

# LLaDA (Large Language Diffusion Models)

拡散モデルによる大規模言語モデルの新パラダイム

後藤 健一郎

## 目次

問題設定	1
新規性	2
核となるアイデア	2
メカニズム: 前方過程 (Forward Process)	2
メカニズム: 逆過程 (Reverse Process)	3
メカニズム: 目的関数 (ELBO)	3
メカニズム: アーキテクチャ	3
実験結果: スケーラビリティ	4
実験結果: 事前学習モデルのベンチマーク	5
実験結果: 事後学習モデルのベンチマーク	6
実験結果: 総合比較	7
実験結果: 頑健性 (Reversal タスク)	7
実験結果: サンプリング過程の可視化	8
制限事項	8
まとめ	9
参考情報	9

## 問題設定

- 現在の大規模言語モデル (LLM) は自己回帰モデル (ARM) が主流
  - 左から右へトークンを逐次生成する
  - 双方向的な文脈の活用や、テキスト全体の整合性確保が構造的に困難
- テキスト生成に拡散モデルを適用する試みはあったが、スケーラビリティに課題
  - 連続拡散モデル (MDLM 等) は ARM の約 64 倍の計算時間要する
  - 大規模化 (数十億パラメータ) での有効性が未検証だった

## 新規性

- 離散的なマスク拡散モデル (Masked Diffusion Model; MDM) を 8B パラメータまでスケール
    - 2.3T トークンで事前学習し、各種ベンチマークで LLaMA3 8B に匹敵する性能を達成
  - 拡散モデルの特性を活かした強みを実証
    - 双向性モーデリング: 文脈の前後を同時に考慮した生成が可能
    - 頑健性の向上: ARM が苦手とする逆順テキスト生成 (Reversal) タスクで GPT-4o や Qwen2.5 を上回る
  - MDM 向けの 事後学習 (SFT) 手法を提案し、指示追従能力を付与
- 

## 核となるアイデア

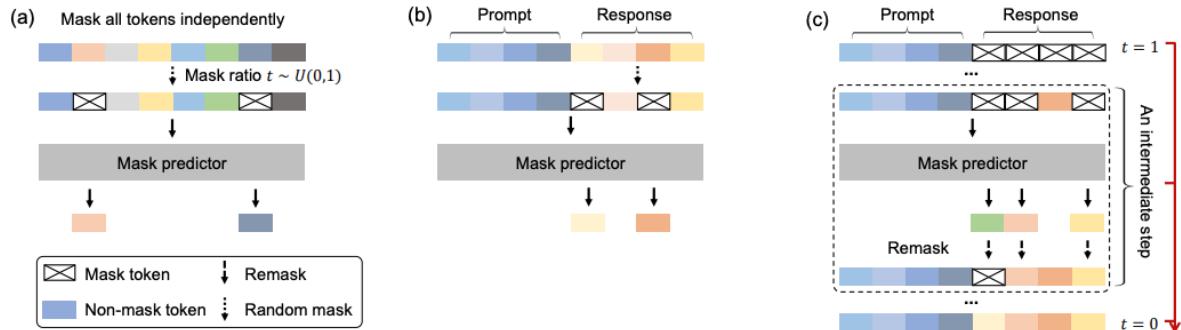


図 1: LLaDA の概要

- (a) 事前学習: トークン列をランダムにマスクし、マスクされたトークンを予測する
  - (b) 指示追従 (SFT): プロンプト部分はマスクせず、応答部分のみをマスクして学習
  - (c) 推論: 全トークンがマスクされた状態から、段階的にマスクを除去して文章を生成
- 

## メカニズム: 前方過程 (Forward Process)

- 元のトークン列  $\mathbf{x}_0$  に対し、マスク率  $t \sim U(0, 1)$  でランダムにマスクを適用
  - 各トークンは独立に確率  $t$  で [MASK] に置換される
    - $t = 0$ : 元のトークン列 (マスクなし)
    - $t = 1$ : 全トークンがマスク (完全ノイズ)
- 遷移確率:  $q(\mathbf{x}_t | \mathbf{x}_0) = \prod_i q(x_t^i | x_0^i)$ 
  - $q(x_t^i | x_0^i) = (1-t) \cdot \delta(x_t^i, x_0^i) + t \cdot \delta(x_t^i, [\text{MASK}])$

---

## メカニズム: 逆過程 (Reverse Process)

- ニューラルネットワーク  $p_\theta(\mathbf{x}_0|\mathbf{x}_t)$  を用いて、マスクされたトークンを予測
  - 推論時の手順:
    - $\mathbf{x}_1$  (全マスク)からスタート
    - 各ステップでマスク位置のトークンを予測
    - 予測の確信度が高いものから順にマスクを解除 (**remasking**)
    - $T$  ステップ後に完全なテキスト  $\mathbf{x}_0$  を得る
  - remasking により、低確信度の予測を後のステップで修正可能
- 

## メカニズム: 目的関数 (ELBO)

- 変分下界 (ELBO) を最大化:

$$\mathcal{L}_{\text{ELBO}} = \mathbb{E}_{t,q(\mathbf{x}_t|\mathbf{x}_0)} \left[ \frac{1}{t} \sum_{i:x_t^i=[\text{MASK}]} \log p_\theta(x_0^i|\mathbf{x}_t) \right]$$

- 各項の役割:
    - $\frac{1}{t}$ : マスク率が低い(ほぼ完成に近い)ステップほど重みが大きい
    - $\sum_{i:x_t^i=[\text{MASK}]}$ : マスクされた位置のみを予測対象とする
    - $\log p_\theta(x_0^i|\mathbf{x}_t)$ : 各マスク位置の交差エントロピー損失
  - ARM の次トークン予測損失の自然な拡張とみなせる
- 

## メカニズム: アーキテクチャ

- Transformer をバックボーンとして使用(LLaMA と同様の構造)
    - 双向 attention を採用(ARM の causal mask は使わない)
    - RMSNorm、SwiGLU、RoPEなどの標準的なコンポーネントを利用
  - モデルサイズ: 0.04B / 0.1B / 0.4B / 1.1B / **8B** の5段階で検証
  - 学習データ: 2.3T トークン (英語・中国語・コード)
-

## 実験結果: スケーラビリティ

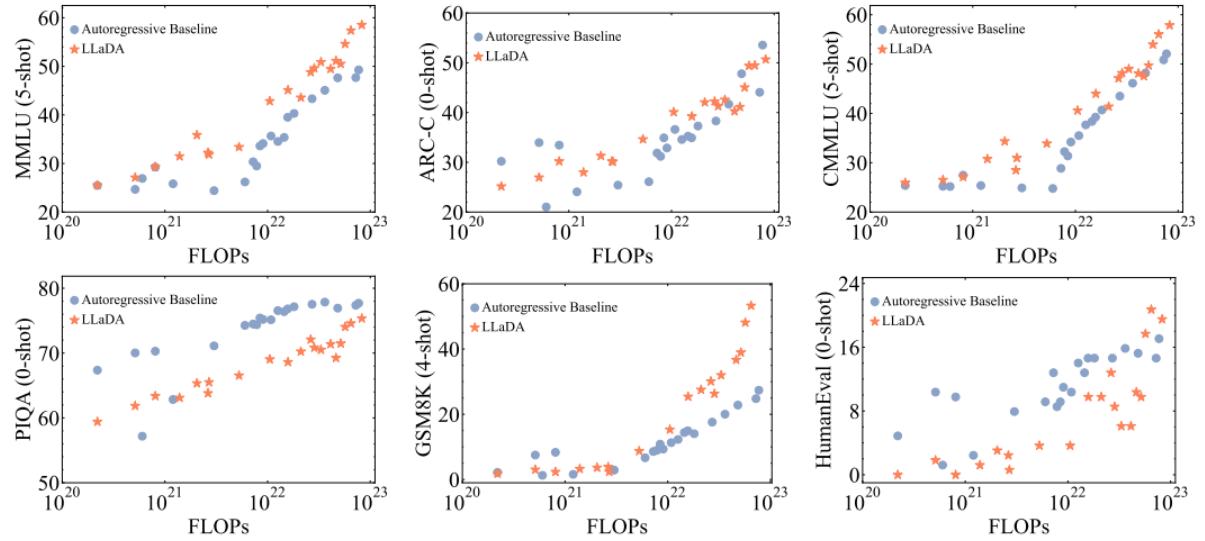


図 2: スケーラビリティ

- FLOPs の増加に伴い、LLaDA の性能は ARM ベースラインと同様のスケーリング則に従う
- MMLU、ARC-C、PIQA 等の一般タスクでは同等のスケーリング傾向
- GSM8K、HumanEval では ARM に比べやや劣るが、スケールに伴い改善傾向

## 実験結果: 事前学習モデルのベンチマーク

	LLaDA 8B*	LLaMA3 8B*	LLaMA2 7B*	Qwen2 7B†	Qwen2.5 7B†	Mistral 7B†	Deepseek 7B†
Model	Diffusion 2.3T	AR 15T	AR 2T	AR 7T	AR 18T	AR -	AR 2T
General Tasks							
MMLU	<b>65.9</b> (5)	65.4 (5)	45.9 (5)	70.3 (5)	74.2 (5)	64.2 (5)	48.2 (5)
BBH	49.7 (3)	<b>62.1</b> (3)	39.4 (3)	62.3 (3)	70.4 (3)	56.1 (3)	39.5 (3)
ARC-C	45.9 (0)	<b>53.1</b> (0)	46.3 (0)	60.6 (25)	63.7 (25)	60.0 (25)	48.1 (0)
Hellaswag	70.5 (0)	<b>79.1</b> (0)	76.0 (0)	80.7 (10)	80.2 (10)	83.3 (10)	75.4 (0)
TruthfulQA	<b>46.1</b> (0)	44.0 (0)	39.0 (0)	54.2 (0)	56.4 (0)	42.2 (0)	-
WinoGrande	74.8 (5)	<b>77.3</b> (5)	72.5 (5)	77.0 (5)	75.9 (5)	78.4 (5)	70.5 (0)
PIQA	73.6 (0)	<b>80.6</b> (0)	79.1 (0)	-	-	-	79.2 (0)
Mathematics & Science							
GSM8K	<b>70.3</b> (4)	48.7 (4)	13.1 (4)	80.2 (4)	85.4 (4)	36.2 (4)	17.4 (8)
Math	<b>31.4</b> (4)	16.0 (4)	4.3 (4)	43.5 (4)	49.8 (4)	10.2 (4)	6.0 (4)
GPQA	25.2 (5)	<b>25.9</b> (5)	25.7 (5)	30.8 (5)	36.4 (5)	24.7 (5)	-
Code							
HumanEval	<b>35.4</b> (0)	34.8 (0)	12.8 (0)	51.2 (0)	57.9 (0)	29.3 (0)	26.2 (0)
HumanEval-FIM	<b>73.8</b> (2)	73.3 (2)	26.9 (2)	-	-	-	-
MBPP	40.0 (4)	<b>48.8</b> (4)	23.2 (4)	64.2 (0)	74.9 (0)	51.1 (0)	39.0 (3)
Chinese							
CMMU	<b>69.9</b> (5)	50.7 (5)	32.5 (5)	83.9 (5)	-	-	47.2 (5)
C-Eval	<b>70.5</b> (5)	51.7 (5)	34.0 (5)	83.2 (5)	-	-	45.0 (5)

図 3: 事前学習モデルのベンチマーク結果

- LLaDA 8B は MMLU (65.9)、TruthfulQA (46.1)、Math (31.4) で **LLaMA3 8B** を上回る
- CMMU (69.9)、C-Eval (70.5) の中国語タスクでも最高スコア
- HumanEval-FIM (73.8) で fill-in-the-middle タスクに特に強い(双方向モデリングの恩恵)

## 実験結果: 事後学習モデルのベンチマーク

	LLaDA 8B*	LLaMA3 8B*	LLaMA2 7B*	Qwen2 7B <sup>†</sup>	Qwen2.5 7B <sup>†</sup>	Gemma2 9B <sup>†</sup>	Deepseek 7B <sup>¶</sup>
Model	Diffusion	AR	AR	AR	AR	AR	AR
Training tokens	2.3T	15T	2T	7T	18T	8T	2T
Post-training Alignment pairs							
	SFT	SFT+RL	SFT+RL	SFT+RL	SFT+RL	SFT+RL	SFT+RL
	4.5M	-	-	0.5M + -	1M + 0.15M	-	1.5M + -
General Tasks							
MMLU	65.5 (5)	<b>68.4</b> (5)	44.1 (5)	-	-	-	49.4 (0)
MMLU-pro	37.0 (0)	<b>41.9</b> (0)	4.6 (0)	44.1 (5)	56.3 (5)	52.1 (5)	-
Hellaswag	74.6 (0)	<b>75.5</b> (0)	51.5 (0)	-	-	-	68.5 (-)
ARC-C	<b>88.5</b> (0)	82.4 (0)	57.3 (0)	-	-	-	49.4 (-)
Mathematics & Science							
GSM8K	69.4 (4)	<b>78.3</b> (4)	29.0 (4)	85.7 (0)	91.6 (0)	76.7 (0)	63.0 (0)
Math	<b>31.9</b> (0)	29.6 (0)	3.8 (0)	52.9 (0)	75.5 (0)	44.3 (0)	15.8 (0)
GPQA	<b>33.3</b> (5)	31.9 (5)	28.4 (5)	34.3 (0)	36.4 (0)	32.8 (0)	-
Code							
HumanEval	49.4 (0)	<b>59.8</b> (0)	16.5 (0)	79.9 (0)	84.8 (0)	68.9 (0)	48.2 (-)
MBPP	41.0 (4)	<b>57.6</b> (4)	20.6 (4)	67.2 (0)	79.2 (0)	74.9 (0)	35.2 (-)

図 4: 事後学習モデルのベンチマーク結果

- SFT のみ (4.5M ペア、RL なし) でも多くのタスクで競争力のある性能
- ARC-C (88.5) で全モデル中最高スコアを達成
- GPQA (33.3)、Math (31.9) で LLaMA3 8B Instruct を上回る

## 実験結果: 総合比較

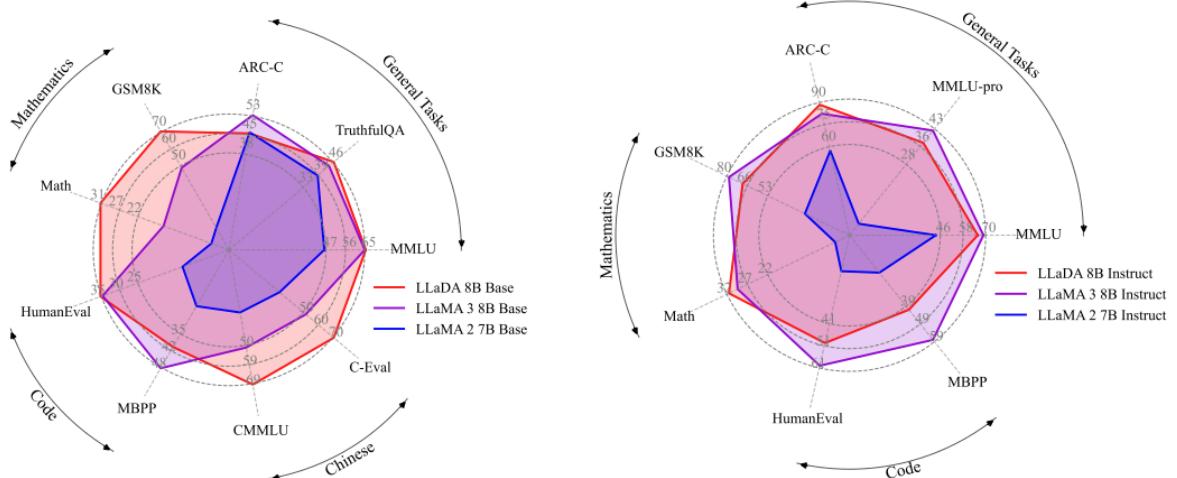


図 5: Zero/Few-shot ベンチマーク総合比較

- 左: Base モデル、右: Instruct モデルのレーダーチャート
- LLaDA 8B は Base・Instruct ともに LLaMA2 7B を大きく上回り、LLaMA3 8B と同等の領域に到達

## 実験結果: 頑健性 (Reversal タスク)

	Forward	Reversal
GPT-4o (2024-08-06)	<b>82.7</b>	34.3
Qwen2.5-7B Instruct	75.9	38.0
LLaDA-8B Instruct	51.8	<b>45.6</b>

図 6: 詩の補完タスクにおける比較

- Forward (通常方向の補完)では GPT-4o が優位
- Reversal (逆順の補完)では LLaDA が GPT-4o、Qwen2.5 を大幅に上回る (45.6)
- ARM は左→右の生成に特化しており逆方向が苦手だが、MDM は双方向的に生成できるため頑健

## 実験結果: サンプリング過程の可視化

### Sampling Process

User	Lily can run 12 kilometers per hour for 4 hours. After that, she runs 6 kilometers per hour. How many kilometers can she run in 8 hours?
LLaDA	 <p>The diagram illustrates the sampling process for a math problem. It shows tokens being selected from a sequence of words and numbers to form the final answer. The tokens are represented by blue boxes, and the sequence is as follows:</p> <pre>     In 4 hours , Lily runs 12 * 4 = 48 kilometers .     After 4 hours she runs 6 * 4 = 24 kilometers .     In total she runs 48 + 24 = 72 kilometers in 8 hours .     Result 72   </pre>

### Multi-round Dialogue Case

User	Could you please share with me what the initial two lines of the renowned poem 'The Road Not Taken'?
LLaDA	Certainly! The initial two lines of the renowned poem "The Road Not Taken" by Robert Frost are: "Two roads diverged in a yellow wood, And sorry I could not travel both." These lines set the stage for the poem, introducing the idea of a speaker facing a choice between two paths.
User	Please help me translate into Chinese.
LLaDA	Sure! The first two lines of "The Road Not Taken" by Robert Frost can be translated into Chinese as: “两条路分岔在黄色的树林中，遗憾我不能同时走”
User	Please translate into German.
LLaDA	Sure! The first two lines of "The Road Not Taken" by Robert Frost can be translated into German as: "Zwei Wege trennten sich im gelben Wald, und leider konnte ich nicht beide tun."
User	Please also write a poem about life choices. I require 5 sentences, and the first word of each sentence starts with C as the first letter.
LLaDA	Certainly! Here's a poem about life choices, with each sentence starting with the letter "C": Choices cascade through life's journey, Creating connections with each decision, Careful consideration leads to better paths, Courageous choices shape our destiny, Choosing wisely guides us through the years.

図 7: サンプリング過程とマルチターン対話の例

- 数学問題に対し、複数ステップで徐々にトークンが確定していく様子
- マルチターン対話も適切に処理可能(翻訳、詩の生成など)

## 制限事項

- 生成長の事前指定が必要: 出力トークン数をユーザーが指定する必要があり、適応的な長さ決定機構がない
- 計算リソースの制約: 最先端 ARM と同規模 (70B 以上)へのスケーリングは未検証
- 推論効率: MDM 専用の attention 機構や KV キャッシュ等の最適化が未導入
  - ARM では標準的な KV キャッシュが使えるが、双向 attention では適用困難
- RL による性能向上が未実施: 既存 ARM ベース LLM では SFT + RL が標準だが、MDM 向けの RL

手法が未確立

- マルチモーダル対応: 画像・音声等への拡張は今後の課題
  - 事後学習手法の発展余地: MDM 固有の alignment 手法(RLHF/DPO 相当)の開発が必要
- 

## まとめ

- LLaDA はマスク拡散モデルを 8B パラメータまでスケールし、ARM ベースの LLM に匹敵する性能を実証した初の研究
  - 双方向モデリングにより、fill-in-the-middle や逆順生成で ARM を上回る独自の強みを持つ
  - 拡散モデルベースの言語モデルが、ARM の代替パラダイムとなり得る可能性を示した
- 

## 参考情報

1. Nie, S., Zhu, F., Liang, Z., Chen, H., Li, D., Lu, Z., & Li, C. (2025). Large Language Diffusion Models. *arXiv preprint arXiv:2502.09992*.