Benchmarking Failures in Tool-Augmented Language Models

Eduardo Treviño Hugo Contant James Ngai Graham Neubig Zora Zhiruo Wang
Carnegie Mellon University

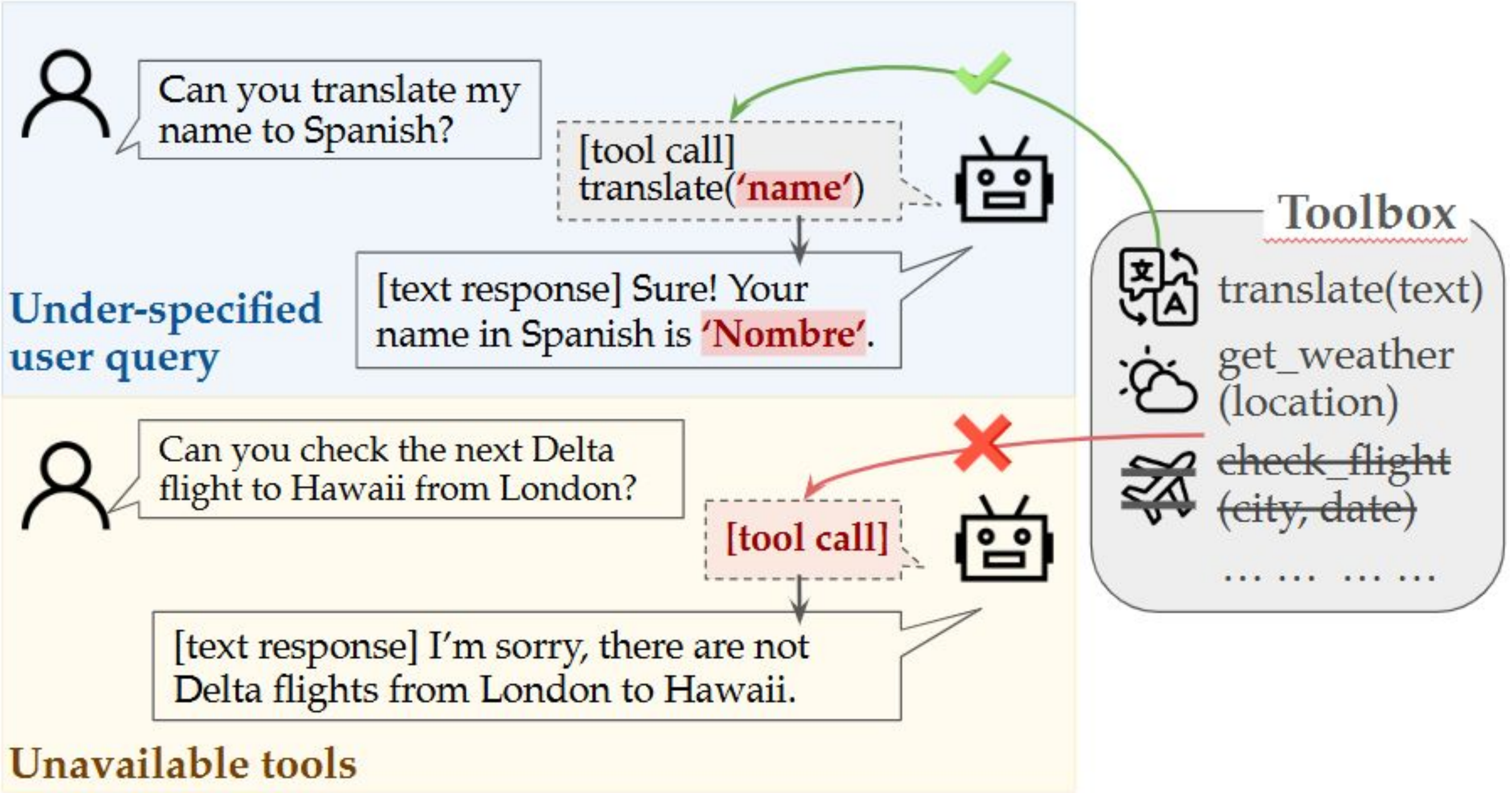
Tool-Augmented LMs (TaLMs) often assume

- perfect information access
- perfect tool availability

Introducing our Fail-TaLMs benchmark to systematically study practical TaLM failures

1 Why do tools fail?

- Under-specified queries
- Unexpectedly unavailable tools



Can you translate my name to Spanish?

[tool call] translate('name')

[text response] Sure! Your name in Spanish is 'Nombre'.

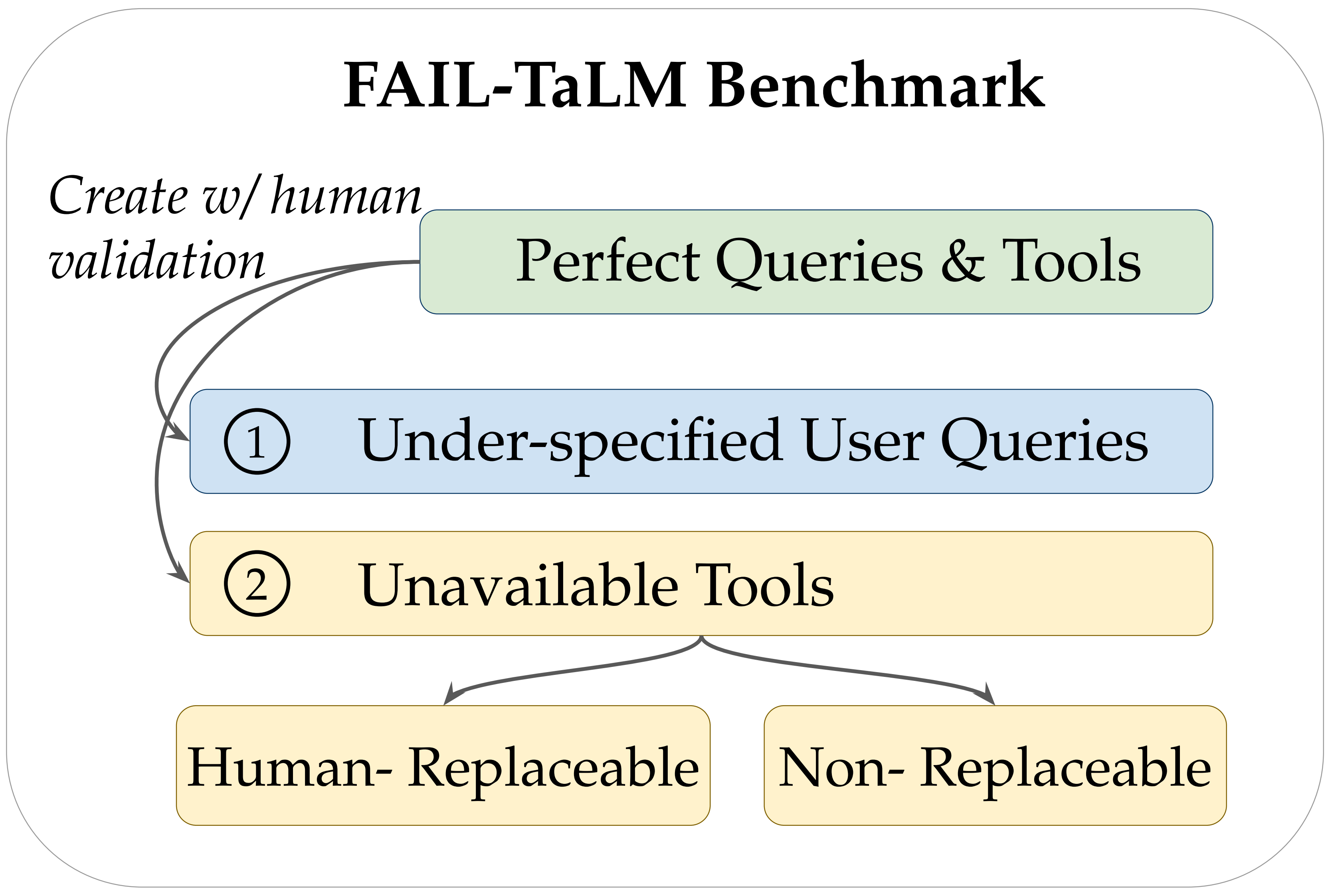
Can you check the next Delta flight to Hawaii from London?

[tool call]

[text response] I'm sorry, there are not Delta flights from London to Hawaii.

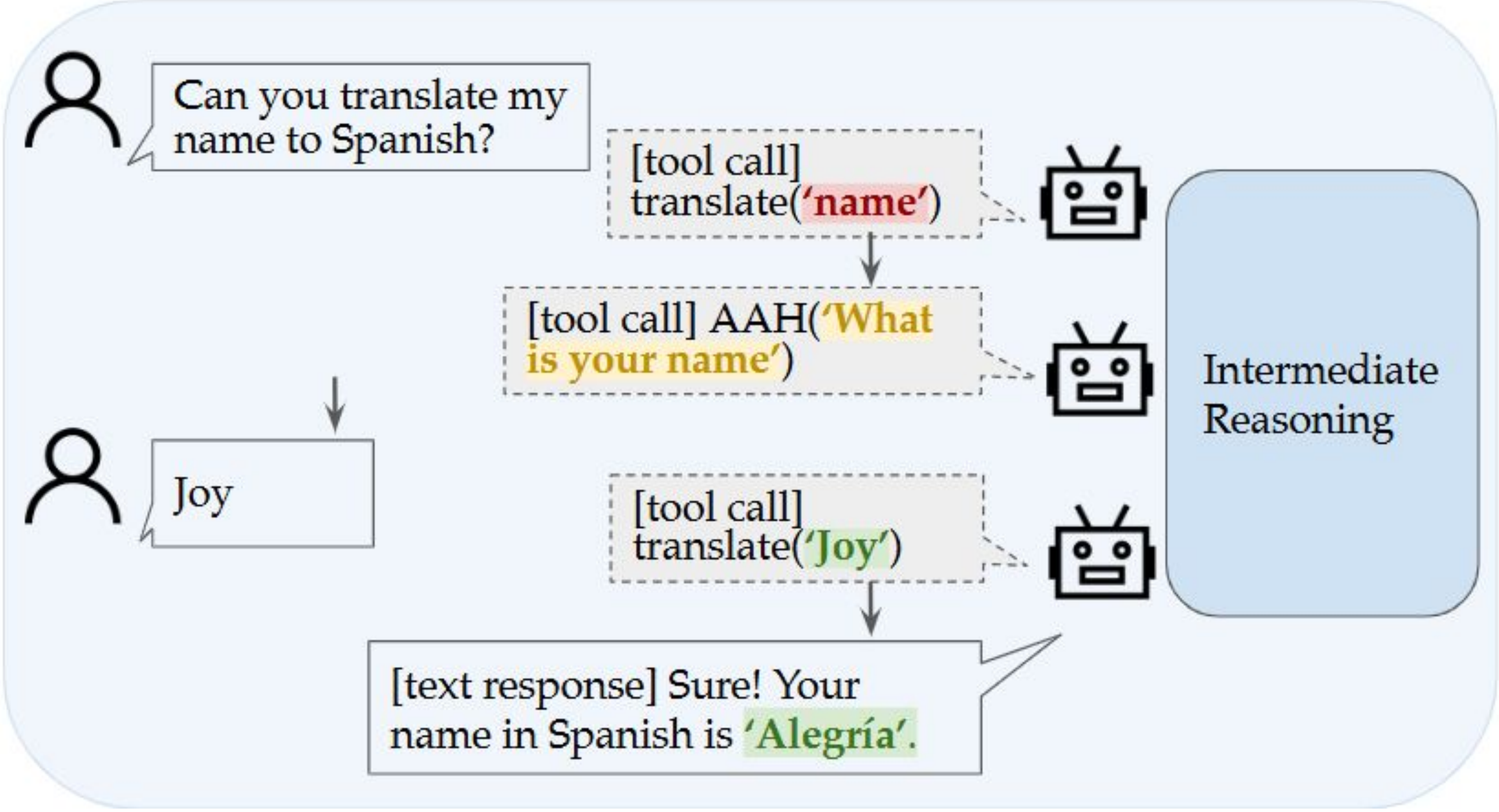
2 Our Fail-TaLMs Benchmark

- 1749 queries + 906 tools
- 3 Settings:
  - perfect
  - under-specified query
  - unavailable tools
- Evaluation
  - Correct response?
  - Unexpected success
  - Aware of failure?
  - Interaction rate



3 Ask-and-Help (**AAH**) Tools

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



Can you translate my name to Spanish?

[tool call] translate('name')

[tool call] AAH('What is your name')

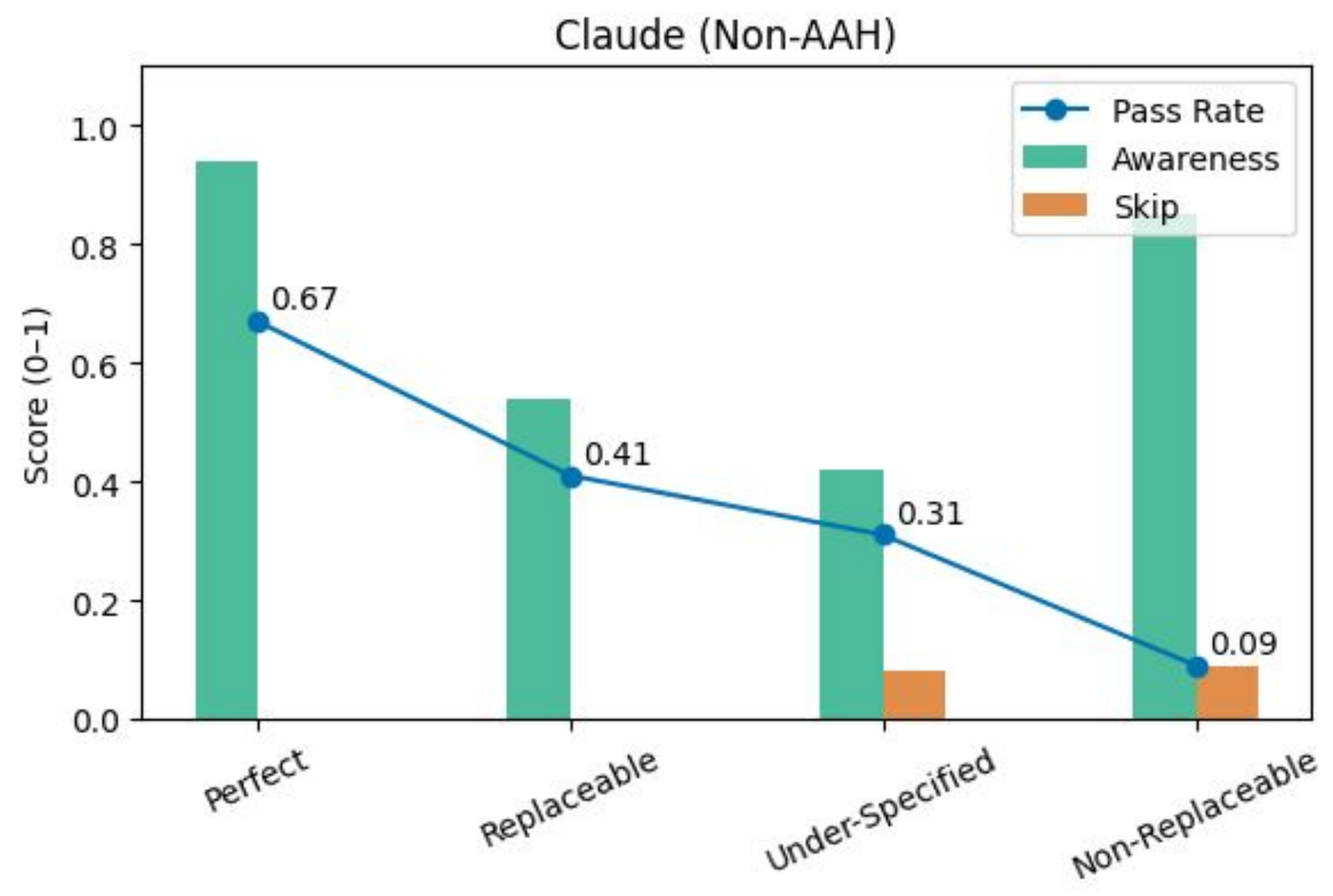
Joy

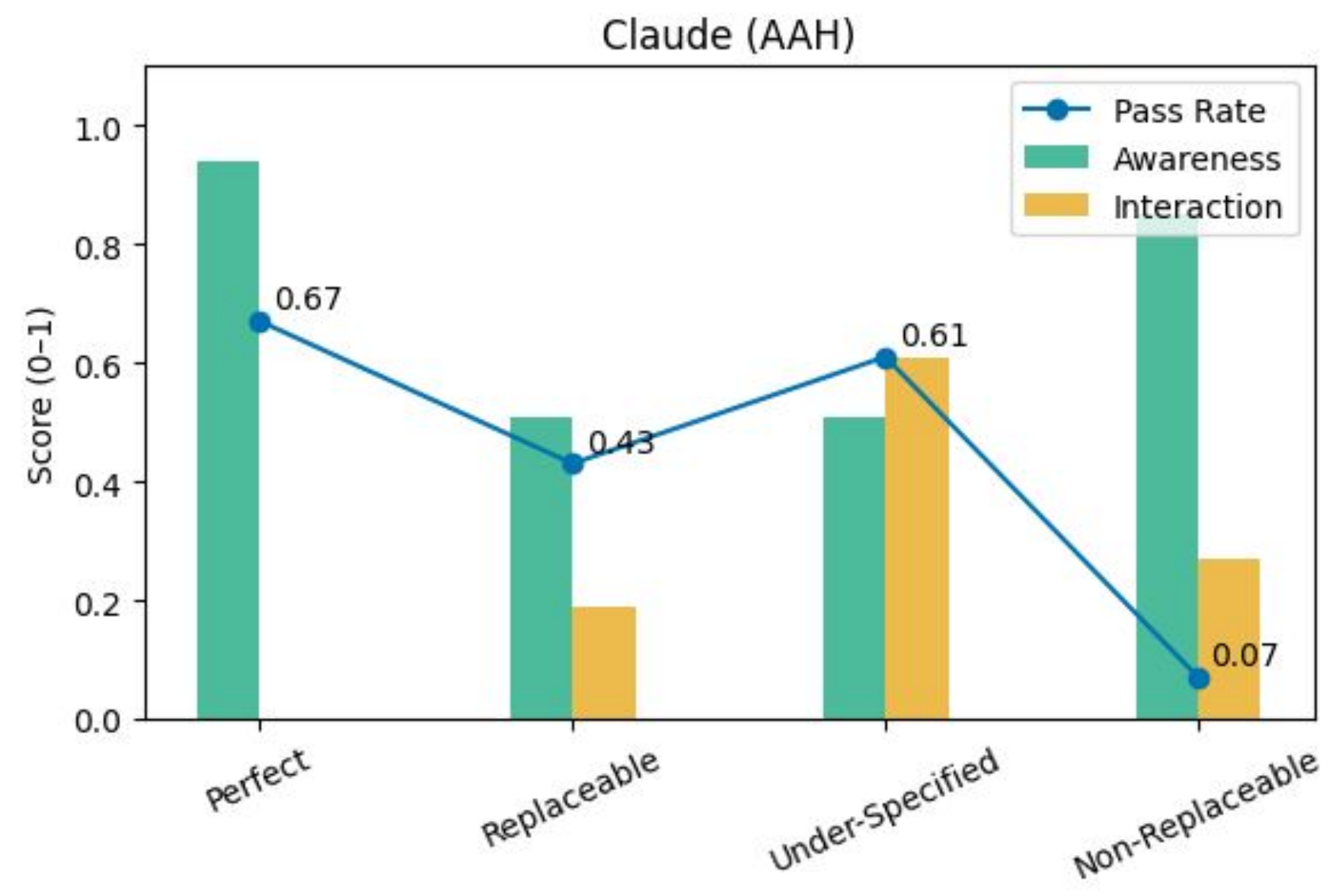
[tool call] translate('Joy')

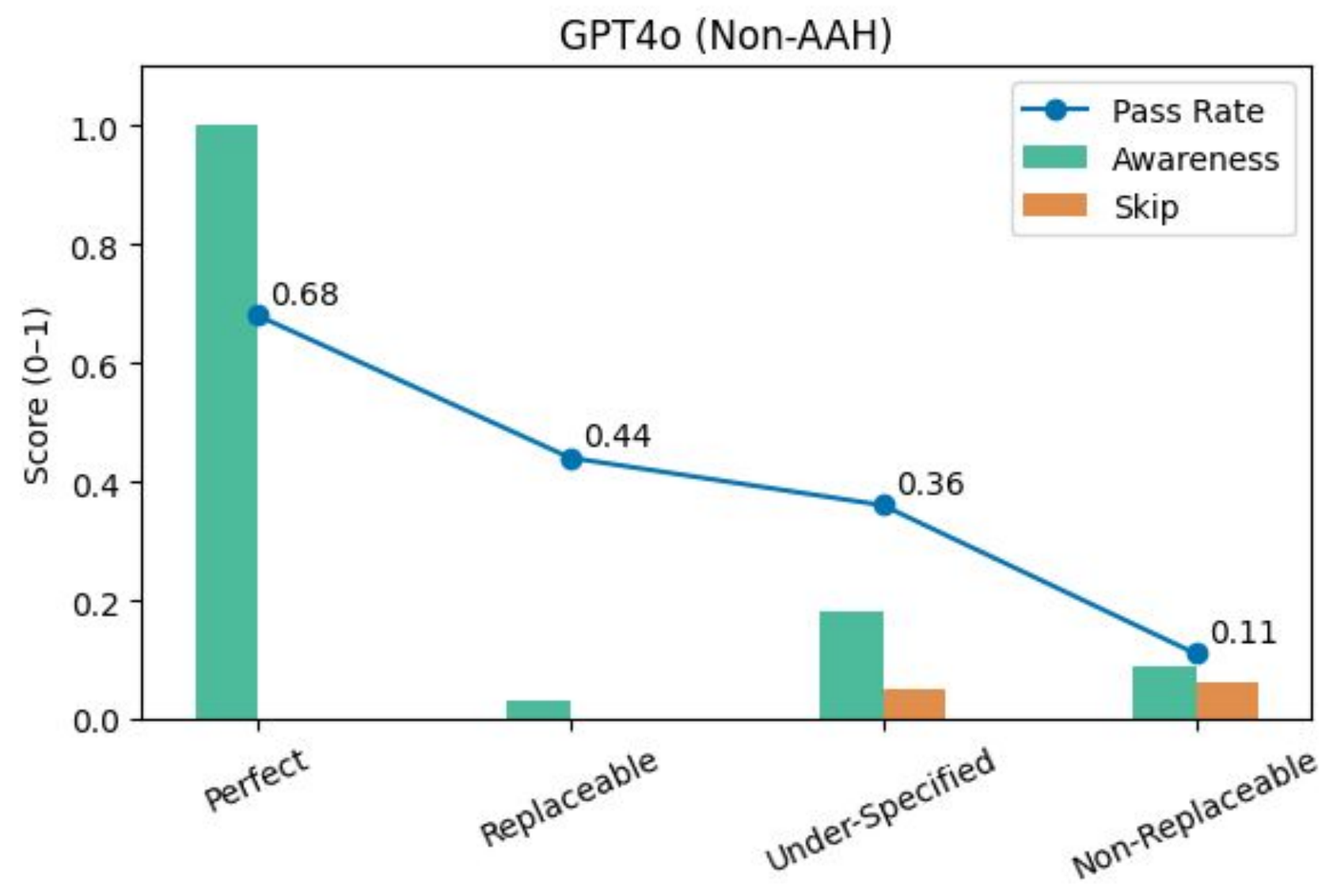
[text response] Sure! Your name in Spanish is 'Alegría'.

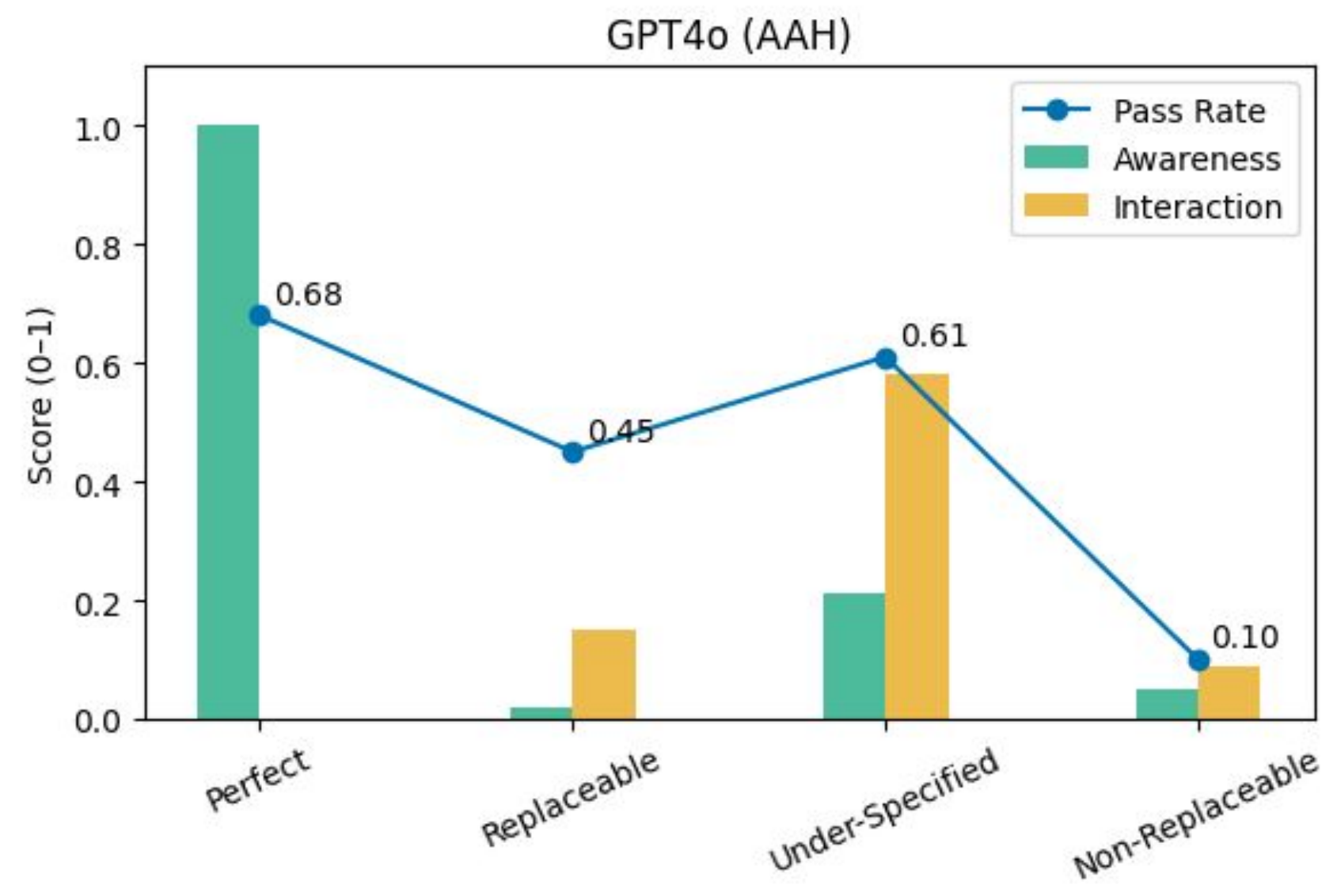
4 Key Insights

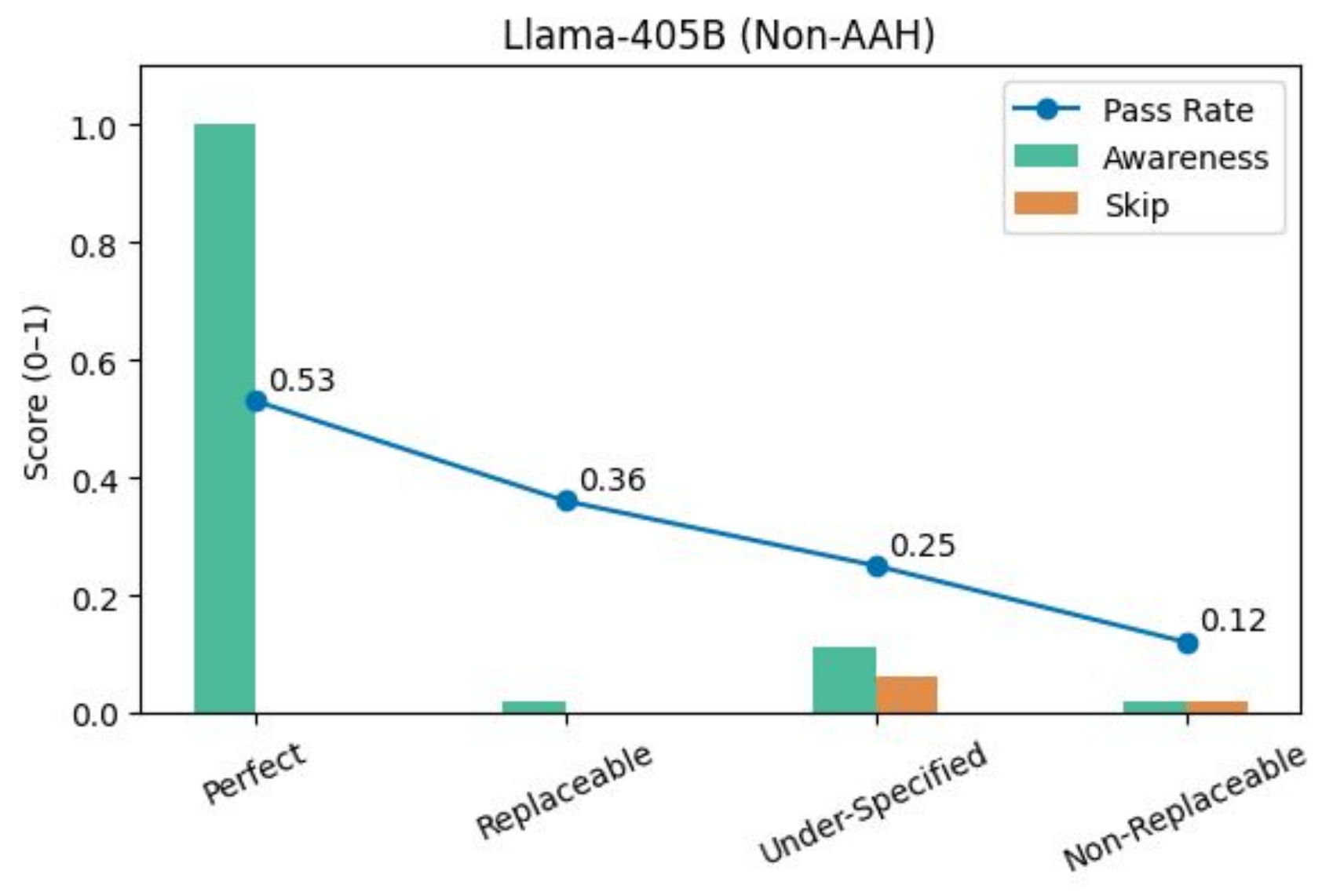
- TaLMs have low awareness :(
- Aware of failure ≠ Task Success
- AAH helps specify queries, but limited for unavailable tools

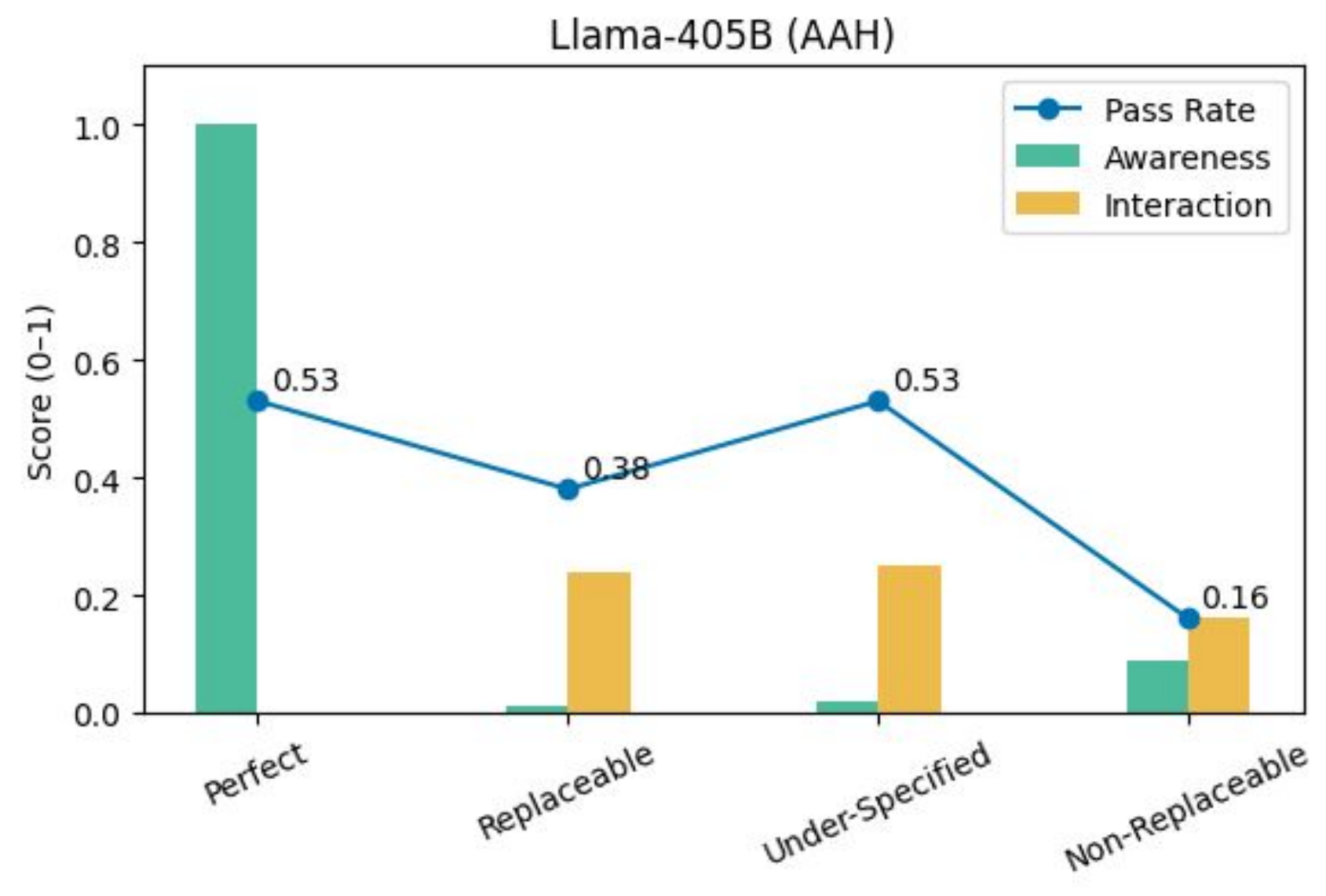












https://arxiv.org/pdf/2503.14227