Gorilla: Large Language Model Connected with Massive APIs

Meenakshi Rajpurohit

January 2026

1 Title

State-of-the-art performance, but highly biased on ML based API and does not explain potential misuse of automatic API execution.

2 Summary

The paper proposes Gorilla, an **LLaMA-7B** instruction fine-tuned large language model that is connected to APIs. It employs the novel **Retriever Aware Training(RAT)** for accurate document retrieval and mitigates the behaviour of hallucination with **Abstract Syntax Tree (AST)** data structure. Gorilla adopts test-time version changes, which results in compatibility with frequent change in API. The APIBench includes varieties of dataset from TensorHub, TorchHub and HuggingFace APIs. Gorilla achieves state-of-the-art performance by achieving 20.43, 51.88, 65.59 percentage accuracy gain than GPT-4 and 29.57 percentage, 3.4x, 14x reduction in hallucination on Torch Hub, HuggingFace and TensorFlow API datasets.

3 Problem Statement

The potential of state-of-the-art LLMs in effective tool use via API calls is unfulfilled because of their unawareness of frequently changed API and tool set. The general purpose LLM behaves unpredictably and hallucinates during the execution of corresponding API in different environments is also a key deployment challenge.

4 Proposed Solution

The proposed methodology is the combination of instruction based fine tuning and retrieval which is called Retriever-Aware Training (RAT). The core idea is to train the LLM to judge the retrieved context, to decide when the retrieved API documentation is useful, and when it should be ignored to support informed and robust decision making. The abstract syntax tree evaluates an LLM generated API call performance by hallucination before even executing it. API calling with a mix of RAT and hallucination measurements leads to much better performance.

5 Strengths

- **State-of-the-Art Performance:** The reported results are numerically impressive. Achieving 65.59 percent accuracy gain on TensorFlow API dataset represents a significant improvement over a strong competitor GPT-4 on accurate retrieval of API documentation. If reproducible, these results establish a new benchmark for API calling quality on ML datasets.

- **Fine tuning and effective retrieval:** The novel approach of Retriever Aware Training(RAT), trains the LLM to retrieve correct API calls. It also trains the model to be robust so that it learns to rely on retrieval when it helps and falls backs to internal knowledge when it does not. It's trained to handle noise gracefully.

- **Hallucination Measurement:** The offline metric given by Abstract syntax tree(AST) evaluates that LLM generated API call is hallucinated without executing it. This offline evaluation is very valuable in reinforcement learning and preference based translation especially when API execution is expensive, requires credentials, or has real-world side effects.

6 Weaknesses

- The APIBench datasets are selected from TensorHub, TorchHub and HuggingFace which are only constrained to perform ML tasks (NLP, Computer Vision) and tool uses but the paper does not explain about other major digital world applications such as cloud and database, etc. This is a significant gap.

- Retrieval aware training allows the model to remain relevant even if API documentation updates (Figure 6) but the dependency on a retrieval is also a key challenge.

- The execution environments including command-line interface, Google Colab, python scripts are generalized that is proposed by the paper; however, it does not demonstrate the practical implementation details. It makes Gorilla hard to deploy across local and cloud environments.