Galaxy Tool Recommendation Agent Benchmark

Benchmark Dataset Overview

December 10, 2025

What is this benchmark?

- Curated question–tool pairs from Galaxy Training Network tutorials.
- Focus: tool recommendation for common genomics workflows.
- Data source: GTN tutorials (e.g., intro, quality control, statistics).
- Goal: evaluate LLMs/agents on picking the right Galaxy tool given a query and dataset context.

Strategy

- Early attempt: feed full tutorial content to an LLM and ask it to write queries plus collect datasets/tools $\rightarrow$ poor relevance and coverage.
- Revised: provide datasets, tool IDs, and tutorial context; only ask the LLM to generate queries matching those assets (see commit 7c47dcb "Strategy changed").
- Result: cleaner questions aligned to the provided tools/datasets with fewer hallucinated resources.

Repository contents

- `examples/v0_items.jsonl`: machine-readable benchmark items.
- `examples/v0_items_readable.md`: human-friendly view of questions.
- `examples/datasets/...`: referenced sample datasets (FASTQ, BAM, TSV).
- `gtn_benchmark/query_generator.py`: generation logic for questions.
- `scripts/export_readable.py`: renders JSONL to Markdown.

Item structure

Each JSONL record contains:

- **id**: stable question ID (e.g., `quality-control-q01`).
- **query**: natural-language question.
- **tools**: Galaxy tool IDs with version.
- **metadata**: topic, tutorial, datasets, workflow name.
- **context**: tutorial-level info (e.g., topic path).

Evaluation ideas

- Accuracy: exact match on recommended tool ID/version.
- Top-k: credit if the correct tool appears in top predictions.
- Justification quality: score clarity of reasoning and dataset use.
- Robustness: perturb question wording and check stability.

Usage

- Mode 1: use query + dataset names to recommend Galaxy tools (text-only).
- Mode 2: give agents query + dataset paths; they load data, then recommend or run Galaxy tools.
- Render a readable brief: `python3 scripts/export_readable.py --input examples/v0_items.jsonl --output examples/v0_items_readable.md`

Next steps

- Improve query quality
  - Manual checks found some LLM items misaligned with GTN intent (e.g., using `fastq_quality_filter` for adapter trimming, or MultiQC for single-sample QC instead of aggregation).
  - Generate more items with commercial LLMs (e.g., GPT-4o) in addition to LLaMA 3.1 4-Scout.
  - Add an evaluator agent to rate queries.
- Expand coverage: add more GTN tutorials (e.g., RNA-seq, variant calling).
- Add automated eval harness for LLM tool selection.
- Publish scores and baselines for reproducibility.

Questions?

`examples/v0_items_readable.md` for full content