2025

Assignment 2 Report:
Analysis of Prompt Engineering Discussions on Reddit and YouTube

GROUP NAME:

- **Assignment2 Group 44**

TEAM MEMBERS:

- **Swayam Mayankkumar Patel(s3994439)**
- **Julian Schmidt-Heron (s4002485)**
- **Shivani Malik (s4042114)**

Team's Code and Data Repository

GitHub shared with lida.rashidi@rmit.edu.au

GitHub repository is titled:

Group-44-Assignment-2-Choose-Your-Own-Analysis

Find the full assignment code, and all the data files here.

Introduction

This project explores the domain of social media analytics by focusing on digital communities that engage in content creation and discussion around artificial intelligence technologies. In particular, the platforms **YouTube** and **Reddit** were selected as primary sources of data due to their complementary characteristics—YouTube offers a rich comment ecosystem attached to video content with high user activity, while Reddit hosts structured, in-depth discussions within topic-specific subreddits. Together, these platforms provide a multifaceted view of public engagement with emerging AI concepts. The investigation centers on user sentiment, engagement patterns, community structures, and the spread of influence in discussions surrounding **prompt engineering**, a technique used to guide the behavior of artificial intelligence models through carefully structured input prompts.

The motivation behind choosing this topic stems from the rising importance of prompt engineering in AI tools such as ChatGPT, Midjourney, and DALL-E. These tools are widely adopted and discussed across social media, where users often exchange tips, critiques, and personal experiences. While YouTube comment sections tend to host spontaneous, emotion-driven reactions to video content, Reddit provides longer, more technical exchanges within niche communities. By analyzing both platforms, this project aims to uncover a more holistic picture of public discourse surrounding prompt engineering—revealing both surface-level sentiment and deeper knowledge-sharing behavior.

The key questions that guide this investigation include: what are the dominant topics discussed on Reddit and YouTube regarding prompt engineering? What kind of sentiment do users express over time? Which users are central or influential in their respective networks? And how does information propagate among users through reply interactions? This report answers these questions through the systematic application of natural language processing, sentiment analysis, network analysis, and influence modeling techniques.

The scope of the project spans both **textual** and **network-based analysis**. Textual analysis involves preprocessing comments from both platforms, identifying frequent terms, modeling discussion topics using LDA, and tracking sentiment over time using rule-based and probabilistic methods. Network analysis involves building reply networks, computing centrality metrics, detecting community structures, and simulating influence spread using Independent Cascade and Linear Threshold Models.

Deliverables for the project include a cleaned and structured dataset of Reddit posts and YouTube comments, visualizations of word frequencies, sentiment trends, topic models, reply network structures, centrality-based rankings, and influence activation maps. Each analytical component is supported with visual outputs and interpreted using accessible language to facilitate understanding across both technical and non-technical audiences.

The data was collected using the **YouTube Data API** and the **Pushshift Reddit API**, and stored in structured JSON formats. The dataset includes top-level comments and nested replies, with each entry timestamped and attributed to an author, allowing for both temporal and social network analysis.

Overall, this report presents a comprehensive and comparative examination of online discussions around prompt engineering on YouTube and Reddit. The structure follows a step-by-step breakdown of methods and results, concluding with key insights that illustrate how communities behave, how influence spreads, and how sentiment evolves in different digital ecosystems.

Data Collection

In order to conduct a thorough analysis on our topic surrounding prompt engineering, we needed to collect a vast amount of data across multiple social media platforms, so that the analysis we undertook was in depth, and was able to uncover meaningful insights, which might be harder to do with a smaller dataset. We set an ambitious goal of collecting a combined total to 100,000 posts and comments across two social media platforms, YouTube and Reddit. We picked Reddit as a data source as we were already familiar with extracting data from there, and as a second social media platform, we picked YouTube comments as the data was freely accessible and offered a different format of text data to analyse alongside Reddit.

We collected data from Reddit using the Pushshift API, which allowed us to gather both posts and their associated comments from several subreddits. When collecting Reddit data, we initially encountered a problem where the data extraction loop would abruptly stop after approaching nearly 1,000 total posts and their associated comments from a specific subreddit. As a result of this difficulty, we had to slightly change our data extraction approach with reddit. To ensure we were able to extract the goal of 50,000 total posts and comments from each platform, we had to do some more research and find more subreddits related to prompt engineering, so that we could work around the reddit API limit while still collecting enough data. We finally settled on four prompt engineering related subreddits, these included rPromptEngineering, rChatGPTPro, rLocalLLaMA, and rChatGPTPromptGenius. With more subreddits now to collect data from, we were able to work around the Pushshift API restriction by creating a loop that iterated over each subreddit, collecting near to the latest 1000 posts and their associated comments from each subreddit, before moving onto the next subreddit to collect data from before the data extraction limit kicked in. This led to us being able to extract a total of 56,342 combined posts and comments from reddit that we saved in a JSON file, exceeding our initial data collection goal. We found that our reddit data spanned the time range from 2025-03-17 to 2025-05-20, covering just over two months of activity. Different prompt

engineering related subreddits had varying degrees of activity, with some containing more data than others. To find out the exact time range and data distribution across subreddits, we created Python code to read through our JSON file out output the number of posts and comments in each subreddit, the corresponding time range, and the total number of posts and comments in the entire dataset, along with the overall time span. See the exact data distribution, size and time range across subreddits in the below **Figure 1** which shows this information printed out by our python code:

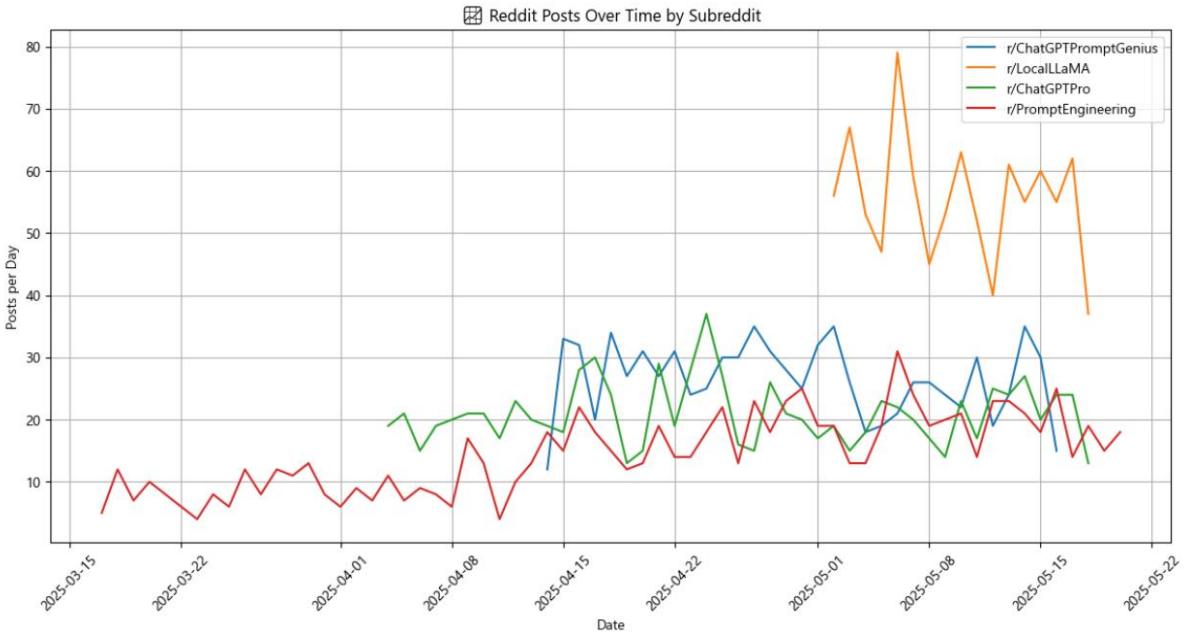
Figure 1: Subreddit Data Distribution and Time Span Summary

```
📊 Subreddit Statistics:
• r/ChatGPTPro: 943 posts | 16730 comments | Range: 2025-04-04 → 2025-05-18
• r/ChatGPTPromptGenius: 877 posts | 6681 comments | Range: 2025-04-14 → 2025-05-16
• r/LocalLLaMA: 944 posts | 21607 comments | Range: 2025-05-02 → 2025-05-18
• r/PromptEngineering: 940 posts | 7620 comments | Range: 2025-03-17 → 2025-05-20

📌 Overall Date Range: 2025-03-17 → 2025-05-20
🧾 Total posts: 3704
💬 Total comments: 52638
📦 Total items (posts + comments): 56342
```

To get a more visual representation of this data, we created a plot of Reddit Posts Over Time by Subreddit, shown in **Figure 2** below:

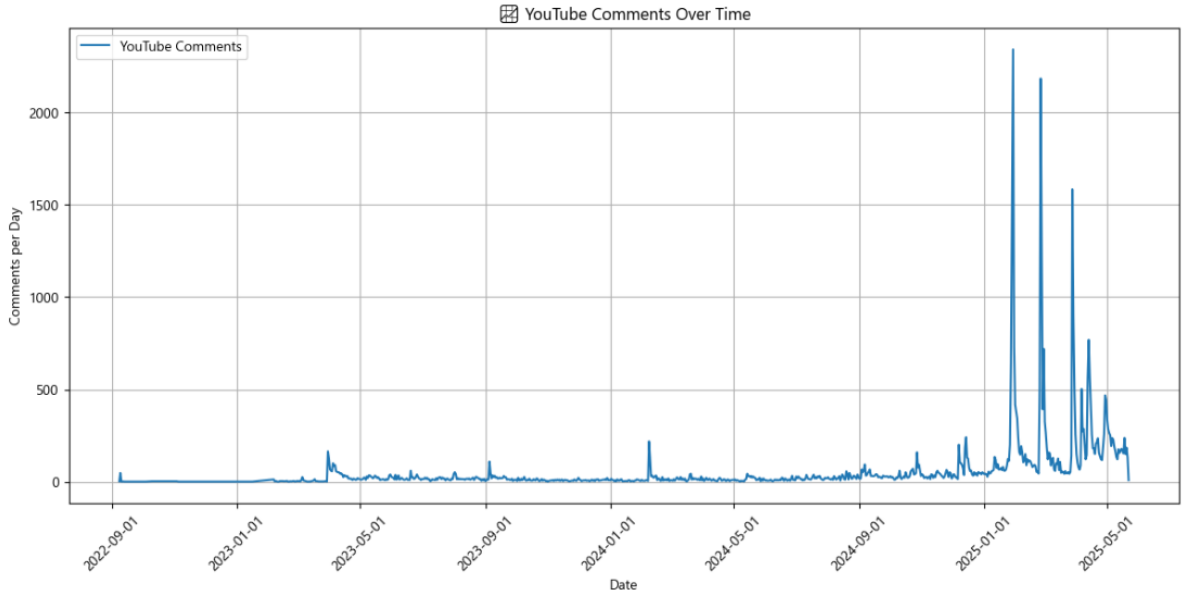
Figure 2: YouTube Comments Over Time



rLocalLLaMA was found to have the most amount of data and spanned a shorter time span of less than a month. The other subreddits spanned a longer time span, with rPromptEngineering spanning the longest. We now had a large amount of reddit data, over 54,000 total combined posts and comments, to conduct our analysis on and gain meaningful insights.

For YouTube, we used the YouTube Data API to gather thousands of video comments related to prompt engineering. These comments were gathered from several video batches that were merged together into one file. Initially we ran into troubles with YouTube data extraction as well, as it was difficult to find YouTube videos that had a large number of comments to extract text data from. Most YouTube videos had a much larger number of views compared to comments, likely because users are generally more passive on the platform—preferring to watch content and move on, rather than engage through comments. So, we had to go to extra lengths in order to extract enough data for YouTube, refining our YouTube searches with different ways of phrasing our search and filtering for the most recent and popular videos, so that many different videos related to our topic of prompt engineering came up, and we were then able to find videos with a larger amount of comments. When extracting our data, as mentioned earlier, we extracted comments from videos in batches, and then merged them together later into one big JSON file that we could use later for our analysis, and to create our reply networks. Due to issues with data mining limits in the YouTube Data API, we needed to create multiple YouTube API accounts to iterate across during data extraction, so that we were able to extract our large amount of data in a timely manner. After completing YouTube data extraction, we were able to extract a total of 48,796 total YouTube comments spanning the time range from 2022-09-08 to 2025-05-22, covering nearly three years of activity. The detailed distribution of YouTube Comments Over Time can be found below in **Figure 3**:

Figure 3: YouTube Comments Over Time



This was a much larger time frame than our reddit data, but was expected due to the smaller amounts of comments found on YouTube videos compared to reddit posts, so we were overall happy that we were able to extract such a large amount of data at all. Also, we didn't quite reach our goal of a combined total of 50,000 YouTube comments, but given the difficulties in extracting YouTube data and the fact that we had gotten even more reddit data than expected, we were satisfied with this amount of YouTube data extracted, which led to us having more

than 105,000 combined posts and comments spread out across two separate social media platforms, more than enough data conduct our analysis on and gain meaningful insights.

Exploratory Data Analysis

For our Exploratory Data Analysis, we began by exploring the amount of data available and its temporal distribution. On Reddit, we observed that the number of posts varied across subreddits. Some subreddits had more frequent discussions than others. We found that the earliest posts dated back several months while most of the activity was concentrated in the recent months. We created line plots showing the number of posts per day in each subreddit. On YouTube, we analysed the number of comments posted each day. We created time series plots showing the daily trend of comment activity. This helped us understand the engagement levels over time. These line plots of posts over time were shown and discussed earlier in the Data Collection section above, in **Figure 2 and 3**.

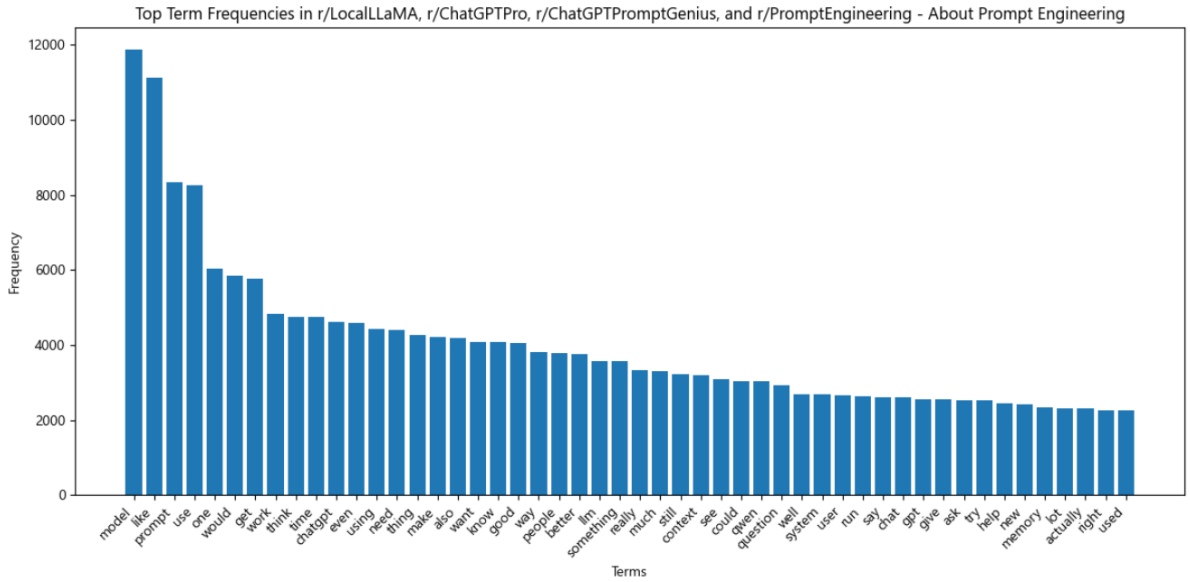
Data Preprocessing

For our Data Preprocessing, we performed a series of preprocessing steps to prepare the text data for analysis. We applied the same preprocessing steps to both the reddit and YouTube data to ensure our final data was consistent across platforms. This consistency was important to make sure the data was treated the same way prior to analysis, allowing for fair comparison and meaningful insights across platforms. We normalized the text to ensure consistency. This included converting all text to lowercase. We removed stopwords such as "the" and "is" which are common words with little analytical value. We removed punctuation marks and symbols. We also removed numeric digits and any links or web addresses. The text was tokenized which means it was split into individual words. Then the words were lemmatized. Lemmatization involves reducing words to their base form. For example, the word "running" becomes "run". This step ensures that similar words are treated as the same during analysis. These preprocessing steps helped clean and standardize the data, ensuring it was in an optimal format for reliable and insightful analysis.

Term Frequency Analysis

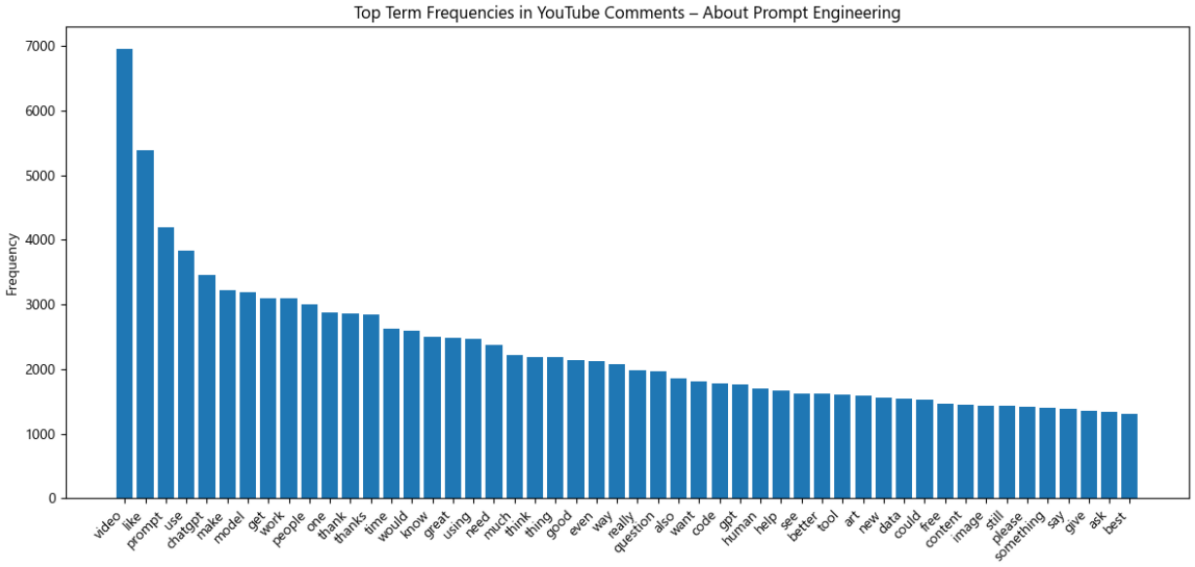
As a first step of simple analysis, we counted the frequency of each word in the text dataset across both social media platforms to identify the most common terms. On Reddit, the most frequently used terms included words like model, prompt, question, user and chatgpt. These terms reflect the central themes in discussions about prompt engineering. “Model” and “prompt” where amongst the top 3 words in the entire dataset, highlighting their central role in conversations and emphasizing that a lot of what is talked about is likely focused on understanding, designing, and interacting with AI models through prompt-based interfaces. We plotted the top 50 words in our reddit text dataset in a visualisation to get a better understanding of how often the most popular word appear and their distribution in comparison to others. The top 50 words plotted together in this visualisation can be seen below in **Figure 4.**

Figure 4: Top Term Frequencies in r/LocalLLaMA, r/ChatGPTPro, r/ChatGPTPromptGenius, and r/PromptEngineering - About Prompt Engineering



On YouTube, a similar trend was observed. The most common words were related to prompt design, language models, and practical usage of artificial intelligence tools. Amongst the top 10 words were included words such as prompt, chatgpt and model, also further reinforcing the idea that user interest and discussion across platforms consistently centred around the mechanics and applications of prompt engineering. We visualized the top fifty terms in the YouTube data as well using bar plots, shown below in **Figure 5**, to give us a clear idea of the vocabulary commonly used in the YouTube communities.

Figure 5: Top Term Frequencies in YouTube Comments – About Prompt Engineering



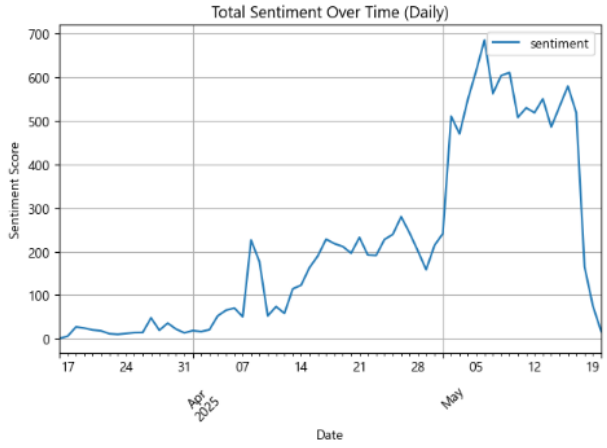
Sentiment Analysis

To properly begin our analysis into prompt engineering discussions across reddit and YouTube communities, we begin with conducting sentiment analysis on both of our social media platforms over time. We used the Vader sentiment analyser to calculate the sentiment of each comment or post. Using this method, we were able to gauge the general mood across platforms, observe how sentiment varied over time, and identify any trends or shifts in public perception related to prompt engineering. Vader provides four sentiment scores which are positive, neutral, negative, and compound. The compound score is a weighted sum that represents the overall sentiment. We plotted the sentiment scores over time to observe how public sentiment changes.

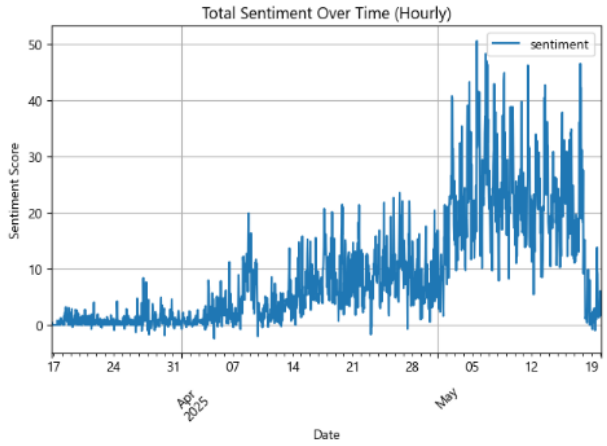
Reddit Data

For our reddit data, we first looked at the total sentiment over time, both daily and hourly, in order to try and capture the overall volume of emotion expressed by users, which reflects the collective sentiment strength in a given period. From **figure 6** below, we can explore the overall volume of emotion expressed by users across the prompt engineering related subreddits across the timeframe, in regard to their attitude towards what they think about prompt engineering at the time.

Figure 6: Total Reddit Sentiment Over Time — Aggregated by Day and Hour



<Figure size 1000x500 with 0 Axes>



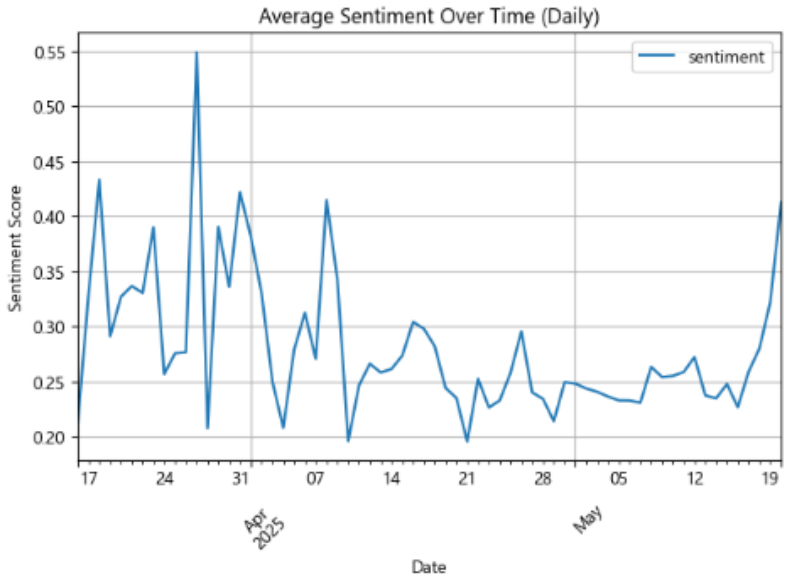
Based on the Total Reddit Sentiment Over Time — Aggregated by Day, we can try and gain some insights about overall volume of emotion expressed by users over time, and try and map the increases or decreases in sentiment to potential real-world events. On Reddit, we observed that the average sentiment fluctuated between neutral and positive, and there were occasional spikes in both positive and negative sentiment. We can see that sentiment begins lower around just over 0, indicating a more neutral initial sentiment. Slowly over time, the sentiment starts to gain an increasing trend, with small ups and downs in sentiment, but slowly increasing. Around the 9th of April, the sentiment spikes by a larger than normal increase, then drops down again, before slowly rising again, with a peak of top sentiment around 5th May, before slowly dropping again and then suddenly plunging from 17th to 19th of May.

Looking at Total Reddit Sentiment Over Time — Aggregated by Hour, we can see that roughly the same trends are still present, but we can see a much more fluctuating sentiment, due to the hour-by-hour changes. The hour-by-hour trends roughly resemble the same shapes as the average, but each hour, the sentiment can change more drastically. One hour the sentiment might be more positive, then it might drastically drop the next hour to negative, due to the fluctuations of different posts, with posts more often being positive, but still a lot of posts can be less positive, neutral, or even slightly negative. Thats why it is useful to see both the daily and hourly trends, to know that while the daily sentiment might look good, it doesn't mean that every hour this positive sentiment is present.
If me try and map these sentiment trends to recent real-world news related to prompt engineering or AI related news, we can see that some of the shifts in sentiment seem to match

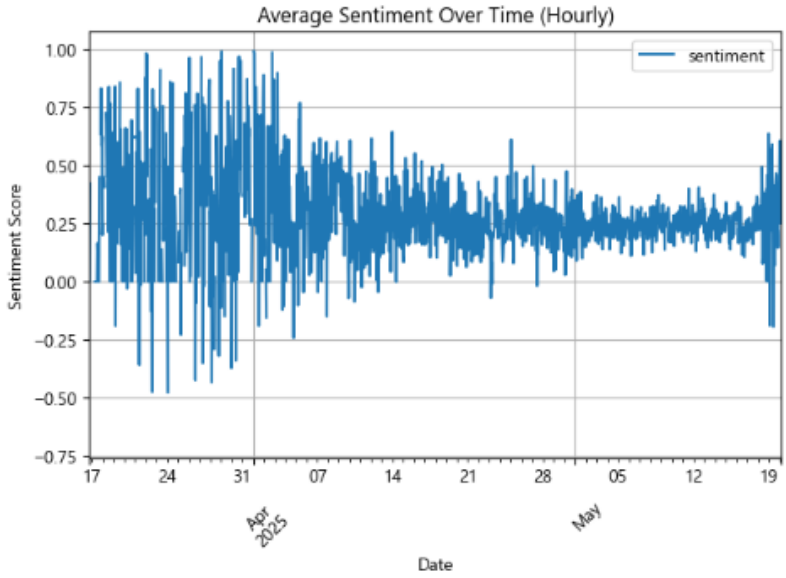
up with real events. The spike around April 9th likely links to Google releasing a big prompt engineering guide that got a lot of attention. The peak on May 5th might be tied to hype around the OnePlus 13s and its AI features. Then the drop from May 17–19 seems to line up with OpenAI's Orb device, which raised privacy concerns and got mixed reactions.

However, there is a potential that the number of reddit posts/comments each day had an impact on the total sentiment each day. To address this, we also look at the average sentiment over time, both daily and hourly, which will indicate the general mood or tone of the community by showing the typical sentiment per post or comment. See **figure 7** below, where we track the Average Reddit Sentiment Over Time — Aggregated by Day and Hour

Figure 7: Average Reddit Sentiment Over Time — Aggregated by Day and Hour



<Figure size 1000x500 with 0 Axes>



Looking at the Average Reddit Sentiment Over Time by day, we can see a noticeable difference in sentiment trends. This is because we are looking at the average sentiment per day, which

indicates the general mood or tone of the community by showing the typical sentiment per post or comment. We can see large changes in sentiment overtime, with one day having a higher sentiment, and then the next dropping by a far margin. But overall, sentiment was higher in the more distant months with a peak in sentiment on the 28th of March, while slowly dropping and then rising again sharply towards the 19th of May, conflicting with the results from the average sentiment.

Looking at the Average Reddit Sentiment Over Time by hour, again we have the same findings that the same trends are still present, but we can see a much more fluctuating sentiment, due to the hour-by-hour changes.

If we again try and map these sentiment trends to recent real-world news related to prompt engineering or AI related news, we can see that the peak in sentiment on the 28th of March may again be related to the release of Google's 69-page prompt engineering guide, which sparked widespread interest and positive discussions across AI communities. The Peak in sentiment on the 19th of May after a period of lower sentiment could be related to Microsoft's announcement of the Prompt API in Edge, potentially generating renewed enthusiasm among users.

YouTube Sentiment Analysis

After performing sentiment analysis on the collected YouTube comment data, a number of interesting observations emerged regarding the emotional tone and engagement patterns of users who participated in discussions related to the chosen topic. The sentiment analysis was carried out using the widely adopted VADER sentiment scoring tool, which is well suited for analyzing social media-style texts due to its ability to handle emojis, slang, and casual expressions. Each YouTube comment was evaluated for its positive, neutral, and negative sentiment components, and a compound score was calculated to represent the overall sentiment polarity.

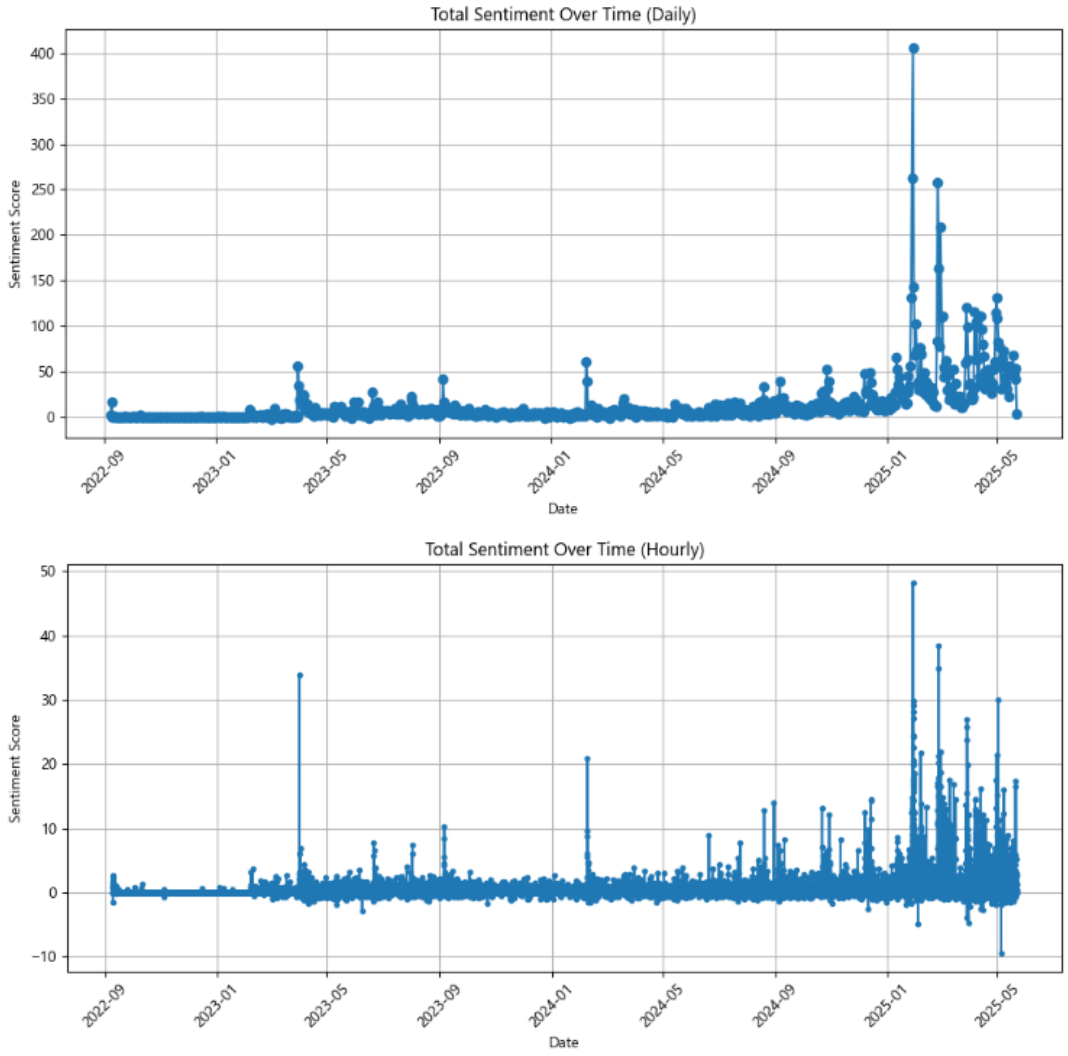
Platform Characteristics and Sentiment Volatility

When comparing the sentiment trends observed on YouTube to those identified on Reddit, it became clear that while both platforms demonstrated fluctuating patterns of emotional response over time, YouTube exhibited relatively greater volatility in its sentiment scores. This increased sentiment variability can potentially be attributed to the inherent characteristics of the YouTube platform. YouTube comments are generally much shorter in length, more casual in tone, and often reflect instant emotional reactions to video content rather than thoughtful, deliberated commentary. On the other hand, Reddit tends to host longer and more technically detailed discussions, especially within topic-specific subreddits, which may contribute to a more stable and consistent sentiment profile over time.

Total Sentiment Over Time

Figure Eight presents the total sentiment score accumulated across all YouTube comments on a daily and hourly basis. This visualization reveals the magnitude and intensity of sentiment expression during different periods of time. While some days demonstrated a noticeable increase in sentiment intensity, others showed sharp drops, likely caused by sudden bursts of emotional or controversial engagement. Peaks in the total sentiment graph often corresponded with periods of higher video engagement or when trending content related to the topic had recently been uploaded. The hourly view, while more granular, highlighted micro-patterns of emotional reaction that may be tied to viewer habits and time zone-driven activity spikes.

Figure 8: Total YouTube Sentiment Over Time — Aggregated by Day and Hour

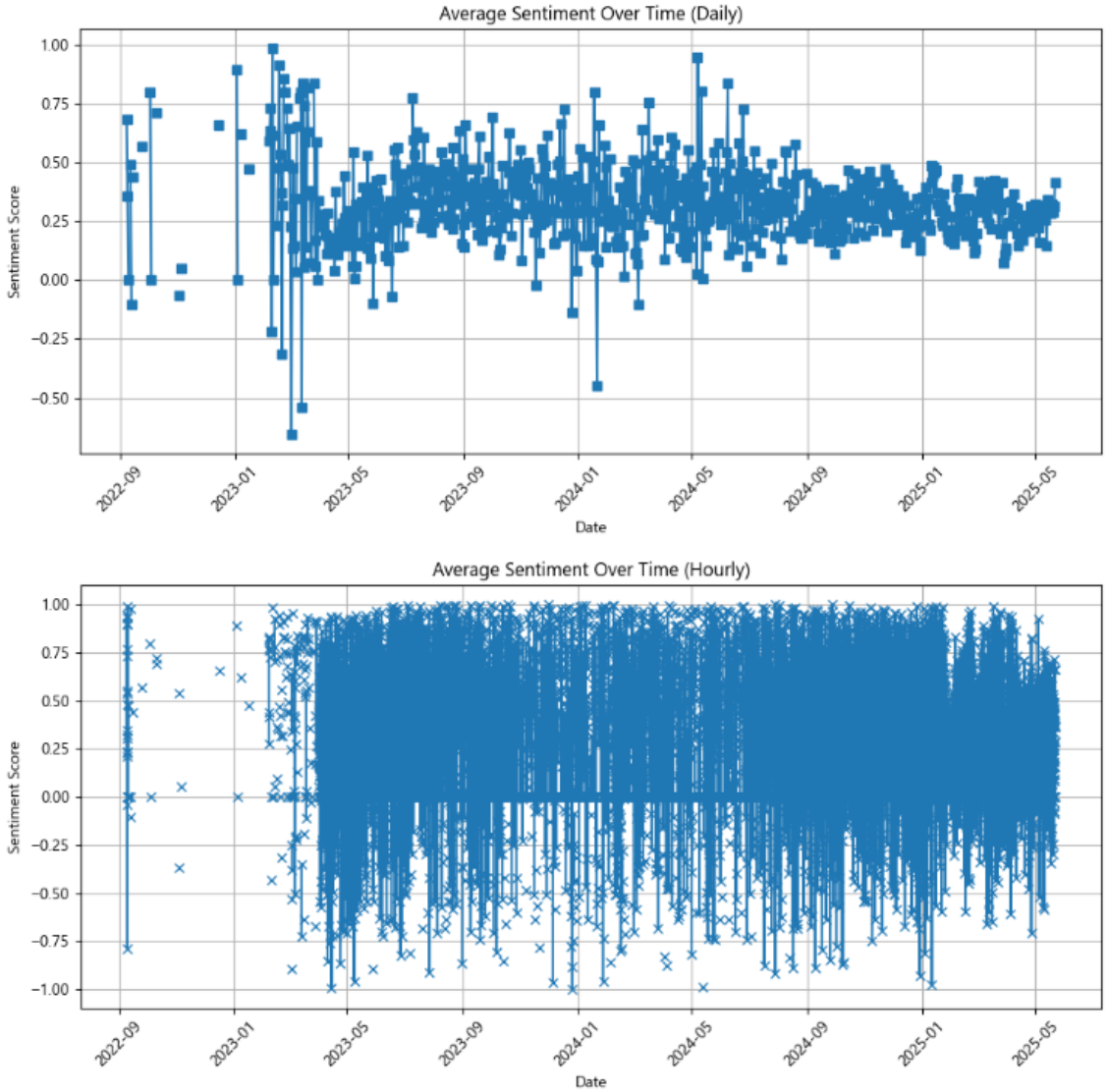


Average Sentiment Over Time

In Figure Nine, the average sentiment over time is shown, once again aggregated both by day and by hour. This visualization smooths out the total comment volume to reveal the underlying

tone of user responses. The average sentiment score across YouTube comments fluctuated above and below a neutral midpoint. On certain days, the overall sentiment skewed more positively, possibly in reaction to new feature updates or community appreciation of educational content. Conversely, negative sentiment spikes may indicate critical responses to perceived failures, unpopular decisions, or miscommunications by the video creator or platform itself.

Figure 9: Average YouTube Sentiment Over Time — Aggregated by Day and Hour



Comparative Summary

The sentiment patterns on YouTube and Reddit, while both informative, illustrated differences in platform dynamics. YouTube's shorter and more spontaneous comment style resulted in a sentiment trend that was more dynamic and sensitive to sudden events. Reddit's more discussion-driven format, in contrast, produced steadier sentiment movements. Despite these differences, both platforms showed evidence of community engagement and emotional investment in the topic of interest, reinforcing the importance of multi-platform analysis when attempting to capture public perception at scale.

Topic Modelling Using Latent Dirichlet Allocation

To analyse the different types of topics that were being discussed in our prompt engineering related social media data from across both reddit and YouTube, we applied topic modelling using Latent Dirichlet Allocation, which is a probabilistic method for discovering hidden themes in a collection of documents. Before running our topic modelling, we defined our choice of hyperparameters to be consistent across social media platforms. We set the number of topics to five so that we were able to find a sizeable number of topics to identify and analyse, while at the same time not being overwhelmed by a very large number of topics. We also limited the vocabulary size to the most frequent fifteen hundred terms per topic because this helps reduce noise from less relevant or overly rare words, making the resulting topics more coherent and easier to interpret. We also set the number of words to display per topic to 15, to have a good representation of what the topic is about while keeping the output concise and manageable for analysis.

The model identified five major topics in both Reddit and YouTube datasets. Each topic was represented by a list of words that frequently occurred together. For example, one topic included words like model, prompt, input, response, and language. Another topic included terms like user, system, token, data, and output. We visualized these topics using word clouds where larger words represent higher importance. We also used an interactive visualization panel created using pyLDAvis to explore the relationships between the topics.

Reddit Data

After running the Latent Dirichlet Allocation topic modelling on our reddit text data, we identified five major topics. As an output of our topic modelling, we were given the top 15 words that best describe our five identified topics. These top 15 words that best describe our five identified topics are:

Topic 1:
like thanks would one work use using really free got good also great llama get

Topic 2:
model get think much want need thing still know time one like would code make

Topic 3:
like chatgpt human real language people question feel one self someone even idea word system

Topic 4:
qwen pro that run gemini stuff lol gpu coding good better vram there cpu look

Topic 5:
prompt model use context user output llm token using example text tool data system research

Using research as well as our own knowledge, we were able to further interpret and describe these topics:

Topic 1:
This topic seems to focus on users expressing appreciation, sharing useful tools or experiences, and discussing how they are using models like LLaMA in practical ways.

Topic 2:
This topic appears to centre around general thoughts and opinions on models, including what users think they need or want from them, as well as some coding-related discussion.

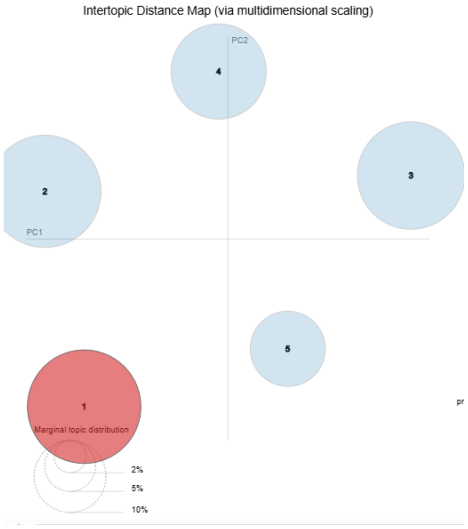
Topic 3:
This topic is more focused on human-AI interaction, with users discussing how ChatGPT responds, how it feels "human," and exploring ideas about self and language.

Topic 4:
This topic mainly discusses hardware and model performance, especially around running models like Qwen or Gemini, and includes some humour and casual comparisons of GPUs and CPUs.

Topic 5:
This topic is clearly centred around the technical aspects of prompt engineering, including prompts, context, tokens, system design, and the use of different tools and models.

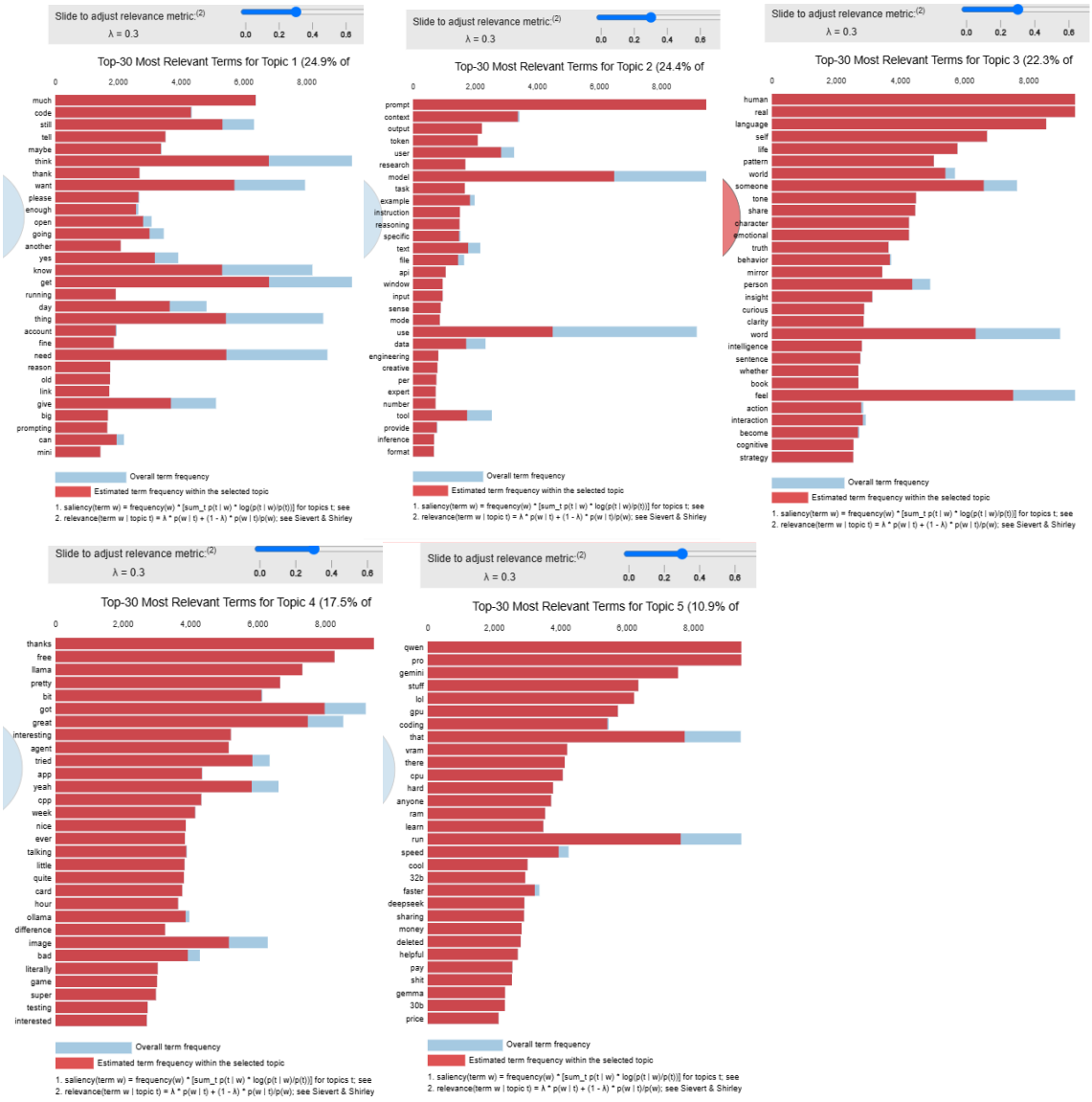
After completing our topic modeling for our reddit data, we use the Interactive pyLDAvis Panel to display our topics and their identified words. In the Intertopic Distance Map section of the display (on the left) each circle represents a topic, and the size of the circle indicates how prevalent (or common) that topic is across the entire dataset. We can see that topic 1 is the largest and most prevalent topic, with the topics slowly decreasing in size from topic 2, 3, 4, with topic 5 being the smallest. See the size distribution of these topics in **figure 10** below.

Figure 10. (Reddit) Topic Size Distribution in Intertopic Distance Map from pyLDAvis



Looking at the right side of the display, the relevance metric, the words in each topic are displayed, with The λ (lambda) slider controlling the relevance metric used to rank words in the topic. At λ = 1, the top words shown are those that are most probable in the topic, and at λ = 0, the top words are those that are more exclusive or unique to that topic (i.e., words that help distinguish it from others). We will select λ = 0.3 to have a balance between most probable and unique words in each topic. The red bars represent estimated term frequency within the selected topic and blue bars show the overall frequency across the entire dataset. These words for each topic can be found in **Figure 11** below.

Figure 11. (Reddit) Top Relevant Words per Topic Identified by LDA Using pyLDAvis (λ = 0.3)



We can also display the word clouds for each topic in our reddit prompt engineering data to get a good visualisation of our topics. In a word cloud, the size of each word reflects its importance or weight within the topic — the more prominent a word is in the topic, the larger it will appear in the cloud. This makes it easy to quickly identify the most representative words associated with each topic. These word clouds can be seen below in **figure 12**.

Figure 12. (Reddit) Word Clouds Representing Top Words in Each Topic from Reddit Prompt Engineering Data

Summary of Reddit Topic Modeling Observations

Reddit discussions on prompt engineering show a stronger focus on technical depth, tool use, and philosophical reflections. Topics one and five highlight practical usage and prompt engineering mechanics, while topics two and three explore user opinions and AI-human interaction. Topic four stands out with detailed hardware comparisons, reflecting Reddit's highly informed and detail-oriented audience.

YouTube Data

After completing the topic modeling on our YouTube prompt engineering dataset, we identified five major topics using Latent Dirichlet Allocation. Each of these topics consists of a set of terms that frequently appear together in comments, indicating underlying themes being discussed by users across the YouTube platform. To visualise and understand these topics, we used the interactive pyLDAvis interface, which shows both topic relevance and term prominence.

Figure 13. (YouTube) Topic Size Distribution in Intertopic Distance Map from pyLDAvis

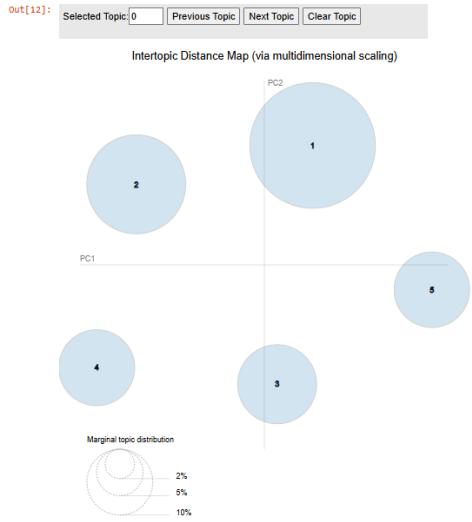
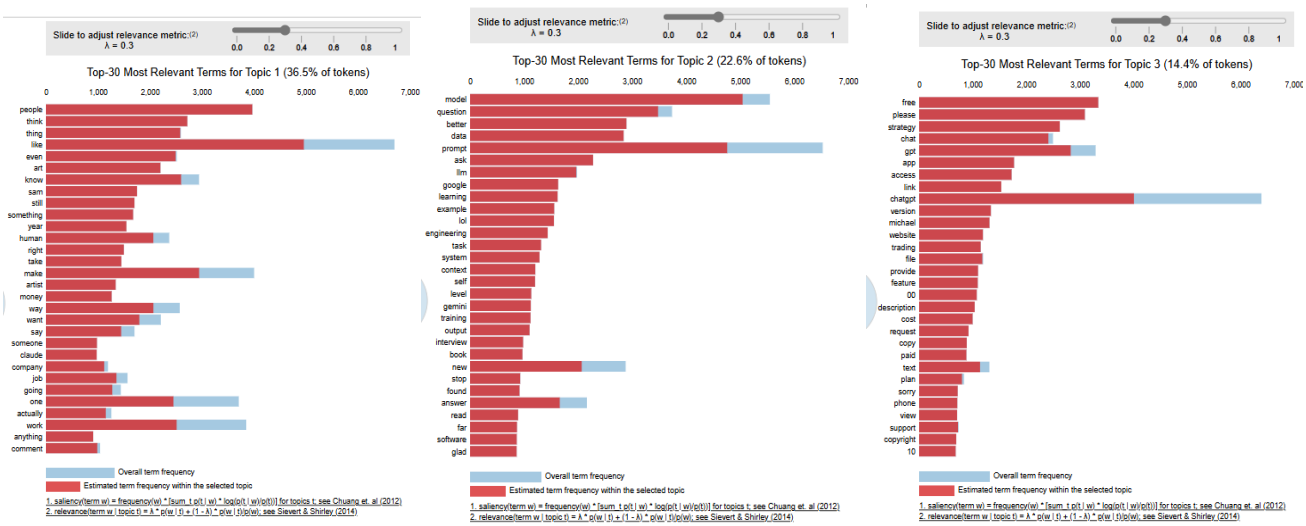
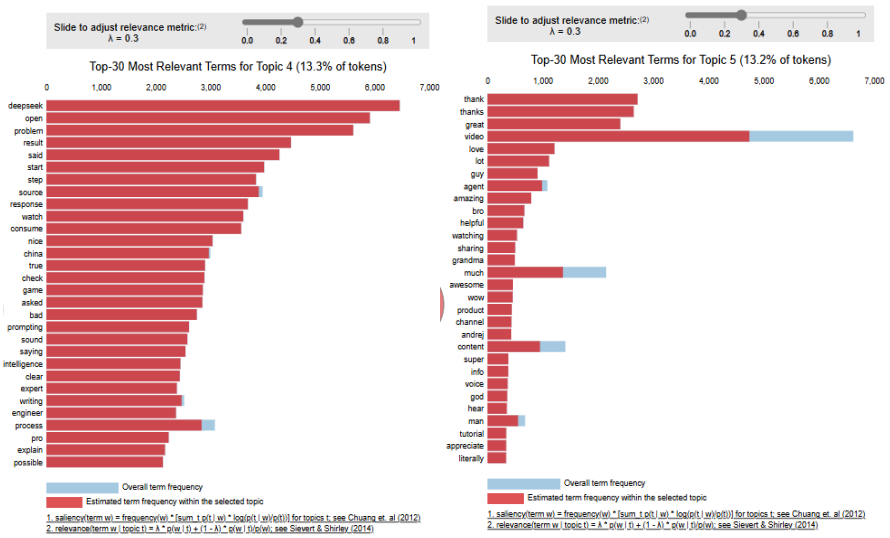


Figure 14. (YouTube) Top Relevant Words per Topic Identified by LDA Using pyLDAvis (λ = 0.3)





The five topics and their most relevant words were extracted and visualised at a λ value of 0.3, which strikes a balance between showing the most frequent words and the most exclusive or unique words per topic. These top relevant terms are shown in **Figure 14. (YouTube) Top Relevant Words per Topic Identified by LDA Using pyLDAvis (λ = 0.3)**. Each panel shows the 30 most relevant terms for each topic, with red bars representing estimated term frequency within the selected topic and blue bars showing the overall frequency across the entire dataset.

Topic 1: Viewer Reactions and General Thoughts

This topic is heavily centered around users expressing opinions and casual reflections. Frequent words such as *people, think, thing, like, know, right,* and *want* suggest that viewers are sharing general ideas, opinions, or feedback. The presence of terms like *actually, going,* and *work* indicate discussions around how well certain prompts or tools function. This topic likely includes viewers reacting to what was said in the video, discussing what they think about the prompt or AI model demonstrated, and sometimes debating broader ideas like fairness, jobs, or ethics.

Topic 2: Prompt Design and Model Engineering

The second topic is clearly technical in nature. Terms like *model, data, prompt, llm, system,* and *context* highlight that this topic covers engineering-level discussions around how prompts are constructed and how models respond to them. Mentions of *task, output, gemini, training,* and *engineering* suggest conversations focused on experimentation, prompt testing, and understanding system behavior. There are also references to *google* and *book*, implying external resources or examples being cited. Overall, this topic focuses on users analysing how AI models work, how to prompt them properly, and the results they generate.

Topic 3: App Access and Feature Requests

Topic three is very focused on access to AI tools, features, and associated limitations. Common terms such as *free, strategy, app, access, link, version, feature,* and *plan* suggest users are frequently discussing how to access a model, whether it is free or paid, and what capabilities are included. Terms like *chatgpt, cost, copy, request,* and *provide* reinforce the idea that this topic relates to functional use of tools, especially with concerns around subscription models or gated features. Users might be commenting to ask others for access help, or to complain about restricted versions.

Topic 4: Model Demos and Explainers

This topic focuses on how videos are presented and the types of content being shared. Common terms such as *deepsseek, open, problem, result, start, said, source,* and *response* show that users are likely responding to walkthroughs, tutorials, or live demos. Additional terms such as *consume, watch, video, expert, engineer,* and *explain* suggest viewers are engaging with explanatory videos that break down complex ideas. The presence of phrases like *sound, prompt, intelligence,* and *writing* also indicates a focus on understanding how models operate during prompts and responses.

Topic 5: Viewer Gratitude and Appreciation

This final topic stands out as being strongly emotional and appreciative in tone. It includes many words of thanks and positivity like *thank, thanks, great, video, love, amazing, awesome, appreciate,* and *literally*. These terms suggest that viewers are thanking the content creators for useful content or helpful tutorials. Phrases like *super, god, tutorial, bro, agent,* and *channel* imply casual, friendly engagement. This topic mostly captures the gratitude and appreciation aspect of YouTube, where viewers commonly leave supportive or thankful comments for informative or entertaining videos.

Summary of YouTube Topic Modeling Observations

From these five identified topics, we can clearly see a combination of technical, practical, and emotional discussions occurring within the YouTube comment sections on prompt engineering. Topics two and three are highly technical, reflecting the audience's interest in prompt structure, system design, and accessibility of tools. Topic one and four serve as more general and interpretive spaces, where viewers reflect on what they learned or how they experienced the content. Topic five stands out for showing high engagement and appreciation, something more unique to YouTube's community culture compared to Reddit.

Compared to Reddit, which had more granular technical topics with sharper distinctions between tool-related and philosophical discussions, YouTube's topics are more conversational and centered around tutorial-style videos. This reflects YouTube's nature as a visual and

explanatory platform, where viewers often come to learn, ask questions, and show support rather than deep-dive into theory.

Figure 15. (YouTube) Word Clouds Representing Top Words in Each Topic from Reddit Prompt Engineering Data

To complement the term relevance visualisations shown earlier in Figure 14, we also created word clouds for each of the five topics discovered in our YouTube LDA model. These are displayed in **Figure 15. (YouTube) Word Clouds Representing Top Words in Each Topic from Prompt Engineering Data**.

In these word clouds, the size of each word reflects its importance or weight within the topic. Larger words are more significant in defining the topic. Word clouds offer a fast and intuitive way to understand what each topic is about at a glance, especially for non-technical audiences.

Observations from Word Clouds (Figure 15)

- **Topic 1** emphasizes technical explanations and demonstrations. The most prominent words such as *deepsseek, open, prompt, source, result,* and *problem* suggest users are reacting to detailed content on how certain models function. The appearance of *said* and *start* reinforces that this topic likely revolves around walk-throughs or step-by-step guides often explained in the videos.
- **Topic 2** features terms like *chatgpt, video, free,* and *access,* which points toward users discussing availability and access to AI tools like ChatGPT. The prominence of the word *free* may reflect common questions about pricing or subscription restrictions when trying to use these tools.
- **Topic 3** highlights conversational and reaction-based words. Large terms such as *people, like, think, know,* and *make* suggest this topic captures general opinions, reactions, and discussions from viewers after watching a video. It reflects a space where viewers reflect on the topic presented or respond to other comments.

- **Topic 4** is focused on prompt-related techniques. Words like *prompt, model, use,* and *question* dominate the cloud, indicating a strong focus on prompt design, model behavior, and possibly experimentation with input and output formatting. This confirms that technical interest in prompts and system responses is a major theme.
- **Topic 5** clearly represents appreciation and positive feedback. Dominant words such as *thank, video, great, thanks,* and *love* show that this topic is full of gratitude. This supports our earlier interpretation that YouTube users often leave thankful messages after learning something useful or enjoying a piece of content.

These word clouds offer a powerful visual confirmation of the main themes identified through our topic modeling. They reinforce the idea that YouTube users engage in a mix of technical exploration, accessibility concerns, general reflection, and enthusiastic support. Compared to Reddit, where discussions are more dense and topic-specific, YouTube's comment sections reflect a wider variety of user emotions and reactions, including casual feedback and high levels of gratitude.

Social Network Analysis

In this section of our analysis, we aimed to conduct social network analysis to identify the most influential users in our network across both social media platforms. We also take a look at the structural properties of our networks as well. Because prompt engineering knowledge is critical for effectively using AI models, identifying influential users helps to find the users that are shaping best practices.

In order to conduct our social network analysis, we constructed reply networks for both Reddit and YouTube. In these networks each node represents a user and each directed edge represents a reply from one user to another. These reply networks were created for both YouTube and Reddit data in python by iterating through the JSON files that contained all the text data for each social media platform, to extract user interactions and identifying authors and their replies. These reply networks where then saved as GraphML files that we would conduct our future SNA on. After creating and saving these files, we found that our 'reddit_reply_network.graphml' file for reddit had 17810 nodes and 43163 edges, indicating a highly interactive and moderately dense network. Our 'youtube_reply_network.graphml' for YouTube had 14505 nodes and 14450 edges, showing that there was almost the same amount as users as replies, and indicating that users on YouTube are much less interactive with way less replies, leading to a much sparser network.

When beginning with our SNA, we removed isolated nodes and retained only the largest connected component for analysis. We calculated three centrality metrics to identify influential users. These metrics were degree centrality, eigenvector centrality, and Katz centrality. Degree centrality is a more basic approach and measures how many direct connections a user has, Eigenvector centrality gives more importance to connections with already influential users, while Katz centrality accounts for both direct and indirect connections with a damping factor. Using these 3 different factors will ensure we can look for influential users from a range of different approaches, as each approach can have its benefits

and limitations. Users that consistently appear amongst the most influential users across all centrality metrics can be flagged and identified as the most influential and active users in the network. In the context of our social media data based on prompt engineering, we can use these insights on the most influential users to indicate the users that are likely the best source of information for the prompt engineering communities across both social media platforms. These users are likely spreading the latest prompt related news throughout our networks, sharing info on the latest way to craft the best prompts amidst the constant changing state of the AI industry, and answering lots of questions from other users about prompt engineering. After running our degree centrality code, we outputted the top 5 users for each centrality metric across social media platforms. The top 5 users per centrality metric for our Reddit reply network can be seen below in **figure 16**, and top 5 users per centrality metric for our Reddit reply network can be seen in **figure 17**.

Figure 16. (Reddit) Top 5 Users for Each Centrality Metric: Degree Centrality, Eigenvector Centrality, and Katz Centrality

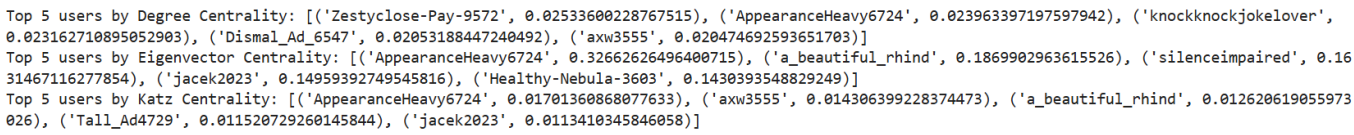
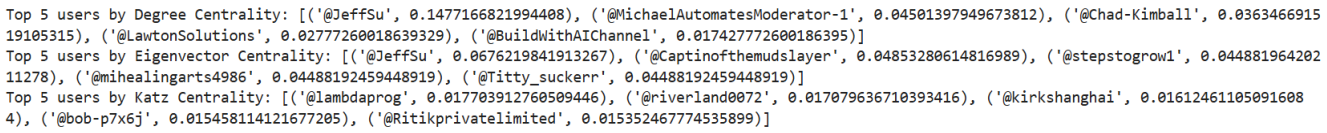


Figure 17. (YouTube) Top 5 Users for Each Centrality Metric: Degree Centrality, Eigenvector Centrality, and Katz Centrality



Looking at the top users in our social networks, we can see that on Reddit, the user AppearanceHeavy6724 stands out as the most influential user, ranking first in both Eigenvector and Katz centrality, and second in Degree centrality. Their high Eigenvector score of 0.3266 suggests they are well-connected to other influential users, positioning them as a central figure in the Reddit prompt engineering discussions.

On the other hand, for YouTube we can see that for YouTube, the user @JeffSu consistently ranks first across both Degree and Eigenvector centrality, with a Degree score of 0.1477, indicating a high volume of direct interactions. However, this user does not appear in the top Katz centrality rankings, suggesting their influence is more direct and less about indirect network reach. This contrast reflects YouTube's overall sparser interaction structure compared to Reddit.

We also plotted histograms (with log-scaled frequencies) to visualize the distribution of centrality scores, to help us gain insight into whether power is equally distributed across users, or if a few users dominate the network. Most users had low centrality scores which means they were not very influential. A few users had very high scores which indicates they played a central role in the network, and that that influence is concentrated in a few users. This was consistent across both YouTube and Reddit data. The YouTube histograms showing centrality

scores (See this in **Figure 18**) show a sharper peak near zero in degree and eigenvector centrality, indicating a highly skewed structure where a few nodes dominate. This could be due to most users on YouTube not ever commenting on videos (including us!) while a select few users like to comment on almost all videos they watch. In contrast, the Reddit histograms showing centrality scores (See this in **Figure 19**) display a slightly flatter distribution, suggesting that the influence of users is a little more evenly spread across the top users, which might be people such as the moderators that input more into their respective subreddit.

Figure 18. YouTube Centrality Score Distributions (Log-Scaled Frequencies): Degree Centrality, Eigenvector Centrality, and Katz Centrality

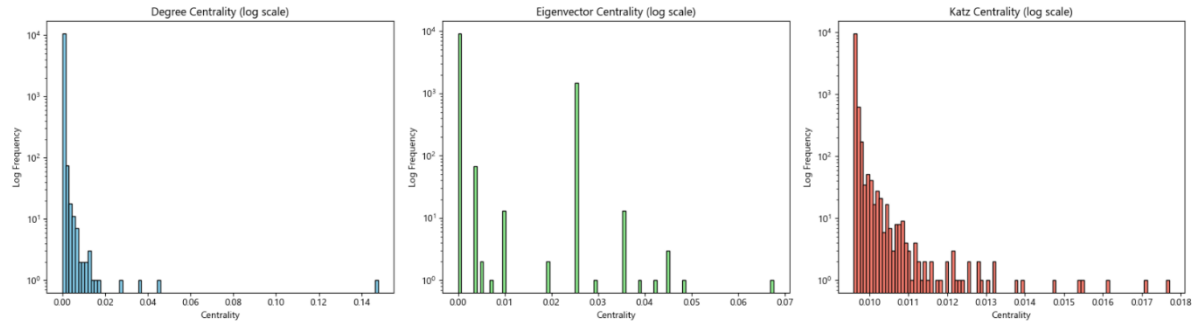
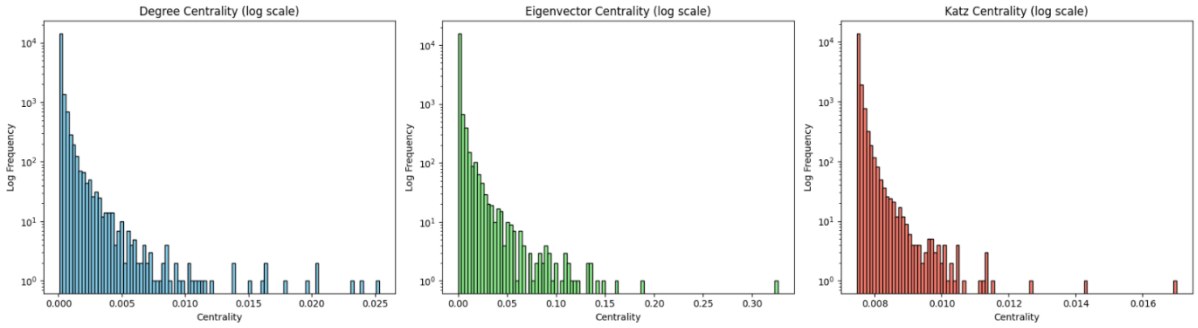


Figure 19. Reddit Centrality Score Distributions (Log-Scaled Frequencies): Degree Centrality, Eigenvector Centrality, and Katz Centrality



We also computed the global clustering coefficient which shows the tendency of users to form tightly connected groups. For Reddit, we got a global clustering coefficient (transitivity) score of 0.00619, indicating less tendency of users to form tightly connected groups. However, we noticed some differences in YouTube, which had a global clustering coefficient (transitivity) score of 5.7056, indicating higher tendency.

We calculated the number of connected components in the network. Reddit was found to have 10006 strongly connected components, while YouTube had 10718, a very similar number, indicating that both platforms have many small groups of users interacting in smaller groups.

Lastly, we identified bridge edges which are links that if removed would break the network into separate parts. These bridges are important for maintaining the cohesion of the network. Both platforms were found to have an extremely large number of bridges, indicating that there a many key users in our networks that are keeping smaller communities connected to each other. These bridges may represent key contributors who connect otherwise isolated prompt

engineering communities, which helps facilitating the spread of new prompting techniques across different user groups.

Community Detection

As part of our social media analysis on discussions related to prompt engineering across both Reddit and YouTube, we wanted to have a deeper look into our reply network graphs, with the aim to identify the different communities present across social media platforms, how many are present in these platforms, and what their sizes are. This will allow us to compare how communities formed around discussing the latest prompt engineering news, insights and advice differ across both YouTube and Reddit. We aim to answer questions such as whether the number of communities present across platforms differ, do communities of similar or different sizes form across platforms, and how community structure differs across platforms.

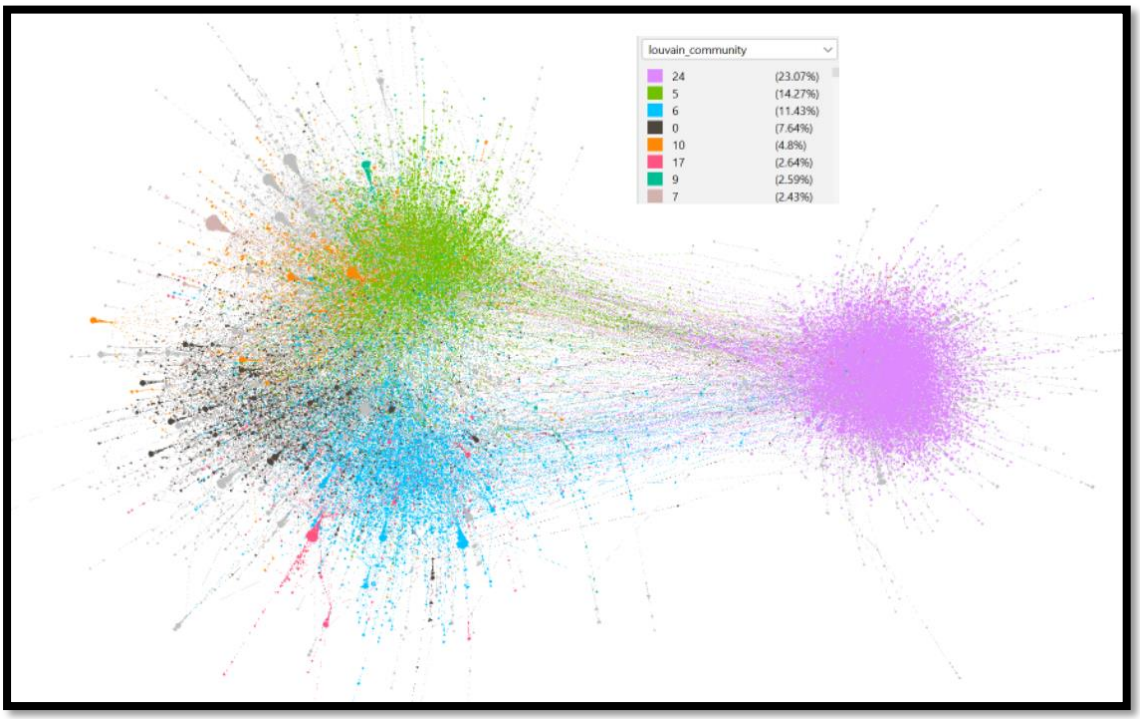
To conduct our analysis, we used two different methods to detect communities in the reply networks. This will allow us to explore the communities present in our networks from different approaches and gain different insights from each approach. The first method was Louvain community detection which optimizes modularity. Modularity measures how well the network is divided into communities. Higher modularity means stronger community structure. On both Reddit and YouTube, we found many communities. The largest community on Reddit had over four thousand users, while the largest on YouTube only had around one thousand six hundred users. We listed the top ten users in these communities. The second method was k-clique percolation which detects overlapping communities. We used a clique size of four. This method allows a user to belong to more than one community. We found many small overlapping communities and calculated the number of users who appeared in multiple groups. We saved the networks with community labels to GraphML files for visualization in Gephi.

Reddit Data

Diving more deeper into our analysis specific to the Reddit reply network, when running Louvain Community Detection on our Reddit reply network, it was found to have a Louvain modularity score of 0.6655, indicating that it had a somewhat high community structure, meaning that the network was fairly well-partitioned into distinct groups with strong internal connections. The Louvain Community Detection approach identified 262 distinct communities, a relatively high number considering that there are around 17,000 total users in our network. To further explore these communities, we also looked at the top 5 largest communities, which had a number of users of 4108, 2542, 2035, 1360, and finally 855. Looking at these top community sizes, we can see that the top 3 communities across Reddit are the dominant ones, with sizes of communities significantly dropping after the main big communities. These findings indicate that prompt engineering discussions likely are primarily dominated by a few large communities, with the majority of the conversation taking place within these bigger groups, then gradually spread out to the other smaller communities by our many identified bridges identified from our earlier social network analysis. Our previously identified most influenceable Reddit user “AppearanceHeavy6724” was found not to be

present in the top 10 users in the largest community, but this could be due to their influence being spread across multiple smaller communities rather than concentrated in the largest one. So, from our Louvain Community Detection on our Reddit reply network, we can gain some insights that larger communities are very important for driving the core discussions and generating high engagement, while influential users such as AppearanceHeavy6724 and other users bridging more smaller communities are important to ensure these discussions on the latest prompt engineering discussions are spread throughout the network. To help visualise these communities, we saved the partition "louvain_community" within our reply network GraphML file for Reddit so that we could visualise our reply network and its communities using the Gephi software introduced to us during the labs. We selected the ForceAtlas2 layout to best highlight the natural structure of the communities and their interactions, while we partitioned our nodes (users) based on their Louvain community which was saved earlier to the graph. This visualisation on our Louvain communities can be found below in **Figure 20**. This helps reinforce our earlier findings, as we can see that the 3 largest communities in **Figure 20** (see the purple, green and blue clusters) are largely dominating the network.

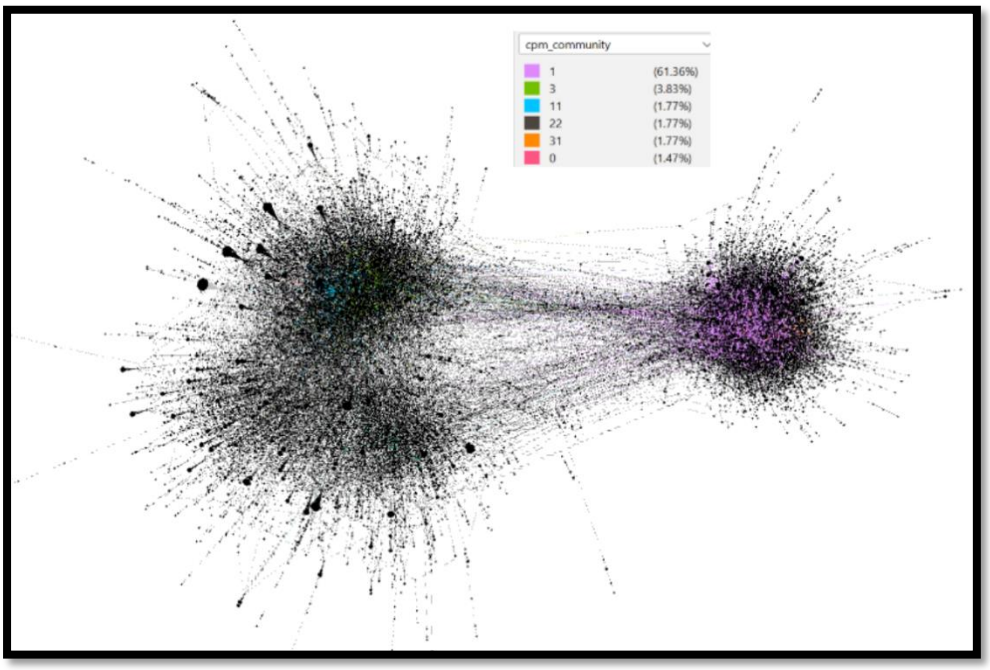
Figure 20. Reddit Reply Network: Louvain Community Visualization



Next, to explore our reddit communities from a different approach, we used the k-clique percolation method (CPM) of community detection, which detects overlapping communities, and allows a user to belong to more than one community. We selected a clique size of four, meaning that users can belong to multiple communities if they are connected to at least four other users within those groups. This size was chosen to capture meaningful overlaps between communities, reflecting users who engage in discussions across different topics or subgroups

within the larger network. After running CPM on our network, we identified only 58 communities, with the largest one having 211 nodes, and the rest all having much smaller numbers of 13 and under. Our CPM communities are shown in a Gephi visualisation below in **Figure 21**. We can see the larger CPM community present within the network (in purple), while the other ones are so small you can barely see them. From our CPM Community detection, we can see that most of the communities are very small and spread out, and likely represent niche discussions or specialized subgroups, where users are potentially engaging with a more specific aspect of prompt engineering.

Figure 21. Reddit Reply Network: CPM Community Visualization

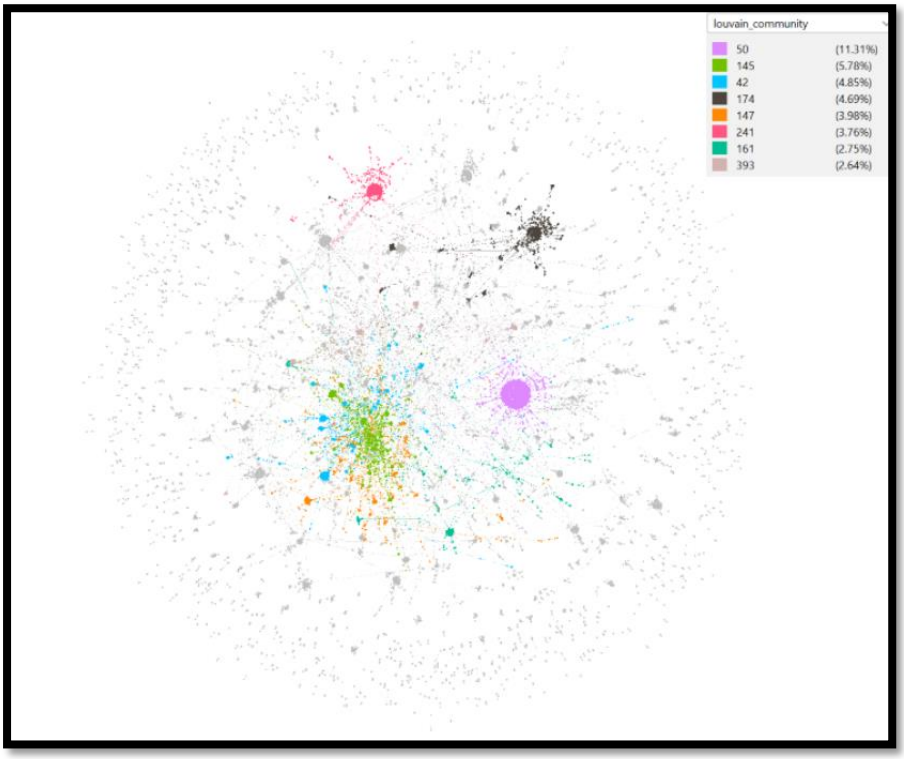


YouTube Data

To explore the community structures present in the YouTube reply network, we first applied the **Louvain Community Detection** algorithm to identify how users organically clustered based on their reply interactions. As mentioned earlier, this algorithm optimizes modularity, which means it tries to find partitions in the network that maximize the density of connections within communities while minimizing the number of connections between different communities.

When running Louvain Community Detection on our YouTube reply network, we identified a total of **182 communities**, with a modularity score of **0.5826**, indicating that the network has a moderate but meaningful community structure. Compared to Reddit's modularity score of 0.6655, the slightly lower score in YouTube suggests that the community structures are present but somewhat looser and less distinct, possibly due to lower user-to-user interaction rates.

Figure 22. YouTube Reply Network: Louvain Community Visualization

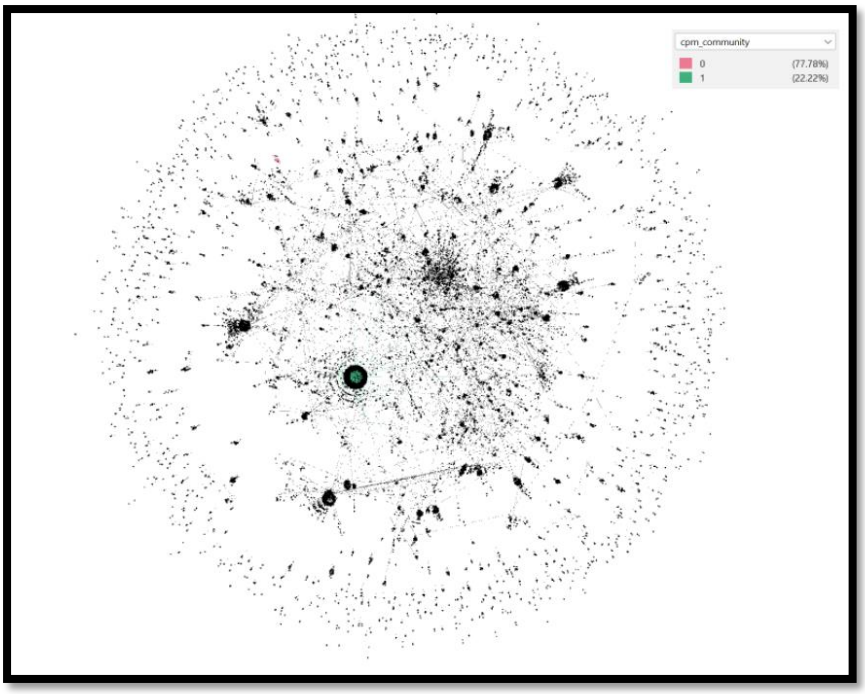


As shown in **Figure 22. YouTube Reply Network: Louvain Community Visualization**, the largest communities are represented by the most dominant colours, such as purple, green, pink, and blue. The most dominant community, labeled **community 50**, accounts for **11.31 percent** of the network. Following that are several mid-sized communities including **community 145** (5.78 percent), **community 42** (4.85 percent), **community 174** (4.69 percent), and others. While these top few communities are clearly visible, a large portion of the network is composed of many smaller communities. This supports the idea that while a few core groups are highly engaged, much of YouTube interaction is fragmented and distributed across smaller, less connected user groups.

Unlike Reddit, where the top three communities held a significant portion of the total user base, YouTube's communities are more evenly distributed among mid-sized clusters, and even the largest community makes up a smaller proportion of the overall network. This reflects a more diverse and decentralized discussion landscape, where no single group overwhelmingly dominates the conversation around prompt engineering.

Next, to analyse overlapping communities, we applied the **k-clique percolation method (CPM)** using a clique size of four. This method is useful for uncovering overlapping groups where users participate in more than one conversation cluster. After running CPM on the YouTube reply network, we discovered **only two major overlapping communities**, shown in **Figure 23. YouTube Reply Network: CPM Community Visualization**.

Figure 23. YouTube Reply Network: CPM Community Visualization



One of these communities, labeled **community 0**, accounts for a significant **77.78 percent** of the users found in overlapping structures, while **community 1** represents **22.22 percent**. This is a much smaller number of overlapping communities compared to Reddit's 58 CPM communities. The results suggest that most YouTube users tend to stay within a single conversational cluster rather than branching out across multiple threads or topics.

Another interesting observation is that the overlapping communities on YouTube are highly centralized, with the majority of interactions anchored within one large community. This may reflect the video-centric nature of the platform, where discussions tend to form around individual content creators or specific video topics, as opposed to the subreddit-based ecosystem of Reddit that fosters cross-topic engagement.

In conclusion, YouTube's community structure is shaped by a few mid-sized Louvain communities and one overwhelmingly dominant overlapping CPM community. The Louvain-based visualization confirms that several tightly connected clusters of users are active, while the CPM results highlight the platform's tendency toward isolated engagement rather than interwoven discussions. These patterns suggest that YouTube's user interaction model favors localized and self-contained discussions, with limited bridging between different clusters of conversation.

Independent Cascade Model Simulation

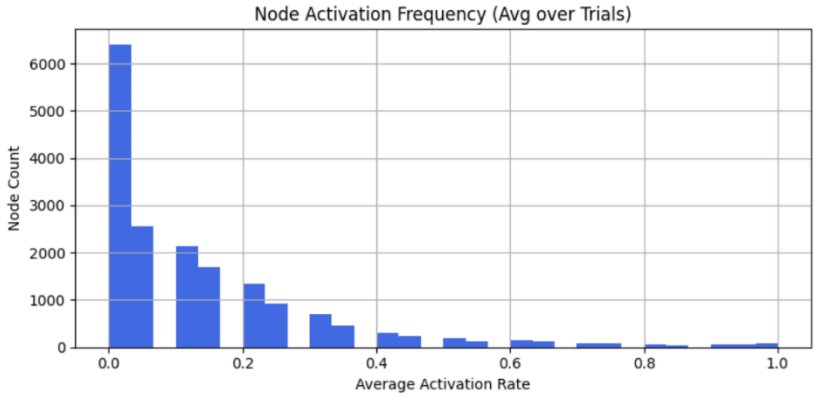
As part of our analysis, we used the Independent Cascade Model to simulate the spread of influence throughout our reply networks created from social media data collected from both YouTube and Reddit. In this model, each active user has a probability of 0.2 (activationProb = 0.2) to activate each of their neighbours, representing the likelihood of influencing others through replies or interactions. We selected five seed users based on the highest degree

centrality, to simulate how information will spread if it originates from just the top 5 influential users in our network. We ran the simulation for twenty trials for each platform to account for randomness and variability in influence spread. Each time we recorded the number of nodes activated, which allows us to measure the average reach and influence potential originating from these key users and help simulate the process of how information would be spread through our Reddit and YouTube network. This can help us understand how information related to the latest prompt engineering news and discussions would likely spread through our networks and communities.

Reddit Data

After running the Independent Cascade Model on our Reddit reply network, we found that the average number of activated nodes (representing users) was 2364 out of the total 17810 users in the network, which is about thirteen percent of the network. These results helped us understand the amplifying effect of central nodes, showing that a few key influencers can spark significant engagement across a social media platform. To get a better understanding of the average activation frequency of nodes across 20 trials, we created a histogram to better visualise it. See in **Figure 24** the average activation frequency of nodes visualised.

Figure 24. Node Activation Frequency (Avg over Trials) for Reddit Social Network



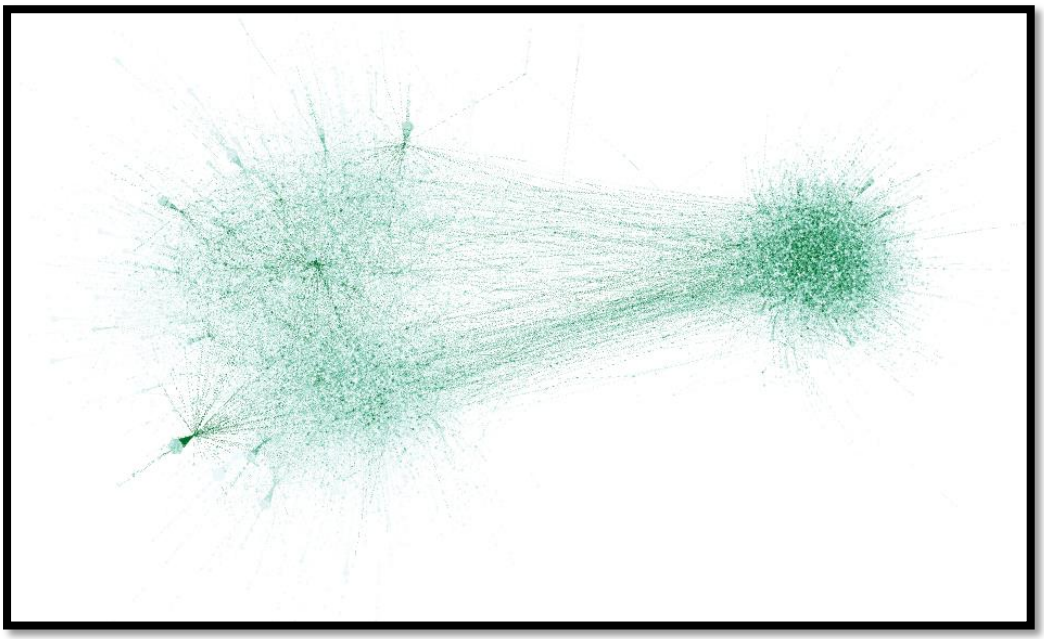
From this Histogram in **Figure 24**, we can see that most nodes had very low average activation rates, with activation sharply declining as rates increased, indicating that only a small subset of users were frequently activated across trials.

The Independent Cascade Model has showcased the influence spread on Reddit, where it reached over 13% of the network, demonstrating the impact of a small set of highly connected users. While many nodes stayed inactive, a significant portion experienced moderate to high activation, suggesting that Reddit's reply structure supports wide-reaching influence, likely due to high connectivity and active user hubs. Even a few central nodes can spark substantial engagement through the network.

To get a visual representation of this spread of influence throughout our reply networks and to visually see how information related to the latest prompt engineering news and discussions would likely spread through our networks and communities, we used Gephi again to visualise this for us, and again selected the ForceAtlas2 to ensure similarity between visualisations with previous sections of analysis. This time we used the ranking feature to rank our nodes by their

average activation, with nodes with a darker green colour representing a higher activation rate, and nodes with a lighter green indicating a lower activation rate. Darker green nodes are those users that were activated more consistently across trials, indicating users who are either highly connected or strategically positioned within the network to receive and propagate information effectively. This visualisation can be seen below in **Figure 25**. Some insights we can get from the visualisation are that the highly activated users are mainly clustered within more denser cores of the network, while less central users tend to have a lower activation rate.

Figure 25. Visualisation of Influence Spread in Reddit Reply Network Using Average Node Activation (Independent Cascade Model)



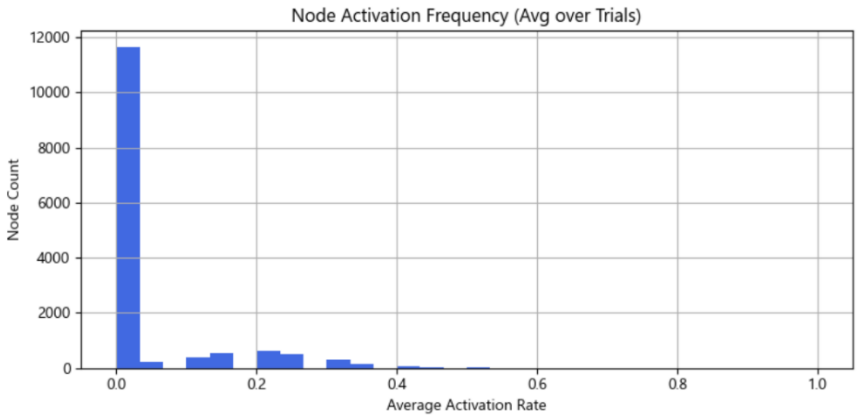
YouTube Data

To better understand how influence spreads across the YouTube social network when prompted by central users, we applied the Independent Cascade Model using the same simulation parameters as used for the Reddit network. This means each user, once activated, had a fixed probability of zero point two to activate each of their direct neighbors in the network. As with Reddit, the top five seed users were selected based on their degree centrality, meaning they had the highest number of direct connections in the YouTube reply network. The simulation was run for a total of twenty independent trials to allow for statistical averaging and to account for the randomness of the cascade process.

After running the model, we observed that the average number of activated nodes on YouTube was approximately five hundred eighty out of a total of fourteen thousand five hundred and five users. This equates to roughly four percent of the entire network, which is significantly lower than the thirteen percent activation rate observed on Reddit. This lower rate of influence spread highlights the more passive and sparse structure of the YouTube comment network, where users tend to interact less frequently and less reciprocally compared to Reddit.

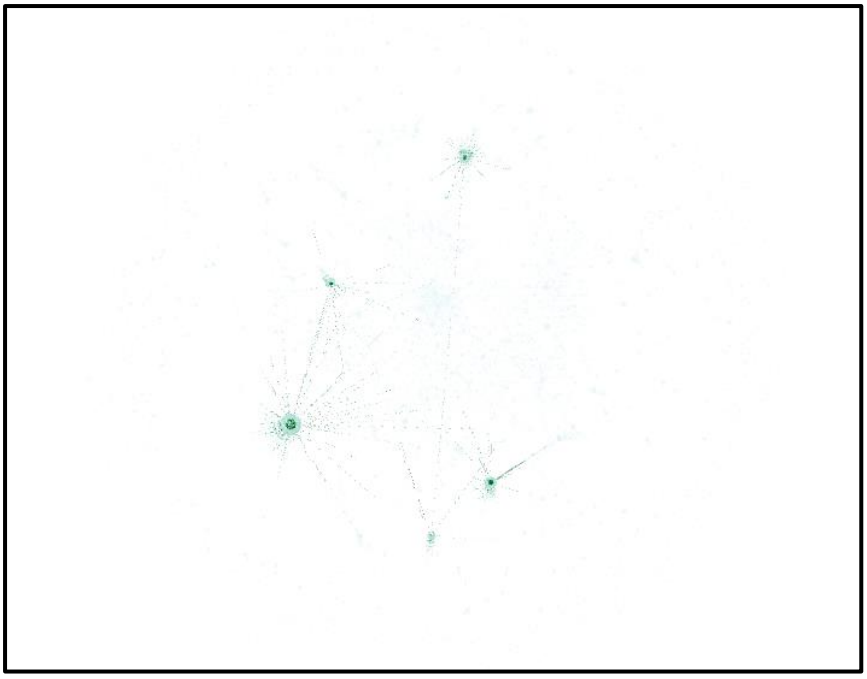
To understand how often users were activated across simulation trials, we generated a histogram representing the **average activation frequency** of nodes. This can be seen in **Figure 26. Node Activation Frequency (Avg over Trials) for YouTube Social Network**. In this histogram, it becomes evident that the majority of nodes had a near-zero activation rate, meaning they were very rarely influenced or activated during simulations. Only a small group of users exhibited moderate to high activation frequency, suggesting a concentration of influence around a few tightly connected hubs or mini-communities.

Figure 26. Node Activation Frequency (Avg over Trials) for YouTube Social Network



This sparse and hub-centric activation pattern is further confirmed by the Gephi visualization of the reply network shown in **Figure 27. Visualisation of Influence Spread in YouTube Reply Network Using Average Node Activation (Independent Cascade Model)**. In this network visualisation, nodes are colored in shades of green based on their average activation rates, with darker green nodes indicating users that were more consistently activated across the twenty trials. These nodes are primarily clustered in isolated, dense patches of the graph, reflecting the relatively insular communication style of YouTube comment threads where users engage more within individual video communities rather than across broader platform-wide discussions.

Figure 27. Visualisation of Influence Spread in YouTube Reply Network Using Average Node Activation (Independent Cascade Model)



From this visualisation, we can conclude that the YouTube reply network exhibits a limited but focused spread of influence under the Independent Cascade Model. Most activation is constrained within local pockets around seed users, with very few ripple effects reaching distant or less connected parts of the network. This finding contrasts with the Reddit network, which had broader and more distributed influence pathways, highlighting how different platform structures affect the dynamics of information dissemination.

Together, these observations reinforce the idea that while central users on YouTube do hold influence, their ability to propagate information or sentiment broadly across the entire platform is limited by structural constraints. For significant influence spread to occur, either a higher number of seed users would be required, or the platform itself would need more inter-user engagement in reply chains to facilitate deeper cascades of information.

On YouTube the average was around five hundred eighty which is about four percent of the network. We plotted histograms of the average activation rates. Most nodes were rarely activated. A small number were activated frequently. This suggests that influence spread is limited unless more seeds or higher probabilities are used.

Linear Threshold Model Simulation

We also simulated influence using the Linear Threshold Model, which, like the Independent Cascade Model, captures how information spreads through a network. However, this model has a key difference to the previous one, as it simulates homophily by requiring a node to be influenced by a collective threshold of its neighbours. In this model each node has a random threshold, and a node activates if the total incoming influence exceeds this threshold. We used the same five seed users as before and selected the same number of trails of 20 so we could compare if homophily has an impact on how information spreads, compared to the

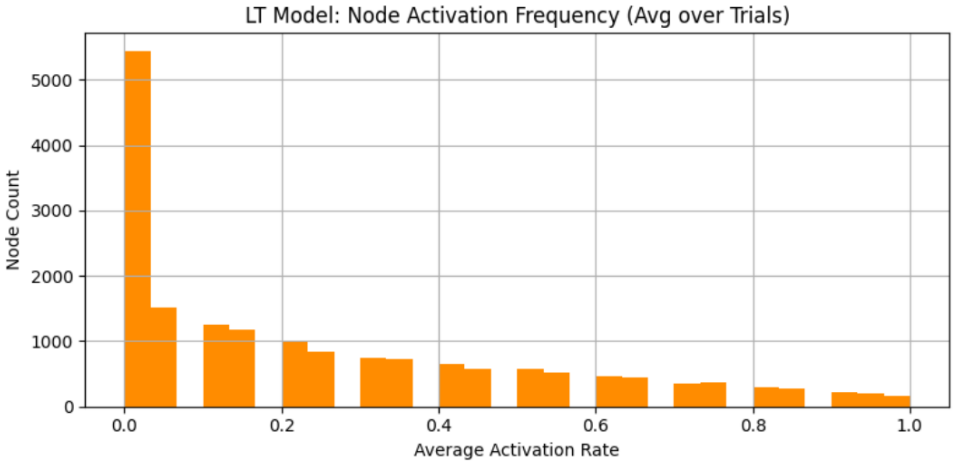
Independent Cascade Model, which relies on individual, probabilistic attempts to activate neighbours. We generated edge weights using a Dirichlet distribution, and the simulation was run for twenty trials.

On Reddit the average number of activations was around four thousand four hundred which is twenty five percent of the network. On YouTube the average was higher than in the Independent Cascade model. This shows that cumulative influence is more effective at activating users than single random attempts. We plotted histograms of the activation frequencies. Many users had low activation, while a large number had medium activation. A few users had very high activation. This pattern suggests that homophily and group reinforcement play a significant role in spreading influence, which is especially relevant for understanding how niche but technical topics like prompt engineering can gain traction when discussed within tightly connected interest groups.

Reddit Data

After running the Linear Threshold Model on our Reddit reply network, we found that the average number of activated nodes (representing users) was 4469 out of the total 17810 users in the network, which is about 25 percent of the network, a significant improvement over the Independent Cascade Model. These results helped us solidify our findings that a few key influencers can spark significant engagement across a social media platform using the power of homophily, where users are more likely to be influenced when multiple similar or connected peers are also engaged. To get a better understanding of the average activation frequency of nodes across 20 trials, we created a histogram to better visualise it. See in **Figure 28** the average activation frequency of nodes visualised.

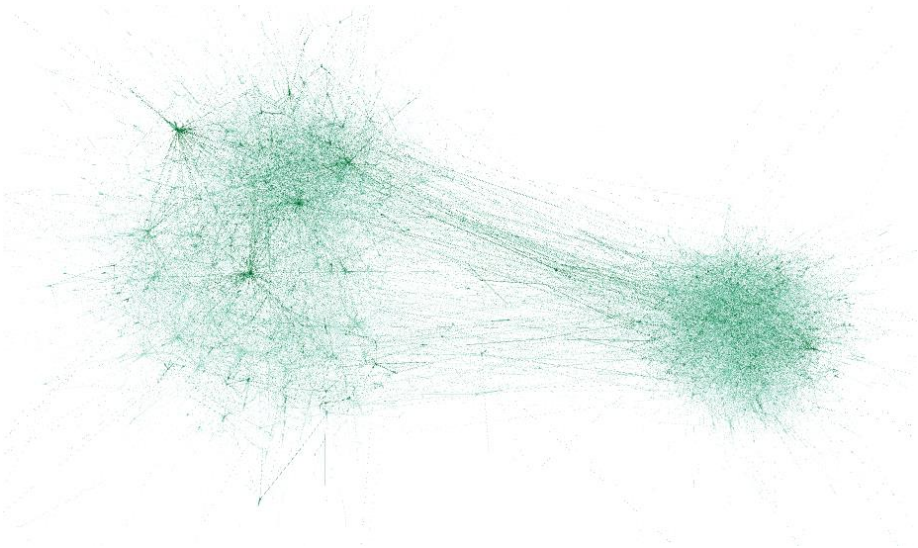
Figure 28. Node Activation Frequency (Avg over Trials) for Reddit Network – Linear Threshold Model



From this Histogram in **Figure 28**, we can see that nodes still folow a similar distribution of activation rate as compared to the Independent Cascade Model, but with users generally having a slightly higher activation rate.

To get a visual representation of this spread of influence throughout our reply networks based on homophily, we used Gephi again to visualise this for us, and once again we selected the ForceAtlas2 to ensure similarity between visualisations with previous sections of analysis. We again used the ranking feature to rank our nodes by their average activation, with nodes with a darker green colour representing a higher activation rate, and nodes with a lighter green indicating a lower activation rate. This visualisation can be seen below in **Figure 29**.

Figure 29. Visualisation of Influence Spread in YouTube Reply Network Using Average Node Activation (Linear Threshold Model)

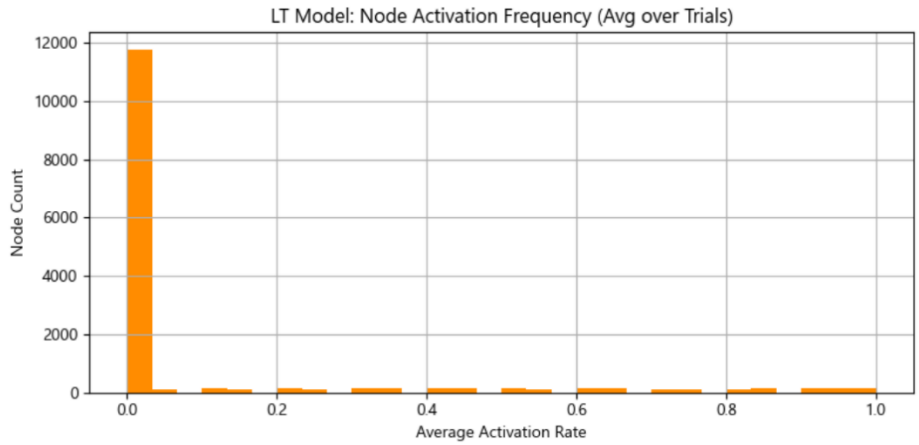


YouTube Data

After completing the Linear Threshold Model simulation on our YouTube reply network, we discovered that the model performed slightly better in activating nodes compared to the Independent Cascade Model. The average number of activated nodes was found to be 937 out of the total 14505 users in the network. This represents around 6.46 percent of the network, which is a notable improvement over the four percent activation we saw earlier from the cascade-based simulation.

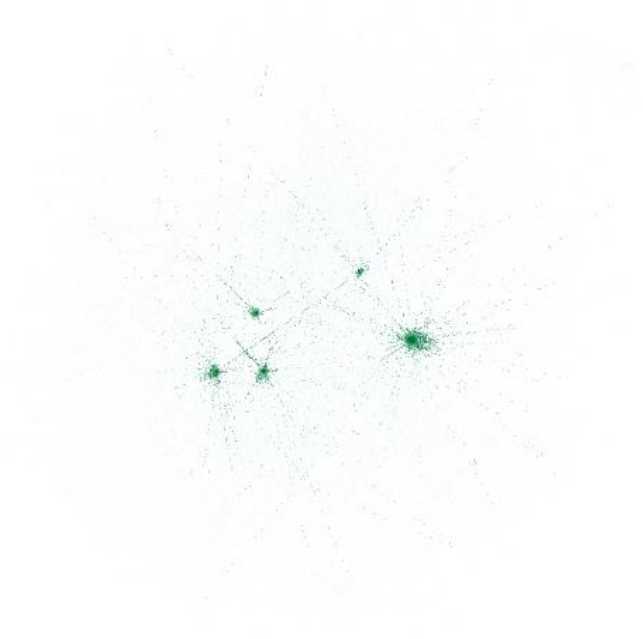
We created a histogram to help visualise the average activation frequency of each node across twenty trials. This histogram is presented in Figure 30 and shows how many times a user was activated on average during the simulations. Most users had very low activation frequencies, while a few users were activated more often, following a long-tail distribution. The spike near the zero end of the histogram suggests that most YouTube users were only influenced occasionally, if at all, unless they were close to or directly connected with the influential seed nodes.

Figure 30. Node Activation Frequency (Avg over Trials) for YouTube Network – Linear Threshold Model



To further explore how influence visually spread through the YouTube reply network, we again used Gephi to create a layout using the ForceAtlas2 algorithm. Each node represents a user and is colored according to its average activation rate across the twenty trials. Nodes with higher activation rates are darker green, showing they were more consistently activated and influenced. Figure 31 below shows this spread pattern across the YouTube reply graph.

Figure 31. Visualisation of Influence Spread in Reddit Reply Network Using Average Node Activation (Linear Threshold Model)



From this visualisation, we can observe that most of the influence spread occurred near the network's denser central regions, similar to what we saw in the Reddit simulation. However, the overall spread was more limited due to YouTube's more sparse interaction structure and

lower network density. Still, we can see that some nodes were activated quite frequently, especially those directly linked to multiple influential users.

Comparative Insights

From comparing both models and platforms, we can draw the following conclusions:

- **Reddit consistently saw broader influence spread** across both models. This is likely because Reddit users engage more deeply in replies and discussions, creating a more connected structure that enables wider influence propagation.
- **YouTube showed lower activation overall**, particularly under the Independent Cascade Model, due to its sparse reply structure. However, the Linear Threshold Model performed better because of its ability to aggregate influence from multiple neighbours.
- **Homophily-driven influence**, as simulated by the Linear Threshold Model, appears more effective on both platforms than single-shot influence as in the Independent Cascade Model.
- The **activation histograms** in both networks suggest a small set of highly central users hold most of the influence, which aligns with the centrality analysis done earlier.

These findings reinforce the importance of social reinforcement in spreading prompt engineering knowledge across online platforms and demonstrate that even sparsely connected platforms like YouTube can benefit from cumulative influence models.

Conclusion

This project provided a detailed and multi-dimensional analysis of discussions surrounding **prompt engineering** across two major social media platforms: **Reddit** and **YouTube**. By collecting over 105,000 combined posts and comments, we were able to perform robust textual and network-based analyses that offer both breadth and depth of insight into how online communities engage with prompt engineering.

Through **text preprocessing and frequency analysis**, we discovered that key terms such as *model*, *prompt*, *chatgpt*, and *use* consistently dominated both platforms, highlighting strong interest in the mechanics and applications of prompting AI systems. Our **sentiment analysis**, performed using the VADER tool, revealed temporal fluctuations in public perception—Reddit showed more stable sentiment trends due to its discussion-heavy format, whereas YouTube's sentiment was more volatile and reactive, likely driven by video content and shorter comment styles.

In our **topic modelling** using Latent Dirichlet Allocation (LDA), we identified five major themes per platform. Reddit topics leaned more technical and theory-driven, discussing model internals, hardware configurations, and philosophical reflections on AI. YouTube, on the other hand, featured more practical, reaction-based topics such as tutorials, tool access, and appreciation for content creators. Word clouds and pyLDAvis visualizations helped us distill and interpret these themes clearly.

Moving into **social network analysis**, we built reply networks for both platforms to uncover how users interact. Centrality metrics revealed influential users—most notably, *AppearanceHeavy6724* on Reddit and *@JeffSu* on YouTube—who played key roles in shaping discussion. Reddit's network was found to be more dense and interaction-rich, while YouTube's was sparse, with far fewer reply chains.

Community detection further reinforced these findings. Louvain modularity revealed that Reddit hosted a large number of tightly-knit communities, with the top three dominating the network. YouTube also had identifiable communities, though they were fewer and less interconnected. The k-clique percolation method showed Reddit users often belonged to overlapping interest groups, whereas YouTube communities were more isolated.

Finally, our **influence simulations** using the Independent Cascade Model and Linear Threshold Model quantified the dynamics of influence spread. Reddit's denser network structure allowed central users to activate up to **25% of the network** under the Linear Threshold Model, demonstrating the power of cumulative influence and community structure. On YouTube, while the absolute spread was lower, the results still showed that strategic influencers could trigger notable reach, especially under models that leverage homophily and repeated exposure.

In summary, this analysis underscores the different ways in which communities engage with prompt engineering across platforms. Reddit favors depth, technical discussion, and structured replies—resulting in more effective influence propagation. YouTube emphasizes breadth, reaction-based feedback, and appreciation, making it a more volatile but emotionally engaged space. Both platforms show that **a small number of central users and community hubs** drive the majority of discourse, and that **social reinforcement is key** to information spread.

Together, these insights contribute to a more nuanced understanding of digital engagement patterns in AI-related domains and highlight the value of cross-platform social media analysis in tracking public discourse and community influence.

Appendix

Location of teamwork documentation

Microsoft Teams channel for groupwork collaboration titled "**Assignment 2 : Choose Your Own Analysis (Group 44)**" shared with lida.rashidi@rmit.edu.au

Find our weekly Timesheets and who completed what task in the document "🗓️ **Assignment 2 Weekly Timesheet**" found in the Microsoft Teams channel files section.

Location of GitHub for team collaboration on the coding

GitHub shared with lida.rashidi@rmit.edu.au

GitHub repository is titled:

Group-44-Assignment-2-Choose-Your-Own-Analysis

Find the full assignment code, and all the data files here.