Exercise 1:

Congressional Record:

- "What did Mr. Flood have to say about Mayor David Black in Congress on January 13, 2026?" (See CR Jan 13, 2026)
    - Manually Found Correct Answer: Mr. FLOOD. Mr. Speaker, I rise today to recognize Papillion, Nebraska’s, Mayor David Black, as a paragon of public service and remarkable stewardship. Through more than two decades of service to the people of Papillion,Mayor Black helped to guide the city into the thriving economic hub that it is today. This month, he announced that he will not seek reelection, which will conclude 17 1⁄2 years as mayor and nearly 5 additional years of serving the city in other roles. Mayor Black leaves behind a strong foundation built on partnership with an exceptional professional staff and the business community. Decades of engagement and community outreach have produced clear results, attracting firms like Google and Facebook to expand to Papillion. b 1015 He will be deeply missed by the people of Papillion, and I am hopeful he will continue public service down the road. Mayor Black has more than earned the opportunity to spend additional time with his family once his final term in office concludes. Papillion, Mr. Speaker, is a special place, with Mayor Black’s legacy being built yet another example of that truth. I thank Mayor Black for his outstanding public service.
    - WITHOUT RAG: The answer is wrong. The model hallucinated the fact that Mr. Flood stated that Mayor David Black was not an effective leader, failed to deliver the promised improvements for the city's residents, was criticized for his handling of public safety issues and corruption allegations, and that he called for new leadership.
    - WITH RAG: Correct!
- What mistake Elise Stefanovic make in Congress on January 23, 2026?" (See CR Jan 23, 2026)
    - Manually Found Correct Answer: I mistakenly voted No when I intended to vote Yes.
    - WITHOUT RAG: The answer is wrong! The model hallucinated about her opinions of the Infrastructure Investment and Jobs Act (IIJA).
    - WITH RAG: Again, the answer is wrong. The model appears to use the wrong context and determined that she was trying to overturn the democratic process and disregarding lawful orders from Capitol Police.
- "What is the purpose of the Main Street Parity Act?" (See CR Jan 20, 2026)
    - Manually Found Correct Answers: Amend the Small Business Investment Act of 1958 to modify the criteria for loans for plant acquisition, construction, conversion or expansion, and for other purposes. H.R. 5763 reduces the minimum equity requirement for special purpose properties to at least 10 percent, in line with the requirement for standard 504/CDC loans. Most importantly, it maintains discretionary authority to require higher equity if deemed necessary. H.R. 5763 allows small businesses on Main Street to have equal access to SBA 504 loans.

        - WITHOUT RAG: The model got the general idea: financial help for small businesses. However, it did hallucinate dates and details (signed into law on December 18, 2017 by President Donald Trump).
        - WITH RAG: Correct!
    - "Who in Congress has spoken for and against funding of pregnancy centers?" (See CR Jan 21, 2026)
        - Manually Found Correct Answers: For: Ms. Dexter, Mr. Bost, Mr. Crank, Mr. Smith, Mrs. Fischbach Against: Mr. Leger Fernandez
        - WITHOUT RAG: The model did not name any specific members.
        - WITH RAG: The answer was wrong. It mentions Dingell, whose name is in that CR, but did not speak on the Act. It also mentions John Hickenlooper, whose name is not in that CR. However, the context appears to be relevant, as all the retrieved chunks are from the correct date.

Model T Ford:
Note: It was difficult for us to verify these queries in the same manner we were for the Congressional Records.

- "How do I adjust the carburetor on a Model T?"
    - WITHOUT RAG: The response was generally relevant and aligned with common mechanical knowledge. While plausible, the answer was not grounded in the specific Model T manual. I was not able to verify it against the source document.
    - WITH RAG: The response incorporated context retrieved directly from the Model T service manual. The explanation was grounded in the original text.
- "What is the correct spark plug gap for a Model T Ford?"
    - Manually Found Potential Correct Answer: Before replacing the plug, check the spark plug points for gap, the gap between the points should measure approximately ~ 1/32"·
    - WITHOUT RAG: The model did hallucinate specific values (0.5mm (20 thousandths of an inch)).
    - WITH RAG: It still hallucinated (0.006 inches to 0.010 inches). The retrieved context appeared to reference the gap between piston rings rather than spark plug points.
- "How do I fix a slipping transmission band?"
    - Manually Found Potential Correct Context: Replacing Transmission Bands (Chapter 12, page 131/149).
    - WITHOUT RAG: The answer seems relevant and the model's general knowledge could be correct.
    - WITH RAG: The model did retrieve the correct context.
- "What oil should I use in a Model T engine?"
    - WITHOUT RAG: The model hallucinated 10W-30 or 15W-40 motor oils, as neither could be found in the manual.
    - WITH RAG: The response referenced "new oil," which technically does not introduce unsupported specifications.

Exercise 2:

Does GPT-4o Mini do a better job than Qwen 2.5 1.5B in avoiding hallucinations?

- Not necessarily. It demonstrated an inaccurate understanding of the Main Street Parity Act, but it did use good general knowledge of republicans and democrats to answer about the support of pregnancy center funding. For the first two of the January 2026 congressional queries, GPT-4o Mini does explicitly state that it does not have access to events after its training cutoff and declines to answer, which is a strong non-hallucination behavior. In contrast, Qwen 2.5 1.5B without RAG had previously fabricated detailed statements about Mr. Flood and Elise Stefanovic. However, on the Model T questions, GPT-4o Mini still provides confident, specific values (e.g., a 0.025-inch spark plug gap and SAE 30 oil) without verifying them against the manual, meaning it can still hallucinate precise specifications.

Which questions does GPT-4o Mini answer correctly? Compare the cutoff date and corpus age.

- GPT-4o Mini correctly handles the Congressional Record (January 2026) questions in the sense that it refuses to fabricate specific details and acknowledges its knowledge cutoff (October 2023). This is appropriate because January 2026 events occur after its training data cutoff. The Model T corpus is related to vehicles produced far before GPT-4o Mini's 2023 cutoff. Given that the Model T information is historical and likely widely documented, the model can plausibly rely on pretraining knowledge and provide generally reasonable answers. That said, while its mechanical explanations are often plausible, the exact numeric specifications (spark plug gap, oil type) are not verified against the manual and may still be inaccurate. I could not manually determine the exact correct answers for those queries.

Exercise 3:

Where does the frontier model's general knowledge succeed?

- The frontier model succeeds on broad, widely documented mechanical knowledge about the Model T. It provides coherent, detailed explanations of carburetor adjustment, transmission band tightening, spark plug gap specifications (0.025"), and historically common oil recommendations such as SAE 30 non-detergent oil. I could not find that oil in the manual, but the answers might still be technically plausible even without direct access to the manual.

When did the frontier model appear to be using live web search to help answer your questions?

- The frontier model clearly appears to rely on live or up-to-date web information when answering the January 2026 Congressional Record questions. It provides specific dates, bill numbers (H.R. 6945), named representatives, and contextual legislative descriptions that extend beyond static historical knowledge. The specificity of these 2026 references suggests real-time web retrieval rather than purely pretraining-based recall. However, it still could not find some important information (claiming that there is no widely referenced federal bill officially called the Main Street Parity Act in the January 2026 congressional session).

Where does your RAG system provide more accurate, specific answers?

- The RAG system seemed to provide more reliable grounding when answering document-specific Congressional Record questions tied to exact dates (e.g., January 13 or January 23, 2026).

What does this tell you about when RAG adds value vs. when a powerful model suffices?

- These results show that powerful frontier models might suffice when questions concern widely documented historical or technical knowledge. However, RAG adds significant value when questions require exact document grounding, especially for recent or corpus-specific material (e.g., the January 2026 congressional proceedings). Frontier models may rely on live search and provide plausible summaries, but RAG guarantees alignment with the specific documents you are analyzing and RAG outputs can be verified much more easily.

Exercise 4:

At what point does adding more context stop helping?

- After k-=5, it seems like the answer does not meaningfully improve in clarity or correctness; instead, it begins to repeat or expand with more trivial procedural details (especially for the query about carburetors and slipping transmission bands). This suggests that additional context stops helping once the most relevant chunk(s) are already retrieved, which typically might be around k=1–5.

When does too much context hurt (irrelevant information, confusion)?

- We see too much context beginning to hurt at k ≥ 10 with the specific question of the spark plug gap. At k=10 and k=20, the model starts mixing spark plug gap with unrelated magneto shim or ring gap information, producing incorrect numerical ranges (e.g., 0.006–0.010 inches). So, the additional retrieved chunks seem to dilute relevance and prompt the model to synthesize mechanical procedures that are distantly related.

How does k interact with chunk size?

- The interaction between k and chunk size needs to be balanced, as they determine the total context volume. If the chunk sizes are larger, there is more context in each chunk, and you might need a smaller k. With smaller chunk sizes, you might have a larger k, but information might span and be broken up across chunks.

Exercise 5:

Does the model admit it doesn't know?

- When asked “What’s the horsepower of a 1925 Model T?” without RAG, it confidently provides a specific horsepower value (20 hp), even though this is not supported by the manual. However, with RAG, the model does admit it does not know. For the synthetic oil question, it also admits with RAG that the context does not provide any specific reason why the manual recommends synthetic oil. For the question, “which section says to use 5W-30 full synthetic oil?” (no section mentions this oil), without RAG the model still answers, but with RAG it does say that the context does not mention it.

Does it hallucinate plausible-sounding but wrong answers?

- Yes, without RAG. For the horsepower question, it invents a detailed engine explanation and numeric value (20 hp), which is not drawn from the corpus. For the synthetic oil and 5W-30 questions, it gives modern oil justifications that are entirely unrelated to the Model T manual. Nevertheless, with RAG, the model sometimes produces answers that are not supported by the retrieved context. For the query, “What is the capital of France?”, it retrieves unrelated Model T manual pages but still answers “Paris,” which implies that the answer is grounded in the context when it is actually completely off-topic. These responses are plausible, and maybe even correct, but not evidence-based.

Does retrieved context help or hurt? (Does irrelevant context encourage hallucination?)

- It does initially seem like irrelevant context can actually strengthen hallucination, but for this specific experiment, the RAG model did already admit in some form that the context lacked the answers to the queries (besides the capital of France). Thus, the context did not seem to interfere with or worsen hallucinations.

Experiment: Modify your prompt template to add “If the context doesn't contain the answer, say ‘I cannot answer this from the available documents.’” Does this help?

- It reinforces more firmly what the RAG model had determined before. After adding the explicit refusal instruction, the model consistently declined to answer questions when the retrieved context does not contain the relevant information (for all queries except the capital of France). This forced the model to align its responses strictly with the retrieved evidence/admit uncertainty, making it more reliable and direct.

Exercise 6:

Which phrasings retrieve the best chunks?

- The best-performing phrasings were the question form (“What gap do I set the spark plugs to?”) and the more specific phrasing (“Spark plug gap in thousandths of an inch for a Model T”). These produced the highest similarity scores (max ≈ 0.590–0.593, mean ≈ 0.545–0.550). The keyword-style query (“Model T spark plug gap”) also performed strongly (max ≈ 0.586, mean ≈ 0.537). In contrast, the indirect phrasing (“Ignition system spark plug spacing requirement”) had noticeably lower similarity scores (mean ≈ 0.464). This suggests that phrasing that closely matches the terminology used in the manual (“spark plug gap”) retrieves more relevant chunks than abstract or indirect wording.

Do keyword-style queries work better or worse than natural questions?

- Keyword-style queries performed comparably to and even sometimes better than natural formal phrasing. The keyword query had higher average similarity scores than the formal phrasing (0.537 vs. 0.517 mean). Additionally, overlap analysis shows strong overlap between casual, keyword, and more specific queries (up to 4/5 shared chunks), indicating they retrieve very similar top results. However, the purely indirect phrasing did reduce retrieval quality.

What does this tell you about potential query rewriting strategies?

- These results tell us that effective query rewriting strategies should consider prioritizing domain-specific keywords that closely match the language of the source documents. The overlap patterns show that small changes in wording can significantly alter the retrieved top-5 set, so retrieval is sensitive to phrasing.

Exercise 7:

Does higher overlap improve retrieval of complete information?

- Yes, but only up to a certain point. 0 overlap results in more fragmented responses and minimal answers. 64 overlap is a noticeable improvement, which helps answers flow better. 128 results in very strong completeness with full lists being cited in the answers. 256 where half of the chunks are overlap makes no substantial improvement for completeness.

What's the cost? (Index size, redundant information in context)

- The overlap and number of chunk pairings are: (0, 2320), (64, 2716), (128, 3277), (256, 5215). 128 overlap increases the index size by about 41% from 0 overlap, but 256 overlap increases index size by about 59% from 128, which is a huge cost. 256 overlap also results in major redundancies in chunk information, and the context window seems wasted because the agent sees repeated sentences.

Is there a point of diminishing returns?

- Yes, 256 has very marginal gains but significant growth in cost, which makes it not a worthwhile decision compared to 128 overlap.

Exercise 8:

How does chunk size affect retrieval precision (relevant vs. irrelevant content)?

- Small chunks result in very precise answers because each chunk is focused on a smaller amount of information with minimal irrelevant content that can be skipped over. Medium chunks are also relatively precise, with more contextual information that may be helpful for a user to know. Large chunks give the least amount of precision because they can include irrelevant procedural references or other noise that is not helpful for the query.

How does it affect answer completeness?

- Small chunks are often incomplete, as it can be difficult to connect related information to the query if the first half of a paragraph is missing. Medium chunks are much more complete as they can preserve multi-step instructions or longer lists. Large chunks are the most complete (but can be a little too complete) because they can retain the full procedure and cross references to other sections, but with a large increase in verbosity (over-answering).

Is there a sweet spot for your corpus?

- The sweet spot seems to be a size of 512 as it preserves structured lists or steps without splitting into separate chunks while avoiding excessive context or irrelevant information.

Does optimal size depend on the type of question?

- Optimal size can change based on the type of question. For example, questions that would only be answered by one specific paragraph or list in the entire document would benefit from smaller chunks. I do not believe there are questions that would benefit from large chunks, as it introduces too much noise.

Exercise 9:

When is there a clear "winner" (large gap between #1 and #2)?

- This happens when asking about noisy time gears and steering gear ratios, which are questions that are drawn primarily from only one chunk. Other chunks may mention these topics but are more loosely related than the first chunk, which gives the direct answer

When are scores tightly clustered (ambiguous)?

- Ambiguous queries, such as why the engine isn't starting, or how to fix the transmission band, means that multiple chunks are relevant to this query. These queries are often more broad, where an answer would have to be pulled from multiple areas in the text, or more open ended questions where there isn't strictly one correct answer

What score threshold would you use to filter out irrelevant results?

- High correlation scores are often above 0.65, relevant chunks still maintain a score above 0.55, and weak or noisy results have scores below 0.5.

How does score distribution correlate with answer quality?

- A high mean combined with a large gap between the top two chunks indicates a specific, high-quality answer. High means but tightly clustered scores are more ambiguous in quality because the topic is covered in many places, which could be a good thing if all chunks are in alignment with each other but could also be confusing for the agent if some chunks are contradicting or relay conflicting information. Low means result in low-quality answers where the agent doesn't have much relevant ground truth information to base the response off of, so it cannot accurately address the question.

Experiment: Implement a score threshold (e.g., only include chunks with score > 0.5). How does this affect results?

- The queries with strong scores (either no chunks removed or only one chunk removed) remained basically the same, as the answers were already high quality. Almost all other queries saw some kind of improvement, either with less hallucinations or by becoming more direct with less loosely related information and an answer more focused on the topic of the question. However, for the question about which oil the engine requires, none of the chunks passed the threshold and so the answer deteriorated because the agent had no context at all.

Exercise 10:

Which prompt produces the most accurate answers?

- Most accurate tends to be strict grounding or encouraging citation because it forces the agent to only use the text provided. Minimal or permissive prompts can be less accurate from hallucinations. The agent struggled to strictly follow the structured output prompt, but if enforced properly this results in accurate answers too.

Which produces the most useful answers?

- Permissive and encouraging citation prompts tend to give the most useful answers because they are rich and actionable. Minimal prompts sometimes are more speculative, and strict prompts can result in incomplete answers when it tells the user that there was not enough context. Structure output prompts can be good for more technical workflows or systematic problems given that the agent is able to follow the prompt correctly, but can be redundant.

Is there a trade-off between strict grounding and helpfulness?

- Yes, strict grounding will be very accurate and faithful to the provided documents with very low risk of hallucination. More helpful prompts such as the permissive or citation prompts will trade some accuracy or hallucinations for more complete and actionable answers that a user might find easier to understand and more meaningful or personable. Structured output can be somewhat unreliable due to how the agent struggles with following exact output instructions, but at its best is a good balance between accuracy and helpfulness.

Exercise 11:

Does retrieving more chunks improve synthesis?

- Not always, increasing the top_k does allow the answers to include more details with more specific tools or lists but it also increases irrelevant content or misattributed steps. Generally it improves coverage but can harm precision.

Can the model successfully combine information from multiple chunks?

- Yes, particularly in the question about front vs rear spring differences it is able to correctly point out several differences in the steps, proving that it is able to correctly retrieve and use information across different non-continuous chunks. However, it is not always 100% perfect, because it can make mistakes with frequency of tasks for the monthly maintenance question, since “overhaul” isn’t technically connected to a specific time frame or frequency.

Does it miss information that wasn't retrieved?

- Yes, the answers for multiple top_k values says there were no explicitly safety warnings. The manual does include safety warnings, but there is no explicit safety section, and no chunk that includes the word “safety”. It does include cautions with the wording “carefully inspect” and other synonyms for safety and precaution, which the agent is not able to retrieve.

Does contradictory information in different chunks cause problems?

- Yes, sometimes chunks can use different wordings, and different sections of the document can contain different maintenance intervals in their recommendations. Usually this results in blending errors for the final answer, as sometimes the model can merge these or treat them as a cumulative.