

Profil

Ingénieur de Recherche en Machine Learning spécialisé dans l'apprentissage profond efficace, les grands modèles de langage (LLMs), l'inférence adaptative et l'entraînement distribué. Expérimenté dans la conception de nouvelles architectures, l'exécution d'expériences à grande échelle, la construction de systèmes multimodaux et l'optimisation de l'infrastructure d'entraînement sur GPU.

Sélection de Recherches et Réalisations

- Entraînement et fine-tuning de LLMs type GPT (70M–1.3B paramètres) sur clusters multi-GPU (V100/A100) avec FSDP, DDP, Accelerate et pipelines distribués personnalisés.
- Développement de méthodes d'inférence adaptative (sorties anticipées, option de rejet, récursion) réalisant des réductions significatives de FLOPs tout en préservant la précision du modèle.
- Construction de pipelines de données à grande échelle : chargeurs streaming, déduplication, filtrage et curation.
- Conception de variantes de Transformers efficaces en calcul et analyse des lois de mise à l'échelle sur les compromis perte–calcul.
- Développement d'outils internes basés sur LLM pour l'automatisation des workflows.
- Gestion de serveurs GPU et infrastructure de recherche ; optimisation du débit, stabilité d'entraînement et reproductibilité.

Formation

- 2022–Présent **Doctorat en Efficacité des Algorithmes d'Apprentissage Profond**, *Université Gustave Eiffel*, Paris, France
- 2021 **Master en Informatique, Big Data et Apprentissage Automatique**, *ECE Paris*, Paris, France
- 2015 **Baccalauréat Scientifique**, *Lycée L'Espérance*, Paris, France

Publications

- 2025 **EERO: Early Exit with Reject Option** — UAI 2025.
- Formalise les sorties anticipées sélectives utilisant les compromis risque–couverture.
 - Respecte des budgets de calcul stricts et améliore le compromis vitesse/précision de l'état de l'art.
- 2024 **Accelerating Large Language Model Inference with Self-Supervised Early Exits**.
- Introduit des extensions de sorties anticipées fine-tunées pour LLMs utilisant des signaux de supervision internes.
 - Permet une inférence efficace en FLOPs avec accélération jusqu'à 50%.
 - Atteint +66% de taux d'acceptation et 14x moins de tokens gaspillés comparé au décodage spéculatif.

Expérience

- 2026–Actuel **Ingénieur de Recherche en IA**, *Fujitsu*, Paris, France
- Développement et adaptation de grands modèles de langage pour diverses applications.
- 2022–2026 **Doctorant et Ingénieur de Recherche**, *Fujitsu – Université Gustave Eiffel*, Paris, France
- Conception d'architectures d'inférence adaptative (sorties anticipées, récursion, prédiction sélective) pour modèles de vision et LLMs.
 - Entraînement de modèles type GPT (70M–1.3B paramètres) avec FSDP, DDP, Accelerate et clusters GPU distribués.
 - Construction de pipelines de données à grande échelle : filtrage, déduplication, scoring qualité et curation multimodale.
 - Recherche sur les architectures efficaces en FLOPs et expériences de mise à l'échelle du calcul.
 - Développement d'outils internes pour évaluation, comparaison de modèles, validation de datasets et visualisation.
 - Gestion de serveurs multi-GPU et infrastructure d'entraînement.
 - Enseignement de Vision par Ordinateur et NLP à des étudiants en Master.
- 2021–2022 **Data Scientist**, *Fujitsu*, Paris, France
- Développement et déploiement de systèmes de vision par ordinateur et d'apprentissage profond pour clients industriels.
 - Travail sur des applications multimodales combinant vision, métadonnées et texte.
- 2018–2021 **Apprenti Ingénieur en Vision par Ordinateur**, *Fujitsu – ECE Paris*, Paris, France
- Application de l'apprentissage profond à la détection, segmentation et systèmes embarqués.
 - Collecte de données, prétraitement et ingénierie de pipelines.
- 2018 **Stagiaire en Deep Learning**, *Fujitsu*, Paris, France
- Développement de démonstrateurs utilisant le deep learning pour tâches de vision.

Compétences

Machine Learning	LLMs, Transformers, entraînement distribué, IA multimodale, inférence adaptative, efficacité modèle, calibration, traitement de données.
Systèmes Distribués	FSDP, DDP, Accelerate, DeepSpeed, Slurm, Docker, Kubernetes, clusters multi-GPU, profiling, monitoring.
Programmation	Python, C, C#, Java, SQL.
Frameworks	PyTorch, TensorFlow, JAX, MLX, écosystème HuggingFace, W&B, Git.
Langues	Anglais (Courant), Français (Natif), Espagnol (Intermédiaire).

Projets

- 2025 **Recursive GPT**. Projet de recherche explorant les lois de mise à l'échelle des Transformers récursifs pour réduire l'usage mémoire et améliorer l'efficacité FLOP ; développement de prototypes à couches récursives et analyse des comportements de mise à l'échelle du calcul.
- 2023 **Inférence Adaptative pour LLMs**. Implémentation de têtes de sorties anticipées et couches de supervision internes sur modèles type GPT. Utilisation de modèles Pythia et Phi ; fine-tuning des têtes intermédiaires ; développement d'outils d'évaluation pour compromis FLOP–perte.
- 2022 **FreshDetect**. Classification en temps réel de produits frais déployée via microservices conteneurisés (PyTorch, Docker).
- 2020 **Handterpret**. Système de détection de position de la main basé sur infrarouge utilisant capteurs embarqués.
- 2017 **AutoCradle**. Détection de pleurs de bébé déclenchant le balancement autonome du berceau.