Fault Localization in LLM-Based Multi-Agent Systems

Scope-guided LLM judging for responsible-agent and failure-step attribution

Yavor Pachedzhiev | Delft University of Technology | Bachelor Research Project – Final Poster
Supervisors: Dr. Burcu Kulahcioglu Ozkan | Dr. Annibale Panichella | M.Sc. Zahra Seyedghorban

Problem and Task

Setting. In an LLM-based multi-agent system, several role-based LLM agents work together by sending messages, using tools, and passing intermediate results to each other. A **trace** is the ordered record of those steps.

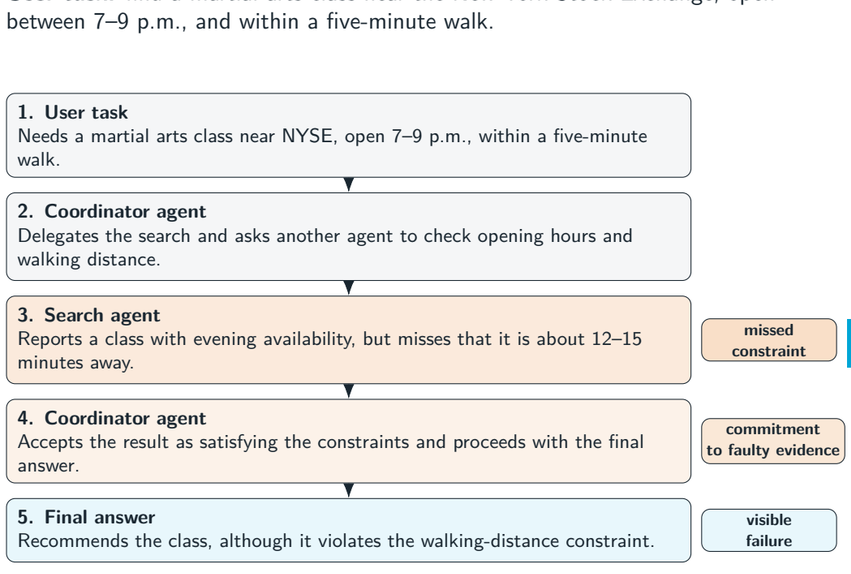
Problem. When the run fails, the final answer often only shows the failure. The step that caused the failure may be earlier, before the final answer is written.

Failure attribution task. For each failed trace, predict:

- the **responsible agent**
- the **decisive failure step**

Example Failed Trace

User task: find a martial arts class near the New York Stock Exchange, open between 7–9 p.m., and within a five-minute walk.



The failure is visible in the final answer, but the important evidence appears earlier: first when a constraint is missed, and then when that incomplete result is accepted.

Dataset and Evaluation

Dataset	Who&When Hand-Crafted [1]
Failed traces	58
Trace steps	2,993 total
Trace length	51.6 mean, 130 max
Model	`gpt-4.1-mini`
Scope budget	up to 5 reference steps

A prediction is counted as correct only if it matches the gold responsible agent, the gold decisive failure step, or both for joint accuracy. Gold labels are not shown to the prompts.

Research Questions

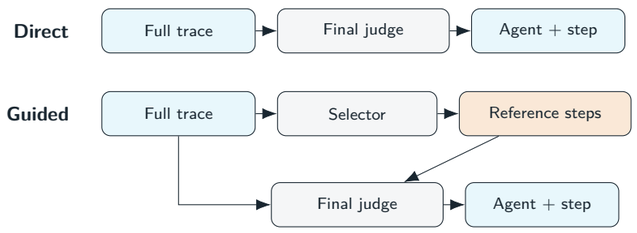
RQ1 – Scope selection. Which selectors recover the gold decisive failure step among the selected reference steps?

RQ2 – Attribution accuracy. Do selected reference steps improve responsible-agent and decisive-step attribution?

RQ3 – Practical trade-off. What extra token cost is introduced by the guided methods?

Judging Setup

All methods use the same final attribution judge. The comparison tests different ways of surfacing trace evidence before that judge makes the final responsible-agent and failure-step prediction.



The selected reference steps are not a candidate list. They are evidence shown alongside the full trace, so the judge can still choose any valid step.

Compared Methods

Method	Role in the comparison
Direct	Whole-trace baseline. No selector is used.
Random	Non-LLM scope baseline. Selects trace steps at random.
Generic LLM	LLM scope baseline. Asks for generally failure-relevant steps.
EASD adapter	Scope-delineation-inspired adapter using overstep-like and loop-like selector calls [2].
Source pool	Project method. Searches for source, commitment, failed-recovery, and premature-termination evidence.

Source-Candidate-Pool Selector

The source-candidate-pool selector tries to make the scope step more source-oriented. Instead of asking only for relevant-looking trace steps, it first collects possible evidence of where the failure was introduced.

Source-candidate-pool stages

1. Build a task brief and visible-failure summary.
2. Split the trace into overlapping chunks.
3. Harvest source, commitment, failed-recovery, and premature-termination candidates.
4. Merge and deduplicate candidates into a bounded pool.
5. Consolidate the pool into up to five selected reference steps.

These selected steps are shown to the same final judge together with the full trace. They guide attention, but the judge still predicts the final responsible agent and decisive step.

Results

Main result: Source pool is the only guided method that improves all final attribution metrics over direct whole-trace judging.

Method	Hit@5	Agent	Step	Joint	Tokens
Direct	–	51.7	24.1	20.7	17.8k
Random	19.0	56.9	22.4	19.0	18.4k
Generic	41.4	51.7	24.1	20.7	38.2k
EASD	22.4	55.2	19.0	15.5	54.9k
Source pool	**44.8**	**60.3**	**29.3**	**25.9**	**75.3k**

Hit@5 means the gold decisive failure step appears among the selected reference steps. Hit@5, Agent, Step, and Joint are percentages. Tokens are mean total tokens per trace.

Answers to the Research Questions

RQ1. Generic and source-candidate-pool selection recover the gold decisive step more often than random scope selection.

RQ2. Only the source-candidate-pool method improves agent, step, and joint accuracy over direct whole-trace judging.

RQ3. The improvement is modest and expensive. Source pool gives the best attribution results, but it also has the highest token cost.

References

[1] Zhang et al. *Who&When: A Failure Attribution Benchmark for Multi-Agent Systems*, 2025.
[2] Sun et al. *Scope Delineation before Localization*, 2026.

Source code: `https://github.com/yavor0/Research-Project`

TU Delft