Research Internship (Master's Level)

Accelerating Flow-Based Language Models through Analytical Denoising

Level	Master's second year (M2) or final year of an engineering degree, or equivalent
Duration	6 months
Location	ENS Lyon & INSA Lyon, France
Start date	01/04/2026 at the latest
Supervision	Ségolène Martin (Inria, LIP) and Christian Wald (Institut Camille Jourdan, INSA)

Scientific context

Generative models for text are usually autoregressive: they produce tokens one after another. *Flow Language Models* (FLMs) provide a different perspective, starting from a noisy continuous representation of a token sequence and progressively transforming it into text. By updating many positions in parallel, they offer a promising alternative to sequential generation. However, their training can be costly, and the factors governing how quickly they learn remain insufficiently understood.

A central issue is *what information a denoising network actually learns*. Predicting a clean token from its noisy representation may rely heavily on information already present at that position, without requiring the model to understand the surrounding words. Yet coherent text generation also depends on learning statistical and semantic dependencies between tokens. The relative importance of these sources of information may vary with the noise level, network architecture and prediction target (e.g., clean-token prediction or a flow velocity field).

Research objectives and methodology

The internship will investigate whether a better understanding of these learning mechanisms can lead to more efficient FLM training. In particular, we will explore the idea of **combining predictions obtained from simple data statistics with a learned neural correction**, so that the network can concentrate on aspects of the task that are not already captured by the analytical component.

The work will be organized around three connected directions:

- **Understanding denoising and contextual learning.** Study which information can be recovered from a noisy token alone and which requires its context. Analyze the influence of noise, architecture and prediction parameterization, initially using controlled synthetic token sequences where relevant conditional probabilities are known.
- **Designing structured predictors.** Investigate ways of incorporating simple statistical knowledge about text into flow-based denoisers. Compare alternative combinations of analytical and neural components, with attention to the information that the network is encouraged to learn and to the computational overhead of the statistical component.
- **Evaluating training efficiency.** Implement promising approaches in an existing FLM framework and compare them with standard training on text datasets (potentially including the One Billion Word Benchmark). Evaluate predictive and generative quality, convergence speed, wall-clock time and GPU memory consumption under comparable computational budgets.

The project combines mathematical analysis, numerical experimentation and deep-learning implementation. A tentative schedule is: months 1–2, analysis and synthetic experiments; months 3–4, method development; months 5–6, larger-scale experiments and evaluation. Results may contribute to a research publication and provide a basis for a possible co-supervised PhD project, subject to mutual interest and funding.

Required profile and skills

- **Education:** M2 student or final-year engineering student in applied mathematics
- **Mathematics and ML:** solid foundations in probability, linear algebra, and deep learning
- **Programming:** Python, PyTorch
- **Research:** interest in mathematical modeling and experimental research, autonomy and critical thinking
- **Additional assets:** familiarity with generative modeling, diffusion/flow matching, language modeling.

Application

Please apply directly on the Inria website and attach the following documents:

- CV
- academic transcripts
- short, personal motivation letter
- (optional) recommendation letter
- (optional) links to relevant projects are welcome.

We will review all applications starting December 7th. The deadline for applying is December 18th. Only complete applications will be considered.