NeurIPS 2025
Workshop on Multi-Turn Interactions in LLMs

Reinforcement Learning for Long-Horizon Multi-Turn Search Agents

Weights & Biases by CoreWeave
Vivek Kalyan
research@vivekkalyan.com
Martin Andrews
martin@redcatlabs.com
red cat labs

Motivation

Goal
Solve complex legal search tasks where answer requires navigating massive corpora over multiple turns

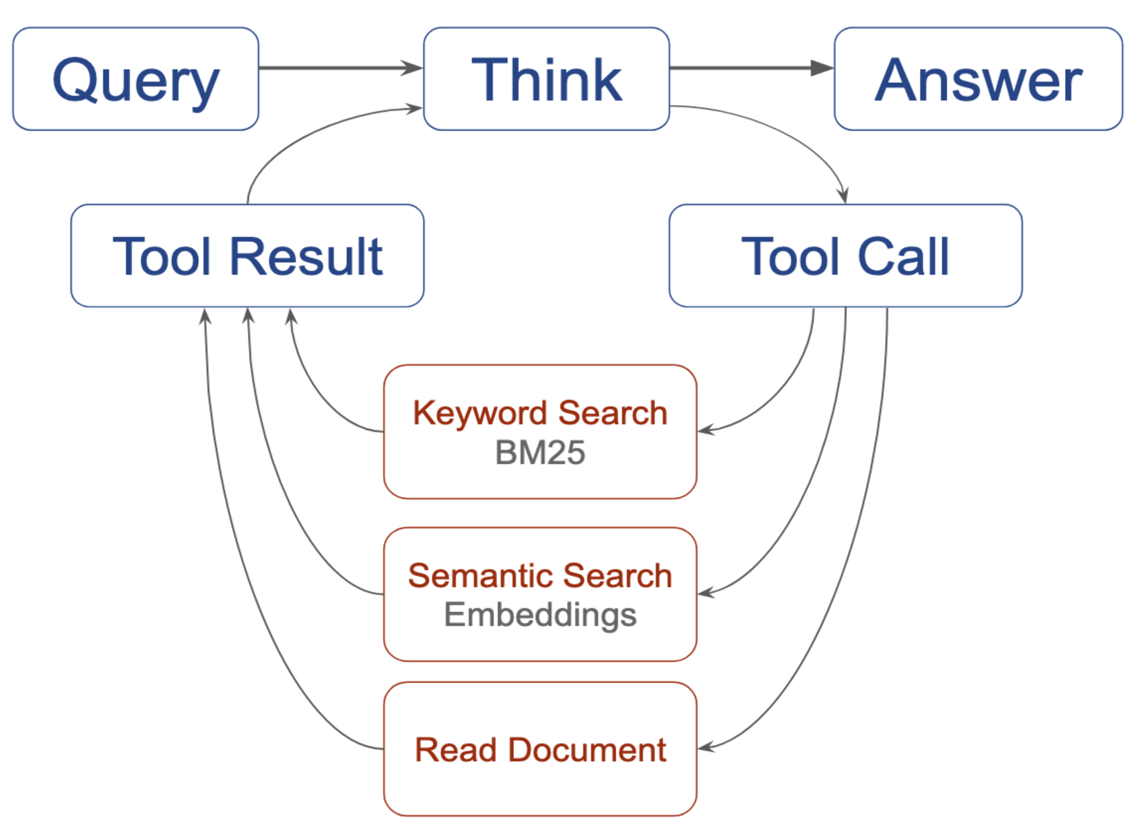
Problem
Prompt-based agents often stall or hallucinate in long-horizons searches

Solution
Treat multi-turn search as a Reinforcement Learning problem. Train a model using verifiable rewards ("did the model find the right document?")

Outcome
Outperform frontier models by training Qwen3-14B: 85% vs 81%

Method and Results

Agent Architecture



Reward

1 to 2	Correct answer
0 to 1	"I don't know"
-1 to 0	Wrong answer
-2 to -1	Formatting errors

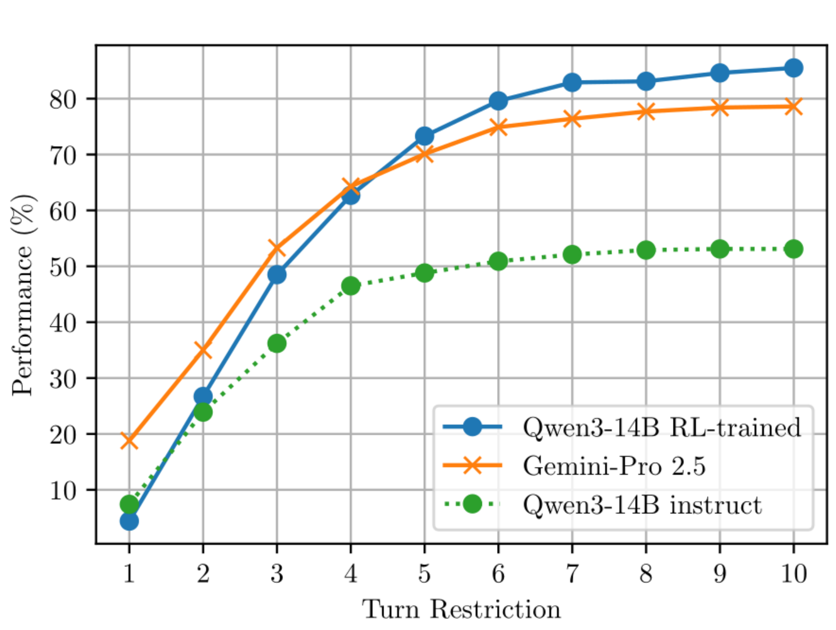
Prefer "I don't know" when unable to find sufficient evidence to hallucination

Model	Accuracy (%)	Avg. Turns
Naïve RAG (Gemini 2.5 Pro)	33	1.0
Qwen3-14B (base)	53	3.7
Gemini 2.5 Flash	66	3.4
Gemini 2.5 Pro	78	5.3
OpenAI o3	81	7.1
Qwen3-14B + RL	**85**	6.2

Multi-Turn Experiments

Turn-restricted inference

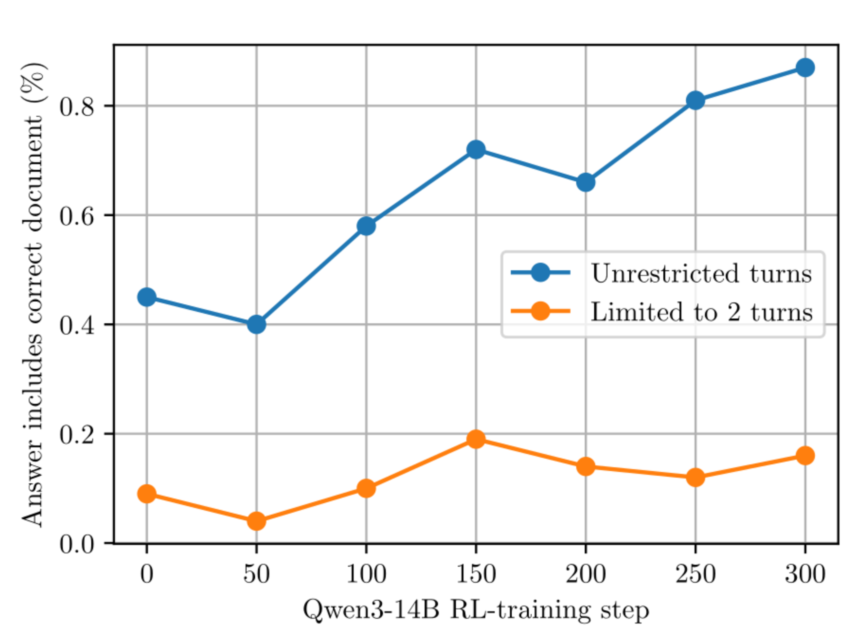
"Does doing more turns improve agent performance?"



All models improve with more turns, but RL-trained Qwen3-14B continues to gain where others plateau

Turn-restricted training

"Is Long-Horizon Training Necessary for Multi-Turn Success?"



Training with ≤ 2 turns prevent the agent from discovering effective long-horizon policies

Discussion

RL turns multiple turns into actual capability: with the same tools and horizon, Qwen3-14B + RL converts additional turns into higher accuracy than both the base model and frontier APIs.

Long-horizon experience is essential: restricting the number of turns during training causes agents to fail on longer-horizon tasks, suggesting that horizon mismatch is a key failure mode for multi-turn agents.

For search, this is a repeatable playbook to create grounded agents

Key References

- "Retrieval-augmented generation for knowledge-intensive NLP tasks" - Lewis *et al* (2020)
- "DeepSeekMath: Pushing the limits of mathematical reasoning in open language models" - Shao *et al* (2024)
- "ART: Agent Reinforcement Trainer" - Hilton *et al* (2025) https://github.com/openpipe/art

Contact

vivekkalyansk
mdda123