PhD Thesis: Toward Fast and Efficient Methods for Modern Natural Language Processing

- **Position location:** Rennes, France
- **Lab:** IRISA, team Archimedia
- **Supervisors:** Caio Corro, Guillaume Gravier
- **Funding is already secured** via the French National Research Agency

Large language models (LLMs) and deep contextual embeddings (i.e. BERT and its variants) are ubiquitous in natural language processing (NLP). Unfortunately, they come at a cost : both fine-tuning and decoding are costly procedures, especially when one does not have access to sufficient GPU resources.

The aim of this PhD project is to propose novel methods that allow to build faster models, or to adapt models at test-time without costly fine-tuning procedures. To this end, we will take inspiration in previous works, for example, but not limited to :

- Speculative decoding, that allows to leverage GPU parallelization capabilities to speed up text generation speed [1,2,3];
- Direct preference optimization, that allows to solve the standard reinforcement learning from human feedback (RLHF) problem with a simpler (and faster in practice) training objective [4];
- Test-time alignment, that relies on a small model to guide a large one [5];
- Approximate inference algorithms that can fully leverage GPU parallelization capabilities [6,7];
- Few-shot adaptation methods [8];
- etc etc.

The specific applications and targeted methods will be decided jointly with the candidate. The PhD is fully funded by the French National Research Agency via the SemiAmor research grant.

The candidate must have an interest for deep learning and natural language processing, but also for algorithmic and optimization.

The candidate is expected to have either a strong computer science background with an interest for applied mathematics, or a strong mathematical background with some knowledge in deep learning and Python+Pytorch. As the goal of the project is to propose novel methods, the candidate is expected to be able to develop his own code and to hack libraries like Pytorch and HuggingFace, i.e. the project will require to go beyond a simple use of tools.

Outcomes of this project are expected to be published in the main natural language processing conferences/journals (*ACL/EMNLP/TACL) and/or main machine learning conferences/journals (NeurIPS/ICLR/ICML/AISTATS/TMLR).

Supervisors:

- Caio Corro (INSA Rennes, IRISA) https://caio-corro.fr/
- Guillaume Gravier (CNRS, IRISA)

To apply, please send an email at caio.corro@irisa.fr with the following documents :

- a CV
- last year of bachelor (licence) and master grades
- a short description of your main interests in NLP and computer science

If you have any question, feel free to send an email to caio.corro@irisa.fr

[1] Fast Inference from Transformers via Speculative Decoding (Yaniv Leviathan, Matan Kalman, Yossi Matias) https://proceedings.mlr.press/v202/leviathan23a.html

[2] QSpec: Speculative Decoding with Complementary Quantization Schemes (Juntao Zhao, Wenhao Lu, Sheng Wang, Lingpeng Kong, Chuan Wu) https://aclanthology.org/2025.emnlp-main.240/

[3] Cactus: Accelerating Auto-Regressive Decoding with Constrained Acceptance Speculative Sampling (Yongchang Hao, Lili Mou) https://openreview.net/forum?id=lpUIkCAy9p

[4] Direct Preference Optimization: Your Language Model is Secretly a Reward Model (Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, Chelsea Finn) https://papers.nips.cc/paper_files/paper/2023/hash/a85b405ed65c6477a4fe8302b5e06ce7-Abstract-Conference.html

[5] KAD: A Framework for Proxy-based Test-time Alignment with Knapsack Approximation Deferral (Ayoub Hammal, Pierre Zweigenbaum, Caio Corro) https://aclanthology.org/2026.eacl-long.179/

[6] Sinkhorn Distances: Lightspeed Computation of Optimal Transport (Marco Cuturi) https://papers.nips.cc/paper_files/paper/2013/hash/af21d0c97db2e27e13572cbf59eb343d-Abstract.html

[7] Bregman Conditional Random Fields: Sequence Labeling with Parallelizable Inference Algorithms (Caio Corro, Mathieu Lacroix, Joseph Le Roux) https://aclanthology.org/2025.acl-long.1430/

[8] Few-shot domain adaptation for named-entity recognition via joint constrained k-means and subspace selection (Ayoub Hammal, Benno Uthayasooriyar, Caio Corro) https://aclanthology.org/2025.coling-main.662/