

# Autoregressive Language Models

—

Methods Program Meetup 11-9-2025

Ruurd Kuiper  
Julius Centrum – Datascience and Biostatistics



# Dutoregressieve Language Models

—

Methods Program Meetup 11-9-2025

Ruurd Kuiper  
Julius Centrum – Datascience and Biostatistics



# Ditorefressin Language Models

---

Methods Program Meetup 11-9-2025

Ruurd Kuiper  
Julius Centrum – Datascience and Biostatistics



# Dftfrerussin Language Models

—

Methods Program Meetup 11-9-2025

Ruurd Kuiper  
Julius Centrum – Datascience and Biostatistics



# Difrefousivn Language Models

—

Methods Program Meetup 11-9-2025

Ruurd Kuiper  
Julius Centrum – Datascience and Biostatistics



# Diffusion Language Models

—

Methods Program Meetup 11-9-2025

Ruurd Kuiper  
Julius Centrum – Datascience and Biostatistics



# Diffusion Language Models

—

Methods Program Meetup 11-9-2025

Ruurd Kuiper  
Julius Centrum – Datascience and Biostatistics



# Contents



DLMs?

- Why?
- What?
- How?



Methods



Results

# Diffusion Language Models (DLMs)



- **Why** try different language models architectures than we use now?



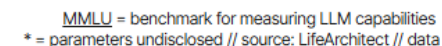
DLMs?

- Why?
- What?
- How?

## No notable progress since GPT-4

ranked by capabilities, sized by billion parameters used for training

show only: all



# Motivation

But... GPT-5 is out now!



# Motivation

But... GPT-5 is out now!



So we've got AGI now right?

# Motivation

But... GPT-5 is out now!

---

So we've got AGI now right?



# Motivation

But... GPT-5 is out now!

---

So we've got AGI now right?



# Motivation

But... GPT-5 is out now!

---

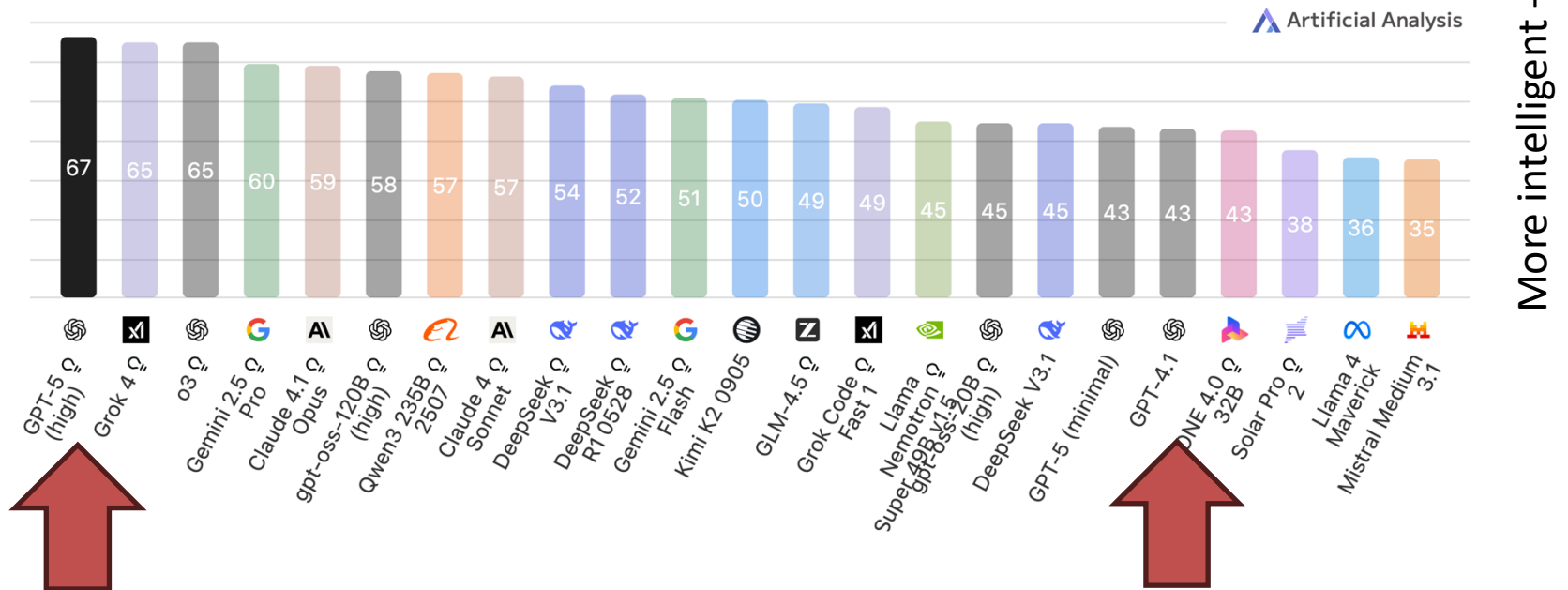
So we've got AGI

SCIENCE NEWS



# Motivation

No notable progress since GPT-4?



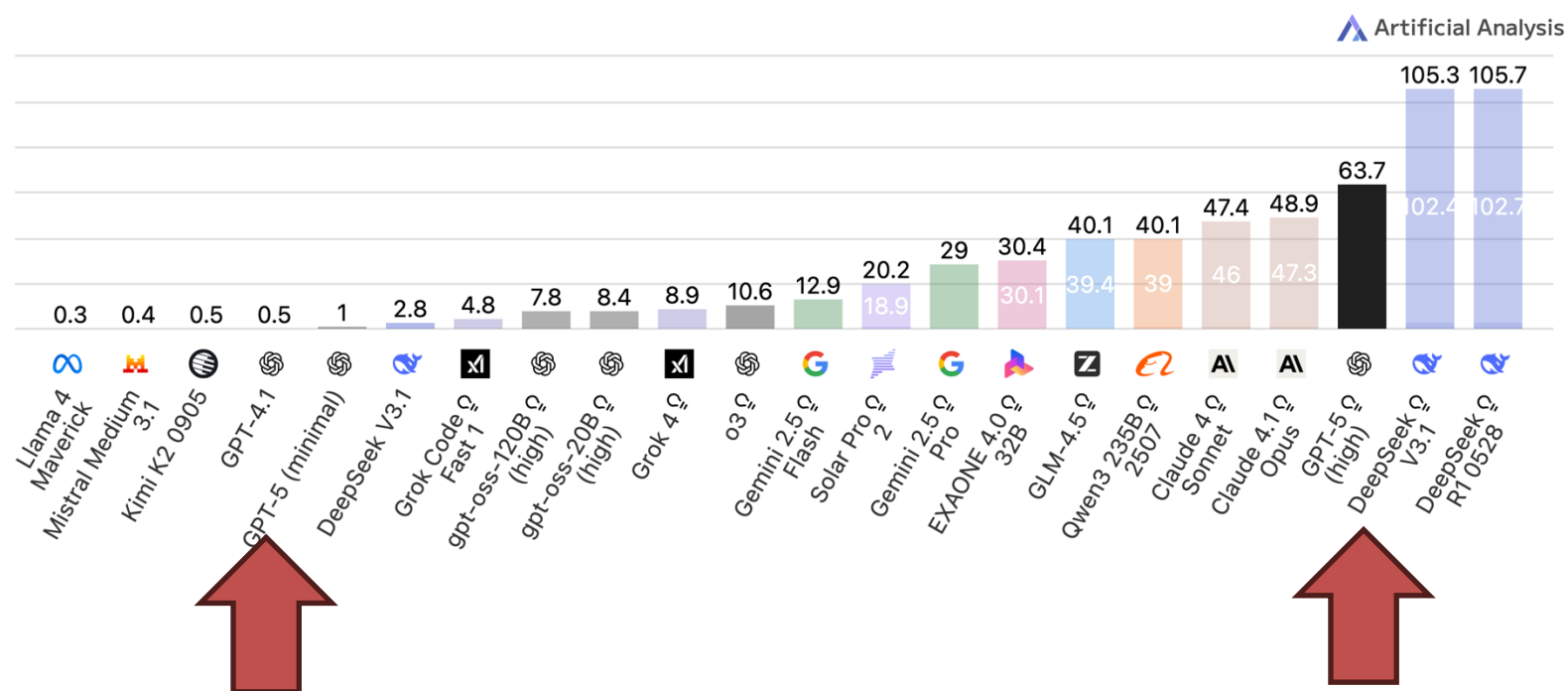
# Motivation

No notable progress since GPT-4?

## Latency: Time To First Answer Token

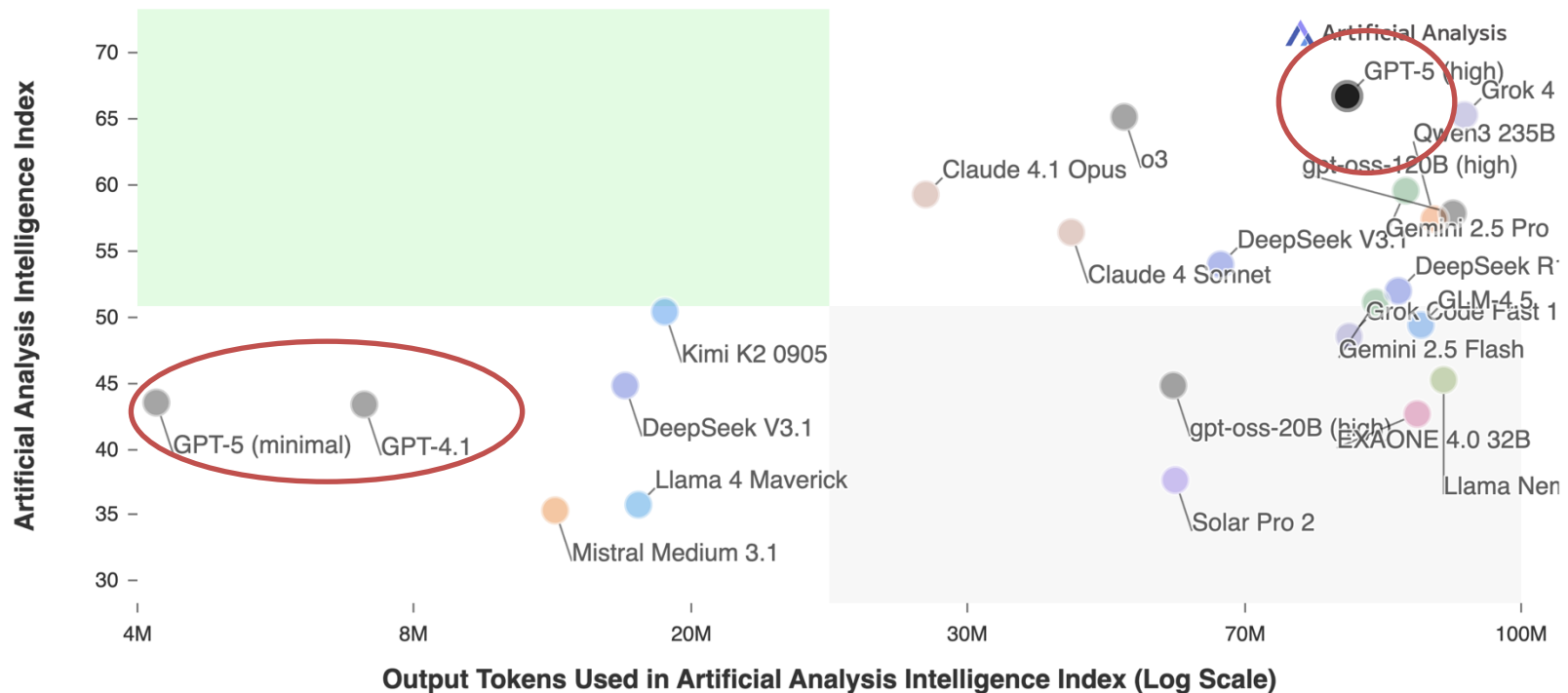
Seconds to First Answer Token Received; Accounts for Reasoning Model 'Thinking' time

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



# Motivation

No notable progress since GPT-4?



# Diffusion Language Models (DLMs)



- **Why** try different language models than we use now?
- **What** are (diffusion) language models?



DLMs?

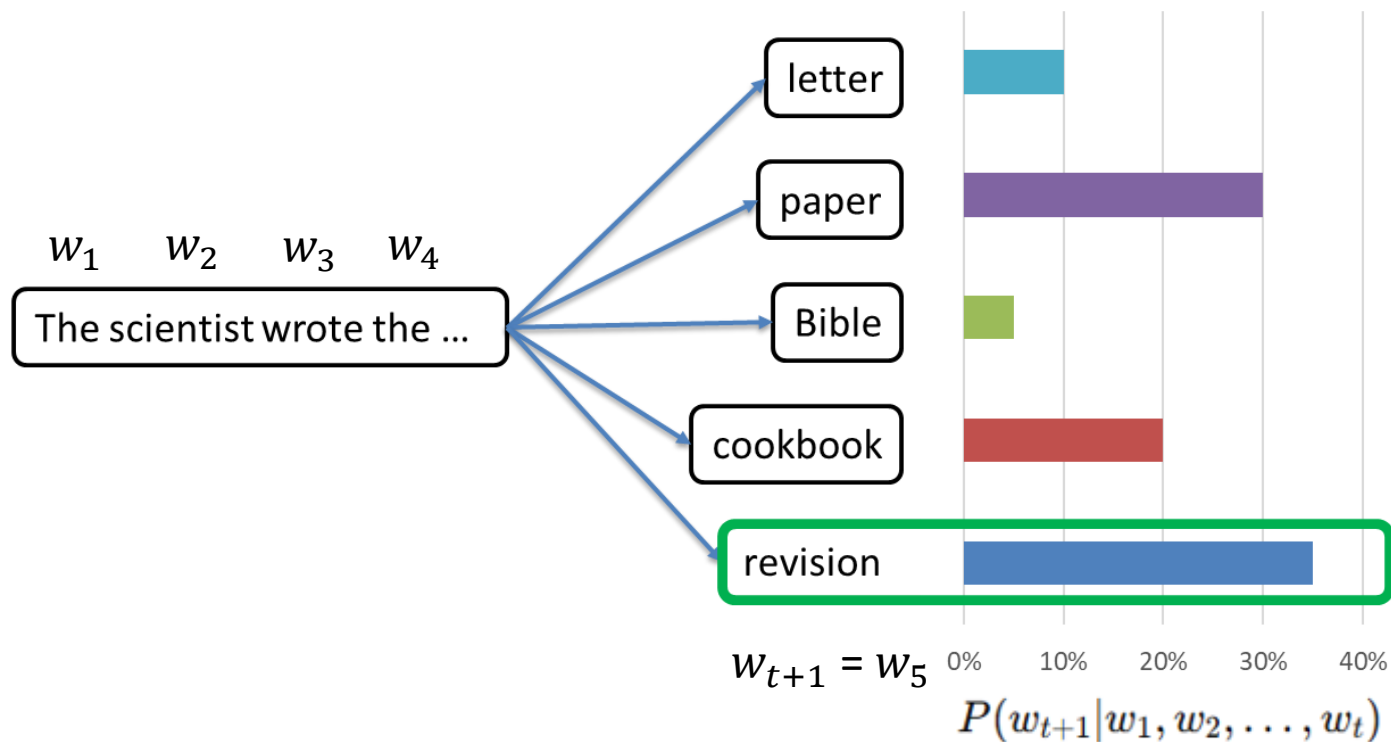
- Why?
- What?
- How?

# What are LLMs?



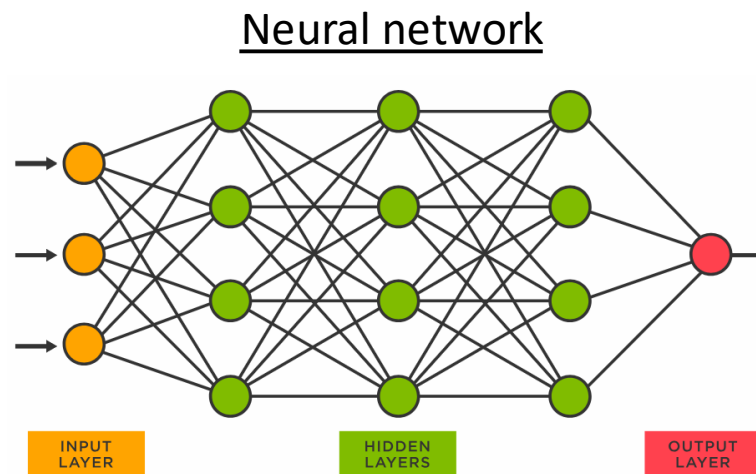
## What do they do: **Next-word-prediction**

"The process of estimating the conditional probability distribution  $P(w_{t+1}|w_1, w_2, \dots, w_t)$ , where  $w_1, w_2, \dots, w_t$  are the preceding words in a sequence, and  $w_{t+1}$  represents the next word to be predicted."



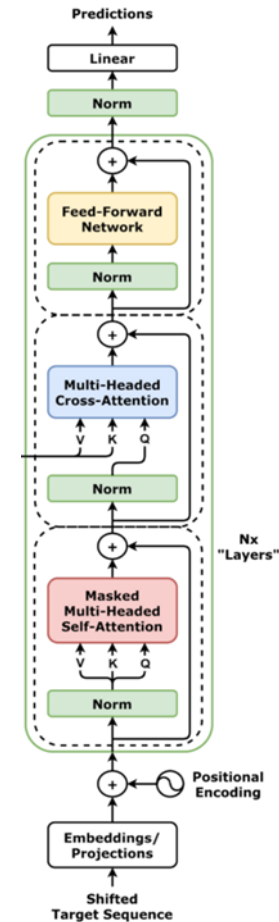
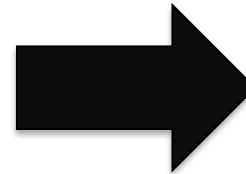
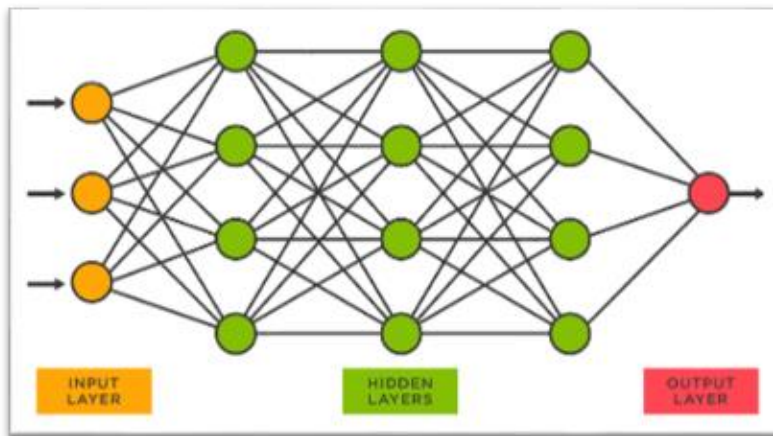
# What are LLMs?

Predictor



# What are LLMs?

## Transformer architecture



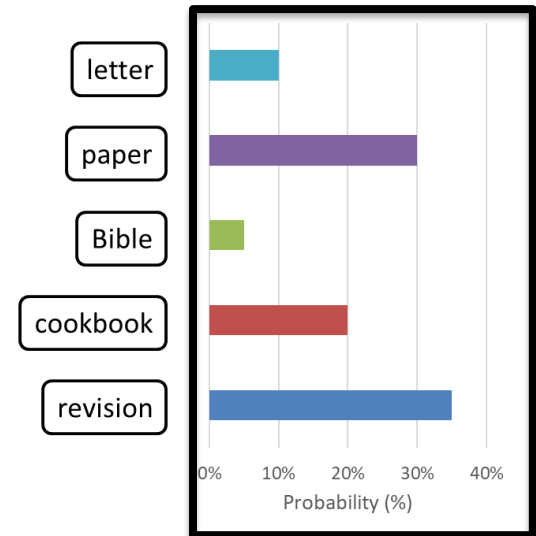
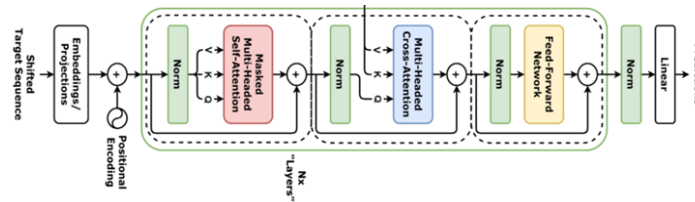
# What are LLMs?

## Prediction



### (Transformer) Neural network

The scientist wrote the ...



### Prediction

# What are LLMs?



## Considerations

---

### Pros:

- Well-studied
- Easy to learn goal (only need to predict the next word)
- Efficient learning

### Cons:

- Speed
- Reasoning
  - Chain-of-thought was promising, but limited and slow
- Error propagation
- Reverse curse (A is B, B is A?)

# What are diffusion models?

---



# What are diffusion models?

You might know it from image generation..



**DALL-E 3**



**MIDJOURNEY 5.2**



**STABLE XL**



# What are diffusion models?

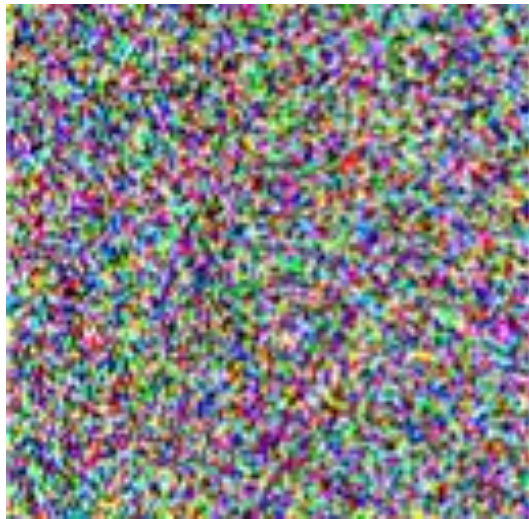
OpenAI Sora



# What are diffusion models?

Generating images from random noise

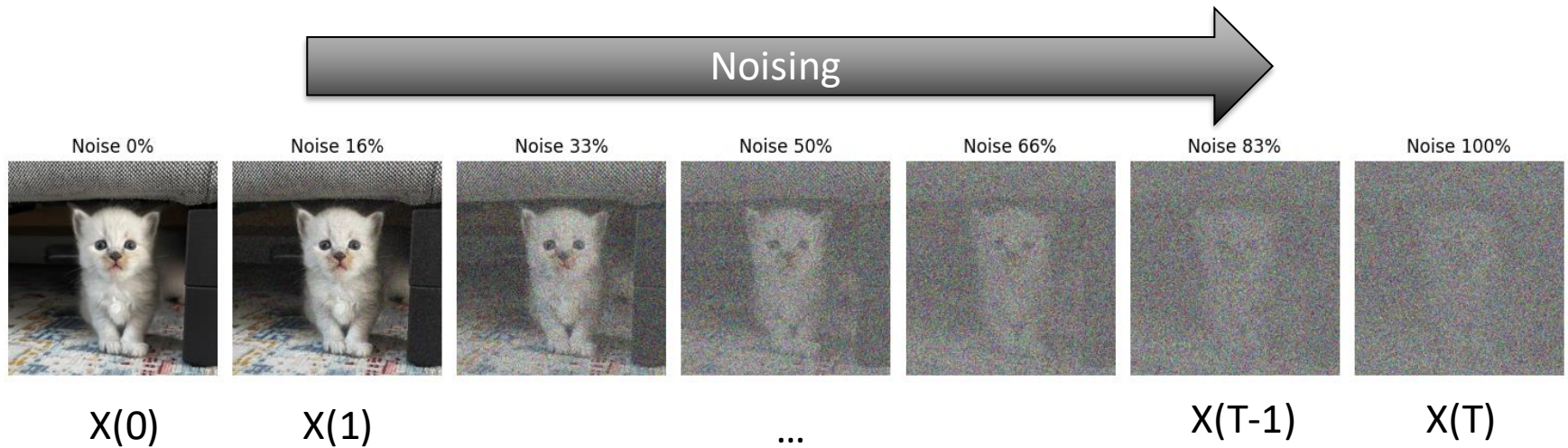
---



# What are diffusion models?



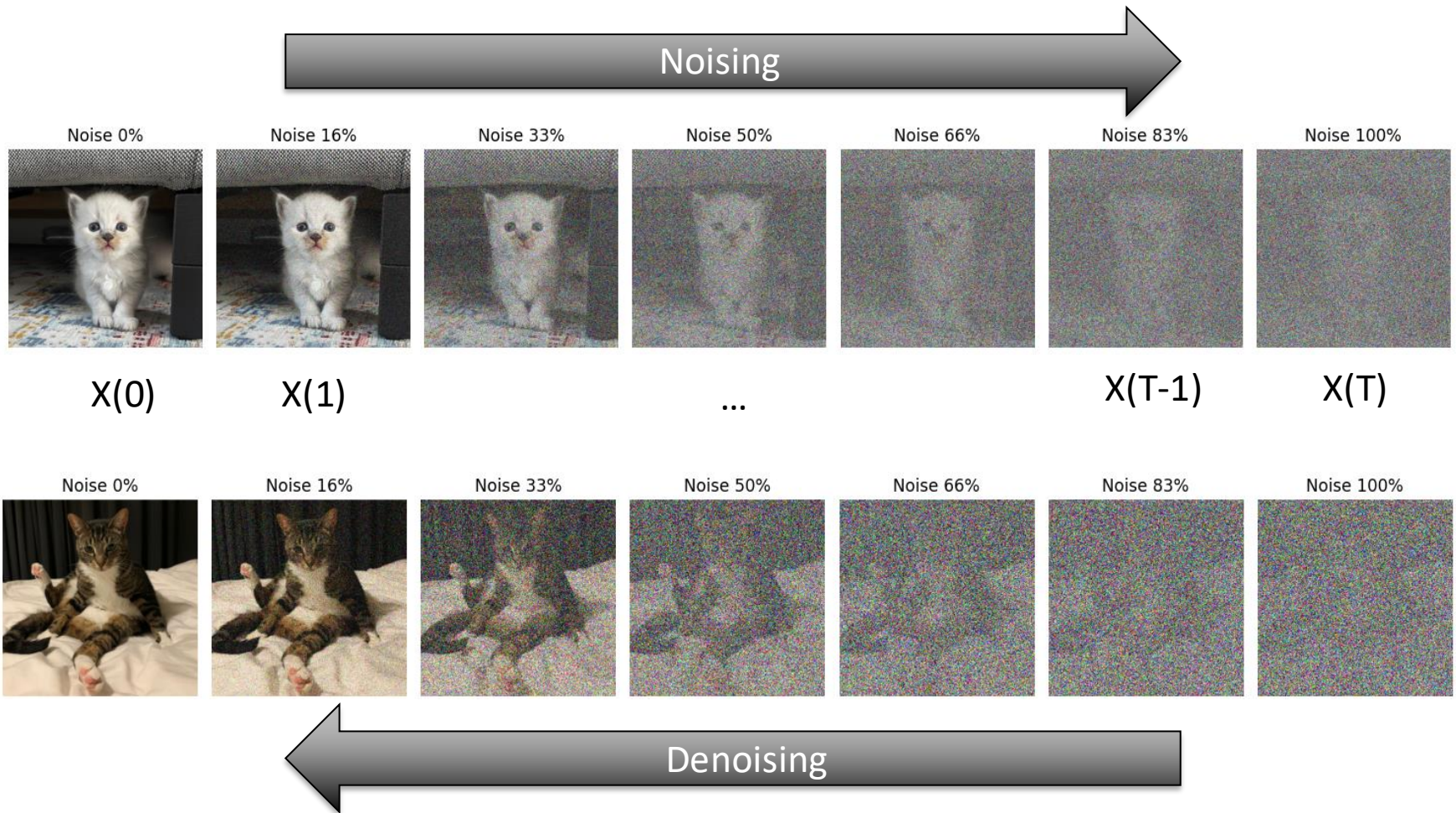
## Noising



# What are diffusion models?

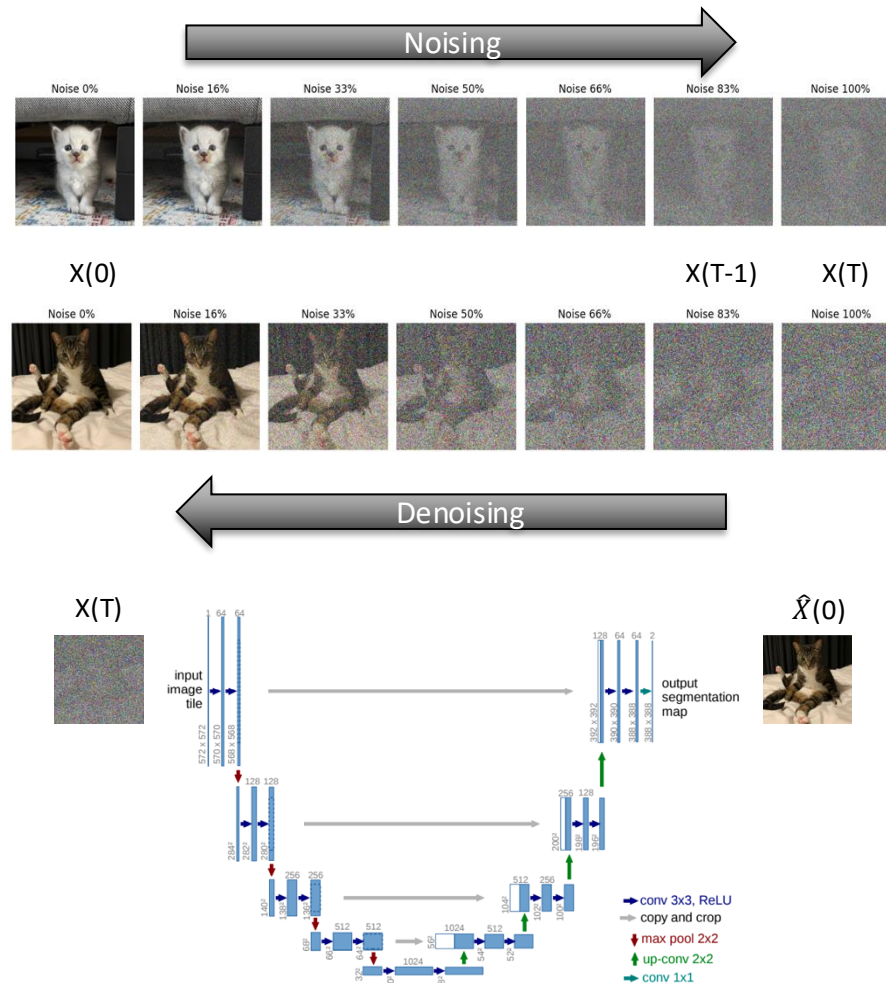


## Denoising



# What are diffusion models?

## Diffusion U-net



# What are Diffusion Language Models (DLMs)



---

## Recap:

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

**Diffusion models** generate images iteratively (refining the whole image in multiple steps)

# What are Diffusion Language Models (DLMs)



---

## Recap:

**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.

# Diffusion Language Models (DLMs)



- **Why** try diffusion language models?
- **What** are (diffusion) language models?
- **How** do diffusion language models work?



DLMs?

- Why?
- What?
- How?

# How do DLMs work?



## 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] |
|------------------------------------------------------------------------------|----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|---------------------------------------------------------|

# How do DLMs work?



## 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] |
|------------------------------------------------------------------------------|----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|----------------------------------------------------------------------------------|------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------|---------------------------------------------------------|

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] |
|---------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------|------------------------------------------------------------------------------------------------|-------------------------------------------------------------------------------------------------------------|--------------------------------------------------------------------------------|---------------------------------------------------------|





# How do DLMs work?



## Considerations

---

### **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
- Difficult to change sentence structure

# Diffusion Language Models (DLMs)

---



- **Why** should we try diffusion language models?
- **What** are (diffusion) language models?
- **How** do diffusion language models work?

# Our study



## LoRA-Adapted Diffusion

Ruurd J.A. Kuiper<sup>1</sup>, Lars de Groot<sup>1</sup>, Bram van Es<sup>2</sup>, Maarten van Smeden<sup>1</sup>, Ayoub Bagheri<sup>2</sup>

<sup>1</sup> University Medical Center Utrecht, Utrecht

<sup>2</sup> 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?

---



Methods

# Methods

What did we do?

---



## **Three main contributions:**

1. Used a novel noising schedule
2. Used LoRA for training
3. Only performed instruction-tuning

# Methods



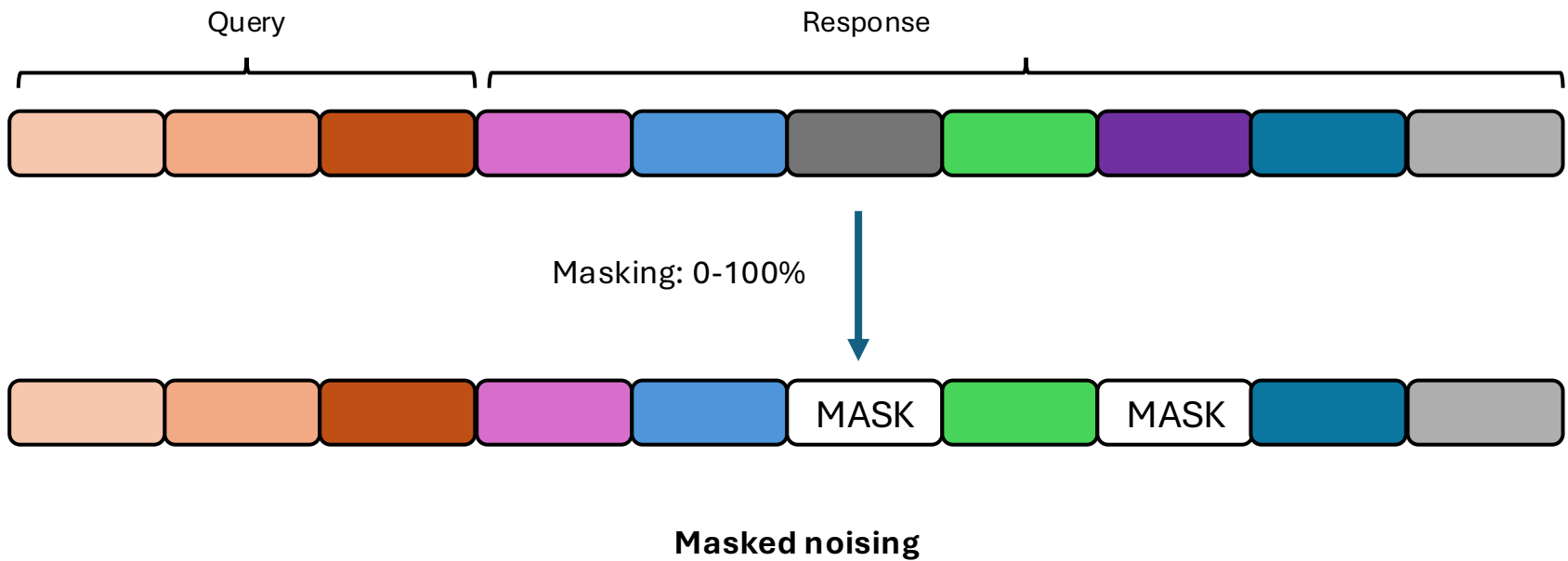
## 1. A novel training noising schedule

---

# Methods



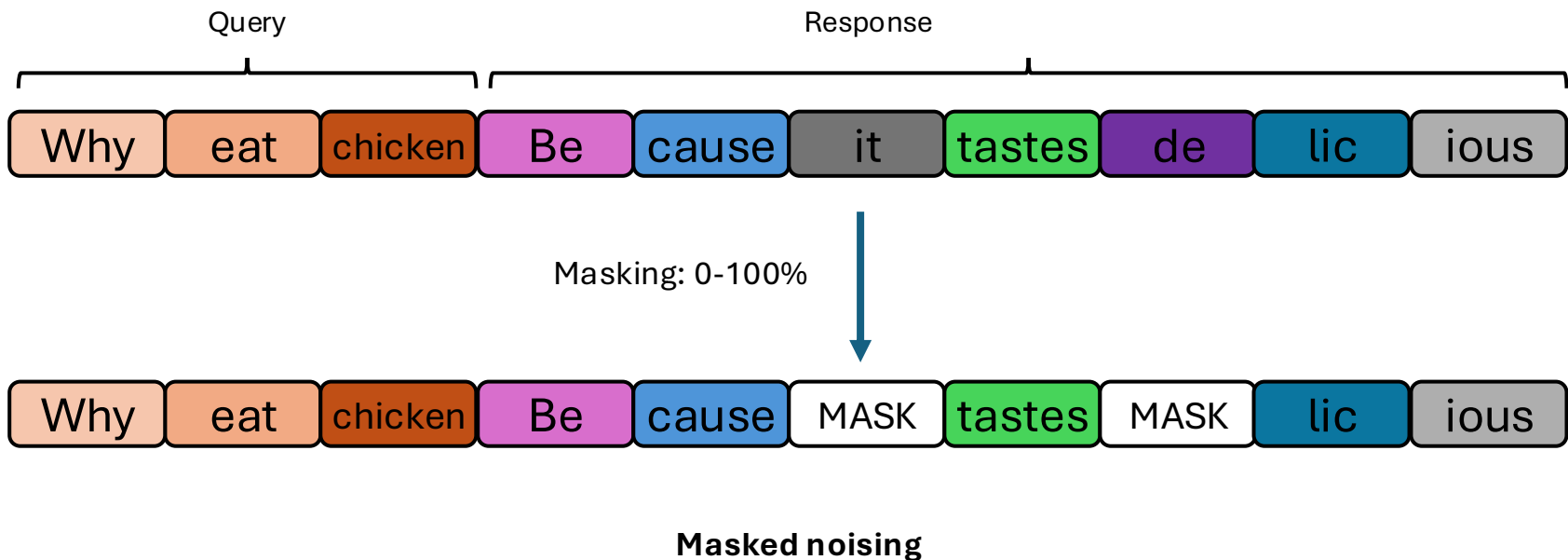
## 1. A novel training noising schedule



# Methods



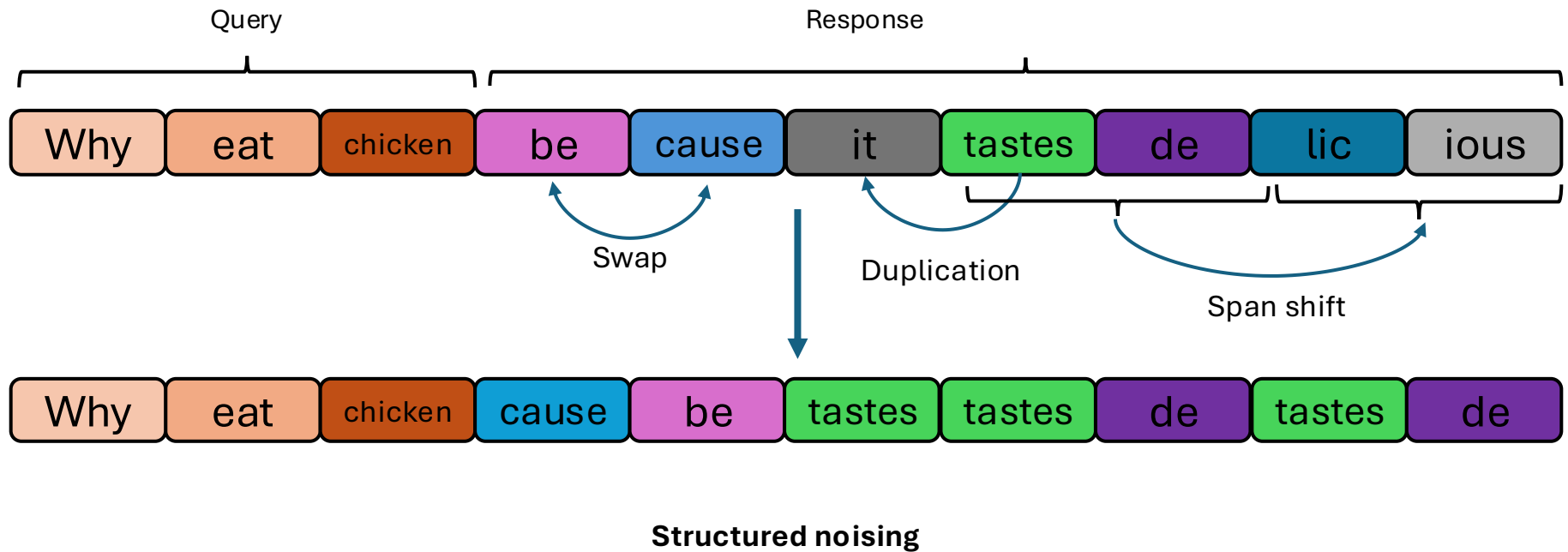
## 1. A novel training noising schedule



# Methods



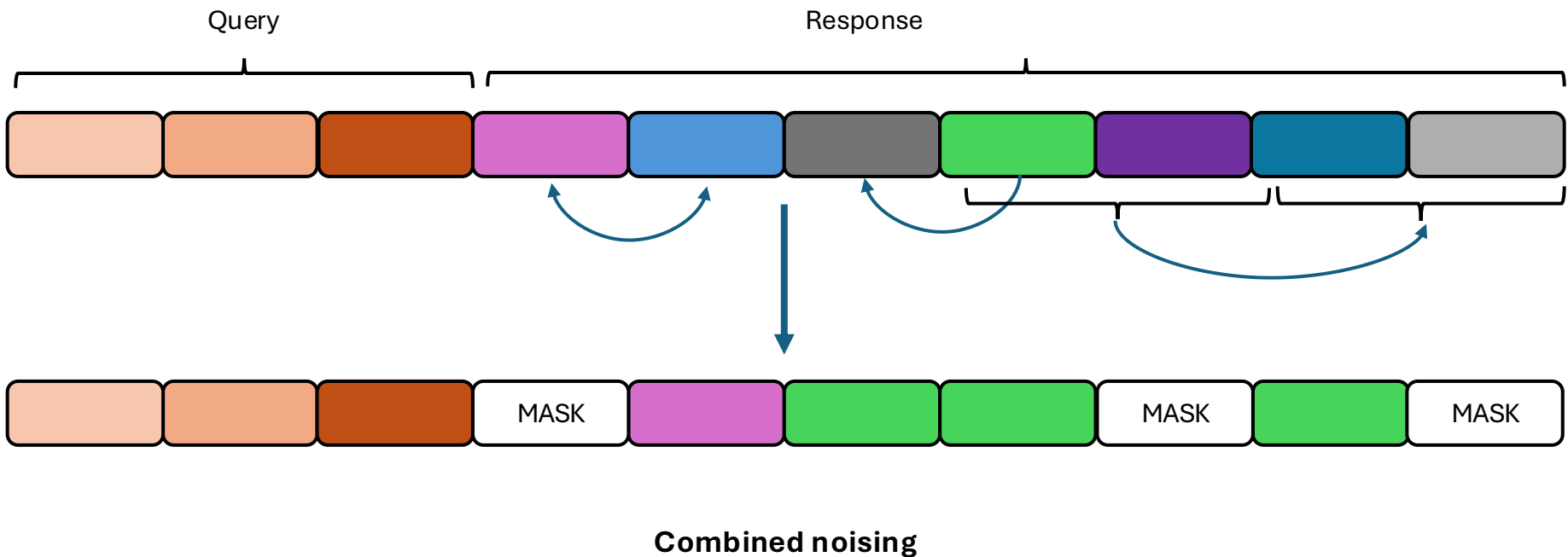
## 1. A novel training noising schedule



# Methods



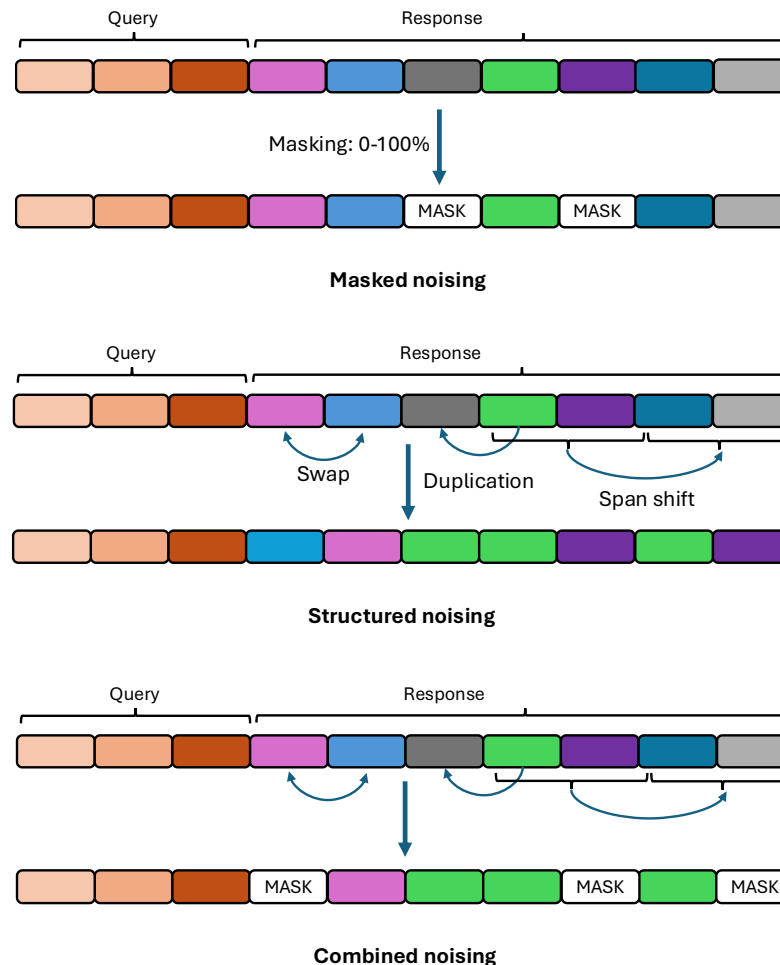
## 1. A novel training noising schedule



# Methods

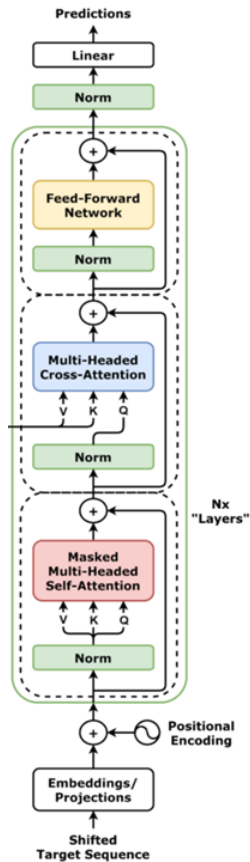


## 1. A novel training noising schedule



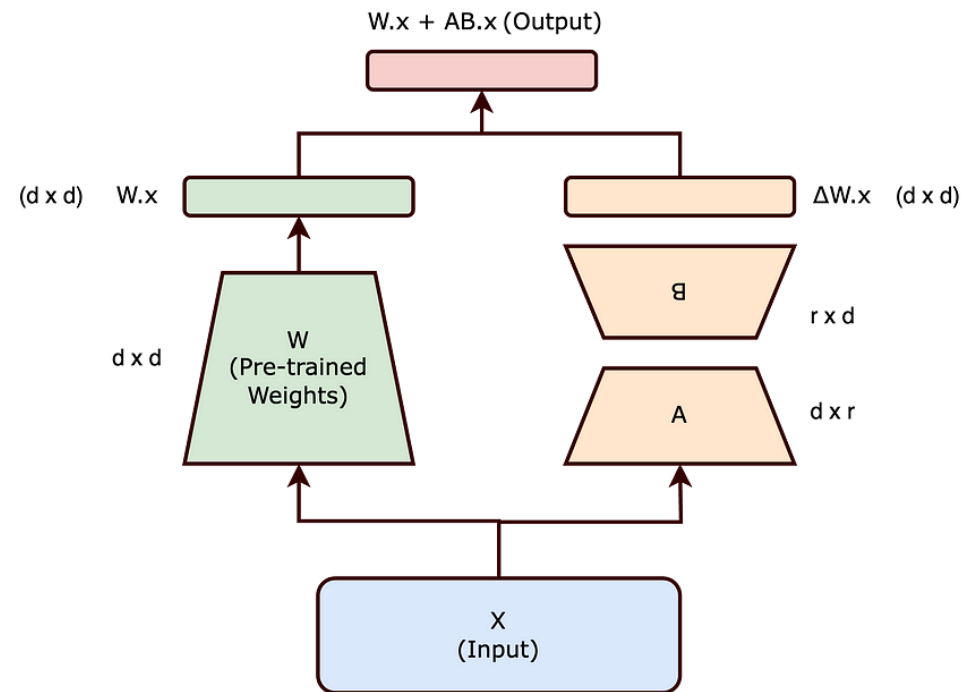
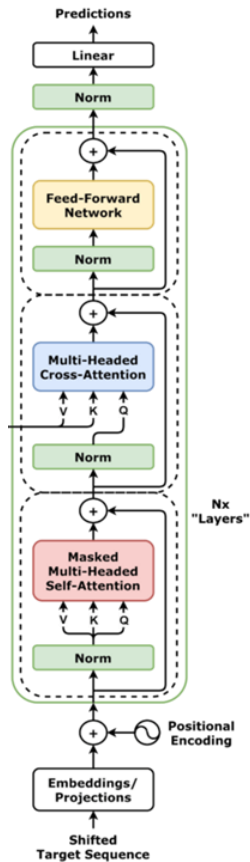
# Methods

## 2. Using LoRA



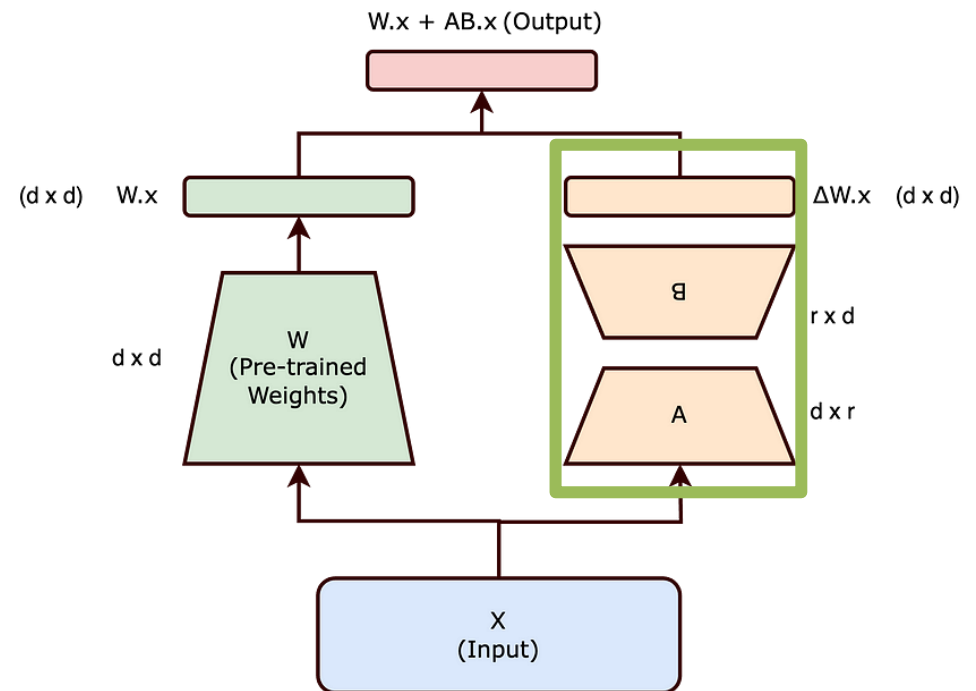
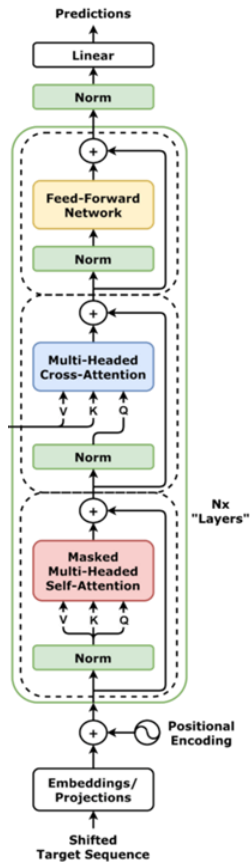
# Methods

## 2. Using LoRA



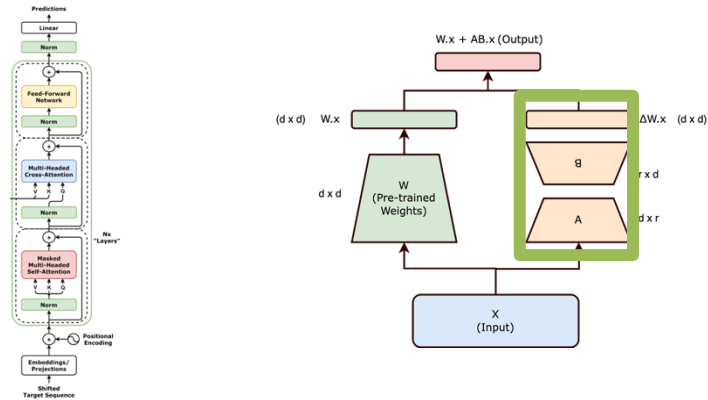
# Methods

## 2. Using LoRA



# Methods

## 2. Using LoRA



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

# Methods



## 3. Only performing 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



## 3. Only performing 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-1024 (Ours)                   | LlaDa 8B                                        | Llama 3.2 8B                                  |
|------------------|--------------------------------------|-------------------------------------------------|-----------------------------------------------|
| <i>Training</i>  |                                      |                                                 |                                               |
| Training* tokens | 200 million<br>$2 \times 10^6$<br>1x | 2.3 trillion<br>$2.3 \times 10^{12}$<br>11,500x | 15 trillion<br>$15 \times 10^{12}$<br>75,000x |
| Trainable params | 436 million                          | 8 billion                                       | 8 billion                                     |

# Results

Does it work?

---



Results

# Results

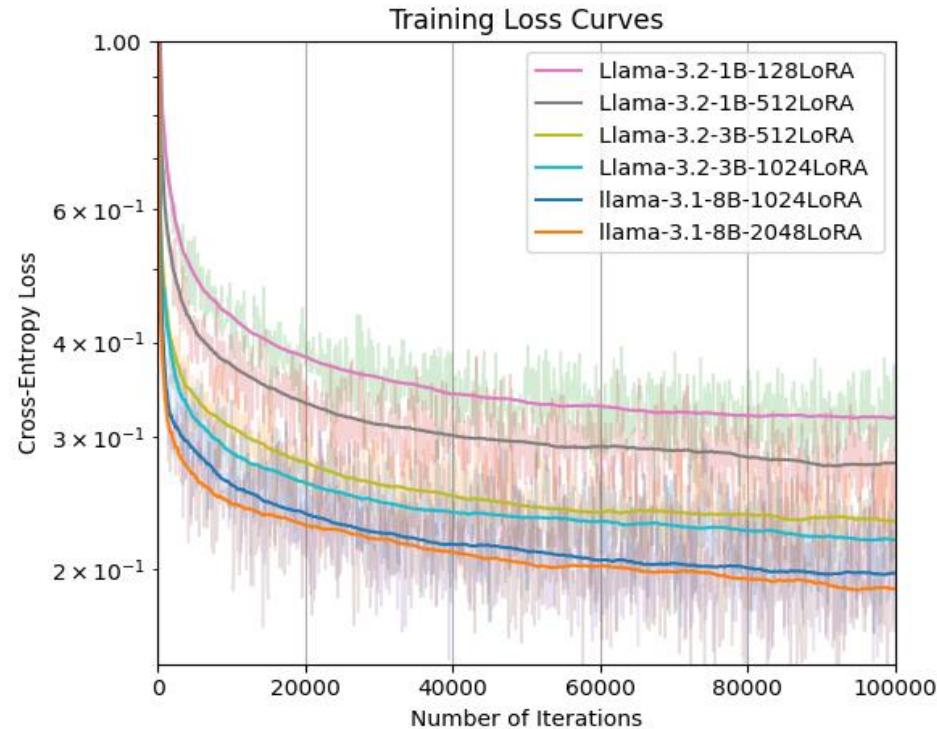
## Benchmark results



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

# Results

## Training loss

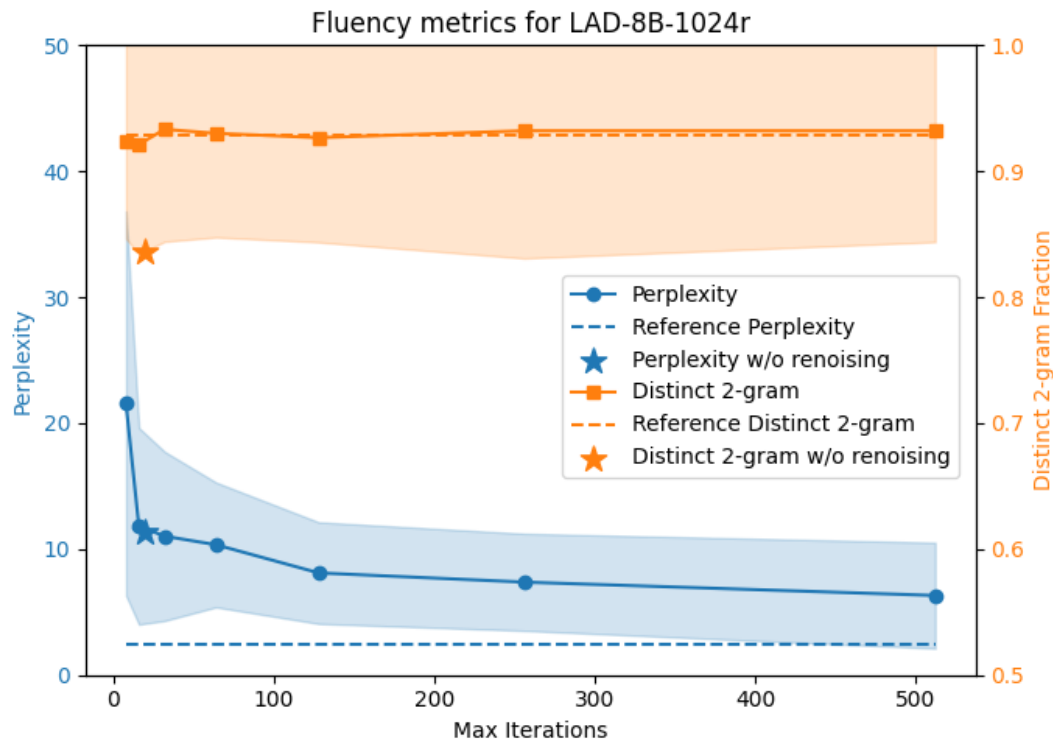


As model parameters increase:

- Training loss decreases

# Results

## Fluency metrics



As inference iterations increase:

- Perplexity decreases
- Distinct 2-gram fraction decreases
- Perplexity without remasking is similar!

## Example: inference with remasking



# Results



## Example: inference without remasking

### 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 Netherlands 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.

# Results

## Demo



**LAD Chat**

User Question

What do you know about Amsterdam?

☒ Enable intermediate noising

☒ Pause between iterations

Increase the maximum number of iterations.  ↻

1  512

Decrease or increase response length  ↻

-5  5

**Clear** **Submit**

<https://ruurdkuiper.github.io/tini-lad/>

# Results



## Goals

---

### Goals

- Train larger models (7B+ parameters)
- Adapt pre-trained autoregressive models
- Fine-tuning using LoRA
- Prediction without masking
- Faster convergence using smarter models
- Finetune on medical data

# Results

## Achievements

---



### Goals

- Train larger models (7B+ parameters)
- Adapt pre-trained autoregressive models
- Fine-tuning using LoRA
- Prediction without masking
- Faster convergence using smarter models
- Finetune on medical data

### Achievements

- 8B parameter Llama model
- Adapt pre-trained autoregressive models
- Fine-tuning using LoRA
- Prediction without masking by using 'structured noising'
- Convergence within less tokens than autoregressive models
- Finetune on medical data

# Results

## Limitations



### Goals

- Train larger models (7B+ parameters)
- Adapt pre-trained autoregressive models
- Fine-tuning using LoRA
- Prediction without masking
- Faster convergence using smarter models
- Finetune on medical data

### Achievements

- 8B parameter Llama model
- Adapt pre-trained autoregressive models
- Fine-tuning using LoRA
- Prediction without masking by using 'structured noising'
- Convergence within less tokens than autoregressive models
- Finetune on medical data

### Limitations

- Grammar
- Coding and math
- Inconsistency

# Future directions



A lot of questions remain...

---

## **Larger (better/faster/stronger)**

- Train for longer
- Train with more than 8B parameters
  - The training and inference objectives are harder, so maybe more parameters will help?

## **Medical finetuning**

- Large (44 million tokens) medical general text dataset available

# Thank you

## Any questions?

---

User Question

Can you tell me something about the Julius Center at the UMCU?

☒ Enable intermediate noising

☐ Pause between iterations

Increase the maximum number of iterations.

64 ↺

1  512

Clear

Submit

### Final Output (47 tokens after 47 iterations)

The Julius Center at the UMCU is an academic center that focuses on interdisciplinary research in the field of health, education, and social sciences. It is considered a key institution for innovative education and research, and collaboration with external partners.