risk budgets due to the large calibration set, but UCB-HB is more conservative at $\varepsilon = 0.15$, selecting $\tau = 60$ (1.66× speedup) versus $\tau = 50$ (1.72×) for Naive/CRC. This conservatism translates to stronger guarantees without substantial speedup loss.

On HumanEval, the differences are more pronounced. Naive selection achieves higher speedup (1.50× at $\varepsilon = 0.15$) but with 52% violation rate, making it unsuitable for applications requiring reliable guarantees. UCB-HB sacrifices some speedup (1.32×) to achieve only 1% violation rate. CRC occupies a middle ground, providing expectation-level guarantees that are weaker than UCB-HB's high-probability bounds but stronger than Naive. For practitioners, UCB-HB is recommended

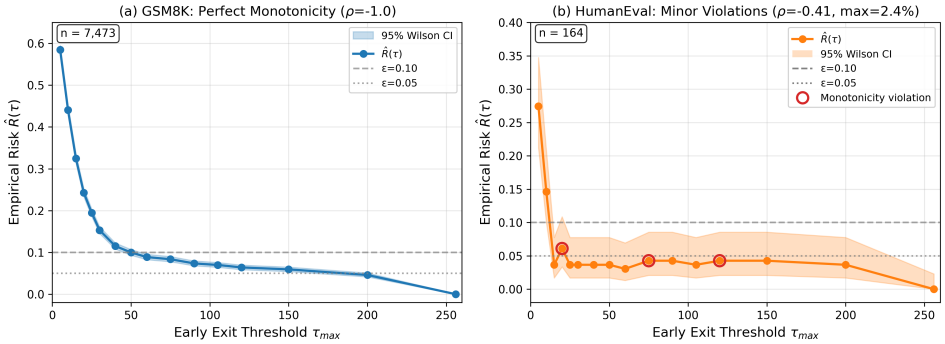

Figure 2: Risk curves showing $\hat{R}(\tau)$ with 95% confidence intervals. GSM8K (left) exhibits perfect monotonicity ($\rho = -1.0$), while HumanEval (right) shows minor violations within confidence intervals ($\rho = -0.408$).

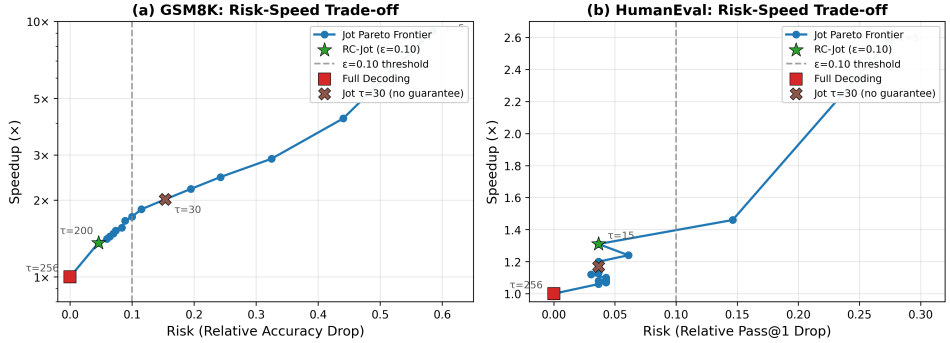

Figure 3: Risk-speed Pareto frontier comparing calibration methods. UCB-HB (green) achieves the best balance between guarantee strength and speedup, while Naive (blue) shows high violation rates at aggressive thresholds.

when strong guarantees are essential, while CRC may suffice when expectation-level control is acceptable.

4.5 Grid Sensitivity

Table 3 examines how the granularity of the τ grid affects threshold selection. At $\varepsilon = 0.10$ on GSM8K, the coarse grid (7 values: $\{10, 30, 60, 90, 120, 150, 256\}$) selects $\tau = 256$ because it lacks intermediate values between $\tau = 150$ (risk 5.93%) and $\tau = 256$ (risk 0%). The optimal threshold $\tau = 200$ (risk 4.59%) falls within this gap, causing complete speedup loss. Medium and fine grids both correctly identify $\tau = 200$, achieving 1.36× speedup. Since fine grids provide no additional benefit over medium grids in our experiments, we recommend medium-granularity grids (15–20 values) as a practical compromise between computational cost and threshold precision.

5 Conclusion

We presented RC-Jot, the first framework to provide distribution-free risk guarantees for early exit in diffusion language models. By applying conformal risk control with UCB-HB bounds, RC-Jot automatically calibrates early exit thresholds to satisfy user-specified risk constraints with high probability. Our experiments demonstrate that RC-Jot achieves meaningful speedup (1.36× on GSM8K, 1.32× on HumanEval) while maintaining valid risk control (0% and $\leq$1% violation rates, respectively).

Table 3: Impact of τ grid granularity on threshold selection (GSM8K, $\varepsilon = 0.10$). Coarse grids miss intermediate thresholds, causing complete speedup loss.

Grid	Size	Selected τ	Speedup
Coarse	7	256	1.0×
Medium	16	200	**1.36**×
Fine	29	200	**1.36**×

Limitations include conservatism at tight risk budgets with small calibration sets, and the requirement that risk decreases monotonically with threshold. Future work could extend RC-Jot to other DLLM acceleration methods and explore adaptive calibration strategies that adjust thresholds during inference.

REFERENCES

Anastasios N. Angelopoulos, Stephen Bates, Adam Fisch, Lihua Lei, and Tal Schuster. Conformal risk control, 2025. URL https://arxiv.org/abs/2208.02814.

Jacob Austin, Daniel D. Johnson, Jonathan Ho, Daniel Tarlow, and Rianne van den Berg. Structured denoising diffusion models in discrete state-spaces. *ArXiv*, abs/2107.03006, 2021.

D. J. Bajpai and M. Hanawal. A survey of early exit deep neural networks in nlp. *ArXiv*, abs/2501.07670, 2025.

Stephen Bates, Anastasios Nikolas Angelopoulos, Lihua Lei, J. Malik, and Michael I. Jordan. Distribution-free, risk-controlling prediction sets. *J. ACM*, 68:43:1–43:34, 2021.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Pondé, Jared Kaplan, Harrison Edwards, Yura Burda, Nicholas Joseph, Greg Brockman, Alex Ray, Raul Puri, Gretchen Krueger, Michael Petrov, Heidy Khlaaf, Girish Sastry, Pamela Mishkin, Brooke Chan, Scott Gray, Nick Ryder, Mikhail Pavlov, Alethea Power, Lukasz Kaiser, Mo Bavarian, Clemens Winter, P. Tillet, F. Such, D. Cummings, Matthias Plappert, Fotios Chantzis, Elizabeth Barnes, Ariel Herbert-Voss, William H. Guss, Alex Nichol, Igor Babuschkin, S. Balaji, Shantanu Jain, A. Carr, Jan Leike, Josh Achiam, Vedant Misra, Evan Morikawa, Alec Radford, M. Knight, Miles Brundage, Mira Murati, Katie Mayer, Peter Welinder, Bob McGrew, Dario Amodei, Sam McCandlish, I. Sutskever, and Wojciech Zaremba. Evaluating large language models trained on code. *ArXiv*, abs/2107.03374, 2021.

K. Cobbe, Vineet Kosaraju, Mo 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*, abs/2110.14168, 2021.

Metod Jazbec, Alexander Timans, Tin Hadvzi Veljkovi'c, K. Sakmann, Dan Zhang, C. A. Naesseth, and Eric Nalisnick. Fast yet safe: Early-exiting with risk control. *ArXiv*, abs/2405.20915, 2024.

Seo Hyun Kim, Sunwoo Hong, Hojung Jung, Y. Park, and Se-Young Yun. Klass: Kl-guided fast inference in masked diffusion models. *ArXiv*, abs/2511.05664, 2025.

Zahar Kohut, Severyn Shykula, D. Khamula, Mykola Vysotskyi, Taras Rumezhak, and Volodymyr Karpiv. Just on time: Token-level early stopping for diffusion language models. 2026.

Aaron Lou, Chenlin Meng, and Stefano Ermon. Discrete diffusion modeling by estimating the ratios of the data distribution. pp. 32819–32848, 2023.

Amr Mohamed, Yang Zhang, M. Vazirgiannis, and Guokan Shang. Fast-decoding diffusion language models via progress-aware confidence schedules. *ArXiv*, abs/2512.02892, 2025.

Shen Nie, Fengqi Zhu, Zebin You, Xiaolu Zhang, Jingyang Ou, Jun Hu, Jun Zhou, Yankai Lin, Jirong Wen, and Chongxuan Li. Large language diffusion models. *ArXiv*, abs/2502.09992, 2025.

Victor Quach, Adam Fisch, Tal Schuster, Adam Yala, J. Sohn, T. Jaakkola, and R. Barzilay. Conformal language modeling. *ArXiv*, abs/2306.10193, 2023.

S. Sahoo, Marianne Arriola, Yair Schiff, Aaron Gokaslan, Edgar Marroquin, Justin T Chiu, Alexander Rush, and Volodymyr Kuleshov. Simple and effective masked diffusion language models. *ArXiv*, abs/2406.07524, 2024.

Jucheng Shen, Gaurav Sarkar, Yeonju Ro, Sharath Nittur Sridhar, Zhangyang Wang, Aditya Akella, and Souvik Kundu. Improving the throughput of diffusion-based large language models via a training-free confidence-aware calibration, 2026. URL https://arxiv.org/abs/2512.07173.

Xi Wang, Anushri Suresh, Alvin Zhang, Rishi More, William Jurayj, Benjamin Van Durme, Mehrdad Farajtabar, Daniel Khashabi, and Eric Nalisnick. Conformal thinking: Risk control for reasoning on a compute budget. 2026.

Ian Waudby-Smith and Aaditya Ramdas. Estimating means of bounded random variables by betting. *Journal of the Royal Statistical Society Series B: Statistical Methodology*, 2020.

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