

Steve Abreu @ AISIG

Background

In September 2025, I left Groningen and started working full-time at a newly founded [AI startup](#) in San Francisco. I ~~am~~ was a PhD student at the AI department of the University of Groningen, working in the [MINDS group](#) of Prof. Herbert Jaeger. During my PhD, I did research visits at ETH Zurich and University of Gent and research internships at Google and Intel Labs.

During my PhD, I worked on new compute paradigms and hardware for AI, especially those taking inspiration from the brain ("**neuromorphic computing**"). I thought about reservoir computing [\[0\]](#), photonic computing [\[1\]](#), programming paradigms for brain-inspired computers [\[2\]](#), and developed an intermediate representation for neuromorphic computers [\[3\]](#). On Intel's neuromorphic Loihi 2 chip, I worked on biomedical signal processing [\[4\]](#), demonstrated the advantage of highly sparse neural networks for real-time machine learning [\[5\]](#), and developed the first large-scale neuromorphic LLM [\[6, 7\]](#). For an introduction to **low-power LLMs**, see [\[8\]](#). I also did some work on LLM-based agents for XR [\[9, 10\]](#).

Towards the end of my PhD, I became interested in **hybrid transformer-recurrent LLMs** which combine the power of self-attention with the more efficient, brain-like recurrent neural networks [\[11\]](#). I supervised two awesome AISIG students on **MechInterp**-style understanding how hybrid LLMs implement retrieval [\[12, 13\]](#).

Together with Joris Postmus, I worked on **representation engineering** using conceptors. We use conceptors as **steering** matrices in LLMs to detect and steer towards different (potentially safety-relevant) behaviors [\[14, 15\]](#).

Before my PhD, I was most interested in **automated machine learning** (AutoML) [\[16\]](#) and with the advent of coding agents and increasingly capable LLMs, I am returning to this line of research in the startup I've joined where we work on agents that build agents. As such, I am interested in methods for **developing and aligning self-improving AI agents**, including [scalable oversight](#), [debate](#), [self-play](#), and [iterated distillation & amplification](#).

Projects I can supervise

We can brainstorm together to define a specific project idea, or you can propose an idea. Topics:

- Anything related to the safety of recursively self-improving AI models
- Anything related to representation engineering and steering methods
- Anything related to recurrent neural networks or hybrid transformer-recurrent models

Find out more about me

[Google scholar for my publications](#), [LinkedIn profile](#), [Personal website](#).