

MAGIK AI Assistant: Measured Evaluation

Release v1.0.4 | production deployment | measured 27 October 2026

Multimodal agentic RAG system over finance documents (PDF, spreadsheet, Word, text, image, audio, video). All numbers below were measured on the live production instance. Where a number rests on a small sample or a partial result, the table says so.

Test setup

Item	Detail
Serving hardware	AWS g6e.xlarge, one NVIDIA L40S GPU (46 GB VRAM), 4 vCPU, 32 GB RAM
Answer model	Qwen2.5-14B-Instruct (Q4_K_M) served by llama.cpp, fully on the GPU
Evaluation judge	Qwen2.5-7B-Instruct (Q4_K_M), local, no hosted model APIs
Corpus	7 finance documents, 1,236 indexed chunks for the evaluation tenant
Isolation	Evaluation ran in its own memory-capped container; the live app was not disturbed

1. Real-time latency (streaming endpoint)

Measure	p50	p95	Mean	Samples
Time to first token	11.55 s	23.70 s	13.16 s	46
Full answer time	12.61 s	24.69 s	14.16 s	46

Method. 48 real questions from the project's gold set were sent one at a time to `/rag/query/stream`, the endpoint the web UI uses, after two untimed warm-up requests. The answer cache was off, each question was scoped to its source document, and each ran in its own chat session. Time to first token is the time from sending the request to the first streamed answer token. The server verifies an answer before it streams, so this figure includes retrieval, generation and verification. 46 of 48 requests produced an answer (2 refused, 0 errors); only answered requests are timed. Percentiles use linear interpolation. Single user, warm models, no concurrent load.

Modality	Questions	TTFT p50	TTFT p95
text	8	9.05 s	19.41 s
pdf	8	12.50 s	31.81 s
xlsx	6	15.19 s	18.34 s
docx	8	11.52 s	14.26 s
image	8	8.40 s	12.18 s
audio	4	11.79 s	15.11 s
video	4	18.75 s	23.28 s

Per-modality p95 values rest on very few samples and are indicative only.

2. Answer quality (RAGAS, 30 questions, local judge)

The run uses the first 30 answerable gold-set questions in fixed order: 14 on a text file, 14 on the Apple 10-K PDF and 2 on a Word report. Spreadsheet, image, audio and video questions were not part of this run.

Metric	Result	Coverage	How to read it
Context recall	0.93	30 questions	Share of reference facts found in retrieved context
Citation accuracy	1.00	16 answers with citations	Citations point to retrieved sources (heuristic check)
Template leak rate	0.00	30 questions	No prompt text leaked into any answer
Finance number fidelity	0.90	30 questions	Figures in answers match source text verbatim
Answer relevancy	0.73	see note below	Question is addressed by the answer
Faithfulness	0.84	11 of 30 gradeable	Statements supported by context; 19 rows the judge could not grade are excluded
Hallucination flag rate	0.33	30 questions	Heuristic flag, 10 of 30 answers flagged; under review

3. Answer quality (DeepEval, 10 questions, local judge)

DeepEval run of 27 October 2026 on the first 10 gold-set questions, all on the text file. Scores are the raw values the library returned. **Three of the six metrics are lower-is-better**, so a high number is a bad result for those.

Metric	Raw score	Better is	Rows scored	Reading
Answer relevancy	0.40	higher	10 of 10	Low; answers only partly address the question as judged
Faithfulness	0.59	higher	7 of 10	3 rows failed on judge output format
Contextual recall	1.00	higher	9 of 10	1 row failed on judge output format
Hallucination	0.87	lower	10 of 10	High value as returned; not consistent with the RAGAS result, see section 4
Bias	1.00	lower	9 of 10	Not credible: judge output did not match the expected format
Toxicity	1.00	lower	10 of 10	Not credible: judge output did not match the expected format

4. What these numbers do not show

- **Judge limits.** Scores come from a local 7-billion-parameter model, not human review. It sometimes returns output the scoring library cannot parse, which drops that question from a metric. The report does not yet record how many questions each RAGAS metric actually scored, so answer relevancy and context recall may rest on fewer than 30.
- **Faithfulness** covers only 11 of 30 answers, so treat it as a partial figure.
- **Hallucination flag rate** is a rule-based check, not a verified error rate. A five-question smoke run flagged none and this run flagged one third, and the flagged answers have not been reviewed yet.
- **DeepEval reliability.** With a local 7-billion-parameter judge, DeepEval's bias, toxicity and hallucination scores are not credible. A toxicity of 1.00 would mean every answer was toxic, and the judge's replies did not match the format the library expects (missing fields, invalid JSON). Treat those three as unresolved, and the other three as indicative. The RAGAS and DeepEval relevancy and faithfulness figures also differ. They use different question subsets (30 against 10) and different judge prompts, so they are not directly comparable.
- **Long sessions.** An earlier latency run put all 48 questions in one chat session, and 17 of them were refused. With a separate session per question only 2 were refused (both on the spreadsheet). Asking many questions about different documents in one long session is a known weak spot that is being investigated.
- **Sample size.** 30 quality questions and 48 latency questions are a first production measurement, not a benchmark.

Reproduce

Quality: `deploy/aws/scripts/run_quality_report.sh ragas 30` and `... deepeval 10` in the project repository (release v1.0.4). Latency: the streaming benchmark described in section 1. Raw reports are kept on the instance under `/opt/magik/quality-reports/`.