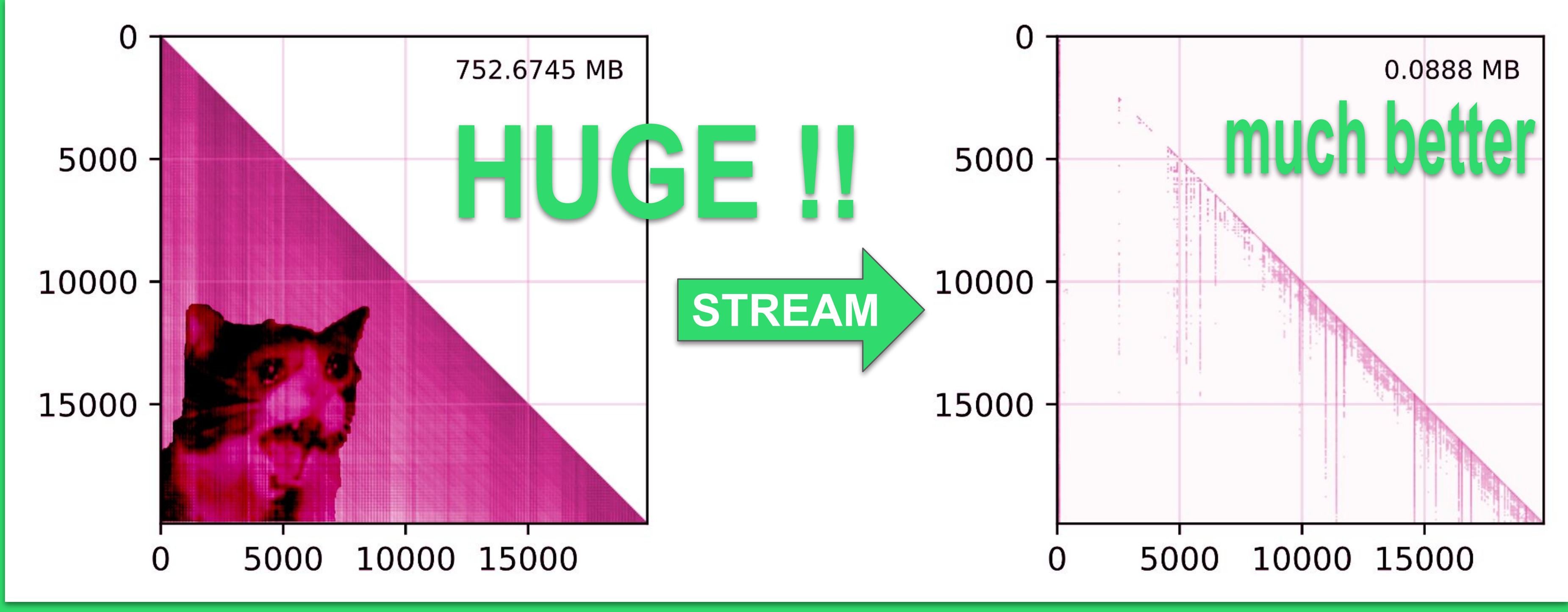
Spotify does Mech Interp !!



Are your attention patterns TOO BIG?

Honestly. Same.

Let's talk :)

Stream:
Scaling up Mechanistic Interpretability to Long Context in LLMs via Sparse Attention

J Rosser, José Luis Redondo García, Gustavo Penha, Konstantina Palla, Hugues Bouchard

1 TL;DR

Main 3 takeaways:

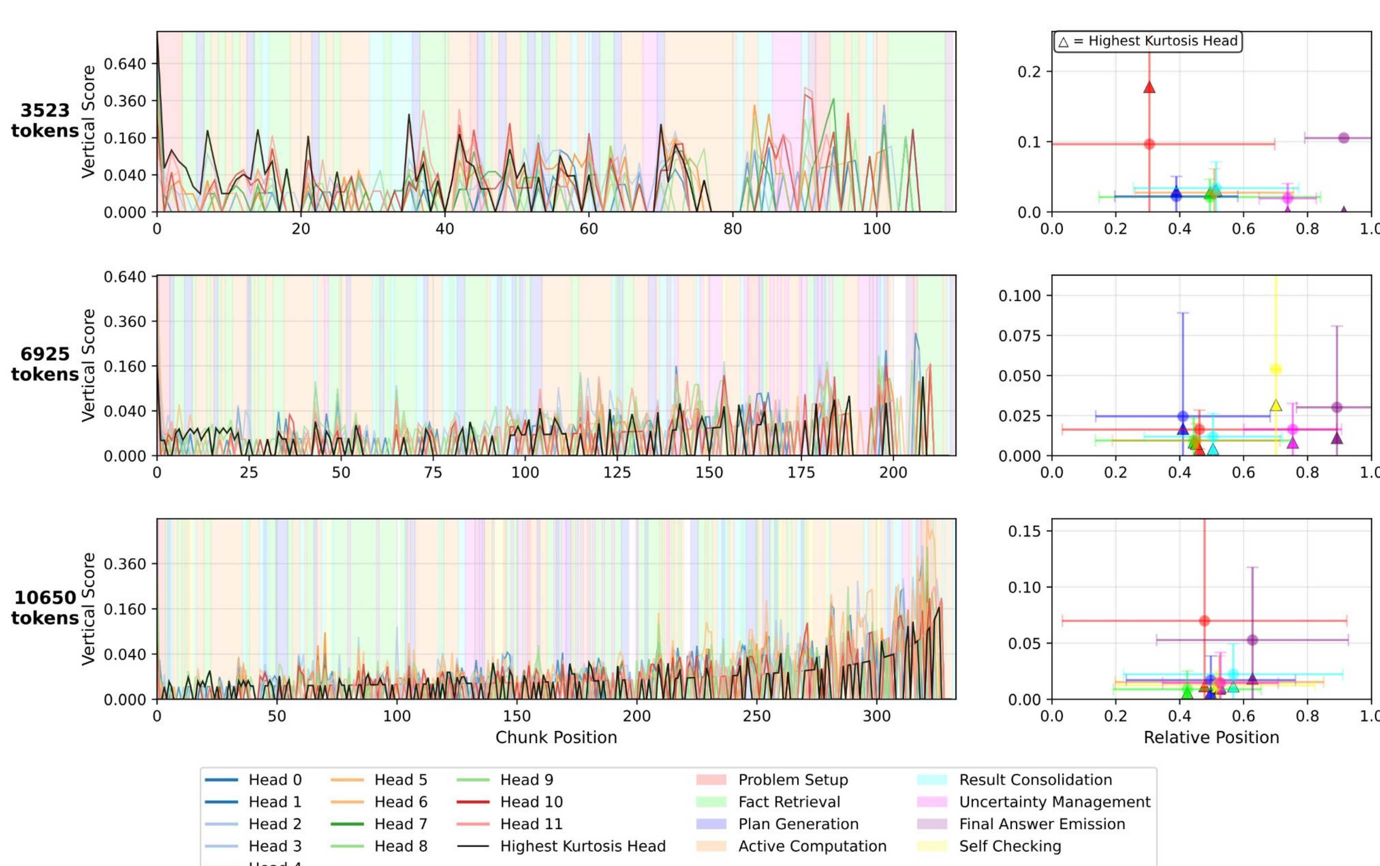
- We introduce **Stream**, an efficient method to analyse **long context** for **reasoning traces**.
- Stream achieves **near-linear time and linear memory complexity.**
- **Stream prunes 90-99% of attention patterns** while highlighting critical thought anchors and retrieval paths

2 Motivation

- Many existing mech interp tools scale quadratically with context length. **E.g. Caching all attention patterns in Gemma 3 4B would require 3.84TB.**
- So, **can we use sparse attention algorithms to preserve model behavior during interpretability experiments?**

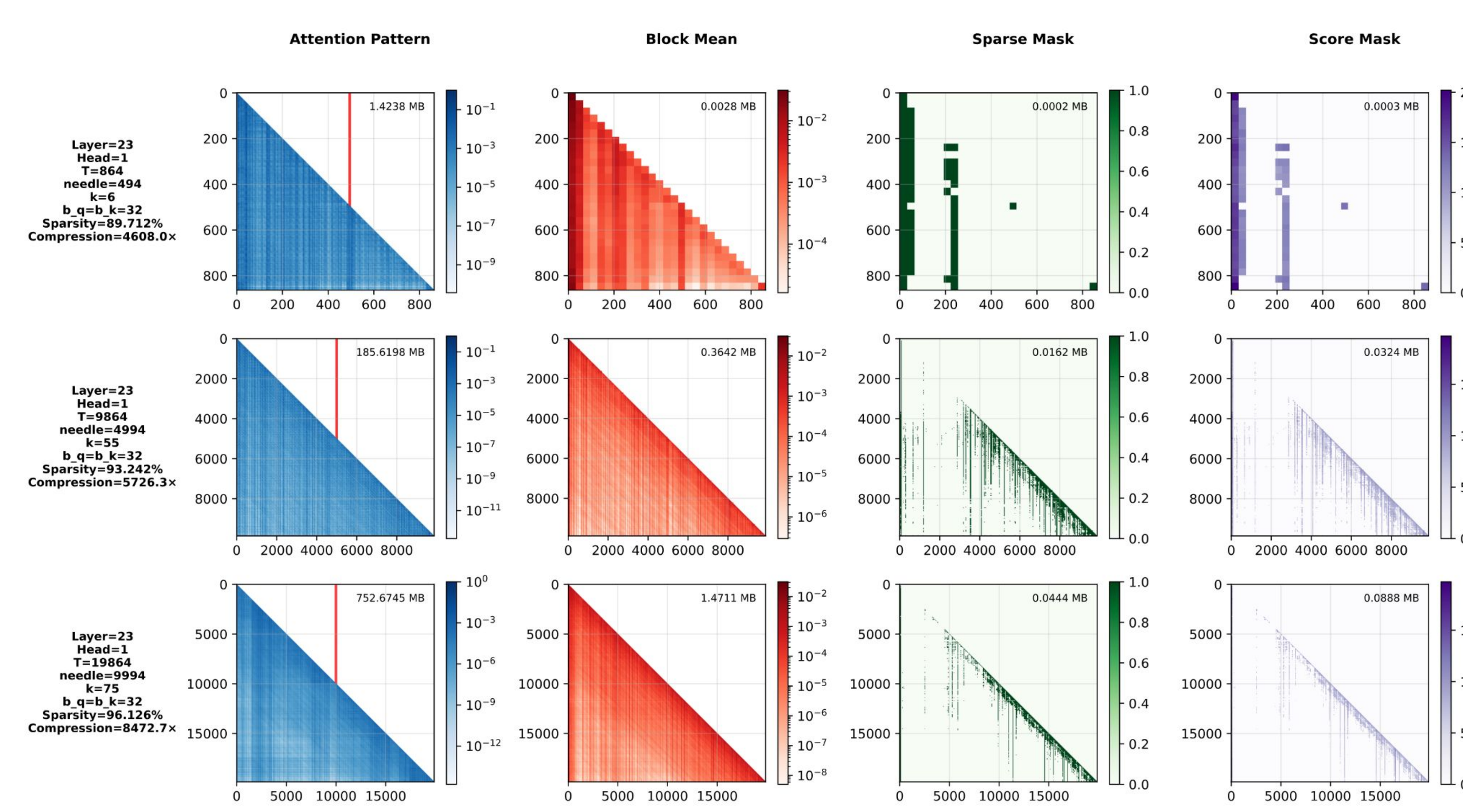
3 Thought Anchors

- **We use Stream to analyze long CoT reasoning traces and we test whether it can recover the same "thought anchors" identified in prior work (Bogdan et al. 2025) - chunks that attract unusually strong downstream attention.**
- Applying Stream across three context lengths, we compute block-level vertical attention scores and examine how attention aligns with reasoning categories. Even after pruning 97-99% of attention links, the method reliably surfaces anchor regions tied to problem setup, planning, and answer emission.
- **As context grows, attention shifts from local computation toward higher-level steps such as planning, uncertainty management, and self-checking, revealing increasingly structured reasoning behavior at scale.**

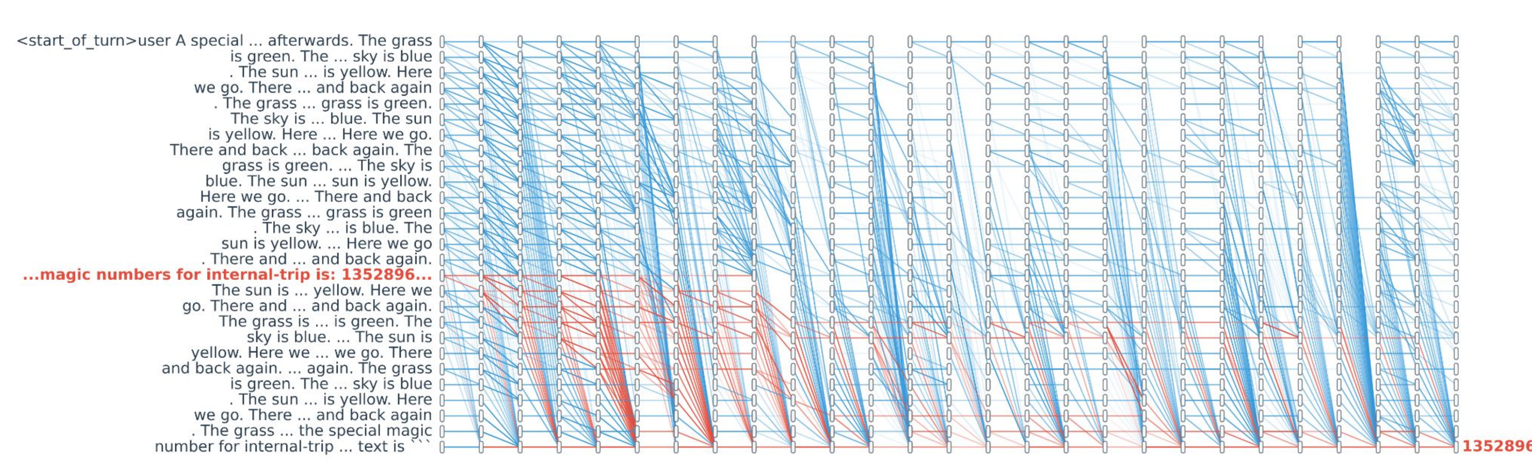


4 Needle in a Haystack

- **To probe information flow more directly, we apply Stream to isolate the specific attention routes that enable successful needle retrieval.** Below we visualize how the pruned mask preserves the critical blocks linking the needle to the final token, while discarding the majority of irrelevant edges.



- The second figure shows the difference between successful and unsuccessful retrieval settings, **revealing a small set of paths that consistently carry the needle signal forward.**



- Together, these visualizations illustrate that **Stream not only compresses the attention graph, but also exposes the sparse structural backbone along which information actually travels.**

5 Limitations and Discussion

- Stream omits key parts of transformer computation and relies on heuristic output-matching rather than formal guarantees.
- Pruning can introduce positional biases, especially near the end of long contexts.
- Future work includes adaptive sparsity and modeling information flow beyond attention alone.

I'm eager to collaborate with researchers interested in applying, evaluating, or extending Stream.

PAPER

CODE

Mechanistic Interpretability Workshop, NeurIPS 2025

NEURAL INFORMATION PROCESSING SYSTEMS

Engineering and Physical Sciences Research Council

AIMS EPSRC Centre for Doctoral Training in Autonomous Intelligent Machines & Systems

jrosser@robots.ox.ac.uk