Benchmarking ReDel

Final list of benchmarks

Benchmark	Task	Tools	Example Query
FanOutQA (Zhu et al., 2024b)	Factual question-answering with retrieval tools	Wikipedia Article Search, Retrieve Wikipedia Article	What is the total number of employees in the five largest banks in the world?
TravelPlanner (Xie et al., 2024)	Creation of a structured travel plan that meets hard and commonsense constraints	Search Flights, Calculate Taxi Cost, Search Accommodations, Search Restaurants, Search Attractions, List Cities in State	Please help me plan a trip from St. Petersburg to Rockford spanning 3 days from March 16th to March 18th, 2022. The travel should be planned for a single person with a budget of $1,700.
WebArena (Zhou et al., 2024a)	Interaction with a web browser – mix of information seeking and artifact creation	Click, Type, Hover, Press Key, Scroll, Open New Tab, Change Focused Tab, Close Tab, Go To URL, Go Back, Go Forward	Show me the ergonomic chair with the best rating

Results