Benchmarking Function-Calling enabled LLMs

Shortcomings of LLMs

1. LLMs do not contain detailed information about many long-tail entities, such as products, events, local businesses, or music recordings
2. information might be outdated

Function calling gives LLMs the ability to decide to invoke user-defined functions to retrieve additional, up-to-date information from APIs or from the Web.

Function calling seems very useful, but does it really work beyond vendor demos?

Benchmarking Function-Calling enabled LLMs

Project Goal

Develop a benchmark for testing the function-calling capabilities of GPT 3.5 and GPT 4.

Involves

- Design LLM test cases that require invoking functions
- Model evaluation and error analysis

Questions to answer:

- Does the LLM decide to invoke the right functions?
- How good is the LLM at translating textual questions into function calls?
- Does it properly combine external answers into an overall result?

Requirements

- programming skills in Python
- relevant courses: Data Mining I, Text Analytics recommended

Organization: 4-9 people, 6 months

Instructors: Keti Korini, Christian Bizer