BibTeX Grammar for Constrained LLM Generation with SynCode

1. Implementation Overview

The solution involved two primary components: the **Lark Grammar** (`bibtex.lark`) and the **SynCode Integration Script** (`bibtex_syncode.py`).

Lark Grammar (`bibtex.lark`)

To implement the `bibtex.lark`, I've learned the Lark grammar from https://github.com/lark-parser/lark/blob/master/docs/grammar.md and followed the example ANTLR BibTex files.

In Lark, all rules are lowercase, and terminals are uppercase. I define 13 standard BibTeX entry types (e.g., `ARTICLE`, `BOOK`), each of them using case-insensitive matching (e.g., `/@article/i`) for flexibility. To simplify the parser, a common structure for all entries was established: `ENTRY_TYPE "{" key "," field_list "}"`.

The `value` rule was designed to accept quoted strings (`QUOTED_STRING`), braced strings BRACED_STRING), or integer literals.

The main implementation was in the `BRACED_STRING` terminal, which uses a complex regex to tokenize the entire field value, including single-level nested braces (e.g., `\{([^{}]|\{[^{}]*\})*\}`), ensuring titles with internal braces are parsed correctly as a single unit.

Besides, the `field_list` rule was defined as `field ("," field)* [","]`. The optional trailing comma (`[","]`) is included to accommodate common but optional BibTeX syntax practices, preventing common parsing failures. Whitespace and BibTeX comments (`%...`) were ignored globally.

SynCode Integration (`bibtex_syncode.py`)

My script uses the SynCode library to enforce the grammar during the LLM's generation process.

I choose the `google/gemma-2-2b-it` model to initialize the `Syncode` class. `max_new_tokens` was increased to `2048` to accommodate the generation of multiple, complex BibTeX entries.

To constrain the LLM's answers, I use a `FEW_SHOT_TEMPLATE` to guide the model's output format, ensuring it starts with the `@` symbol and adheres to a specific BibTex structure. The `messages` list is initialized with the `FEW_SHOT_TEMPLATE` combined with the user's request.

The `syn_llm.infer(messages)` method ensures that every token generated by the model is checked against the grammar's accepted next tokens, guaranteeing a valid BibTeX output for the three different prompts tested.

Here's the result of the three sample prompts:

```
amazinguknow@wirelessprv-10-194-132-8 syncode-bibtex % /usr/local/bin/python3 bibtex_syncode.py
Initializing SynCode with model: 'google/gemma-2-2b-it' and grammar: 'bibtex.lark'
[2025-11-18 18:02:52,293-root] - Loading model google/gemma-2-2b-it with device:cuda, device_map:auto, torch_dtype:torch.bfloat16
Loading checkpoint shards: 100%|██████████| 2/2 [00:04<00:00,  2.46s/it]
Initialization complete. Running prompts...

============================================================
PROMPT 1:
Generate a BibTeX entry for 3 recent papers on LLM security

SYNCODE-CONSTRAINED OUTPUT:

@inproceedings{chen2022robust,
  title={Robustness of Large Language Models to Adversarial Examples},
  author={Chen, Y. and Liu, Y. and Wang, J. and Zhang, Y. and Zhang, Y.},
  booktitle={Proceedings of the 2022 Conference on Fairness, Accountability, and Transparency in Machine Learning},
  year={2022},
  publisher={Association for Computing Machinery},
  doi={10.1145/3514493.3514503}
}
@inproceedings{liu2023adversarial,
  title={Adversarial Attacks on Large Language Models: A Survey},
  author={Liu, Y. and Chen, Y. and Wang, J. and Zhang, Y. and Zhang, Y.},
  booktitle={Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency in Machine Learning},
  year={2023},
  publisher={Association for Computing Machinery},
  doi={10.1145/3714493.3714503}
}
@inproceedings{wang2023security,
  title={Security and Privacy Challenges of Large Language Models},
  author={Wang, J. and Chen, Y. and Liu, Y. and Zhang, Y. and Zhang, Y.},
  booktitle={Proceedings of the 2023 Conference on Fairness, Accountability, and Transparency in Machine Learning},
  year={2023},
  publisher={Association for Computing Machinery},
  doi={10.1145/3714493.3714503}
}
============================================================
```

```
============================================================
PROMPT 2:
Create a BibTeX citation for a conference paper by authors Smith and Johnson

SYNCODE-CONSTRAINED OUTPUT:

@inproceedings{smith_johnson_2023,
  author = {Smith, John and Johnson, Jane},
  title = {Title of the Paper},
  booktitle = {Conference Name},
  year = {2023},
  publisher = {Publisher Name},
  address = {City, Country}
}
============================================================


============================================================
PROMPT 3:
Provide the BibTeX entry for the RL book by Barto and Sutton

SYNCODE-CONSTRAINED OUTPUT:

@book{bart2018 reinforcement,
  title = {Reinforcement Learning: An Introduction},
  author = {Barto, A. G., & Sutton, R. S.},
  year = {2018}
}
============================================================
```