Saibo Geng

(+41) 79-520-77-49 | saibo.geng@epfl.ch | Google Scholar Profile

Research Interests

- Formal Grammar-Constrained Decoding, CFG, Regular Expressions, EBNF
- Efficient Decoding Methods for Large Language Models, low-memory beam search
- LLM for Domain-Specific Language Generation, Structured Text Generation, Information Extraction

Education

Swiss Federal Institute of Technology, Lausanne (EPFL) *Lausanne, Switzerland*
PhD in Computer Science *Sep. 2022 - Present*

- Supervisor: Prof. Robert West (EDIC PhD Fellowship)

Swiss Federal Institute of Technology, Lausanne (EPFL) *Lausanne, Switzerland*
M.S. in Electrical Engineering *Sep. 2019 - Mars. 2022*

- Minor in Data Science

University Paris-Saclay *Orsay, France*
B.S. in Physics *Sep. 2017 - Jun. 2019*

- Paris-Saclay Excellence Scholarship

Publications

Sketch-Guided Constrained Decoding for Boosting Blackbox Large Language Models without Logit Access *ACL 2024 Main [Paper]*
Saibo Geng, Berkay Doner, Chris Wendler, Martin Josifoski, Robert West *Jan. 2024*

- We propose a novel method to boost the performance of blackbox large language models without logit access.
- Our method extends the scope of constrained decoding to blackbox models and achieves strong performance

Grammar-Constrained Decoding for Structured NLP Tasks without Finetuning *EMNLP 2023 Main [Paper]*
Saibo Geng, Martin Josifoski, Maxime Peyrard, Robert West *Oct. 2023*

- We formulate a series of NLP tasks as **constrained text generation** problems described by a **formal grammar**.
- Our method **doubles** the performance of LLaMA models on various tasks without finetuning.

Flows: Building Blocks of Reasoning and Collaborating AI *Preprint [Paper]*
Martin Josifoski,Lars Klein, Yifei Li, Maxime Peyrard, **Saibo Geng** et al. *Nov. 2023*

- Introduces the conceptual framework of Flows, a novel approach for modeling complex interactions in AI systems.
- Our experiments suggest that structured reasoning and collaboration substantially improve generalization, adding **54%** absolute improvement in competitive programming solving rate.

Industry Experience

Microsoft Research *Redmond, WA, USA*
Research Intern *June. 2024 - Sept. 2024*

- Working with Dr. Harsha Nori and Dr. Eric Horvitz on the Guidance Project at Microsoft Research.
- Researching and developing calibration methods for constrained text generation methods with LLM
- Develop backtracking and classifier-in-the-loop methods for improving the safety of LLM

Invited Talks

Topic: Grammar-Constrained Decoding for Reliable generation with LLM *Prilly, Switzerland*
NexThink *June 2024*

- NexThink is a Swiss Unicorn company that provides end-user experience management solutions with a focus on digital employee experience.
- Presented the Grammar-Constrained Decoding method for reliable generation with LLM and shared insights on the potential applications in query language generation.

Honors

2023	**Stack Overflow Reputation: 2K+**, Top 0.5%
2022	**EPFL EDIC PhD Fellowship**, EPFL
2021	**Finalist**, ACM SIGMOD Programming Contest
2019	**Paris-Saclay Excellence Scholarship**, Paris-Saclay University

Open Source Contributions

Transformers-CFG(main author)

- A library for integrating context-free grammars (CFG) in EBNF with the Hugging Face Transformers.
- Features: Prefix Tree based sampling, Unicode support for CFG, Dynamic Programming based parsing, and more.

Huggingface Transformers

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

Text-generation-webui

- PR 4953: Context-Free Grammar Constrained Text Generation

LMQL

- PR 336: add support for torch compile with HF models
- PR 334: add a basic QueryBuilder, test and documentation