Supervision: Hugo Larochelle (Mila), Ulrich Aïvodji (ÉTS/Mila)

Duration: 24 months

Keywords: AI research assistants, trustworthy AI, peer review, LLM evaluation

Context Artificial agents equipped with natural language processing and tool-use capabilities, able to plan and execute sequences of actions on behalf of users, are rapidly changing the daily work of AI researchers. They are now used to summarize papers, generate hypotheses, assist with coding, write reviews, polish manuscripts, check mathematical arguments, and support scientific decision-making. These capabilities could help AI researchers work faster, verify claims more rigorously, and reduce the burden of tasks such as literature review and peer review [2]. However, they also raise important risks. If they introduce hidden biases [3], homogenize scientific feedback, enable paper laundering [1], or generate unverifiable claims, they may weaken rather than strengthen the research ecosystem.

Goals The goal of this postdoctoral project is to study how AI systems can support the work of AI researchers while preserving scientific rigor, fairness, diversity of perspectives, and verifiability. It will focus on two complementary work streams:

Evaluate AI-assisted peer-review systems The first work stream will study AI reviewers and AI-assisted reviewing systems as high-stakes decision-support tools. The project will ask which parts of the peer-review pipeline can be safely assisted by AI, under which constraints, and with which evaluation criteria. The project will build on recent findings showing that AI reviewers can exhibit affiliation bias, hidden status-based preferences, excessive agreement, and vulnerability to stylistic manipulation. It will also study more constructive designs, including reviewer personas, ensembles, metareviewing, literature-grounded reviewing, and human-AI collaboration protocols.

Develop AI assistants for verifiable mathematical reasoning in AI/ML. The second will study the use of LLMs to formalize typical mathematical claims in AI/ML papers, identify the classes of claims that are currently amenable to AI-assisted theorem proving, and characterize the claims that remain beyond the reach of existing systems. It will also examine how retrieval from formal libraries can improve proof search in domain-specific AI/ML settings, and how verification-aware research assistants can help authors detect missing assumptions, ambiguous theorem statements, and gaps in reasoning before submission.

Settings and environment The postdoc will join a research environment focused on trustworthy AI, privacy, fairness, and explainability. The project will benefit from connections with Mila, ÉTS, and broader research communities working on AI evaluation and governance. The position offers significant freedom to shape the research direction. A successful candidate may focus more strongly

on AI reviewer evaluation, formal verification for AI/ML mathematics, or the intersection of both. In all cases, the project will emphasize open science, reproducible evaluation, and tools that can be useful to the research community.

Requirements We are seeking a highly motivated candidate with a strong interest in trustworthy AI and AI-assisted scientific research. We are particularly looking for candidates with:

- Strong background in machine learning, natural language processing, trustworthy AI, formal methods, or AI for science.
- Good programming skills in Python and modern machine-learning frameworks.
- Experience with LLM evaluation, benchmarking, fairness, or peer review systems is an asset.
- Experience with theorem proving, Lean, formal verification, or mathematical reasoning is an asset.
- Strong written communication skills and the ability to formulate research ideas clearly.
- A PhD in computer science, machine learning or a related field is required or should be close to completion.

How to apply Applications consisting of the following elements should be sent by email to ulrich.aivodji@etsmtl.ca and hugo.larochelle@mila.quebec as soon as possible and not later than September 15th:

- Detailed CV
- Letter of motivation
- Two references or recommendation letters

References

[1] Joachim Baumann, Jiaxin Pei, Sanmi Koyejo, and Dirk Hovy. Stop automating peer review without rigorous evaluation. *arXiv preprint arXiv:2605.03202*, 2026.

[2] Gaurav Sahu, Hugo Larochelle, Laurent Charlin, and Christopher Pal. Reviewertoo: Should ai join the program committee? a look at the future of peer review, 2025. *URL https://arxiv.org/abs/2510.08867*.

[3] Sai Suresh Macharla Vasu, Ivaxi Sheth, Hui-Po Wang, Ruta Binkyte, and Mario Fritz. Justice in judgment: Unveiling (hidden) bias in llm-assisted peer reviews. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 307–330, 2026.