

Florian Valade

Machine Learning Research Engineer

France
+33 6 17 57 19 12
florian_val@outlook.fr
fvalade.fr
florian-valade
FlorianVal

Profile

Machine Learning Research Engineer specializing in efficient deep learning, large language models (LLMs), adaptive inference, and distributed training. Experienced in designing new architectures, running large-scale experiments, building multimodal systems, and optimizing GPU training infrastructure.

Selected Research & Engineering Highlights

- Trained and fine-tuned GPT-like LLMs (70M–1.3B parameters) on multi-GPU clusters (V100/A100) using FSDP, DDP, Accelerate, and custom distributed pipelines.
- Developed adaptive-inference methods (early exits, reject option, recursion) achieving significant FLOP reductions while preserving model accuracy.
- Built large-scale dataset pipelines: streaming loaders, deduplication, filtering, and curation.
- Designed compute-efficient Transformer variants and performed scaling-law analysis on loss–compute trade-offs.
- Built internal LLM tools for workflow automation.
- Managed GPU servers and research infrastructure; optimized throughput, training stability, and reproducibility.

Education

- 2022–Present **PhD in Efficiency of Deep Learning Algorithms**, *University Gustave Eiffel*, Paris, France
2021 **Master's Degree in Computer Science, Big Data and Machine Learning**, *ECE Paris*, Paris, France
2015 **High School Diploma in Science**, *L'Espérance High School*, Paris, France

Publications

- 2025 **EERO: Early Exit with Reject Option** — UAI 2025.
○ Formalizes selective early exits using risk-coverage trade-offs.
○ Respects strict compute budgets and improves the state-of-the-art speed/accuracy trade-off.
- 2024 **Accelerating Large Language Model Inference with Self-Supervised Early Exits**.
○ Introduces fine tuned early-exit extensions for LLMs using internal supervision signals.
○ Enables FLOP-efficient inference with up to 50% speedup.
○ Achieves +66% acceptance rate and 14x fewer wasted tokens compared to speculative decoding.

Experience

- 2026–Current **AI Research Engineer, Fujitsu, Paris, France**
○ Developing and adapting large language models for various applications.
- 2022–2026 **PhD Candidate & Research Engineer, Fujitsu – University Gustave Eiffel, Paris, France**
○ Designed adaptive-inference architectures (early exits, recursion, selective prediction) for vision models and LLMs.
○ Trained GPT-style models (70M–1.3B parameters) with FSDP, DDP, Accelerate, and distributed GPU clusters.
○ Built large-scale data pipelines: filtering, deduplication, quality scoring, and multimodal curation.
○ Conducted FLOP-efficient architecture research and compute scaling experiments.
○ Built internal tooling for evaluation, model comparison, dataset validation, and visualization.
○ Managed multi-GPU servers and training infrastructure.
○ Taught Computer Vision and NLP to Master's students.
- 2021–2022 **Data Scientist, Fujitsu, Paris, France**
○ Developed and deployed computer vision and deep learning systems for industry clients.
○ Worked on multimodal applications combining vision, metadata, and text.
- 2018–2021 **Apprentice Computer Vision Engineer, Fujitsu – ECE Paris, Paris, France**
○ Applied deep learning to detection, segmentation, and embedded systems.
○ Data collection, preprocessing, and pipeline engineering.
- 2018 **Deep Learning Intern, Fujitsu, Paris, France**
○ Built demonstrators using deep learning for vision tasks.

Skills

- Machine Learning LLMs, Transformers, distributed training, multimodal AI, adaptive inference, model efficiency, calibration, data processing.
- Distributed Systems FSDP, DDP, Accelerate, DeepSpeed, Slurm, Docker, Kubernetes, multi-GPU clusters, profiling, monitoring.
- Programming Python, C, C#, Java, SQL.
- Frameworks PyTorch, TensorFlow, JAX, MLX, HuggingFace ecosystem, W&B, Git.
- Languages English (Fluent), French (Native), Spanish (Intermediate).

Projects

- 2025 **Recursive GPT**. Research project exploring scaling laws of recursive Transformers to reduce memory usage and improve FLOP efficiency; developed recursive-layer prototypes and analyzed compute scaling behaviors.
- 2023 **Adaptive Inference for LLMs**. Implemented early-exit heads and internal supervision layers on GPT-like models. Used Pythia and Phi models; fine-tuned intermediate heads; built evaluation tools for FLOP-loss trade-offs.
- 2022 **FreshDetect**. Real-time produce classification deployed using containerized microservices (PyTorch, Docker).
- 2020 **Handterpret**. Infrared-based hand-position detection system using embedded sensors.
- 2017 **AutoCradle**. Baby-cry detection triggering autonomous cradle rocking.