Progress Report

Data Collection and Processing

As one of the representative applications of Web3, Dogecoin follows the same laws as other virtual currencies. Therefore, I selected two datasets: an introduction to cryptocurrency from a Github project and relevant news from RootData.

- **Tutorial Dataset:** This dataset consists of nine white papers on Blockchain and Crypto Basics, covering popular cryptocurrencies like Bitcoin and Ethereum. It contains a total of 148,533 tokens, which are segmented into 79 text chunks, each containing 2,000 tokens with a 100-token overlap between consecutive chunks.
- **News Dataset:** This dataset comprises five recent English news reports on Dogecoin, spanning from February 2021 to September 2024. The texts were scraped from a website using Firecrawl. The dataset contains a total of 4,243 tokens, which are divided into three text chunks of 2,000 tokens each, with a 100-token overlap between consecutive chunks.

Before feeding the data into the GraphRAG model, a preprocessing step is required. For the *Tutorial* dataset, we simply converted the data into a text file by concatenating the individual files, as they were already clean and well-formatted PDFs. For the *News* dataset, we utilized an HTML tag-selector to remove hyperlinks and images. Additionally, we ensured that each article included its title preceding the main content, as shown below.

```
[ARTICLE 1: Developers Compete To Bring Smart Contracts To Dogecoin Ecosystem - "The Defiant"]

An increasing number of projects are working to bring smart contract functionality to the Dogecoin ecosystem. On Sunday, Laika, a Layer 2 solution for Dogecoin, ...
```

Knowledge Graph Construction and Optimization

~~TODO: Not implemented yet. Use vanilla prompts.~~

Question Set

Typically, RAG is employed to extract high-level claims and assertions from given texts. To avoid focusing on specific details within the texts, we prompt the LLM to generate a set of questions based solely on a brief description of the Dogecoin datasets. This approach reduces the likelihood of generating questions that are overly specific to a particular article. After providing an introduction, we instruct the LLM to generate N potential user profiles, each associated with M potential tasks. For each task assigned to each user, the LLM generates K questions exploring various aspects of that topic. In our experiment, we set $N = 4$, $M = 3$ and $K = 5$, resulting in a total of 60 questions for evaluation. An example is shown in Table 1.

User	Task	Question Set
Investors	Understanding Dogecoin Basics	What is Dogecoin, and how does it differ from other cryptocurrencies? Who created Dogecoin, and what was the original purpose behind its creation? How has the value of Dogecoin fluctuated over time? Can you explain the technology that powers Dogecoin? What are some key events in the history of Dogecoin that have significantly impacted its development or popularity?
Tech Enthusiasts	Studying Technical Foundations	What specific blockchain technology does Dogecoin utilize? How does the consensus mechanism employed by Dogecoin (i.e., proof-of-work) work? Can you explain the concept of mining Dogecoin? What measures are in place to ensure the security and integrity of the Dogecoin network? How does the supply distribution of Dogecoin (with no hard cap) differ from other cryptocurrencies?

Table 1: Example of evaluated questions on users, tasks and question suites.

Evaluation Metrics

The primary goal in selecting evaluation metrics for comparing different RAG models is to enable intuitive and efficient comparisons. Since there is no standard or predefined correct answer, the chosen metrics should provide objective insights into the quality of the generated responses. Therefore, selecting pairs of conflicting metrics helps to avoid bias. In our evaluation, we use the following pair-wise metrics:

- **Conciseness vs. Information Coverage:** Conciseness requires responses to be brief and to the point, while information coverage demands that all relevant aspects of the query be addressed. These metrics are in conflict because a highly concise answer may omit important details, whereas a fully comprehensive response can become excessively lengthy.
- **Relevance vs. Diversity:** Relevance emphasizes providing answers that are directly aligned with the query, whereas diversity encourages the inclusion of additional information. Striking a balance between these two can be challenging, as extending a response to incorporate diverse information may reduce its focus on the core query.

For each evaluation, we provide a question along with responses from each model, apply the evaluation metrics to the LLM, and ask it to select a winner based on the given criteria. If there is a tie, the evaluation is skipped. ~~To ensure fairness, each test suite is evaluated three times.~~ A sample evaluation prompt is attached below:

```
I am comparing the responses generated by different models for a given question. I need you to evaluate these responses based on specific criteria and determine the winner for each criterion. Please avoid ties unless the responses are very similar.

Original Question: {original_question}

Model A Response: {model_a_response}

Model B Response: {model_b_response}

Evaluation Criteria:
1. Conciseness: The response that is more concise and to the point.
2. Information Coverage: The response that covers the most relevant information.
3. Relevance: The response that is most relevant to the original question.
4. Diversity: The response that provides a more diverse range of information or perspectives.

Please evaluate the responses based on the above criteria and provide the winner for each criterion in the following format. If there is a tie, indicate "Tie":

Conciseness: [Winner (Model A, Model B, or Tie)]
Information Coverage: [Winner (Model A, Model B, or Tie)]
Relevance: [Winner (Model A, Model B, or Tie)]
Diversity: [Winner (Model A, Model B, or Tie)]
```

Results

In this section, we compared the performance of the Nano GraphRAG method and the naive RAG method on question sets. To make the result more straight-forward, we combined the corpus of the tutorial and the news. ~~The global-searched method beats the naive model on most criteria~~, as shown in Figure 1.

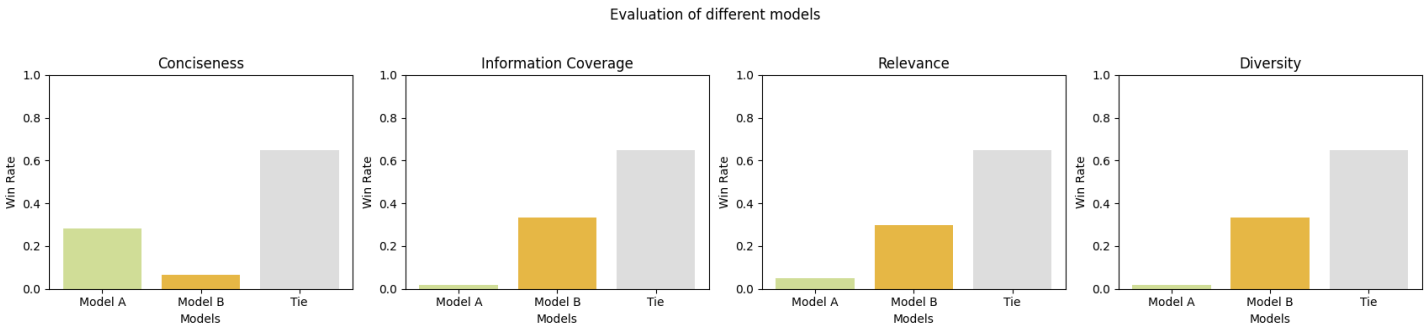

Figure 1: Results of graph and naive RAG models under different criteria