Automating Franchise Document Analysis with Large Language Models

Marissa Inga, Kush Malhotra, Victoria Lin
Columbia University, Department of Industrial Engineering and Operations Research

Abstract

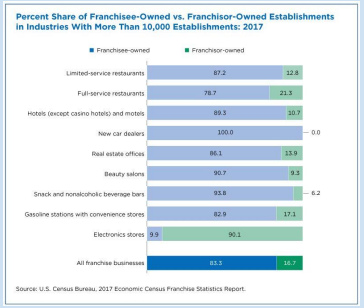
This project explores the usage of Large Language Models to systematically extract information from Franchise Disclosure Documents. This project was led by an industry partner at the Morgan Stanley Equity Research Division. Seven franchises were selected and their Franchise Disclosure Documents were downloaded from official websites. We then explored Google Gemini and Mistral AI APIs and created a pipeline to read and preprocess the data, query the API, and save down results in a uniform JSON format. The project necessitated developing an understanding of Franchise Disclosure Document structure, prompt engineering, and strategic data preprocessing. We were able to achieve significant accuracies across our set of Franchise Disclosure documents, especially through the use of Mistral AI which had a powerful model compared to the free Gemini version. This work builds on existing research into automating document extraction using Large Language Models.

Background & Introduction

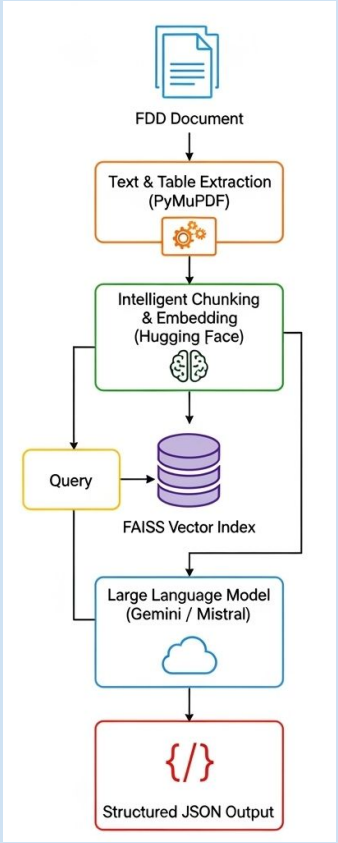
- Franchise Disclosure Documents (FDDs) exceed 300-900 pages of legal data that is often inconsistently formatted
- FDDs important for financial analysts to understand industry trends
- Manual extraction is time-consuming and costly

Project Goal

Determine if large language models (LLMs) could systematically extract structured data from complex Franchise Disclosure Documents (FDDs). By developing an automated pipeline to systematically extract key information from FDDs, we could cut down significantly on time consuming manual tasks performed by analysts at Morgan Stanley.



Methodology



Results

Our pipeline demonstrates that LLMs, when combined with smart chunking and embeddings, can extract structured insights (fees, locations) from 300+ page legal FDDs with ~75–80% accuracy, paving the way for scalable franchise data analysis.

High Level Results

Information Type	Accuracy (Approx.)	Notes
Royalty Fees	~80%	Usually consistent across FDDs
Advertising Fees	~75%	Extracted well but sometimes mislabeled
Misc. Fees	~60–65%	Hardest to standardize
Technology Fees	~70%	Lower due to variation in wording
Projected Location Counts	~70%	Struggles with table formatting

Example FDD Evaluation

Section	Gemini Output	Mistral Output
Franchise Name	Correct: "Taco Bell"	Correct: "Taco Bell"
Document Issuance Da	Present: "March 26, 2025"	Present: "March 26, 2025"
Investment Details	Missing	Multiple unit types with ranges captured
Fees (Items 5–7)	Partial list, some fees missing	Some fees listed, but null
Financial Performance	Missing (no sales data extracted)	Sales Information Correctly Extracted
Location Info (Item 20)	Total locations, projected openings mis	Total locations, projected openings corre
Agreement Term (Year	Present (20)	Present (20)
Renewal Options (Years	Present (10)	Present (10)
Parent Company	Present: "Yum! Brands"	Present: "Yum! Brands"

Conclusion & Next Steps

We conclude it is feasible to use LLMs to extract structured information from Franchise Disclosure Documents with generally high accuracy. Mistral AI and smart chunking and embedding were crucial for achieving our level of performance. Next steps include exploring further LLM APIs, integrating further franchise documents into the working set, optimizing the pipeline for improved runtime, and fine-tuning the LLMs on our data for accurate data retrieval.

Huge thanks to Bas Jaspers, Jim Strugger, Professor Ali Hirsa, and Miao Wang!