

# Memory mapped Memory for transformers

Bastian Apportin\*  
Joomo GmbH

July 1, 2025

## Abstract

The success of LLM's is mainly based on the astonishing capabilities of the Transformer Architecture. Which does a great job in catching long range dependencies.

The positional encoding used in Transformers solves the problem of context awareness, but limits the distance a transformer can look. What we want is actually to keep the context awareness of the architecture, like distance memories coming back, influencing your current thoughts.

We've tried a lot of different approaches like ALiBi or placing cnn's in the key/query calculations, but none of them worked like we expected. All the limited experiments we did worked significantly better when using RoPE like position encodings. But maybe, we don't need to remove them.

The RoPE embedding can be seen as an offset plus content addressing. The offset is the rotation and the mixture of the frequencies can address sharper or broader position back from the current position.

## 1 Introduction

The transformer architecture is self attention as its core. The worst and best part of that is, that there is no local context. Therefore most Transformer architectures apply some kind of positional encoding. This positional encoding solves the problem of the missing local context but introduces the problem that for long distances the 'context' or distance matters as well. Therefore the view of site can not be infinite, which effectively limits memorizing things by that mechanism.

The RoPE embedding can be seen as an offset plus content addressing. The offset is the rotation and the mixture of the frequencies can address sharper or broader position ranges back from the current position and the content are the original values of the queries and keys.

When we have positional addressing an idea came to mind, the memory mapped IO, where registers to access secondary hardware are mapped to main memory. One would write the block id of a harddrive to a memory position and then the content of that block will be mapped to a specific position in memory.

## 2 Methodology

we use the llama 3.2 1b model for our experiment, since that is a size we can still handle.

the llama model uses RoPE. RoPE implements an influence of relative distance. So it's not important what actual position(id) the token has. The result will be the same if we shift the whole sequence.

We think of the current position as the anchor and place the mapped memories at a fixed negative offset. For ease of understanding we place the current position at 4096 and the memory map at 0. This way we

---

\*Corresponding author: [bastian.apportin@joomo.de](mailto:bastian.apportin@joomo.de)

can store the unrotated keys and query them using a kind of LSH. The algorithm is an approximation and the model is not trained for that explicitly. We'll see if that actually works.

We can now store unrotated older key,value pairs in an LSH like structure and search with the rotated queries. The queried results can be appended to the key and value tensors before the attention calculation. This way relevant memories from the storage can 'come to mind' and influence the generated sequence.

This mechanism can probably be improved when the model is trained (another fine tuning step maybe) with that mechanism in place. In addition this mechanism can be used to memory map other mechanisms. For example we have a transformer model that generates limb movements from camera inputs and an llm to understand and generate text. We can now memory map those models into each other - potentially with a linear transformation in between. So the text model has a limited view on some key/values from the vision/limb model and vice versa.

For now we want to place memories in the storage and ask questions where the answer can show that knowledge from the memory was actually used.

### 3 Experiments

To do useful experiments we will use fairy tales that we will put to the memory and a set of commands we ask the model like versions of 'tell me a story about ...'

for the dataset we use the

- <https://www.kaggle.com/datasets/cuddletail/fairy-tales/data>

dataset from kaggle.

For the Questions we asked chatgpt to produce context free questions that would be influenced when reading one of the fairy tales before. The resulting Questions:

- What kind of person do you think is most likely to succeed in a difficult situation?
  - A reader of To Your Good Health! might say "a trickster," while Lovely Ilonka might lead to "a brave and clever person," and Lucky Luck could prompt "someone with good fortune."
- If someone has a secret, should they always tell it? Why or why not?
  - Answers would differ strongly between The Boy Who Could Keep a Secret, The Language of Beasts, or someone who hasn't read any tale.
- Do you think it's better to be lucky or smart?
  - A Lucky Luck reader may favor luck, while The Hairy Man or Seven Simons might push someone toward cleverness or teamwork.
- What kinds of help are most valuable — magical, practical, or social?
  - A reader of The Seven Simons may argue for teamwork (practical), while Lovely Ilonka or The Hairy Man could lead someone to defend magical help.
- Is silence ever more powerful than speaking?
  - The Language of Beasts or The Boy Who Could Keep a Secret readers might say yes; others may disagree.
- How should someone deal with people in power who are being unfair?

- Readers of To Your Good Health! may lean toward rebellion or trickery, while others might suggest patience, service, or outwitting.
- If you could have one unusual skill, what would it be and why?
  - The Seven Simons readers might recall specific skills (e.g., building towers, spying from afar) and answer accordingly.
- What makes someone truly trustworthy?
  - Answers might reflect how characters like the boy in The Hairy Man or The Boy Who Could Keep a Secret prove themselves.
- Do you believe people can change their fate? Why or why not?
  - Some tales lean toward predestination (Lucky Luck), others toward personal effort (Lovely Ilonka, The Hairy Man).
- Can disobedience ever be a good thing?
  - Characters in The Hairy Man or The Language of Beasts disobey to their benefit — this could spark a nuanced response from someone who read those stories.

we will use this question for now together with the fairy tails (Lovely Ilonka, Lucky Luck, The Hairy Man, To Your Good Health!, The Story of the Seven Simons, The Language of Beasts or The Boy Who Could Keep A Secret) or none and see if the generated answers lead to results that depend on the memory.

## **4 Results**

## **5 Discussion**

## **6 Conclusion**

## **7 Related Work**

## **Acknowledgments**

## **A Appendix Title**