Postdoc offer

Theoretical Analysis of Large Context Transformers

Nicolas Keriven (CNRS, IRISA) and Samuel Vaiter (CNRS, LJAD)

Location	Rennes or Nice, France
Funding	ERC MALAGA or 3IA Côte d'Azur chair BOGOL, depending on the starting date and location
Duration	2 years
Salary	CNRS postdoctoral salary based on experience

Motivation

Chain-of-thought and, more generally, reasoning models write long traces, so the context keeps growing as they compute the output. Yet we know little about what a Transformer computes at these lengths: whether extra tokens bring new information, or are simply averaged. The goal of this postdoc is to understand fine properties of LLMs at this scale.

Theory is starting to answer this. Attention on n tokens can be seen as an empirical version of an infinite-token limit. Bohbot et al. [2] show that output moments can converge slower than $1/\sqrt{n}$, which suggests diminishing returns. Boursier and Boyer [3] show that softmax behaves like linear attention on long Gaussian prompts. Chen et al. [4] show that attention weights collapse toward uniform as n grows, and identify $\log n$ as the critical scaling of the inverse temperature.

These settings remain far from a real reasoning trace, whose tokens are dependent, generated by the model itself, and processed through many layers. This position aims to close that gap.

Research directions

We aim to describe the fine properties of Transformers at large context, for instance:

- **Depth:** multilayer limits and the interplay of depth and context length.
- **Fine behaviour:** rates, fluctuations, transition behaviors.
- **Realistic contexts:** dependent tokens and contexts generated by the model itself.
- **Looped transformers:** these models reason by repeating layers in latent space [5]. Mechanistic interpretability [1] provides a fixed point interpretation.
- **Reasoning:** when a longer trace adds information, and when it stops helping.

Profile and application

We seek a PhD in mathematics or computer science with a taste for concentration phenomena, LLMs and theoretical machine learning. Send a CV and a **short** research statement to samuel.vaiter@cnrs.fr and nicolas.keriven@cnrs.fr.

References

[1] H. Blayney, A. Arroyo, J. Obando-Ceron, P. S. Castro, A. Courville, M. Bronstein, X. Dong. A mechanistic analysis of looped reasoning language models. arXiv:2604.11791, 2026.

[2] L. Bohbot, C. Letrouit, G. Peyré, F.-X. Vialard. Token sample complexity of attention. ICML, 2026.

[3] E. Boursier, C. Boyer. Softmax as linear attention in the large-prompt regime: a measure-based perspective. ICML, 2026.

[4] S. Chen, Z. Lin, Y. Polyanskiy, P. Rigollet. Critical attention scaling in long-context transformers. ICLR, 2026.

[5] L. Yang, K. Lee, R. Nowak, D. Papailiopoulos. Looped transformers are better at learning learning algorithms. ICLR, 2024.