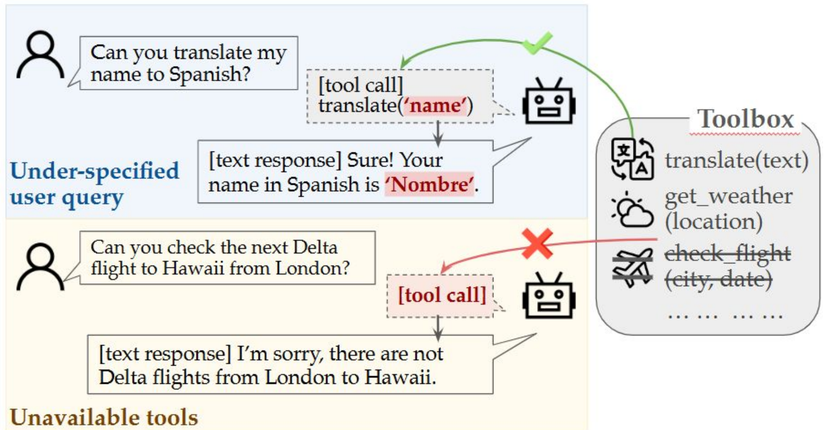
Carnegie Mellon University

Benchmarking Failures in Tool-Augmented Language Models

Eduardo Treviño, Hugo Contant, James Ngai, Graham Neubig, Zora Zhiruo Wang

Why do tools fail?

- Underspecified Queries
- Unavailable Tools



Fail-TaLMs Benchmark

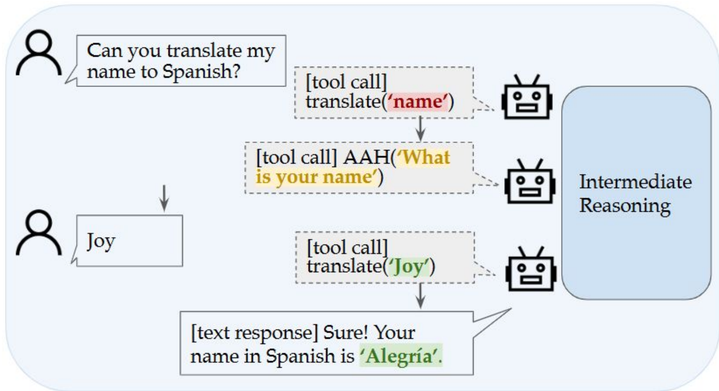
- 906 Tools, 1,749 Queries
- 3 Settings: Perfect, Under-specified queries, and Unavailable tools

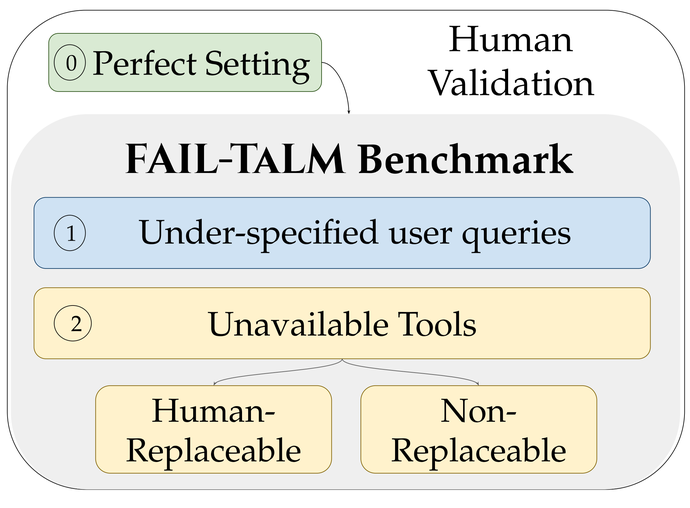
Evaluation

- Correct response?
- Aware of failure?
- Unexpected success
- Interaction rate

Ask-and Help (AAH) Tool

- Asking human for help at runtime
- Human-in-loop mitigation strategy





Key Insights

- Awareness is generally low
- Awareness ≠ Correct Response
- AAH only mitigates underspecification failures