

Mimi president

aishift

March 21, 2025

Contents

1	Problem	1
2	Solution	2
2.1	Announce	2
2.2	Implementation details	2
2.2.1	Embedding model	2
2.2.2	LLM chat bot	2
2.2.3	Data store	3
2.2.4	Programming language	3
2.2.5	Parsing	3
3	Feature improvements	3

1 Problem

The rapidly developing Cyber Valley project has diverse sources of truth represented in the following resources:

- X.com tweets
- Telegram group chat messages
- Logseq knowledge base git repositories
- GitHub issues

Searching all of them becomes a time-consuming process and requires a simple way of querying all of them at one time.

2 Solution

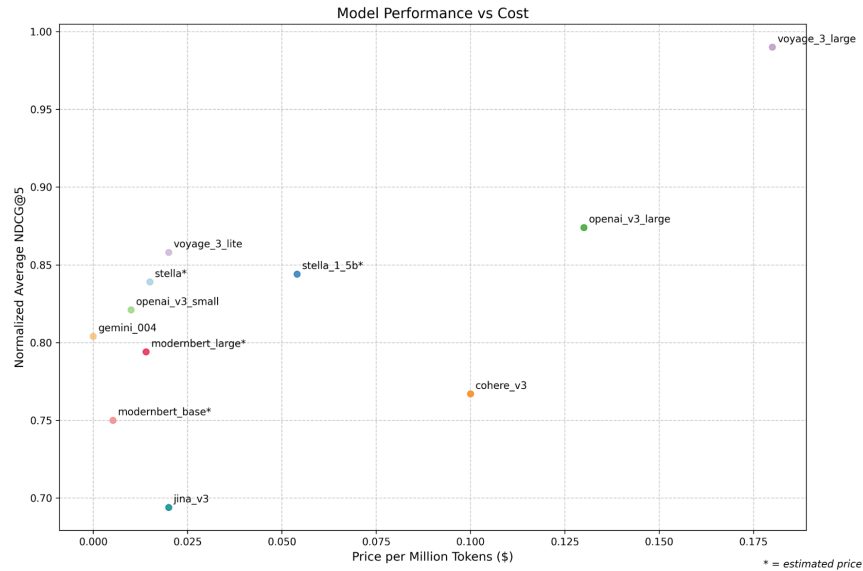
2.1 Announce

Develop RAG over all resources mentioned in the problem statement and provide an LLM-driven chat bot, which allows interactive and free-form querying of all of them at once.

2.2 Implementation details

2.2.1 Embedding model

We keep in mind that in the future it could be great to change the chosen model, but it requires complete recalculation for the whole dataset (because of different dimensions and algorithms in general). To handle this, we will store all source data "as is", so making embeddings will be a question of computation. For the POC we will stick to the OpenAI text-embedding-3-small which is pretty cheap and should work well enough.



2.2.2 LLM chat bot

Our solution is completely model-agnostic, so any provider could be used and switched on the fly.

2.2.3 Data store

We choose Turso as our DBMS; it works perfectly with vector search, scales greatly on HDD drives, and has zero network latency because it's built on libSQL.

2.2.4 Programming language

We will use Python & LangChain for the project because it'll just need glue between IO operations. Rust wouldn't make a visible difference in speed or durability and lacks ready-to-use packages for fast idea testing.

2.2.5 Parsing

1. X.com

We don't know for sure the general required number of accounts, their requirements, and their publicity. So for the start and completely for free, it's possible to use Google news RSS. As an example, here is the RSS feed generated for @levelsio - <https://news.google.com/rss/search?q=site:twitter.com/levelsio+when:7>

2. Telegram groups

We offer to use the Telegram Client API. It requires its own Telegram account but in exchange has access to the whole history of messages (in super groups where it's allowed). The algorithm for adding support for a new group will be the same as adding a new participant to the group. Then we will download all message history (with a given threshold or fully), then listen to new messages and process them as well.

3. GitHub

We can use the Webhooks API to get updates on commits to the LogSeq files and issues.

3 Feature improvements

- Embed media (pictures, video, and audio) as well
- Query and embed provided URLs in the text info
- Include URLs to the initial sources found with RAG

- Allow querying only given resources e.g., "What are the statuses of the current projects with aishift in GitHub issues"