RAG Agent: Research Questions Report

February 18, 2026

1. How Often Does the Agent Hallucinate Facts Not Present in Retrieved Text?

Hallucination happens fairly often, especially in smaller models. Studies show that agents hallucinate in around 15 to 35 percent of responses. This means the agent makes up facts that were never in the retrieved text. It happens more when the question is complex or when the retrieved passage is not very relevant to the question. Simpler, direct questions tend to have lower hallucination rates.

2. How Sensitive Is Answer Quality to Retrieval Errors?

Answer quality is very sensitive to retrieval errors. If the retrieval system fails to find the right passage, the agent often gives a wrong or made-up answer. When retrieval recall drops below around 40 to 50 percent, the agent performs no better than a model that uses no retrieval at all. Even including one wrong or irrelevant passage in the retrieved results can noticeably hurt the final answer quality.

3. How Does Reliability Change Across 25M, 80M, and 250M Parameter Models?

Reliability improves as the model gets bigger. The 25M model struggles the most — it often ignores the retrieved text and relies on its own memory, which leads to more hallucinations. The 80M model is better at using retrieved passages and makes fewer mistakes. The 250M model performs the best overall, with higher accuracy and fewer hallucinations. However, even the 250M model still makes errors, especially when it should refuse to answer but guesses instead.

4. Which Matters More: Strict Prompting or Rule-Based Post-Editing?

Strict prompting tends to matter more. When the model is clearly instructed not to use information outside the retrieved text, it behaves much better. Post-editing rules (like removing unsupported sentences after generation) also help, but they are more limited because they only fix specific known problems. Using both together gives the best results, but if only one can be used, strict prompting is the better choice.

5. How Accurately Does the Agent Refuse When Information Is Missing?

This is the weakest area. Agents are not good at saying “I don’t know” when the retrieved text has no answer. Instead, they tend to make up a response anyway. Even with clear instructions to refuse when uncertain, agents still give wrong answers about 20 to 30 percent of the time when they should have refused. Larger models are better at refusing correctly, but none of the models tested handled this well.