

Diffusion Language Models

Efficient and flexible discrete diffusion using pretrained LLMs

***Ruurd J. A. Kuiper¹, Lars de Groot², Bram van Es²,
Maarten van Smeden¹, Ayoub Bagheri²***

¹University Medical Center Utrecht. ²Utrecht University.



UMC Utrecht



Universiteit
Utrecht

LoRA Adapted Diffusion

Efficient and flexible discrete diffusion using pretrained LLMs

***Ruurd J. A. Kuiper¹, Lars de Groot², Bram van Es²,
Maarten van Smeden¹, Ayoub Bagheri²***

¹University Medical Center Utrecht. ²Utrecht University.



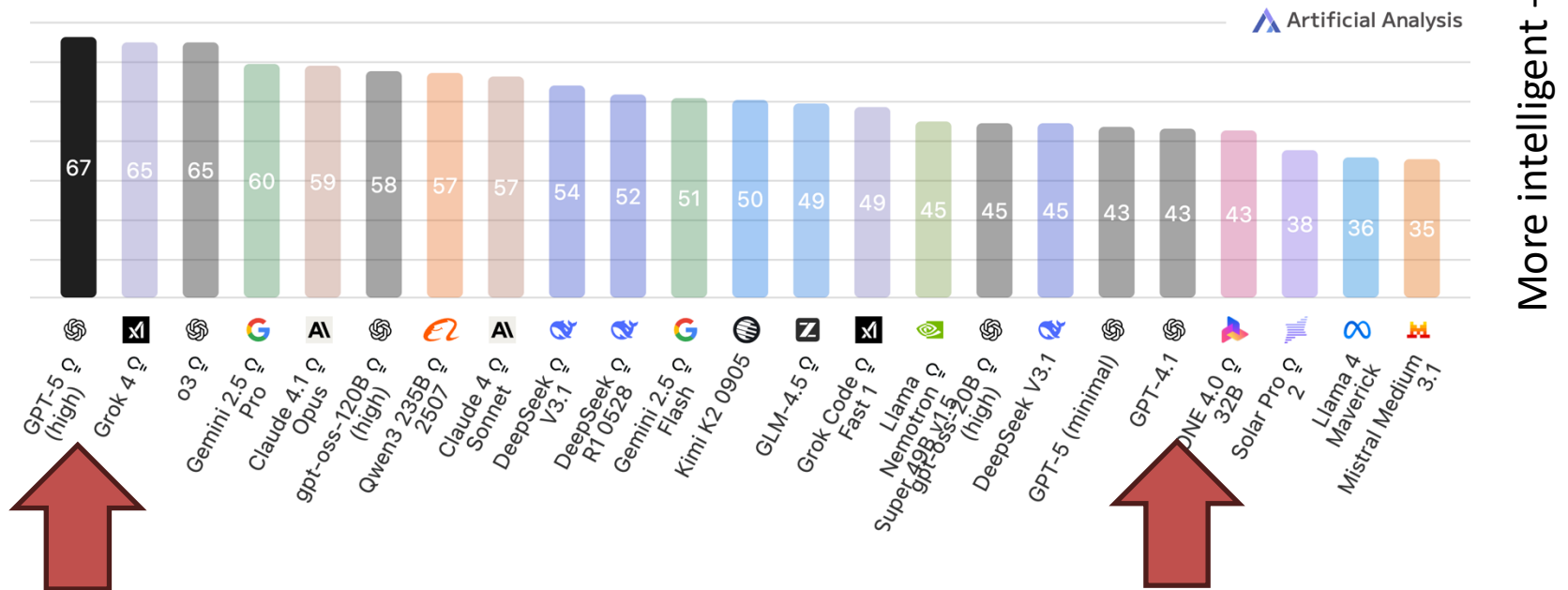
UMC Utrecht



Universiteit
Utrecht

Motivation

Little actual progress since GPT-4



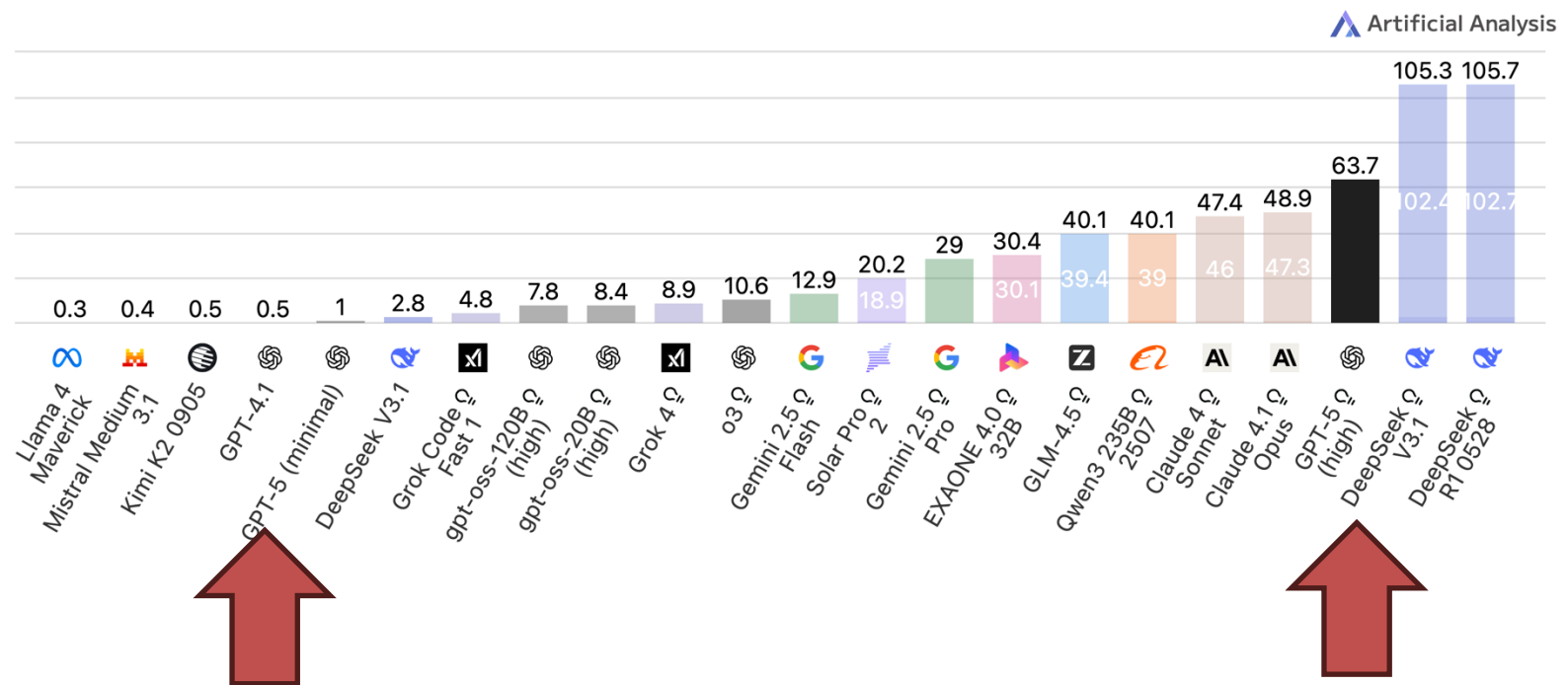
Motivation

Little actual progress since GPT-4

Latency: Time To First Answer Token

Seconds to First Answer Token Received; Accounts for Reasoning Model 'Thinking' time

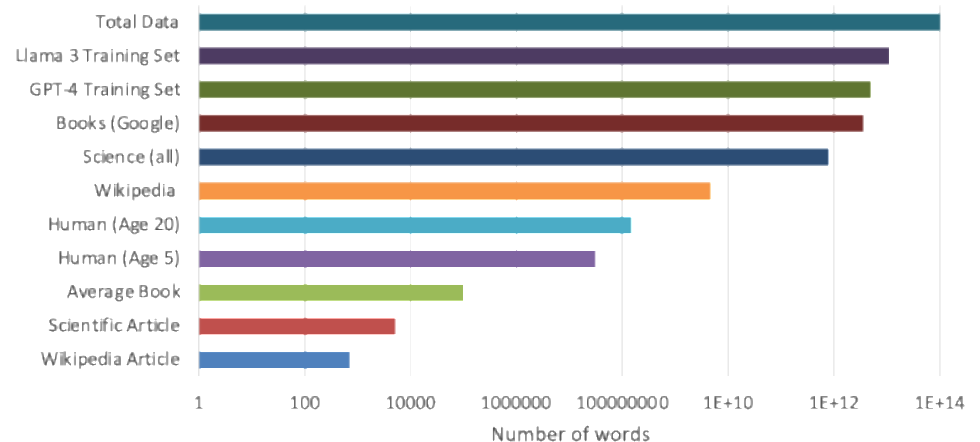
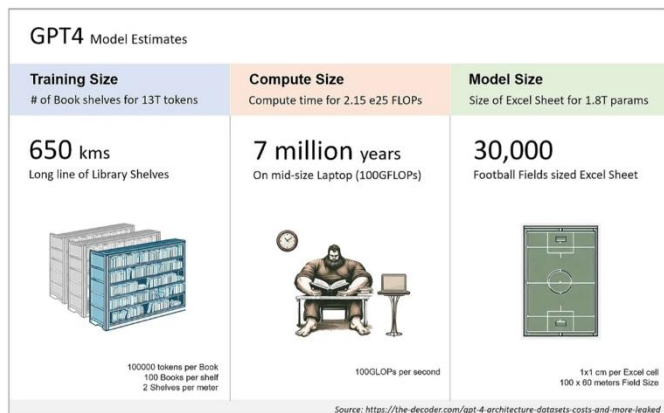
■ Input processing ■ Thinking (reasoning models, when applicable)



Motivation

A call to (academic) **action**

Scaling data, compute or model size is not going to save us...



So it's time to get creative!

Our proposal: diffusion for text

Why?

Pros:

- Speed
- Reasoning
- Error propagation
- Reverse curse
- Flexible inference time compute scaling

Cons:

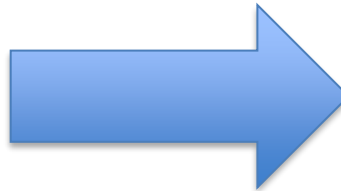
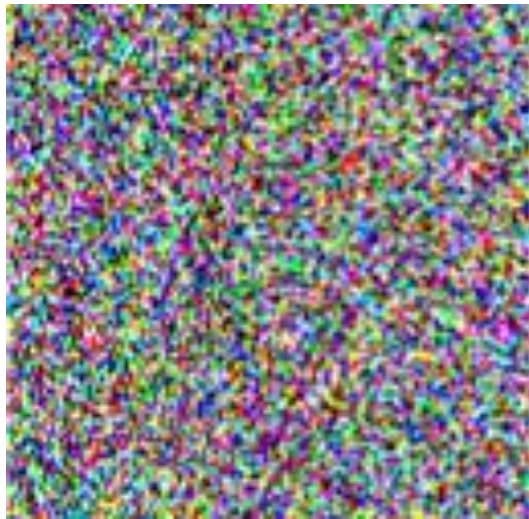
- Poorly studied and understood
- Fixed context length
- Can also get stuck on early guesses
- Suffer from grammatical errors

What are diffusion models?

Generating images from random noise

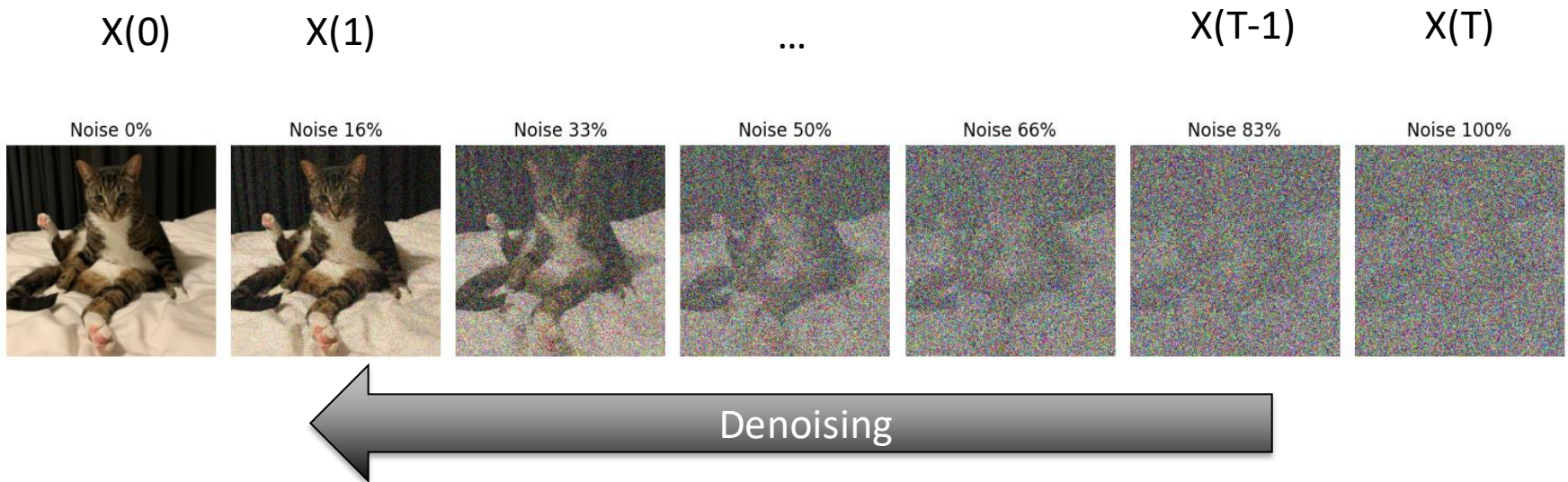
Prompt: Cat

+



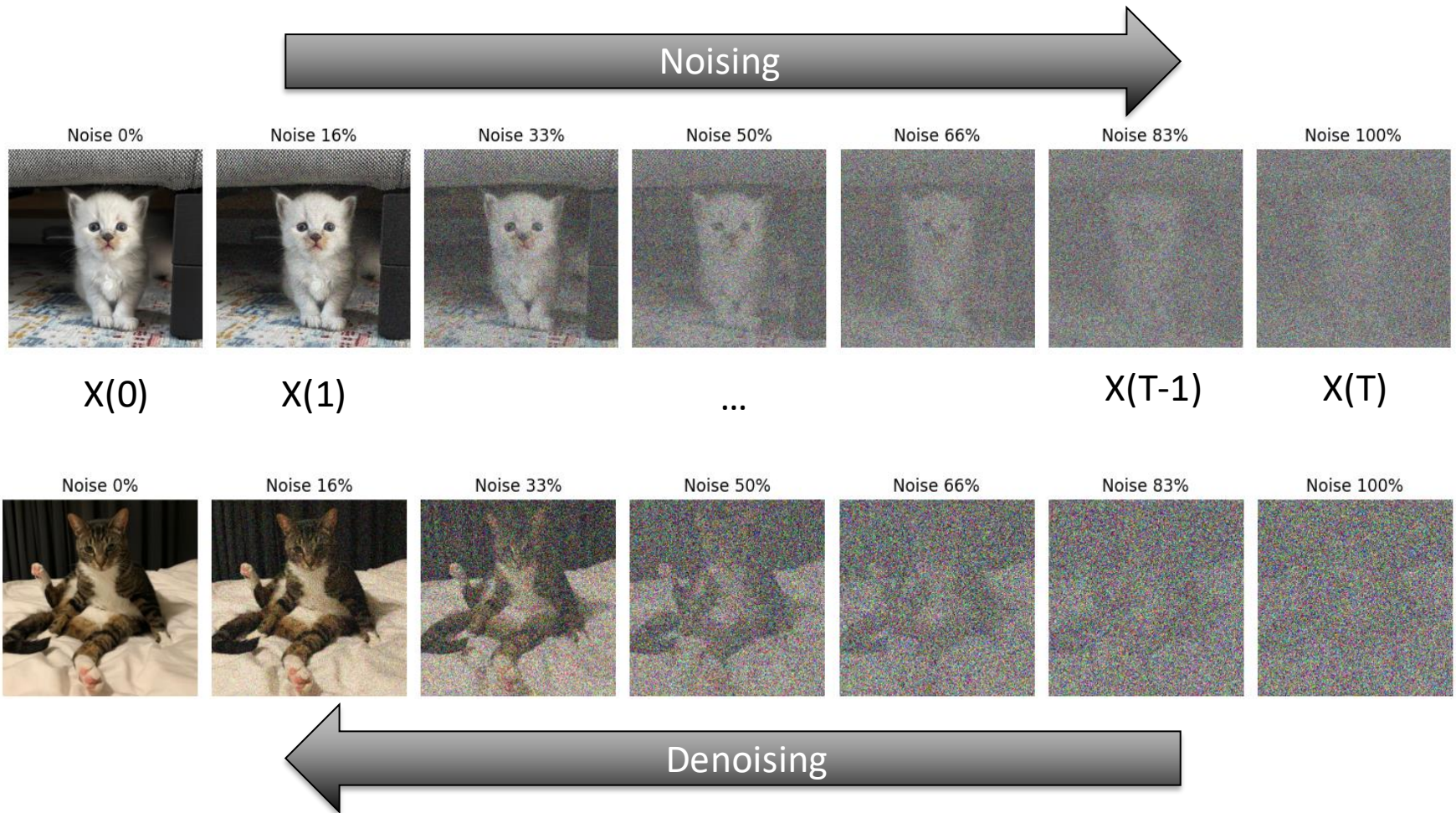
What are diffusion models?

Inference: Denoising



What are diffusion models?

Generating training data: Noising



What are Diffusion Language Models (DLMs)

Simply put:

Large language models generate text autoregressively (word-for-word)

Diffusion models generate images **iteratively** (by refining noisy images in multiple steps)

What are Diffusion Language Models (DLMs)

Simply put:

Large language models generate text autoregressively (word-for-word)

Diffusion models generate images **iteratively** (by refining noisy images in multiple steps)

Diffusion Language Models use the **iterative generation** of diffusion models to **refine noisy text** in multiple steps.

How do DLMs work?

Generating training data: noising text



Noise 0%

Noise 16%

Noise 33%

Noise 50%

Noise 66%

Noise 83%

Noise 100%

Once upon a time, there was a research group studying large language models.	Once upon a time, there [MASK] a research [MASK] studying large language models.	Once [MASK] [MASK] time, [MASK] was a research group studying large [MASK] models.	[MASK] upon a [MASK] [MASK] was a [MASK] [MASK] studying [MASK] language models.	[MASK] [MASK] a time, [MASK] [MASK] a research [MASK] [MASK] [MASK] [MASK] models.	[MASK] [MASK] [MASK] [MASK] [MASK] was [MASK] [MASK] [MASK] [MASK] large [MASK] models.	[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
--	--	--	--	--	---	--

How do DLMs work?

Inference: Denoising text



Noise 0%

Noise 16%

Noise 33%

Noise 50%

Noise 66%

Noise 83%

Noise 100%

Once upon a time, there was a research group studying large language models.	Once upon a time, there [MASK] a research [MASK] studying large language models.	Once [MASK] [MASK] time, [MASK] was a research group studying large [MASK] models.	[MASK] upon a [MASK] [MASK] was a [MASK] [MASK] studying [MASK] language models.	[MASK] [MASK] a time, [MASK] [MASK] a research [MASK] [MASK] [MASK] [MASK] models.	[MASK] [MASK] [MASK] [MASK] [MASK] was [MASK] [MASK] [MASK] [MASK] large [MASK] models.	[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
--	--	--	--	--	---	--

Noise 0%

Noise 16%

Noise 33%

Noise 50%

Noise 66%

Noise 83%

Noise 100%

A team of scientists have improved the capabilities of massive AI models through diffusion.	A team of scientists have [MASK] the [MASK] of massive AI models through diffusion.	[MASK] team [MASK] scientists have improved the capabilities of massive AI [MASK] [MASK] diffusion.	A [MASK] [MASK] [MASK] have improved [MASK] [MASK] of massive [MASK] [MASK] through diffusion.	[MASK] [MASK] [MASK] scientists have [MASK] [MASK] capabilities [MASK] massive [MASK] [MASK] through [MASK]	A [MASK] of [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] diffusion.	[MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK] [MASK]
---	---	---	--	---	--	--



So what does that look like?

Iteration 1/64 (after generation)

Am

Autoregressive (word-for-word) generation

Iteration 1/32 (before noising)

Amsterdam is the capital city the in in in the the Netherlands, the,, the. the.. the.,,,.,,,., the.,,,.,,,.,,,.,,,.,,,.

Diffusive generation

Our study

LoRA-Adapted Diffusion

Ruurd J.A. Kuiper¹, Lars de Groot¹, Bram van Es², Maarten van Smeden¹, Ayoub Bagheri²

¹ University Medical Center Utrecht, Utrecht

² Utrecht University, Utrecht

`r.j.a.kuiper@umcutrecht.nl`

Abstract. Autoregressive models dominate text generation but suffer from left-to-right decoding constraints that limit efficiency and bidirectional reasoning. Diffusion-based models offer a flexible alternative but face challenges in adapting to discrete text efficiently. We propose LAD (LoRA-Adapted Diffusion), a framework for non-autoregressive generation that adapts LLaMA models for iterative, bidirectional sequence refinement using LoRA adapters. LAD employs a structural denoising objective combining masking with text perturbations (swaps, duplications and span shifts), enabling full sequence editing during generation. We aim to demonstrate that LAD could be a viable and efficient alternative to training diffusion models from scratch, by providing both validation results as well as two interactive demos directly available online via: <https://ruurdkuiper.github.io/tini-lad/>

Keywords: Large Language Models, Diffusion

Methods

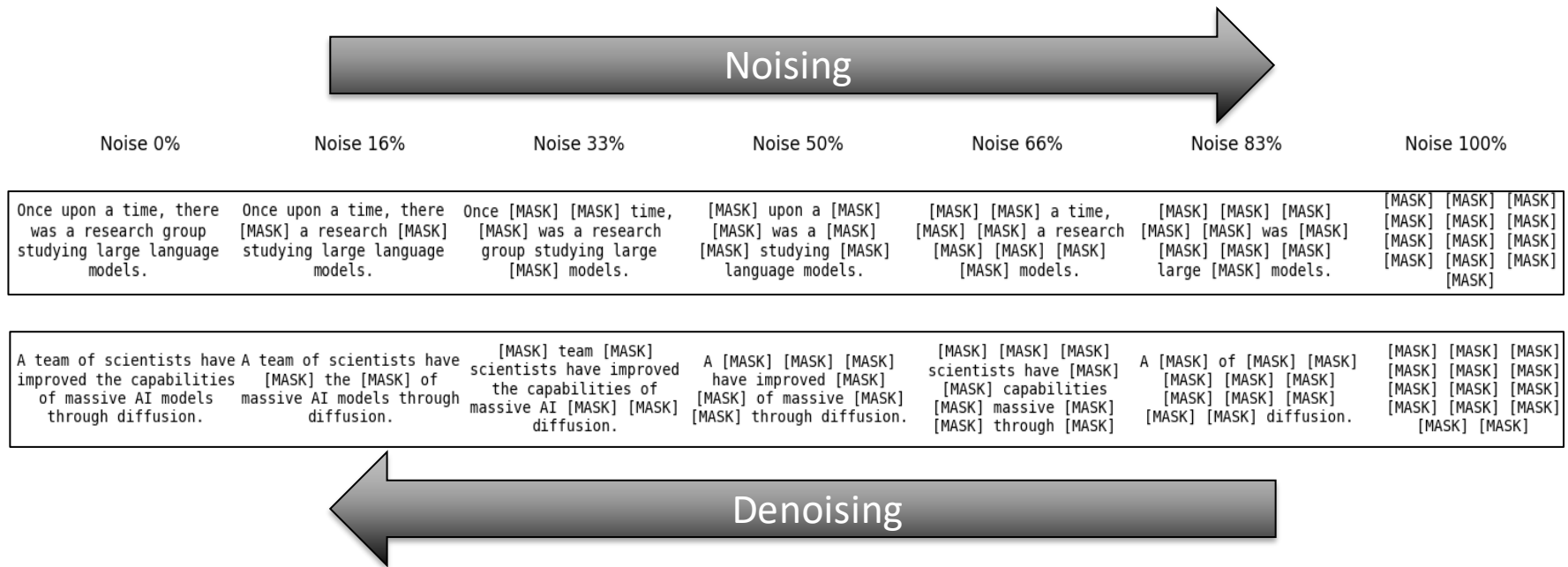
What did we do?

Two main contributions:

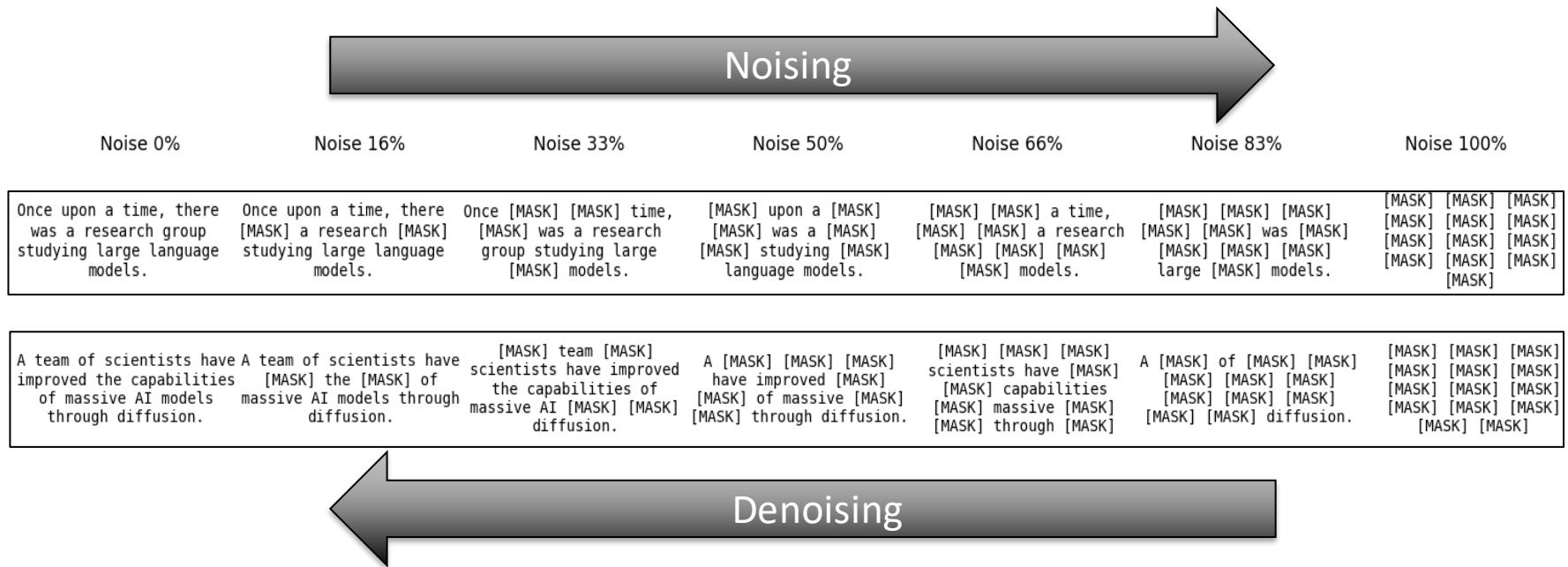
1. Using a novel noising schedule
2. Using LoRA for instruction fine-tuning

Methods

1. A novel training noising schedule



1. A novel training noising schedule



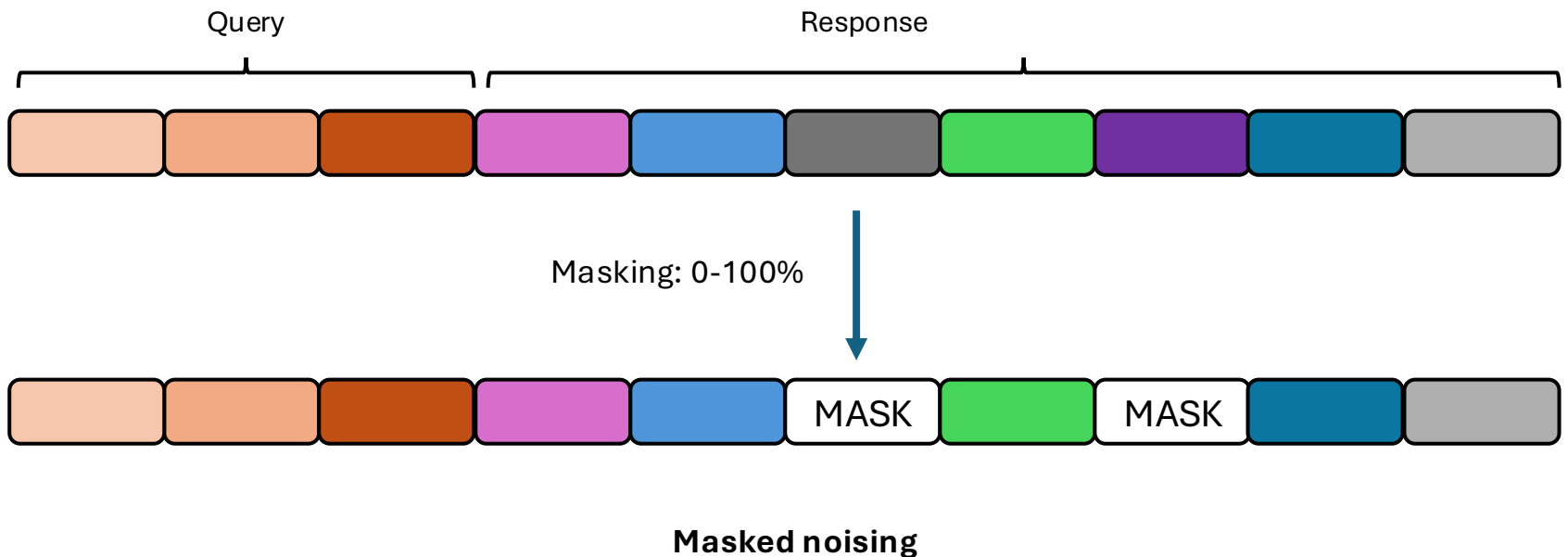
Simple masked diffusion has some shortcomings...

1. Many refinement iterations needed
2. Tokens that are set cannot be changed
3. A lot of information is in each step discarded

Methods

1. A novel training noising schedule

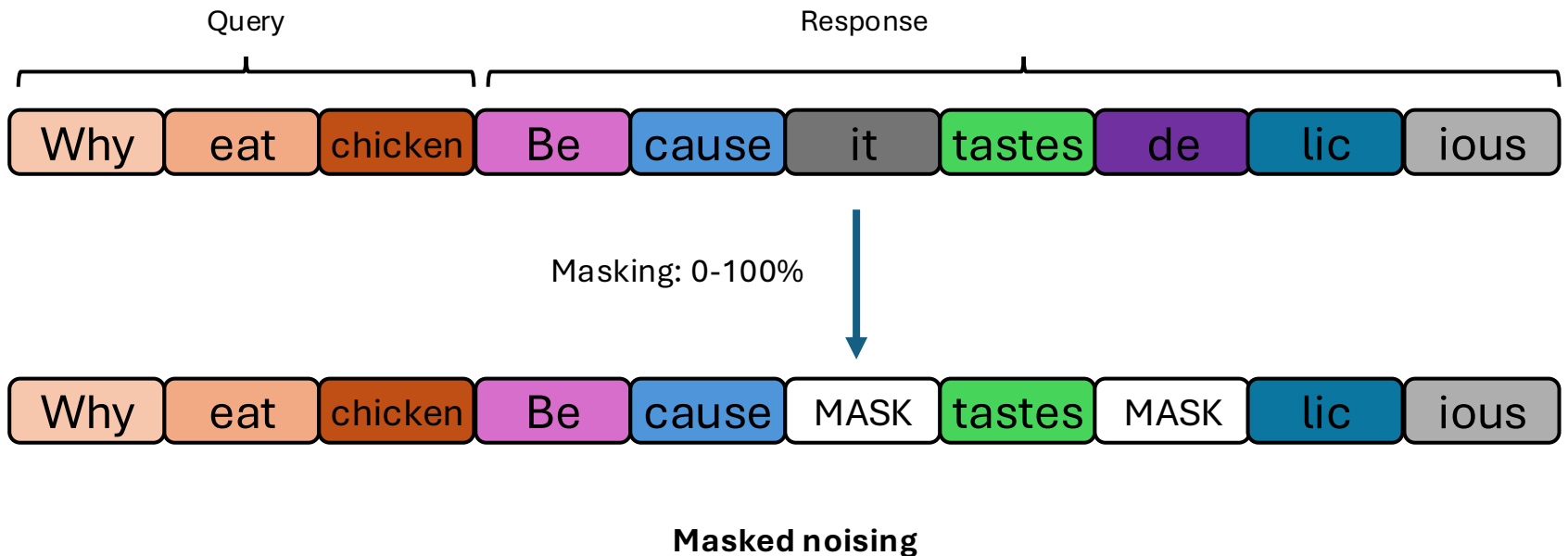
Generating training data: noising



Methods

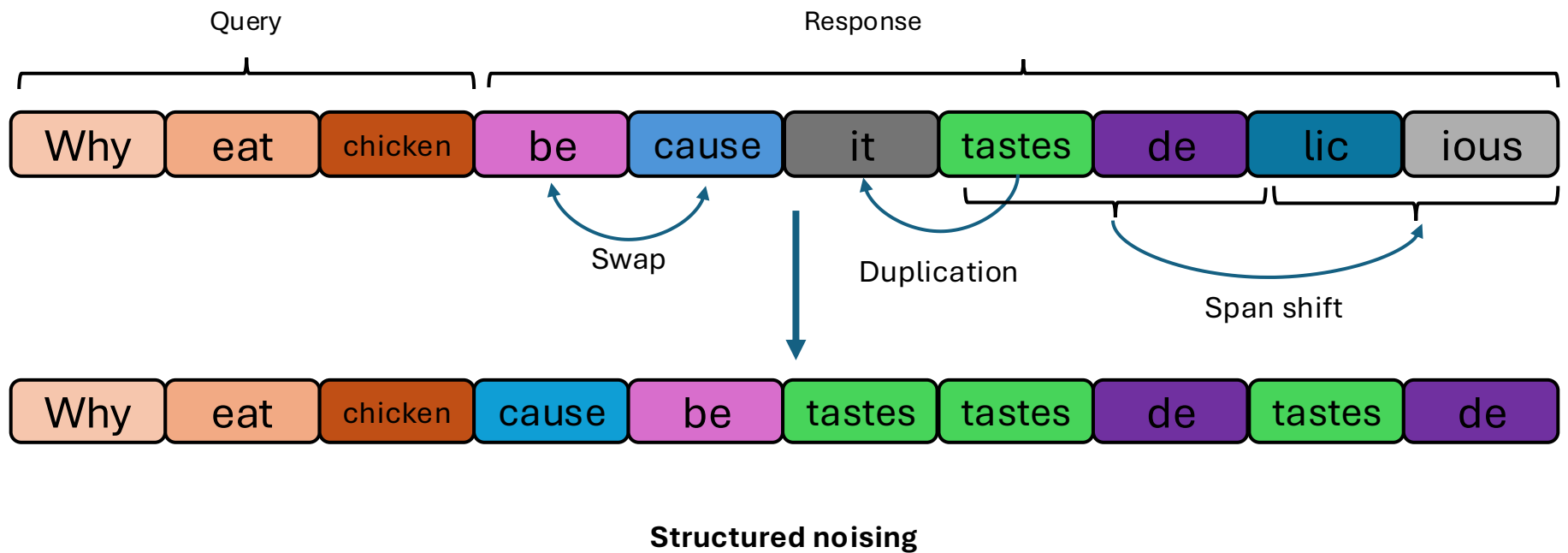
1. A novel training noising schedule

Generating training data: noising



Methods

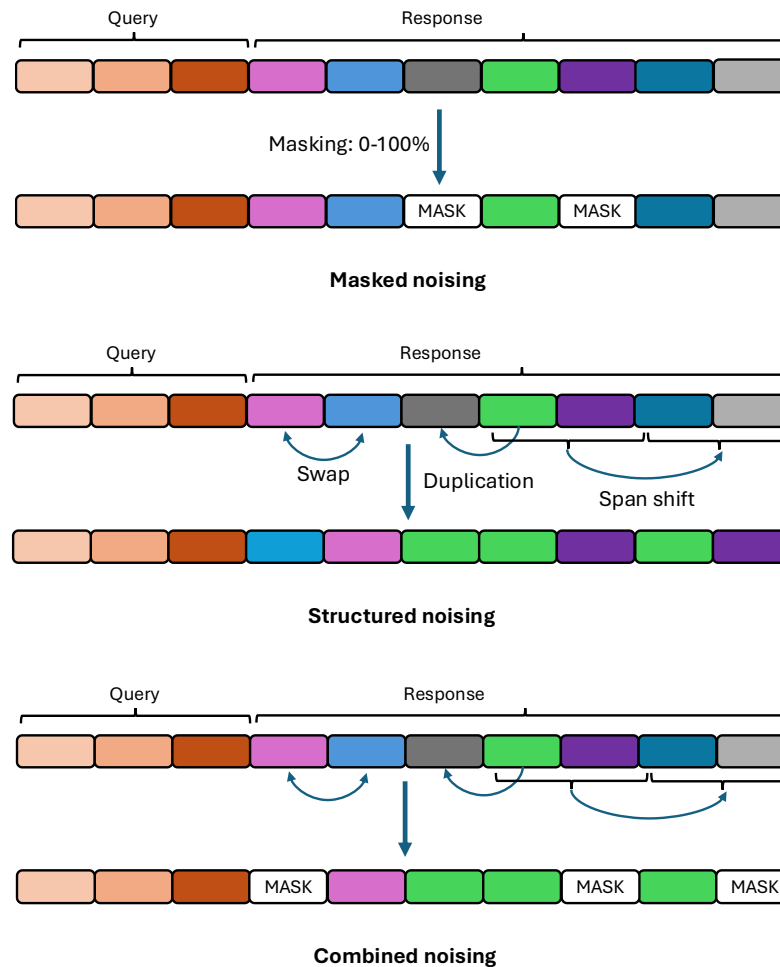
1. A novel training noising schedule



Methods



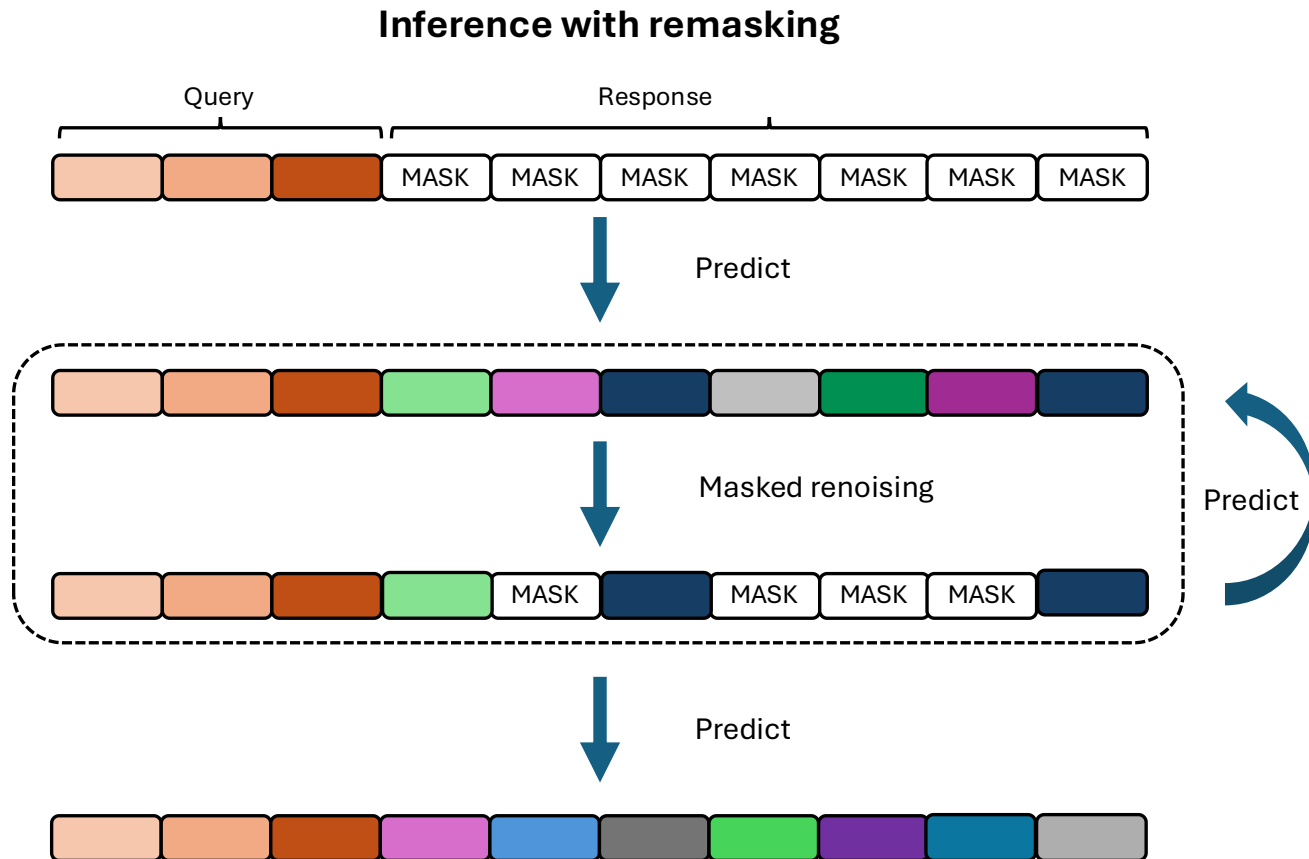
1. A novel training noising schedule



Methods



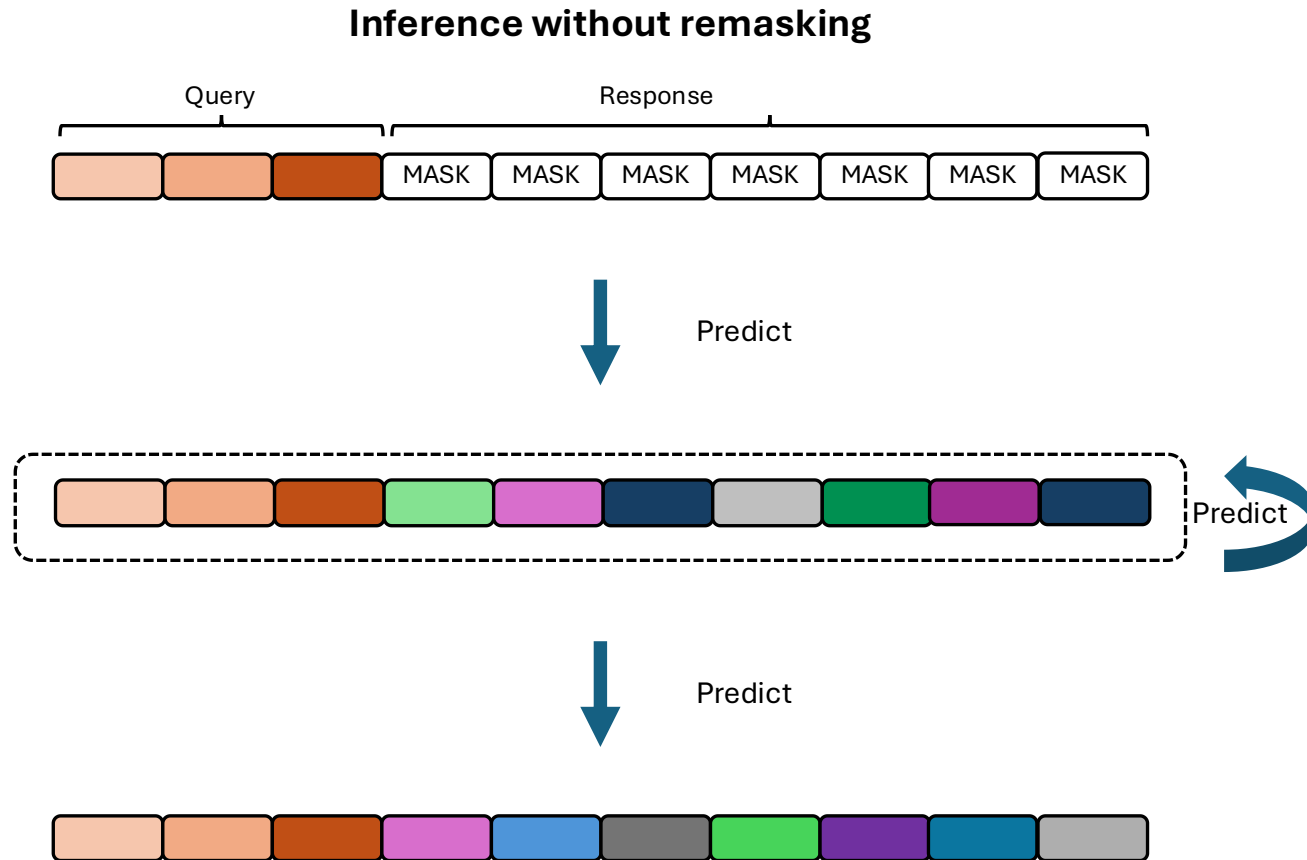
1.b. A novel inference method



Methods

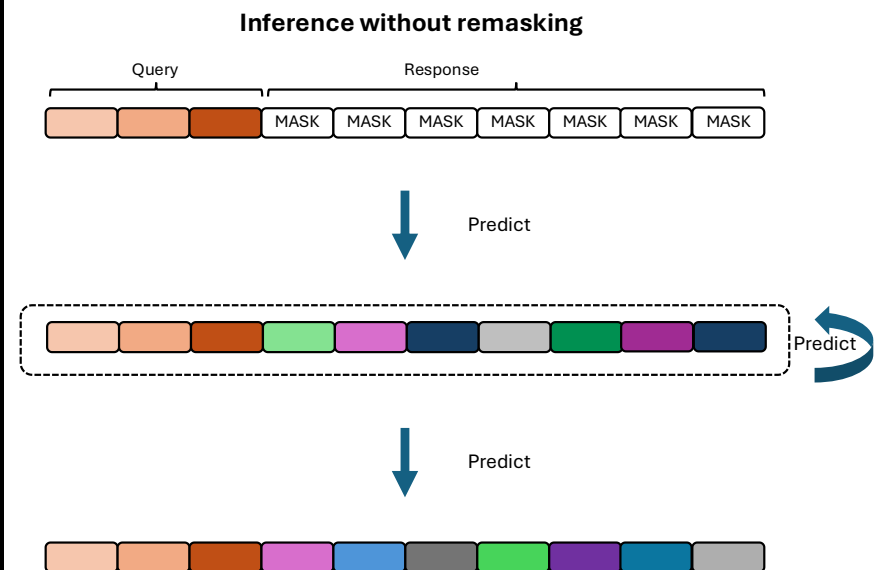
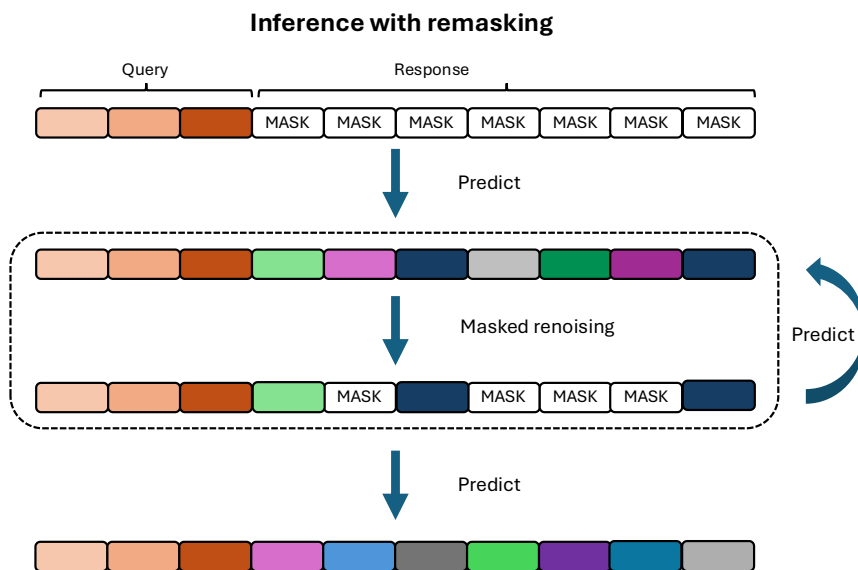


1.b. A novel inference method



Methods

1.b. A novel inference method



Methods

2. Using LoRA instruction fine-tuning

Pre-training:

- General next-word-prediction
- Billions of tokens
- Thousands of GPU hours (millions of euros)

Fine-tuning:

- Only question answering/instruction following
- A few hours on a single GPU

Methods

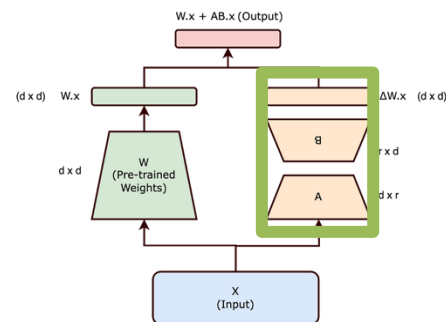
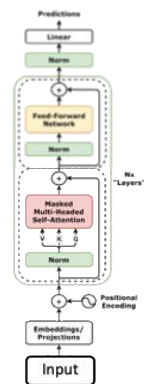
2. Using LoRA instruction fine-tuning

Pre-training:

- General next-word-prediction
- Billions of tokens
- Thousands of GPU hours (millions of euros)

Fine-tuning:

- Only question answering/instruction following
- A few hours on a single GPU



Methods

2. Using LoRA instruction fine-tuning

Pre-training:

- General next-word-prediction
- Billions of tokens
- Thousands of GPU hours (millions of euros)

Fine-tuning:

- Only question answering/instruction following
- A few hours on a single GPU

Model (LAD-)	Total params	Trainable params	Trainable (%)	LoRA Rank	Training (hours)
<i>1B-r128</i>	1.2B	13.6M	1.1	128	2.5
<i>1B-r512</i>	1.2B	54.4M	4.2	512	2.8
<i>3B-r512</i>	3.4B	146.8M	4.4	512	7.5
<i>3B-r1024</i>	3.5B	293.6M	8.4	1024	8.1
<i>8B-r1024</i>	8.5B	436.2M	5.1	1024	16.0
<i>8B-r2048</i>	8.9B	872.4M	9.8	2048	16.5

Methods

2. Using LoRA instruction fine-tuning

Pre-training:

- General next-word-prediction
- Billions of tokens
- Thousands of GPU hours (millions of euros)

Fine-tuning:

- Only question answering/instruction following
- A few hours on a single GPU

	LAD 8B (Ours)	LlaDa 8B	Llama 3.2 8B
Training* tokens	200 million 1x	2.3 trillion 11,500x	15 trillion 75,000x
Trainable params	436 million	8 billion	8 billion

* Both LAD (ours) and LlaDa only performed fine-tuning

Results

Does it work?



Results

Results

Benchmark results

	LAD 8B-1024 (Ours)	LlaDa	Llama 3.2 8B
ARC-Easy	93%	-	94%
GSM8K	44%	79%	80%
ARC-C	79%	89%	80%
MMLU	54%	66%	65%
HellaSwag	80%	74.6%	62%
<i>Perplexity</i>	<i>9.1</i>	-	<i>2.7</i>

Example: inference with remasking



	Masked	Output
Iteration 0	MASKMASKMASKMASKMASKMASKMASKMASK KMASKMASKMASKMASKMASKMASKMASKMASKMA SKMASKMASKMASKMASKMASKMASKMASKMASK ASKMASKMASKMASKMASKMASKMASKMASKMASK	Amsterdam is largest city city of, in the the in of famous..., the most of. famous isth known park, the, History the a9 famous., old
Iteration 4	Amsterdam is a MASK in the MASK MASK MASK MASK MASK history attractions, culture MASK MASK MASK It li MASK MASK MASK,, park MASK MASK architecture architecture and all MASK MASK MASK MASK li MASK the as for world MASK It beautiful city, with MASK house Anne Frank MASK MASK and MASK of as MASK MASK.. MASK	Amsterdam is a historic in the Netherlands of known for history history,, culture, and architecture. It is its,, park,, vibrant architecture, and all in. activities It's known the world world for its's city city, with the house of Anna Frank museum many its of as a as and.
Iteration 8	Amsterdam is MASK city in the Netherlands, known MASK its bikes MASK beautiful culture MASK and architecture. It has MASK als, MASKS and and MASK, and rich cuisine. MASK is considered one of world world for MASK its cityscape, including the Van of Frank MASK MASK and and attractions MASK art MASK..	Amsterdam is a city in the Netherlands, known for its beautiful, beautiful culture, and architecture. It has canals, canss and museums, and rich cuisine. Amsterdam is considered one of the world cities for its cityscape, including the Van Van Franklin Museum and numerous attractions and in art.
Iteration 16	Amsterdam is a MASK in the Netherlands, known for its art, literature, culture and culture. li MASK stunning architecture, can MASK MASK museums, and a diverse cuisine. Amsterdam is MASK one of the MASK cities for its nightlife MASK, including the famous area area MASKij, and famous for its MASK history.	Amsterdam is a city in the Netherlands, known for its art, literature, and and culture. It has stunning architecture, canals can museums, and a diverse cuisine. Amsterdam is also one of the popular cities for its nightlife scene, including the famous Red area Redij, and famous for its liberation history.
Iteration 32	Amsterdam is a city in the Netherlands, known for its architecture, history, and cultural heritage. It features stunning architecture, beautiful water canals, and MASK delicious cuisine. It is also one of the popular cities for its nightlife scene, with the famous Red Light District Quarter, and its famous LGBTQ+ liberation movement.	Amsterdam is a city in the Netherlands, known for its architecture, history, and cultural heritage. It features stunning architecture, beautiful water canals, and hearty delicious cuisine. It is also one of the popular cities for its nightlife scene, with the famous Red Light District Quarter, and its famous LGBTQ+ liberation movement.

Results



Example: inference without remasking

Query: What do you know about Amsterdam?

Without masking

Iteration 0

Amsterdam is largest city in Netherlands Netherlands country in capital,s the to, features of famous is. capital known stunning R is7 the, the architecture iconic the in famous countryt is

Iteration 4

Amsterdam is the largest city in the Netherlands Netherlands a country among Netherlandsss or capital capital and most most famous the the Rijple Street and its architecture architecture. The The Amsterdam canal is Amsterdam is and nightlife and as as its the art art. Additionally, remains a home home of of museums museums and many of beautiful houses of historical.

Iteration 8

Amsterdam is the largest city in the Netherlands, a country known among the popular capitals. It is also famous for the Rijks Square Square and its historic architecture. The city is also famous for its shopping, nightlife well as the art scene. Additionally, it remains a home for many artists, and is home of significant significant historical significance.

Iteration 16

Amsterdam is the largest city in the Netherlands, a country known among the popular capitals. It is also famous for the Rijksm Museum and its historic architecture. The city is also famous for its shopping, as well as the art scene. Additionally, it remains a home for many artists, and is home to sites of historical significance.

Conclusion

Current autoregressive methods:

- Inherent limitations due to left-to-right reasoning
- Diminishing returns on scaling compute, data and model size
- Novel architectures or paradigms are necessary

Diffusion has possible benefits:

- Faster inference
- Iterative refinement/error correction
- Scalable test-time compute

Current limitations:

- Grammatical errors
- Inconsistency
- Better evaluation is needed

Take home message

Please try novel network architecture, paradigms etc.!

Industry can polish and optimize current models by spending millions.

In academia we have the freedom to try things that are not directly profitable, so get creative!

Thank you

And now try it for yourself!



Interactive demo available:

www.tinyurl.com/tini-lad

or scan the QR-code

Iteration 1/32 (before noising)

Amsterdam is the capital city the in in in the the Netherlands, the,, the. the.. the.,,,.,,,.,,, the.,,,.,,,.,,,.,,,.