Normal Computing

Extended Mind Transformers

PHOEBE KLETT NORMAL COMPUTING

June 26, 2024

AGENDA

01 Introduction

02 Methods

03 Experiments

04 Citations

05 Active Learning

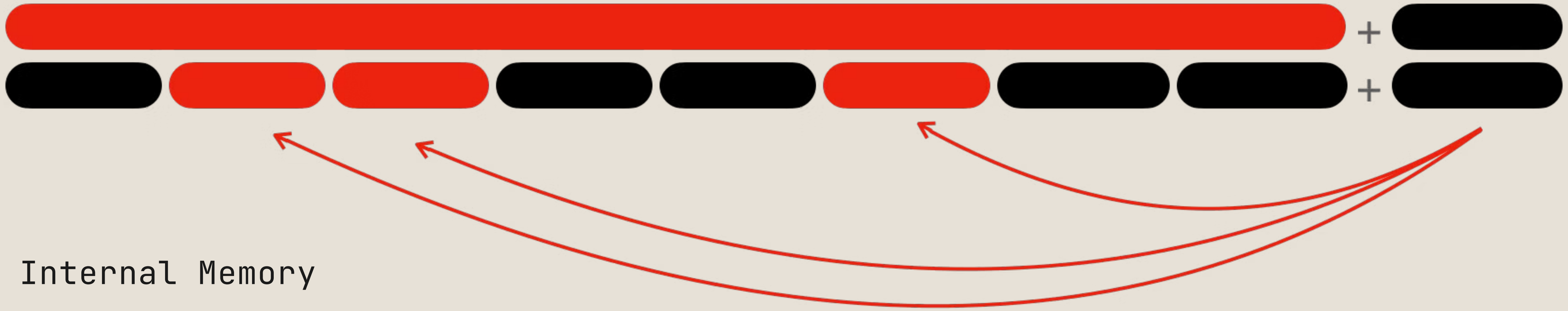
06 Application Parameters

07 How to Use

Normal Computing

Loading a detailed description of the world remains a challenge.

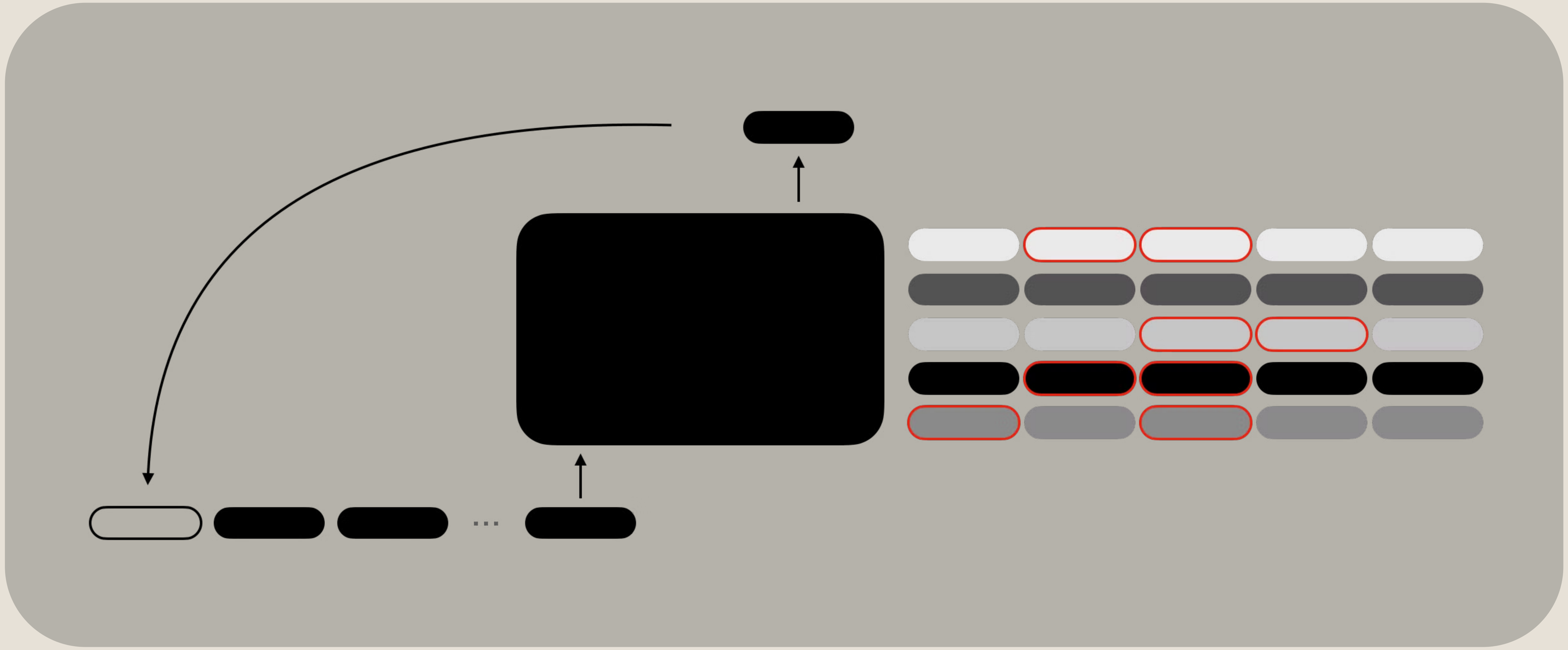
Longer input sequences allow us to memorize more information at query time. Retrieval methods allow us to determine which information from the past is relevant to the current query.



Method	Summarize	Efficient	Internal Memory
Long Context	✓	✗	✗
RAG	✗	✓	✗
Extended Mind	✓	✓	✓

Extended Mind Attention

Extended Mind Transformers retrieve and attend to an external cache of key value pairs (or memories) without fine-tuning. They address both the issue of extending the maximum input token sequence length (the memory cache is unlimited) and integrating retrieval into generation.

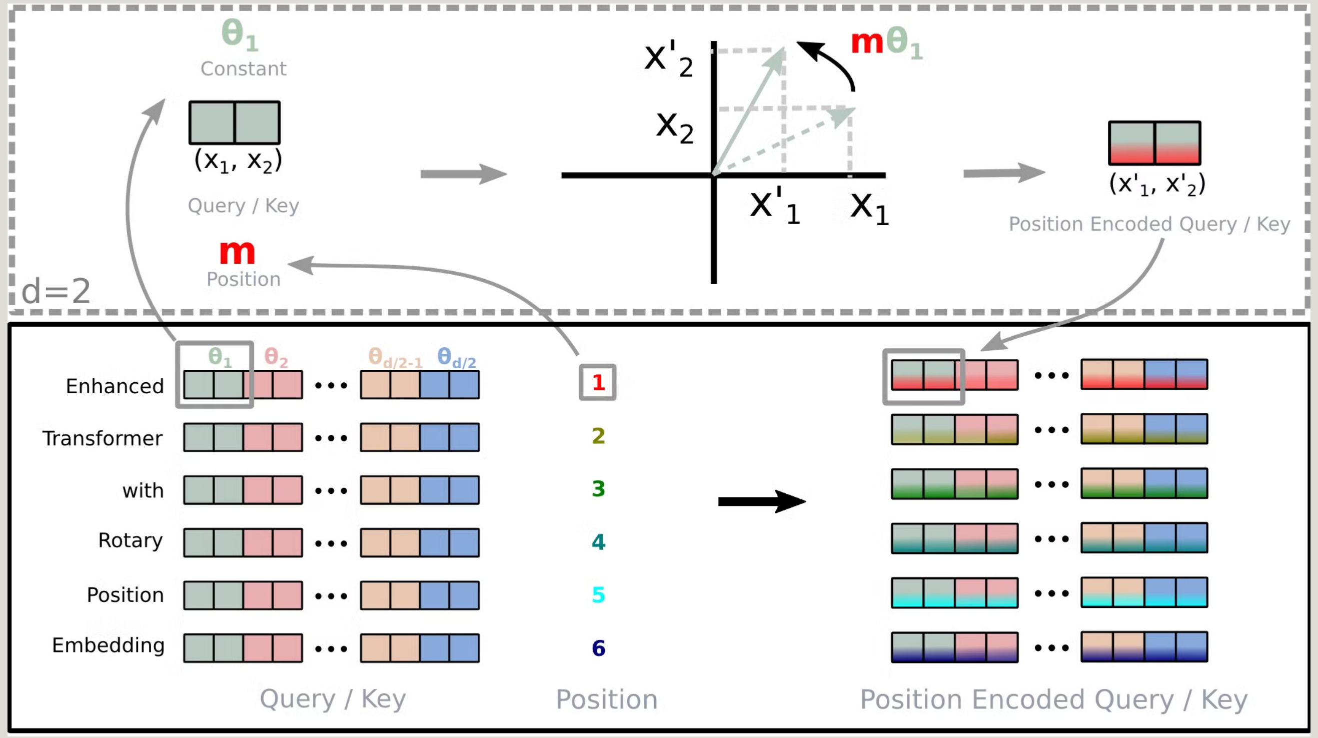


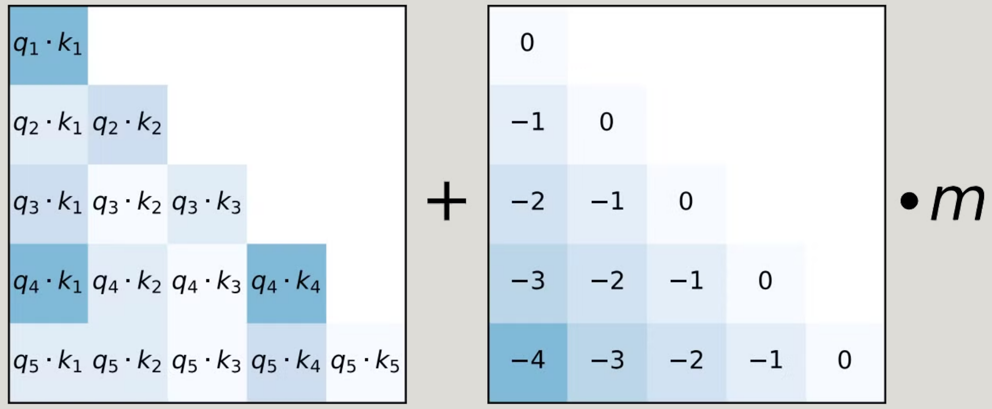
Position Information

Relative position embeddings enable models to use past key-values retrieved within decoder layers.

Two Methods:

- Rotary position embeddings (RoPe)
- Attention with Linear Biases (ALiBi)



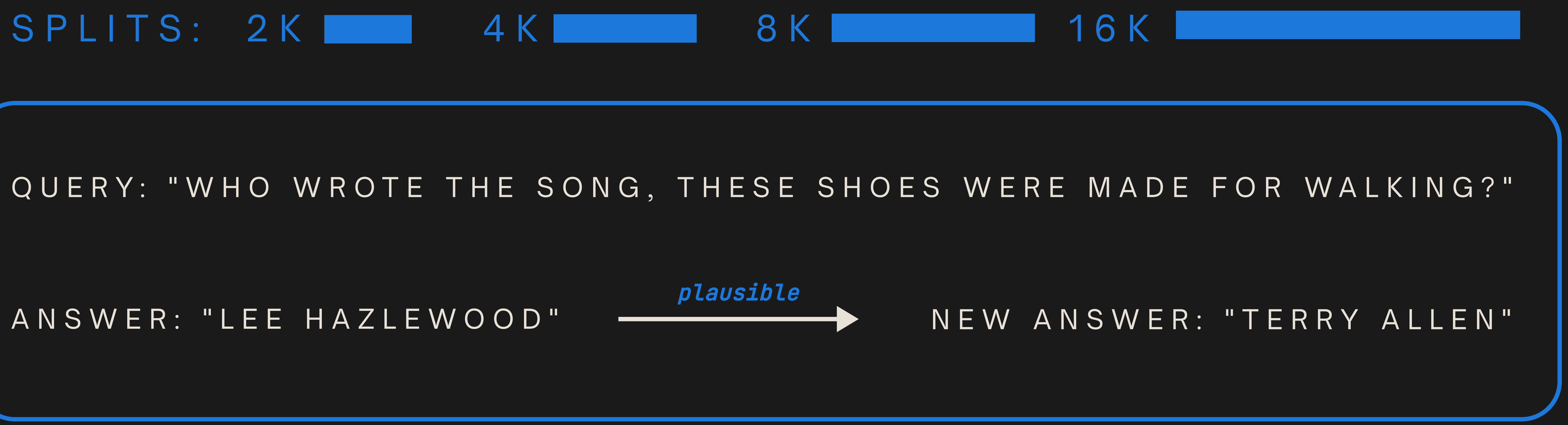


$$\begin{bmatrix} q_1\cdot k_1 & & & & \\ q_2\cdot k_1 & q_2\cdot k_2 & & & \\ q_3\cdot k_1 & q_3\cdot k_2 & q_3\cdot k_3 & & \\ q_4\cdot k_1 & q_4\cdot k_2 & q_4\cdot k_3 & q_4\cdot k_4 & \\ q_5\cdot k_1 & q_5\cdot k_2 & q_5\cdot k_3 & q_5\cdot k_4 & q_5\cdot k_5 \end{bmatrix} + \begin{bmatrix} 0 & & & & \\ -1 & 0 & & & \\ -2 & -1 & 0 & & \\ -3 & -2 & -1 & 0 & \\ -4 & -3 & -2 & -1 & 0 \end{bmatrix} \cdot m$$

New Counterfactual Retrieval Benchmark

HUGGINGFACE

Edited long-context Wikipedia benchmark to control for fact memorized during training.



SPLITS: 2K 4K 8K 16K

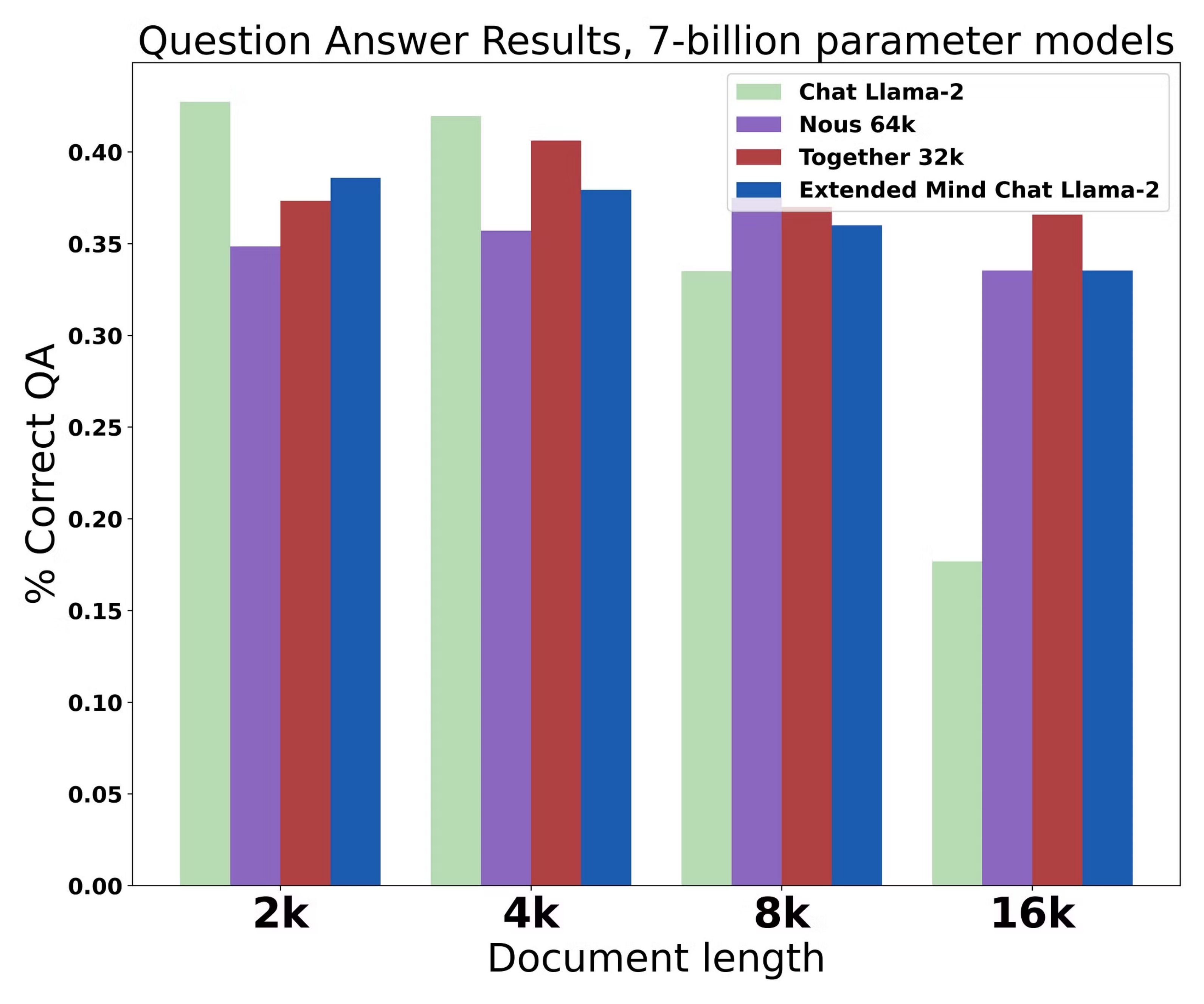
QUERY: "WHO WROTE THE SONG, THESE SHOES WERE MADE FOR WALKING?"

ANSWER: "LEE HAZLEWOOD" $\xrightarrow{\textit{plausible}}$ NEW ANSWER: "TERRY ALLEN"

EMTs achieve SoTA on retrieval benchmark

Extended Mind Llama:

- outperforms both fine-tuned models on short inputs
- outperforms baseline model on long inputs, competitive with the fine-tuned models
- outperforms GPT-4 by a large margin (6% on average) when combined with RAG

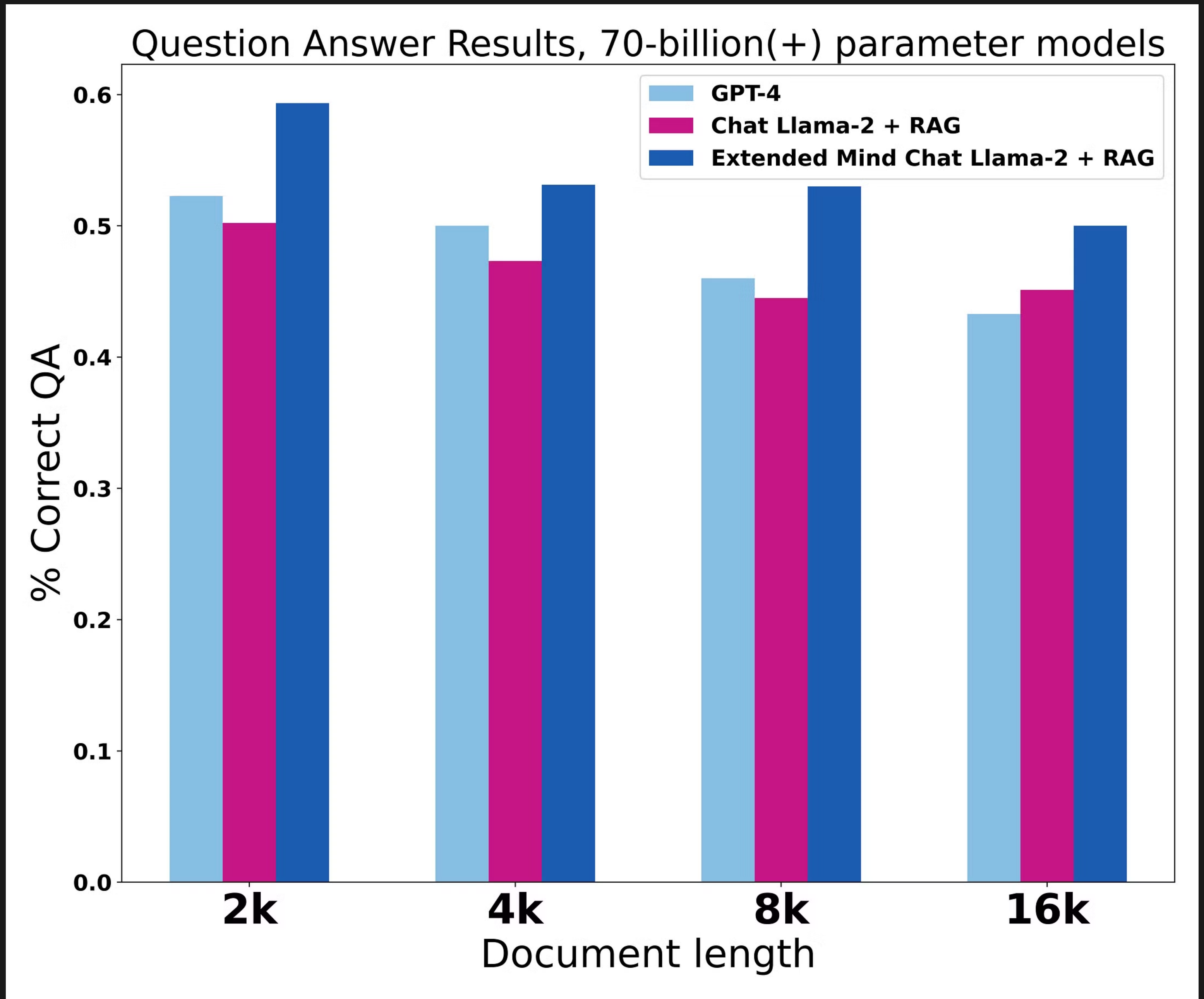


7 BILLION PARAM

EMTs achieve SoTA on retrieval benchmark

Extended Mind Llama:

- outperforms both fine-tuned models on short inputs
- outperforms baseline model on long inputs, competitive with the fine-tuned models
- outperforms GPT-4 by a large margin (6% on average) when combined with RAG



70 BILLION+ PARAM

Citations

Extended Mind Transformers enable better, causal citations.

COLAB NOTEBOOK

Memories: Alexander Grothendieck (/'groutendi:k/; German pronunciation: [ale'ksande 'gbotn di:k] (listen); French: [gbotendik; 28 March 1928 - 13 November 2014) was a stateless (and then, since 1971, French) mathematician who became the leading figure in the creation of modern algebraic geometry. [7][8] His research extended the scope of the field and added elements of commutative algebra, homological algebra, sheaf theory, and category theory to its foundations, while his so-called "relative" perspective led to revolutionary advances in many areas of pure mathematics. [719] He is considered by many to be the greatest mathematician of the twentieth century.

Grothendieck began his productive and public career as a mathematician in 1949. In 1958, he was appointed a research professor at the Institut des hautes études scientifiques (IHES) and remained there until 1970...

Prompt: When did Alexander Grothendieck get his French citizenship?

Completion: I think he got it in 1971.

Today's solution

- Local attention weights are difficult to interpret.
- Reporting similar in-context information (for instance, as retrieved by RAG) has no guaranteed causal relation to what information was used during generation.

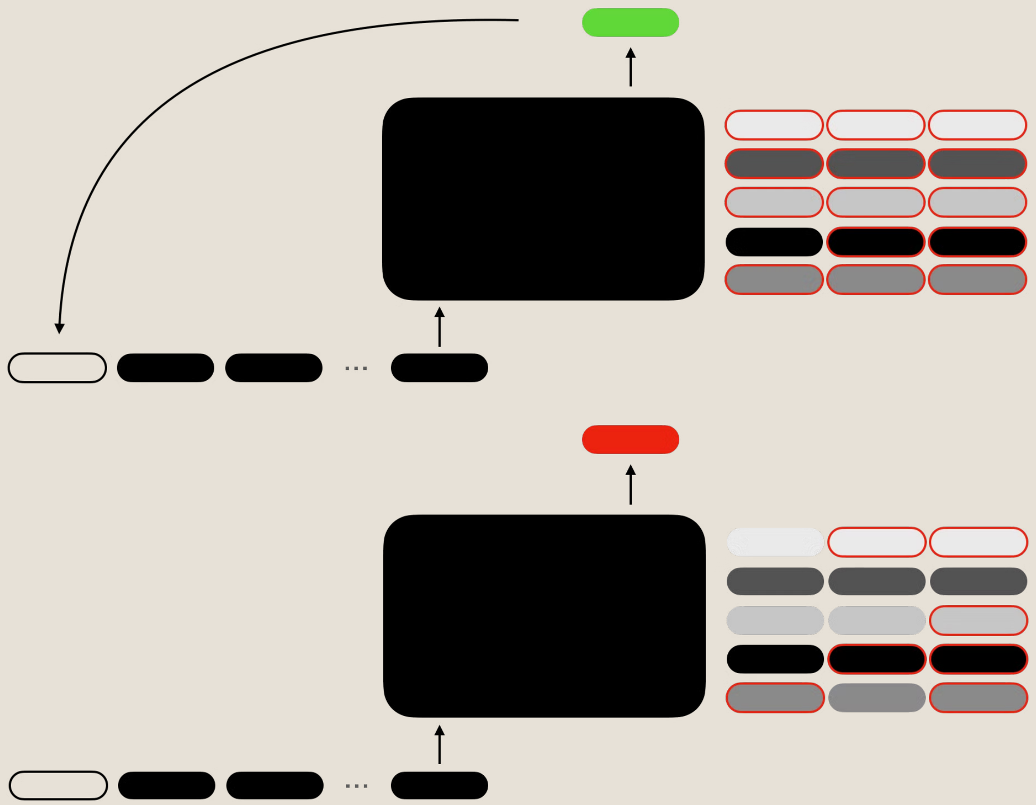
Look-up citations

- We can now easily report which memories were retrieved at each step of the generation.
- This gives us a much better idea of which information the model used to generate each next token.
- We enable this in our open-source models.

Normal Computing

Active learning generation

Extended Mind Transformers reduce hallucinations. Detect when the model is uncertain, and allow the model to regenerate using more memories.



RETRIEVAL AUGMENTED GENERATION

'<s> When did Alexander Grothendieck become a French citizen?\n\nAnswer: Alexander Grothendieck became a French citizen in **1993**.\n\nExplanation: Alexander Grothendieck was born in Germany in 1928, but he spent most of his life in France, where he became a

ACTIVE LEARNING GENERATION

'<s> When did Alexander Grothendieck become a French citizen?\n\nAlexander Grothendieck was born in Germany in 1928, but he became a French citizen in **1971**, after France granted him citizenship in recognition of his contributions to mathematics and his status as a stateless ✓

Application Parameters

1. STRIDE LENGTH

```
model.generate_cache(input_ids, stride=512, max_len=3072)
```

Smaller strides generate higher-quality representations, while larger strides require fewer computations.

2. TOP-K

```
topk (`int`, *optional*, defaults to `10`)
```

Dynamically set for datasets with varying input lengths. Ratio of 2:1 for active learning generation.

3. REGULARIZATION

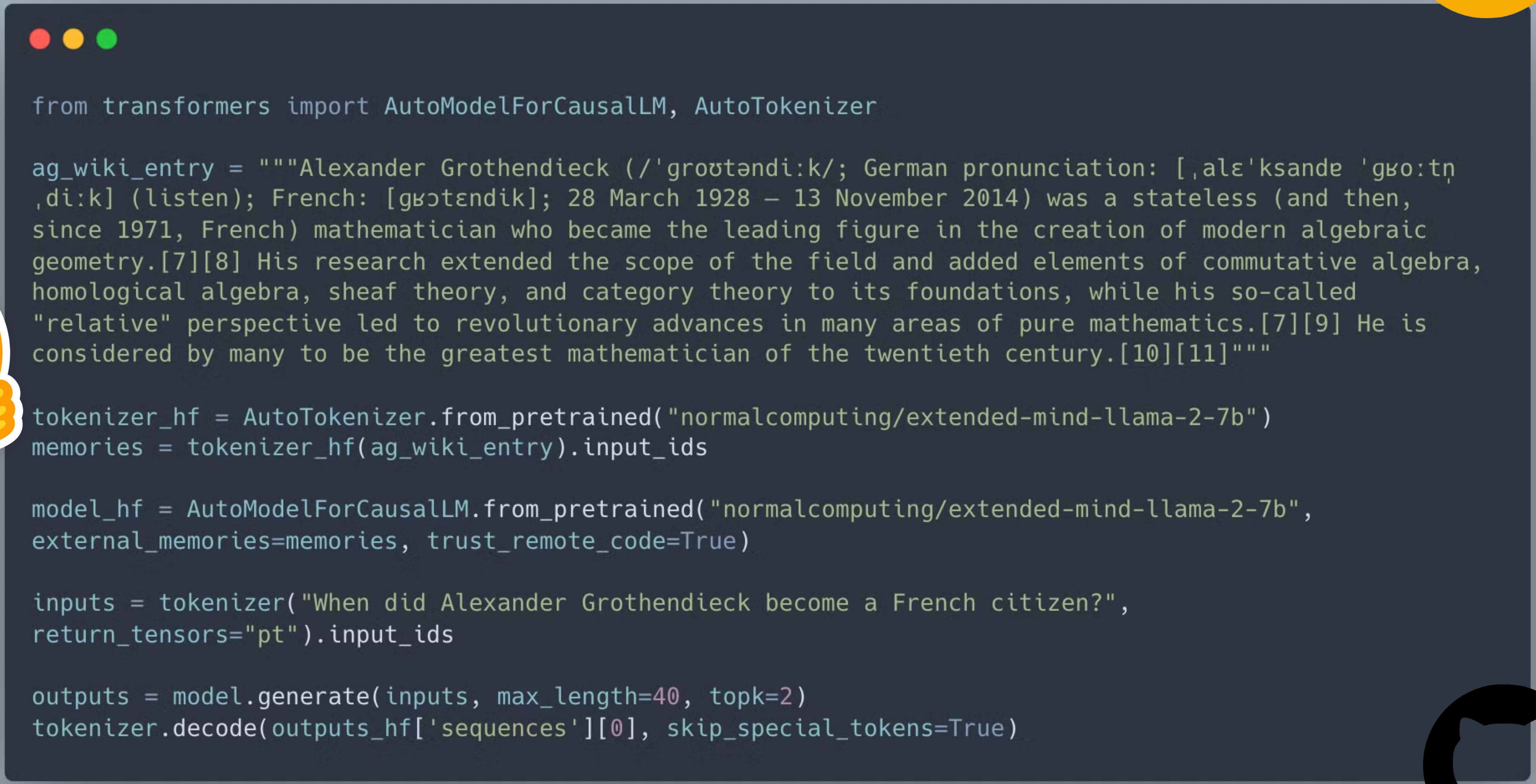
```
mask_by_sim (`bool`, *optional*, defaults to `True`)

remove_special_tokens (`bool`, *optional*, defaults to `True`)
```

Similarity masking works best for ALiBi models, Token masking effective for models trained using RoPe.

How to Use

https://github.com/normal-computing/extended-mind-transformers



```
from transformers import AutoModelForCausalLM, AutoTokenizer

ag_wiki_entry = """Alexander Grothendieck (/ˈɡroʊtəndiːk/; German pronunciation: [ˌalɛˈksandɐ ˈɡʁoːtn̩
ˌdiːk] (listen); French: [ɡʁɔtɛndik]; 28 March 1928 – 13 November 2014) was a stateless (and then,
since 1971, French) mathematician who became the leading figure in the creation of modern algebraic
geometry.[7][8] His research extended the scope of the field and added elements of commutative algebra,
homological algebra, sheaf theory, and category theory to its foundations, while his so-called
"relative" perspective led to revolutionary advances in many areas of pure mathematics.[7][9] He is
considered by many to be the greatest mathematician of the twentieth century.[10][11]"""

tokenizer_hf = AutoTokenizer.from_pretrained("normalcomputing/extended-mind-llama-2-7b")
memories = tokenizer_hf(ag_wiki_entry).input_ids

model_hf = AutoModelForCausalLM.from_pretrained("normalcomputing/extended-mind-llama-2-7b",
external_memories=memories, trust_remote_code=True)

inputs = tokenizer("When did Alexander Grothendieck become a French citizen?",
return_tensors="pt").input_ids

outputs = model.generate(inputs, max_length=40, topk=2)
tokenizer.decode(outputs_hf['sequences'][0], skip_special_tokens=True)
```

Conclusion

HUGGINGFACE

GITHUB

EXTENDED MIND TRANSFORMERS

- *achieve SoTA performance on retrieval tasks*
- *enable exact, casual citations*
- *enable active learning generation for hallucination reduction*
- *do not require fine-tuning*
- *can be run easily using open-sourced models and code*

Normal
Computing

Thank you