1. **Root Cause Analysis**

   Based on your analysis of logs.json, identify the two most critical problems with the chatbot. State a clear hypothesis for the cause of each problem, connecting it to a specific system component (Ingestion, Retrieval, or Generation) and a specific data source (Wiki, PDFs, or Confluence). Justify your hypotheses with evidence from the logs. You may attach a separate analysis document of the logs.

```
all_sources_down = [source for sources_list in df_negative['sources'] for source in sources_list]
source_counts = Counter(all_sources_down)
print(source_counts)
```

✓ 0.0s

```
Counter({'Archived Design Docs (PDFs)': 25, 'Confluence': 20, 'Engineering Wiki': 19})
```

```
df_positive = df[df['user_feedback'] == 'thumb_up']
```

✓ 0.0s

```
all_sources_up = [source for sources_list in df_positive['sources'] for source in sources_list]
source_counts_up = Counter(all_sources_up)
print(source_counts_up)
```

✓ 0.0s

```
Counter({'Engineering Wiki': 108, 'Confluence': 88, 'Archived Design Docs (PDFs)': 64})
```

	user_query	sources	response_latency_ms
36	What is the plan for migrating from Python 2 t...	[Archived Design Docs (PDFs), Engineering Wiki...	4950
68	What is the company's policy on using personal...	[Archived Design Docs (PDFs), Engineering Wiki...	4900
41	What is our cloud provider's SLA for virtual m...	[Archived Design Docs (PDFs), Engineering Wiki...	4700
44	What is the company's social media policy?	[Archived Design Docs (PDFs), Confluence, Engi...	4400
33	What is the company's stance on remote work?	[Archived Design Docs (PDFs), Confluence, Engi...	4350
78	What is the company's policy on employee educa...	[Archived Design Docs (PDFs), Engineering Wiki...	4250
23	What is the company's travel and expense polic...	[Archived Design Docs (PDFs), Confluence, Arch...	4200
11	What are the performance benchmarks for the ne...	[Archived Design Docs (PDFs), Confluence, Arch...	4150
14	What is the company's policy on open source so...	[Archived Design Docs (PDFs), Engineering Wiki...	3900
20	What version of Java is approved for new backe...	[Engineering Wiki, Archived Design Docs (PDFs)...	3650

Mainly PDFs are mentioned as sources for queries with negative feedbacks.

The first problem: Responses take longer than required
System component: Generation
Data Source: PDFs

Reasons:

- Retrieval itself is relatively fast since vector search is efficient and usually lightweight.

- The bottleneck lies in generation, as the LLM must process large input contexts.
- PDFs contribute large text chunks that are difficult to segment into small, coherent pieces. This results in more tokens being passed to the LLM, which increases both latency and cost.
- PDFs often lack clear structure (e.g., proper headings or clean sections), making chunking and embedding less precise and further inflating token usage.

The second problem: The responses contain outdated and false information
System component: Retrieval
Data source: PDFs and Confluence

Reasons:

- The retrieval process prioritizes vector similarity during vector search, rather than content freshness or accuracy.
- PDFs are static and archived, so they often contain outdated information but may still match user queries closely by keywords or embeddings.
- Confluence content is updated frequently, but it may not be re-ingested or re-indexed promptly, or newer versions may not be prioritized in retrieval. As a result, older or incorrect chunks can be retrieved and used in the final answer.

2. Quantitative Trade-off Analysis

The engineering team proposes two mutually exclusive options to improve relevance:
Option A (Re-ranker): Add a Cohere Re-ranker. This adds a fixed 600ms to latency and a new cost of $1.00 per 1,000 queries.
Option B (Increase Context): Increase retrieval from k=4 to k=10. This adds 250ms to retrieval latency and increases the number of tokens sent to the generator. Choose one option.

First, calculate the estimated monthly generation cost increase for Option B, assuming an average of 400 tokens per chunk and a volume of 100,000 queries per month.

K = 10
Tokens = 400
Total retrieved tokens = 4000 (previously 1600)
We increase the tokens by 4000 - 1600 = 2400
Monthly increased number of tokens = 2400 x 100000 = 240M
Monthly cost difference = 240 x $3 = $720

Then, write a recommendation to your product manager arguing for either Option A or B, justifying your decision by balancing relevance improvement, impact on the latency SLA, and monthly cost.

Option A keeps the number of tokens to 1600, but selects during vector search top 20 matching and using Cohere reranks them considering semantic similarity.
K = 4 (not changed)
1 extra dollar for each 1000 queries
Cost change = 100 x $1 = $100 (less than for b)

Suggestion: Option A
Advantages:

- Significantly lower additional cost Only $100/month extra, compared to $720/month for Option B. This keeps operational expenses under control and makes the solution more scalable as query volumes grow.
- Higher precision and accuracy Instead of simply adding more context (which risks diluting relevance), the re-ranker intelligently selects and prioritizes the most semantically relevant chunks from a larger candidate set (top 20). This directly improves answer quality and reduces the chance of including noisy or irrelevant information.
- Predictable and controlled latency impact Option A adds a fixed 600 ms latency, making it easier to plan for and optimize. In contrast, Option B's stated +250 ms retrieval latency does not account for the additional generation latency caused by the larger context (more tokens), making total latency unpredictable and likely higher.