COrAL: Order-Agnostic Language Modeling for Efficient Iterative Refinement

Context-Wise Order-Agnostic Language Modeling

Yuxi Xie, Anirudh Goyal, Xiaobao Wu, Xunjian Yin, Xiao Xu,
Min-Yen Kan, Liangming Pan, William Yang Wang

Follow me on X 👉 @sigrid_xie

Paper 👉

Motivation and Contribution

Background

- **Iterative Refinement** and **Self-Correction** have emerged as an effective paradigm for enhancing the capabilities of large language models (LLMs).
- Existing approaches typically implement iterative refinement at the application or **prompting** level, as a **multi-turn** process relying on **next-token** prediction based on **autoregressive** (AR) modeling.

Pros and Cons of Next-Token based AR Language Modeling

Pros:

Simplicity → Scalability;
Zero-shot Generalization;
Cheap Training Cost; etc.

Cons:

- **Sequential generation** limits the ability to capture **dependencies** spanning beyond the immediate next token, especially when requiring **backward** context.
- The **sequential** nature of AR models leads to **high inference latency**, resulting in computational inefficiency for long sequences

🤔 Research Question

Can we unify the strengths of denoising techniques with order-agnostic modeling to enhance the capabilities of AR-LLMs while mitigating their respective limitations?

- **VL**:varying-length generation
- **BT**: backtrack / look-ahead
- **MV**:multi-variable generation
- **MD**:multi-dependency (inter-sample connection) modeling
- **FS**: fitting feasibility
- **EF**: inference efficiency
- **IT**: mechanism of iterative refinement

Architectures	VL	BT	MV	MD	FS	EF	IT
Next-Token AR (Uria et al., 2016)	✓	✗	✗	✗	✓	✗	✗
Permutation-Based AR (Uria et al., 2014)	✗	✓	✓	✓	✗	✓	✗
NAR (Gu et al., 2018)	✗	✓	✓	✓	✓	✓	✓
Diffusion (Ho et al., 2020)	✗	✓	✓	✓	✓	✗	✓
Consistency Model (Song et al., 2023)	✗	✓	✓	✓	✓	✓	✓
COrAL (Ours)	✓	✓	✓	✓	✓	✓	✓

Decoding: Sliding Blockwise Order-Agnostic Decoding

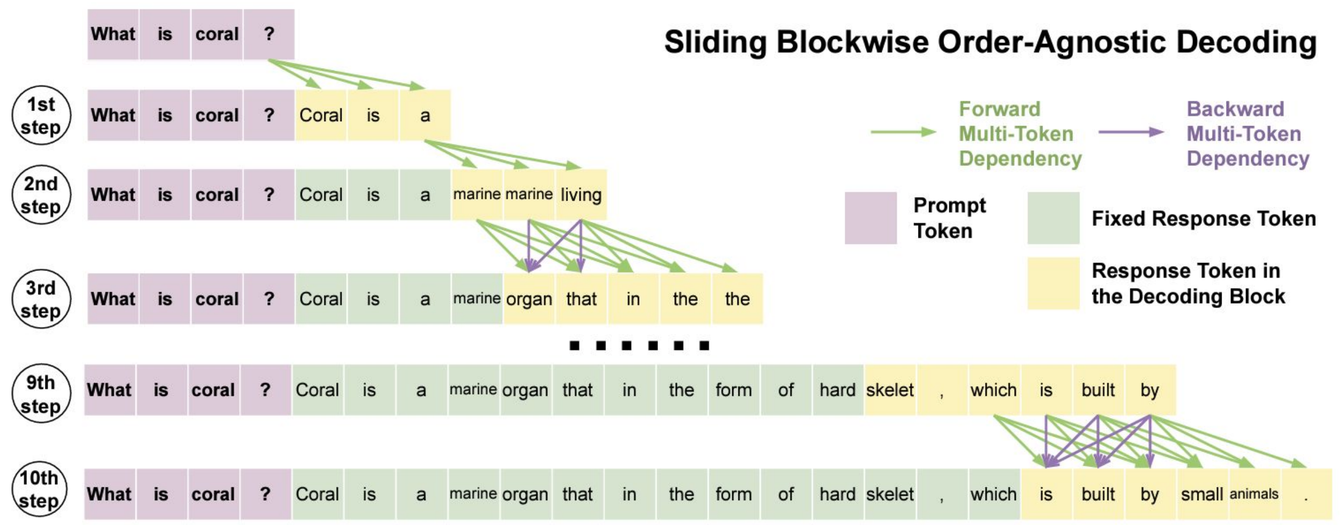


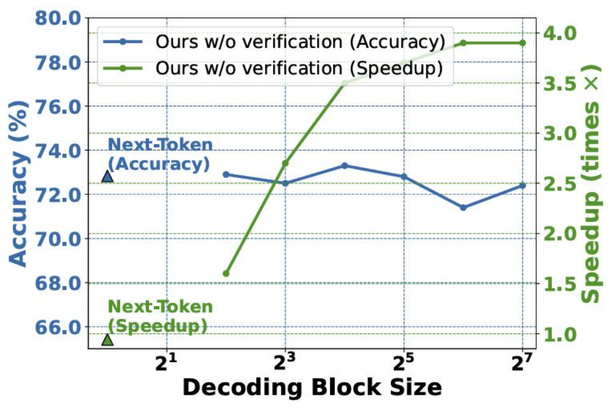
Figure 2: **Sliding Blockwise Order-Agnostic Decoding**. COrAL performs multi-token prediction and refinement in the sliding block with context window size $k = 3$ and block size $b = 6$.

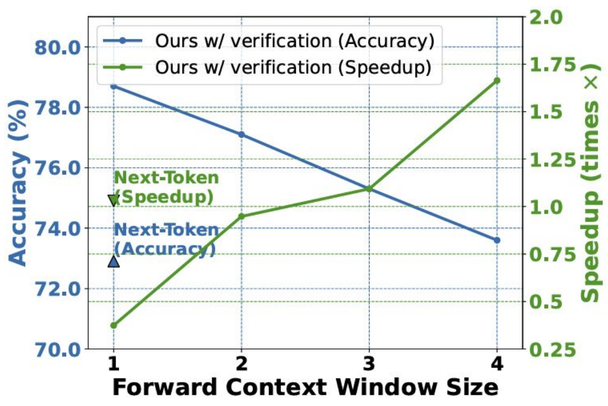
Experiment Result

Result comparison of performance (accuracy%), speed (tokens per second), and cost (seconds per sample) on GSM8K.

Approach	GSM8K			
	Accu.	Speed	Speedup	Cost
NT	74.1	39.7	1.0×	3.67
SC@4	76.2	37.8	–	15.5
Ours	75.3↑1.2	43.4	1.1×	3.35
Ours w/o verifier	72.4↓1.7	156.8	**3.9**×	0.96
Ours w/o multi-forward	**78.7**↑4.6	14.9	–	9.81

Quality–Speed Trade-off





Modeling: Context-Wise Order-Agnostic Language Modeling

Training Objective

$$\log p_\theta(\boldsymbol{y} \mid \boldsymbol{x}) \geq \sum_{t=1}^{T} \mathbb{E}_{i \in [t-k, t+k]} \mathbb{E}_{l \geq 0} \log p_\theta(y_t \mid \boldsymbol{y}_{<i}^{(l)}, \boldsymbol{x})$$

Target-aware Rotary Position Embedding

$$f(\boldsymbol{q}_m, \mu)^\top f(\boldsymbol{k}_n, n) = g(\boldsymbol{q}_m, \boldsymbol{k}_n, \mu - n), \quad \mu \in [m-k, m+k]$$

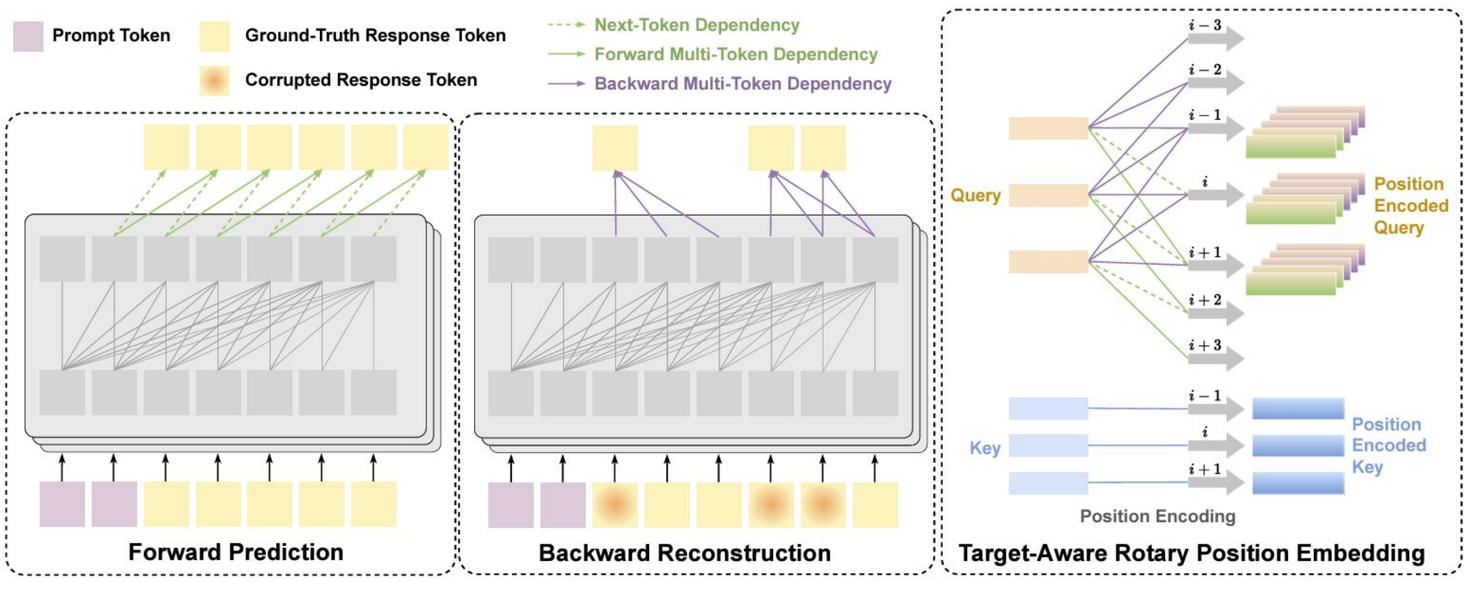
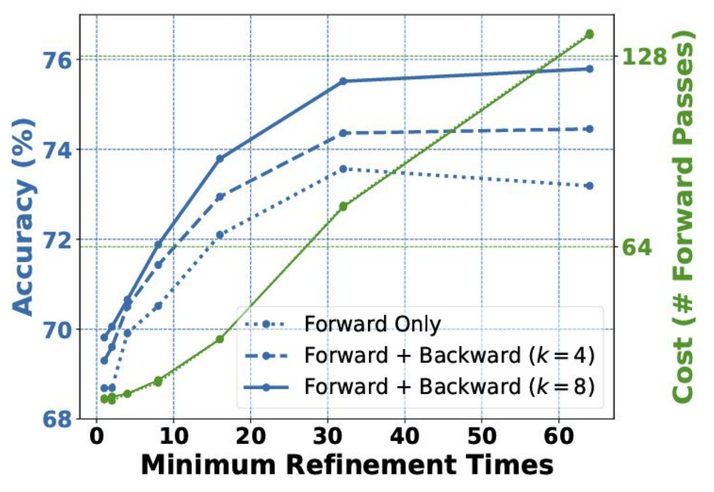


Figure 3: **Context-Wise Order-Agnostic Language Modeling**. We visualize the order-agnostic dependencies within a context window size $k = 2$. For target-aware position encoding, we show how COrAL obtains query representations for multiple positions within a context window size $k = 2$.

Performance Scaling



Token-level accuracy