Document Extraction Stress Test Report 1 (Initial Exploration)

1. Experiment Overview

Date: 2025-09-14
Objective: Assess system behavior when processing non-SOW documents (MD&A; and Proxy/Notice), highlight shortcomings, and reveal stress points under schema mismatch.
Documents Used:
- Management Discussion and Analysis.pdf (Minaurum Gold Inc.)
- Notice of Annual Meeting of Stockholders.pdf

2. Input Documents

Document Name	Type	Key Characteristics
Management Discussion and Analysis.pdf	MD&A	Risk Factors, Liquidity, Operations, Financials. Long narrative financial disclosure.
Notice of Annual Meeting of Stockholders.pdf	Proxy/Notice	Meeting details, record date, proposals, board recommendations, voting methods.

3. Extraction Results

3.1 Management Discussion and Analysis
- Sections Produced: None (all fell back to `{}`)
- Notable Gaps: Executive Summary, Risks, Assumptions all reported 'No extracted content found.'
- Unexpected Artifacts: Drafting Agent rejected with 'insufficient information' for every section.

3.2 Notice of Annual Meeting of Stockholders
- Sections Produced: None (all fell back to `{}`)
- Notable Gaps: Meeting overview, proposals, and voting instructions were present but not mapped
to schema slots.
- Unexpected Artifacts: Same fallback pattern → 'No extracted content found' for every SOW section.

4. Observed Shortcomings

- Rigid Schema Dependence: Only SOW-style labels supported.
- No Semantic Mapping: Couldn’t map alternative section names.
- Empty Outputs: Passing `{}` to the drafting agent caused total section failure.
- Poor Graceful Degradation: No 'Not applicable' fallback.

5. Stress Test Findings

- Document Complexity Handling: Could not handle alternate document genres.

- Consistency: Both MD&A; and Proxy failed identically.
- Scalability: Every non-SOW doc type will fail as-is.
- Robustness: No fallback strategy for alternate labels.

6. Key Insights

- Current pipeline cannot generalize beyond SOW/Proposal style documents.
- Failures stem from hard-coded section expectations.
- Real business documents (regulatory filings, shareholder notices) use different headings and structures; without semantic mapping, the system cannot extract meaningful content.

7. Suggested Stress Test Directions

- Introduce a **universal meta-schema** (sections, facts, risks, lists) instead of fixed SOW slots.

-Implement a **doc-type router** to detect MD&A vs Proxy vs SOW.

-Add **semantic header mapping** (e.g., “Risk Factors” → Risks, “Election of Directors” → Deliverables/Decisions).

-Ensure **graceful N/A fills** rather than empty {} outputs.

-Re-test with additional Proxy, MD&A and several genre documents to confirm improved resilience.