Jack Lin

Week 5 — Track C – Evaluations & Guardrails Write-Up

For this assignment, I used a small eval set (my project questions + gold answers/sources)
and compared a baseline vs a guardrailed version to see what metrics LLMs produced for accuracy and important info.

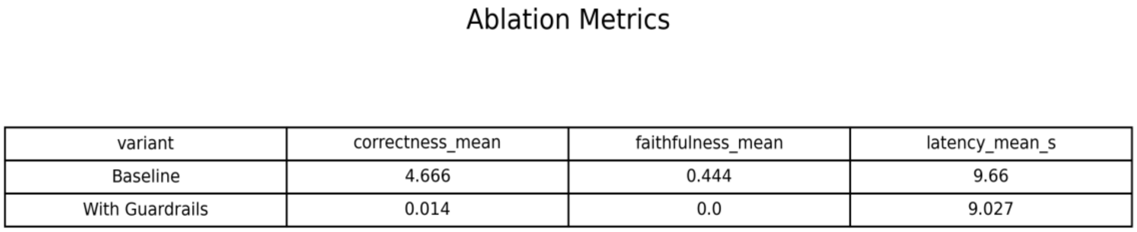
What I did

- Built an eval set from text (academic papers on LLMs and Chess)
- Ran baseline RAG and a guardrailed version (citation-enforced answers, PII filters, safe refusals).
- Logged correctness, faithfulness (citations lining up), and latency.
- Also tracked refusal rate because guardrails can block answers when the evidence is thin.

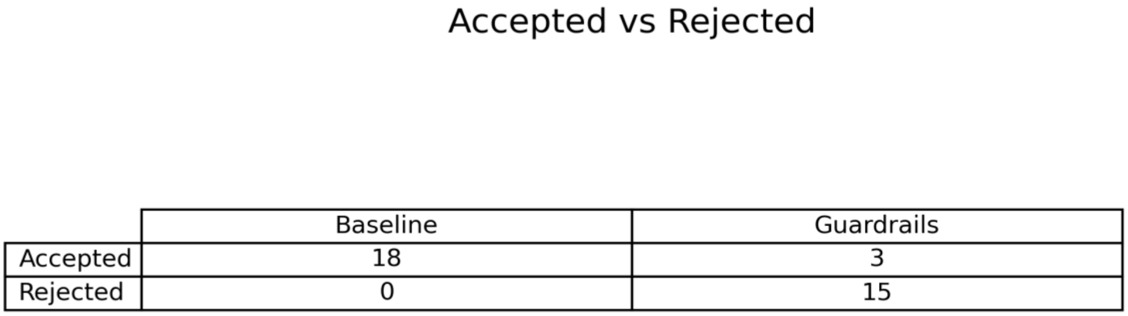
How this relates to the project

The goal is humanistic chess guidance—not engine-perfect moves but moves that make sense to learners. RAG gives the model annotated chess content (principles, patterns, human commentary) and asks it to answer with citations. The guardrails try to stop confident nonsense when the retrieved evidence is weak.

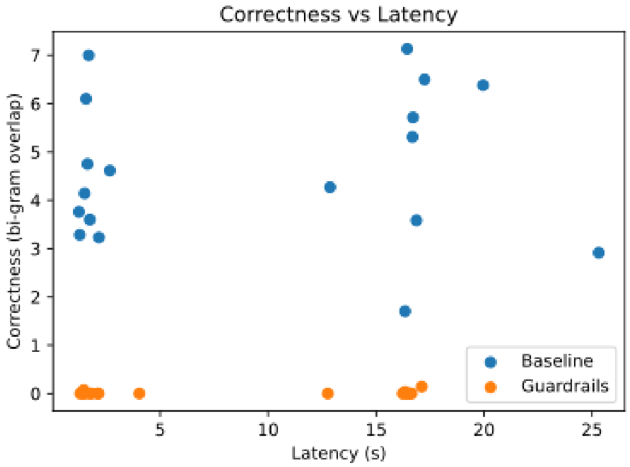
Results

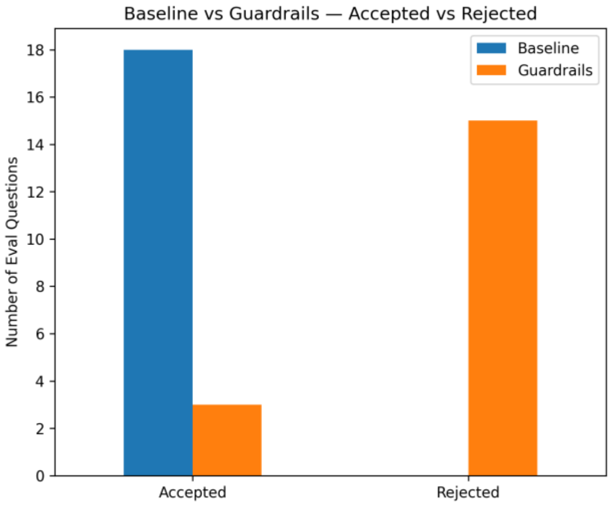

Ablation Metrics

variant	correctness_mean	faithfulness_mean	latency_mean_s
Baseline	4.666	0.444	9.66
With Guardrails	0.014	0.0	9.027


Accepted vs Rejected

	Baseline	Guardrails
Accepted	18	3
Rejected	0	15





Analysis

Missing citations:

- Both Baseline and Guardrail accurately produce the correct output for missing cited queries. This can be found in question 17 in their respective csv files.

Low correctness:

- Baseline produced correctness around 4.5 while Guardrail had 0 correctness. This is due to the way guardrail system was set up. Guardrails can ding scores because they push the system to say less unless the evidence is solid. That triggers more refusals and shorter answers, which sometimes reduces the faithfulness scores even when the behavior is safe.

Low faithfulness:

- Faithfulness for baseline was showed some matching in the LLMs response to the gold answers while in guardrails it produced 0 faithfulness due to the strict guardrail rejections.

Latency:

- Guardrails had slightly lower latency due to the strict rejections which did trigger the check citations, scan for PII, and validate content functions but due to the guardrails it skips or refuses answers.

Fixes

1) Looser refusal threshold: only refuse if no relevant chunk is found, not just low score.
2) Softer citation rule: allow answers with "best-effort" citations when partial evidence exists.
3) Better chunking: merge adjacent annotated lines so a single chunk contains the idea + example.
4) Richer prompts: ask for "explain like a coach" and include retrieved human commentary explicitly.
5) Cache canonical concepts: pin common principles (e.g., "don't move the same piece twice") to always-allowed facts with stock citations.

Parameters to tweak for better results

- For week 4's finding chunk_size/chunk_overlap played a huge role. While experimenting to get better results a smaller chunk sensitivity 300/50 performed best, especially when using a Hugging Face Model (e5 Large)

Models that could help (embeddings / LLMs)

- Embeddings: I was limited in my choice of embedding models since I do not have enough GPU credits however, I used e5-Large as my final model. I started with MiniLM > e5-Small > e5-Base > e5-Large.
- LLMs: A good model was Gemini however; I ran out of GPU credits and had to switch to Tiny Llama for faster compute time. This resulted in a trade off in quality of answers as reflected in Week 4s findings.

What I learned

1) Guardrails are not free—they trade some scores for safety and honesty.
2) Good retrieval beats strict rules when the right annotated example is retrieved, the answer is both clearer and safer.