Question: Which system performed better, on what metrics, and why?

Answer:
The expectations that were set for us was that reranking would improve results, overall.
My follow up research suggests that is the correct expectation

The two tables below show the 'expected' results of reranking, and then the actual comparison between the two runs.

This was my first reaction:
The reranking results are not better, and maybe worse (initially I didn't realize that the noise-sensitivity is scaled as lower=better).
Can's quick response during breakout was: this example is using a very small dataset, so that might impact the results.

Here is my further reflection:

Factual Correctness and Noise Sensitivity: both showed improvement.
This is good, and expected, indicating that the reranker provided more relevant context.

Answer Relevancy: stayed essentially the same. This suggests that the original context was "relevant enough", even before re-ranking.

Faithfulness decreased, but only slightly. Not significant.

Context recall and Context Entity Recall both dropped. A little anyway. Not what I expected. ChatGPT rationalized it like this:

"Reranking changes the composition of those chunks — typically favoring high relevance over broad coverage.

Reranking selected chunks that were more topically aligned, but possibly missed peripheral details that the baseline retriever included"

This is a plausible explanation, but if I wanted to explore further, I might try rerunning with a larger test set.
Because, "once doesn't make a trend"

Q: Should I run the eval again one or more times?
A: Yes, but I might get similar results, because the dataset is "too small"

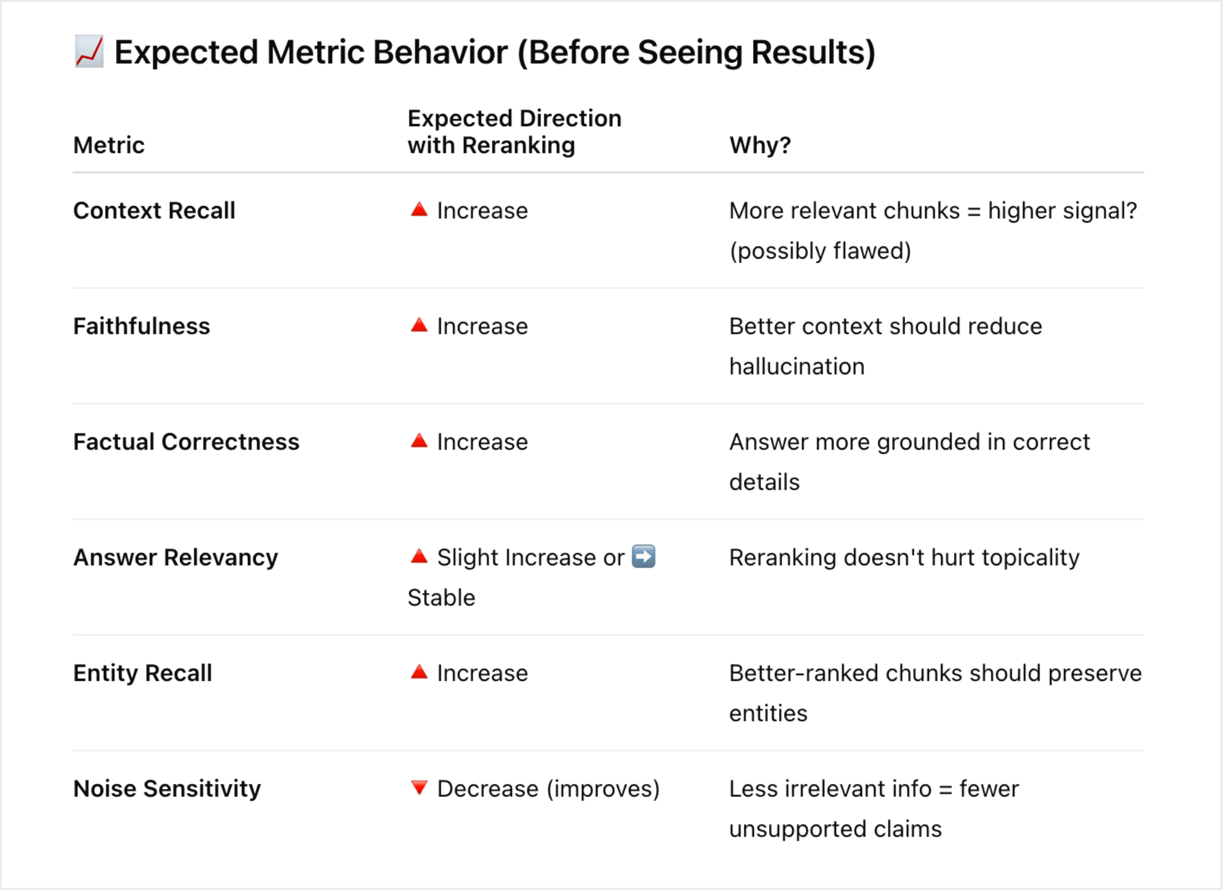
Q: what exactly is it about our data that is "too small"

A: we only have 10 question-answer-referene triplets

Q: how to make it bigger
A: change the testset_size parameter here:
dataset = generator.generate_with_langchain_docs(docs[:20], testset_size=10)

Q: why does that impact the results?
A: because with only 10 data points, one anomaly will have outsized impact

So, if I really want to get a better comparison, I would need to increase the testset_size (to maybe 20 or 25 probably) and run it that way.



📈 Expected Metric Behavior (Before Seeing Results)

Metric	Expected Direction with Reranking	Why?
Context Recall	🔺 Increase	More relevant chunks = higher signal? (possibly flawed)
Faithfulness	🔺 Increase	Better context should reduce hallucination
Factual Correctness	🔺 Increase	Answer more grounded in correct details
Answer Relevancy	🔺 Slight Increase or ➡️ Stable	Reranking doesn't hurt topicality
Entity Recall	🔺 Increase	Better-ranked chunks should preserve entities
Noise Sensitivity	🔻 Decrease (improves)	Less irrelevant info = fewer unsupported claims

📊 RAGAS Metric Comparison — With vs Without Reranking

Metric	No Rerank	With Rerank	Absolute Change	% Change
Context Recall	0.868	0.724	−0.145	−16.6%
Faithfulness	0.891	0.879	−0.012	−1.3%
Factual Correctness	0.628	0.643	+0.014	+2.3%
Answer Relevancy	0.956	0.954	−0.002	−0.2%
Context Entity Recall	0.415	0.403	−0.012	−2.9%
Noise Sensitivity	0.181	0.170	−0.011	−6.1%