Data Mining for Mining Data

Extracting Structured Data for Competitive Intelligence

MIT MANAGEMENT BUSINESS ANALYTICS

Faculty Advisor
Prof. Dimitris Bertsimas

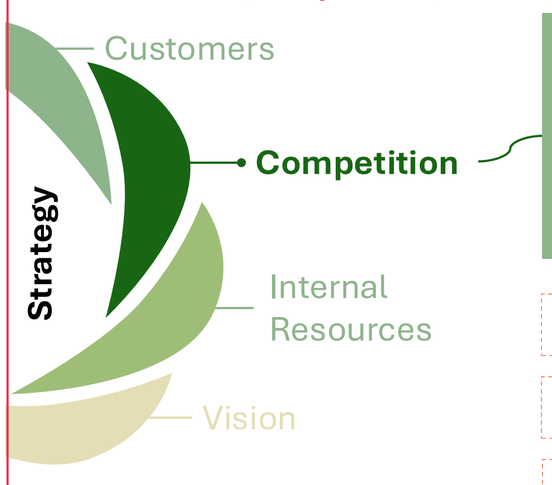
Company Advisor
Agnès Gorge

Daniel Rapoport

Ronan McCooey

Problem Statement

Competitive intelligence is a key input to OCP's strategy, but processes are **bottlenecked**, **subjective**, and **disconnected**.



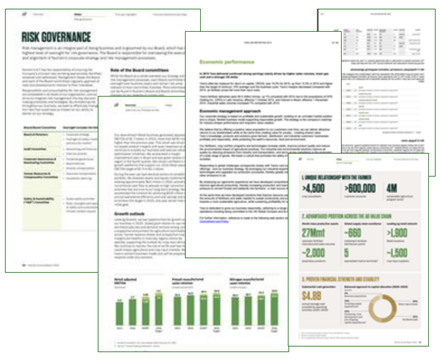
Current Process

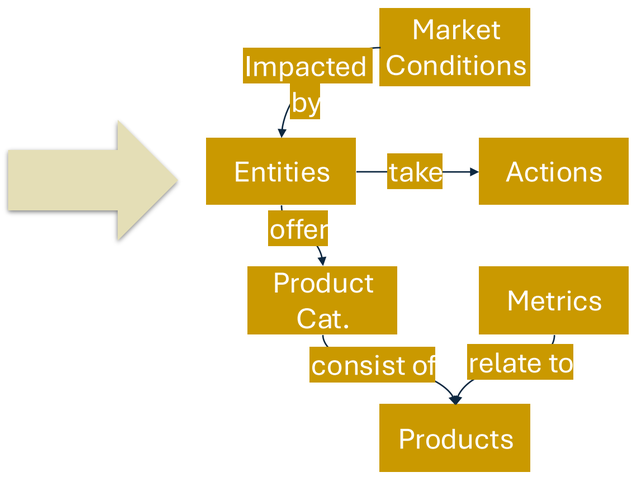
Competitor Data → Corporate Dev Team → Insights

- **Bottlenecked by resources**
- **Potential human biases**
- **Disconnected from tech stack**

Objective

Build a **generative AI system** that transforms unstructured documents into **structured data** aligned with OCP's existing infrastructure.

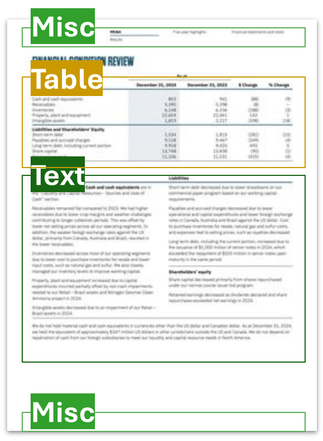




Methodology

Ingestion

PDF reports are converted to **page-level objects** that LLMs can ingest, using specialized loaders capable of handling complex layouts, embedded tables, and text artefacts.




```
"formatted_tables":
"<Page Number: 62>\n**Table 1:**\n|
||December 31, 2024|December 31, 2023|$ …
</Page Number: 62>",
```

```
"formatted_text":
"<Page Number: 62>\nOverview Explanations
for changes in Cash …</Page Number: 62>"
```

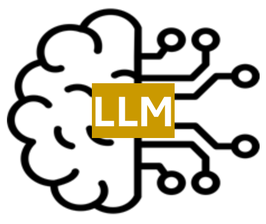
Chunking Strategy

- **Dual-stream segmentation:** text and tables are chunked and processed independently, allowing each stream to receive context-specific prompts.
- **Token budget for chunking (L)** calculated as

$$L = \lfloor \alpha \cdot C_{max} \rfloor - P$$

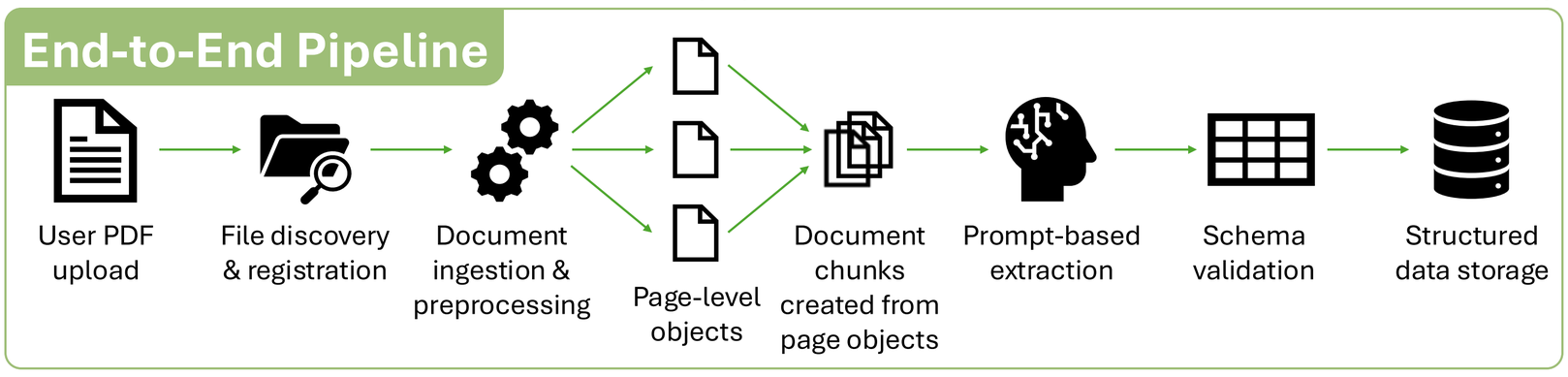
Where α is the inputted context fraction, C_{max} is the model context window, and P is the prompt tokens before context.

Extraction



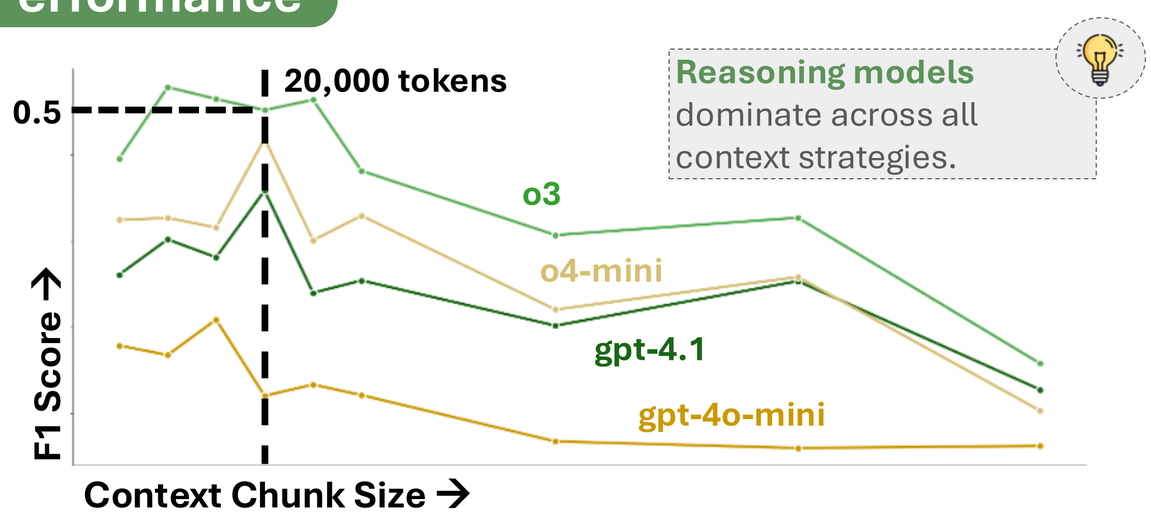
LLM extracts data from page objects using **domain-specific system prompt** and **schema-specific user prompt**, combined with careful **context engineering**.
Stores data with **source passages** for validation.

End-to-End Pipeline



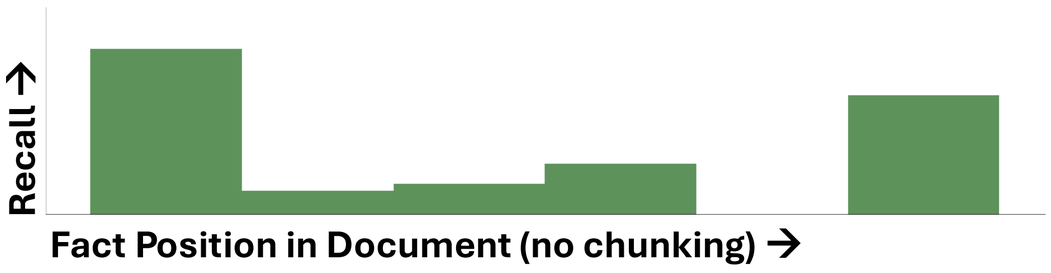
Results

Performance



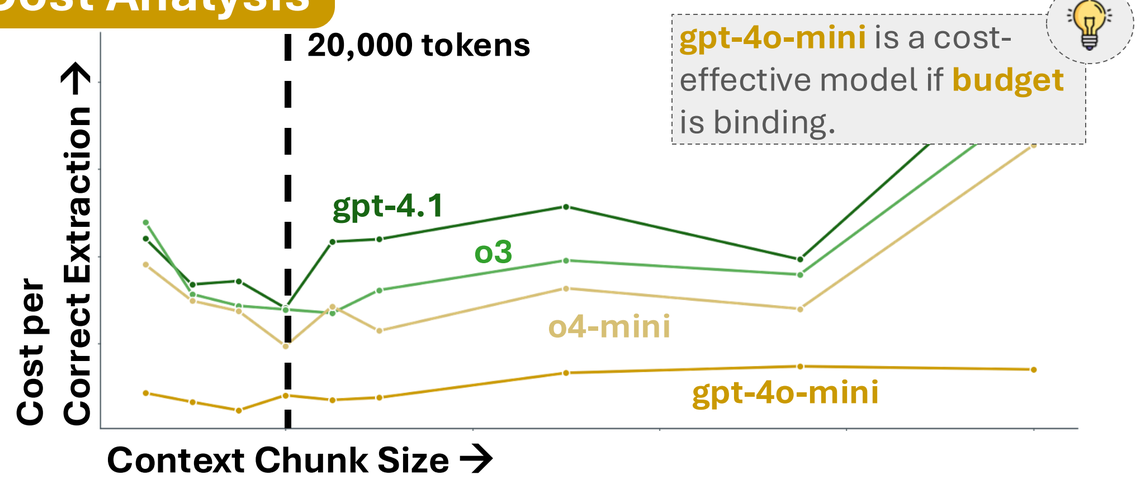
Reasoning models dominate across all context strategies.

- **Precision** increases while **recall** decreases as chunk size increases.
- **15,000 – 25,000 tokens** is the optimal chunk size to strike the balance between precision and recall, regardless of the size of the context window of the model used.
- **Extracted data format validity** is consistently high (above 95%), independent of the amount of context provided.



- Sharp decline in recall with large chunks can be explained by the **lost in the middle** effect.

Cost Analysis



gpt-4o-mini is a cost-effective model if **budget** is binding.

- **Larger chunk sizes** lead to significantly less costs as prompt overheads reduce and less duplicating output tokens are produced.
- **15,000 – 25,000 tokens** is again optimal when it comes to cost efficiency.
- Efficiency differences between models should not be considered, given business context in this case:
  - Total cost per 100-page document processed with o3 is roughly **$2**.
  - OCP need to process around **100 documents per month**.
  - Therefore, cost is not a binding constraint and **efficiency is dominated by performance** when it comes to model selection

Best Configuration

Ingestion
Custom Page-Level Loaders

Chunking
20,000 Tokens

Model
OpenAI o3

Impact

Direct: System reduces processing times by up to **95%**, freeing up **100 hrs** of work **monthly**.

Indirect: System enables **better strategy decisions** for company making **$10B/year**.

Outlook: Experimentation crucial for configuration of **production system**.