Any-Order GPT as Masked Diffusion Model: Decoupling Formulation and Architecture

Shuchen Xue[1, 2] Tianyu Xie[3] Tianyang Hu[4] Zijin Feng[5]
Jiacheng Sun[5] Kenji Kawaguchi[4] Zhenguo Li[5] Zhi-Ming Ma[1, 2]

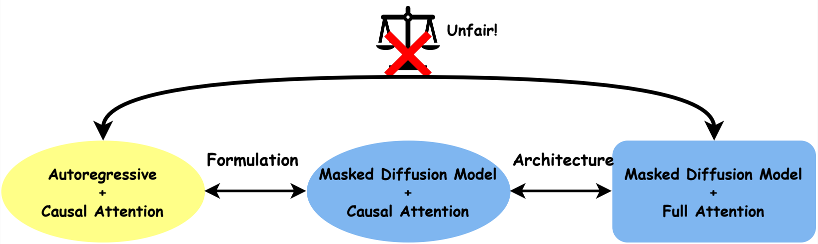
ES-FoMo @ ICML 2025
July 19, 2025

[1]University of Chinese Academy of Sciences
[2]Academy of Mathematics and Systems Science, Chinese Academy of Sciences
[3]Peking University
[4]National University of Singapore
[5]Huawei Noah's Ark Lab

Speaker: Brian K Chen[4]

Motivation and Background

Motivation: An Unfair Comparison



Autoregressive Model (AR)

- Causal Attention
- Decoder-Only

Masked Diffusion Model (MDM)

- Full Attention
- Encoder-Only

Motivation: Efficiency Comparison

Training Efficiency (Token Utilization)

- **Decoder-only AR**: near 100%
- **Encoder-only BERT**: 15%$\sim$25%
- **Encoder-only MDM**: 50% on average
- **Decoder-only MDM**: near 100%

Density Estimation Efficiency (on a fixed L2R order)

- **Decoder-only AR**: $O(n)$
- **Encoder-only MDM**: $O(n^2)$ (but more flexible)
- **Decoder-only MDM**: $O(n)$ (also flexible)

Generation Efficiency

- **Decoder-only AR with KV-cache**: $O(n)$
- **Encoder-only MDM**: $O(Tn)$, where T is the generation steps
- **Decoder-only MDM with KV-cache**: $O(n)$

Training MDM with Causal Attention: MDM is Equivalent to Any-Order AR

$$
\begin{aligned}
\mathcal{L}_{\text{MDM}} &= \int_0^1 \frac{1}{t} \mathbb{E}_{q_{t|0}(\boldsymbol{x}_t|\boldsymbol{x}_0)} \left[\sum_{i:\boldsymbol{x}_0^i=[\text{MASK}]} -\log p_{\boldsymbol{\theta}}(\boldsymbol{x}_0^i|\boldsymbol{x}_t) \right] \mathrm{d}t \\
&= n \cdot \mathbb{E}_{l\sim U(1,\ldots,n)} \frac{1}{n-l+1} \mathbb{E}_{\sigma\sim\mathcal{U}(S_n)} \sum_{r=l}^{n} -\log p_\theta(\boldsymbol{x}_{\sigma_r}|\boldsymbol{x}_{\sigma_{<l}}) \\
&= \mathbb{E}_{\sigma\sim\mathcal{U}(S_n)} \left[\sum_{i=1}^{n} -\log p_{\boldsymbol{\theta}}\left(\boldsymbol{x}_{\sigma_i}|\boldsymbol{x}_{\sigma_{<i}}\right) \right] = \mathcal{L}_{\text{AO-AR}}
\end{aligned}
$$

Two existing decoder-only architectures for training any-order autoregressive models: XL-Net and σ-GPT.

- XL-Net incorporates the target position using two-stream attention, which differs from current architectures.
- σ-GPT incorporates the target position through an additional target positional encoding on a standard GPT architecture.

AO-GPT: A Decoder-Only Model for Flexible Order Modeling

Our AO-GPT incorporates two key enhancements upon σ-GPT:

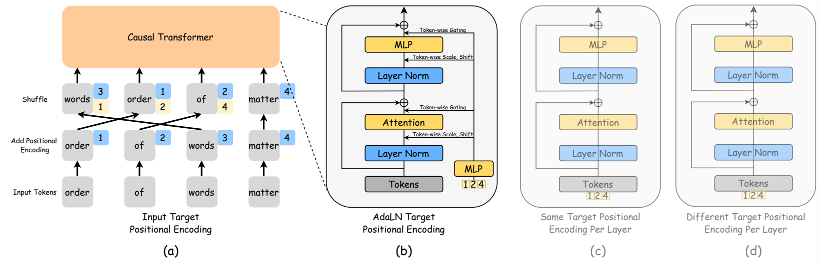


Figure 1: Target position information injections: (a) σ-GPT (b) ours.

1. **Adaptive layerNorm (adaLN) for target position information.**
2. **Exponential Moving Average (EMA).**

AO-GPT: A Decoder-Only Model for Flexible Order Modeling

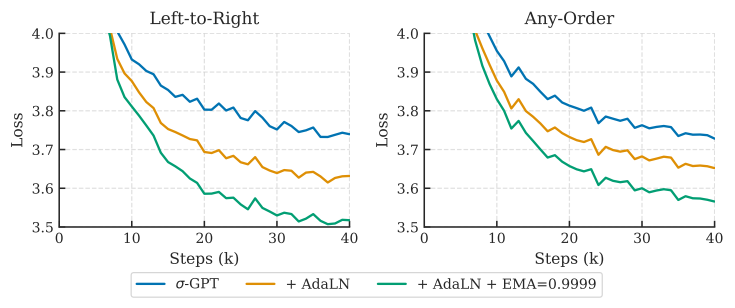


Figure 2: Combined effect of AdaLN and EMA.

Experiment Settings in this paper:

Base Repo: NanoGPT   Data: OpenWebText

Evaluation: LAMBADA, WikiText, PTB, LM1B

Context Length: 1024   Model Size: Small (125M), Medium (355M)

Part 1: AR vs. MDM
(in Decoder-Only Setting)

Finding 1: Any-Order Training Converges Slower

Finding 1

Any-Order GPT converges significantly slower in the initial training stages compared to its standard GPT counterpart, even with the same decoder-only architecture.

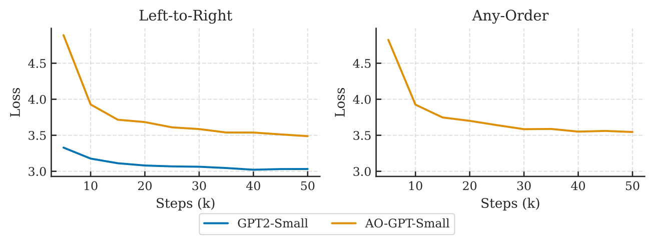


Figure 3: Training loss curves for standard AR (GPT2-Small) and Any-Order AR (AO-GPT-Small).

Finding 2: Language Has a Strong Left-to-Right (L2R) Bias

Finding 2.1 & 2.2

Even when trained on a single fixed order, the standard L2R order converges much faster than an arbitrary, randomly selected fixed order.
A fixed block-wise random order interpolates between L2R and purely random order in terms of convergence speed.

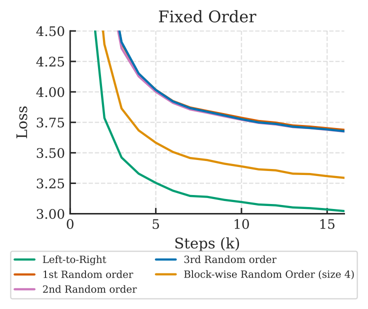


Figure 4: Convergence speed with different fixed prediction orders.

Finding 3: A Little L2R Guidance Goes a Long Way

Finding 3

Incorporating a small fraction (10%) of L2R ordered data into AO-GPT training drastically improves performance on **both** L2R evaluation and Any-Order evaluation.

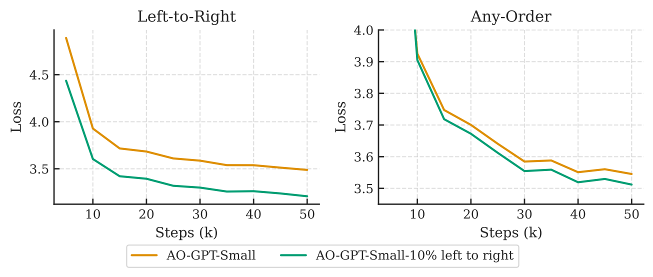


Figure 5: Adding 10% L2R data improves convergence and final loss for both tasks.

Summary of Part 1

- MDM is equivalent to training on **uniform** order distributions, which is not aligned with language's **inherent left-to-right** structure.
- Future MDM research could explore **non-uniform** order distributions to better align with language structure.

Part 2: Encoder vs. Decoder (for MDMs)

Finding 4: A Massive Difference in Modeled Conditionals

Encoder-only and Decoder-only models learn fundamentally different conditional probability spaces.

Encoder-Only: Order-Invariant
The prediction $p(\boldsymbol{x}_j | \boldsymbol{x}_E)$ is conditioned on an *unordered set* of context tokens. The model learns $n \cdot 2^{n-1}$ unique conditional probabilities.

Decoder-Only: Order-Dependent
The prediction $p(\boldsymbol{x}_j | \boldsymbol{x}_E, \sigma_E)$ is conditioned on an *ordered sequence* of context tokens. The model must learn $\approx e \cdot n!$ distinct conditional probabilities.

Key Insight
The decoder's task is combinatorially larger, which may explain performance differences.

Finding 5: Ensembling Context Order Fills the PPL Gap

Finding 5

Decoder-only AO-AR initially shows higher perplexity than its Encoder-only counterpart. However, ensembling predictions over random permutations of the context order bridges this performance gap.

$$p_{\text{ens}}(\boldsymbol{x}_{\sigma_i} | \boldsymbol{x}_{\sigma_{<i}}) = \frac{1}{M} \sum_{j=1}^{M} p_{\boldsymbol{\theta}} \left(\boldsymbol{x}_{\sigma_i} | \boldsymbol{x}_{\text{perm}_j(\sigma_{<i})}, \text{perm}_j(\sigma_{<i}) \right).$$

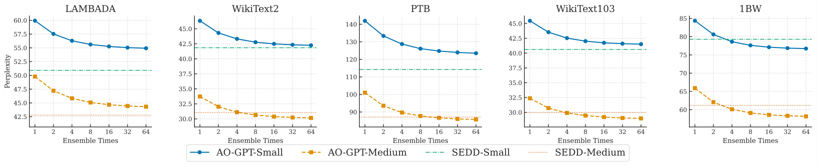


Figure 6: Zero-shot perplexity improves as the number of context order ensembles increases, approaching encoder-only (SEDD) performance.

Findings 6 & 7: Decoders Offer Massive Generation Speedups

Finding 6: Complexity

- Encoder: $O(n^2)$
- Decoder: $O(n)$

(with KV-cache & efficient sampling)

Finding 7: Speed

AO-GPT can achieve 25× speedup on generation compared with SEDD.

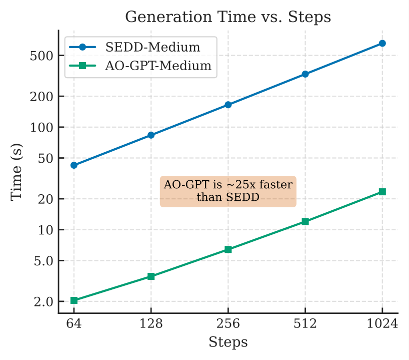


Figure 7: AO-GPT is $\sim 25\times$ faster in generation time than SEDD.

Summary of Part 2

- Different generation complexity (**linear vs. quadratic**).
- Many advantages attributed to MDMs may, in fact, stem from the powerful full attention mechanism they employ, rather than the modeling formulation itself.
- The flexibility of causal MDM gives the potential to **unify AR and MDM paradigms in a single model**

Thank You!

Code is available at:

`https://github.com/scxue/AO-GPT-MDM`