Brainstorming Problem: Semantic Search for Spreadsheets

At **Superjoin**, we believe spreadsheets are more than grids—they're documents full of business logic, calculations, and meaning. Users think semantically ("Where are my profit calculations?" "Show me customer metrics," "Find efficiency ratios"), but spreadsheets mostly support structural queries (exact text matches, formula types, filters).

The Challenge

How might we design a semantic search system for spreadsheets that lets users ask natural-language questions and returns **conceptually relevant** results with **useful context** (not just cell references)?

Your solution should consider how to:

1. **Understand Business Concepts**
    - Synonyms/aliases (e.g., "sales" ≈ "revenue", "profit" ≈ "earnings").
    - Context signals (e.g., "Marketing Spend" → cost; "Marketing ROI" → efficiency metric).
    - **Formula semantics** (e.g., =(Revenue-COGS)/Revenue → margin %).
2. **Process Natural-Language Queries**
    - Conceptual (e.g., "Find all profitability metrics").
    - Functional (e.g., "Show percentage calculations").
    - Comparative (e.g., "Budget vs actual analysis").
3. **Rank & Explain Results**
    - Relevance scoring (semantic match, context importance, formula complexity, data recency).
    - Output design: concept name, human-readable location, value/formula, explanation ("why this matched"), business context.
4. **(Bonus) Multi-Sheet Understanding**
    - Link related concepts across tabs (Budgets, Actuals, Forecasts), and recognize relationships.

> **Note:** You are **not** expected to build anything for this round. We're assessing your **critical thinking** and ability to pressure-test a solution.

Solution Expectations

Your single axis of optimization is accuracy. Focus on surfacing the factors that impact the accuracy of your search engine.

Preparation Expectations (~4-5 Hours)

To get the most from our discussion, please **prepare thoroughly**:

- Mentally run your proposed workflow on realistic spreadsheets; identify gaps, edge cases, and failure modes. There are n nuances of data on a spreadsheet, capture those as well.
- **Alternatives**: Present **two or more distinct approaches** (e.g., rules + heuristics vs. embeddings/LLM-assisted vs. hybrid index). We will discuss trade-offs on the call if you identify them.
- Pick some example queries that you can run on sample sheets.
- **Artifacts**: Create **diagrams** (Excalidraw or Whimsical preferred) that illustrate your approach. Screenshots/PDFs are fine.

What to Bring to the Call (We'll walk through these)

1. **Short brief** covering:
- Problem framing & assumptions
- What Data structure(s) are you using and why
- Suggested Architecture/ Workflow Diagram (Excalidraw/Whimsical) - Optional but will help in the conversation.
- Approach alternatives & trade-offs

Constraints & Hints (useful guardrails)

- **No coding required** for this round.
- Assume access to typical spreadsheet artifacts: sheet/tab names, headers, ranges, formulas/ASTs, cell formats, sample values.
- Be explicit about **confidence** and fallback behavior.
