Hell or High Water: Evaluating Agentic Recovery from External Failures

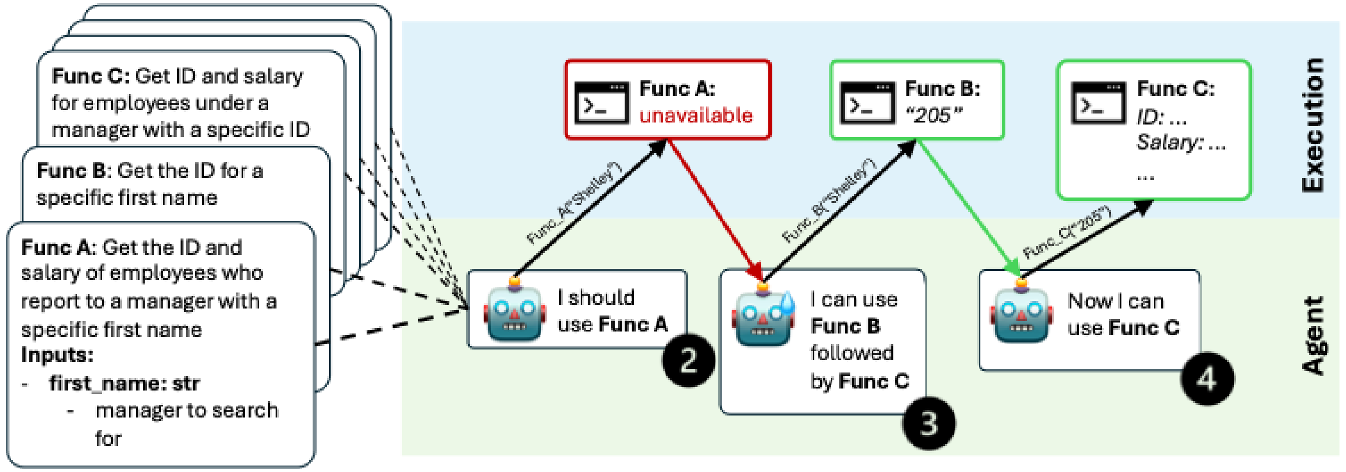
Andrew Wang, Sophia Hager, Adi Asija, Daniel Khashabi, Nicholas Andrews

Johns Hopkins University

Tools break in the real world — how well do LLMs deal with tool failures?

Agentic tool use research has focused on analyzing and correcting **agent errors**. A few works study whether LLMs can *identify* **tool errors**. Our work goes one step further and measures how well LLM agents find backup solutions if tools fail.

1 What are the employee ids of employees who report to Shelley, and what are their salaries?

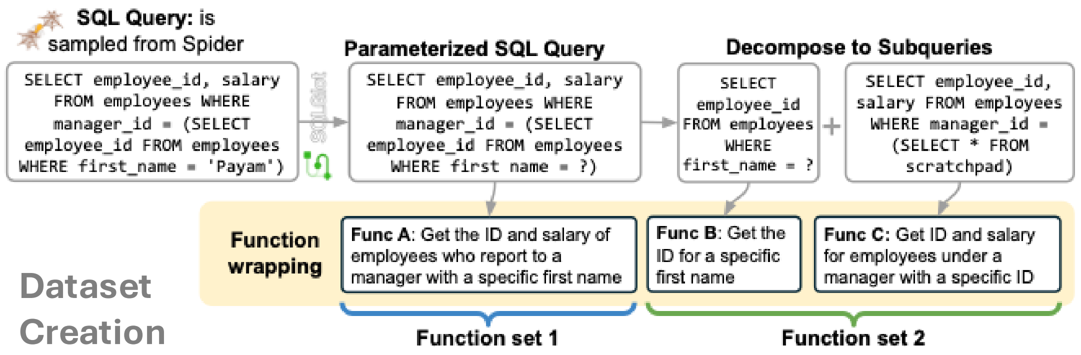

Func C: Get ID and salary for employees under a manager with a specific ID
Func B: Get the ID for a specific first name
Func A: Get the ID and salary of employees who report to a manager with a specific first name
Inputs:
- first_name: str
- manager to search for
Func A: unavailable
Func B: "205"
Func C: ID: ... Salary: ... ...
Func_A("Shelley")
Func_B("Shelley")
Func_C("205")
I should use Func A
2
I can use Func B followed by Func C
3
Now I can use Func C
4
Execution
Agent

Large set of functions + docs

Agent Workflow

We introduce HOHW, a tool-use benchmark where problems remain solvable even when tools break adversarially. HOHW consists of **830 problems** and **4450 available tools**

- The agent correctly tries to use Func A, but encounters an external error outside its control: Func A is unavailable. Therefore, it must form a backup plan to use the composition of Func B and Func C.



```
SELECT employee_id, salary
FROM employees WHERE
manager_id = (SELECT
employee_id FROM employees
WHERE first_name = 'Payam')
```

```
SELECT employee_id, salary
FROM employees WHERE
manager_id = (SELECT
employee_id FROM employees
WHERE first name = ?)
```

```
SELECT
employee_id
FROM employees
WHERE
first_name = ?
```

```
SELECT employee_id,
salary FROM employees
WHERE manager_id =
(SELECT * FROM
scratchpad)
```

Func A: Get the ID and salary of employees who report to a manager with a specific first name

Func B: Get the ID for a specific first name

Func C: Get ID and salary for employees under a manager with a specific ID

Dataset Creation

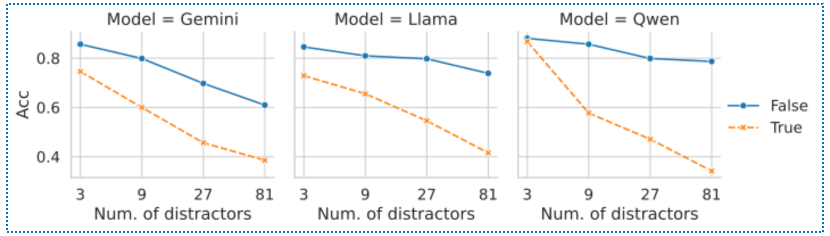
Process:
Input: text-to-SQL (Spider)
Output: task + python tools

1. Find SQL with subqueries
2. Wrap subqueries individually (func. set 2)
3. Wrap original query (func. set 1)

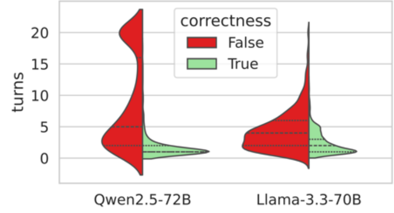
"function sets" refer to groups of functions that together can be used to solve a given question

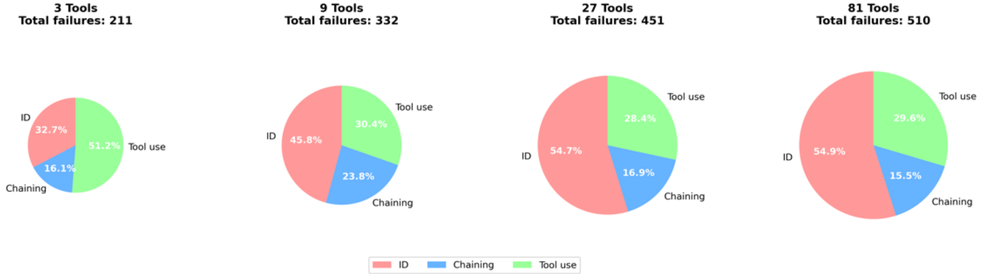
In-context

Tool schemas provided in prompt



Taking more turns doesn't help.





The share of tool identification errors increases with the number of tool schemas in the prompt

RAG

RAG with a vector DB of 4450 tool schemas

Model	External errors (N)	External errors (Y)	Acc. decrease
Gemini 2.0 (Flash)	71.4 ± 1.6	41.1 ± 1.7	42.4%
GPT 4o	60.5 ± 1.7	38.4 ± 1.7	36.5%
Llama 3.3 (70b, Instruct)	64.0 ± 1.7	38.9 ± 1.7	39.2%
Llama 3.1 (70b, Instruct)	42.3 ± 1.7	23.3 ± 1.5	44.9%
Qwen 2.5 (72B, Instruct)	64.1 ± 1.6	35.3 ± 1.7	44.9%

Model	Total failures	Search	ID	Chaining	Tool use
Gemini 2.0 (Flash)	489	57.1	25.5	12.3	6.14
GPT 4o	511	52.6	24.6	8.61	14.1
Llama 3.3 (70b, Instruct)	496	66.3	11.9	13.9	7.86
Llama 3.1 (70b, Instruct)	495	59.3	18.0	11.7	10.9
Qwen 2.5 (72B, Instruct)	637	60.5	26.4	7.06	6.12

What's Next?

- **Agent memory and test-time learning:** Can agents remember which tools to avoid? Can memory modules improve performance?
- **Inference-time compute**: Does thinking more help agents identify backup plans?

Paper

Follow on X