

KHUONG T G HIEU

ktgiahieu@gmail.com
(+33) 7 64 88 28 09
linkedin.com/in/hieu-khuong

PhD Student — Evaluation Methodologies for Generative AI

Summary

I build the infrastructure and develop methodologies for automated evaluation: reproducible evaluation formats ([Croissant Tasks](#), Google Research & MLCommons), benchmark platforms ([Codabench](#), 50k+ users), [positional studies](#) (ICML 2026) and [user experiments](#) (NeurIPS 2024). I am looking to extend these methodologies to multimodal quality evaluation, out-of-distribution robustness, and interpretability of automated judges.

Experience

03/2025 – present

Doctoral Researcher

Région Île-de-France & Codabench project

LISN — Orsay, France

Research on evaluation methodologies for LLMs.

- Organized the [NeurIPS 2025 social “The Role of AI in Scientific Peer Review”](#)
- Built a declarative metadata format for reproducible ML evaluations, prioritizing “conceptual reproducibility” over “technical reproducibility”, from which LLM agents can regenerate full evaluation pipelines ([NeurIPS 2026 submission](#), Google Research & MLCommons).
- Proposed a framework to help peer review with unprecedented submission volumes: automated LLM-as-judge evaluation handles verification (reproducibility, claim checking, compliance, citation integrity), freeing expert human judgment for more important assessments ([ICML 2026 workshop](#)).
- Created a randomized-control-trial-based evaluation methodology to host an Ile-de-France challenge for the national job search platform, France Travail ([Codabench & Région IdF](#))
- Built the evaluation framework for LLM-assisted grading in [CorrectExam](#), a tool that helps university professors grade assignments (AMUE–Mistral, ILaaS)
- Built and deployed the [NeurIPS’24 LLM Checklist Assistant](#): automated compliance evaluation of scientific papers at a real venue, with user studies on failure-mode and gameability analysis (Microsoft, Carnegie Mellon University, NeurIPS program chairs)

03/2024 – 09/2024

AI Researcher Intern (LLM/RAG)

BNP Paribas CIB — Paris, France

- Fine-tuned an E5-Mistral retriever on synthetic persona-diverse data with partial LoRA updates
- Instruction-tuned open-source LLMs with hard-negative samples to mimic out-of-distribution domain
- Built an automated end-to-end RAG evaluator to score the LLM pipeline correctness

05/2023 – 09/2023

AI Research Intern (collaboration with Google)

Inria (TAU) & Google — France

- Created the Auto-Survey Challenge, a benchmark evaluating LLMs as both generators and reviewers of scientific surveys, and presented at AutoML 2023 (Googler Initiated Grant 2023, supervised by Isabelle Guyon & Marc Schoenauer)

Education

Ph.D.
2025 – 2028
(expected)

Université Paris-Saclay — LISN / Inria TAU, ED STIC

Advisors: Isabelle Guyon & Nicolas Thiéry

Thesis: benchmark & evaluation methodologies for large language models

Funded by the Région Île-de-France & Codabench project

M.Sc. in Artificial
Intelligence
2022 – 2024

Université Paris-Saclay

SaclAI Scholarship of Excellence · ChaLearn scholarship

My Publications

O. Benjelloun, L. Martins Bianco, I. Guyon, **T.G.H. Khuong**, J. Lebensold, S. Lobentanzer, L. Oala, B.K. Rachmat, I. Ullah, P. Vahidi, J. Vanschoren (2026). Croissant Tasks: A Metadata Format for Reproducible Machine Learning Evaluations. *arXiv:2605.29786*.

P. Singh, **T.G.H. Khuong**, V. Sikimić, B.K. Rachmat, K. Ayonrinde, I. Ullah, C. Lioma, L. Oala, K.Q. Lin, L. Cao, N.F. Abernethy, H. Weerts, J. Vanschoren (2026). Position: AI Should Verify, Not Judge, Scientific Work. *AI for Science Workshop @ ICML 2026 (Oral)*.

A. Goldberg, I. Ullah, **T.G.H. Khuong**, B.K. Rachmat, Z. Xu, I. Guyon, N. Shah (2024). Usefulness of LLMs as an Author Checklist Assistant for Scientific Papers: NeurIPS'24 Experiment. *arXiv:2411.03417*.

R. Mussard, A. Gauffre, I. Ullah, **T.G.H. Khuong**, M.-R. Amini, I. Guyon, L.H. Sun (2024). Stylized Meta-Album: Multi-domain computer vision meta-dataset. *DMLR*.

T.G.H. Khuong, B.K. Rachmat (2023). Auto-Survey Challenge: Advancing the Frontiers of Automated Literature Review. *JDSE 2023*.

L.D. Hanh, **K.T.G. Hieu** (2021). 3D matching by combining CAD model and computer vision for autonomous bin picking. *IJIDeM 15*, 239–247.

Teaching & Mentoring

2025-2026 Taught the M1 Artificial Intelligence course at Université Paris-Saclay, and mentored a team of M.Sc. students to fine-tune deep vision models for wheat varietal estimation with INRAE, leading to a publication at CAP 2026.

2023-2024 Taught the M1 Artificial Intelligence course at Université Paris-Saclay, and mentored a team of M.Sc. students to design a benchmark of peer-reviewed paper, leading to a publication at CAP 2024.

Publications as supervisor

O.A. Houmed, Q.-P. Ho, A. Moujar, O. Oloruntobi Paul, G.D.M. Okou, R. Lu, **T.G.H. Khuong**, M. Ecarnot, S. Chevallier, T. Flutre, L. Sun-Hosoya, F.P. Landes. Estimating Varietal Proportions in Wheat Grain Mixtures with Fine-Tuned Deep Vision Models. *Conférence sur l'Apprentissage automatique (Cap 2026)*.

P.H. Couto, Q.-P. Ho, N. Kumari, B.K. Rachmat, I. Ullah, L. Sun-Hosoya, **T.G.H. Khuong**, (2024). RelevAI-Reviewer: A Benchmark on AI Reviewers for Survey Paper Relevance. *Conférence sur l'Apprentissage automatique (Cap 2024)*.

Conferences & Community

Jul 2026 **Oral: Position: AI Should Verify, Not Judge, Scientific Work** — AI for Science Workshop, ICML 2026, Seoul

Dec 2025 **Co-organizer:** NeurIPS 2025 social “*The Role of AI in Scientific Peer Review*”, San Diego — with I. Guyon (Google DeepMind); panelists from CMU, Stanford, Sakana AI, TTIC, Oregon State & Google DeepMind; authored the official session summary

Sep 2023 **Challenge presentation:** Auto-Survey Challenge — AutoML 2023 & JDSE 2023

Awards & Prizes

2023 SacAI Scholarship of Excellence, Université Paris-Saclay

2023 & 2022 ChaLearn scholarship

2023 1st place & Creativity Award, Hi! Hackathon Paris (energy forecasting)

2021 – 2022 [Kaggle](#) Top 4 Kaggle Days Nowcasting (invited to Kaggle Days event, Barcelona) · Top 1% CommonLit Readability · Top 2% chai QA (Google Research) · Top 3% Feedback Prize · Top 4% Ventilator Pressure (Google Brain)

Additional Skills

Evaluation & benchmarking — LLM/VLM-as-judge design & validation, benchmark construction, Codabench infrastructure · **ML engineering** — PyTorch, computer vision, LoRA/instruction fine-tuning, pretraining, RAG, agentic pipelines, W&B

Programming — Python, C++, C · **Languages** — English, French, Vietnamese

References

Prof. Isabelle Guyon

Google DeepMind & ChaLearn USA
guyon@chalearn.org

Prof. Nicolas Thiéry

Université Paris-Saclay, LISN
nthiery@liscn.fr

Prof. Lisheng Sun Hosoya

Université Paris-Saclay & DatalA
lisheng.sun@universite-paris-saclay.fr