ZEKUN ZHAO

zzhao99@ucsc.edu | zekunzhao.github.io

SUMMARY

My interests focus on model **inference** with an emphasis on **efficiency and effectiveness**.

SKILLS: PyTorch, SGlang, vLLM, Kubernetes, LLM training/Inference, RL, GRPO, synthetic data, Lean 4 verification

EDUCATION

University of California, Santa Cruz | Anticipated graduation date: 10/27
Ph.D. in Natural Language Processing, Computer Science Engineering

Nankai University, Tianjin
Bachelor of Engineering in Intelligent Science and Technology

RESEARCH PROJECTS

Fast LLM Inference with Parallel Prompting | Aug 2025 – Present

- Designed **Parallel Prompting**, an inference method for Transformer LLMs that decodes many questions about one long document as parallel streams of a single request. All streams share the document's KV cache, so each decode step reads the document once instead of once per question. A stream-id attention mask keeps each answer independent.
- Implemented it in the **SGLang** serving engine with about 80 lines across three files plus a small engine package. The changes are an in-kernel stream mask in SGLang's Triton attention kernel, CUDA-graph decoding and continuous batching. Ported it to the latest SGLang releases.
- Compared with SGLang's fastest stock configuration on 8k–16k-token documents, it was 3.7× faster on an A100 and **5.9–6.0× faster** on an RTX A6000. It was faster in all 144 paired benchmark runs and 1.3–1.9× faster on Qwen2.5 models from 7B to 32B.
- On QuALITY, a long-document question-answering benchmark (2,086 questions about 115 articles), answered the full benchmark 1.81× faster than stock SGLang on an A100. Continuous batching added a further 1.22× speedup when answer lengths varied.
- **Website&Code:** https://jlab-nlp.github.io/parallel-prompting/

Reinforcement Learning for Efficient Formal CoT Verification | Aug 2025 – Apr 2026

- Built a **GRPO** reinforcement-learning pipeline that trains **Qwen2.5-Coder** to translate natural-language chain-of-thought reasoning into Lean 4 theorem statements, running distributed jobs on **Kubernetes** GPUs
- Designed a neuro-symbolic reward that combines Lean 4 compiler checks with an LLM judge of semantic faithfulness, plus guards against reward hacking (copying the example, forbidden commands, extra proofs)
- Designed a CoT-to-Lean translator that produces gold statements (100% compile coverage on 52.7K problems) for an SFT warm-up for a 0.5B model, raising the share of test statements that compile from 6% to 39%
- Made each training step **3× faster** (180s → 60s) by deduplicating and caching judge verdicts, running judge requests in parallel, and turning off judge reasoning, which also removed truncated, verdict-less replies

PUBLICATIONS

- **Zhao Zekun**, Zilong Li, Jon Cai, Federico Ortega-Riba, Martha Palmer, James Martin, and Jeffrey Flanigan. 2026. Autoformalizing Chain of Thought: Representation, Verification, and Logical Shortcuts. *NeurIPS 2026 Workshop on Linguistic Principles for Foundation Models*.
- **Zekun Zhao**, Jeffrey Flanigan. 2025. Parallel Prompting: Fast LLM Inference for Shared-Context, Short-to-Moderate Output. *2025 BayLearn Symposium*.
- Kristin Wright-Bettner, Jon Cai, **Zekun Zhao**, James H. Martin, Jeffrey Flanigan, Martha Palmer, Stelios Piperidis, Núria Bel, Henk van den Heuvel, and Nancy Ide. 2026. Building Bridges between Student and Curricular Language: Creating a Corpus of Abstract Meaning Representations for the Classroom. In *Proceedings of the Fifteenth Language Resources and Evaluation Conference*, pages 11960–11971.
- Jon Cai, Kristin Wright-Bettner, **Zekun Zhao**, Shafiuddin Rehan Ahmed, Abijith Trichur Ramachandran, Jeffrey Flanigan, Martha Palmer, and James H. Martin. 2025. LiDARR: Linking Document AMRs with Referents Resolvers. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 426–435.