

Master 2 Research Internship and Ph.D. Proposal

Laboratory: Centre de Nanosciences et de Nanotechnologies (C2N)
Université Paris-Saclay/CNRS.

Address: 10 Boulevard Thomas Gobert, 91120 Palaiseau, France

Advisor: Damien QUERLIOZ

Website: <https://sites.google.com/site/damienquerlioz/>

E-mail: damien.querlioz@universite-paris-saclay.fr

université
PARIS-SACLAY



Deep Learning Meets Emerging Memories for Energy-Efficient Generative AI

You will develop new ways for language models to manage their memory, and investigate how these mechanisms can be implemented and tested using emerging electronic memories. The project welcomes students in computer science, electrical engineering or applied physics. Depending on your interests and background, your work can emphasize algorithms and simulation, memory architectures, or experiments on devices and circuits. The necessary background in deep learning and Bayesian methods can be acquired during the project.

Generative AI can write text, answer questions and assist scientific research, but running these systems requires considerable energy and computing resources. Reducing these demands is an important challenge for making AI more environmentally sustainable and accessible. One part of the problem is how AI remembers information. Language models use an architecture called a transformer: a neural network whose “attention” mechanism identifies which parts of a text are relevant to generating the next part of a response. To work efficiently, the model keeps intermediate representations of the text it has already processed in a working memory called the key-value cache, or KV cache. As a conversation or document becomes longer, this memory grows, and storing and accessing it becomes increasingly costly.

This project will explore a more economical way for AI to remember. Our starting point is recent work presented at ICML 2026 that uses Bayesian principles to manage a compact working memory [1]. The model tracks uncertainty about the information it holds and uses it to control how strongly new information changes what is already stored. You will investigate how these Bayesian mechanisms can be designed together with their electronic implementation. Emerging memories, particularly ferroelectric devices, would store the model’s working-memory state, and we will explore how devices and circuits can efficiently perform its read and update operations. The physical behavior of emerging memories will help shape the algorithms, while the algorithms will guide the design of the memory system. The project will combine language-model simulations with device measurements and selected experimental tests in collaboration with CEA-Leti.

The project builds on our group’s expertise in Bayesian electronics: electronic systems designed to compute with probabilities and uncertainty [2]. Our previous work includes experimental Bayesian computing systems [3], neural networks that estimate their uncertainty [4], and algorithms that balance learning and forgetting [5]. Our recent work on combined ferroelectric-memristor memories provides an additional foundation for exploring new hardware implementations [6].

Funding: The Master’s internship and Ph.D. thesis are fully funded through the PEPR CAMELIA program

During the Master’s internship, you will learn how attention and Bayesian memory work, starting from existing models and research code. You will develop a first prototype in simulation and investigate a focused question: for example, how much information the model can retain when its memory stores numbers with limited precision, or how a particular memory technology affects its energy requirements. Depending on your background, the emphasis can be on learning algorithms, hardware modeling or computer architecture.

During the PhD, you will develop new mechanisms for storing and updating context in language models, taking advantage of the capabilities of emerging memories. The objective is to demonstrate how these mechanisms can preserve useful context while reducing memory requirements and energy consumption. You will evaluate them on increasingly demanding language tasks and connect their performance to hardware questions such as numerical precision, update frequency and the energy cost of memory operations. Your main contribution can emphasize algorithm development, memory architecture or experimental implementation, according to your interests and strengths.

This project takes place within the interdisciplinary INTEGRANO research group at the Centre de Nanosciences et de Nanotechnologies. This group associates research on nanodevices, bioinspired computing and artificial intelligence, with researchers and students of very diverse backgrounds, making it an exciting environment that fosters interdisciplinary thinking. This work will involve collaboration with Elisa Vianello's group at CEA-Leti. You will develop skills in modern deep learning, emerging memories and energy-efficient computing, relevant to research careers in academia, AI software and hardware startups, and industry. For more technical details on the project, contact us!

Selected references

- [1] D. Bonnet et al., "Learning to Remember, Learn, and Forget in Attention-Based Models", ICML 2026; arXiv:2602.09075. Introduces Palimpsa.
- [2] D. Querlioz and E. Vianello, "Bayesian electronics for trustworthy artificial intelligence", **Nature Reviews Electrical Engineering** 2, 846–855 (2025).
- [3] K.-E. Harabi, ..., D. Querlioz, "A memristor-based Bayesian machine", **Nature Electronics** 6, 52–63 (2023).
- [4] D. Bonnet, ..., D. Querlioz, E. Vianello, "Bringing uncertainty quantification to the extreme-edge with memristor-based Bayesian neural networks", **Nature Communications** 14, 7530 (2023).
- [5] D. Bonnet, ..., D. Querlioz, "Bayesian continual learning and forgetting in neural networks", **Nature Communications** 16, 9614 (2025).
- [6] M. Martemucci, ..., D. Querlioz, E. Vianello, "A ferroelectric–memristor memory for both training and inference", **Nature Electronics** 8, 921–933 (2025).