Evaluating Recursive Multi-Agent Systems

What kind of complex tasks can we use?

- To evaluate ReDel, we started with two agentic benchmarks:
    - **TravelPlanner** (Xie et al., 2024): Use tools to search flight and other DBs, create travel plans
    - **WebArena** (Zhou et al., 2024): Interact with a web browser to do complex tasks like commenting on GitLab

- But these benchmarks don't cover a type of question we want these systems to be able to answer...

24

Evaluating Recursive Multi-Agent Systems

What kind of complex tasks can we use?

- To evaluate ReDel, we started with two agentic benchmarks:
  - **TravelPlanner** (Xie et al., 2024): Use tools to search flight and other DBs, create travel plans
  - **WebArena** (Zhou et al., 2024): Interact with a web browser to do complex tasks like commenting on GitLab
- But these benchmarks don't cover a type of question we want these systems to be able to answer...

Fan-Out Questions