RAG on Rails

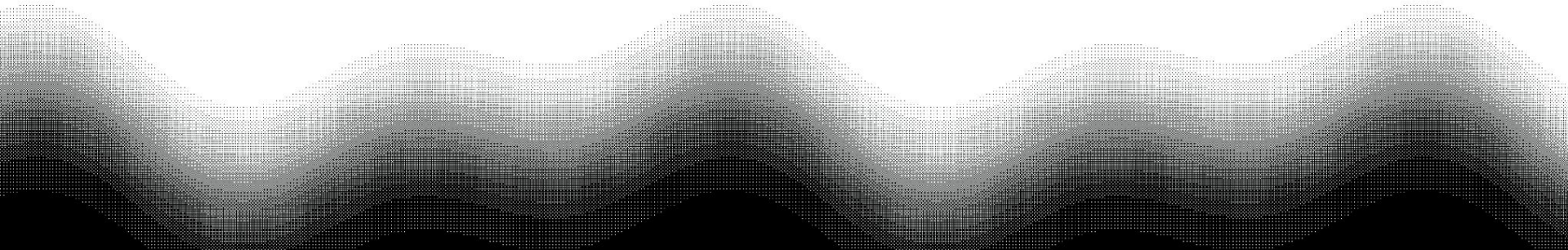
What? Why? How?

Neil Cameron

neilcameron.me

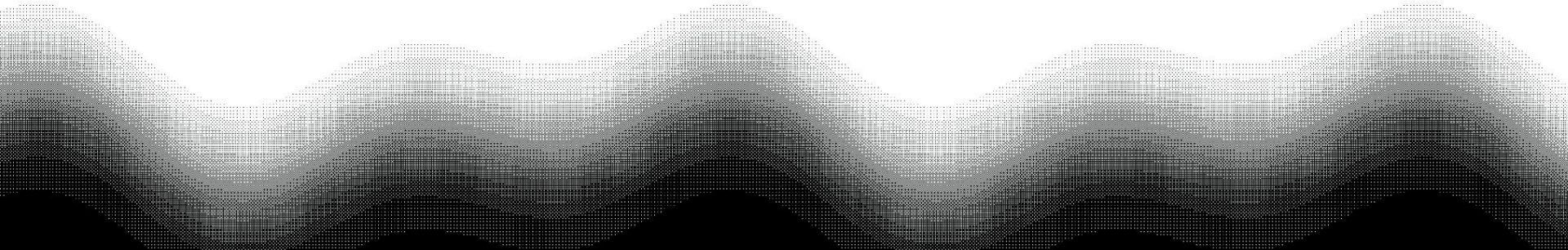
What we'll cover

1. What is RAG?
2. Building retrieval pipeline
3. Constructing the prompt



Hi, I’m Neil

- Building ResponseHub, an AI automation tool for security questionnaire responses
- Previously co-founder and CTO at Progression
- Using Rails since 2016, integrating LLMs since GPT-3 (Nov 2021)



What is RAG?

What is RAG?

- Retrieval Augmented Generation
- Generate text using specific context
- Examples:
  - Support tickets: customer question + help docs + prompt
  - Security questionnaire question: security question + policy docs + prompt
  - Coding assistants: code base + task + prompt

Question

Do you encrypt tenant data at rest (on disk/storage) within your environment?

Explain this

Applicable View rationale

Answer

Yes | No | N/A

The organisation encrypts tenant data at rest within its environment using industry-standard encryption protocols. Full disk encryption is implemented on all user endpoints that store company data.

Regenerate Answer

Sync enabled

Last updated: 03 Mar 2026

Confidence: Very good

References (10)

Sufficient Context

Used in answer

Knowledge Base Item

Relevance: Excellent

Do you encrypt tenant data at rest (on disk/storage) within your environment?

Yes, the organisation encrypts tenant data at rest within its environment. Confidential data, which includes personal identifiable information (PII), financial records of clients, authentication credentials, financial data and forecasts, source code for proprietary software products, client contracts and agreements, and detailed security architecture information, must be encrypted at rest using industry-standard encryption protocols. Additionally, full disk encryption is implemented on all user endpoints (laptops, desktops, mobile devices) that store company data. All devices used to access the organisation's systems and data must have disk encryption enabled as part of the device and endpoint security requirements.

Used in answer

Data Management Policy for Secure Information Handling

Page 3 , Section 5.2 Restricted Data Handling

Relevance: Low

5.1 Confidential Data Handling Confidential data requires the highest level of protection against unauthorized access or disclosure. Access must be strictly limited to individuals with a clear business need. Requirements for confidential data: - Must be encrypted at rest and in transit using industry-standard encryption protocols- Access requires multi-factor authentication- Must not be stored on personal devices under any circumstances- Must not be shared via email without encryption- Physical documents must be stored in locked cabinets when not in use- Must not be discussed in public places or over unsecured communication channels- When working remotely, confidential data must only be accessed through secure VPN connections- Special handling procedures apply for finance sector clients' data, including enhanced logging and access controls- Must be backed up according to the schedule in Appendix B

Used in answer

Knowledge Base Item

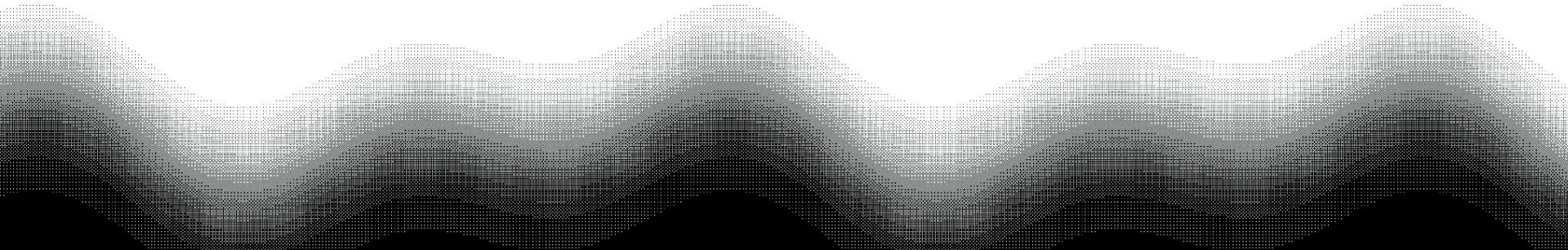
Relevance: Low

Is full disk encryption implemented on all user endpoints (laptops, desktops, mobile devices) that store company data?

Full disk encryption is implemented on all user endpoints that store company data.

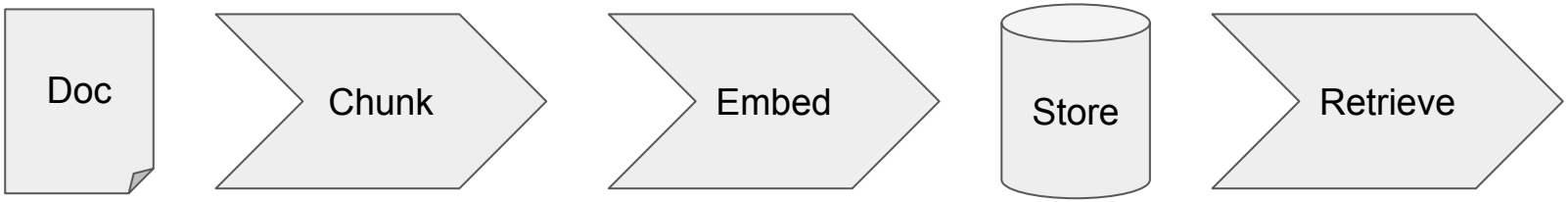
Why use RAG?

- Large context
- Output needs to be true
- Output requires citations to source
- Access control to sources



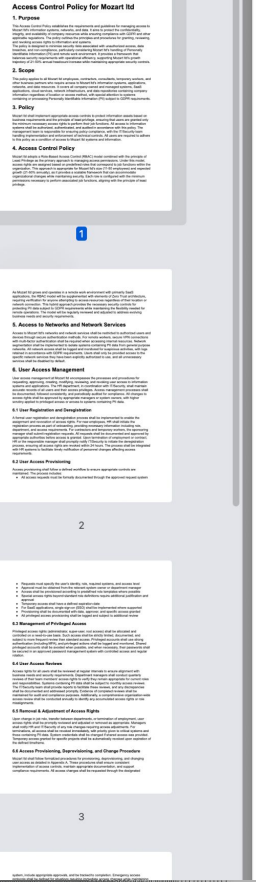
The Retrieval Pipeline

The Pipeline



Chunking

- Splitting a doc into chunks, how hard could it be?
- Chunk size (1000 chars)
- Overlap (20%)
- Sentence boundaries (U.S., e.g., !, $1.99)
- ♦️ Pragmatic_segmenter
- Don’t forget the meta data (source, location)
- Chunks become DB records



Access Control Policy for Mozart ltd

1. Purpose

This Access Control Policy establishes the requirements and guidelines for managing access to Mozart ltd's information systems, networks, and data. It aims to protect the confidentiality, integrity, and availability of company resources while ensuring compliance with GDPR and other applicable regulations. The policy outlines the principles and procedures for granting, reviewing, and revoking access rights to information and systems.
The policy is designed to minimize security risks associated with unauthorized access, data breaches, and non-compliance, particularly considering Mozart ltd's handling of Personally Identifiable Information (PII) and remote work environment. It provides a framework that balances security requirements with operational efficiency, supporting Mozart ltd's growth trajectory of 21-50% annual headcount increase while maintaining appropriate security controls.

2. Scope

This policy applies to all Mozart ltd employees, contractors, consultants, temporary workers, and other business partners who require access to Mozart ltd's information systems, applications, networks, and data resources. It covers all company-owned and managed systems, SaaS applications, cloud services, network infrastructure, and data repositories containing company information regardless of location or access method, with special attention to systems containing or processing Personally Identifiable Information (PII) subject to GDPR requirements.

3. Policy

Mozart ltd shall implement appropriate access controls to protect information assets based on business requirements and the principle of least privilege, ensuring that users are granted only the minimum necessary access rights to perform their job functions. All access to information systems shall be authorized, authenticated, and audited in accordance with this policy. The management team is responsible for ensuring policy compliance, with the IT/Security team handling implementation and enforcement of techni
to this policy as a condition of access to Mozart ltd

4. Access Control Policy

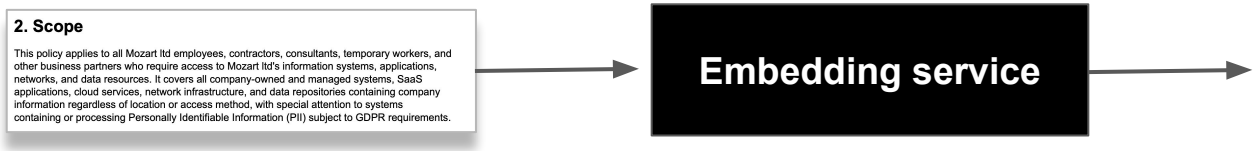
Mozart ltd adopts a Role-Based Access Control (RB
Least Privilege as the primary approach to managir
access rights are assigned based on predefined rol
organization. This approach is appropriate for Moza
growth (21-50% annually), as it provides a scalable
organizational changes while maintaining security.
permissions necessary to perform associated job fu
privilege.

2. Scope

This policy applies to all Mozart ltd employees, contractors, consultants, temporary workers, and other business partners who require access to Mozart ltd's information systems, applications, networks, and data resources. It covers all company-owned and managed systems, SaaS applications, cloud services, network infrastructure, and data repositories containing company information regardless of location or access method, with special attention to systems containing or processing Personally Identifiable Information (PII) subject to GDPR requirements.

Vectorize

- An array of numbers that describes a thing
- x,y coordinates but, n dimensions
- Meaning is encoded



Store

- PGVector
- neighbor

Create a migration

```ruby
class AddEmbeddingToItems < ActiveRecord::Migration[8.1]
  def change
    # cube
    add_column :items, :embedding, :cube

    # pgvector, MariaDB, and MySQL
    add_column :items, :embedding, :vector, limit: 3 # dimensions

    # sqlite-vec
    add_column :items, :embedding, :binary
  end
end
```

Add to your model

```ruby
class Item < ApplicationRecord
  has_neighbors :embedding
end
```

Update the vectors

```ruby
item.update(embedding: [1.0, 1.2, 0.5])
```

Retrieve

- ♦️Neighbor
- Embed the search term
- Pro tip: consider re-writing the query
  - “Could I possibly cancel my account please?”
  - “Can cancell acount pls?”

```
Item.nearest_neighbors(:embedding, [0.9, 1.3, 1.1], distance: "euclidean").first(5)
```

Retrieve bonus: keyword search

- Sometimes embeddings are not enough
- E.g. internal acronyms “NIST CSF 2.0”
- ♦ engtagger (English Part-of-Speech Tagger)
- Extract nouns
- ♦ pg_search

Searching against one column

To search against a column, pass a symbol as the :against option.

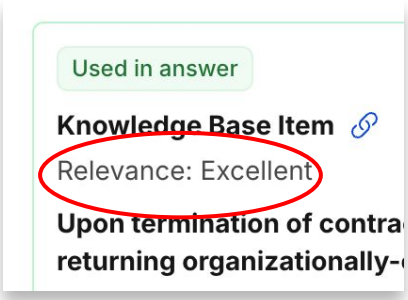
```
class BlogPost < ActiveRecord::Base
  include PgSearch::Model
  pg_search_scope :search_by_title, against: :title
end
```

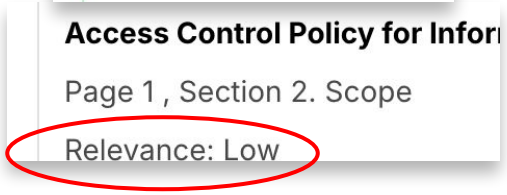
We now have an ActiveRecord scope named search_by_title on our BlogPost model. It takes one paramete search query string.

```
BlogPost.create!(title: "Recent Developments in the World of Pastrami")
BlogPost.create!(title: "Prosciutto and You: A Retrospective")
BlogPost.search_by_title("pastrami") # => [#<BlogPost id: 2, title: "Recent Developments in
```

Rerank

- Infer relevancy (cosine similarity ok, keyword not so much)
- Retrieve x items, use top N items
- Can use LLM or dedicated service, I like Cohere





PYTHON cURL

```python
import cohere

co = cohere.ClientV2()

query = "What is the capital of the United States?"
docs = [
    "Carson City is the capital city of the American state of Nevada. A
    "The Commonwealth of the Northern Mariana Islands is a group of isl
    "Charlotte Amalie is the capital and largest city of the United Sta
    "Washington, D.C. (also known as simply Washington or D.C., and of
    "Capital punishment has existed in the United States since before t
]

results = co.rerank(
    model="rerank-v4.0-pro", query=query, documents=docs, top_n=5
)
```

```
V2RerankResponse(
    id="2104ccd0-74b5-4951-9bb1-cc543b26720f",
    results=[
        V2RerankResponseResultsItem(
            index=3, relevance_score=0.943264
        ),
        V2RerankResponseResultsItem(
            index=2, relevance_score=0.62209207
        ),
        V2RerankResponseResultsItem(
            index=1, relevance_score=0.6054258
        ),
        V2RerankResponseResultsItem(
            index=0, relevance_score=0.59040135
        ),
        V2RerankResponseResultsItem(
            index=4, relevance_score=0.4664567
        ),
    ],
```

The Prompt

The System Prompt

Initial instructions

Response guidance

JSON structure rules

Guardrails

Example

System Prompt details

- **Initial instructions:** “you are an expert security analyst...”
- **Response guidance:** Only info from context items, answer length, prioritise high relevance items, prioritise certain context
- **JSON structure:** “only reply in JSON, using these keys...”, “array of citations with verbatim text, context item ID”
- **Guardrails:** “do not refer to context items”, “do not justify”, “do not use lists or tables”, “only answer no follow up questions”
- **Example:** JSON response example

User Prompt details

- **User question:** consider rewriting
- **Context items:** inc. meta info

```
<<~PROMPT
  ----Question----
  #{@question}

  ----Available Context Items----
  #{format_context_items(context_items)}
PROMPT
```

```
----Available Context Items----
[Knowledge Base Entry - Relevance: 96.7%]
ID: 125457
Question: Upon termination of contract or business relationship, are
Answer: Upon termination, employees and business partners are informe


[Knowledge Base Entry - Relevance: 5.1%]
ID: 125486
Question: Does your organization have a process to promptly deactivat
Answer: The organization has a process to promptly deactivate supplie
```

Bonus: Confidence rating

Question

Has your company (including subsidiaries, parent, affiliate, etc.) suffered a data loss or security breach within the last five (5) years?

Explain this

Applicable View rationale

Answer

No

No, the organisation has not suffered a data breach in the last 36 months.

Used in answer

Knowledge Base Item

Relevance: Good

Has there been any data breach in the last 36 months?

No

Requires Attention

- Missing citation: Documentation only covers last 36 months, but question requires confirmation for the full 5-year (60-month) period - 24 months of the requested timeframe lack documented evidence

Recheck Answer

Confidence: Very good

The answer is partially supported by the citation but has a temporal mismatch issue.

The citation confirms 'No' data breaches in the last 36 months, which the answer correctly references.

However, the question explicitly asks about the last 5 years (60 months), while the available documentation only covers 36 months.

The answer acknowledges this gap transparently, stating there is 'no documented evidence' for the full 5-year period.

While the answer is technically accurate about what can be confirmed from the citations, the 24-month gap in coverage (months 37-60) represents a significant limitation.

For a small organization (1-10 employees) with cloud-based infrastructure, this level of documentation might be reasonable, but the incomplete timeframe coverage prevents a higher confidence score.

The answer doesn't make unsupported claims, but the partial temporal coverage is a material limitation for fully answering the question as asked.

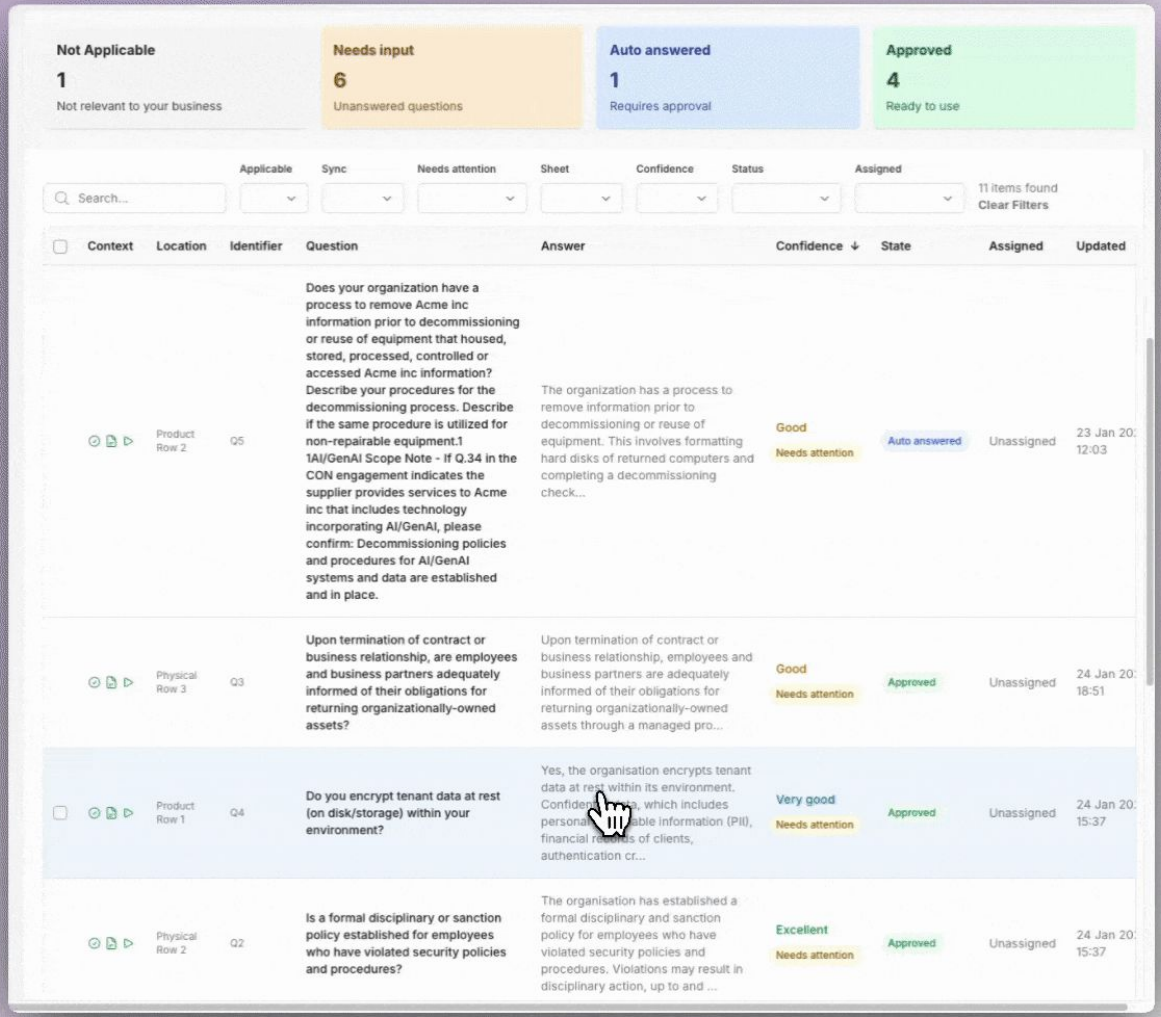
System prompt

```
  "system": [
    {
      "text": "You are an expert evaluator assessing the confidence level of AI-generated answers based on source citations.

CONFIDENCE SCALE:
- 0.9-1.0: Very high confidence - Answer is directly and clearly supported by citations with strong textual matches.
- 0.8-0.9: High confidence - Citations strongly support the answer with relevant, specific evidence.
- 0.6-0.8: Good confidence - Citations support the answer but with some minor gaps or less direct matches.
- 0.4-0.6: Moderate confidence - Some relevant citations exist but coverage is incomplete or tangential.
- Below 0.4: Low confidence - Citations don't adequately support the answer, or there are factual errors.

EVALUATION CRITERIA:
1. Relevance of citations to the question asked
2. Quality and specificity of the cited text
3. Coverage of the answer by the citations
4. Accuracy of the answer relative to the source material

GUIDELINES:
- Evaluate relative to the organisation's size and infrastructure profile
- Stay within scope - don't flag adjacent concerns the question didn't ask about
- Be pragmatic: if the answer reasonably addresses the question with supporting evidence, confidence should be 0.8+
",
      "type": "text"
    }
  ],
```

User prompt

```
You are evaluating the confidence level of an AI-generated answer based on the quality and relevance of the source citations.

QUESTION: Do you encrypt tenant data at rest (on disk/storage) within your environment?
ANSWER: Yes, the organisation encrypts tenant data at rest within its environment using industry-standard encryption protocols.


CITATIONS:
Citation 1:
FullText: # 5.1 Confidential Data Handling Confidential data requires the highest level of protection against[...]
CitedText: Must be encrypted at rest and in transit using industry-standard encryption protocols
Document: Data Management Policy for Secure Information Handling
Page: 3
Section: 5.2 Restricted Data Handling

Citation 2:
[...]

Citation 3:
[...]

Organization Profile:
[{\"Q\":\"How many employees are there in your company?\",\"A\":\"11-50\"}[...]]

Based on the question, answer, citations and organization profile provided, evaluate the confidence level using the scale provided in the system
prompt.

Provide a complete JSON response:
{
  \"confidence\": <float between 0.0 and 1.0>,
  \"rationale\": \"<explanation of confidence assessment and how well citations support the answer>\",
  \"requires_attention\": <true/false - see rules below>,
  \"points_to_address\": [\"<see categorisation below>\"]
}

REQUIRES_ATTENTION RULES:
- Set to FALSE if confidence is 0.8+ and there are no factual errors
- Set to TRUE only if: confidence is below 0.8, OR there are factual errors, OR critical requirements from guidance are unmet

POINTS TO ADDRESS - only include if genuinely problematic:
1. Answer quality issues: Information EXISTS in citations but wasn't used well
2. Knowledge gaps: Required information DOES NOT EXIST in citations


IMPORTANT:
- Consolidate similar points into one
- Do not flag \"nice to have\" improvements - only genuine gaps
```

Pulling it on together



Thanks!

Slides, gems, video

Available for consulting