PL-CUB Homework 3

Perplexity Team

October 2025

1 State your goal in 1-2 sentences

Determine the impact of syntactic features of a language on performance of LLMs on it.

2 State three research questions that would be interesting for your team

1. Current benchmarks comparing the productivity of different programming languages do not account for differences in standard libraries and language features. We want to mitigate this bias and more fairly compare the impact of syntax. Is there a bias of pretrained LLMs towards any specific existing language syntax (Python, C, Lisp, etc.)?
2. Is it possible to obtain code generation metrics like 'pass@k' comparable to existing languages on the same benchmark?
3. How different can we get results by changing only the syntax (changing grammar), but not the language features, LLM, or benchmark?

3 Describe your research methodology

We take test-based HumanEval benchmark (Python subset), where the LLM is tasked with generating a function that passes unit tests. We create a wrapper that for each test case directly invokes Evaluator with the parameters from the test and obtains the result. We prompt LLM to generate the function in the new syntax from textual description, and then validate the generated function using wrapped unit tests. Thus, we can obtain pass@k metrics as in original benchmark. This way, we can compare pass@k metrics between different syntaxes and the original Python language.

3.1 What changes will you make to your language to investigate the research question?

Answer: Adding multiple syntaxes (with their respected parsers) for the same AST and Evaluator. The syntaxes will be each similar to some existing language, like LISP, C or Python semantics. We also plan to experiment with boolean features (contains / not contains) for syntax search space, such as comments.

3.2 How will those changes inform your answer to the research question?

Answer: We are going to evaluate each syntax. The resulting code generation metrics will answer the research question.

3.3 What resources will you require, aside from your time?

Answer: We plan to evaluate the languages on relatively small 7B models, which is achievable without any costs. However, if we were to evaluate GPT-4, we would probably require no more than 5-10 million tokens. HumanEval contains 164 samples. The prompt will be primarily dominated by language description, which we estimate to be about 2000 tokens. So, we get: $2000 \times 164 = 328k$ tokens for a single syntax. We plan to have about 10-15 languages, so we get: $328k \times 15 \approx 5M$ tokens.

3.4 Why is this methodology a reliable way to answer the question?

Answer: It is the only reasonable way of comparing syntaxes that we could think of.