Name-Match LLM General Search

# Objectives:

1. Find an appropriate large language model to
   1. Perform sentiment analysis of news articles
   2. Identify names in negative news articles
   3. Based on positivity/negativity of news articles, build risk profiles with scores/tags.
   4. Compare the performance of existing models *(Spacy: en\_core\_web\_sm or lg/trf; Dslim/Bert-base-NER, ref. Hugging Face)* in terms of efficiency and accuracy.
2. Explore advanced NLP techniques to match names
   1. Vector representation of names
   2. Clustering names in 3D space

## Plan of next week

* Look into data
* Research
* Implement 1–2 LLMs if there is more capacity

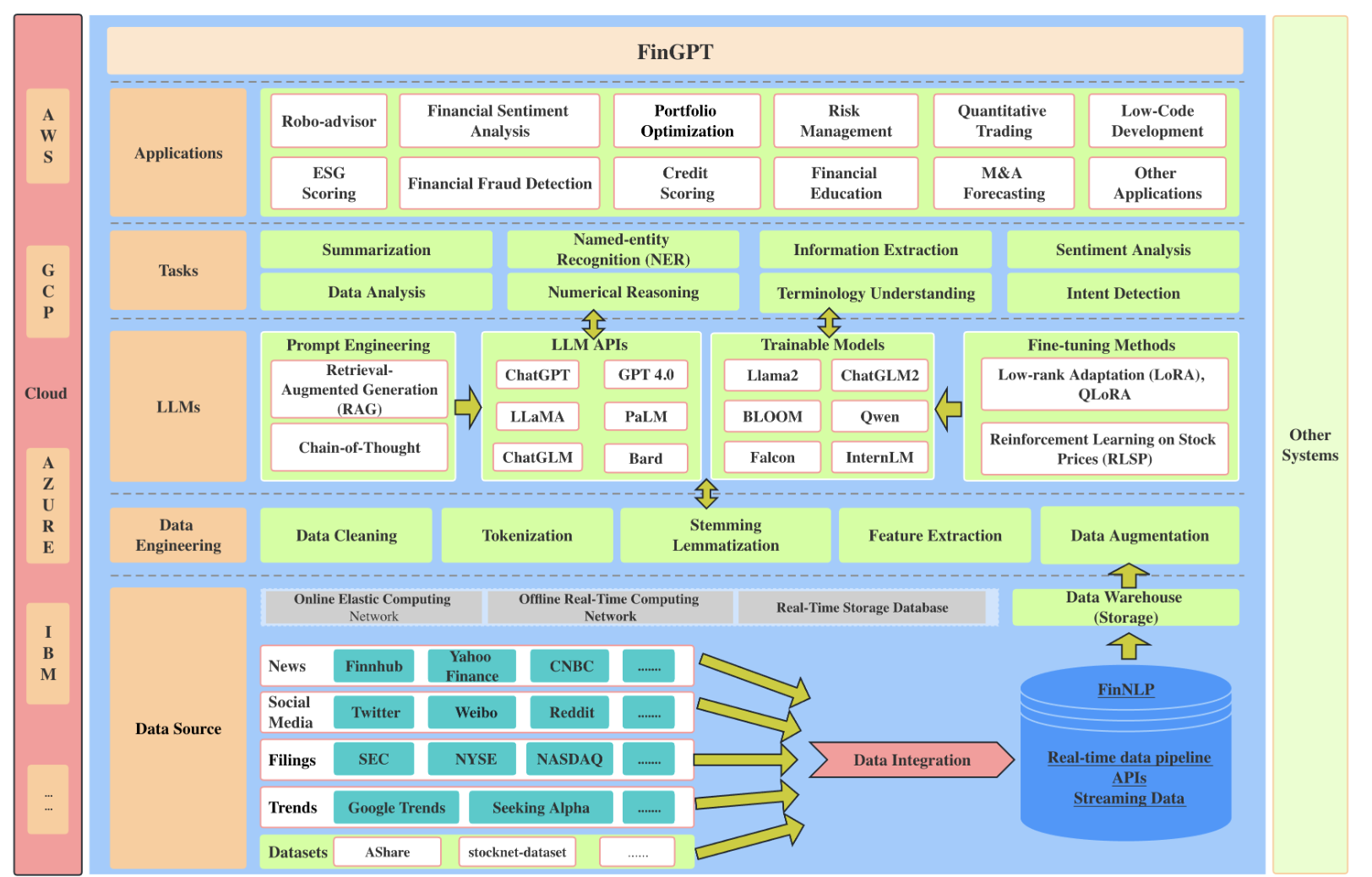
# Open-source LLMs

Start with small LLMs first (less than 10B). LLMs with more than 10B parameters are difficult to scale up on Greene clusters.

**Requirements:**

* One-stop shop model
* Open-source
* Scalable to handle large dataset
* Multi-lingual support
* Hopefully performs NER better than existing NLP models

| **LLM** | **Organization** | **# of Parameters** | **Characteristics** |
| --- | --- | --- | --- |
| [BloombergGPT](https://www.bloomberg.com/company/press/bloomberggpt-50-billion-parameter-llm-tuned-finance/) (not open-source) | Bloomberg | 50B | Specializes in finance; input source consists of public data + financial data; good at sentiment analysis and NER. |
| [FinGPT](https://github.com/AI4Finance-Foundation/FinGPT) | AI4Finance | 3M | Lightweighted LLM for finance which can be fine-tuned easily and costs much less than BloombergGPT. |
| [MPT](https://www.mosaicml.com/blog/mpt-30b) | MosaicML | 7B, 30B | MosaicML provides users with multiple sizes and fine-tuned variants that can be trained for a specific industry; licensed for commercial use; fast processing speed. |
| [Claude](http://claude.ai) (not open-source) | Anthropic | ~175B | Top competitor of ChatGPT; one of the best LLMs.  Processes inputs up to 100k tokens; can be accessed through API or designated platform. |
| [Falcon](https://falconllm.tii.ae/) | Technology Innovation Institute | 7B, 40B, 180B | The 180B version is the most parameter-rich open-source LLM to date; ranked first on Hugging Face Open LLM Leaderboard; can be fine-tuned for specific tasks. |
| [Llama 2](https://ai.meta.com/llama/) | Meta | 7B, 13B, 70B | Ranked second on the leaderboard. |



**Name matching Candidates:**

1. Weighted Fuzzy Logic

Pros:

* Help with typos and mis-spelling
* Scalability
* Increased accuracy compared to Fuzzy logic

Cons:

* Unable to handle high levels of precision

1. Multilayer perceptron

Pros:

* Scalability
* Suitable for large data set

Cons:

* Computationally intensive
* Lack of context

1. Transformers

Pros:

* Contextual understanding
* Multilingual support
* Model sequential data efficiently
* Capture dependencies between input tokens

Cons:

* Need pre-process data labeling
* Lack of scalability

1. Salesforce AI - [fuzzy first name matching](https://engineering.salesforce.com/ai-based-identity-resolution-the-key-for-linking-diverse-customer-data/)

* allows companies to select their level of rigor of matching per business case
* Includes nicknames: William - Bill
* MLP, align name similarity scores and name embeddings
* Multilingual DistilBERT
* Soft matching scoring examples

