Statement of Purpose

Chiung-Yi Tseng ctseng@luxmuse.ai https://ctseng777.github.io

My research centers on advancing **AI-assisted mathematical proof, autoformalization, formal verification, and interpretable reasoning** as interrelated approaches to aligning AI systems with human values. These directions directly address the pressing challenge of LLM opacity and the need for provably beneficial AI. While large language models (LLMs) exhibit remarkable capabilities, they largely remain "black boxes"—deep, inscrutable systems whose reasoning processes are hidden [Jiang et al., 2023b]. This opacity poses serious risks in safety-critical domains [Han et al., 2021], eroding trust and allowing subtle errors to go unnoticed. My recent research on *LLM-as-judge* confirms the unreliability of LLM-based evaluations widely adopted in research settings [Feuer et al., 2025]. Another recent project, *StreetMath*, investigates approximation behaviors in LLMs and reveals their divergence from human reasoning patterns in everyday contexts (`https://ctseng777.github.io/assets/pdf/StreetMath_2025.pdf`).

One pillar of my work focuses on **AI-assisted mathematical proof** and **autoformalization**—leveraging AI to translate informal human reasoning into formally verifiable logic. Mathematical proofs represent the gold standard of rigor [Aggarwal et al., 2024]. Recent work demonstrates that models such as GPT-4 can solve challenging mathematical problems yet often generate plausible-sounding outputs that contain reasoning errors or hallucinations [Azerbayev et al., 2024]. By translating such reasoning into formal systems (e.g., Lean or Coq), we can verify each step against strict logical standards [Yang et al., 2023]. Autoformalization not only enhances verifiability but has also been shown to improve reasoning quality and trustworthiness [Jiang et al., 2023a]. In particular, LLMs have been successfully used to generate formal Lean 4 proofs step by step, allowing proof assistants to flag incorrect inferences [Azerbayev et al., 2023]. These methods unite the creativity of natural language models with the reliability of formal reasoning, achieving substantial gains in performance and correctness [Jiang et al., 2024].

Another central interest of mine lies in integrating **formal verification** into AI reasoning loops to ensure **interpretable and trustworthy behavior**. Chain-of-thought prompting encourages LLMs to produce intermediate reasoning steps [Yang et al., 2023], but these explanations are not guaranteed to be correct. Most existing benchmarks evaluate only final answers, omitting scrutiny of intermediate reasoning and offering no formal guarantees [Lample et al., 2022]. To overcome these limitations, I aim to design hybrid systems in which each reasoning step is verified against a formal specification. Formal verification provides principled, provable correctness rather than relying on heuristics [Hilton, 2024], and it is essential for developing aligned AI that behaves safely and predictably.

Beyond technical contributions, I bring strong self-motivation and a breadth of experience in research ideation, experimentation, and scientific communication. My background also includes extensive industry experience in multimodal models, inference optimization, and large-scale infrastructure. In parallel, I place great value on collaboration and community engagement. I have served as a reviewer for AAAI 2026 and the MathAI Workshop at NeurIPS 2025, and volunteered at the HackHer hackathon. I deeply enjoy working with inspiring peers to pursue impactful research, and I seek a mission-driven community where I can contribute my skills while advancing the foundations of beneficial AI.

References

Pranjal Aggarwal, Bryan Parno, and Sean Welleck. AlphaVerus: Bootstrapping formally verified code generation through self-improving translation and treefinement. *arXiv preprint arXiv:2412.06176*, 2024.

Zhangir Azerbayev, Bartosz Piotrowski, Hailey Schoelkopf, Edward W. Ayers, Dragomir Radev, and Jeremy Avigad. ProofNet: Autoformalizing and formally proving undergraduate-level mathematics. *arXiv preprint arXiv:2302.12433*, 2023.

Zhangir Azerbayev, Hailey Schoelkopf, Keiran Paster, Marco Dos Santos, Stephen McAleer, Albert Q. Jiang, Jia Deng, Stella Biderman, and Sean Welleck. Llemma: An open language model for mathematics. *arXiv preprint arXiv:2310.10631*, 2024.

Benjamin Feuer, Chiung-Yi Tseng, Astitwa Sarthak Lathe, Oussama Elachqar, and John P Dickerson. When judgment becomes noise: How design failures in llm judge benchmarks silently undermine validity, 2025.

Jesse Michael Han, Jason Rute, Yuhuai Wu, Eric W. Ayers, and Stanislas Polu. Proof artifact co-training for theorem proving with language models. *arXiv preprint arXiv:2102.06203*, 2021.

Jacob Hilton. Formal verification, heuristic explanations and surprise accounting. *AI Safety Journal*, 2024.

Albert Qiaochu Jiang, Wenda Li, and Mateja Jamnik. Multilingual mathematical autoformalization. *arXiv preprint arXiv:2311.03755*, 2023a.

Albert Qiaochu Jiang, Sean Welleck, Jin Peng Zhou, Timothée Lacroix, Jiacheng Liu, Wenda Li, Mateja Jamnik, Guillaume Lample, and Yuhuai Wu. Draft, sketch, and prove: Guiding formal theorem provers with informal proofs. In *Proceedings of the 11th International Conference on Learning Representations (ICLR)*, 2023b.

Albert Qiaochu Jiang, Wenda Li, and Mateja Jamnik. Multi-language diversity benefits autoformalization. In *Advances in Neural Information Processing Systems, vol. 37 (NeurIPS 2024)*, pages 83600–83626, 2024.

Guillaume Lample, Marie-Anne Lachaux, Thibaut Lavril, Xavier Martinet, Amaury Hayat, Gabriel Ebner, Aurélien Rodriguez, and Timothée Lacroix. HyperTree proof search for neural theorem proving. In *Advances in Neural Information Processing Systems, vol. 35 (NeurIPS 2022)*, pages 26337–26349, 2022.

Kaiyu Yang, Aidan M. Swope, Alex Gu, Rahul Chalamala, Peiyang Song, Shixing Yu, Saad Godil, Ryan Prenger, and Anima Anandkumar. Leandojo: Theorem proving with retrieval-augmented language models. In *Advances in Neural Information Processing Systems, vol. 36 (NeurIPS 2023)*, 2023.