Last updated in Jan 2025

Saibo Geng

INN 315, EPFL, Lausanne, Switzerland | saibo.geng@epfl.ch | github.com/Saibo-creator

About

I'm interested in (1) LLM inference efficiency (2) Structured output with LLM (3) LLM domain adaptation

Expected to graduate in 2026

Education

Ecole Polytechnique Fédérale de Lausanne (EPFL), PhD in Computer Science – Lausanne, Switzerland — Sept 2022 – present

- Working in Data Science Lab under the supervision of **Prof. Robert West**
- Research focus: Large Language Models (LLM) and neural symbolic approaches

Ecole Polytechnique Fédérale de Lausanne (EPFL), MSc in Electrical Engineering – Lausanne, Switzerland — Sept 2019 – June 2022

- Computer Science Minor

University Paris-Saclay, BSc in Physics – Orsay, France — Sept 2016 – June 2019

- Ranked top 5% in the department

Work Experience

Student Researcher, Microsoft – Redmond, WA, USA — Oct 2024 – present

- Developed the first systematic benchmark for Structured Output with LLM, covering three aspects: structure coverage, runtime efficiency, and generation quality.

Research Intern, Microsoft – Redmond, WA, USA — June 2024 – Oct 2024

- Working with the Guidance team under the supervision of Harsha Nori and Dr. Eric Horvitz
- Modelling the out-of-distribution and calibration issue with constrained decoding in LLM

Peer-Reviewed Publications

Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning (EMNLP 2023) — 2023

Saibo Geng, Martin Josifoski, Maxime Peyrard, Robert West

aclanthology.org/2023.emnlp-main.674

Sketch-Guided Constrained Decoding for Boosting Blackbox Large Language Models without Logit Access (ACL 2024) — 2024

Saibo Geng, Berkay Döner, Chris Wendler, Martin Josifoski, Robert West

aclanthology.org/2024.luhme-short.23

Preprints

Generating Structured Outputs from Language Models: Benchmark and Studies — 2025

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, Harsha Nori

arxiv.org/abs/2501.10868

Byte BPE Tokenization as an Inverse string Homomorphism — 2024

Saibo Geng, Sankalp Gambhir, Chris Wendler, Robert West

arxiv.org/abs/2412.03160

Flows: Building blocks of reasoning and collaborating ai — 2023

Martin Josifoski, Lars Klein, Maxime Peyrard, Nicolas Baldwin, Yifei Li, ***Saibo Geng***, Julian Paul Schnitzler,, Yuxing Yao,, Jiheng Wei,, Debjit Paul, Robert West
arxiv.org/abs/2308.01285

Open Source Projects

Transformers-CFG 2024

- A library for Context-Free Grammar constrained decoding with LLM (100+ stars on GitHub). Available at epfl-dlab/Transformers-CFG

AIflows 2024

- framework for collaborative AI agent (200+ stars on GitHub). Available at epfl-dlab/aiflows

Skills

Programming: Proficient with Python; good understanding of Python Virtual Machine, meta-programming

Languages: English: Fluent (TOEFL: 110/120), French: Fluent (C1), Chinese: Native

Extracurricular Activities

Awesome-LLM-Constrained-Decoding 2024 – present

- Collection of papers, blogs, and tools for LLM constrained decoding (100+ stars on GitHub). Access at Link

Stackoverflow Python Question contributor 2022 – present

- Top 0.5% of the year, 2000+ reputation

Open Source contributor to Huggingface Transformers 2023 – present

- PR #26304: Low-Memory Beam Search Optimization
- PR #27797: Constrained Beam Search Issue Fix
- PR #27557: Grammar-Constrained Decoding

Other Open Source contribution 2023 – present

- TEXT-GENERATiON-WEBUI PR #4953: Context-Free Grammar Constrained Text Generation
- LMQL PR #336: add support for torch compile with HF models
- LQML PR #334: add a basic QueryBuilder, test and documentation

Awards

EPFL Doctoral Fellowship 2022

ACM SIGMOD Programming Contest Finalist 2021

Paris-Saclay Excellence Scholarship 2019