Investment Analysis: An AI-Powered Chatbot for Mining Investment Insights

Soham Agarwal, Durga Madhab Dash, Anto Frederic Henry Mohan Dass, Sai Bheeshma Ramaraju Pagilla, Chaitanya Varma Sanaboina, Matthew A. Lanham

Purdue University, Krannert School of Management
agarw402@purdue.edu; dashd@purdue.edu; ahenrymo@purdue.edu; spagilla@purdue.edu; csanaboi@purdue.edu; lanhamm@purdue.edu

BUSINESS PROBLEM FRAMING

All of us have used ChatGPT, Gemini, or other LLMs to sift through the large amounts of information on our screens that bombard us. The result is a neatly packaged output displayed
But how much of that output is accurate?
.

THE ISSUE
Financial investors have difficulty in going through large document like SEC 10k filings. [1]

THE DEEPER ISSUE
SEC filings provide detailed audited reports. Reading 100 page documents isn't easy!

RESOURCE CONSTRAINTS
Sorting through SEC data for additional information requires considerable time and involves high opportunity costs.

WHY SOLVE THIS PROBLEM?
Shift investment analysis from days to a hours to save costs, time, and improve decision-making.

[1] The Importance of SEC Filings for Securities Investors. **Yale University**. Retrieved from https://tenthousandrooms.yale.edu/project/importance-sec-filings-securities-investors

ANALYTICS PROBLEM FRAMING

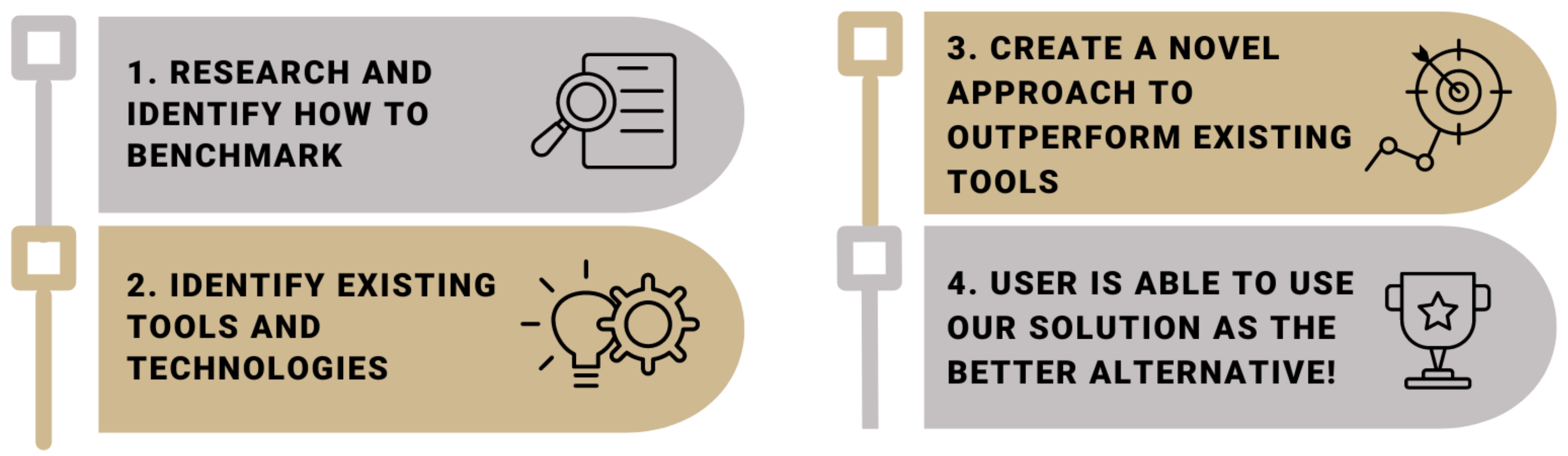

1. RESEARCH AND IDENTIFY HOW TO BENCHMARK
2. IDENTIFY EXISTING TOOLS AND TECHNOLOGIES
3. CREATE A NOVEL APPROACH TO OUTPERFORM EXISTING TOOLS
4. USER IS ABLE TO USE OUR SOLUTION AS THE BETTER ALTERNATIVE!

Fig 1. User (Investor) Journey

Problem: With Exponential Growth of Data

- Time Intensive
- Resource Intensive
- Financial expertise

Success Metrics: AI vs Human experts

- Accuracy
- Conciseness
- Efficiency

Assumptions: Of our Solution

- SEC Data: Reflect accurate information
- Web scraping: Gather representative reports.
- Gen AI: Is a competent tool to summarize key information

Justification of Approach

- Leverage public SEC data.
- Advanced Gen AI: Extraction and summarization.
- Focus on success metrics to determine value

DATA

- Our data source consists of 10-K filings in text format (unstructured). There is no inherent relationship in the data extracted.

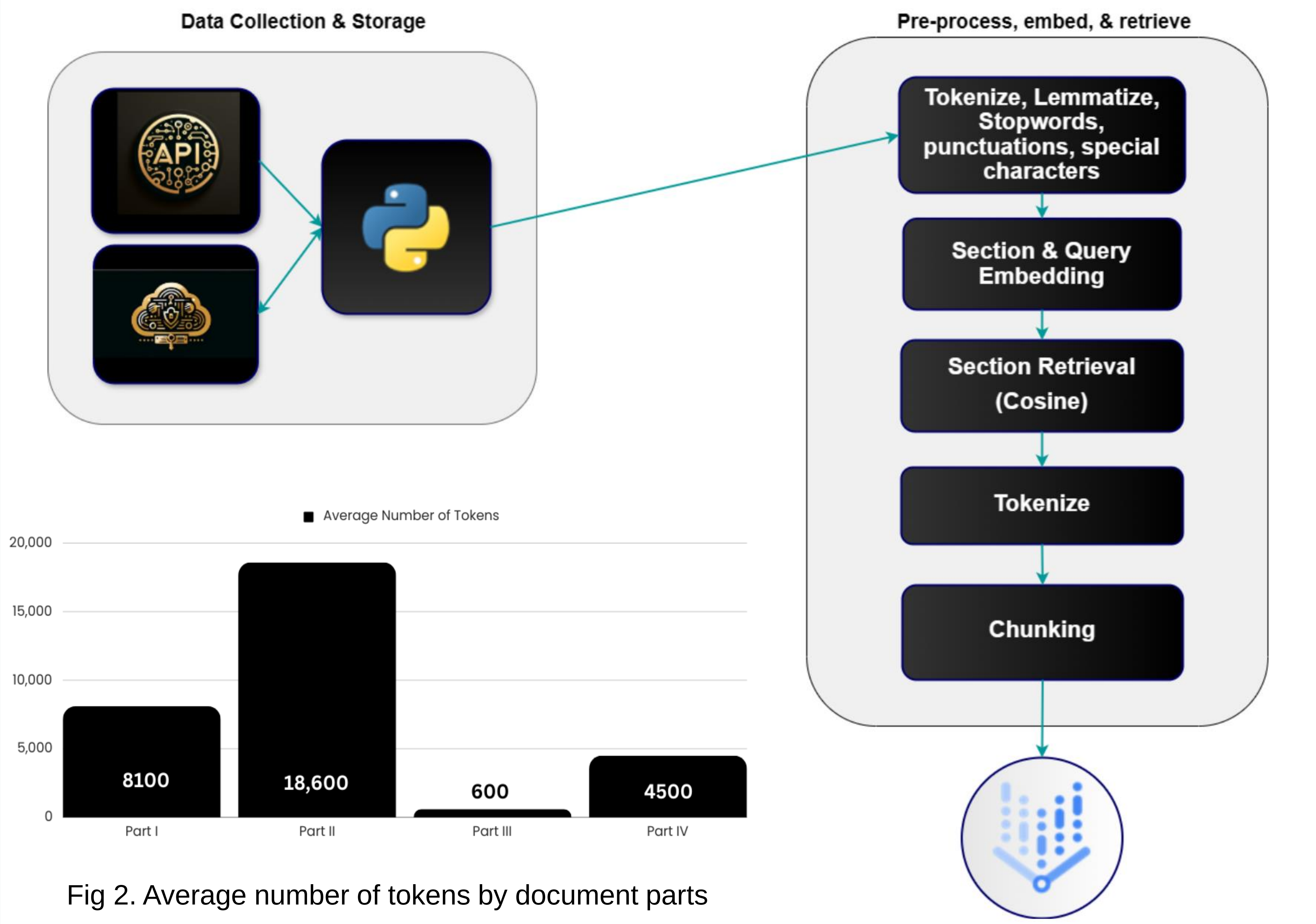


Fig 2. Average number of tokens by document parts

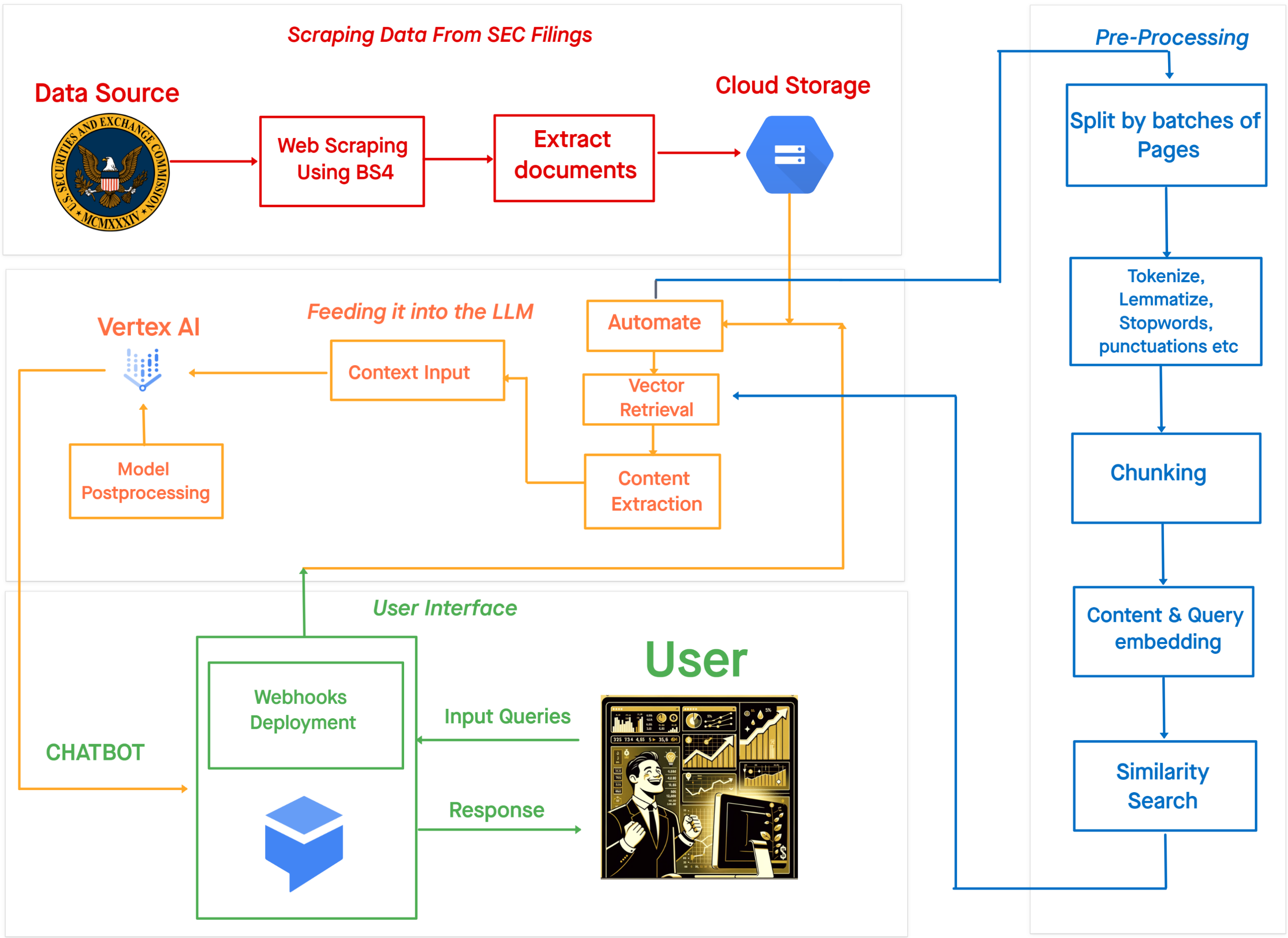
PURDUE UNIVERSITY | Mitchell E. Daniels, Jr. School of Business

METHODOLOGY

Scraped 10-K filings are stored in cloud storage, acting as the data source for our solution. Our solution uses RAG (Retrieval-Augmented Generation) which focuses on utilizing relevant content from PDFs based on user queries, addressing specific information needs without relying on the LLM's pre-trained knowledge or the need to train a model for this specific task (saving resources). This task is achieved by using a combination of tools like Python, NLTK, BeautifulSoup, LLM/Embedding models on Vertex AI etc.

- **Data Cleaning**: Ensured data quality through stopword & punctuation cleaning, tokenization, & lemmatization.
- **LLM Fine-tuning**: Enhanced model performance using one-shot prompting.
- **Query Rephrasing:** Optimized user queries to improve retrieval accuracy.
- **Chunking & Retrieval:** Divided PDFs into manageable sections (chunks) & retrieved the top-ranked chunks for information extraction.

MODEL



Model Evaluation:
The model is assessed against FinanceBench, with a focus on quantitative inquiries.

Interpretations & Utilization of the Model:
Our approach utilizes textbison/Gecko models for interpretation & embeddings and RAG for answering. Models are integrated within a chatbot interface for user interaction.

Enhancement Opportunities:
Integration with mathematical libraries is proposed to augment the accuracy of responses to quantitative queries

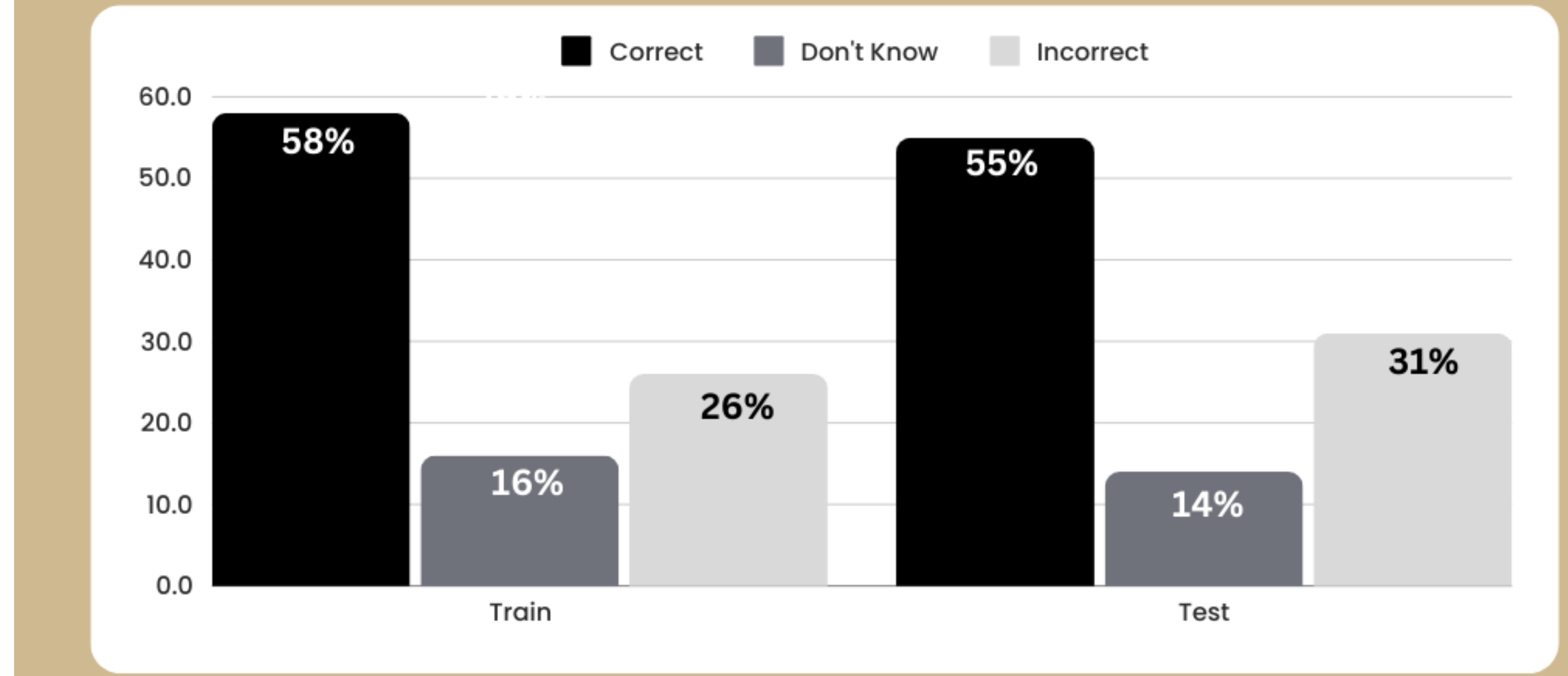
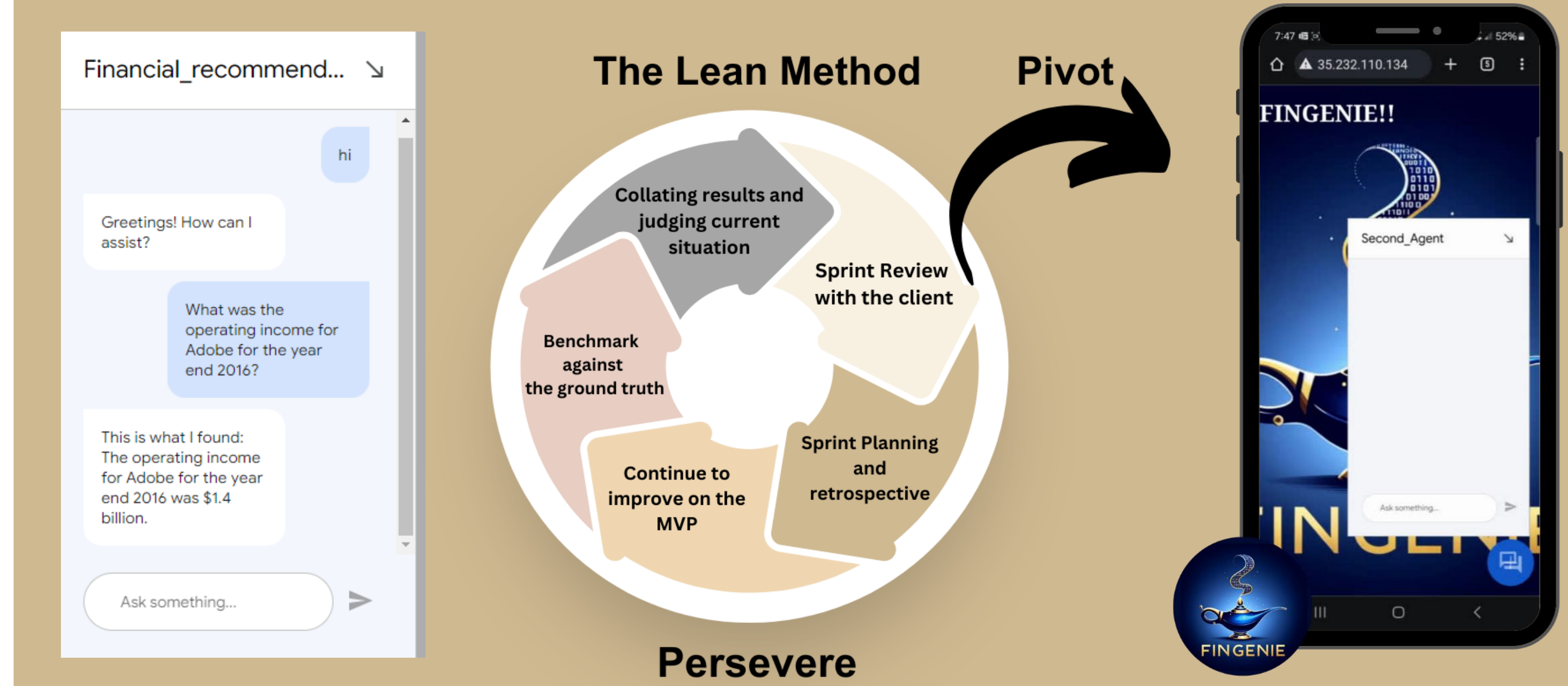


Fig 4. Best Model Performance (Train vs Test)

DEPLOYMENT & LIFE CYCLE MANAGEMENT


Financial_recommend...
The Lean Method
Pivot
Collating results and judging current situation
Sprint Review with the client
Sprint Planning and retrospective
Continue to improve on the MVP
Benchmark against the ground truth
Persevere
FINGENIE!!
Second_Agent
Ask something...
FINGENIE

hi

Greetings! How can I assist?

What was the operating income for Adobe for the year end 2016?

This is what I found: The operating income for Adobe for the year end 2016 was $1.4 billion.

Fig 4. Iterative lean methodology SDLC followed

- Development process involves enriching the product backlog, planning, and resuming work on Minimum Viable Product (MVP).
- A/B testing, in-depth research, and benchmarking against FinanceBench Data are conducted to calibrate model performance.
- If the MVP is not achieving substantial results, we pivot
- Areas of improvement include tuning context window, domain specific fine tuning, hugging face pipelines are being explored.
- As future scope, enhance the bot's accuracy by integrating multiple data interpretation models and pairing them with mathematical libraries for improved quantitative results

ACKNOWLEDGEMENTS

We would like to thank our industry partner Ashwin Mishra and Professor Matthew Lanham for their guidance and support on this project as well as the Purdue MS BAIM program for partially funding this work.

informs ANALYTICS CONFERENCE