Large Language Diffusion Models

NeurIPS 2025, Nie, et. al
Presented by: Andrew Morgan, Joshua Foster

Overview & Introduction

- Large Language Diffusion with mAsking (LLaDA) is an **8-billion parameter** generative language model **trained entirely from scratch.**
- **Abandons** standard Autoregressive Modeling (ARM) “next-token” prediction framework
- Employs a **forward process** that gradually masks text

Motivation

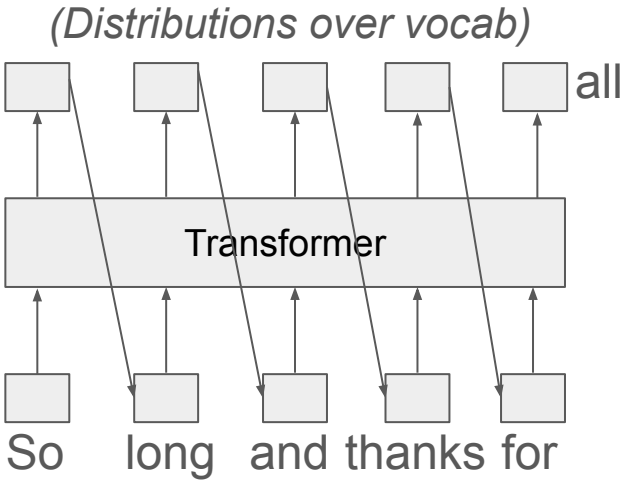
- Question: Is the left-to-right (LTR) autoregressive paradigm the *only* path to achieving SOTA language modeling?
- Success of modern LLMs stems from **optimization** of true data distribution and **generative likelihood.**
  - Can be achieved through multiple probabilistic frameworks
- ARMs possess LTR inductive bias
  - Restricts generalization → **Reversal Curse**, failure at bidirectional reasoning
- LLaDA's Solution: Masked Diffusion Models **naturally** construct distributions with **bidirectional dependencies**
  - Unexplored perspective for large-scale language modeling

Main Claims of the Paper

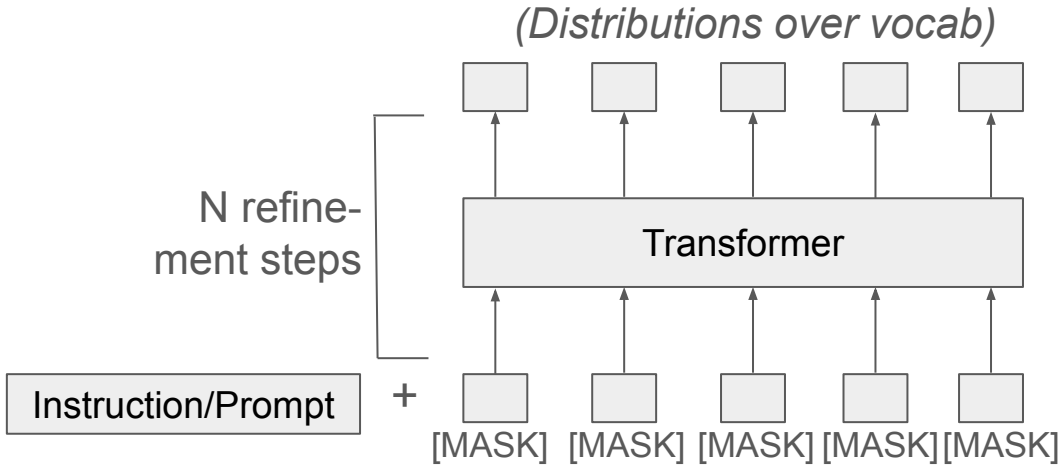
1. Pure masked diffusion models **can match** zero-shot/few-shot capabilities of SOTA ARMs.
2. Diffusion models can **efficiently scale** to large parameter counts and massive compute budgets.
3. Bidirectional training natively **solves** the left-to-right reasoning failures in standard LLMs.
4. LLaDa is built on a principled training objective that is bounded by negative log-likelihood.
5. Core LLM capabilities emerge from massive scale and **generative modeling** principles.
    a. Autoregression is **not** the only path forward!

High-level: Diffusion for Language Modeling

- **Autoregressive** language modeling architectures are dominant
- (Discrete) **Diffusion** is one non-autoregressive approach to language generation, where all tokens in a sequence are **predicted in parallel**



Autoregressive

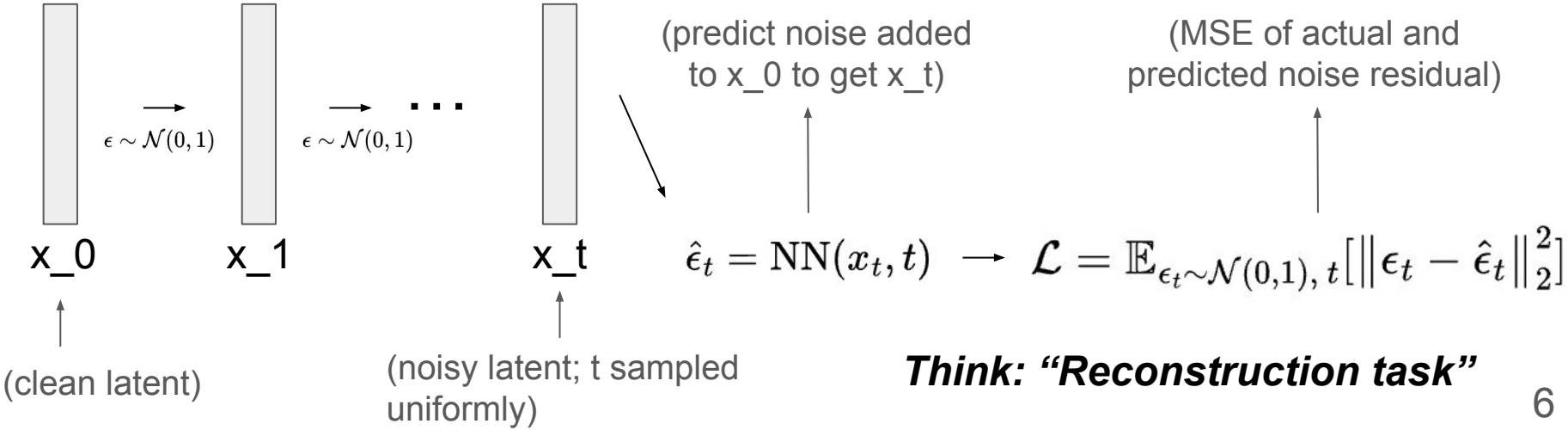


Diffusion (predict all at once)

Background: Diffusion Basics (**L**atent **D**iffusion **M**odeling)

- Dominant architecture in image and video generation
- Training (below):
  - **Forward Diffusion:** incrementally add noise over *t* timesteps
  - **Reverse (Denoising) Process:** train a neural network to predict the noise added at each *t*

- Inference:
  - **Use the Denoising Process to iteratively remove predicted noise from a randomly initialized Gaussian**
  - Can use **cross-attention** to condition denoising output with, e.g., a prompt



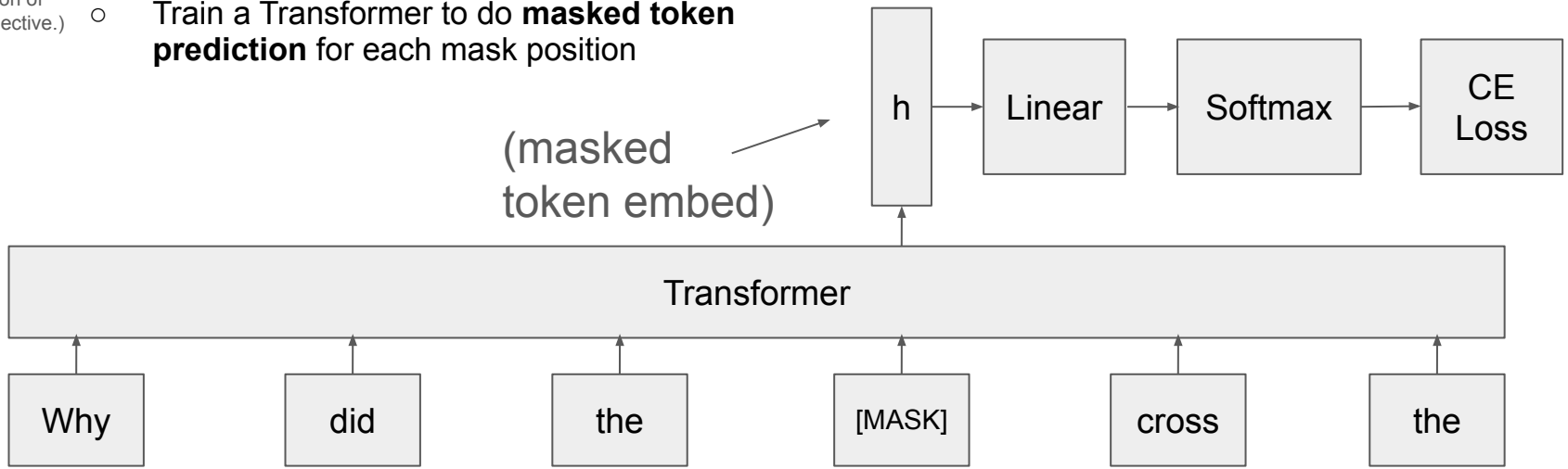
Think: "Reconstruction task"

Background: Bi-directional Encoder Representations (BERT)

- **Conceptually similar** to LLaDa and discrete diffusion
- Training:
  - **Mask some percentage of tokens** in a sequence
  - Train a Transformer to do **masked token prediction** for each mask position

(For brevity, excluding discussion of NSP objective.)

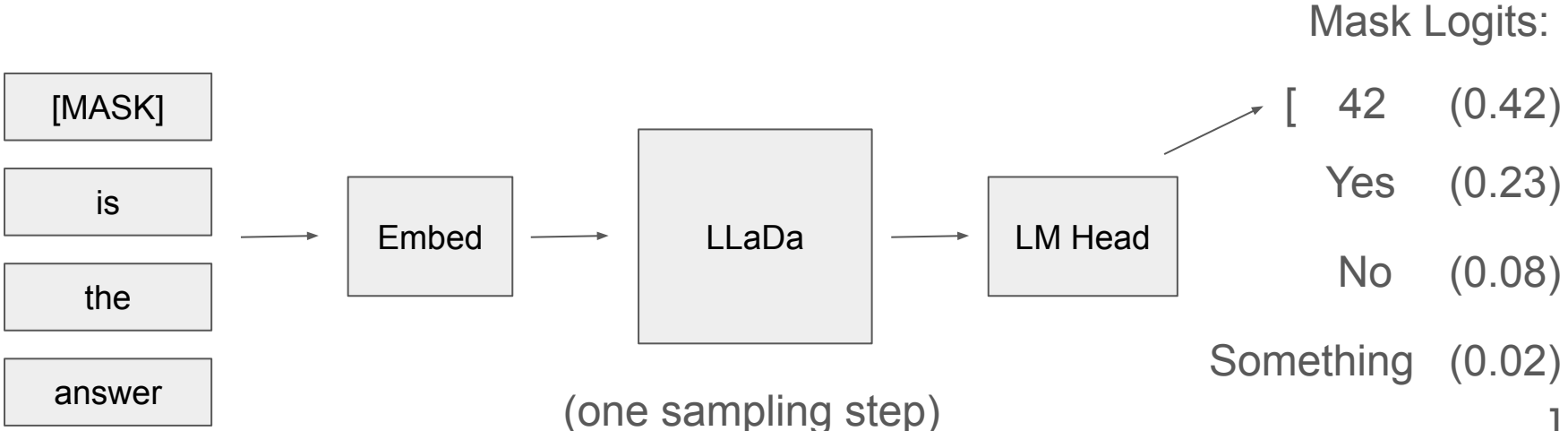
- **LLaDa is like a fusion of BERT and LDM.**



LLaDa Inputs and Outputs

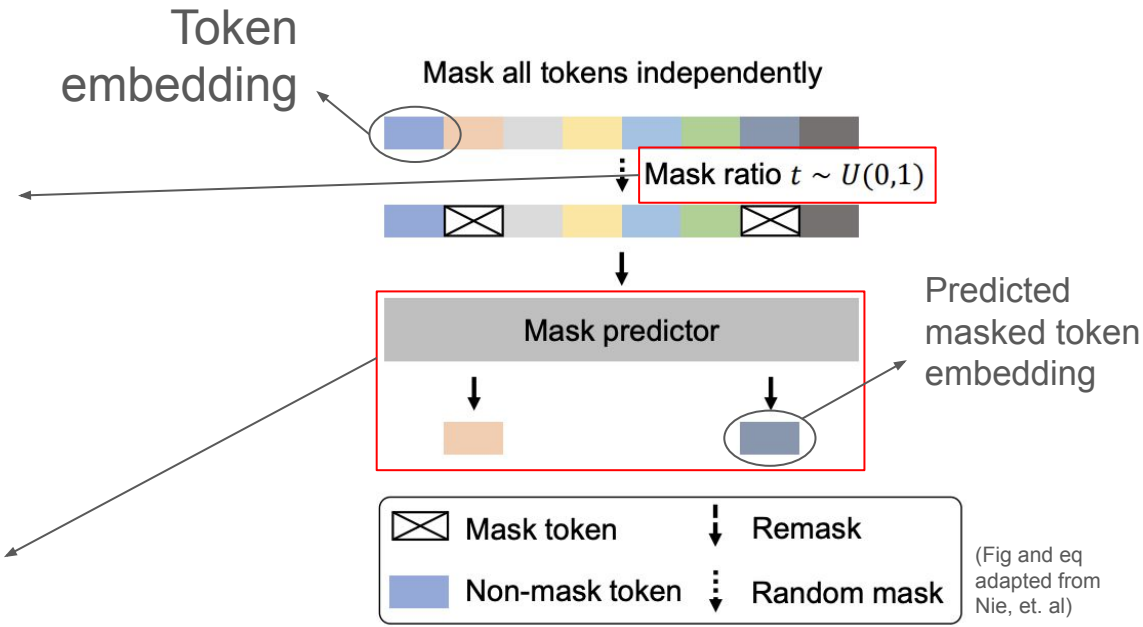
- **What actually goes in and comes out of LLaDa?**
  - **Embeddings, logits,** like in autoregressive LLMs.

- As well, [MASK] is just another learned embedding in the Transformer's embedding matrix.



Pretraining: Large LAnguage Diffusion with MAsking (LLaDa)

- Equivalent to *t* in LDM, but in a discrete setting.
- This value is fixed in BERT but **is sampled uniformly in LLaDa and LDM.**
  - **Discretized "forward process"**
- **Bi-directional Transformer that predicts (reconstructs) tokens at masked positions,** like BERT.
  - **"Reverse/denoising process"**



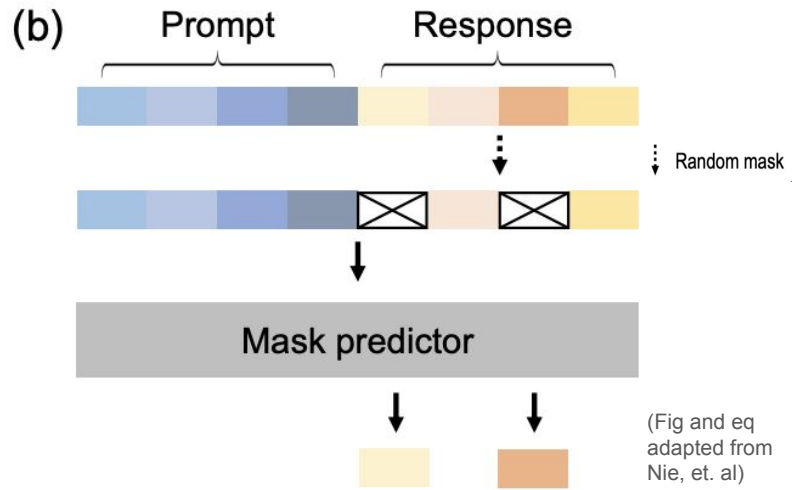
(Fig and eq adapted from Nie, et. al)

$$\mathcal{L}(\theta) \triangleq -\mathbb{E}_{t,x_0,x_t}\left[\frac{1}{t}\sum_{i=1}^{L}\mathbf{1}[x_t^i = \mathrm{M}]\log p_\theta(x_0^i|x_t)\right]$$

Masked token — **Unmasked** token — **Masked** sequence

Post-training: Instruction-tuning and SFT with LLaDa

- Like autoregressive LMs, LLaDa can be post-trained using **supervised fine-tuning (SFT) / instruction-tuning.**
- **Prompt/instruction is unmasked**, and target response is masked using random masking schedule *t,* same as in pretraining.
- Identical training objective as pretraining, **but conditioned on an unmasked prompt sequence.**



(Fig and eq adapted from Nie, et. al)

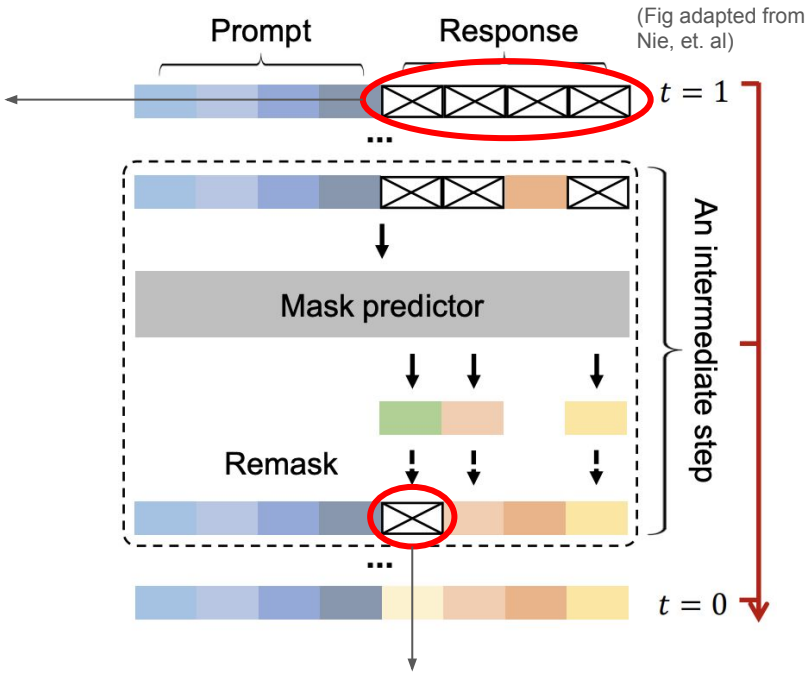
$$-\mathbb{E}_{t,p_0,r_0,r_t}\left[\frac{1}{t}\sum_{i=1}^{L'}\mathbf{1}[r_t^i=\mathrm{M}]\log p_\theta(r_0^i|p_0,r_t)\right]$$

Unmasked resp. token

Unmasked prompt and **masked** sequence

Inference

Fixed-length output, though predicted tokens after <EOS> are discarded.



- **The "diffusion-y" part:**
  - **Start from N fixed, uninformative [MASK] embeddings**
  - **Perform the iterative reverse/denoising process from LDM** in discrete token space.
  - Instead of incrementally denoising using the predicted noise at a timestep (LDM), **incrementally denoise by predicting all masked tokens, and then take a small step by remasking low-confidence predictions.**

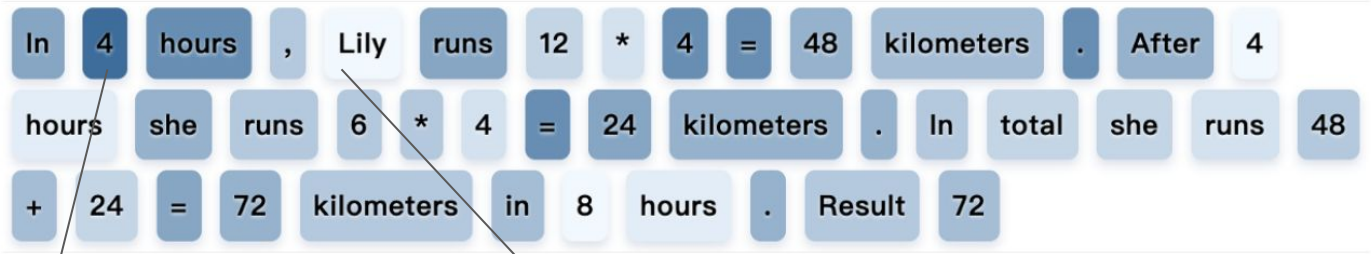
- **Lowest-confidence tokens** are remasked at each prediction step.
  - With probability $\frac{s}{t}$
- Like in LDM, we **don't** go straight to a prediction in one-step (one-step denoising) - yields poor output.

Inference Example (from paper)

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:



Darker: tokens "locked-in" later in inference

Lighter: …earlier in inference

The amount of time spent "thinking" about a token is dynamic

LLaDa varieties and pre-training/post-training details

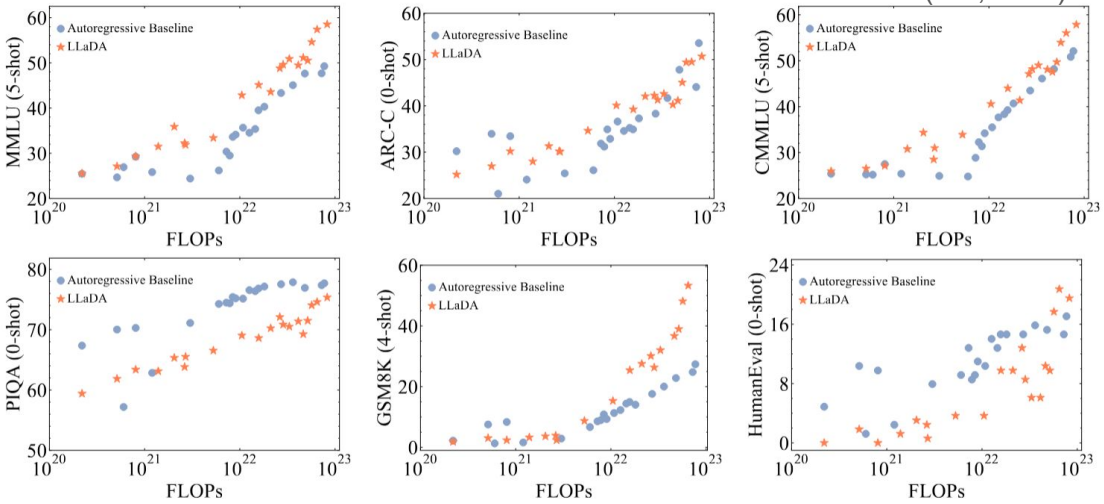
- The authors train **LLaDa-8b and LLaDa-1b** varieties using:
  - **Pretraining**: **2.3T tokens (internet corpora)** and a sample sequence length of **4096**
    - AdamW with a learning rate scheduler using warmup and decay and weight decay of **0.1**
  - **SFT: 4.5 million prompt/response pairs** covering coding, mathematics, and general instruction tasks
    - AdamW with similar LR/decay settings
  - Number of inference sampling steps is a hyperparameter (e.g. 512 in Math benchmark)

- **Sampling Tradeoff -** how to choose sampling steps hyperparameter?
  - Smaller iterative steps (more sampling steps) → higher quality output → inference takes longer!

Experiments: Scalability

- At the 1B scale, LLaDA **matched** custom ARM baselines trained on identical data up to 10^{23} FLOPs.
- LLaDA exhibited stronger curves on **reasoning-heavy tasks** such as MMLU and GSM8K.

(Nie, et. al)



- Even on weaker tasks, the performance gap narrows as scale increases.

Experiments: Zero/Few-Shot Benchmarking

- Surpassed LLaMA2 7B on **nearly all** tasks
- Beat LLaMA3 8B:
  - MMLU (5-shot): **65.9** vs 65.4
  - GSM8K (4-shot): **70.3** vs 48.7
  - HumanEval (0-shot): **35.4** vs 34.8
- Strong overall performance **without** reinforcement learning alignment!

$$-\mathbb{E}_{x_0,\pi\sim \mathrm{U}_\pi}\left[\sum_{i=1}^{L}\log p_\theta\left(x_0^{\pi(i)}\right)|x_0^{\pi(<i)};\pi\right]$$

Minimizing expected NLL over all possible permutations (π) gives **bidirectionality.**

	LLaDA 8B*	LLaMA3 8B*	LLaMA2 7B*	Qwen2 7B†
Model	Diffusion	AR	AR	AR
Training tokens	2.3T	15T	2T	7T
General Tasks				
MMLU	**65.9** (5)	65.4 (5)	45.9 (5)	70.3 (5)
BBH	49.7 (3)	**62.1** (3)	39.4 (3)	62.3 (3)
ARC-C	45.9 (0)	**53.1** (0)	46.3 (0)	60.6 (25)
Hellaswag	70.5 (0)	**79.1** (0)	76.0 (0)	80.7 (10)
TruthfulQA	**46.1** (0)	44.0 (0)	39.0 (0)	54.2 (0)
WinoGrande	74.8 (5)	**77.3** (5)	72.5 (5)	77.0 (5)
PIQA	73.6 (0)	**80.6** (0)	79.1 (0)	-
Mathematics & Science				
GSM8K	**70.3** (4)	48.7 (4)	13.1 (4)	80.2 (4)
Math	**31.4** (4)	16.0 (4)	4.3 (4)	43.5 (4)
GPQA	25.2 (5)	**25.9** (5)	25.7 (5)	30.8 (5)
Code				
HumanEval	**35.4** (0)	34.8 (0)	12.8 (0)	51.2 (0)
HumanEval-FIM	**73.8** (2)	73.3 (2)	26.9 (2)	-
MBPP	40.0 (4)	**48.8** (4)	23.2 (4)	64.2 (0)
Chinese				
CMMLU	**69.9** (5)	50.7 (5)	32.5 (5)	83.9 (5)
C-Eval	**70.5** (5)	51.7 (5)	34.0 (5)	83.2 (5)

(Nie, et. al)

Experiments: Instruction Following & Reversal

- Post-trained strictly via SFT
  - Improved LLaDA performance on most downstream tasks
  - Native multi-turn dialogue abilities w/o RL
- Evaluated on 496-pair Chinese poem dataset testing forward vs backward reasoning
  - Other models **degraded**, LLaDA more balanced

Comparison on Poem Completion Task

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

(Nie, et. al)

Classifier-Free Guidance (CFG):

$$\tilde{p}_\theta(r_0|p_0, r_t) \propto \frac{p_\theta(r_0|p_0, r_t)^{1+w}}{p_\theta(r_0|m, r_t)^w}$$

- m is a masked sequence of the same length as p_0
- w is a tunable parameter that controls strength of p_0
- Unsupervised CFG balances prompt alignment and text diversity

Ablation on CFG

	ARC-C	Hellaswag	TruthfulQA	WinoGrande	GPQA	PIQA
w/o CFG	45.9	70.5	46.1	**74.8**	25.2	73.6
w/ CFG	**47.9**	**72.5**	**46.4**	**74.8**	**26.1**	**74.4**

(Nie, et. al)

Related Work: Diffusion in NLP

- Early attempts mapped text to continuous embeddings.
  - Not scalable → 1B continuous model requires **~64x compute** of ARM for similar performance
- Discrete Masked Diffusion Models (MDMs) replaced continuous tracking with discrete forward/reverse processes.
  - Theoretically viable and matched ARM perplexity, but empirical success had a **GPT-2 ceiling**
  - Parallel studies fine-tuned existing ARMs within MDM framework, limited improvements
- LLaDA is the first pure discrete MDM scaled natively to **8B** from scratch.
  - Core LLM capabilities are **not tied** exclusively to **autoregression.**

Limitations & Future

- Inference & Architecture:
  - Generation sequence length is a **user-defined** hyperparameter → **No adaptive length control**
  - Iterative global token prediction → No **system-level optimizations** like KV caching
  - More efficient and controllable sampling remains **preliminary**
- Training:
  - Direct comparisons to ARMs were **restricted** to $<10^{23}$ FLOPs computational budget
  - Sole **reliance** on SFT → No alignment via RL **caps** performance and alignment with human intent
- LLaDA has not been tested on multi-modal data
- Impact on prompt tuning & agent integrated systems not completely understood
- LLaDA post-training study needed to further explore diffusion LM potential

References

Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2019, June). Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)* (pp. 4171-4186).

Nie, S., Zhu, F., You, Z., Zhang, X., Ou, J., Hu, J., ... & Li, C. (2025). Large language diffusion models. *arXiv preprint arXiv:2502.09992*.

Rombach, R., Blattmann, A., Lorenz, D., Esser, P., & Ommer, B. (2022). High-resolution image synthesis with latent diffusion models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition* (pp. 10684-10695).

Emiel Hoogeboom, Alexey A Gritsenko, Jasmijn Bastings, Ben Poole, Rianne van den Berg, and Tim Salimans. Autoregressive diffusion models. arXiv preprint arXiv:2110.02037, 2021

Shen Nie, Fengqi Zhu, Chao Du, Tianyu Pang, Qian Liu, Guangtao Zeng, Min Lin, and Chongxuan Li. Scaling up masked diffusion models on text. arXiv preprint arXiv:2410.18514, 2024.