Planned Diffusion

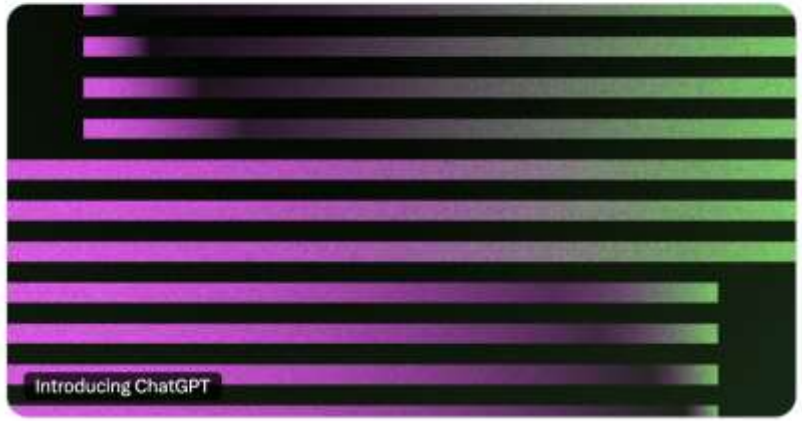
Daniel Israel*, Tian Jin*, Ellie Cheng, Guy Van den Broeck,

Aditya Grover, Suvinay Subramanian, Michael Carbin

The state of LLM

OpenAI
@OpenAI

Try talking with ChatGPT, our new AI system which is optimized for dialogue. Your feedback will help us improve it.



Anthropic @AnthropicAI · Feb 24, 2025

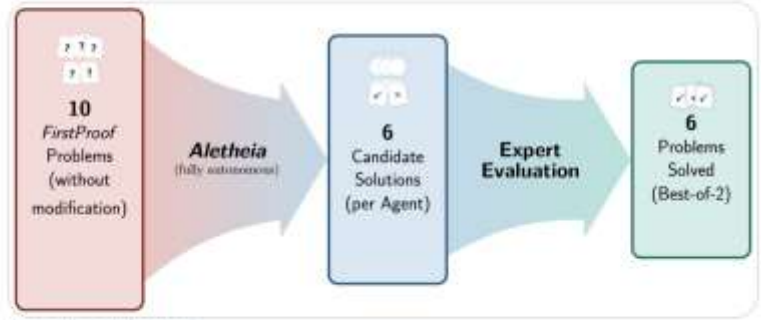
Claude Code has become indispensable for our team. In early testing, Claude completed tasks in a single pass that would normally take 45+ minutes of manual work.

Join the limited preview: docs.anthropic.com/en/docs/agents...

Thang Luong @lmthang · Feb 25

Thrilled to share: #Aletheia, our math research agent, just solved 6/10 notoriously hard FirstProof problems autonomously, the best result in the inaugural challenge! To me, this is even bigger than our historic IMO-gold achievement last year; these problems challenge even top

Show more



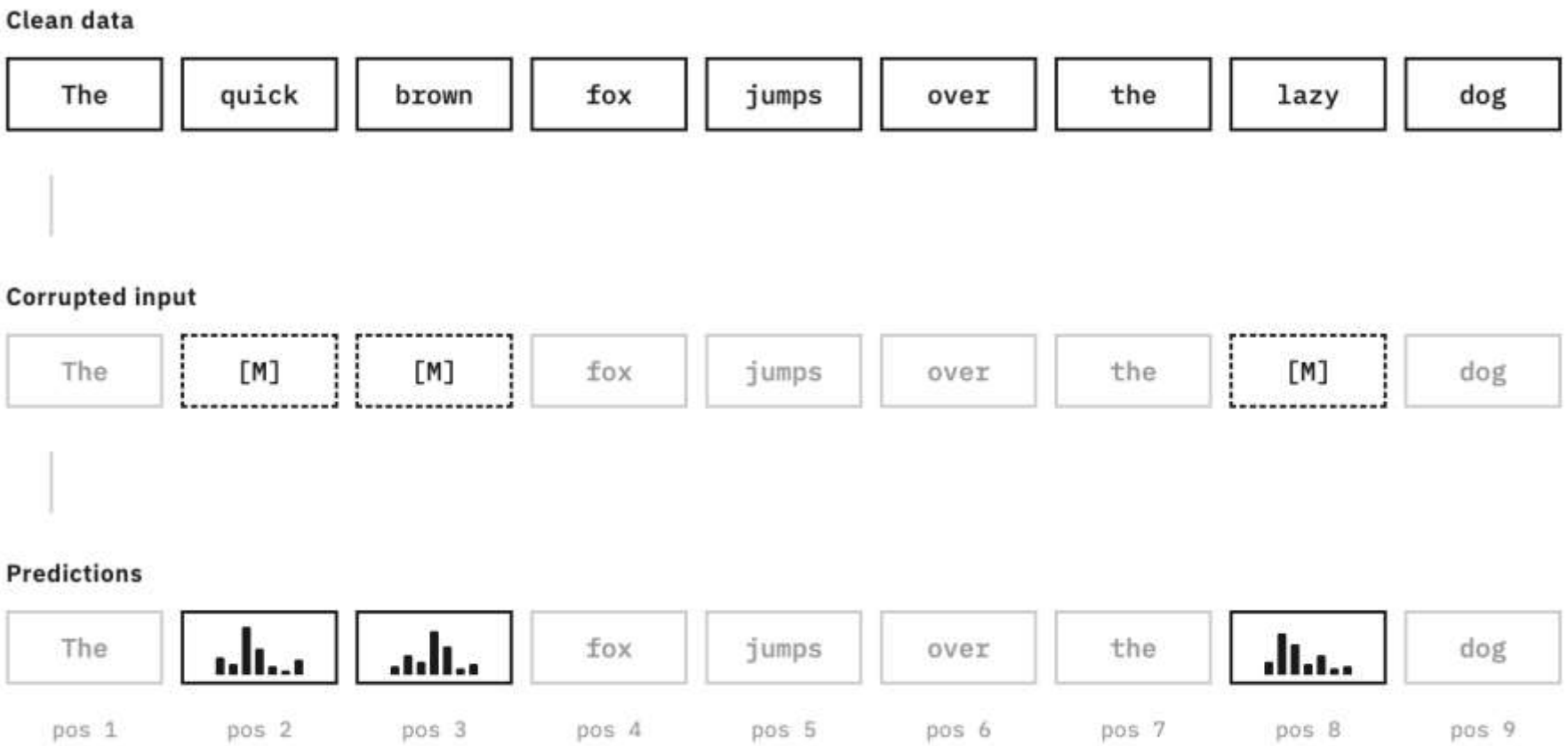
Quoc Le and 9 others

21 184 871 131K

The challenge

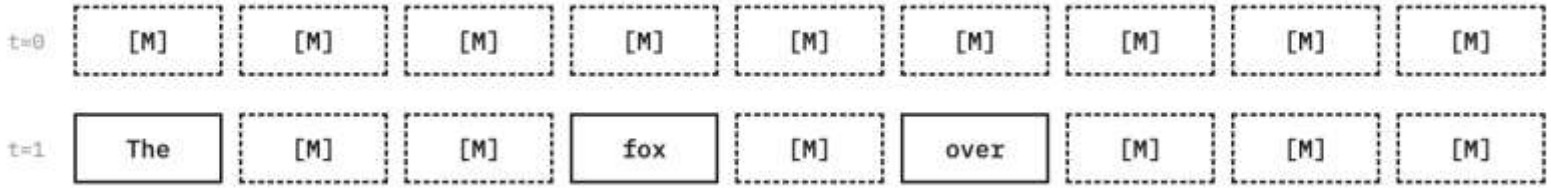
Sequential Autoregressive Decoding

Discrete Diffusion LLM

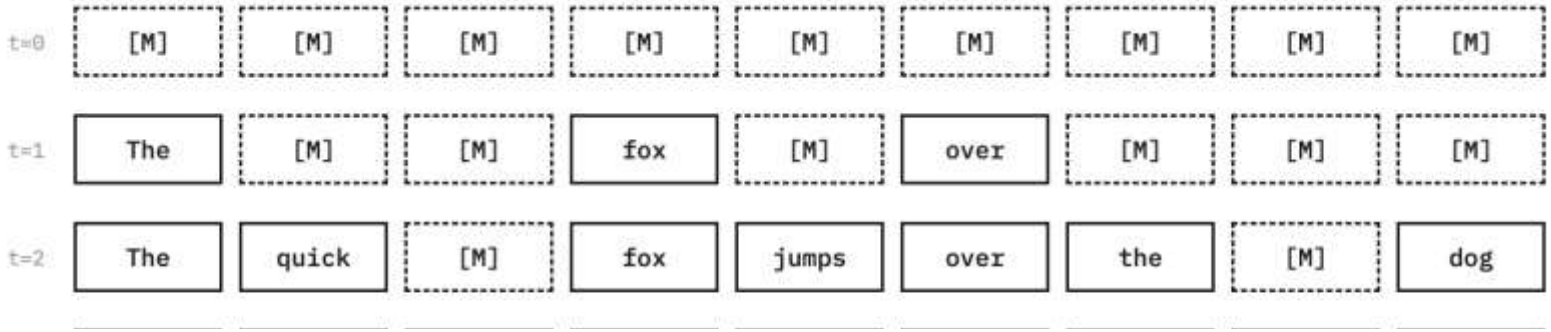


Discrete Diffusion LLM

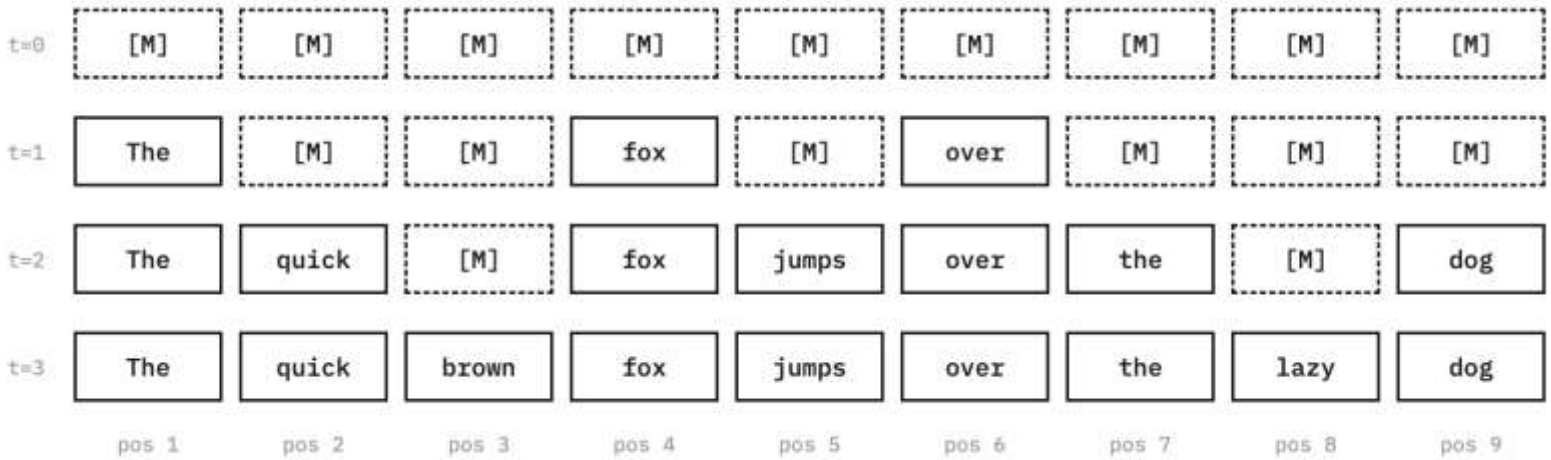
Discrete Diffusion LLM



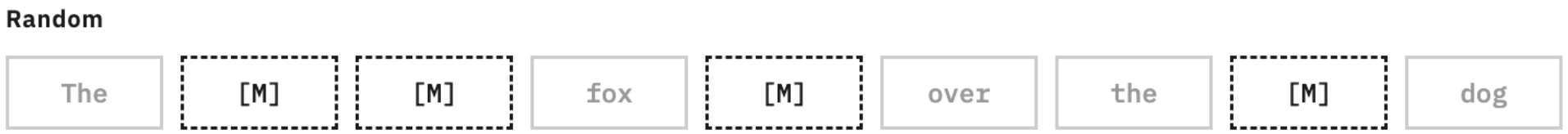
Discrete Diffusion LLM



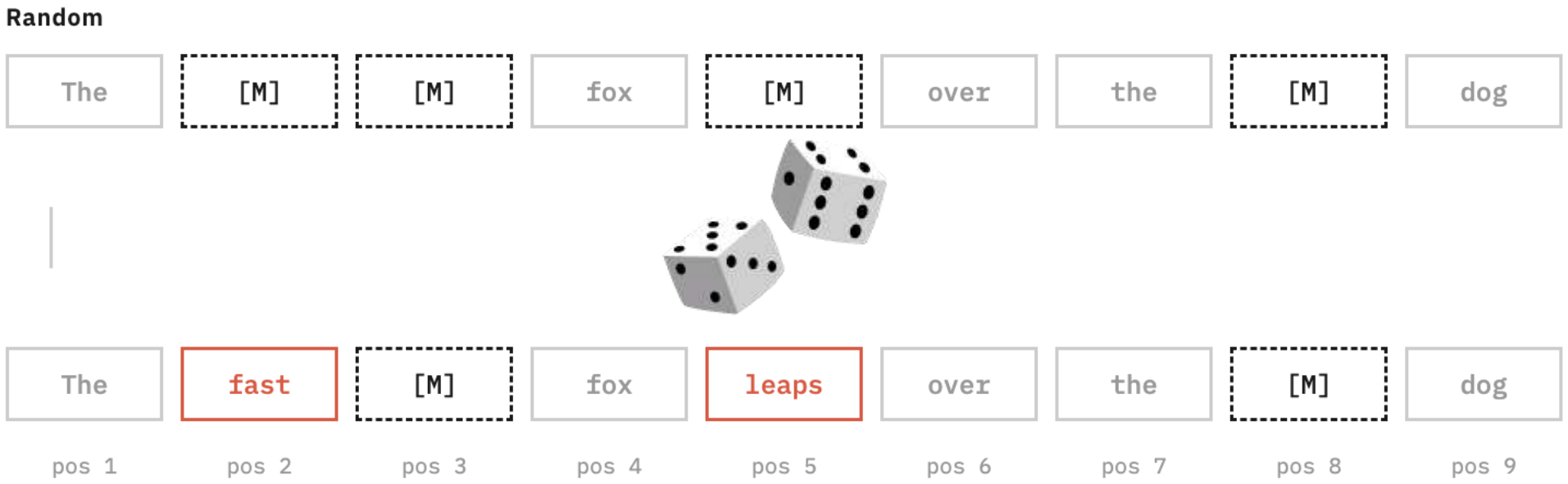
Discrete Diffusion LLM



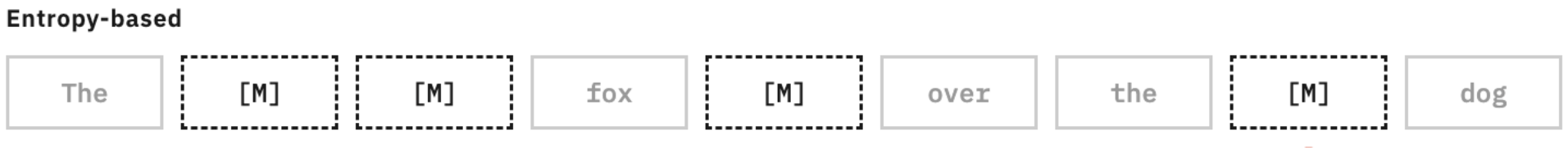
Random denoising order



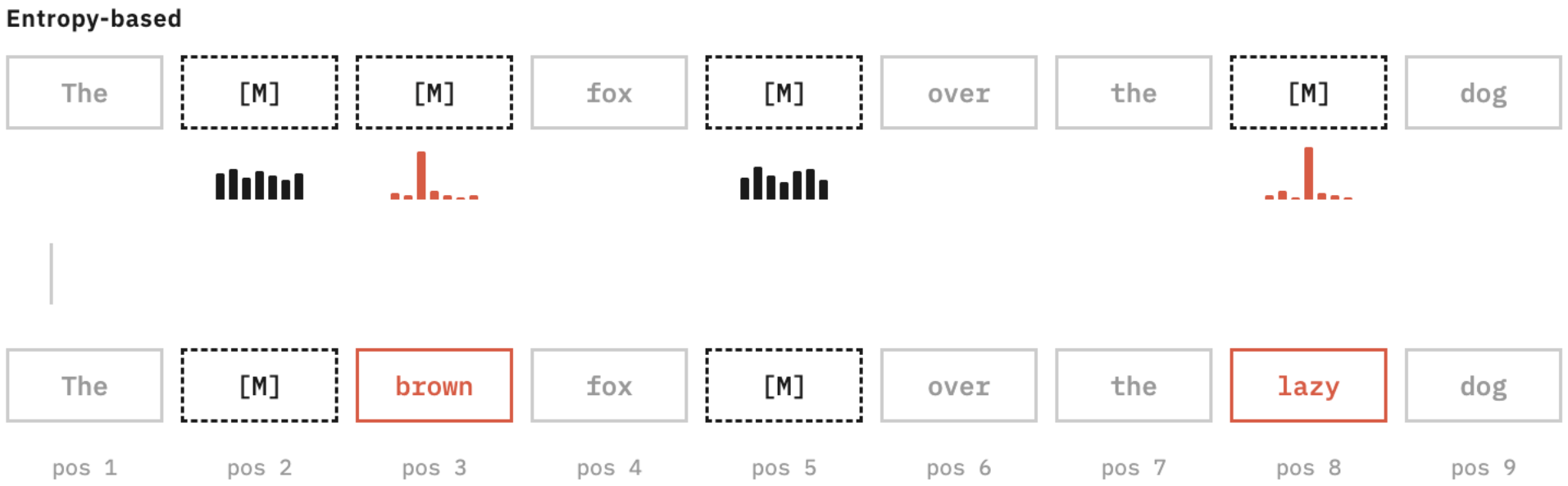
Random denoising order



Entropy denoising order



Entropy denoising order



Discrete Diffusion Large Language Model (dLLM)

Any-Order Generation

Goes beyond left-to-right; naturally supports infilling

What is Aurora Borealis?

1 token/step, 1.8s

[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M]

Discrete Diffusion Large Language Model (dLLM)

Any-Order Generation

Goes beyond left-to-right; naturally supports infilling

Natively Parallel

Generates multiple tokens at once

What is Aurora Borealis?

2 token/step, 0.7s

[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M]

What is Aurora Borealis?

3 tokens/step, 0.4s

[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M]

The catch

Designing good parallel denoising order is hard.

The catch

Designing good parallel denoising order is hard.

Planned Diffusion is a parallel denoising order.

It achieves a1.3x-1.8x speedup over AR with -1-5% quality.

Existing solutions – Skeleton-of-Thought/APAR/Pasta

Prompt

What are the most effective strategies for conflict resolution in the workplace?

prompting

Outline

1. Active listening
2. Identify issues
3. Compromise

Existing solutions – Skeleton-of-Thought/APAR/Pasta

Prompt

What are the most effective strategies for conflict resolution in the workplace?

prompting

Outline

1. Active listening
2. Identify issues
3. Compromise

API

Parallel Expansion

1. Active listening. Active listening involves fully concentrating on …
2. Identify issues. Look into the root causes of …
3. Compromise. Look for a middle ground …

Planned Diffusion

- Model determines its own denoising order.
- Model predicts and annotates semantic dependences.
- Runtime interprets & implements corresponding denoising order.
- Annotations serve as a **denoising plan**.

What is Aurora Borealis?

What is Aurora Borealis?

What is Aurora Borealis?

`definition` `appearance` `analogy`

Planned Diffusion

Annotation Language

Runtime System

Training Algorithm

What is Aurora Borealis?

definition appearance analogy

The Aurora Borealis is a northern light. It appears in the sky as a colorful, glowing light that can be seen near the North Pole. It's like a pretty light show in the sky that happens during the winter season.

What is Aurora Borealis?

<topic>definition 1</topic> <topic>appearance 2</topic>
<topic>analogy 2</topic>

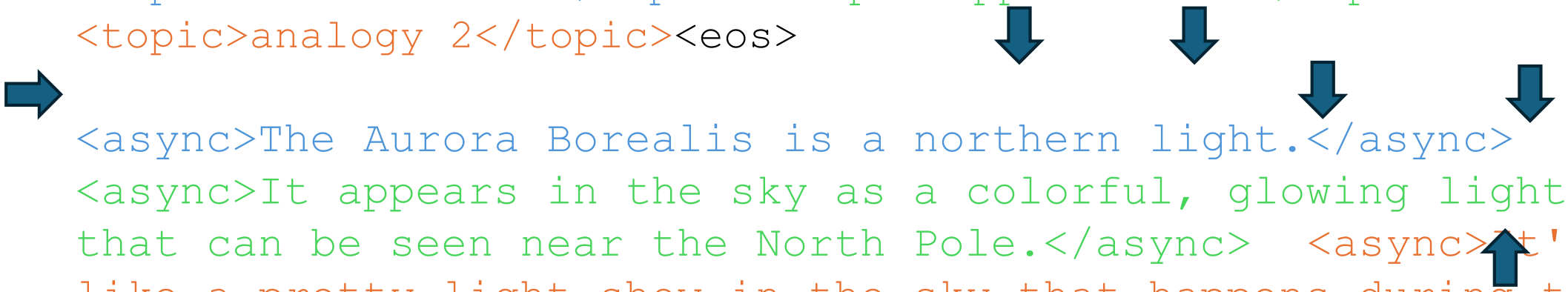
The Aurora Borealis is a northern light. It appears in the sky as a colorful, glowing light that can be seen near the North Pole. It's like a pretty light show in the sky that happens during the winter season.

What is Aurora Borealis?

<topic>definition 1</topic> <topic>appearance 2</topic>
<topic>analogy 2</topic><eos>

The Aurora Borealis is a northern light. It appears in the sky as a colorful, glowing light that can be seen near the North Pole. It's like a pretty light show in the sky that happens during the winter season.

What is Aurora Borealis?



What is Aurora Borealis?

<topic>definition 1</topic> <topic>appearance 2</topic>
<topic>analogy 2</topic><eos>

What is Aurora Borealis?

<topic>definition 1</topic> <topic>appearance 2</topic>
<topic>analogy 2</topic><eos>

<async>[M] [M] [M] [M] [M] [M] [M] [M] [M] [M]</async>
<async>[M] [M] [M] [M] [M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M]</async> <async>[M] [M] [M] [M]
[M] [M] [M] [M] [M] [M] [M] [M] [M] [M] [M] [M] [M] [M]
[M] </async>

What is Aurora Borealis?

<topic>definition 1</topic> <topic>appearance 2</topic>
<topic>analogy 2</topic><eos>

<async>The Aurora Borealis is a northern light.</async>
<async>It appears in the sky as a colorful, glowing light that can be seen near the North Pole.</async> <async>It's like a pretty light show in the sky that happens during the winter season. </async>

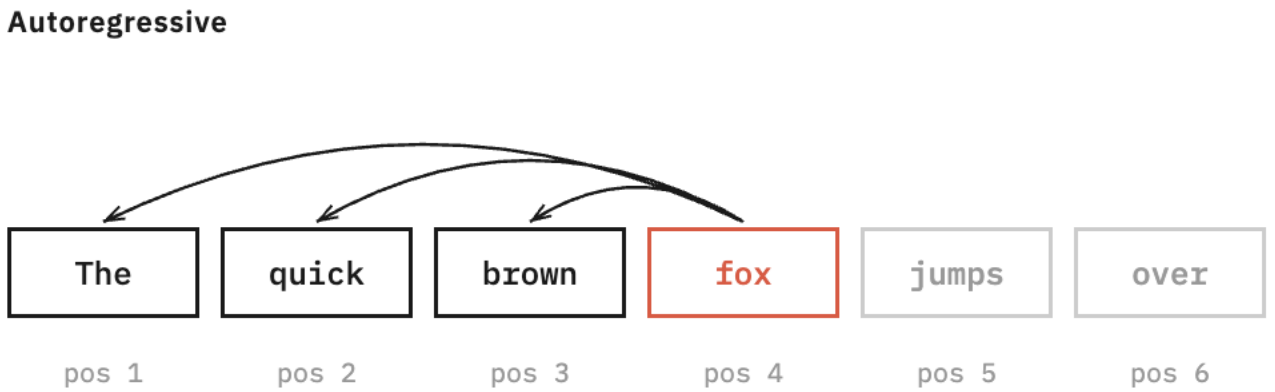
What is Aurora Borealis?

<topic>definition 1</topic> <topic>appearance 2</topic>
<topic>analogy 2</topic><sync>

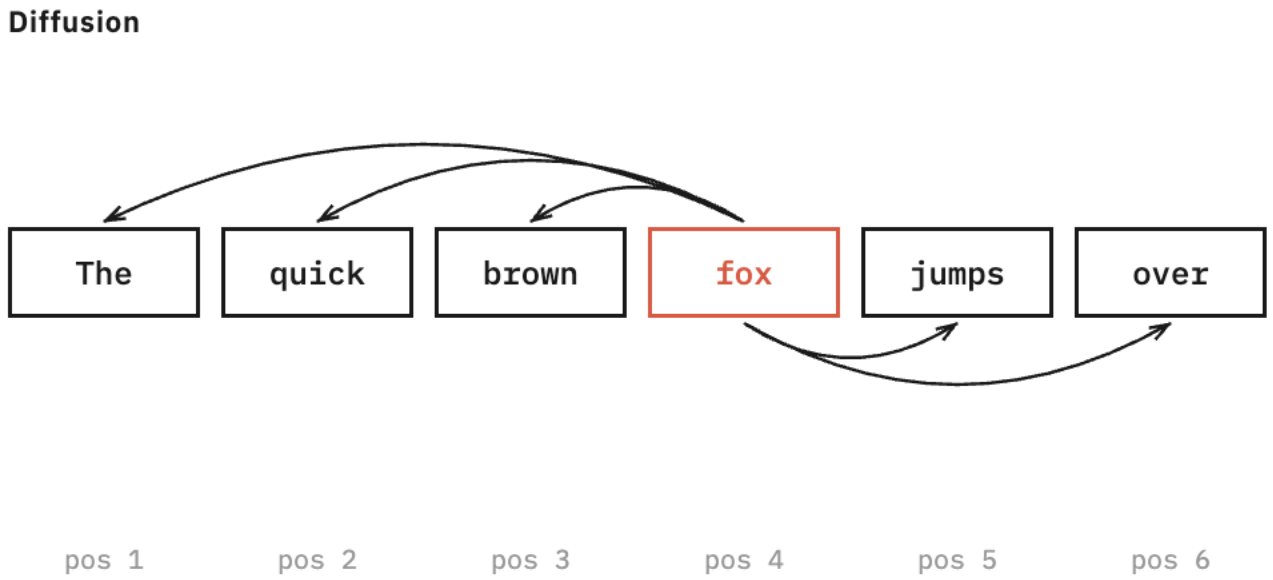
<async>The Aurora Borealis is a northern light.</async>
<async>It appears in the sky as a colorful, glowing light that can be seen near the North Pole.</async> <async>It's like a pretty light show in the sky that happens during the winter season. </async>

<topic>location 1</topic> <topic>science 2</topic>

KV Cache - AR


Autoregressive
The
quick
brown
fox
jumps
over
pos 1
pos 2
pos 3
pos 4
pos 5
pos 6

KV Cache - Diffusion



What is Aurora Borealis?

```
<topic>definition 1</topic> <topic>appearance 2</topic>
<topic>analogy 2</topic><sync>

<async>The Aurora Borealis is a northern light.</async>
<async>It appears in the sky as a colorful, glowing light
that can be seen near the North Pole.</async>  <async>It's
like a pretty light show in the sky that happens during the
winter season. </async>

<topic>location 1</topic> <topic>science 2</topic>
```

What is Aurora Borealis?

<topic>definition 1</topic> <topic>appearance 2</topic>
<topic>analogy 2</topic><sync>

<async>The Aurora Borealis is a northern light.</async>
<async>It appears in the sky as a colorful, glowing light that can be seen near the North Pole.</async> <async>It's like a pretty light show in the sky that happens during the winter season. </async>

<topic>location 1</topic> <topic>science 2</topic>

What is Aurora Borealis?

<topic>definition 1</topic> <topic>appearance 2</topic>
<topic>analogy 2</topic><sync>

<async>The Aurora Borealis is a northern light.</async>
<async>It appears in the sky as a colorful, glowing light that can be seen near the North Pole.</async> <async>It's like a pretty light show in the sky that happens during the winter season. </async>

<topic>location 1</topic> <topic>science 2</topic>

Training - Data Curation

01

Start with ~100K subsample of SlimOrca – a popular instruction following dataset.

02

Prompt Gemini-Flash with syntax/semantics of annotation tags.

03

Use prompted Gemini-Flash to annotate SlimOrca responses.

04

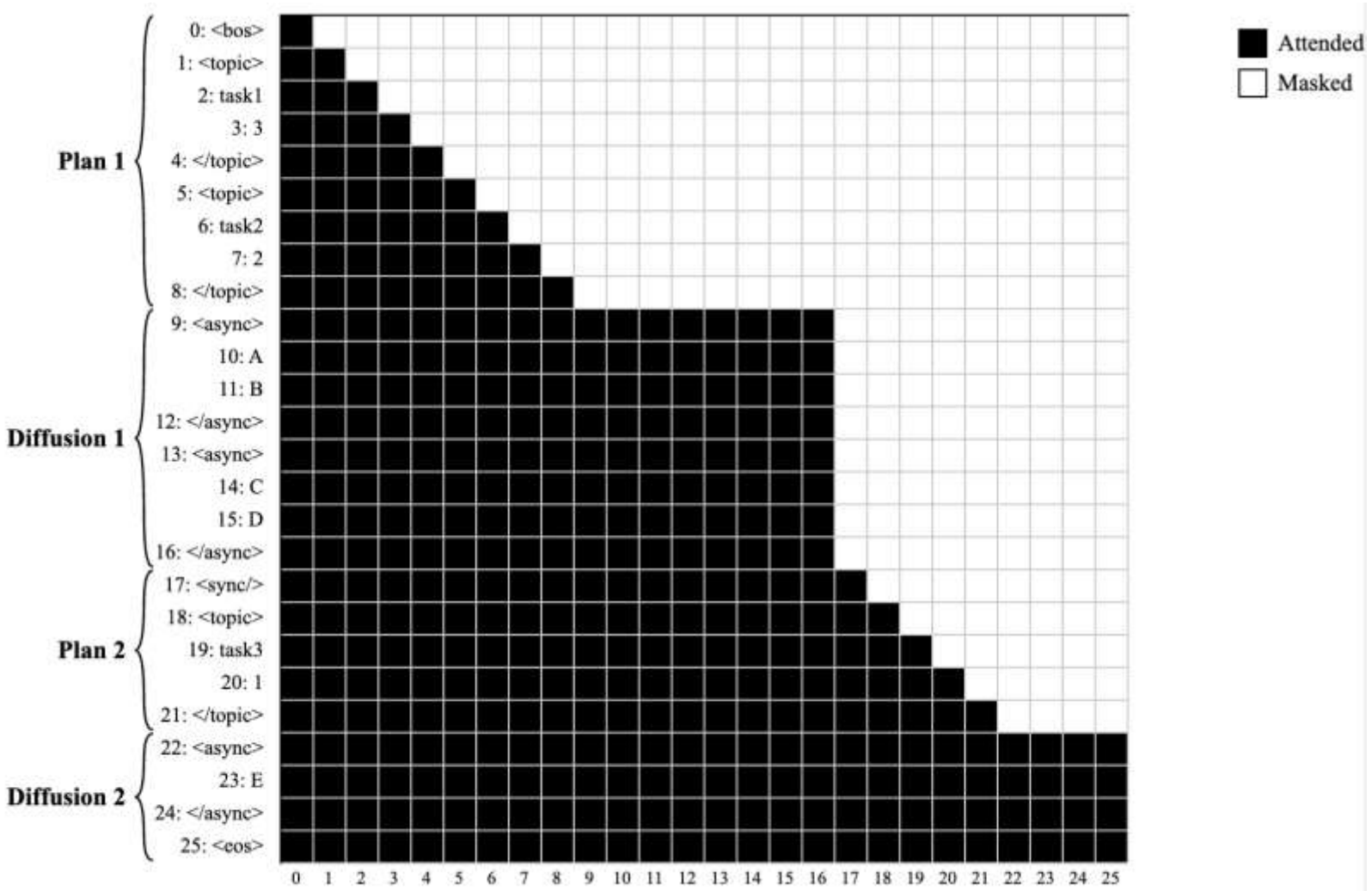
Validate annotations, discard malformed examples.

Hybrid AR/Diffusion Training Loss

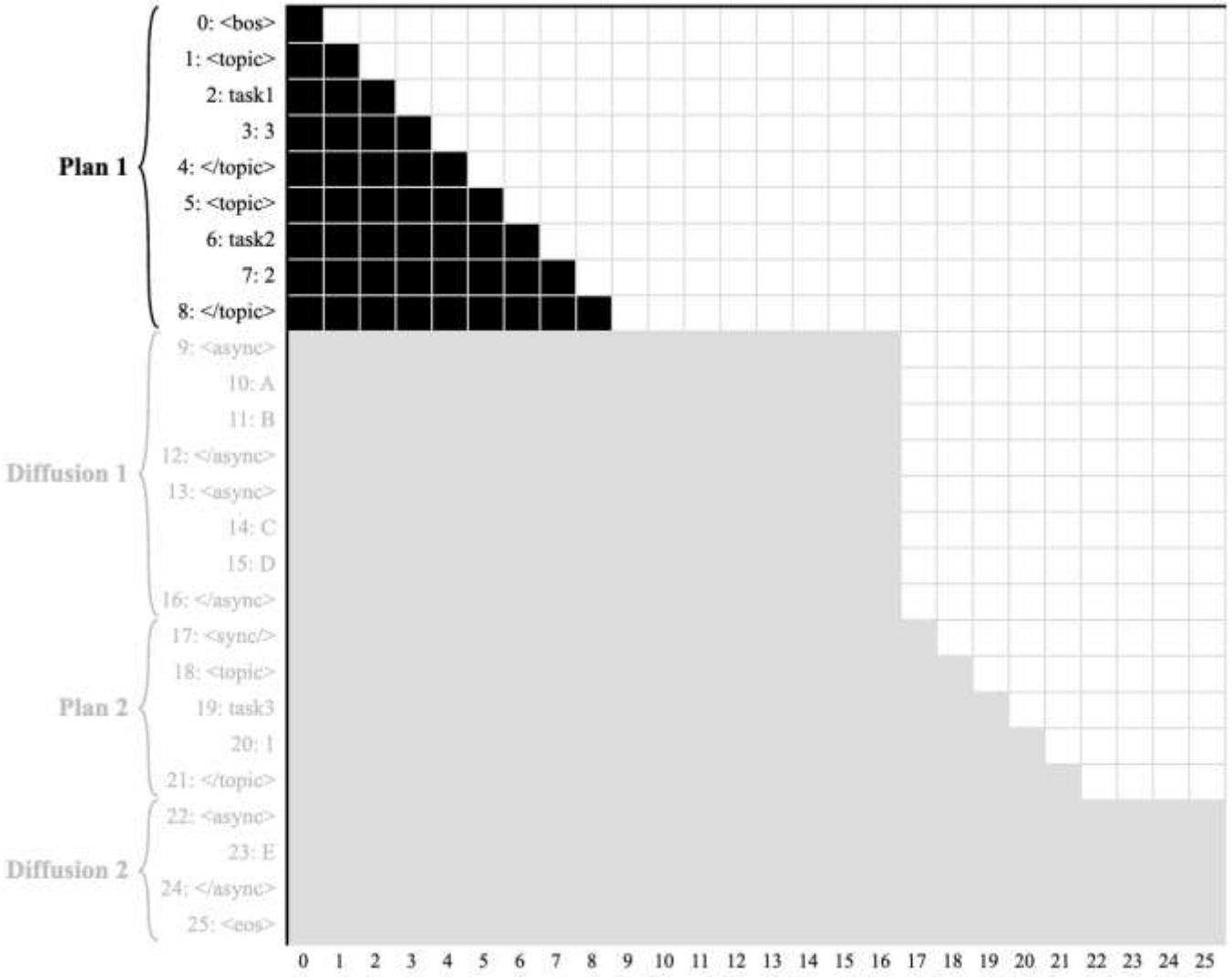
- Challenge: Planned diffusion requires learning both **AR planning** and **diffusion generation**
- Hybrid Loss = **AR Loss** + **Diffusion Loss**

Component	Tokens	Loss
AR Planning	Plan Annotation (<topic>, <sync/>,..)	Next-token prediction
Diffusion Denoising	Content inside <async> chunks	Masked token prediction

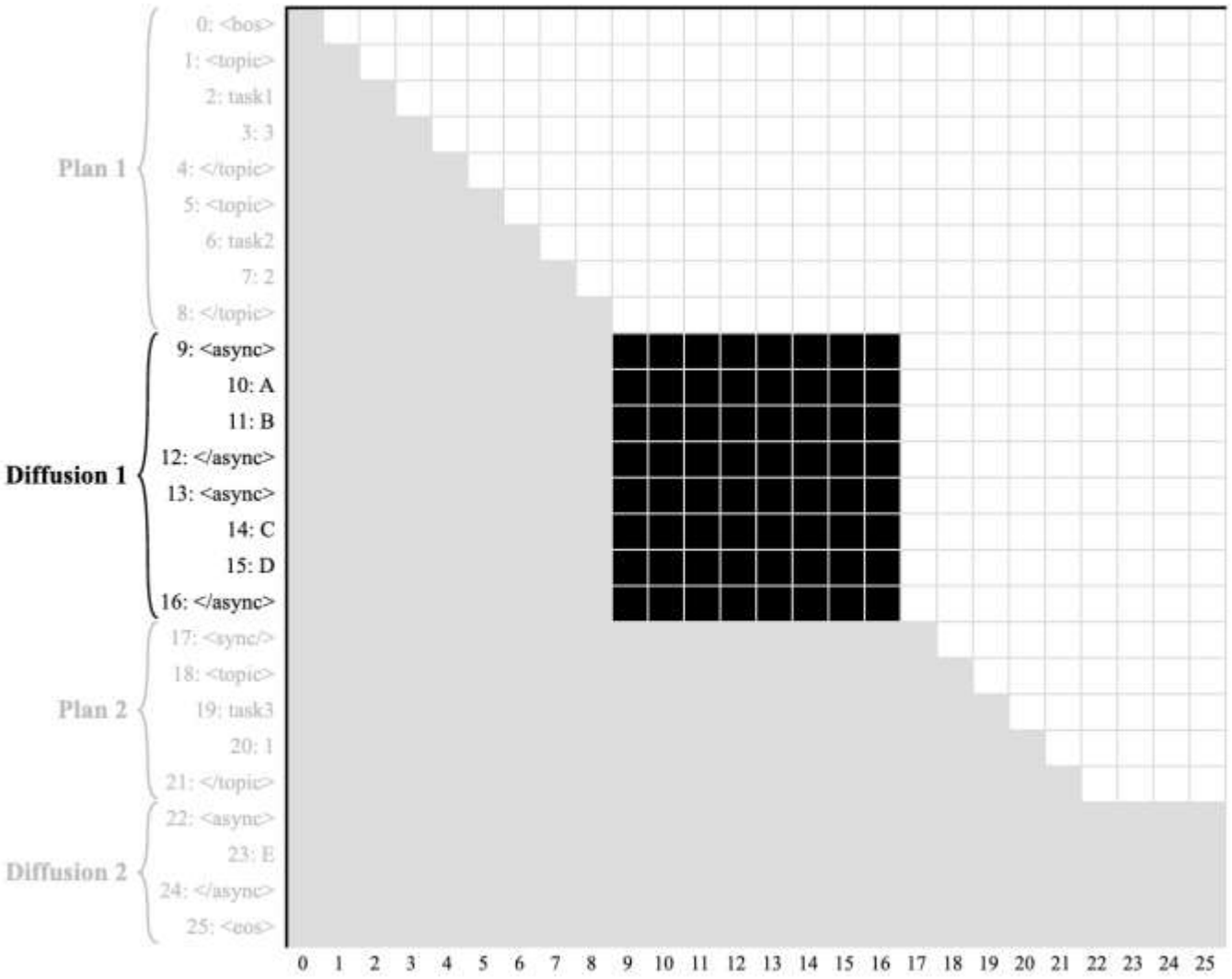
Attention Mask



Attention Mask



Attention Mask



Evaluation - Base Model

All experiments start from **Dream 7B** base model,

which is an AR LLM finetuned with a **diffusion objective.**

Model Variants

AR baseline

- Finetune Dream-7B with AR objective.

Diffusion baseline

- Finetune with same data as AR.
- Except using diffusion objective.

Planned diffusion

- Finetune using data with plan annotations.
- Use hybrid AR + diffusion objective.

Planned diffusion sparse attn variant

- Enforces parallel chunks do not attend to each other.
- Otherwise, the same as Planned Diffusion.

Pasta-SFT

- Supervised-finetuned version of PASTA

Skeleton-of-Thought (SoT)

- Prompting based async parallel decoding.
- Orchestrated from AR baseline.

Quality vs Latency

Baseline

Measure quality and speedup relative to autoregressive decoding baseline.

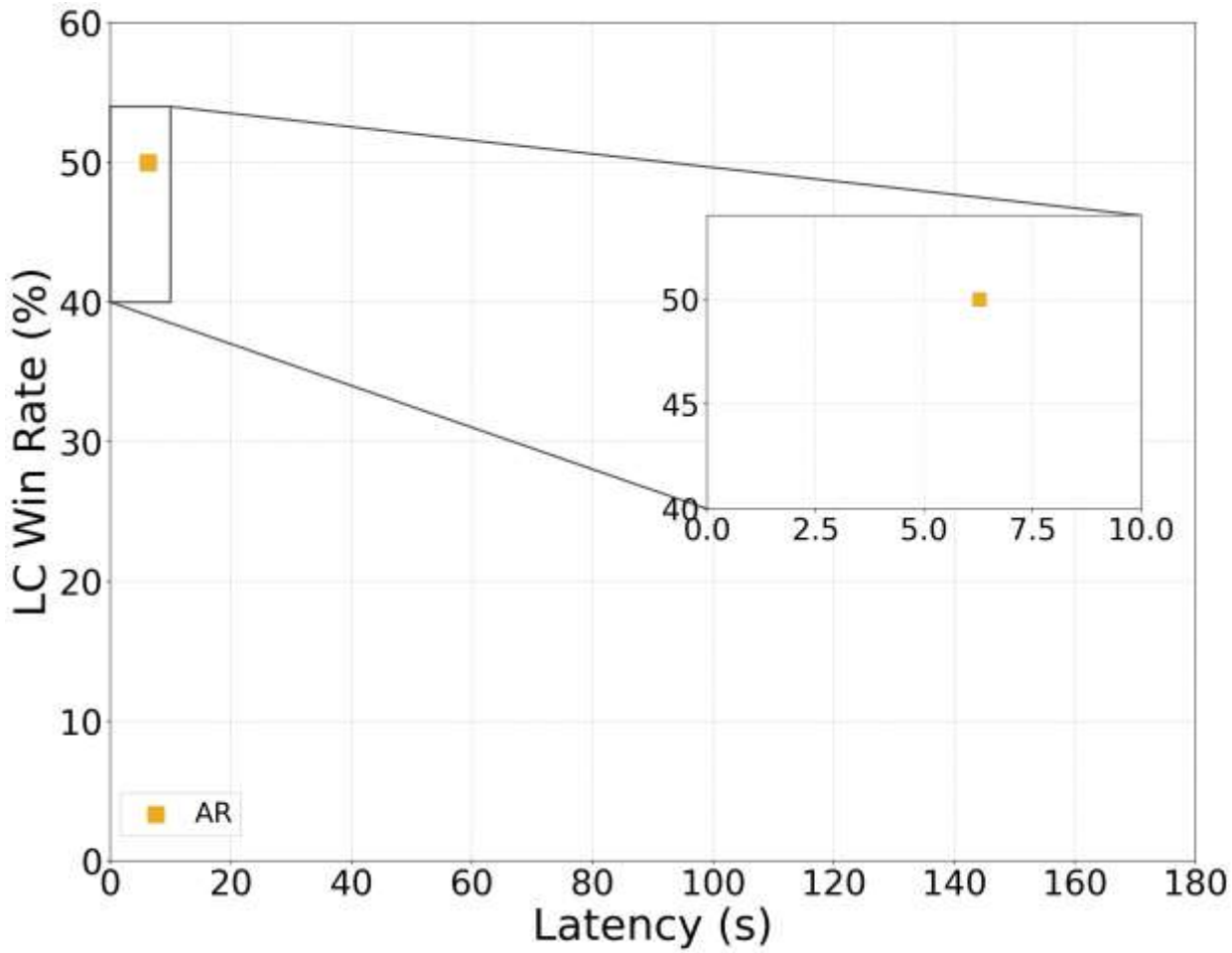
Quality

Length-controlled win rate on AlpacaEval, LLM-as-a-judge with GPT4.

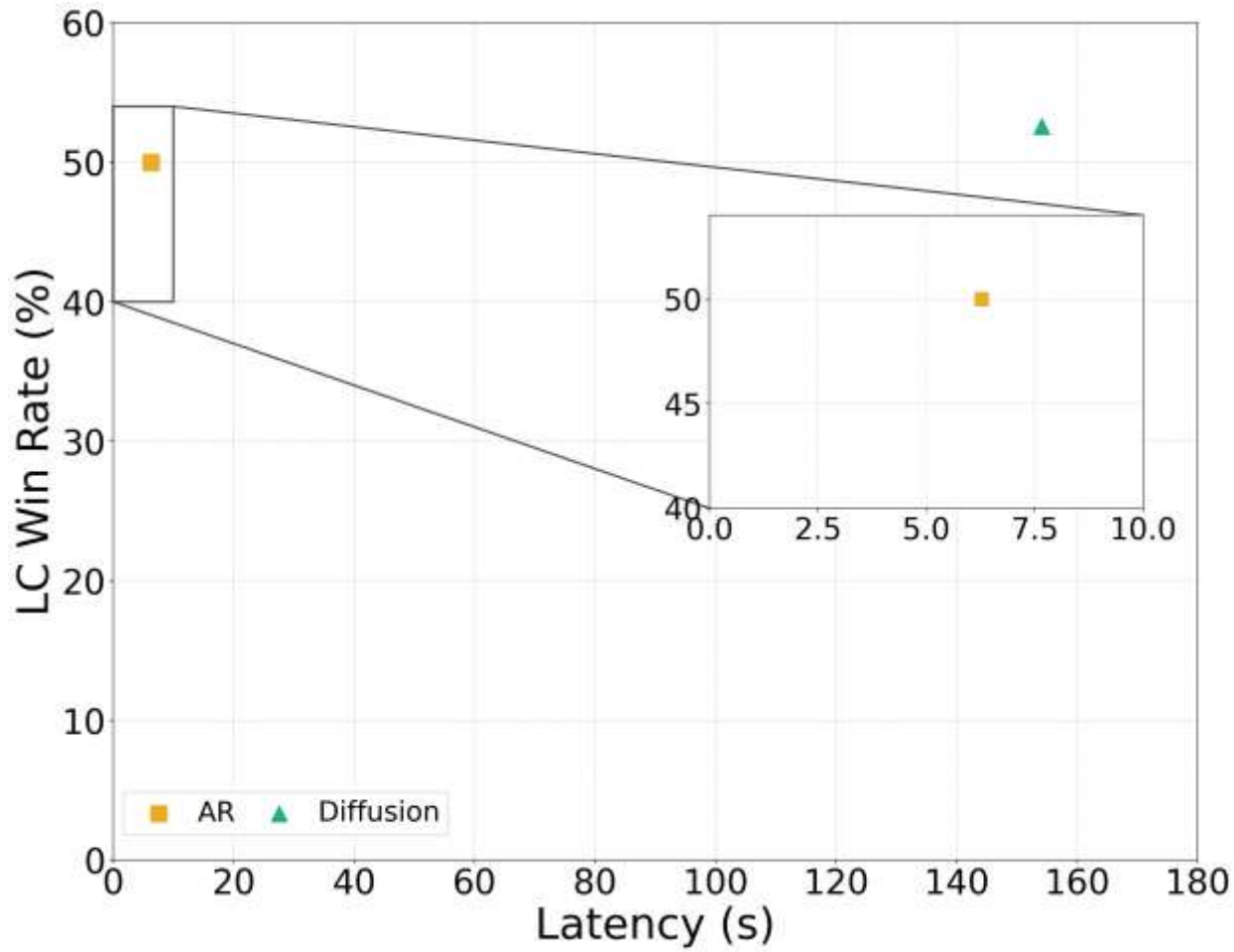
Latency

Decoding time with batch size 1.

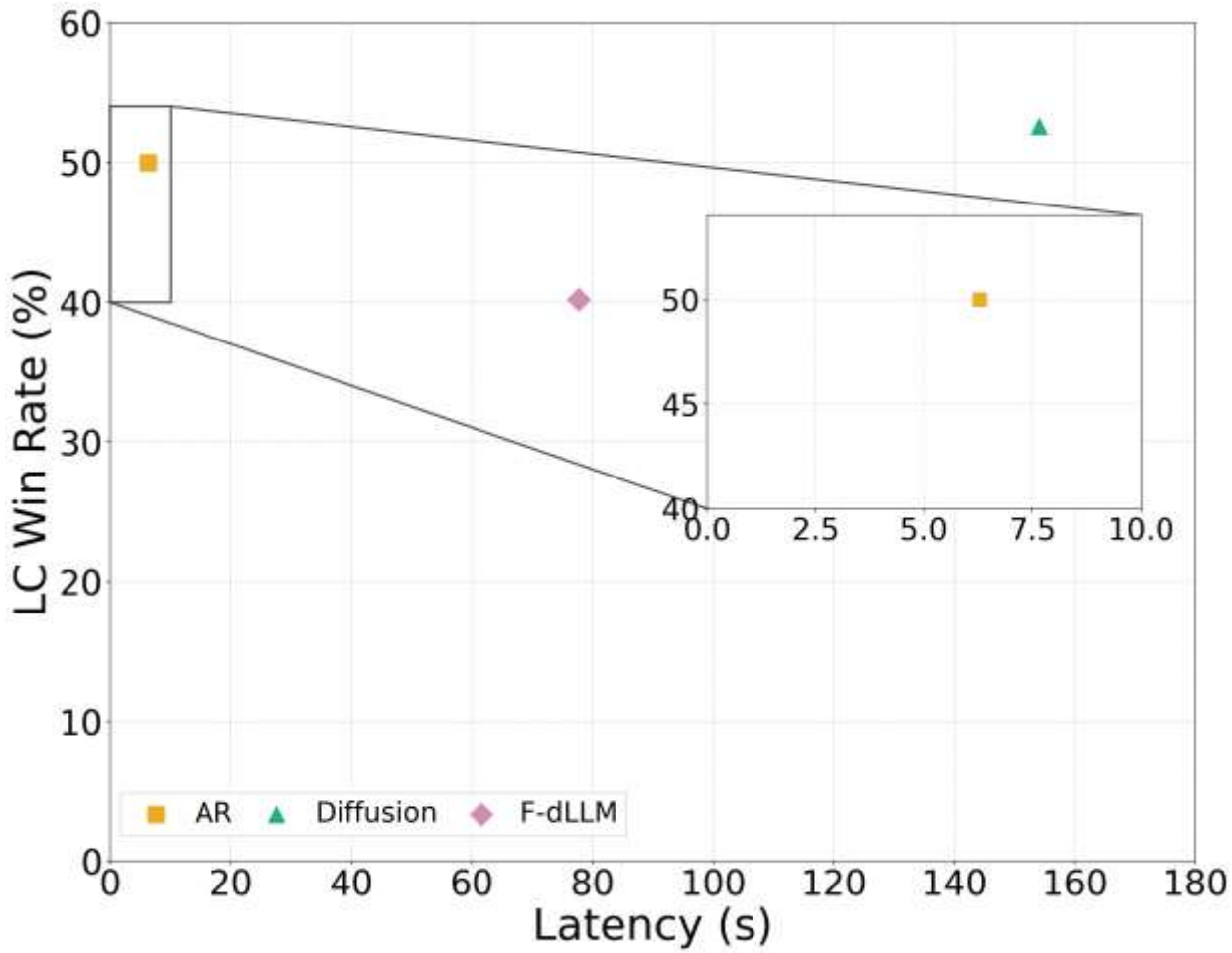
Results


LC Win Rate (%)
Latency (s)
AR
0
10
20
30
40
50
60
0
20
40
60
80
100
120
140
160
180
40
45
50
0.0
2.5
5.0
7.5
10.0

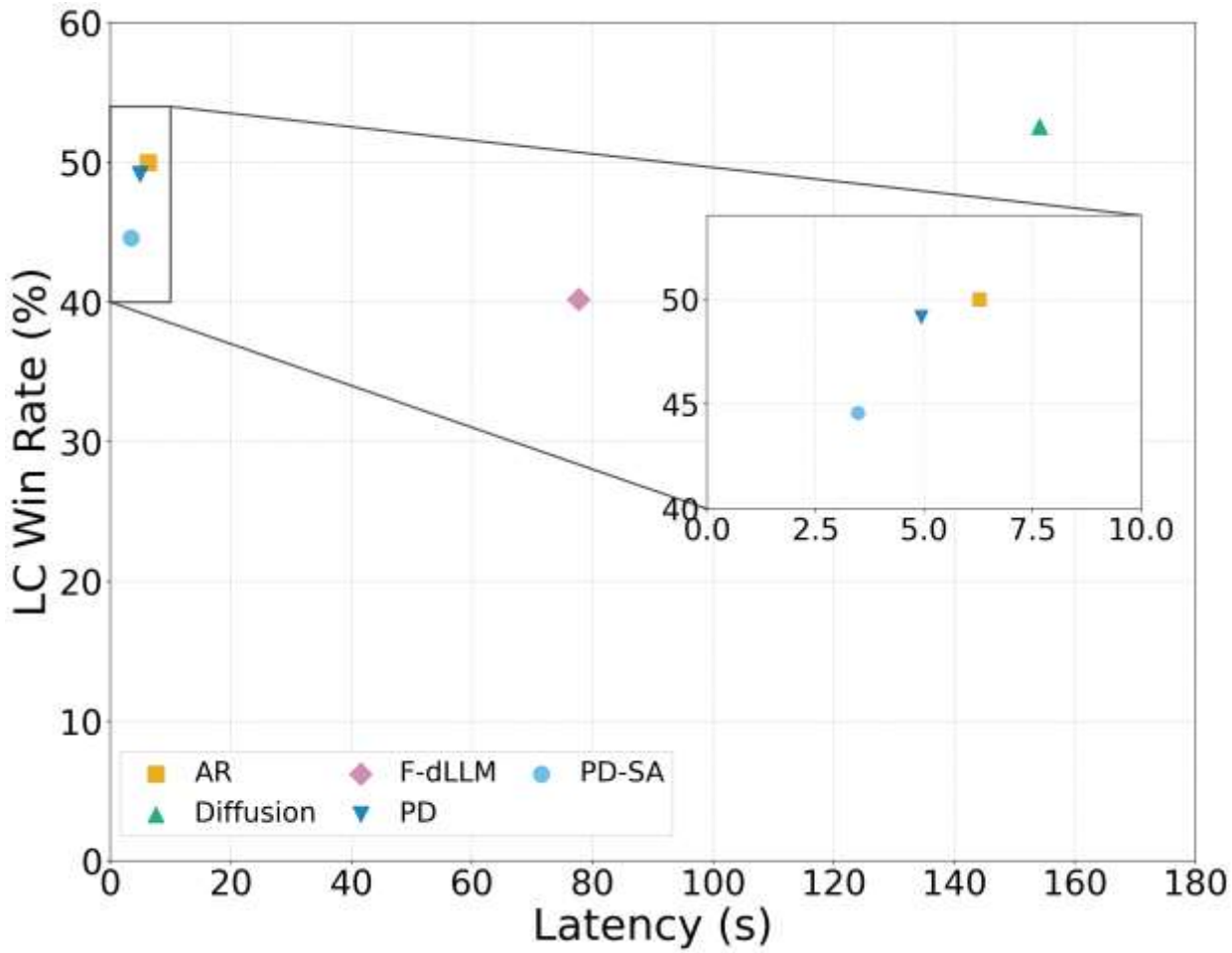
Results


LC Win Rate (%)
Latency (s)
AR
Diffusion
0
10
20
30
40
50
60
0
20
40
60
80
100
120
140
160
180
40
45
50
0.0
2.5
5.0
7.5
10.0

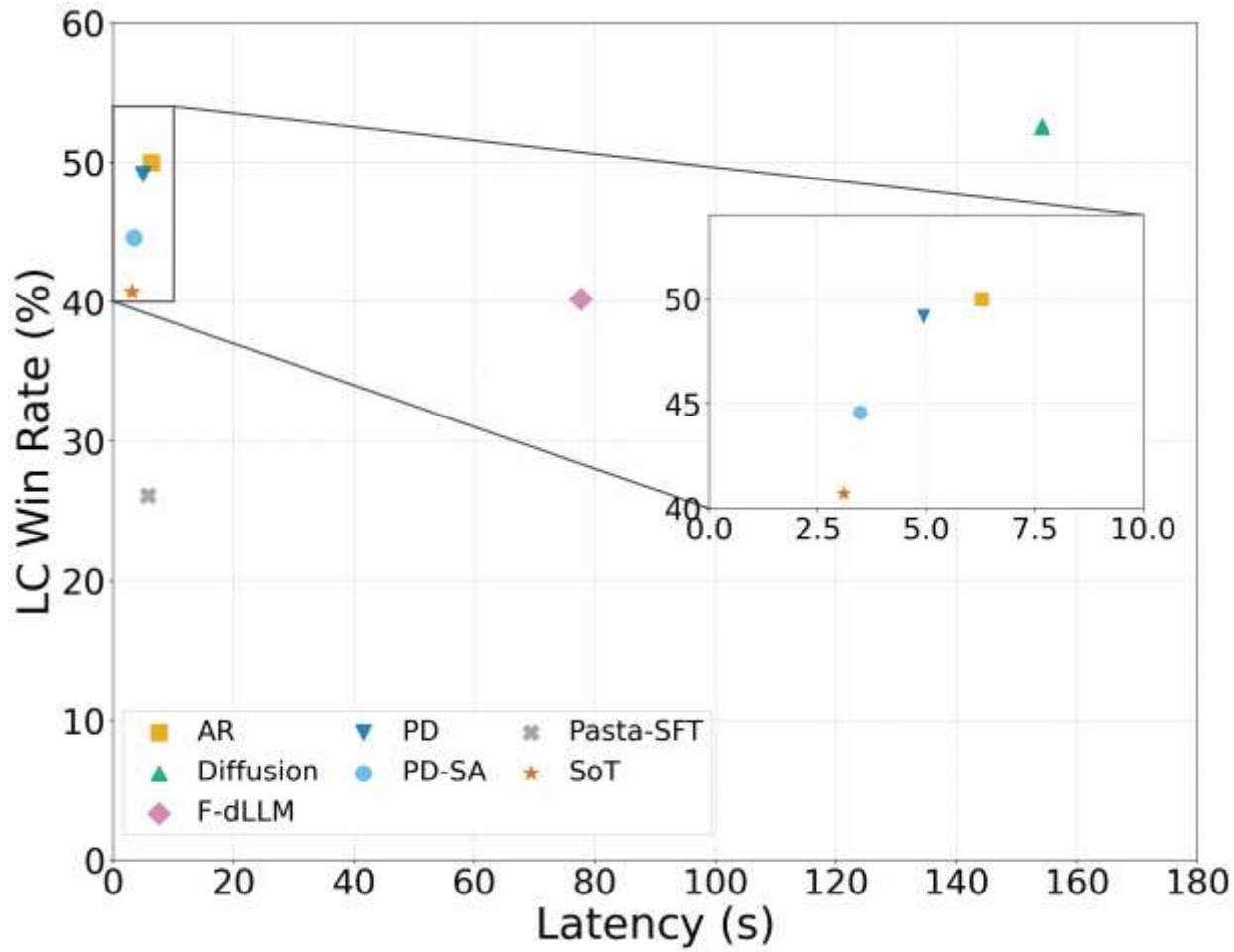
Results


LC Win Rate (%)
Latency (s)
AR
Diffusion
F-dLLM
0
10
20
30
40
50
60
0
20
40
60
80
100
120
140
160
180
40
45
50
0.0
2.5
5.0
7.5
10.0

Results



Results


LC Win Rate (%)
Latency (s)
AR
Diffusion
F-dLLM
PD
PD-SA
SoT
Pasta-SFT

Conclusion

- Designing good parallel denoising order is hard.

Conclusion

- Designing good parallel denoising order is hard.
- PD let dLLM decide its own denoising order.

Conclusion

- Designing good parallel denoising order is hard.
- PD let dLLM decide its own denoising order.
- PD achieves SoTA quality-latency trade-off.