Project Ideas and Datasets

*Please include project ideas and dataset links in this doc - to be discussed on Tuesday.*

**Zainab**

AI Compliance Officer

Utilize RAG over regulatory AI policy documents (NIST, EU AI Act) to analyze an AI use case, classify risk level, and generate a citation-backed compliance checklist.

Corpus:

1. Policy Documents *(Can also use summary docs 8-10 pages - easier for chunking + can include more policies to diversify use case risk analysis)*
    a. EU AI Act: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=OJ:L_202401689

    b. NIST: https://nvlpubs.nist.gov/nistpubs/ai/nist.ai.100-1.pdf

2. Use Case Datasets: (need to research or create synthetic 5-10 use cases)
3. Ground truth to determine accuracy/see how well the model is performing

Requirements:

1. Vector DB For Document Chunking
2. LLM

Pipeline:

1. **User Scenario Input** (user enters AI use case) →
2. **Intake Agent** (LLM extracts key fields from user input i.e., industry, users, stakeholders etc.) →
3. **RAG Policy Retrieval** (get all policy chunks based on scenario key words) →
4. **Determine Risk Level** (Classifier uses retrieved chunks to determine risk levels) →
5. **Report Generator Agent** (LLM generates final report (pdf) including citations)

---

**Nazgul**

**Medical RAG assistant:**

**Combination of:**

1. Data Storage (user uploads medical records anytime)

2. Medical knowledge retrieval (from trusted sources)
3. Natural-language generation (using an LLM)
4. Reasoning + explanation (summaries, guidelines, evidence-based answers)

**Architecture:**

User -> Personal Medical History + Medical Info from Resources -> LLM -> Answer

**Ayon**
https://www.kaggle.com/code/pcbreviglieri/predicting-smart-grid-stability-with-deep-learning

Predicting Smart Grid Stability with Deep Learning

In a smart grid, consumer demand information is collected, centrally evaluated against current supply conditions and the resulting proposed price information is sent back to customers for them to decide about usage. As the whole process is time-dependent, dinamically estimating grid stability becomes not only a concern but a major requirement.

Put simply, the objective is to understand and plan for both energy production and/or consumption disturbances and fluctuations introduced by system participants in a dynamic way, taking into consideration not only technical aspects but also how participants respond to changes in the associated economic aspects (energy price).

**Manikandan**
https://www.kaggle.com/code/ayaabdalsalam/hair-loss-analysis/input

Predict Hair loss based on Genetic and medical conditions as input features

This CSV file contains data on various factors that may contribute to baldness in individuals. The dataset is structured with rows representing individual records and columns representing different attributes related to potential contributors to baldness.

—-

Final Project

Financial Analyst Assistant with RAG and News Intelligence

Build a financial analyst assistant that ingests financial documents (10-K, shareholder meetings, investor presentations, earning calls etc) for which a business user (analyst, PM, or strategy lead) could ask about company's risks, performance drivers etc. Additionally, we would integrate a news API to pull relevant recent articles, press releases, and sentiment about the company and to avoid scope creep, keep the focus on a single domain. Overall, the end user may ask questions to determine company trajectory.

End-to-end Pipeline:

Part A) Preparing for RAG: Ingestion and Indexing

1. Determine Scope: Select approx 5 companies from a specific domain/industry e.g. semiconductors (NVIDIA, Intel etc.) *might be interesting due to their impact on bursting the AI bubble*
2. Collect financial documents for each organization:
    a. 10K filings, annual reports etc (what is readily available for all selected companies)
3. Document Extraction and Cleaning: Extract text and tables from pdfs via Docling or other similar tools
4. Chunking and metadata extraction:
    a. Determine VDB, embedding model, and chunking strategy
    b. Attach metadata to chunks e.g. company name, document type, doc year, section etc.
5. News Ingestion:
    a. Ensure keywords from the user's query are extracted to feed to news API for retrieving relevant articles (more info below)

Part B) User Query and Answer Generation

1. User inputs query:
    a. How has company x been performing over the last year?
    b. How might recent news about growing competition with China affect the company's strategy?
2. Perform RAG over financial Docs:
    a. Understand the query (company name, query intent etc) to filter through the vector DB and retrieve relevant chunks
3. Fetch relevant news:
    a. Call news API to fetch 1-3 relevant articles within a given time frame (7 days)
    b. Get headline, a 1-2 sentence summary, and hyperlink to actual article
4. LLM reasoning and output generation:

a. Get 1. Retrieved relevant financial docs chunks and 2. Relevant news info from steps 2 and 3 and feed into LLM with structured prompt to output the following:
    i. Company snapshot
    ii. Key risks
    iii. Recent developments (relevant news)
    iv. Trajectory
    v. Suggested next best action

When to output full report vs specific to the query:
- Broader user input: full report
- Narrowed down question: stick to relevant information ONLY

Additional Considerations:
Answer relevance evaluation: ?
Fine-tuning?

Must include:
citing sources
End-User Experience: web/chat UI interface

News API Options:

NewsAPI.org
- can do keyword search and/or date range to extract relevant news articles
- Outputs: Author, title, description, URL (can we make it into a hyperlink?), date published, content (don't think we need to be this detailed)
- Has a free tier with 100 requests per day - should be usable for this project

Sprint Planning

Week 1:
- Domain definition
    - Semiconductor industry:
        - Nvidia - Ayon
        - Intel - Brenden
        - Broadcom - Mani
        - TSMC - Naz
        - Samsung - Zainab
- Pipeline definition - done (input from whole team)
- Baseline architecture diagram - Zainab
- Resource/Data gathering - 10-K filing, Quarterly Reports (All quarters 2024 and 2H 2025)
- Document parsing (PDFs/tables) - Mani possible option: https://www.docling.ai/

- Vector Database research - (current option: Chroma) - Zainab
- News API research and implementation - Zainab
- Small embedding model - (Nazgul)
- Lightweight LLM model research and possible fine-tuning - Brendan
- Prompt engineering (RAG model prompts research) - Ayon

Week 2 - primarily testing and UI setup:

- Parsing continued - Mani
- Bring results together and filling gaps
- UI
- Prompt fine tuning for improved results
- Internal testing + with ChatGPT, Gemini, and other platforms
- Answer relevance evaluation (brainstorm and implement)
- Citing sources

Week 3:

- Slide deck
- Demo build - 10 min
- Report - 8 pages
- Github repo cleanup

| Week One Sprint Planning - Initial Setup and Implementation<br>**Expected Completion Date: 12/01/2025** | | | |
|---|---|---|---|
| **Tasks** | **Sub-Tasks** | **Asignees** | **Status** |
| Domain Definition - semiconductor industry | | All team members | Done |
| Pipeline definition | | All team members | Done |
| Resource + Data Gathering: 10K filings, quarterly reports (all quarters 2024 and 2H 2025) | Nivedia | Ayon Roy | Done |
| | Intel | Brandon Rodrig… | Done |
| | Broadcom | Manikandan Kar… | Done |
| | TSMC | Nazgul Maksutk… | Done |
| | Samsung | Zainab Makhdum | Done |
| Baseline architecture diagram | | Zainab Makhdum | Done |
| Document parsing (pdfs and tables)<br><br>possible option: https://www.docling.ai/ | | Manikandan Kar… | Done |
| News API research and initial implementation | | Zainab Makhdum | Done |
| Vector Database research - (current option: Chroma) | | Zainab Makhdum | Done |
| Small embedding model - (Nazgul) | | Nazgul Maksutk… | Done |
| Lightweight LLM model research and fine-tuning | | Brandon Rodrig… | Done |
| Prompt engineering (RAG model prompts research and implementation) | | Ayon Roy | Done |

| Week Two Sprint Planning - Continue Building<br>Expected Completion Date: Saturday 12/6/2025 | | | |
|---|---|---|---|
| **Tasks** | **Deliverables** | **Asignees** | **Status** |
| Enhance News API results | Saturday | Zainab Makhdum | Done |
| Continue on document parsing (text only) | Thursday | Manikandan Kar… | Done |
| Enhance chunking strategy and work on ingesting chunks into VDB<br><br>Convert text to embeddings | Friday Morning | Zainab Makhdum<br>Nazgul Maksutk… | Done |
| Draft code for LORA and Full Training - pass over to Ayon for running in cloud | Friday night/Saturday morning | Brandon Rodrig…<br>Ayon Roy | Done |
| Prompt enhancements | Saturday | Brandon Rodrig… | Done |
| GCP and TPU connectivity check - system setup and test runs | Friday Night<br><br>Saturday | Ayon Roy | Done |
| UI Research | | Zainab Makhdum | Done |

| Week Three Sprint Planning - Continue Building<br>Expected Completion Date: Wednesday | | | |
|---|---|---|---|
| **Tasks** | **Deliverables** | **Asignees** | **Status** |
| Enhance News API results | Wednesday | Zainab Makhdum | Done |
| Template chunking and final prompt creation | Friday Morning | Zainab Makhdum<br>Nazgul Maksutk… | Done |
| UI Implementation Draft | | Zainab Makhdum | Done |
| Parse remaining company documents into text files (10K and Q3 2025 filing) | Friday | Manikandan Kar… | Done |
| Virtual Machine and TPU | | Ayon Roy | Pending - continue to next sprint |
| Testing and example dataset generation (pending) | Thursday | Brandon Rodrig… | Pending - continue to next week |
| Initial report draft | Later | Brandon Rodrig… | Done |

| Week Three Sprint Planning - Continue Building<br>**Expected Completion Date: Friday** | | | |
|---|---|---|---|
| **Tasks** | **Deliverables** | **Asignees** | **Status** |
| Draft chat conversation flow | | Zainab Makhdum | Done |
| Enhance News API results | Wednesday | Zainab Makhdum | Done |
| Template chunking and final prompt creation<br><br>(jot down python packages versions) | Friday Morning | Zainab Makhdum<br>Nazgul Maksutk… | Done |
| UI Implementation Draft - clean up coloring (blue and white), ensure whole page is the chat, Implement initial Assistant message (please see below), find sample professional UI template | Friday | Ayon Roy | Done |
| Parse docs to json with file name and page no and text | Thursday | Manikandan Kar… | Done |
| Virtual Machine and TPU | | Ayon Roy | In Progress - continue to next sprint |
| Testing and example dataset generation (pending) | Friday evening | Brandon Rodrig… | In Progress - continue to next sprint |
| Initial report draft | Later | Brandon Rodrig… | Done |

| Week Four Sprint Planning - Continue Building<br>Expected Completion Date: Tuesday | | | |
|---|---|---|---|
| **Tasks** | **Deliverables** | **Asignees** | **Status** |
| Clean up prompt template (remove mention of Broadcom, remove chunk ID from citation) | | Zainab Makhdum | Done |
| Modifying the architecture | Wednesday | Zainab Makhdum | Done |
| Creating slides | Friday Morning | Zainab Makhdum | Done |
| UI Implementation - keep the entire page as the chat interface, remove welcome tab, keep names in footer, include contrast between user prompt and chat response, make background color darker to highlight chat box, hard code text to avoid API endpoints usage | Friday | Ayon Roy | |
| Final Report Draft | Thursday | Brandon Rodrig…<br>Manikandan Kar…<br>Nazgul Maksutk… | |
| Virtual Machine and TPU | | Ayon Roy | |
| Testing and example dataset generation (pending) | Friday evening | Brandon Rodrig… | In progress |

| Week Four Sprint Planning - Continue Building<br>**Expected Completion Date: Tuesday** | | | |
|---|---|---|---|
| **Tasks** | **Deliverables** | **Asignees** | **Status** |
| UI Implementation<br>- Brandon to share model integration details with Mani for UI (Monday) | Tuesday | Manikandan Kar…<br>Ayon Roy | URGENT |
| Modifying the architecture | Tuesday | Zainab Makhdum | |
| Slides<br>- Brendon to modify the technical aspects | Tuesday | Zainab Makhdum<br>Nazgul Maksutk…<br>Ayon Roy<br>Brandon Rodrig…<br>Manikandan Kar… | URGENT |
| Model creation<br>- Brandon to share model integration details with Mani for UI (Monday) | Tuesday | Brandon Rodrig… | |
| Final Report Draft | Thursday AFTERNOON | Brandon Rodrig…<br>Manikandan Kar…<br>Nazgul Maksutk…<br>Zainab Makhdum<br>Ayon Roy | |
| GitHub Repo<br>- Add files and code<br>- Add instructions on how to run it on readme file | Thursday AFTERNOON | Ayon Roy<br>Nazgul Maksutk…<br>Brandon Rodrig…<br>Manikandan Kar…<br>Zainab Makhdum | |
| Demo Script | Tuesday | Brandon Rodrig…<br>Zainab Makhdum | URGENT |

| Final Week Sprint Planning<br>Expected Completion Date: Wednesday 10:00 pm | | | |
|---|---|---|---|
| **Tasks** | **Deliverables** | **Asignees** | **Status** |
| Fine tuning model | Tue | Brandon Rodrig… | Done |
| Testing step 1 and 2 | Wed | Nazgul Maksutk… | Done |
| UI LLM integration | | Brandon Rodrig… | In progress |
| Slideshow technical slides - every topic covered and within time limit | Wednesday<br>Tuesday | Zainab Makhdum<br>Nazgul Maksutk… | |
| Script | | Zainab Makhdum | Done |
| Finish testing and add results to slideshow | Tuesday<br>Tuesday/Wednesday | Brandon Rodrig… | |
| Final Report | Friday | Person | |
| Github | Friday | Person | |
| Hardcoded UI (query + answer) | | Manikandan Kar…<br>Ayon Roy | |
| 10 am EST Friday 19th for demo - 1 hour at least<br>- Test run | | | |

**Accuracy (Groundedness/Faithfulness)**
**Definition:** The generated answer must be factually correct and directly supported by the external knowledge retrieved from the database, not fabricated by the LLM.
**What it checks:** Whether the model uses the provided context accurately (faithfulness) and if the information aligns with the source data (correctness).
**Why it matters:** Prevents the system from "hallucinating" or providing misinformation, building user trust.

**Coherence (Fluency/Logical Flow)**
**Definition:** The response should read naturally, with smooth transitions and logical connections between sentences and ideas, even when combining retrieved snippets.
**What it checks:** The grammatical correctness, readability, and overall organization of the output, ensuring it's not disjointed or jumbled.
**Why it matters**: A coherent response is easier to understand and interact with, improving the user experience and perceived quality of the system.

**Relevance (Contextual Appropriateness)**
**Definition:** The generated answer must directly address the user's original question or task, using the most pertinent retrieved information.
**What it checks:** Both the quality of the initial retrieval (is the context relevant?) and the final generation (does the answer stay on topic?).
**Why it matters:** Ensures the system provides a useful answer rather than just factually correct but irrelevant information, keeping the conversation focused.

Sunday

| |
|---|
| Slides generation |
| Final Architecture diagram |

| |
|---|
| Script for demo |
| Upload on Github |
| Final Report |

Aiming to finish: 17th