Thesis Dissertation

POLITICAL IDEOLOGY AND CONSPIRACY THEORY ENGAGEMENT ON REDDIT

Konstantinos Sofokleous

UNIVERSITY OF CYPRUS

COMPUTER SCIENCE DEPARTMENT

May 2026

UNIVERSITY OF CYPRUS
COMPUTER SCIENCE DEPARTMENT

Political Ideology and Conspiracy Theory Engagement on Reddit

Konstantinos Sofokleous

Advisor: Dr. Demetris Paschalides
Supervisor: Dr. George Pallis

Thesis submitted in partial fulfilment of the requirements for the award of degree of Bachelor in Computer Science at University of Cyprus

May 2026

Declaration of Originality

I declare that this dissertation and any accompanying software are my own original work, conducted in full compliance with the University of Cyprus Code of Conduct for Research (https://www.ucy.ac.cy/legislation/kodikas-deontologias-kai-politikes/), and with full accountability on my part for the accuracy, integrity, and validity of all content.

Acknowledgments

I would like to express my thanks to my supervisor, Dr. George Pallis, for creating my interest in this field in the first place and for his support throughout the project.
I would also like to give a special thank you to my advisor, Dr. Demetris Paschalides, for his constant guidance and support during the project from start to finish, especially giving me constant feedback and new ways to improve my work.
Finally, I would like to thank my close friends and fellow students for their support, ideas and suggestions for the project.

Abstract

This thesis investigates whether political ideology predicts conspiracy theory engagement on Reddit during the 2024 U.S. presidential election cycle (May-November 2024). A modular Python pipeline is developed that infers user ideology from political subreddit behaviour, discovers conspiracy topics through unsupervised clustering, constructs interaction networks, and tests the ideology-conspiracy link with proper statistical controls across eight research questions.

The pipeline identifies 55,154 users active in both conspiracy and political subreddits, computes continuous three-dimensional ideology participation vectors (left, centre, right) for each, and selects the top 14.1% by ensemble bridging score as high-value users (7,798 HVU) for deeper analysis. BERTopic with Google's EmbeddingGemma-300M embeddings clusters HVU posts into 14 main conspiracy topics, further refined to 20 subtopic-level categories through hierarchical sub-clustering. Four network types are constructed (user-post, user-user co-activity, user-subreddit bipartite, and user-topic) alongside a null model that tests whether the observed interaction patterns exceed chance. A fine-tuned RoBERTa classifier (76.9% accuracy, macro F1 0.716 on 21 classes) and a cosine-similarity centroid pipeline jointly power a Chrome extension for real-time classification of Reddit posts.

The results present a layered picture across three levels of analysis. At the topic and user level, no conspiracy topic shows a statistically significant ideological skew after Benjamini-Hochberg correction (0/14 main topics, 0/20 subtopic-level categories), and all effect sizes are negligible to small. At the network level, a null model based on 500 ideology-vector shuffles confirms significant ideological homophily in the user co-activity network. The observed mean cosine edge distance (0.4425) falls at the 0th percentile of the null distribution ($p < 0.001$), meaning users interact preferentially with ideologically similar peers. At the community level, subreddit choice is linked to ideology ($p = 0.035$), even though topical engagement within those communities is not. Political diversity, meaning how broadly someone participates across ideological subreddits, is a stronger predictor of conspiracy engagement than ideological direction.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

1.1 Motivation

Social media has changed the way people come across and discuss conspiracy theories. Reddit, one of the biggest online forums, has communities like r/conspiracy where users talk about everything from political cover-ups to public-health scepticism. At the same time, there are plenty of politically-oriented subreddits where users express and reinforce their ideological views. The fact that many users are active in *both* types of communities raises an interesting question about whether a person's political leaning actually predicts what conspiracy theories they engage with.

This question became especially relevant during the 2024 U.S. presidential election, a period with a lot of political polarisation and conspiracy-related activity. Major events like the assassination attempt on former President Trump (July 13, 2024), President Biden dropping out of the race (July 21, 2024), and Election Day itself (November 5, 2024) all triggered waves of conspiratorial discussion online. Understanding whether ideology either drives conspiracy engagement or just happens to co-exist with it matters for content moderation, media literacy, and research on online radicalisation.

1.2 Problem Statement

The original objective of this thesis, as defined in the project proposal, was to automatically identify, classify, and analyse conspiracy theories related to *specific presidential candidates* (Donald Trump, Joe Biden, Kamala Harris) within ideologically polarised Reddit communities, and to build a browser plugin to detect these discussions in real time. The initial research questions focused on which candidate-specific conspiracy theories appeared in left vs. right subreddits, how conspiracy-candidate networks differed across ideological groups, and whether a classifier could reliably detect these candidate-targeted narratives.

As the project progressed, the scope evolved in two important ways. First, unsu-

pervised topic modelling (BERTopic) revealed that the conspiracy landscape on Reddit is not neatly organised around individual candidates. The topics that emerged were thematic (vaccines, election integrity, geopolitics, occult theories, etc.) rather than candidate-centred, although candidate-related content remained prominent, with the single largest topic by volume being *Trump, Biden, and Election Manipulation* ($N = 628$ posters) and a second topic directly addressing *Election Integrity and Political Dynamics* ($N = 230$ posters), together accounting for roughly a third of the clustered corpus. Trying to force a candidate-specific framing onto all 14 discovered topics would have been artificial and would have discarded large portions of the data without theoretical justification. Second, the more interesting and testable question turned out to be the broader one of whether political ideology predicts conspiracy engagement *at all*, regardless of which candidate is implicated. This reframing allowed for a more rigorous statistical analysis across the full topic landscape rather than a narrower descriptive study. Candidate-specific threads are interpreted in the Discussion (Chapter 7) in relation to the original supervisory objectives.

Most previous work on conspiracy theories and political ideology has relied on surveys, small-scale qualitative studies, or keyword-based content filtering. These methods have clear drawbacks, as they are limited in scale, prone to selection bias, and struggle to capture the full range of topics that conspiracy discussions cover on a platform like Reddit. On top of that, existing studies usually assign users a single ideological label based on self-reports or manual annotation, which ignores how political identity is really more of a spectrum than a binary choice.

What is needed is a data-driven approach that can (a) infer ideology from actual user behaviour rather than self-reports, (b) discover conspiracy topics automatically instead of relying on predefined keyword lists, and (c) test the ideology-conspiracy relationship using proper statistical controls. This thesis aims to fill that gap.

1.3 Research Questions

The main question this thesis tries to answer is the following.

Does political ideology predict conspiracy theory engagement on Reddit?

This breaks down into the following sub-questions.

RQ1. Which ideological groups dominate each conspiracy topic?

RQ2. Do left-leaning and right-leaning users differ in how intensely they engage with each topic?

RQ3. How ideologically diverse is each conspiracy topic's audience?

RQ4. Can user ideology predict which conspiracy topics they engage with?

RQ5. Do composite user bridging scores differ by ideology?

RQ6. Is the conspiracy-subreddit co-activity network more ideologically assortative than chance?

RQ7. Does the ideological composition differ across conspiracy subreddit ecosystems?

RQ8. How accurately can the pipeline classify a short-form Reddit post into one of the conspiracy topic categories?

1.4 Contributions

This thesis makes the following contributions.

- A **behavioural ideology inference pipeline** that builds continuous three-dimensional ideology participation vectors for over 55,000 Reddit users from their political subreddit activity, using a six-tier confidence scheme instead of hard labels.
- An **unsupervised topic discovery framework** using BERTopic with engagement-weighted EmbeddingGemma-300M embeddings, finding 14 main conspiracy topics with a sub-clustering step that expands them to 20 subtopic-level categories.
- A **null model for network ideology mixing** based on 500 ideology-vector shuffles, confirming significant ideological homophily ($p < 0.001$) and showing that ideologically mixed users hold no structural advantage in the network.
- A **multi-layer statistical evaluation** of the ideology-conspiracy link using chi-squared tests, Mann-Whitney U tests, Cohen's d, odds ratios, logistic regression, Spearman correlation, and Benjamini-Hochberg correction across eight research questions.
- A **Chrome extension** (Conspiracy Identifier) backed by a fine-tuned RoBERTa classifier (76.9% accuracy, macro F1 0.716 on 21 classes) that classifies Reddit posts in real time into the discovered topics.

1.5 Thesis Outline

The rest of this thesis is organised as follows.

- **Chapter 2** covers the background on conspiracy theory research, Reddit as a platform, and the NLP and network analysis techniques used.

- **Chapter 3** describes the data collection process, the subreddit sources, and the exploratory data analysis.
- **Chapter 4** goes through the full pipeline, covering ideology vectors, user scoring, BERTopic clustering, network construction, polarisation metrics, and the Chrome extension.
- **Chapter 5** presents the findings for all eight research questions, covering topic-level composition and engagement tests (RQ1-RQ5), a null model for network ideology mixing (RQ6), subreddit ecosystem comparison (RQ7), and classifier evaluation (RQ8).
- **Chapter 6** reviews related work in conspiracy analysis, ideology inference, topic modelling, and polarisation measurement, and positions this thesis within that literature.
- **Chapter 7** discusses the findings in relation to the original research objectives, summarises the main results, and suggests future directions.

Chapter 2

Background

This chapter covers the main concepts and tools that this thesis builds on. Section 2.1 discusses conspiracy theories and their relationship with political ideology. Section 2.2 introduces Reddit as a research platform. Section 2.3 explains the NLP techniques used for topic modelling. Section 2.4 covers the network analysis concepts, including the ideological homophily and null model framework used in this work.

2.1 Conspiracy Theories and Political Ideology

A conspiracy theory can be defined as an explanation that attributes events to a secret, coordinated effort by some group of people working toward a hidden goal [16]. Researchers have long debated whether conspiracy beliefs lean toward one side of the political spectrum or whether they are spread evenly. Goertzel [21] argued that conspiracy beliefs form a kind of *monological belief system*, once someone believes one conspiracy, they are more likely to believe others, regardless of their politics. Brotherton et al. [11] built on this idea with the Generic Conspiracist Beliefs Scale, a tool for measuring how prone someone is to conspiratorial thinking in general.

There is disagreement in the literature on whether ideology matters at all. Uscinski and Parent [50] argue that conspiratorial thinking is a general trait that does not depend on left or right leanings. On the other hand, van Prooijen et al. [53] found that political extremism on *both* sides, not just one, predicts conspiracy belief. Miller et al. [33] showed that conspiracy endorsement depends on political knowledge and trust, while Enders et al. [17] found that the ideology-conspiracy link varies depending on which specific conspiracy is being considered.

The 2024 U.S. election cycle is a useful setting for studying this, since both left and right-leaning communities were actively producing and sharing conspiracy content, from election fraud claims to vaccine scepticism to geopolitical cover-up theories.

2.2 Reddit as a Research Platform

Reddit is a social platform organised into user-created communities called *subreddits*, each focused on a specific topic or interest. Subreddits can be about anything, from broad topics like r/politics to niche interests like r/ModelTrains. Users can post text, links, or media, and other users vote the content up or down. The Pushshift project [7] has made large-scale Reddit data available for research, leading to many studies on community dynamics and political discourse. Because Reddit is pseudonymous and has such fine-grained community structure, where someone posts can tell you a lot about their interests and political leanings [47, 55].

For this thesis, two types of subreddits matter. The first are *politically classified subreddits*, sorted into left-leaning, right-leaning, or centre based on established classification lists. The second are *conspiracy subreddits* (r/conspiracy, r/conspiracy_commons, r/conspiracytheories, and r/ConspiracyII). Previous studies have looked at r/conspiracy's user base and content [25, 44], finding that it acts as a gateway community with a lot of user overlap with both political and fringe subreddits. The users who are active in *both* political and conspiracy subreddits are the main population of interest for this thesis.

2.3 NLP Techniques for Topic Modeling

2.3.1 Traditional Approaches

The classic methods for topic modelling include Latent Dirichlet Allocation (LDA) [10] and Non-negative Matrix Factorization (NMF) [27], both of which work on term-frequency representations of documents. These work reasonably well on clean, well-structured text, but they struggle with the kind of short, noisy, context-heavy posts found on social media. NMF with TF-IDF was actually used in an earlier iteration of this thesis before being replaced with a better approach.

2.3.2 BERTopic

BERTopic [23] is a newer topic modelling framework that uses pre-trained transformer-based sentence embeddings, UMAP for dimensionality reduction, and HDBSCAN for clustering. Unlike LDA, it does not require you to specify the number of topics in advance, and it captures meaning that simple word-counting methods miss. After clustering, it uses class-based TF-IDF (c-TF-IDF) to extract the most representative keywords for each topic. Top2Vec [3] is a similar approach, though it uses a joint document-word embedding instead of the modular pipeline that BERTopic uses.

2.3.3 Sentence Embeddings

Sentence embeddings are vector representations of text that capture semantic meaning, two sentences about the same thing will end up close together in the vector space, even if they use different words. Reimers and Gurevych [42] introduced Sentence-BERT, which fine-tunes BERT [15] with siamese networks to produce good sentence-level representations. This thesis uses Google's EmbeddingGemma-300M [22], a 300-million parameter model that works well on short to medium-length English text. The choice of embedding model matters a lot for clustering quality, since the whole approach depends on similar posts being placed near each other in embedding space.

2.4 Network Analysis Concepts

Network analysis gives us a way to study relationships between entities as graphs [35]. In this thesis, several types of networks are built.

- **Bipartite networks**, where edges connect two different types of entities (e.g., users and subreddits, or users and topics).
- **Co-activity networks**, where edges connect users who posted in the same conspiracy thread, weighted by how many threads they shared.

Some key concepts used in this work are *homophily*, the tendency for similar people to connect with each other, which is one of the most well-studied patterns in social networks [32], and *modularity*, a measure of how well a network splits into distinct communities [35]. Polarisation is measured using a suite of custom metrics that look at ideological mixing, cross-participation, and bridging behaviour, defined in detail in Chapter 4.

A key question in this thesis is whether the observed ideological interaction patterns in r/conspiracy and related subreddits are genuine or merely a consequence of the network's ideological composition. This is addressed using a *null model* that randomly permutes ideology labels across users while keeping the graph structure fixed, allowing the observed patterns to be compared against what would be expected by chance. This approach follows standard practice in network science for separating structural effects from compositional ones.

2.4.1 Gephi

Gephi [6] is an open-source tool for visualising and analysing networks. In this thesis, node and edge lists are exported from the Python pipeline in Gephi's CSV format, but as explained in Chapter 4 the visualisation itself proved unreadable at the final corpus

size, so the exported files are consumed by the polarisation metrics and heatmap pipelines instead of by Gephi.

Chapter 3

Dataset

This chapter describes the data sources, how they were collected, and the exploratory data analysis that helped guide the methodology. Data collection started in September 2025 and was later expanded with additional conspiracy subreddits in March 2026.

3.1 Data Sources

Two categories of Reddit data were collected, both covering the period from May 5, 2024 to November 5, 2024. This is a six-month window around the 2024 U.S. presidential election.

3.1.1 Conspiracy Subreddit Data

Posts and comments were collected from four conspiracy-focused subreddits using Arctic Shift [4]. Initially, only r/conspiracy was scraped on the supervisor's recommendation, as it is the largest and most-studied conspiracy community on Reddit [25, 44]. Three more subreddits were added way later during the RoBERTa training phase to broaden the corpus and improve both the centroid and RoBERTa classification performance. Table 3.1 shows the breakdown.

Table 3.1: Conspiracy subreddit corpus breakdown.

Subreddit	**Posts**	**%**	**Unique Authors**
r/conspiracy	44,411	76.2%	17,814
r/conspiracy_commons	9,698	16.6%	2,002
r/conspiracytheories	3,660	6.3%	2,456
r/ConspiracyII	515	0.9%	265
Total	**58,284**	**100%**	**20,688**

Why these four subreddits: The selection followed a deliberate principle. The thesis is interested in *general-purpose* conspiracy discussion, meaning communities where any conspiracy narrative can surface organically, rather than communities that are already

organised around a single topic or political position. r/conspiracy is the canonical example, being the oldest and largest conspiracy community on Reddit, covering everything from election fraud to vaccines to occult theories, and having been the subject of multiple prior studies on user overlap and gateway behaviour [25, 44]. r/conspiracy_commons and r/conspiracytheories emerged as overflow communities as r/conspiracy grew larger and its moderation became more contested. r/ConspiracyII was created explicitly as a less partisan alternative. Together, the four cover the main general-purpose conspiracy ecosystem on Reddit.

Several other communities were considered during the conspiracy data expansion but were deliberately excluded. Topic-specific communities like r/UFOs, r/ElectionIntegrity, and r/IsraelPalestine were excluded because they pre-filter their content around a single narrative, which would have biased the topic model toward whatever theme that community is built around. Debunking-oriented communities like r/Qult_Headquarters were excluded because their content is primarily *about* conspiracy theories rather than *endorsing* them, which would have muddied the signal. Purely political communities like r/politics and r/Conservative were excluded from the conspiracy side because they already appear in the political subreddit data used for ideology inference, and including them on both sides would have conflated the two signals. Finally, fringe or paranormal communities like r/HighStrangeness, r/C_S_T, and r/conspiracyNOPOL were excluded either because they lean heavily toward paranormal or spiritual content rather than political conspiracy, or because they explicitly prohibit political discussion, which would have introduced a selection effect against the election-related content this thesis focuses on.

Selection bias: The choice of these four English-language subreddits introduces several biases that should be acknowledged. First, because the thesis targets the 2024 U.S. presidential election, the data is inherently U.S.-centric and English-language. Conspiracy discourse in other languages and about non-American politics is not captured. Second, users who discuss conspiracy theories outside of these four subreddits (for example, in comment threads on r/politics or in niche communities not included in the selection) are invisible to the pipeline unless they also posted in one of the four. Third, the political subreddit classification list itself introduces a labelling bias, as borderline subreddits like r/PoliticalDiscussion (42.8% left, 42.6% right in the data) could reasonably be classified differently, and any such reclassification would shift the ideology vectors of users who post there. Fourth, the four conspiracy subreddits differ in size by two orders of magnitude (r/conspiracy has 86$\times$ more posts than r/ConspiracyII), so the corpus is heavily dominated by r/conspiracy's culture, moderation norms, and user base. These biases do not invalidate the analysis, but they do mean that the findings should be read as describing *this specific ecosystem* during *this specific period*, not as universal claims about conspir-

acy engagement on Reddit as a whole.

3.1.2 Political Subreddit Data

Political subreddit data (submissions and comments) was obtained from Academic Torrents [1]. The subreddit ideology classification list, which determines which subreddits count as left-leaning, right-leaning, or centre, was sourced separately from the Harvard Dataverse [46], with its lineage drawing on established Reddit-ideology resources such as those used by [47, 55]. The full classification list covers 424 subreddits (157 left, 39 centre, 228 right). Two filters then reduce this list to the operational working set used downstream. First, only 212 of the 424 subreddits had at least one post or comment in the May-November 2024 window in the Academic Torrents dump. Second, only 161 of those 212 had at least one of the 55,154 overlap users (users active in both political and conspiracy subreddits) posting in them, the rest are politically active but their participants do not appear in any conspiracy community. This 161-subreddit subset (86 left, 27 centre, 48 right) is the set that actually drives ideology inference and is the one referred to as the working classification set throughout the rest of the thesis. The political data is only used to infer user ideology, its content is not analysed for topics.

3.2 Data Collection Timeline

The data collection and preparation happened in several stages.

- **September 2025:** Compiled the lists of left, right, and centre subreddits. Scraped posts and comments from all of Reddit for the May-November 2024 window, and separately from r/conspiracy. Filtered everything into per-ideology CSV files, keeping relevant fields (`id`, `subreddit`, `author`, `created_utc`, `title`, `selftext`, `score`, `num_comments`, etc.). Also did initial keyword matching on r/conspiracy content with regular expressions to find election-related discussion.
- **March 2026:** Added r/conspiracy_commons, r/ConspiracyII, and r/conspiracytheories to the corpus. Merged all four subreddits into a unified pipeline.

3.3 Exploratory Data Analysis

Before building the methodology pipeline, an exploratory analysis was run across several dimensions to understand the dataset.

3.3.1 Corpus Overview

Table 3.2 summarises how the dataset narrows through each processing step, and Figure 3.1 visualises this funnel.

Table 3.2: Data processing pipeline stages.

Stage	Count	Notes
Raw conspiracy posts	58,284	All 4 subreddits
Users with ideology vectors	55,154	Matched to political activity
High-value users (HVU)	7,798	Top 14.1% by ensemble score
Topic-clustered posts (incl. noise)	13,194	BERTopic on HVU posts
Topic-clustered posts (excl. noise)	7,212	54.7% retained, 14 topics
Users in topic clusters	1,354	Used for topic-level analysis

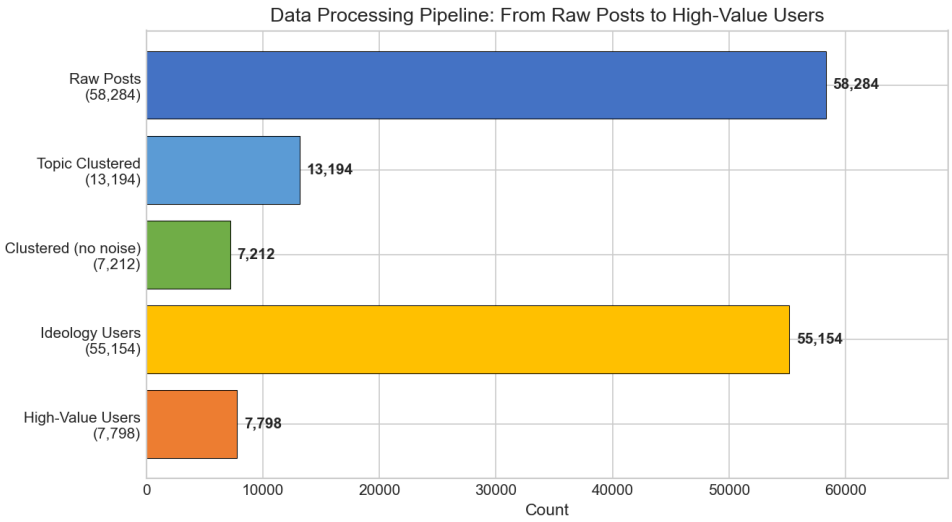


Figure 3.1: Pipeline funnel: from raw posts to the final analysis population.

The post-level funnel above answers the question "how much content survives each stage". The complementary question, "how many *users* survive each stage", is worth charting separately because every RQ in Chapter 5 operates on a user-level population. Figure 3.2 shows this progression. After the HVU selection step, the pipeline branches into two independent filters rather than remaining a single linear chain, where one branch retains users with a clear partisan lean regardless of whether they posted in named topics (the RQ5 population of 6,299), while the other retains users with at least one named topic post regardless of lean (1,354). The final RQ1/RQ2/RQ4 population of 1,106 sits at the intersection of both filters, requiring both a clear lean and a topic post. *The HVU selection criteria are formally defined in Section 4.3, the partisan-lean labelling in Section 4.2.2, and the named topic clusters in Section 4.4.4.*

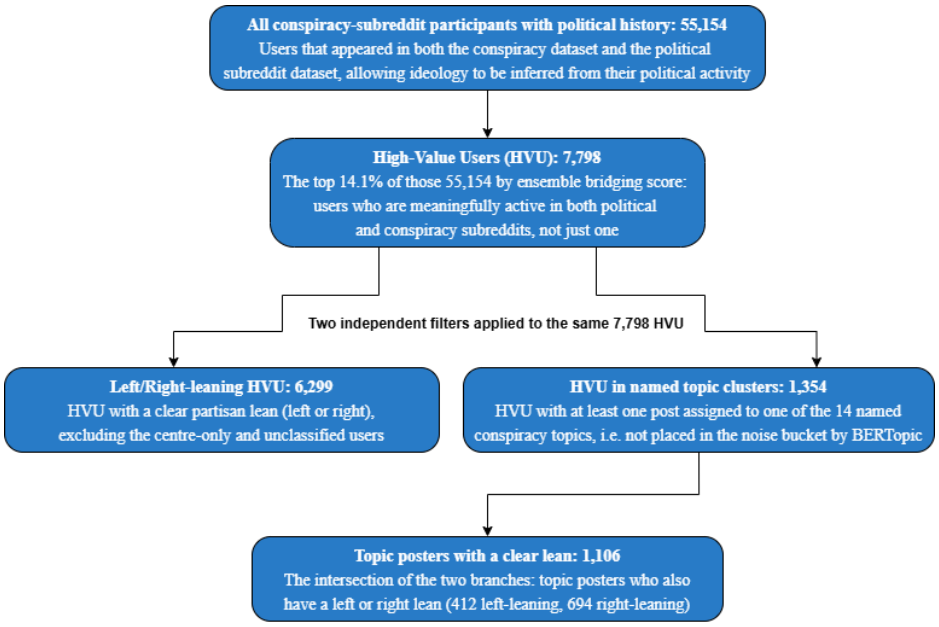


Figure 3.2: User-level distillation funnel. After HVU selection (7,798 users), the pipeline branches into two independent filters, where the left branch retains users with a clear partisan lean for RQ5's bridging-score analysis (6,299), while the right branch retains users with at least one named conspiracy topic post for the topic-level tests (1,354). The final RQ1/RQ2/RQ4 population (1,106) sits at the intersection of both filters.

Figure 3.3 shows the raw post volume across the four conspiracy subreddits. r/conspiracy dominates the corpus with 76.2% of all posts.

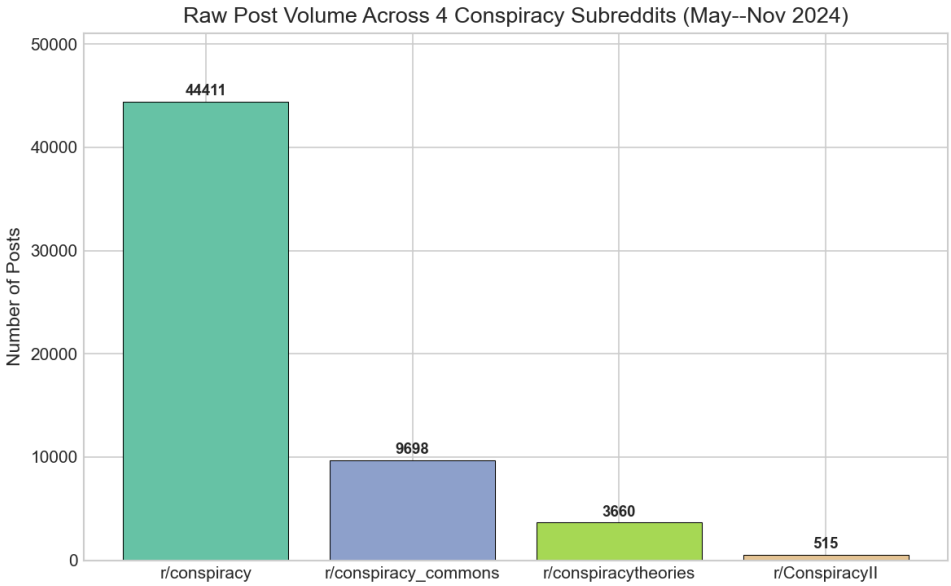


Figure 3.3: Post volume per conspiracy subreddit. r/conspiracy dominates the corpus.

Looking beyond posts, the conspiracy corpus also includes 1,836,935 comments. Figure 3.4 shows the posts-vs-comments breakdown across both the conspiracy and political

datasets. Comments vastly outnumber posts in all cases, with r/conspiracy alone accounting for over 1.6 million comments. The political dataset is an order of magnitude larger, with over 33 million comments and 929,000 posts across the three ideology categories. *The left/centre/right grouping used here is the Harvard Dataverse subreddit classification introduced in Section 3.1.2, not yet the per-user behavioural ideology defined later in Section 4.2.*

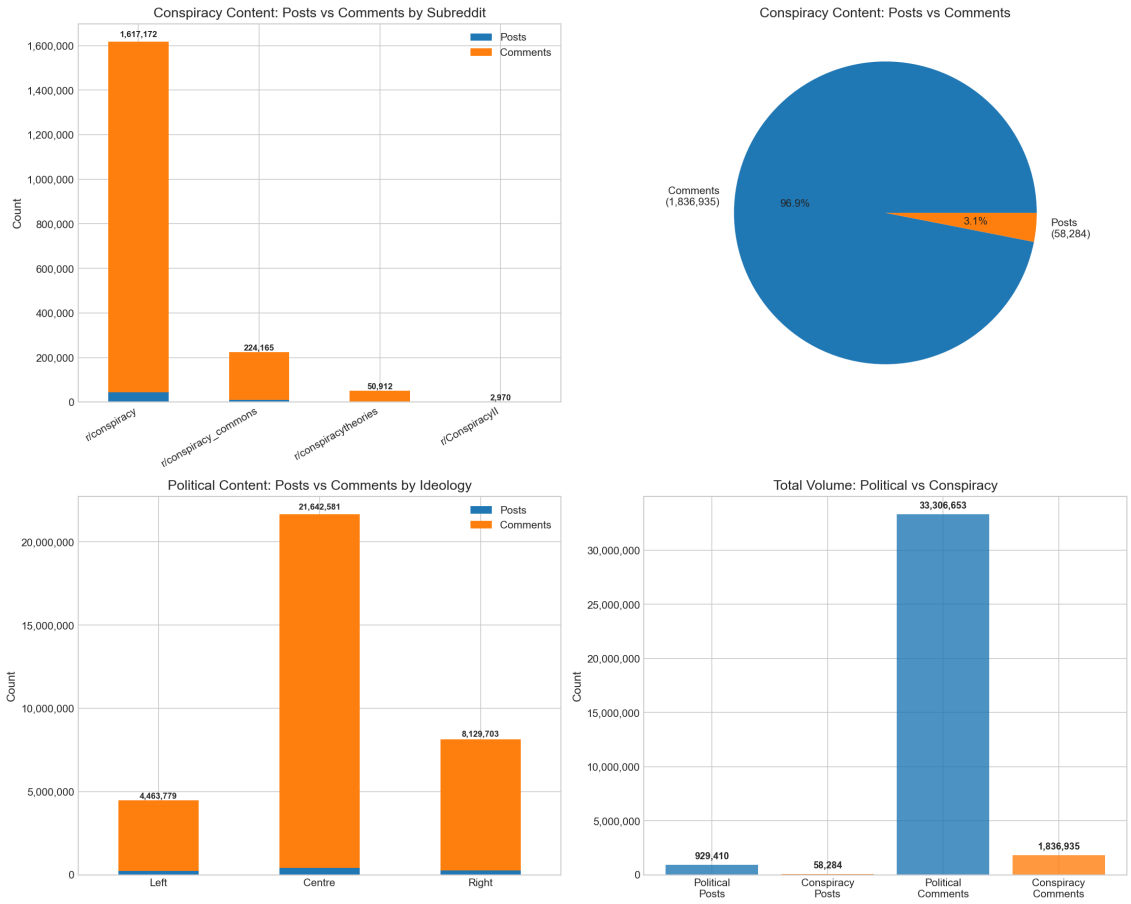


Figure 3.4: Posts vs. comments breakdown across conspiracy subreddits (top) and political subreddits by ideology (bottom), with total volume comparison.

3.3.2 Author Activity Distribution

Author activity follows a power-law distribution with a slope of -1.35 (Figure 3.5), meaning a small number of users post a lot while most barely contribute. 68.2% of authors made just a single post, while the most active author posted 1,971 times (3.4% of the total corpus). Only 3.5% of authors (728 users) made 10 or more posts.

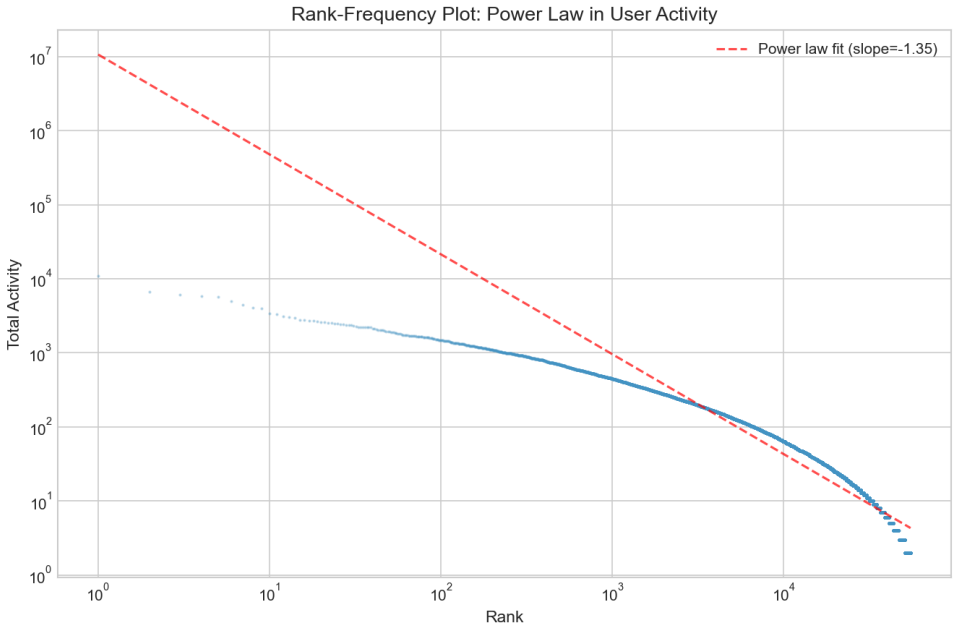


Figure 3.5: Rank-frequency plot of user activity (log-log scale) with power-law fit ($\alpha = -1.35$).

Figure 3.6 shows the distribution of each user's *conspiracy-to-political activity ratio*, which is simply the total number of conspiracy posts and comments divided by the total number of political posts and comments. In plain terms, this ratio answers the question "for every action this user takes in a political subreddit, how many do they take in a conspiracy subreddit?" A ratio of 0.1 means the user posts roughly ten times more in political communities than in conspiracy ones, a ratio of 1.0 means they split their time evenly, and a ratio of 5.0 means they are five times more active on the conspiracy side. The x-axis uses a log scale because the ratios span several orders of magnitude. The vast majority of users are politically dominant (ratio well below 0.5, shown by the blue dashed line), which is expected given the much larger volume of political data. Only a small fraction exceed the 2.0 threshold (red dashed line), indicating conspiracy-heavy users.

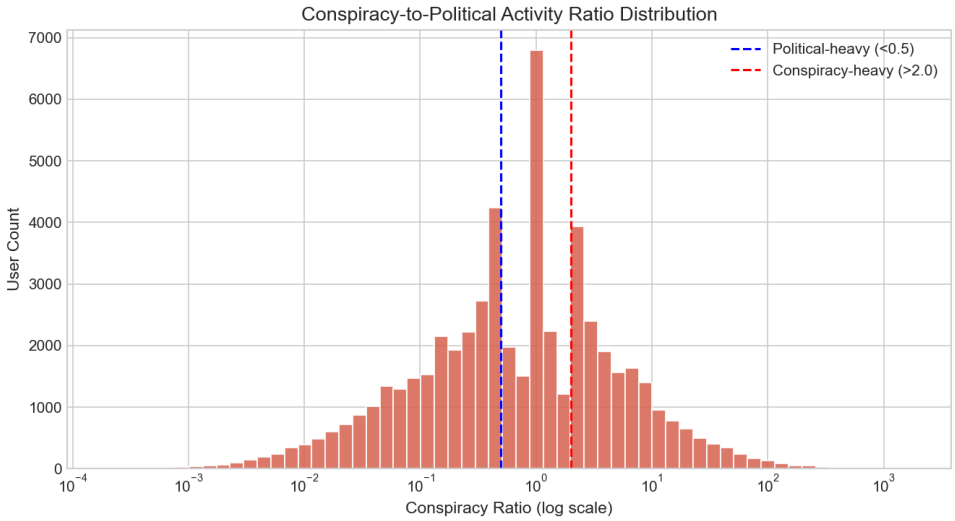


Figure 3.6: Conspiracy-to-political activity ratio distribution. Most users are politically dominant, with a small portion being conspiracy-focused.

3.3.3 Content Composition

Table 3.3 summarises the key post-level statistics across the corpus. For context, Reddit exposes three engagement metrics for every post, specifically the *score* (net upvotes minus downvotes), the *number of comments* (total comment count), and the *upvote ratio* (fraction of votes that are upvotes, where 1.0 means unanimously upvoted and 0.5 means evenly split). The subreddits differ quite a bit in content type (Figure 3.7). r/conspiracytheories is mostly text-based (56.0% self-posts), while r/conspiracy_commons is mostly links (80.5% link posts). Post scores are heavily right-skewed (skewness = 10.58), with 65.3% of posts having a score of 1 or less while the top 10% (score $\geq$ 98) average 515.6 points. For context, self-posts are strictly text without any embedded media, while link-posts can include anything a text-post has along with any external link or media (e.g., images, videos, etc.).

Table 3.3: Post-level statistics across the full corpus.

Metric	**Mean**	**Median**	**Std**	**Max**
Score	57.3	1	261.7	8,733
Comments	29.1	5	93.8	2,762
Upvote ratio	0.61	0.59	0.27	1.00
Text length (chars)	493	129	1,411	40,034

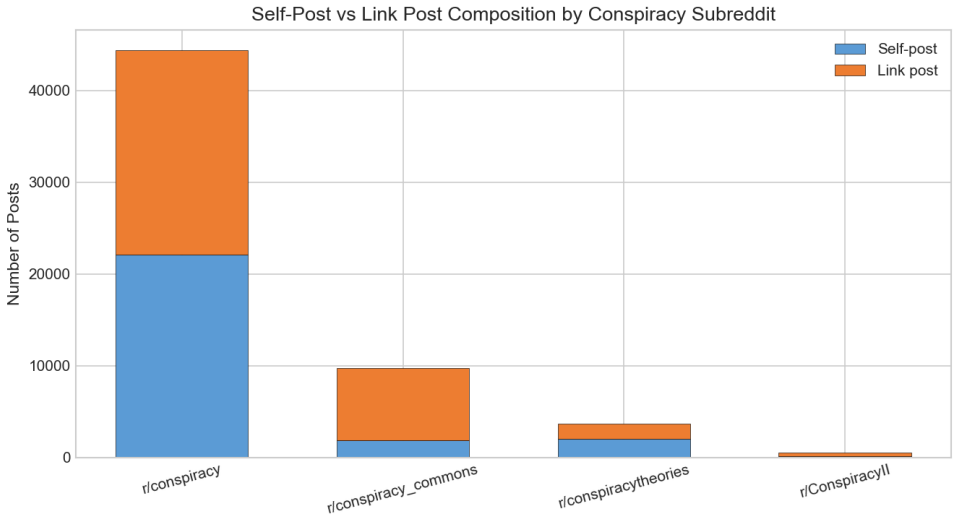


Figure 3.7: Content type breakdown (self-post vs. link-post) per subreddit.

Figure 3.8 compares the four subreddits across four engagement metrics (post score, number of comments, upvote ratio, and text length). r/conspiracy_commons has the highest median scores, while r/conspiracytheories posts tend to be longer, consistent with its higher self-post ratio.

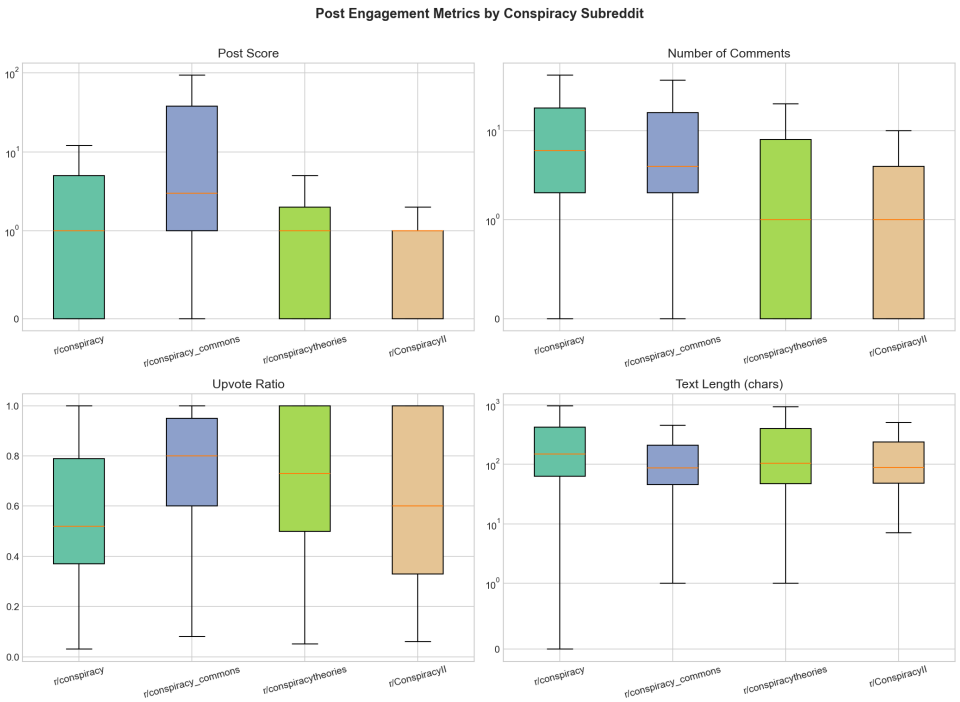


Figure 3.8: Post engagement metrics (score, comments, upvote ratio, text length) per conspiracy subreddit.

Post score and number of comments are strongly correlated (Pearson $r = 0.703$), as shown in Figure 3.9. This suggests that higher-scoring posts generate more discussion, making both metrics useful proxies for engagement in the user scoring pipeline.

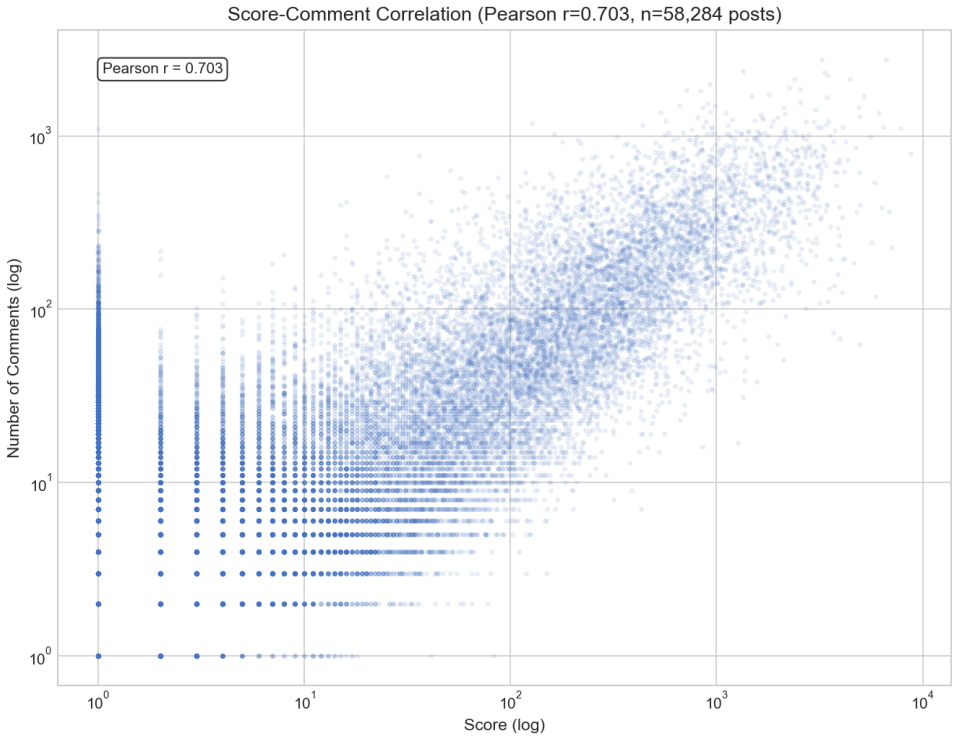


Figure 3.9: Score vs. comments per post (log-log scale). Pearson $r = 0.703$.

3.3.4 Temporal Patterns

Figure 3.10 shows the monthly post volume across the six-month window. July 2024 is the dominant month with 13,416 posts (23.0% of the total), possibly driven by the Trump assassination attempt on July 13. November 2024 only has 1,245 posts since the collection window ends on November 5.

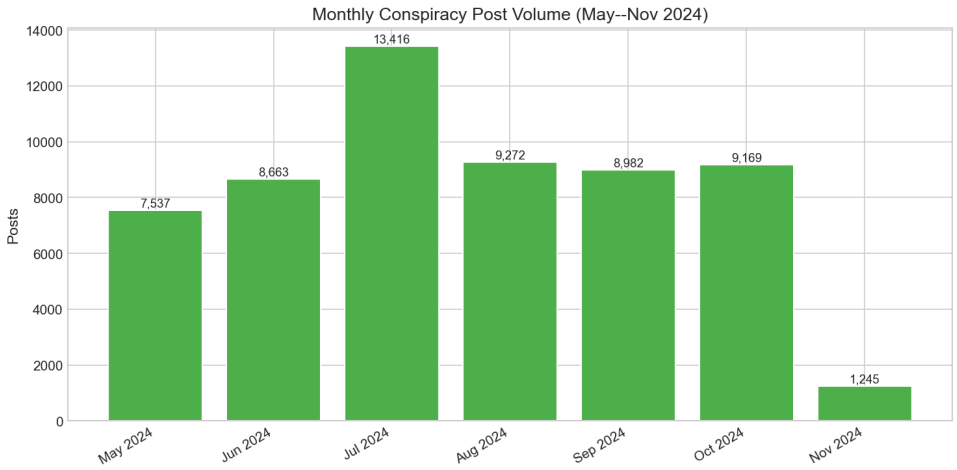


Figure 3.10: Monthly conspiracy post volume. July 2024 dominates, driven by the Trump assassination attempt.

At the day-of-week level (Figure 3.11), posting activity is fairly uniform. Sunday is

the busiest day (8,916 posts) and Saturday the quietest (7,657 posts), but the difference is only about 15%.

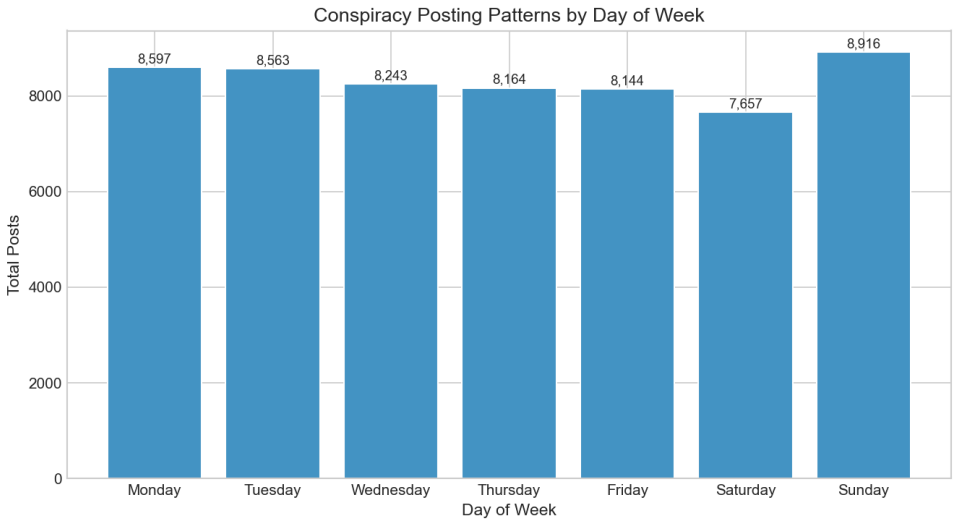


Figure 3.11: Conspiracy posting volume by day of week. Activity is fairly uniform, with a slight dip on Saturdays.

At the hour-of-day level (Figure 3.12), activity peaks at 17:00 UTC (12:00-13:00 US Eastern) and drops to its lowest around 08:00-09:00 UTC (03:00-04:00 Eastern), confirming that the user base is predominantly US-based.

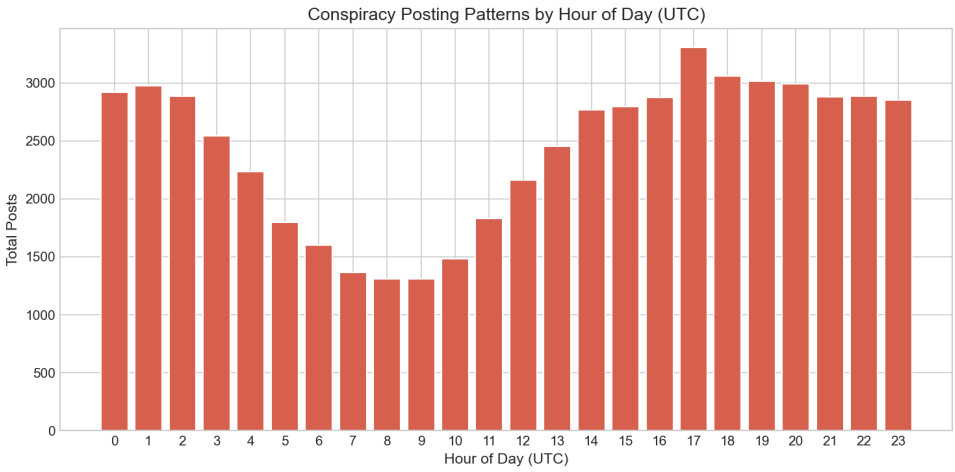


Figure 3.12: Conspiracy posting volume by hour of day (UTC). The pattern confirms a predominantly US-based user base.

The bigger story is in the event-driven spikes shown in Figure 3.13. The busiest single day was July 14, 2024 (1,518 posts), the day after the Trump assassination attempt.

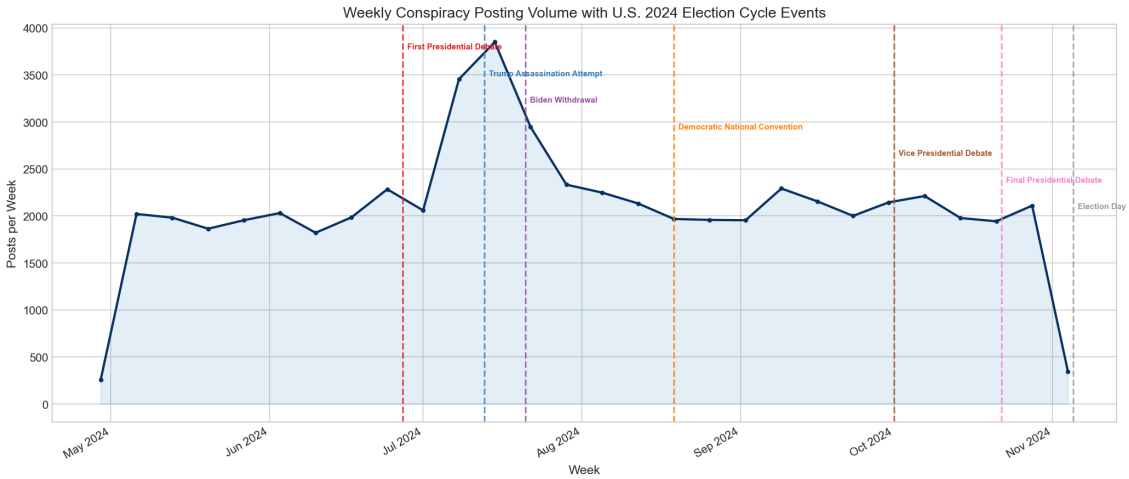


Figure 3.13: Daily posting volume with key political events annotated.

Interestingly, all ideology groups follow the same temporal patterns. The spikes happen at the same time regardless of user ideology, as shown in the stacked area chart in Figure 3.14. *The per-user ideology labels used to colour the bands are derived from the behavioural inference procedure formally defined in Section 4.2.*

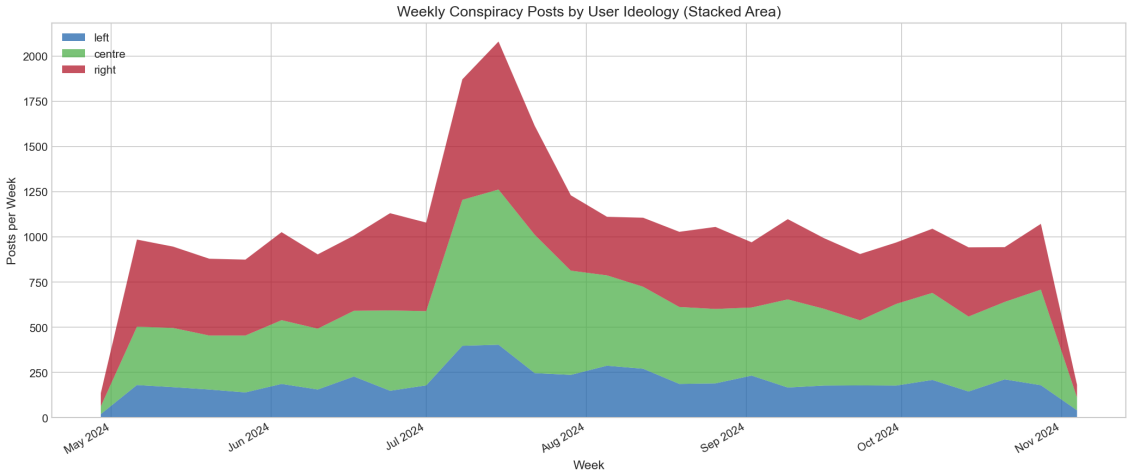


Figure 3.14: Weekly conspiracy posts by user ideology (stacked area). All groups spike together around the same events.

Figure 3.15 overlays weekly political activity (left, right, and centre subreddits) alongside conspiracy activity. The left y-axis shows political volume (hundreds of thousands per week), while the right y-axis shows conspiracy volume (thousands per week) on a separate scale so both trends are clearly visible. Despite the roughly 100× difference in absolute volume, all four lines spike together around the same events, particularly the Trump assassination attempt in mid-July 2024. The key takeaway is the temporal alignment, where conspiracy posting responds to the same real-world events that drive political discussion. *Left, right, and centre here refer to the Harvard Dataverse subreddit grouping*

introduced in Section 3.1.2.

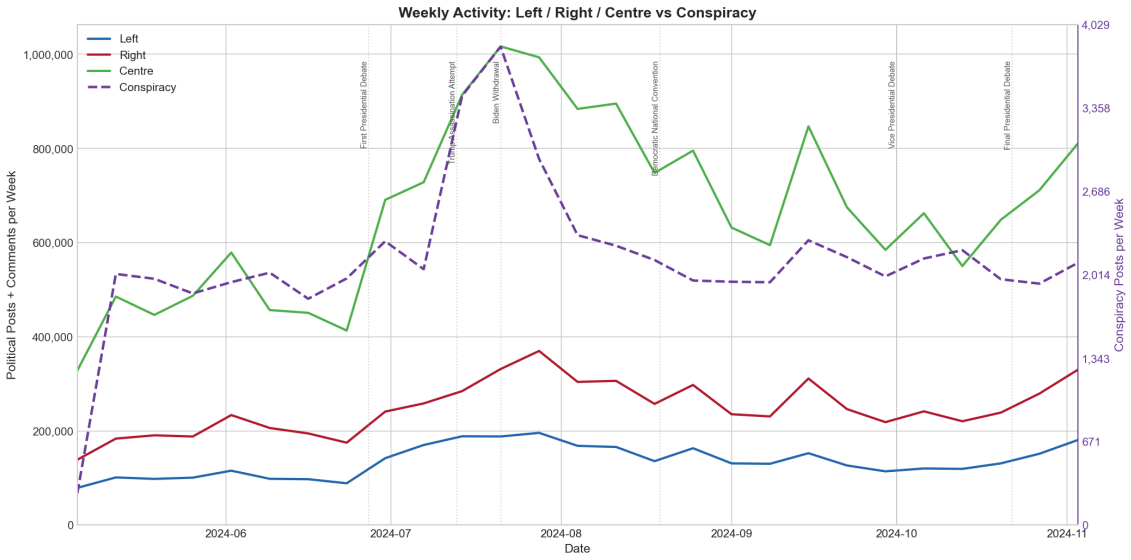


Figure 3.15: Weekly activity across political ideology categories (left axis, solid lines) and conspiracy subreddits (right axis, purple dashed). All groups respond to the same events.

Figure 3.16 breaks down the temporal patterns by conspiracy topic (top 8 topics by volume). Topic T1 (Trump, Biden, and Elections) shows the sharpest spike around July 2024, while other topics like T2 (Vaccine Data and Covid-19) and T6 (Russia, Ukraine, and Geopolitics) remain relatively stable throughout the period. *The topic labels and assignments are produced by the BERTopic pipeline detailed in Section 4.4.4.*

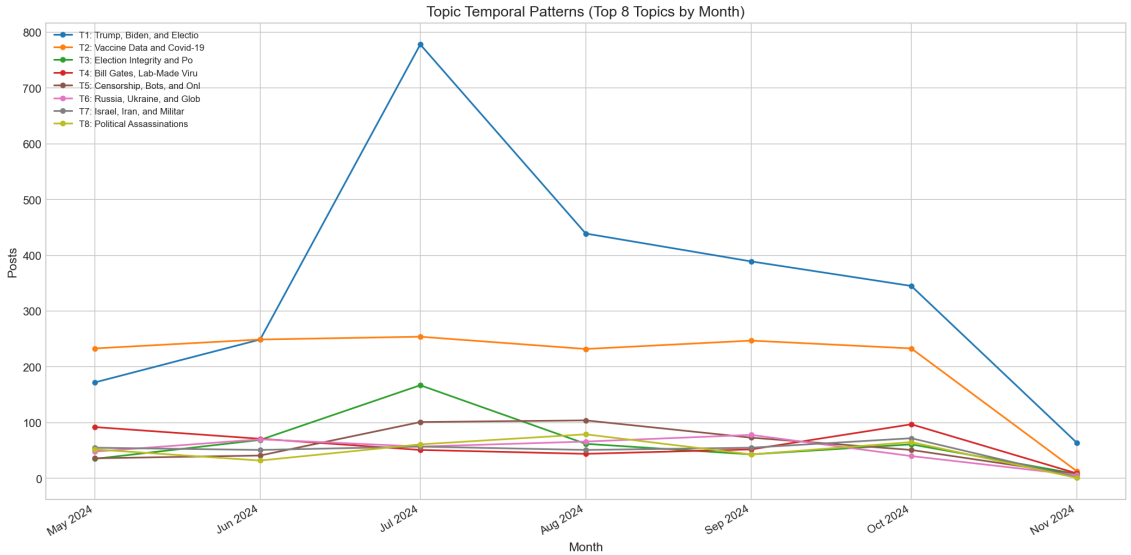


Figure 3.16: Monthly posting volume by conspiracy topic (top 8). T1 (Trump/Elections) dominates the July spike.

3.3.5 Political Subreddit Landscape

Figure 3.17 shows the 20 most popular political subreddits among conspiracy-posting users. General-purpose subreddits like r/politics and r/worldnews dominate, but explicitly partisan communities like r/Conservative, r/LateStageCapitalism, and r/AskThe_Donald also appear. Figure 3.18 breaks down each of these subreddits by the ideology of their users. *The per-user ideology assignment used in this breakdown is the behavioural inference scheme formally defined in Section 4.2, with the "Other" slice corresponding to balanced users whose activity does not concentrate in any one ideology group.*

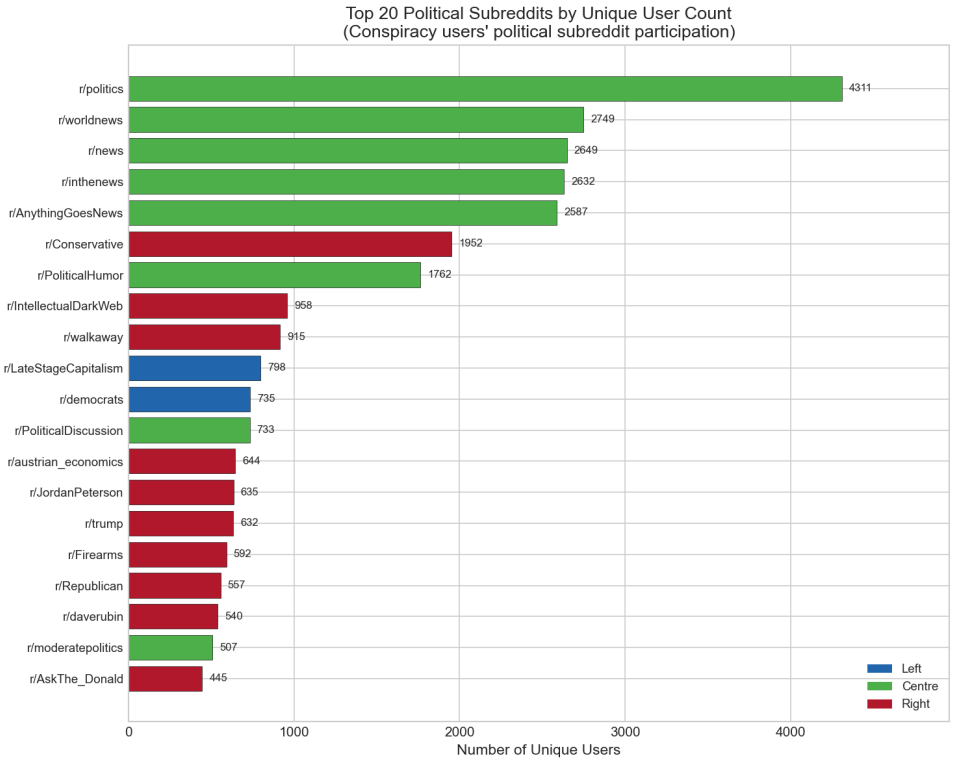


Figure 3.17: Top 20 political subreddits by unique conspiracy-posting user count.

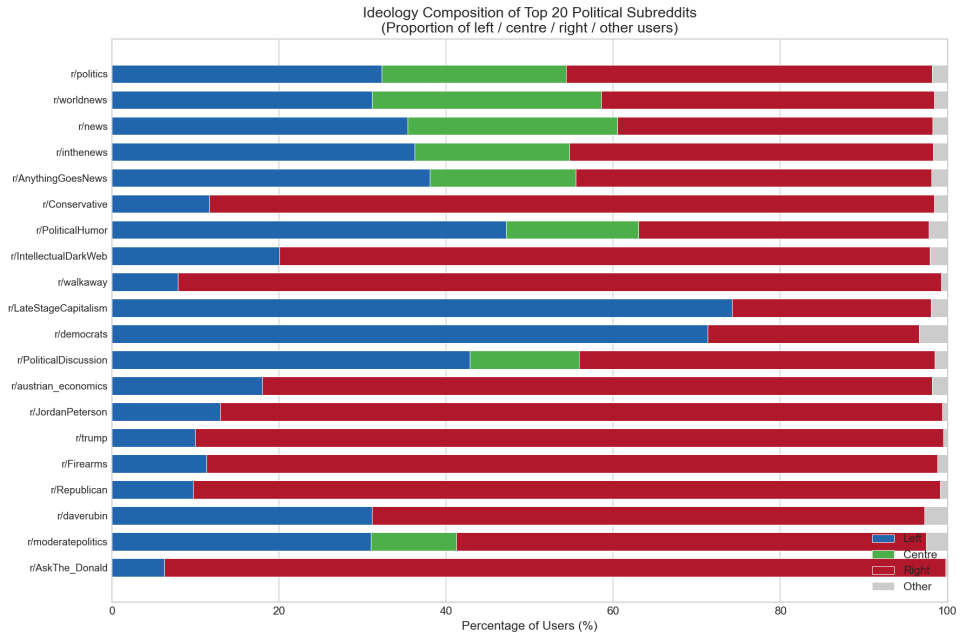


Figure 3.18: Ideology composition of the top 20 political subreddits. Colours show left/centre/right proportions, with the "Other" slice capturing balanced users whose activity is split roughly evenly across ideology groups and therefore do not fall into a single primary lean.

For a broader view, Figure 3.19 shows the 50 most active political subreddits (out of the 212 with data in the study window), colour-coded by their primary ideology under the Harvard Dataverse classification. As described in Section 3.1.2, the full Harvard list covers 424 subreddits, 212 of which have at least one post or comment in the May-November 2024 window, and 161 of which (the working subset used for ideology inference) have at least one of the 55,154 overlap users posting in them. The remaining 51 of the 212 have political activity in the window but contribute no edges to the user-subreddit network because no overlap user posts in them. The distribution is heavily right-skewed, a handful of large subreddits like r/politics and r/worldnews dominate the volume, while the long tail of smaller communities collectively covers a wide range of ideological niches.

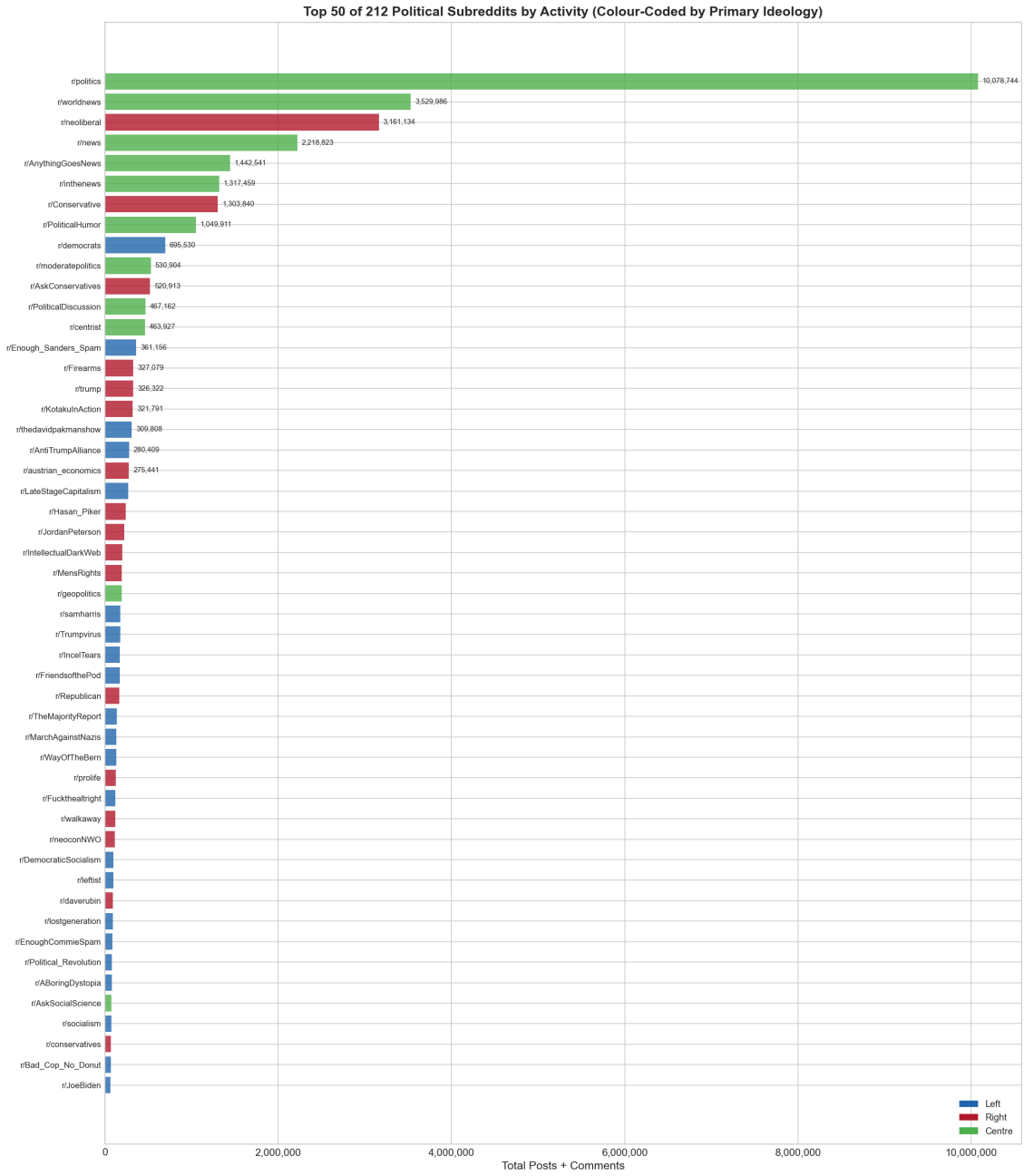


Figure 3.19: Top 50 political subreddits by total activity, colour-coded by primary ideology under the Harvard Dataverse classification (blue = left, red = right, green = centre). The Harvard list covers 424 subreddits in total, of which 212 have data in the study window and 161 (the working subset used for ideology inference) host at least one post by a user in our 55,154-user overlap set.

Table 3.4 shows the ideology split for a few notable subreddits, ranging from strongly right-leaning (r/AskThe_Donald: 93.5% right) to strongly left-leaning (r/LateStageCapitalism: 74.2% left). Table 3.5 gives the category-level breakdown of the full 161-subreddit classification set used for ideology inference.

Table 3.4: Ideology composition of selected political subreddits.

Subreddit	Left %	Centre %	Right %
r/AskThe_Donald	6.3%	0.0%	93.5%
r/walkaway	7.9%	0.0%	91.4%
r/Conservative	11.7%	0.0%	86.7%
r/politics	32.3%	22.1%	43.8%
r/PoliticalDiscussion	42.8%	13.1%	42.6%
r/LateStageCapitalism	74.2%	0.0%	23.8%

Table 3.5: Category-level breakdown of the political subreddit classification list. The first column shows the original 424-subreddit Harvard Dataverse list, the second shows the 161-subreddit working subset that actually contributes to ideology inference (those with at least one post by a user in our 55,154-user overlap set). The full subreddit list is available in the project repository.

Category	Full list	Working subset	Examples
Left-leaning	157	86	r/LateStageCapitalism, r/SandersForPresident, r/socialism
Right-leaning	228	48	r/Conservative, r/AskThe_Donald, r/walkaway
Centre	39	27	r/politics, r/PoliticalDiscussion, r/worldnews
Total	**424**	**161**	

Figure 3.20 shows how many unique political subreddits each user participates in. The median is 4 subreddits and the mean is 5.1, with a right-skewed distribution, most users stick to a handful of political communities, but some are active in 15 or more. This breadth measure is later used as the basis for the political diversity (entropy) score discussed in Chapter 4.

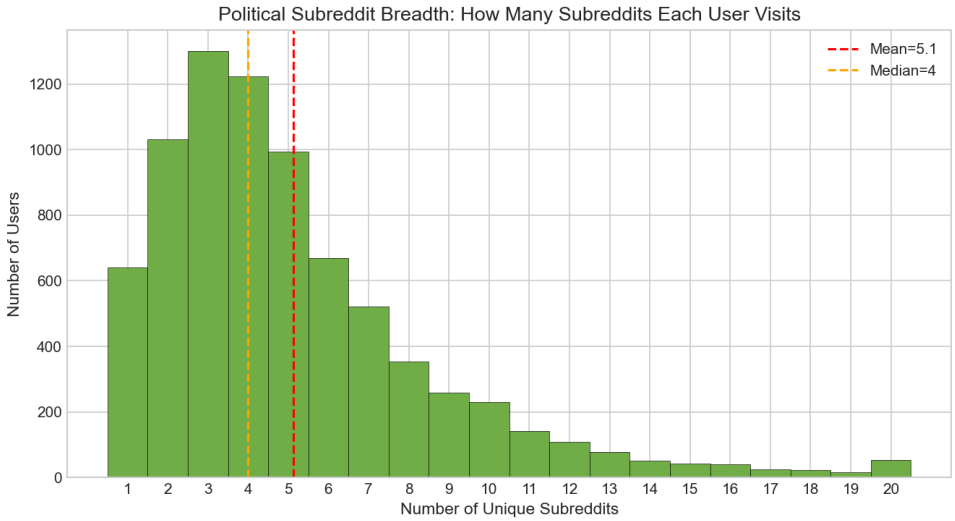


Figure 3.20: Distribution of unique political subreddits per user (mean = 5.1, median = 4).

3.3.6 Ideology Distribution

Each of the 55,154 overlapping users is assigned a *primary ideology* (left, centre, or right) based on whichever ideology group accounts for the most activity in their political subreddit history. The underlying data is the user's *ideology vector*, a three-number summary (p_L, p_C, p_R) that records what fraction of their political subreddit activity falls in left, centre, and right communities respectively. Think of it as a fingerprint of where someone spends their political time on Reddit. A user with $(0.05, 0.10, 0.85)$ is almost exclusively active in right-leaning subreddits, while one with $(0.40, 0.15, 0.45)$ spreads their activity across both sides with a slight right tilt. The primary ideology label is just the coarsest reading of this vector, whichever of the three components is largest. The full construction of these vectors, including derived metrics like entropy and dominance, is detailed in Section 4.2. For the purposes of this exploratory section, the primary label is sufficient. Figure 3.21 shows the result: 60.1% centre, 25.4% right, and 14.5% left. Among users who have a clear non-centre lean, 59.8% lean right and 40.2% lean left.

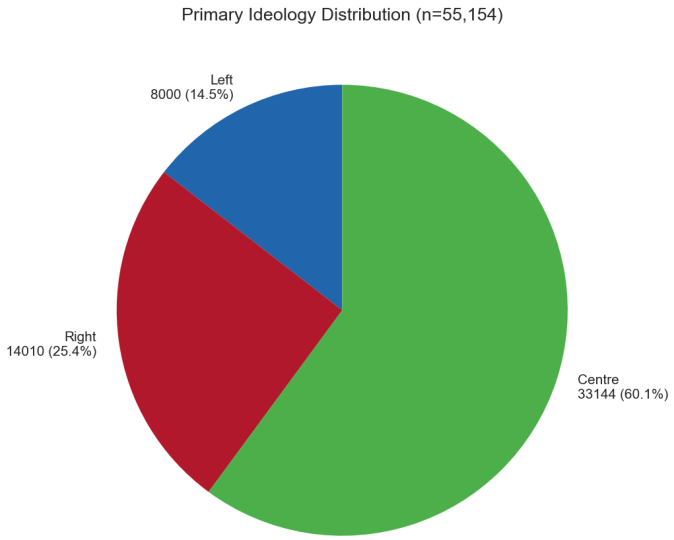


Figure 3.21: Primary ideology distribution of users with ideology vectors ($n = 55{,}154$).

For finer-grained analysis, users with a non-centre primary ideology are further divided into six confidence tiers (strong, lean, and weak for each direction) based on how dominant that ideology is in their activity profile. This six-tier scheme groups all strong/lean/weak left users into the "left" hard label and likewise for the right, so the hard labels are simply the collapsed version of the six tiers. Centre users are excluded from the six-tier scheme because they do not lean in either direction. Figure 3.22 shows the distribution across the 33,567 non-centre users: peaks appear at the extremes (strong_left: 11.5%, strong_right: 28.3%) and in the weak tiers (weak_left: 26.1%, weak_right: 26.4%), forming a bimodal

pattern. The remaining 21,587 users (the 60.1% centre population from Figure 3.21) are not shown here. *The exact thresholds that define strong, lean, and weak tiers are given in Section 4.2.2.*

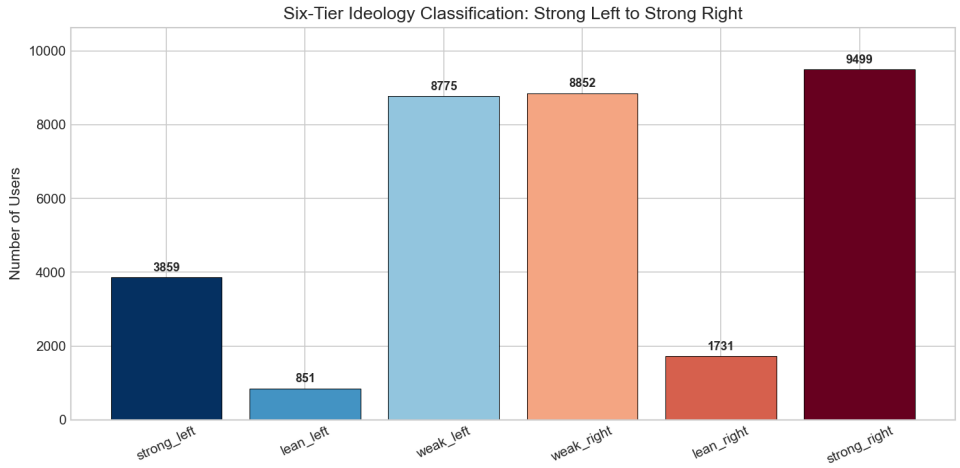


Figure 3.22: Six-tier ideology distribution across the 33,567 non-centre users. Strong/lean/weak left collapse into the "left" hard label, and likewise for right.

Figure 3.23 plots each user's total political activity against their conspiracy activity on a log-log scale, coloured by primary ideology. The "None" category in the legend refers to a small number of users whose ideology vector is exactly tied between two groups, so no single primary ideology could be assigned, in practice this affects 0% of users. The scatter shows that users span a wide range of activity levels in both domains, and no ideology group clusters in a distinct region, left, right, and centre users are intermixed throughout. *Primary ideology here, and in the following three figures, is derived from the per-user ideology vectors defined in Section 4.2.*

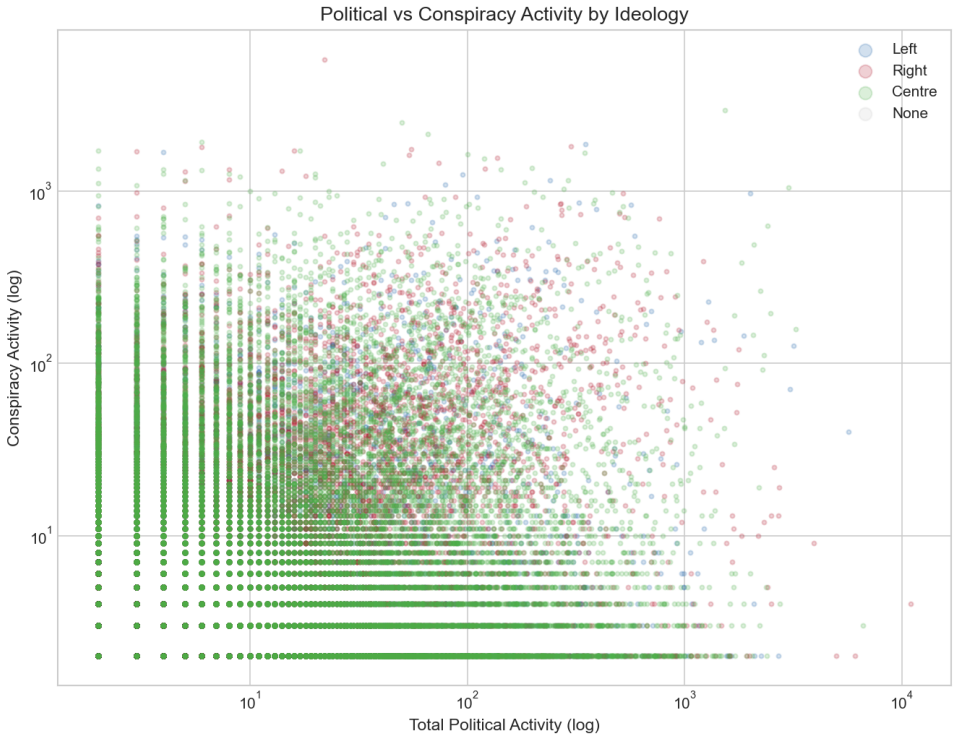


Figure 3.23: Political vs. conspiracy activity per user (log-log scale), coloured by primary ideology. All groups are intermixed.

Nearly half of all users (48.5%) participate in at least one political subreddit aligned with a different ideology than their own (Figure 3.24). This high rate of cross-ideology exposure suggests that the conspiracy-posting population is not strongly siloed by political affiliation.

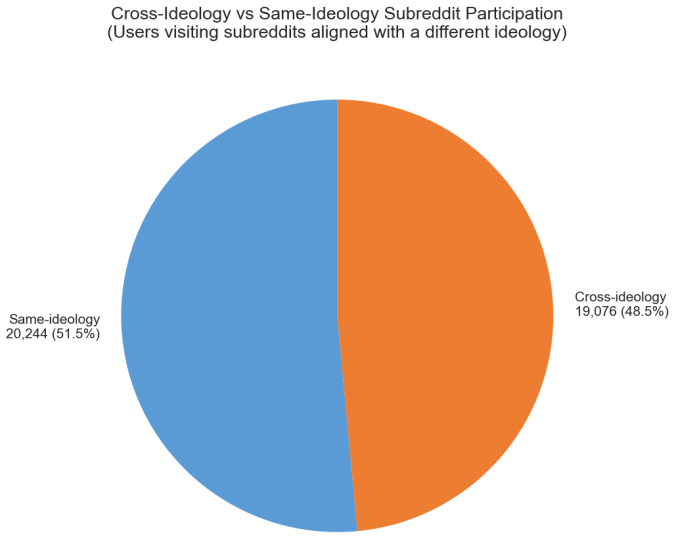


Figure 3.24: Cross-ideology vs. same-ideology subreddit participation. Nearly half of users post in subreddits aligned with a different ideology.

To quantify how evenly a user spreads their activity across ideological groups, we compute *normalised Shannon entropy* [45], which is referred to as the *political diversity score* throughout the rest of the thesis. In intuitive terms, this score answers the question "does this user stick to one side, or do they participate across the political spectrum?" A score of 0 means the user posts exclusively in subreddits of one ideological leaning, for example only in left-leaning communities. A score of 1 means their activity is split perfectly evenly across left, centre, and right subreddits. Most real users fall somewhere in between, with lower values indicating a more politically concentrated footprint and higher values indicating someone who engages broadly. Figure 3.25 shows this distribution across all users. It is heavily zero-inflated, most users participate in subreddits of only one ideological leaning (entropy near 0), but a secondary peak around 0.5-0.6 captures users who spread their activity more evenly. The mean score is 0.194, which reflects the fact that most conspiracy-posting users are politically concentrated rather than omnivorous. This metric turns out to be more predictive of conspiracy engagement than ideological direction itself, as discussed in the RQ4 analysis in Chapter 5. *The formal definition of the normalised Shannon entropy used here is in Section 4.2.*

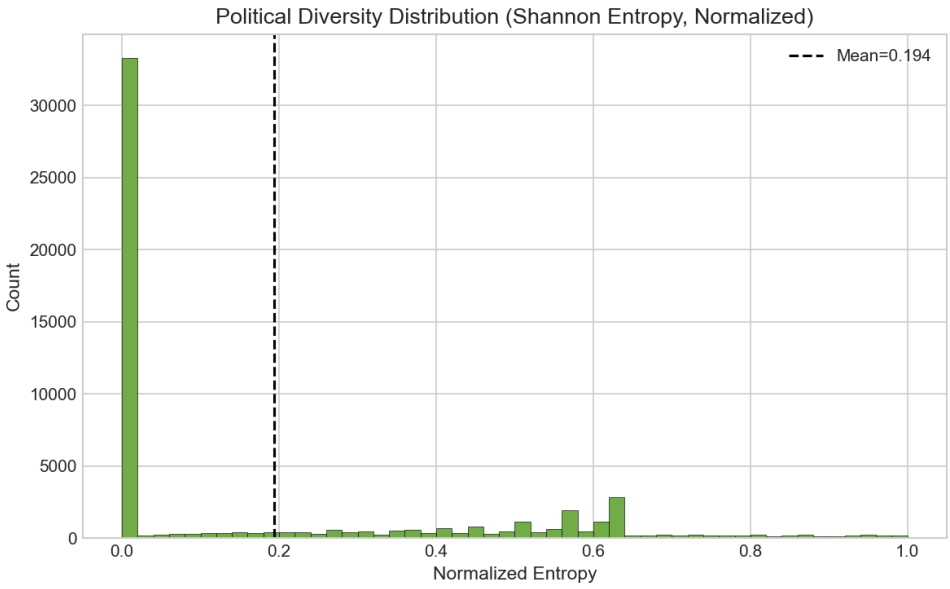


Figure 3.25: Distribution of political diversity (normalised Shannon entropy). Most users are ideologically concentrated (entropy near 0), with a secondary peak around 0.5-0.6.

Figure 3.26 examines whether ideological isolation predicts conspiracy engagement. Each point represents a user, with political diversity (normalised entropy) on the x-axis and conspiracy-to-political activity ratio on the y-axis. There is no clear trend, users with both low and high diversity show a wide range of conspiracy engagement levels, and all three ideology groups are intermixed throughout the scatter.

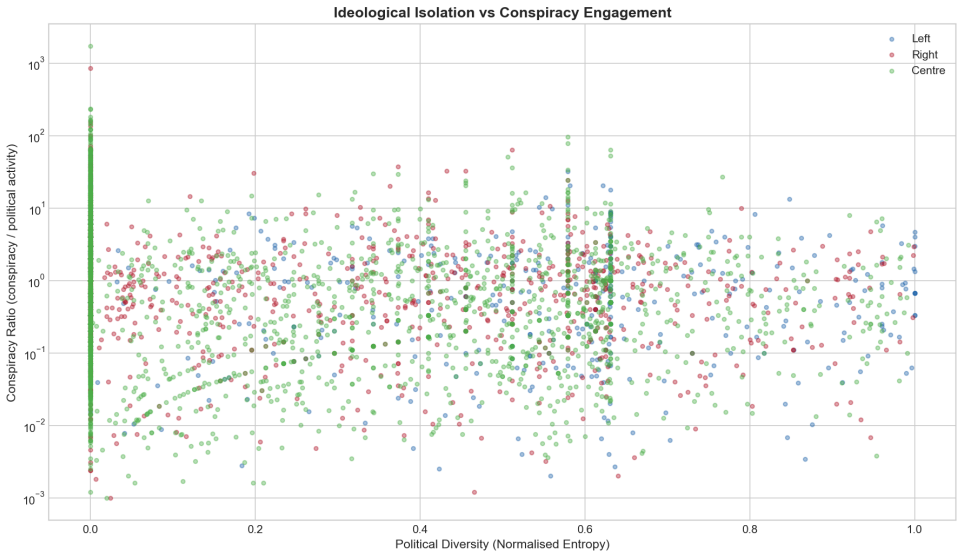


Figure 3.26: Ideological isolation (entropy) vs. conspiracy engagement ratio, coloured by ideology. No clear relationship is visible.

Finally, Figure 3.27 classifies all 55,154 users into four behavioural archetypes based on their ideology vectors and engagement patterns. *Highly Isolated* users (63.8%) have low entropy and high dominance in a single ideology. *Cross-Ideology* users (21.9%) participate substantially across ideological boundaries. *Conspiracy-Focused* users (3.0%) have more conspiracy activity than political activity. The remaining *Moderate* users (11.3%) fall between these categories. These archetypes come from an early exploratory categorisation used mainly to sanity-check the data, and were later superseded by the continuous ensemble scoring described in Section 4.3, they are kept here only as an intuitive summary of the population shape.

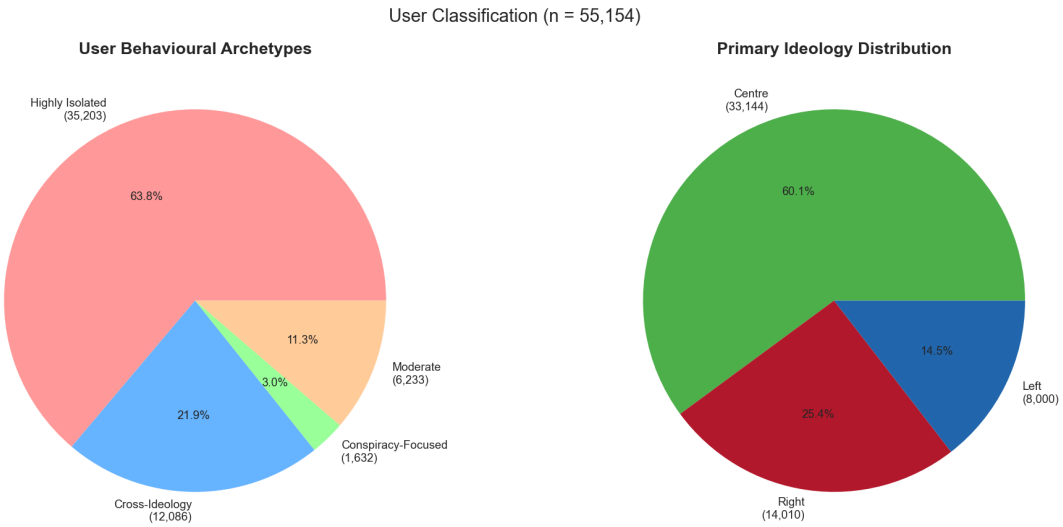


Figure 3.27: User behavioural archetypes (left) and primary ideology distribution (right). Most users are ideologically isolated, but nearly a quarter participate across ideological lines.

Chapter 4

Methodology

This chapter walks through the full pipeline built to answer the eight research questions from Chapter 1. As discussed in Section 1.2, the methodology evolved from an initial candidate-focused approach, identifying conspiracies about Trump, Biden, and Harris in left, centre, and right leaning subreddits, to a broader topic-level analysis after early experiments with unsupervised clustering showed that the conspiracy landscape does not organise neatly around individual candidates, even before a major event such as the 2024 presidential election. The pipeline initially started as a set of Jupyter notebooks and was then refactored into a modular Python codebase for easier maintenance and extension as the project evolved.

There are six main stages, each covered in its own section. Table 4.1 summarises each stage, its primary output, and which research questions it feeds into. Figure 4.1 gives a top-level view of how the six stages connect before each is described in detail.

Table 4.1: Pipeline stages, their outputs, and the research questions each one supports.

Stage	Section	Primary output	RQs served
1	Overlapping users (4.1)	55,154 users with dual activity	All
2	Ideology vectors (4.2)	(p_L, p_C, p_R) per user	RQ1-RQ7
3	User scoring (4.3)	7,798 high-value users	RQ1-RQ6
4	Topic clustering (4.4)	14 topics, 20 subtopics	RQ1-RQ4, RQ8
5	Networks (4.5)	Co-activity and bipartite graphs	RQ5-RQ7
6	Chrome extension (4.6)	Real-time classifier	RQ8

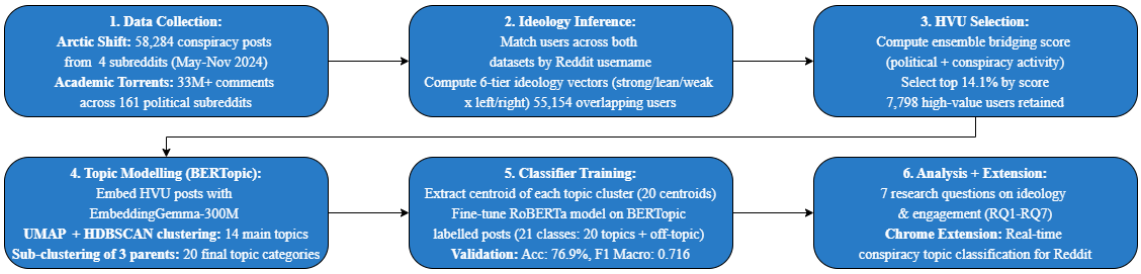


Figure 4.1: Top-level overview of the six-stage methodology pipeline. Stages 1-3 build the user population and ideology labels. Stages 4-5 produce the topic model and classifiers. Stage 6 applies them to the research questions and the Chrome extension.

All stages are wired together by a top-level orchestrator script (`run_all.py`) that executes them in order and writes intermediate artefacts to a shared `pipeline_results/` directory, so the full pipeline can be reproduced end-to-end with a single command. A companion script (`verify_outputs.py`) checks that every expected artefact is present and consistent with the previous stage's outputs before the next stage runs, which prevents silent corruption when a partial rerun is attempted after a code change.

4.1 Overlapping User Identification

This stage takes the raw conspiracy and political CSV files described in Chapter 3 and produces the set of users who appear in both domains. Every downstream stage depends on this overlap set, so it is the foundation of the entire pipeline.

The first step is to find users who were active in *both* at least one political subreddit and at least one conspiracy subreddit during the study period. These are the users for whom we can measure both ideology and conspiracy engagement. A user who only shows up in r/politics is not useful for this thesis, because we have no conspiracy posts of theirs to study. The same goes for someone who only posts in r/conspiracy with no political activity, we cannot place them on a left-right spectrum.

The very first prototype of this step did not even use subreddit membership as the signal. It scanned post titles and bodies for a hand-written list of regular expressions, and used Wikipedia and KeyBERT, covering candidate names and conspiracy keywords (e.g. "Trump", "Biden", "Harris", plus phrases like "rigged election" or "deep state") and flagged any user whose posts matched both a political and a conspiracy pattern. This worked for a quick sanity check but was fragile, since as the keyword lists kept growing, edge cases like sarcasm or negation were missed, and a user who discussed conspiracies in their own words without using any of the listed triggers was invisible to the filter. It was later replaced by the membership-based approach used here, which simply trusts that posting inside a classified subreddit is itself a stronger signal of domain activity than keyword matching on free text.

The membership-based version that followed initially only used r/conspiracy. After the corpus was expanded on the conspiracy side, the script was updated to handle all four conspiracy subreddits. The final overlap set contains **55,154 users** with activity in both domains.

4.2 Ideology Vector Computation

This stage takes the 55,154 overlapping users from Stage 1 and computes a continuous ideology profile for each one based on their political subreddit activity. The output, a three-dimensional probability vector per user, is the ideological representation used in every RQ from RQ1 through RQ7 and in the null model of RQ6.

Initially, users were given a single hard label of left, centre, or right based on which group of subreddits they posted in the most. This approach was simple and easy to understand, but it threw away a lot of nuance. Instead, the final approach, developed on the supervisor's suggestion, involved building *continuous ideology vectors* that show how a user's political subreddit activity is distributed across ideological categories. Think of it as a fingerprint of where someone spends their political time on Reddit, rather than stamping a user as "right-leaning", we record what fraction of their political activity lives in left, centre, and right subreddits. The reason continuous vectors were preferred over hard labels is straightforward. A hard label treats a user with 999 left-leaning actions and 1,000 right-leaning actions as simply "right-leaning", even though their activity is essentially balanced and they arguably belong in neither camp. The continuous version preserves that nuance, a user with 60% right-leaning and 40% left-leaning activity looks different from one with 95% right-leaning activity, but the hard-label system would call both of them "right". This matters for every downstream analysis, because the statistical tests in RQ1-RQ5 rely on the continuous vectors to measure effect sizes rather than just counting heads in binary buckets, and the null model in RQ6 permutes the full vectors rather than single-bit labels.

4.2.1 Participation Vectors

For each user, we count their posts and comments across the 161-subreddit working set defined in Section 3.1.2 and normalise by ideology group. This gives a three-number vector $\mathbf{p}_u = (p_L,\ p_C,\ p_R)$ that sums to one, where p_L, p_C, and p_R are the shares of the user's political activity in left, centre, and right subreddits respectively. For example, a user with $\mathbf{p}_u = (0.10,\ 0.05,\ 0.85)$ puts 85% of their political activity in right-leaning subreddits, 10% in left, and 5% in centre.

4.2.2 Confidence Tiers

The continuous vectors are also mapped to a discrete labelling scheme based on how dominant the strongest direction is, so plots and tables that need a single label can still fall back on one without losing the strength of the lean. Users whose primary activity is in centre subreddits or who are perfectly balanced across left and right are assigned the dedicated labels *centre-only* and *balanced* respectively, and the remaining users (those with a clear left or right lean) are placed into one of six strength tiers summarised in Table 4.2. An earlier draft of this mapping used a richer nine-label scheme with the labels *Strong Left*, *Lean Left*, *Centrist*, *Lean Right*, *Strong Right*, *Mixed*, *Partisan Mixed*, and *Inactive*, each with its own thresholds on dominance, entropy, and minimum activity. In practice most of those extra buckets were either redundant ("Mixed" and "Partisan Mixed" differed only by a thin dominance margin) or empty ("Inactive" duplicated the high-value filter from Section 4.3), so the L/R partition was simplified to the six strength tiers in Table 4.2 while *centre-only* and *balanced* were retained as standalone categories for users with no clear partisan lean.

Table 4.2: Six-tier scheme for mapping continuous vectors back to labels. Dominance is the share of the user's most active ideology ($\max(p_L, p_C, p_R)$), and entropy is the normalised Shannon entropy of (p_L, p_C, p_R) measuring how spread out the user's activity is across ideology groups (0 = all in one group, 1 = evenly split).

Tier	Condition	Interpretation
Strong left/right	entropy < 0.2 and dominance > 0.8	Near-pure single-ideology user
Lean left/right	entropy < 0.5 and dominance > 0.6	Clearly leaning, not pure
Weak left/right	any remaining non-centre lean	Slight tilt but mixed

When the thesis refers to hard left/centre/right labels in a plot, that is the labelling scheme collapsed: the three strength tiers on each side roll up into a single left or right bucket, and the *centre-only* and *balanced* categories roll up into centre (or, in left-vs-right tests, are excluded). This is made explicit wherever both versions appear so the reader knows whether a figure is using the fine-grained tiers or the coarse hard labels.

We also compute some derived metrics that describe the *shape* of the vector rather than a specific direction.

- **Normalised entropy:** how spread out a user's activity is across the three ideology groups. An entropy near 0 means the user posts entirely in one group, while an entropy near 1 means their activity is evenly split. It is computed from the Shannon entropy [45] of (p_L, p_C, p_R) divided by $\log 3$ so that it always falls in $[0, 1]$.
- **Dominance:** the share of the user's single most active ideology, that is, $\max(p_L, p_C, p_R)$. A dominance of 1.0 means the user only posts in one side, while a dominance of 0.33 means the activity is perfectly balanced across all three.

- **Left-right balance:** a score in $[-1,+1]$ computed as $p_R - p_L$. Negative values mean left-leaning, positive means right-leaning, and zero means balanced (or purely centre). It ignores the centre component on purpose, so that a user who is $(0.4, 0.2, 0.4)$ still reads as balanced.

4.3 User Scoring and High-Value User Filtering

This stage takes the 55,154 users with ideology vectors from Stage 2 and filters them down to the subset worth studying in depth. Its output, the high-value user (HVU) set of 7,798 users, defines the population for the topic clustering in Stage 4, the network construction in Stage 5, and the per-topic analyses of RQ1 through RQ5. Users who do not pass the filter still appear in the broader ideology statistics but are excluded from the topic-level and network-level tests where a minimum data density is required.

Not all users are equally useful for analysis. Someone who left one comment in r/politics and one post in r/conspiracy does not give us much signal compared to a user with hundreds of actions in both domains. To deal with this, a scoring system was developed that ranks users by their *bridging value*, meaning how much political activity, conspiracy activity, and balance between the two they have. The users at the top of this ranking are the ones worth studying in detail, because they have enough data on both sides to make claims about.

An earlier version of this step used four separate hand-built metrics instead of a single ensemble score. an **Isolation Score** measuring how concentrated a user's activity was inside a small set of subreddits, an **Ideology Dominance** score capturing how one-sided their political posting was, a **Conspiracy Ratio** describing the share of their posts that landed in conspiracy subreddits, and an aggregate **User Value Score** that combined the three into a bucket label of *Highly Isolated*, *Cross-Ideology*, or *Conspiracy-Focused*. The categorical bucketing made early exploration easier but also hid a lot of variation inside each group, and the thresholds for moving between buckets were chosen by eye rather than driven by the data. The later redesign kept the spirit of combining multiple signals but replaced the hand-tuned buckets with the continuous ensemble score described below, so that downstream filters like "top k% of users" become well-defined instead of sitting on a somewhat arbitrary cut-off.

4.3.1 Scoring Functions

Three scores are computed for each user, each capturing a slightly different notion of "active in both domains". All three take the user's normalised political activity a_P, and normalised conspiracy activity a_C (both in $[0,1]$) as inputs.

- **Weighted score:** a simple weighted average of a_P and a_C, balancing breadth across both domains. Rewards users who are active overall, even if mostly on one side.
- **Geometric score:** the geometric mean $\sqrt{a_P \cdot a_C}$. This rewards users who are active in both rather than just one, because if either side is zero the whole score goes to zero.
- **Harmonic score:** the harmonic mean $\frac{2a_P a_C}{a_P + a_C}$. Similar idea to the geometric mean, but penalises imbalance even more aggressively. A user with a lot of political activity and a tiny bit of conspiracy activity still scores poorly here.

An **ensemble score** averages all three into a single ranking (Figure 4.2), which is the number used for the final filtering. Using three functions and averaging them is a form of ensembling, where each score has its own bias (the weighted score rewards breadth, the geometric and harmonic scores reward balance), and averaging smooths over the edge cases where any one of them might misrank a user.

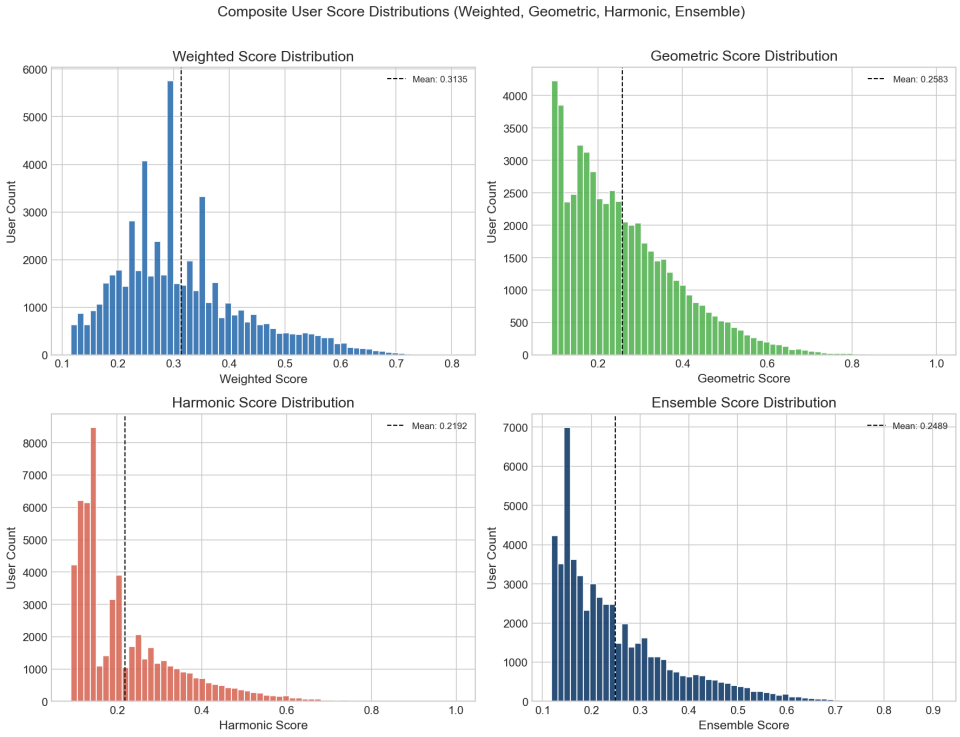


Figure 4.2: Distribution of composite user scores (weighted, geometric, harmonic, ensemble). The harmonic mean is the most conservative, with most users scoring below 0.2.

Users that pass a minimum activity threshold in both domains are marked as *high-value users* (HVU). The idea behind the filter is that downstream analyses, particularly the topic clustering and per-topic statistical tests, need users who have enough activity on both the political and conspiracy sides to produce a meaningful ideology vector *and*

enough conspiracy posts to actually land in a topic cluster. Without filtering, the majority of the 55,154 overlap users are one-post-one-comment accounts that add noise to the clustering and dilute the per-topic samples below useful sizes.

The threshold was tuned iteratively over the course of the project. Early versions were strict (ensemble ≥ 0.5, 25+ actions per side), which gave a clean but very small population that left many topics with too few users for a chi-squared test. Later versions were relaxed to the current values to reach a study population large enough for all eight research questions while still excluding low-signal accounts. The final settings, summarised in Table 4.3, require an ensemble score of at least 0.3 combined with at least 10 political and 10 conspiracy actions (posts or comments) per user. The ≥ 10 activity floor ensures that a user's ideology vector is computed from at least 10 data points rather than one or two, which is the minimum needed for the vector to reflect a pattern rather than a coincidence. The ≥ 0.3 ensemble floor ensures that both sides of the user's activity contribute meaningfully to the score, since a user with 500 political actions and 10 conspiracy actions will have a high weighted score but a low geometric and harmonic score, dragging the ensemble below 0.3. Using these cut-offs yields **7,798 HVU** (14.1% of all overlap users).

Table 4.3: Final HVU filtering thresholds. All three conditions must be satisfied.

Condition	**Minimum**	**Meaning**
Ensemble score	≥ 0.30	Combined weighted/geometric/harmonic score
Political activity (posts + comments)	≥ 10	Activity in any of the 161 political subreddits in the workin
Conspiracy activity (posts + comments)	≥ 10	Activity in any of the 4 conspiracy subreddits

Figure 4.3 compares HVU to non-HVU across key metrics, showing that HVU have substantially higher activity in both political and conspiracy domains.

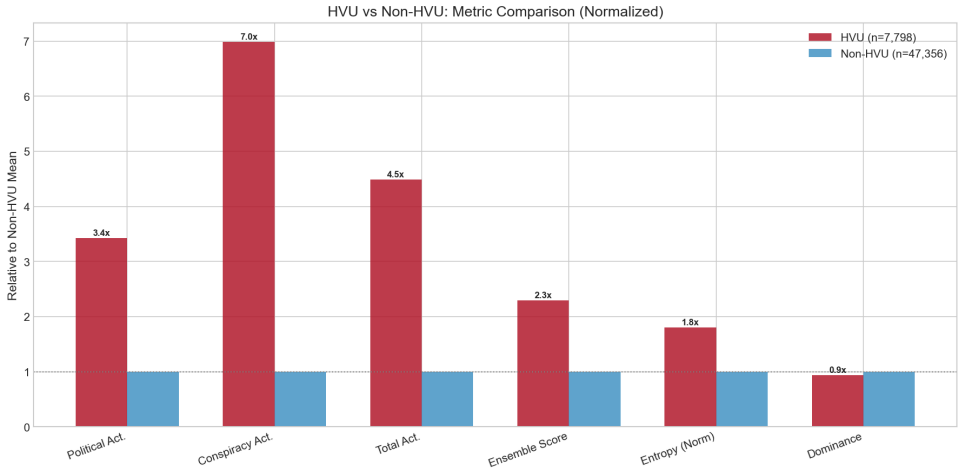


Figure 4.3: HVU vs. non-HVU: normalised metric comparison. HVU are roughly 3$\times$ more active politically and 7$\times$ more active in conspiracy subreddits than the non-HVU baseline.

4.4 Topic Clustering

This stage takes the conspiracy posts from the 7,798 HVU and groups them into topics automatically. The topic assignments are the basis for the per-topic composition, intensity, entropy, and odds-ratio tests of RQ1 through RQ4, and the topic centroids computed here are what the Chrome extension uses for real-time classification in RQ8.

The idea is to automatically group posts that are "about the same thing" without manually tagging them, so that instead of studying 13,194 individual posts we can talk about a handful of coherent topics like "vaccines", "elections", or "Russia and Ukraine". The clustering approach went through several iterations during the project, each time fixing a shortcoming of the previous one.

Before any of the formal clustering pipelines below, an informal exploration phase early in the project was made and just embedded every conspiracy post with a Sentence-Transformer model and plotted the embeddings as a 2D scatter, colouring each point by the user's (hard) ideology label. The goal at that stage was only to check whether the semantic space showed any visible ideological structure before committing to a clustering algorithm. It did show broad thematic regions, but without an actual clustering step on top there was no way to get topic assignments, keywords, or sizes out of it, which is why the next iteration moved to a real clustering pipeline.

4.4.1 Iteration 1: TF-IDF + NMF

The first pipeline used TF-IDF [41] vectorisation followed by NMF. TF-IDF, short for *term frequency-inverse document frequency*, turns each post into a vector of word counts weighted so that common words like "the" or "and" count for less and rare, distinctive words count for more. NMF, short for *Non-negative Matrix Factorisation*, then factorises the matrix of post vectors into a small number of "topic" axes, each defined by a set of characteristic words. It produced interpretable keywords, but the results were sensitive to the choice of k (the number of topics, which has to be set manually), and did not handle the noisy, short-text nature of Reddit posts very well. NMF produced 9 topics from 3,441 documents.

4.4.2 Iteration 2: Sentence Embeddings + HDBSCAN

The second approach swapped TF-IDF for sentence-transformer embeddings (all-mpnet-base-v2), the kind of semantic vector representation introduced in Section 2.3.3, which places posts about the same thing close together in vector space even when they use different words. UMAP [31] (Uniform Manifold Approximation and Projection) was then used to compress those high-dimensional embeddings down to a handful of dimensions while preserving neighbourhood structure, and HDBSCAN [30] was used for clustering.

HDBSCAN (Hierarchical Density-Based Spatial Clustering of Applications with Noise) is a density-based clustering algorithm that discovers the number of clusters on its own and can mark unclear points as noise instead of forcing them into a cluster. Semantic grouping was better than in iteration 1, but there was a high noise rate and too many tiny clusters.

4.4.3 Iteration 3: BERTopic with all-mpnet-base-v2

The third iteration adopted BERTopic [23], a framework that wraps the embedding + dimensionality reduction + density clustering idea from iteration 2 and adds automatic topic labelling on top. BERTopic was chosen over the manual pipeline of iteration 2 for three reasons that emerged from the iteration history above. First, it does not require a pre-specified number of topics (unlike NMF, which needs k set manually), so the clustering can discover however many coherent narratives actually exist in the data. Second, it builds on transformer-based sentence embeddings rather than bag-of-words representations, which handles the short, noisy, context-dependent text of Reddit posts far better than TF-IDF (the lesson from iteration 1). Third, it wraps UMAP, HDBSCAN, and c-TF-IDF keyword extraction into a single modular pipeline with well-tested defaults, which solved the fragmentation and high noise rate of the ad-hoc embedding + HDBSCAN combination in iteration 2. The supervisor recommended BERTopic specifically after reviewing the iteration 1 and 2 results.

The embedding model carried over from iteration 2 was all-mpnet-base-v2 (110M parameters), the standard sentence-transformer default. This iteration produced more coherent clusters than the manual pipeline in iteration 2, validating the BERTopic framework as the right approach. However, mpnet still produced relatively close-together cluster centroids on the short, noisy Reddit text in the corpus, with several topics that were thematically related not separating cleanly. This suggested the embedding step rather than the clustering one was the bottleneck, which motivated the next iteration's upgrade to a larger embedding model.

4.4.4 Iteration 4: BERTopic with EmbeddingGemma

The fourth iteration kept the BERTopic framework from iteration 3 but swapped the embedding model for Google's EmbeddingGemma-300M [22], a 300-million parameter sentence transformer that produces 768-dimensional vectors. It was chosen, again on the supervisor's recommendation, for two reasons. First, it performs well on short-to-medium English text, which is exactly the kind of input the pipeline processes (Reddit post titles and bodies, typically a few sentences long). Second, at 300M parameters it is small enough to run locally on a consumer-grade NVIDIA GPU (the development machine used

a laptop RTX 3060 with 6 GB VRAM), which meant the full embedding step could be done without relying on an external API or cloud compute. The larger and more recent embedding model produced markedly more separable clusters than mpnet on the conspiracy corpus, likely because its training data includes more short-form social media text. This iteration is the final flat clustering used throughout the thesis.

The full pipeline consists of the following steps.

1. **Corpus preprocessing:** Custom stopword removal (dropping function words like "the" and "and"), named-entity recognition (NER) for proper nouns so that "Donald Trump" is kept as one token instead of being split into two, synonym normalisation (e.g. "covid" and "coronavirus" mapped to the same form), and lemmatisation (reducing words to their base form so "running" and "ran" collapse to "run").

2. **Embedding:** each post is turned into a 768-dimensional vector using the EmbeddingGemma-300M model described above.

3. **Input filtering:** Only posts from high-value users are kept, comments are excluded, and any post shorter than 30 characters is dropped so that near-empty submissions cannot form spurious clusters.

4. **Dimensionality reduction:** UMAP with $n\_neighbors = 20$, $n\_components = 5$, $min\_dist = 0.0$, and cosine distance, compressing the 768-dimensional embeddings down to 5 dimensions so that density-based clustering can work effectively.

5. **Clustering:** HDBSCAN with $min\_cluster\_size = 75$ (any group smaller than 75 posts is not a topic), $min\_samples = 10$ (a post needs at least 10 dense neighbours to count as a core member of a cluster), and leaf-style cluster selection so that tight sub-groups are preferred over broader super-clusters.

6. **Topic representation:** Class-based TF-IDF (c-TF-IDF) for keyword extraction. Regular TF-IDF ranks words per document, while c-TF-IDF treats each cluster as a single "super-document" and ranks words per cluster, giving the distinctive keywords for each topic.

7. **Engagement weighting:** Posts with higher community engagement (score, comments, upvote ratio) get more influence during clustering, while zero-engagement posts are down-weighted. This stops ignored spam from dragging cluster centroids around.

8. **Topic validation:** Each candidate topic has to pass a set of post-clustering quality gates before it is kept. The main one is a user-diversity check that stops topics from being dominated by a handful of prolific posters, plus a minimum-engagement gate that filters out low-interest clusters. The full set of thresholds is listed in Table 4.4.

9. **LLM-generated labels:** The OpenAI API was used to generate human-readable topic names from keyword lists and representative posts.

10. **Centroid extraction:** Once the topics are finalised, the *centroid* of each topic, the mean of its engagement-weighted document embeddings, is computed by the code `compute_centroids.py` and saved to disk as `topic_centroids.npz`, together with a noise centroid built from the outlier bucket. These centroids are the artefact the Chrome extension's primary classifier loads at inference time (Section 4.6).

Table 4.4: Topic viability thresholds. A candidate cluster has to satisfy every gate to be kept in the final topic set, otherwise its posts are moved to the noise bucket.

Gate	**Threshold**	**Purpose**
HDBSCAN minimum cluster size	≥ 75 posts	Filters out tiny, noisy groups
HDBSCAN minimum samples	≥ 10 core neighbours	Keeps cluster cores genuinely dense
Minimum post text length	≥ 30 characters	Excludes near-empty posts
Unique user floor	≥ 10 users	Blocks single-user spam clusters
User diversity ratio	$\geq \max(10, \lceil 0.5\sqrt{n} \rceil)$ users	Scales the user floor with topic size
Weighted engagement	≥ 5.0 total	Drops clusters with no real discussion
Cross-topic keyword exclusivity	Enabled (whitelist: Trump, government, conspiracy)	Stops generic keywords from blurring topics

The final BERTopic run produced **14 topics** from **13,194 documents**, with a 45.3% noise rate (posts in topic 0, which is the catch-all bucket for posts that did not fit anywhere well). After removing noise, 7,212 posts across 1,354 users remain for analysis. Table 4.5 lists all discovered topics, Figure 4.4 shows their relative sizes, and Figure 4.5 shows the keyword word clouds for each.

Table 4.6 lists the top c-TF-IDF keywords for each topic, which informed the LLM-generated labels.

Table 4.5: The 14 conspiracy topics discovered by BERTopic, with a brief description of the dominant narrative in each.

#	Label	Core narrative	Posts	Users
1	Trump, Biden, and Election Manipulation	Electoral fraud, candidate-integrity claims, Trump shooting incident, media manipulation around 2024 cycle (Trump, Biden, Harris).	2,436	628
2	Vaccine Data and Covid-19 Concerns	Vaccine safety/cardiac risks, bioweapon framing, distrust of pharmaceutical companies and public-health policy.	1,461	442
3	Election Integrity and Political Dynamics	Narrower ballot-handling and debate-framing strand, distrust of electoral institutions rather than specific candidates.	445	230
4	Bill Gates, Lab-Made Viruses, and Elon Musk	Alleged ties between Gates/Musk and Covid-19 origin, EcoHealth and lab-leak accusations.	416	105
5	Censorship, Bots, and Online Discourse	Moderation bias, shadow-banning, and bot-driven manipulation on Reddit and other platforms.	413	269
6	Russia, Ukraine, and Global Military Tensions	Scepticism about Western narratives on the Russia-Ukraine war, NATO, 2014 coup, geopolitical motivations.	365	155
7	Israel, Iran, and Military Conflicts	Middle-East military operations, hidden geopolitical motives behind Israeli/Iranian actions.	343	116
8	Political Assassinations and Government Control	Assassination/attempt speculation (Kennedy, Trump), CIA involvement, deep-state cover-up.	333	165
9	Occult Practices and Elite Control	Secret societies, esoteric rituals, hidden-elite framing of mainstream politics.	300	151
10	Trump, Elections, and Spiritual Warfare	Religious and apocalyptic framing of the election, Trump as part of a divine or spiritual struggle.	190	111
11	Diddy, Drake, and Alleged Trafficking	Celebrity-industry trafficking allegations, Epstein-adjacent narratives.	149	82
12	Zodiac Killer and Unsolved Mysteries	Zodiac case, cipher discussion, possible military/government involvement in cold cases.	129	25
13	9/11 Truth and Government Involvement	Inside-job narratives, hijacker-passport claims, CIA/foreign-intelligence angles.	116	49
14	Project 2025 and Election Manipulation	Heritage Foundation's Project 2025 as a threat to electoral fairness and democratic norms.	116	72
	Total (excl. noise)		**7,212**	**1,354**

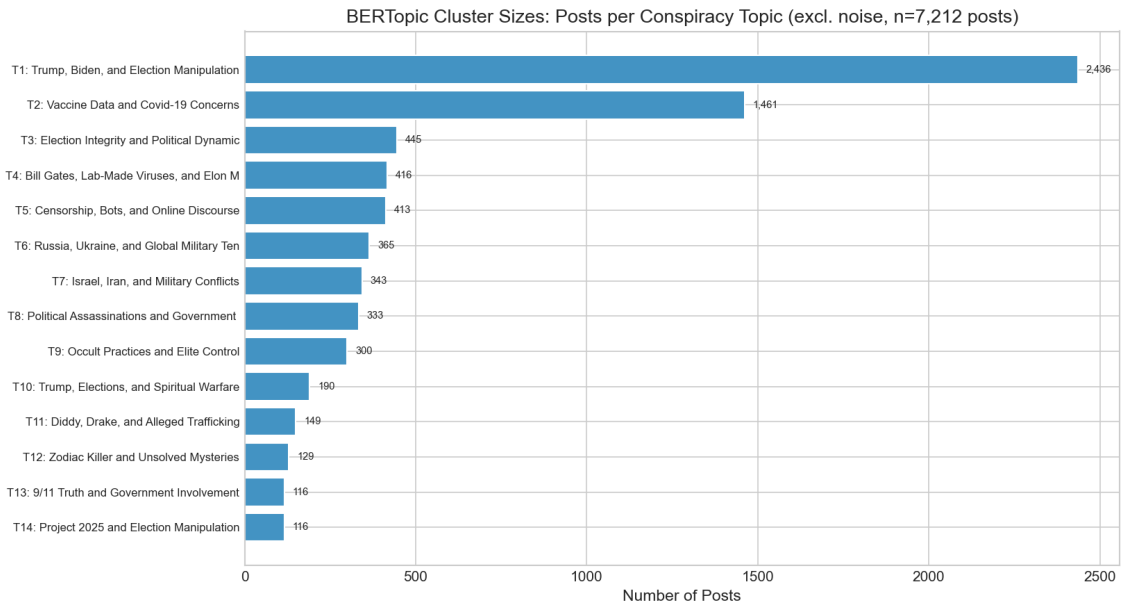


Figure 4.4: BERTopic cluster sizes (excluding the noise cluster). Topic 1 (Trump/Biden/Election) is the largest, the long tail of smaller topics reflects the diversity of conspiracy narratives.

Table 4.6: Top c-TF-IDF keywords for each conspiracy topic discovered by BERTopic.

#	Topic Label	Top Keywords
1	Trump, Biden, and Election Manipulation	Trump, Kamala, Biden, Democrat, vote, Joe Biden, FBI, president, shooter
2	Vaccine Data and Covid-19 Concerns	Covid-19, temperature, vaccine, study, data, cardiac, dose, placebo, NASA
3	Election Integrity and Political Dynamics	Biden, vote, election, Trump, ballot, count, debate, Democrat, poll
4	Bill Gates, Lab-Made Viruses, and Elon Musk	virus, EcoHealth, Covid-19, Bill Gates, cleavage site, lab, sequence
5	Censorship, Bots, and Online Discourse	bot, account, conspiracy, political, fake, propaganda, censorship, content
6	Russia, Ukraine, and Global Military Tensions	Russia, Ukraine, Russian, NATO, Putin, China, nuclear, Ukrainian, military
7	Israel, Iran, and Military Conflicts	Israel, attack, Iran, Hamas, Iraq, Netanyahu, Israeli, Bin Laden
8	Political Assassinations and Government Control	Sirhan, government, evidence, political, Kennedy, data, CIA, vaccine, Trump
9	Occult Practices and Elite Control	religion, Lucifer, conspiracy, occult, ancient, light, universe, ritual, Illuminati
10	Trump, Elections, and Spiritual Warfare	Trump, Jesus, light, Egypt, Antichrist, fight, universe, race, evil, election
11	Diddy, Drake, and Alleged Trafficking	Diddy, Drake, rapper, party, Epstein, sex, rap, trafficking
12	Zodiac Killer and Unsolved Mysteries	letter, Zodiac, murder, code, map, FBI, famous, military
13	9/11 Truth and Government Involvement	attack, passport, hijacker, Bin Laden, Atta, CIA, New York, Osama
14	Project 2025 and Election Manipulation	project, CERN, election, steal, program, agency, Trump

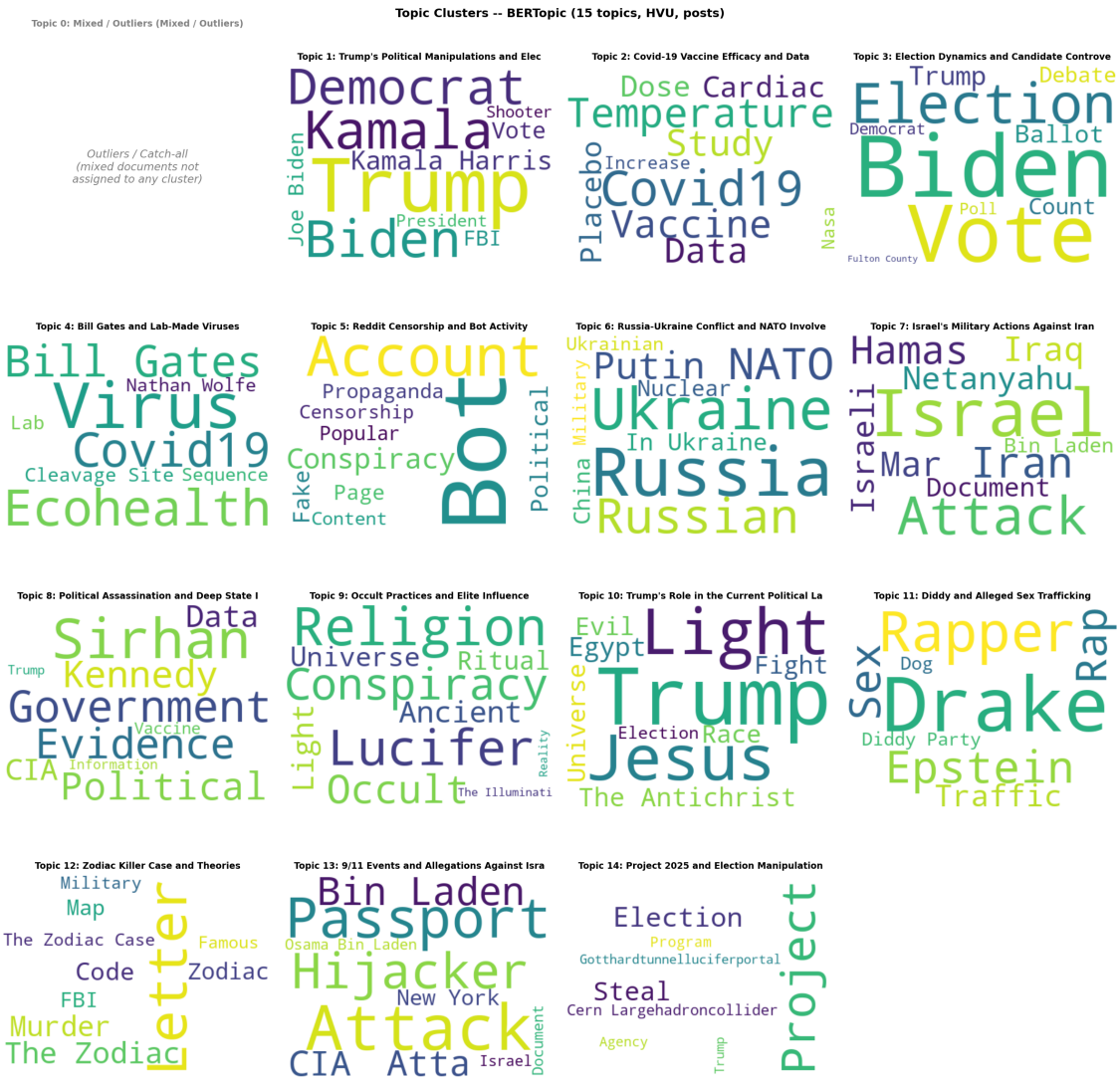


Figure 4.5: Word clouds for the discovered conspiracy topics (BERTopic + EmbeddingGemma). The top-left panel (T0) shows the noise/outlier bucket and is not used in the analysis.

4.4.5 Iteration 5: Sub-clustering within Dominant Topics

The BERTopic run of Iteration 4 is the final *flat* clustering used throughout the thesis, but a fifth iteration was added on top to split the largest topics into finer-grained subtopics. It is treated as its own stage rather than a post-hoc tweak because it has a separate module on disk (`subclustering/`), its own embeddings-to-assignments pipeline, and it produces the 20-category scheme used in the results chapter alongside the 14-topic scheme.

Three of the largest topics turned out to contain distinct sub-narratives. For example, Topic 6 (Russia/Ukraine) mixed discussions about Zelensky's role in Western strategy with content about the military-industrial complex and much older historical framings, which are related but genuinely different conversations. A sub-clustering step was added late in development to split these into finer-grained subtopics using the same EmbeddingGemma embeddings, HDBSCAN for discovering how many subtopics each parent contains, and K-Means for the final clean assignment. K-Means is a classic clustering algorithm that assigns every point to exactly one of k clusters, which is useful here because HDBSCAN would leave some posts as noise. The result was **9 subtopics** across 3 parent topics (Table 4.7), giving 20 topic-level categories in total (11 unchanged topics + 9 subtopics, replacing the 3 parents).

Table 4.7: Sub-clusters discovered within the three largest parent topics.

Parent	Subtopic Label	Posts
T2 Vaccine Data and Covid-19 Concerns (1,461 posts)		
T2.a	Government Distrust and Social Commentary	387
T2.b	Vaccine Efficacy and Health Data	1,074
T4 Bill Gates, Lab-Made Viruses, and Elon Musk (416 posts)		
T4.a	Gates, Musk, and Lab Virus Accusations	159
T4.b	EcoHealth and Wuhan Virus Connection	65
T4.c	Elon Musk's Military and Government Ties	192
T6 Russia, Ukraine, and Global Military Tensions (365 posts)		
T6.a	Russian Influence and American Perception	48
T6.b	Zelensky's Role and Western Response	131
T6.c	Proxy War and Military Industrial Complex	85
T6.d	Historical Context and Elite Manipulation	101

Most users engage with only one or two topics (Figure 4.6), but a subset of users participates across many. Figure 4.7 shows the topic co-occurrence matrix, which counts the number of users shared between each pair of topics: the darker a cell, the more users have posted in both of those topics. The largest overlaps are between Topics 1 and 2 (Trump/Biden + Vaccines), which are also the two biggest clusters.

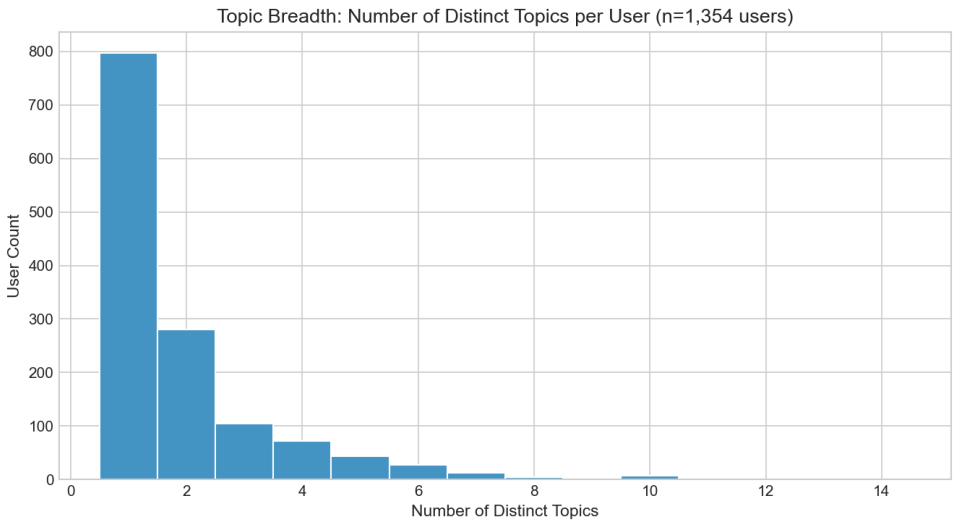


Figure 4.6: Distribution of topic breadth per user. 58.9% of users engage with only one topic.

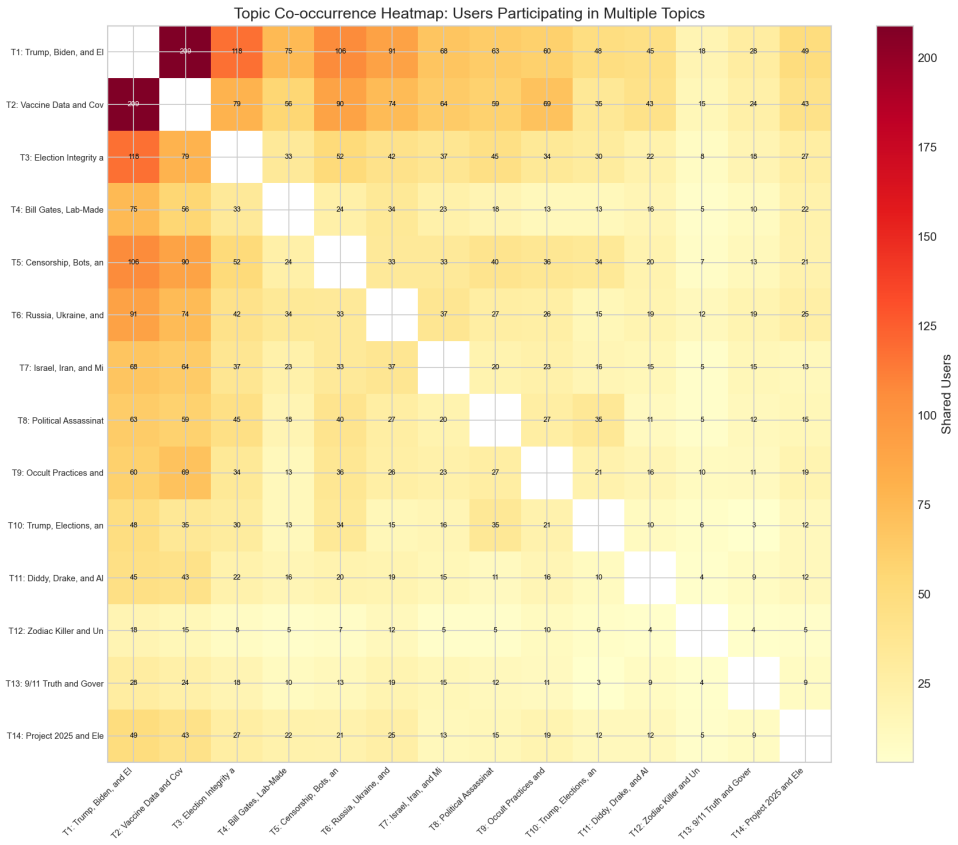


Figure 4.7: Topic co-occurrence heatmap: number of users participating in each pair of topics.

4.5 Network Construction and Polarisation Analysis

This stage takes the HVU set, their ideology vectors, and their topic assignments and builds four network representations that encode different kinds of relationships between users, subreddits, posts, and topics. The user-user co-activity network is the input for the null model in RQ6, the user-subreddit bipartite network feeds the subreddit ecosystem comparison of RQ7, and the polarisation metrics computed here inform the bridging-score analysis of RQ5.

Four types of networks were built to look at the structural relationships between users, subreddits, posts, and topics. A *network*, in this context, is just a graph where nodes represent entities (users, subreddits, posts, topics) and edges represent some relationship between them (e.g. "user posted in this subreddit").

The original plan was to visualise all of these directly in Gephi [6], an open-source network analysis and visualisation tool. The first prototype was a single user-subreddit bipartite network with no weighting or ideology colouring, used mostly to confirm that the overlap set produced a visually meaningful structure. A second iteration added ideology colouring, edge weighting, a user-user co-activity view, and an early set of five polarisation metrics. At the final corpus size, though, the graphs turned out to be too dense to read as pictures. The user-subreddit bipartite network alone has 7,959 nodes and nearly 40,000 edges, and standard Gephi layouts produced an unreadable hairball rather than anything informative about ideology. The Gephi-side visualisation was therefore dropped, and the node and edge CSVs originally generated for it were repurposed as the input to the quantitative polarisation analysis in Section 4.5.2 and the heatmap and workbook outputs in Section 4.5.3. The four networks described below are the final redesign, which standardised the node/edge schemas, added the user-topic and user-post views that depend on the BERTopic output, and replaced the ad-hoc early metrics with the eight-metric suite in Section 4.5.2.

4.5.1 Network Types

The four networks, summarised in Table 4.8, each capture a different slice of the data. A *bipartite* graph is one where nodes split into two disjoint sets and edges only go between the two sets (e.g. users and subreddits, never user-user).

Table 4.8: The four constructed networks and what they capture.

Label	Type	Edge meaning	Scale
User-Post (4.1)	Bipartite	User posted this post, post → topic	7,212 posts
User-User (4.2)	Undirected	Both users posted in same conspiracy thread, weighted	4,368 users, 193,876 edges
User-Subreddit (4.3)	Bipartite	User posted in subreddit, weighted by activity volume	7,798 + 161 nodes, 39,984 edges
User-Topic (4.4)	Bipartite	User posted in topic, weighted by post count	Derived from 4.1

All networks are exported as CSV node and edge lists, the format originally chosen for Gephi but now consumed directly by the polarisation and heatmap pipelines. Table 4.9 summarises the scale and density of the two main networks used for polarisation analysis. *Density* is the fraction of possible edges that actually exist, so a density of 0.032 for the user-subreddit network means only 3.2% of all possible user-subreddit pairings are realised, which is typical for a social graph.

Table 4.9: Summary of the two primary networks.

Network	Nodes	Edges	Density
User-Subreddit bipartite	7,798 + 161	39,984	0.032
User-User co-activity	4,368	193,876	0.020

Figure 4.8 visualises the scale of each network, and Figure 4.9 shows the user degree distribution in the user-subreddit bipartite network, where *degree* means the number of political subreddits a user posts in. Most users post in only a handful of political subreddits, but a long tail is active in over 40.

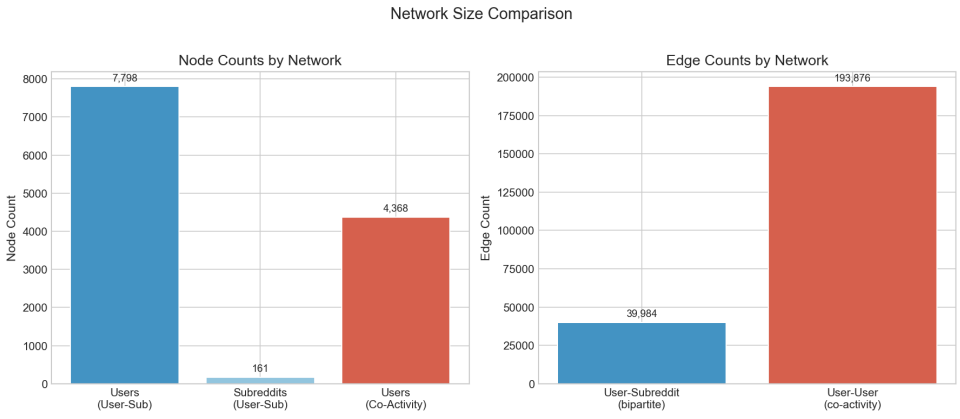


Figure 4.8: Size comparison of the constructed networks (node and edge counts).

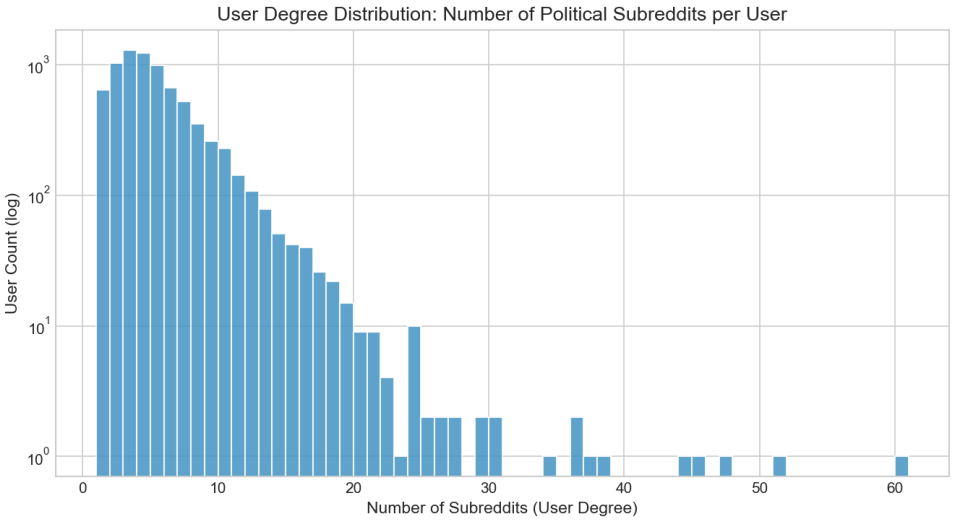


Figure 4.9: User degree distribution in the user-subreddit bipartite network (log scale). Most users post in 3-5 political subreddits.

4.5.2 Polarisation Metrics

Eight metrics were computed to measure how ideologically polarised the conspiracy ecosystem is (Table 4.10). Each metric looks at the question from a slightly different angle so that one anomaly cannot dominate the conclusion. Seven of the metrics are normalised to $[0,1]$, and one (the bridge score) is shown as an unbounded ratio alongside its normalised form. Each metric is defined below with its formula, range, and concrete reading for the HVU population.

Homophily Index: Measures the tendency of users to connect with others who share their ideology in the user-user co-activity network. Let o be the observed fraction of same-ideology edges and e the fraction expected under random pairing given the ideology distribution. The raw index is $H = (o-e)/(1-e)$, which lies in $[-1,1]$, with negative values indicating *heterophily* (cross-ideology mixing above chance) and positive values indicating homophily. It is then rescaled to $[0,1]$ as $H' = (H+1)/2$. A value of 0.5 means "exactly as expected by chance", above 0.5 means same-ideology clustering, and below 0.5 means cross-ideology mixing. Here $H' = 0.298$, clearly below the chance line, which already says the conspiracy co-activity network is more cross-ideological than a random graph with the same ideology mix would be.

Cross-Engagement Rate: Measures how evenly a user's engagement is spread across conspiracy topics, using the Shannon entropy of their topic distribution. For each user, topic engagement counts are normalised to a probability vector p over the K topics, and entropy $E = -\sum_i p_i \log_2 p_i$ is computed and divided by $\log_2 K$ to give a value in $[0,1]$.

The ecosystem-level score is the mean over all HVU. A value near 1 means users spread their engagement almost uniformly across topics, a value near 0 means each user concentrates on a single topic. Here the score is 0.935, meaning HVU touch a wide variety of conspiracy topics rather than specialising.

Subreddit Cross-Participation Rate: Measures the fraction of users who participate in subreddits from more than one ideological group (left, centre, right). Formally, SCR $= |\{u : \text{ideology groups of } u\text{'s subreddits} \geq 2\}|/|U|$, which lies in $[0,1]$. A value of 1.0 does *not* mean "everyone on Reddit mixes ideologies". It is measured over the HVU population, which by construction (ensemble score ≥ 0.3 with non-trivial political activity, Table 4.3) already requires a breadth of political participation. In practice, essentially every retained user ends up touching at least two of the three ideology groups, so SCR $= 1.000$ is a *property of the filter*. It is included because it documents how ideologically broad the HVU set is, and anything below 1 would flag that a subset of the HVU is actually 100% ideologically homogeneous.

Narrative Clustering Coefficient: Measures how tightly posts on the same topic cluster together in the user-post-topic network, using the average local clustering coefficient of topic nodes. It lies in $[0,1]$, where higher values mean tighter per-topic communities and lower values mean topics share many users with each other. Here 0.681 indicates that topics have moderately distinct user bases, but with a noticeable amount of user sharing across topics.

Top-3 Concentration: The share of total topic engagement captured by the three largest topics combined. Formally, T3 $= \sum_{i \in \text{top } 3} e_i / \sum_i e_i$, in $[0,1]$. A value near 1 means the ecosystem is dominated by a handful of narratives, a value near $3/K$ would mean uniform spread. The metric is computed on the user-topic network built from the main BERTopic output (before the sub-clustering step of Section 4.4.5), so the denominator is the set of main topics that actually received user-topic edges rather than the post-sub-clustering 20 categories. Here 0.780 means three main topics account for 78% of the engagement, pointing to a highly skewed, long-tailed topic distribution that is exactly what motivated breaking the largest topics into subtopics in the first place.

Ideological Bridge Score: Measures how much more connected HVU are in the user-subreddit bipartite network than the average Reddit user in the full dataset. It is defined as the ratio BS $= \bar{d}_{\text{HVU}}/\bar{d}_{\text{all}}$, where $\bar{d}_{\text{HVU}}$ is the mean political-subreddit degree of HVU and $\bar{d}_{\text{all}}$ is the mean over all overlapping users. Because it is a ratio, it is *unbounded above* and is reported with the multiplier unit "$\times$". A value of 1.0 would mean HVU are no more connected than an average user, the observed $1.791\times$ means HVU post across 79% more

politically classified subreddits on average than a typical user. For the aggregate overall-polarisation score, this ratio is mapped into $[0,1]$ via the normalised *specialist tendency* $\mathrm{BS}' = \max(0, \min(1, 1-(\mathrm{BS}-1)/2))$, so the aggregate does not inherit the unbounded scale.

Ideology Imbalance: Measures how skewed the HVU population is toward a single ideology group. Let f_L, f_C, f_R be the fractions of HVU classified as left, centre, and right. Imbalance is the total variation distance from the uniform distribution over $\{L, C, R\}$, rescaled so that 0 means perfectly balanced and 1 means all users in one group. Here 0.257 indicates a mild skew but a population that is still clearly tri-modal.

Overall Polarisation: The mean of the other seven metrics after each is mapped into $[0,1]$ (using BS' rather than the raw bridge ratio). This aggregate should be read with care, as each underlying metric uses a different convention for what "high polarisation" looks like, and for some of them a value near 0.5 is ambiguous rather than neutral. For instance, a homophily index of 0.5 means "exactly chance", while a cross-engagement rate of 0.5 already indicates narrow engagement per user. Two ecosystems with very different profiles can therefore land on the same mean, which is why the per-metric table above is the primary artefact and the aggregate is used only as a single-number summary. Here 0.400 is at the low-moderate end of the scale, consistent with the per-metric picture of an ideologically mixed space.

Table 4.10: Polarisation metric values for the conspiracy ecosystem.

Metric	Value	Reading
Homophily Index	0.298	Below chance, cross-ideology mixing
Cross-Engagement Rate	0.935	Users span many topics each
Subreddit Cross-Rate	1.000	Property of the HVU filter
Narrative Clustering	0.681	Moderately distinct topic communities
Top-3 Concentration	0.780	Three topics carry 78% of engagement
Ideological Bridge Score	1.791×	HVU are 1.79× as connected as average
Ideology Imbalance	0.257	Mild skew, still tri-modal
Overall Polarisation	0.400	Low-moderate

The low homophily index (0.298) and high cross-engagement rate (0.935) already suggest that the conspiracy ecosystem is not strongly polarised along ideological lines. Figure 4.11 shows the breakdown of connection types in the co-activity network, with 45.5% of connections between right-leaning users, 43.5% cross-ideology (left-right), and 11.0% between left-leaning users. Two things are worth pointing out here. First, the high cross-ideology rate (43.5%) hints at the main finding that conspiracy communities are not strongly segregated by ideology. Second, the same-ideology connections are strikingly asymmetric, right-right edges outnumber left-left edges by roughly 4× (45.5% vs 11.0%),

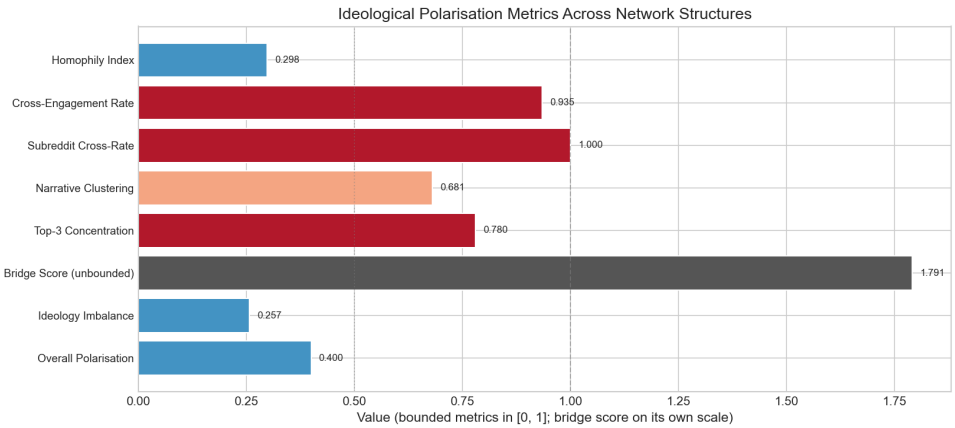


Figure 4.10: Polarisation metric values across the conspiracy ecosystem. The bridge score is on an unbounded scale and is shown here for completeness, the remaining seven metrics are in $[0, 1]$.

which is more lopsided than the underlying HVU ideology split would predict on its own. This says that right-leaning HVU not only make up a larger share of the population but also co-activate with each other far more densely than left-leaning HVU do, a within-group clustering effect that the single homophily number averages away.

These descriptive proportions are a useful first look, but they do not on their own establish whether the observed pattern is statistically meaningful or simply a consequence of the network's ideological composition. A null model that permutes ideology labels across users while keeping the graph structure fixed provides the formal test, and is reported as RQ6 in Section 5.6.

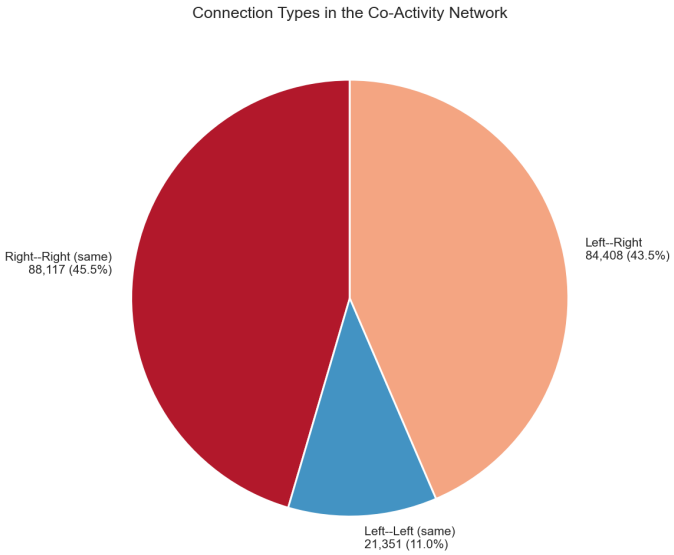


Figure 4.11: Connection types in the co-activity network. Cross-ideology connections (43.5%) are nearly as common as same-ideology ones.

4.5.3 Heatmap and Workbook Outputs

Because the Gephi visualisations at full scale were unreadable, the same node and edge data was reshaped into a set of 2D heatmaps that let the reader scan patterns by row and column instead of by graph layout. Five heatmap views were produced from the user-subreddit bipartite edges.

- **User-subreddit activity heatmap:** rows are users, columns are political subreddits, cell colour encodes activity volume on a log scale. Users are sorted left → centre → right, subreddits are grouped by ideology, so the three-block diagonal pattern is what one would expect if ideology strongly predicted subreddit choice.
- **Ideology-difference heatmap (non-centre subs and users):** both axes drop centre, and the colour encodes whether the subreddit leans left (blue) or right (red), sorted by the absolute magnitude of the lean.
- **Ideology-difference heatmap (all subs, non-centre users):** keeps all subreddits but drops centre users, with centre subreddits rendered green, giving a fuller subreddit landscape.
- **Signed R-L heatmaps (two variants):** the same two views sorted by the *signed* right-minus-left score rather than absolute lean, so right-leaning subreddits appear first and left-leaning ones last.

Each view was first prototyped as a standalone interactive Plotly HTML page, and then all five plus a set of reference sheets were combined into a single Excel workbook, `ideology_heatmaps.xlsx`, which became the final artefact used during analysis. The reference sheets include subreddit splits with per-label counts, per-user ideology and scoring information, topic-by-conspiracy-subreddit cross-tabs, a cross-posting overlap matrix, and the full eight-category ideology breakdown per subreddit. The workbook supports sorting, filtering, and cross-sheet lookups that the standalone Plotly pages do not, and on top of that it provides both a more user-friendly interface and easier navigation. The HTML versions are kept only as legacy artefacts, and the workbook is used primarily as a scan-and-sort tool when investigating specific users or subreddits, rather than for embedding static snapshots in the thesis.

4.6 Chrome Extension: Conspiracy Identifier

This stage takes the topic centroids from Stage 4, the ideology vectors from Stage 2, the fine-tuned RoBERTa classifier, and the user scoring data from Stage 3, and wraps them

into a real-time classification tool. The section covers both the classification methodology and the software that delivers it, and the two play different roles in the thesis. The **methodological contributions**, marked with [METHODOLOGY] below, define *how* posts are classified and are the substance evaluated in RQ8. The **implementation details**, marked with [IMPLEMENTATION] below, describe the engineering that makes the methodology accessible in a browser but are not themselves research contributions.

As a practical application, a Chrome extension (*Conspiracy Identifier*) was built for real-time use of the topic model and ideology profiles within a user's browser. The extension classifies live Reddit posts against the topic scheme from Section 4.4 and, when the post author is a user the pipeline already profiled, attaches the ideology and scoring information from the ideology stage as well. The goal is transparency rather than moderation, so a reader browsing Reddit sees whether its content looks like an existing conspiracy narrative in the thesis corpus, and whether the account behind the post has a measurable political footprint on the platform.

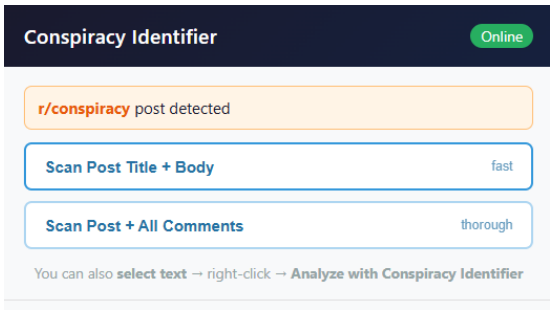
4.6.1 Architecture [IMPLEMENTATION]

The system has two halves, a Manifest V3 Chrome extension on the browser side and a local FastAPI inference server on the Python side, connected over localhost HTTP. Keeping the model server local was a deliberate choice, as no post content ever leaves the machine, the EmbeddingGemma-300M model runs on the user's own GPU when CUDA is available, and the ideology lookup table for known users is never exposed outside the machine running the backend.

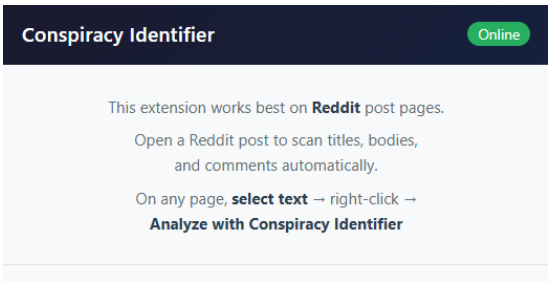
The original target for the extension was actually a fully on-device deployment, with no backend at all, so that a user could install it and browse Reddit without having to run a local Python server in the background. The obstacle was the embedding model: running EmbeddingGemma-300M directly inside a Chrome extension is not feasible with the stock Manifest V3 runtime, and the only way to make the centroid pipeline work in-browser would have been to downgrade the representation back to something like TF-IDF, which would have meant giving up exactly the semantic signal that motivated moving past NMF in Section 4.4. An alternative suggested during supervisor feedback was to port the model to the `transformers.js` JavaScript library, which can load a subset of Hugging Face models directly in the browser via WebGPU. That route was left as future work (Section 7.5) rather than pursued in this thesis, because it would require re-validating the embeddings against the Python pipeline and the centroid pipeline to confirm that the in-browser embeddings line up numerically with the ones used to build the centroids, and the localhost FastAPI backend was sufficient for demonstrating the topic model end-to-end within the thesis scope.

Browser extension: The extension registers three surfaces in its `manifest.json`.

- A **popup** (`popup.html`) with a textarea and an "Analyze" button for classifying arbitrary text outside of Reddit, plus a small settings panel for toggling the inline overlays and the optional image/link enrichment.
- A **content script** (`content.js`, roughly 1,000 lines) injected into every `*.reddit.com` page, which walks the DOM to locate post containers across the three Reddit layouts (old-Reddit, redesigned-Reddit, and the newer shreddit Web Components), extracts the title, self-text, visible top-level comments, author handle, and any attached image URLs, and renders the classification panel inline next to each post via a small CSS overlay (`content.css`).
- A **background service worker** (`background.js`) that owns the HTTP calls to the backend, registers a context-menu entry ("Analyze with Conspiracy Identifier") on any text selection, caches recent classifications keyed by the post's permalink so that scrolling back through a feed does not re-embed the same post, and serialises requests so a scroll-heavy feed does not flood the backend with parallel inference calls.



(a) Reddit post detected: the popup offers a fast "title + body" scan and a thorough "post + all comments" scan.



(b) Non-Reddit page: popup directs the reader to open a Reddit post or use the right-click "Analyze with Conspiracy Identifier" entry on any selected text.

Figure 4.12: Popup entry states. The green indicator in the top-right confirms the local backend is reachable.

Inference backend: The backend is a FastAPI server (`app.py`) that binds to `127.0.0.1:8000` and holds all heavy state in memory after a one-off warm-up on startup. On boot it loads the EmbeddingGemma-300M sentence transformer (onto CUDA if available, otherwise CPU), the saved topic centroids from `topic_centroids.npz`, the per-topic ideology profile table from `topic_profiles.json`, the fine-tuned RoBERTa classifier from

`bert_classifier/`, and the per-user ideology/scoring lookup table computed in Section 4.2. Table 4.11 lists the HTTP API that the extension consumes.

Table 4.11: HTTP endpoints exposed by the local inference backend.

Method	Path	Purpose
GET	`/health`	Liveness check and confirmation that all models loaded.
GET	`/topics`	Full list of topic labels, keywords, and ideology profiles, used by the popup to render the topic browser.
GET	`/user/{username}`	Returns the pipeline's ideology vector, six-tier label, activity counts, and scoring percentiles for a known HVU, or 404 if the user is not in the dataset.
POST	`/analyze`	Classifies a text payload using the centroid pipeline and the RoBERTa head, returning top-k topic matches with confidences and the associated ideology profile of the winning topic.
POST	`/analyze-images`	Optional. Takes a list of image URLs and returns text descriptions produced by a GPT-4o vision call, used to enrich image-only posts before re-classification.
POST	`/analyze-links`	Optional. Fetches and extracts the readable text of a linked article so link-only posts have something for the classifier to embed.

4.6.2 Classification Pipeline [METHODOLOGY]

When the content script picks up a post, it sends the extracted fields to `/analyze` as a structured payload (title, body, top-level comments, and when enabled, the image and link context produced by the relevant endpoints). The backend applies the same text normalisation as the clustering pipeline so inference-time text is distributed the same way training text was, and then runs two classifiers in parallel.

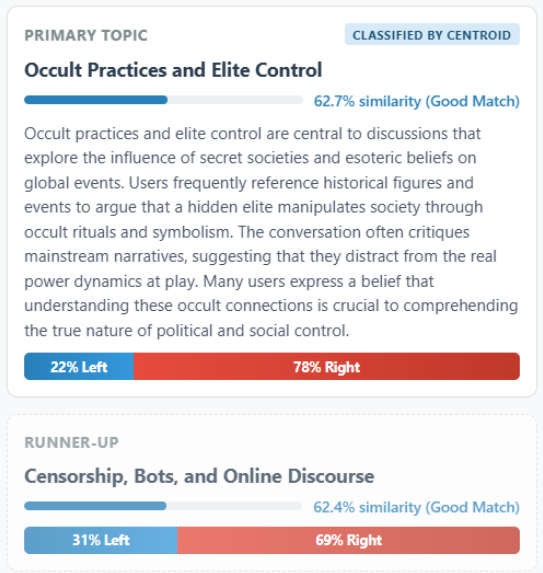
- **Primary: component-weighted topic centroids + cosine similarity.** During the BERTopic run, each topic's *centroid*, the mean of its engagement-weighted document embeddings, is saved to disk via `compute_centroids.py`. At inference time the post is split into up to four components (*title*, *body*, optional *image context* from `/analyze-images`, and optional *link context* from `/analyze-links`). Each component is embedded separately and compared against every centroid using cosine similarity $\cos(\theta) = (\mathbf{x} \cdot \mathbf{c})/(\|\mathbf{x}\| \, \|\mathbf{c}\|)$. The per-component similarities are fused using a weighted combination that reflects how diagnostic each component is, with base weights of $40/30/12/18$ for title / body / image / link when all four are present, renormalised to sum to one whenever a component is missing. Without image or

link context the scheme falls back to a 55/45 title / body split. The top-k most similar topics are returned together with their scores, and the top assignment is only surfaced if its fused similarity exceeds a specific confidence floor, otherwise the post is labelled *unclassified* rather than being forced into the nearest topic. Alongside the real topics, the backend also keeps a *noise centroid*, which is the mean embedding of the posts BERTopic put in the outlier bucket (topic 0 in Section 4.4). Two quantities derived from it are used later as decision thresholds. The *noise margin* is the best real-topic similarity minus the noise similarity, where a positive value means the post is closer to a real cluster than to the noise cloud. The *noise baseline* is the similarity to the noise centroid treated as a rough lower bound on how generic a post looks. Together they let the classifier distinguish a confident match from a post that only happens to be marginally closer to one cluster than to everything else. This method is the primary classifier because it was validated directly against the clustering pipeline and because it generalises to future topic re-runs without any retraining, a new centroid file is the only thing the backend needs.

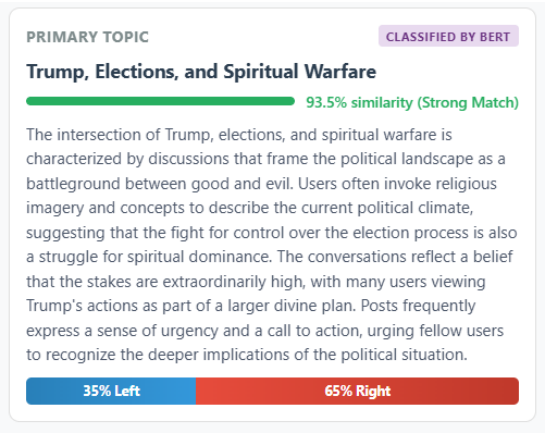
- **Secondary: fine-tuned RoBERTa classifier.** In parallel, a RoBERTa [28] backbone was fine-tuned on the labelled HVU corpus. Two additions were made to stabilise training on the small, imbalanced label set. First, *data augmentation* via paraphrasing multiplies the number of training examples per topic and reduces the class imbalance between the largest and smallest topics. Second, *k-fold cross-validation* is used during training so every labelled example is used for validation exactly once across folds, which gives a more stable estimate of per-topic accuracy than a single held-out split and exposes topics that the model systematically confuses. At inference time the same concatenated text (with image and link context appended when available, so RoBERTa sees what the centroids see) is passed through the fine-tuned head, and the softmax top-1 plus its confidence is returned alongside the centroid result. Having both methods in parallel is useful when the centroid similarity is close to the confidence floor, since agreement between the two classifiers confirms a marginal match and disagreements trigger the RoBERTa rescue path rather than a manual inspection.

The response payload returned to the extension includes the winning topic's label, its keyword list, the confidence score, and the ideology profile attached to that topic (left / right percentages, dominant lean, and entropy over the six-tier scheme), so the classification panel rendered in the page can show both *what* the post is about and *who typically posts about it* in the HVU corpus. The extension shows these values as raw composition only, without a categorical label attached, for the same reason RQ3 drops the echo-chamber classification (Section 5.3): the underlying composition deviations do

not survive multiple-comparisons correction, so collapsing them into a category would overstate what the data supports.



(a) Centroid-selected primary topic with a runner-up shown underneath. Both cards display their raw left/right composition bars, with no categorical label.

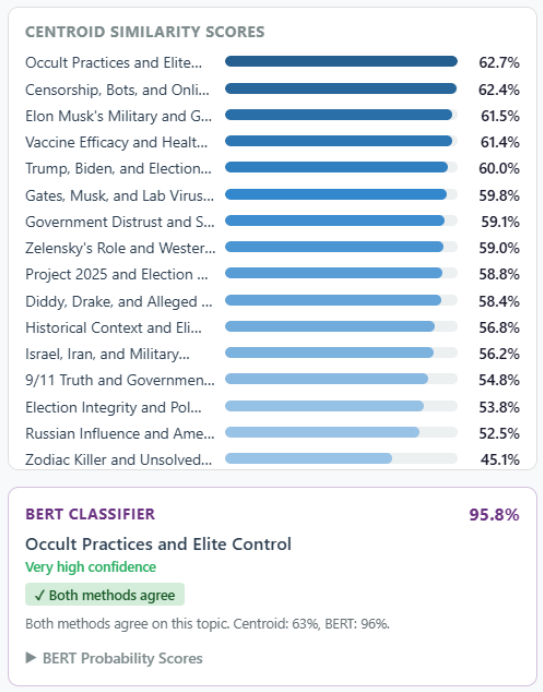


(b) BERT-selected primary topic (the RoBERTa head overrode the centroid because of its much higher confidence and clear margin).

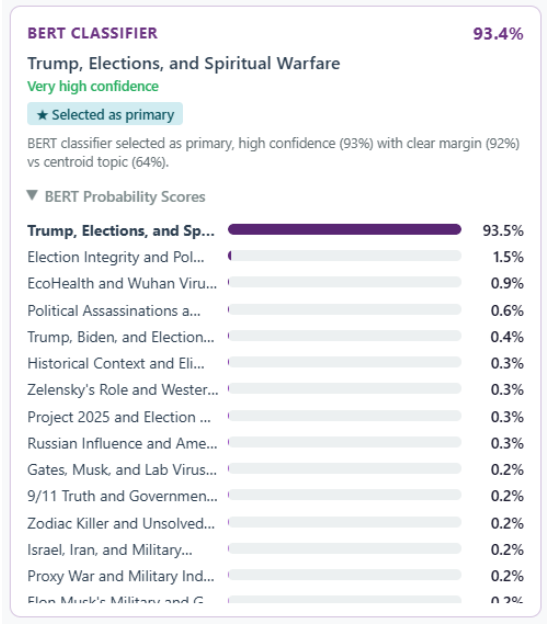
Figure 4.13: Primary topic cards rendered by the extension. The badge in the header indicates which classifier produced the assignment, and the coloured bar shows the fused similarity relative to the strong / good / fair match thresholds.

The two panels in Figure 4.14 report scores on different scales and should not be compared value-for-value. The centroid panel (a) shows cosine similarities between the post's embedding and each topic centroid, each score is an independent geometric match in embedding space, bounded in $[-1, 1]$, and the values across topics do *not* sum to 100%. The BERT panel (b) shows softmax probabilities from the fine-tuned RoBERTa head, these are normalised class probabilities over the topic set and therefore *do* sum to 100% across all topics. A centroid score of 60% and a BERT probability of 60% thus carry different meanings, where the former says the post lies fairly close to that centroid in absolute terms while the latter says the classifier assigns 60% of its total confidence to that topic relative to all others.

Because the clustering pipeline in Section 4.4 splits several of the dominant BERTopic topics into finer sub-clusters, the extension also surfaces that structure on a match. This means that when the post is assigned to a sub-topic, the primary card shows both the sub-topic label and the name of its parent topic, so the reader can see the coarser category the narrative belongs to as well as the specific variant the model latched onto.



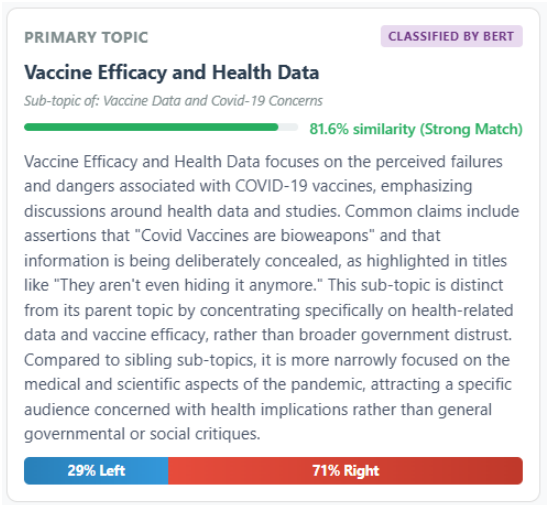
(a) Ranked centroid similarities for the candidate topics, plus the BERT agreement badge for cross-validation.



(b) BERT per-topic probabilities, with the "Selected as primary" badge indicating that the RoBERTa head drove the final decision.

Figure 4.14: Transparency panels showing both classifiers' per-topic scores side by side. This lets a reader verify whether the chosen topic is a confident winner or simply the least-bad option among close competitors.

(a) Centroid-selected sub-topic (*Zelensky's Role and Western Response*) under its parent *Russia, Ukraine, and Global Military Tensions*.

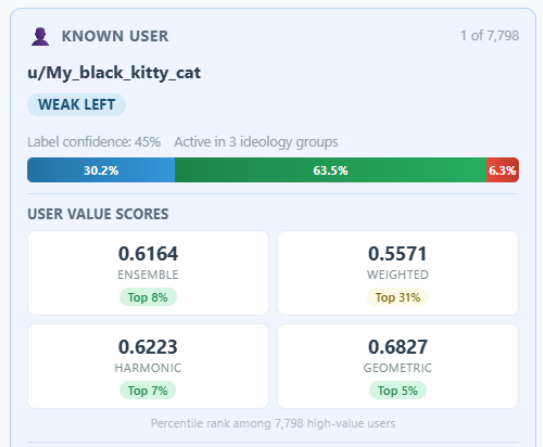


(b) BERT-selected sub-topic (*Vaccine Efficacy and Health Data*) under its parent *Vaccine Data and Covid-19 Concerns*.

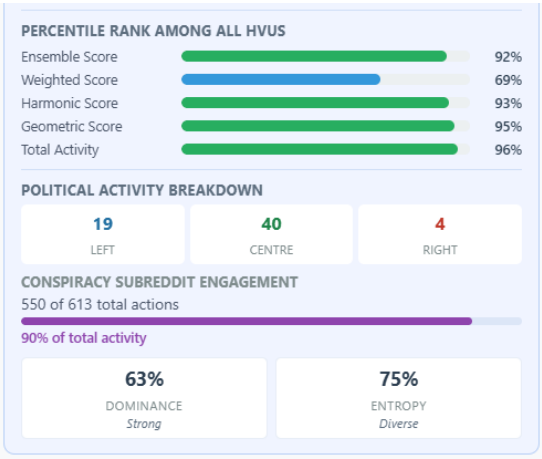
Figure 4.15: Sub-topic assignments. The "Sub-topic of:" badge connects fine-grained narrative clusters back to the coarse topic scheme.

4.6.3 User Identification [IMPLEMENTATION]

If the post author is a known HVU, the extension makes a second call, this time to `/user/{username}`, and augments the classification panel with the user's six-tier ideology label, their political and conspiracy activity counts, their ensemble bridging score, and their percentile rank against the full HVU population. This connects the per-post view produced by the topic classifier to the per-user profiles from Section 4.2, so a reader can see both the content classification and the behavioural footprint of the account behind the post in a single panel. For users who are not in the dataset, the endpoint returns a 404 and the extension cleanly omits the user-profile block rather than guessing an ideology from the post alone, since inferring ideology from a single post would contradict the behavioural-inference principle the rest of the pipeline is built on.



(a) Header block: six-tier ideology label (consolidated from the label/lean/detail fields), label confidence, the (P_L, P_C, P_R) probability bar, and the four value-scores (Ensemble, Weighted, Harmonic, Geometric) with percentile badges.



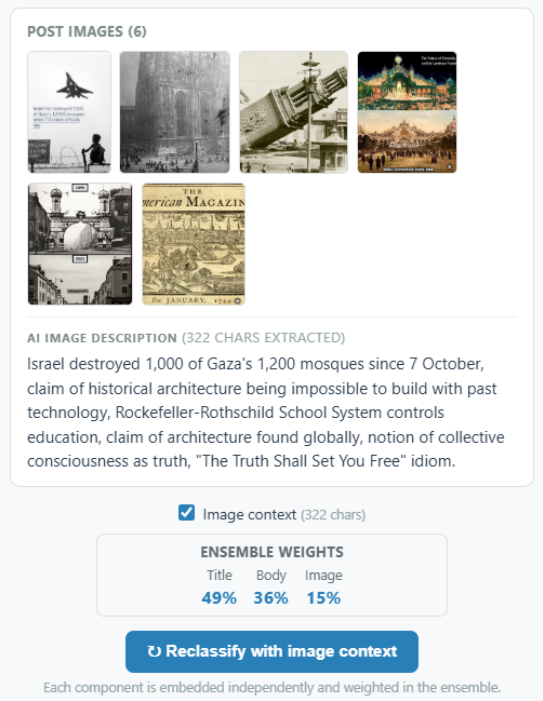
(b) Lower block: percentile ranks across all five metrics, ideology-wise activity breakdown, conspiracy engagement share of total activity, and dominance / entropy descriptors.

Figure 4.16: Known-user panel for an HVU in the pipeline's lookup table. The panel stitches the per-post classification together with the per-user ideology and scoring information computed in Section 4.2.

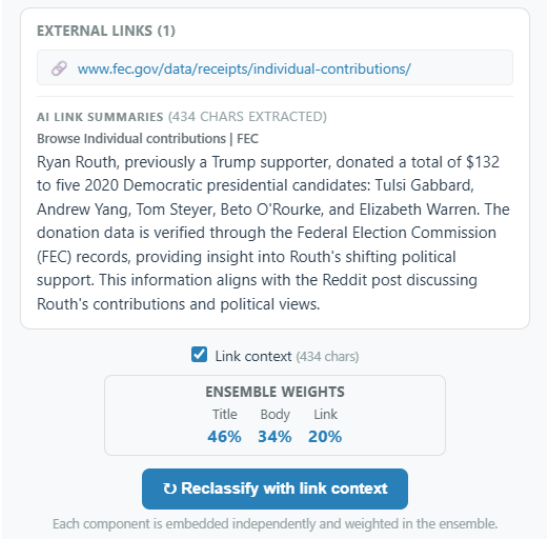
4.6.4 Multimodal Enrichment [IMPLEMENTATION]

Link-only and image-only posts carry little embeddable text on their own, so a strict text pipeline would either refuse to classify them or fall back on the title alone. To keep those posts within the topic scheme, the extension optionally resolves external context through two side endpoints. `/analyze-images` calls a GPT-4o vision model to produce a short textual description of each attached image. `/analyze-links` fetches the linked page,

strips boilerplate, and returns a compact summary of the underlying article, with handlers for PDFs, X/Twitter, and Google redirectors, plus HTTP-code-aware error messages for blocked publishers. Both calls are opt-in, and once the context is fetched the user can toggle each component on or off and re-submit, with the ensemble weights shown in the reclassification card updating accordingly. Without enrichment the classifier falls back to the original 55/45 title / body split. When all four components are present the base weights 40/30/12/18 (title / body / image / link) are used as declared. When only one extra component is enabled the weights of the active components are renormalised to sum to one, which explains the two asymmetric weight distributions the interface reports, where image-only gives 49/36/15 (40/82, 30/82, 12/82) and link-only gives 46/34/20 (40/88, 30/88, 18/88). The two splits are deliberately different because the base weights themselves differ. A linked article usually contains a larger and more diagnostic slab of text than a vision-model caption of an image, so links are given a higher base weight (18 vs. 12), and renormalisation preserves that ratio when one side is missing.

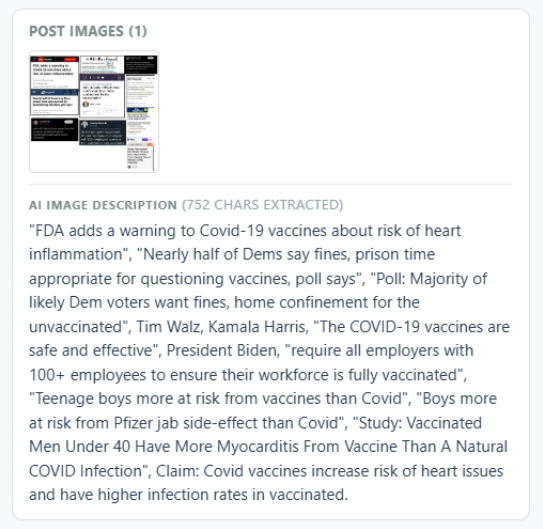


(a) Image-only post: the vision model produces a caption of the attached headline collage, which is then fed to the classifier (weights renormalise to 49/36/15).

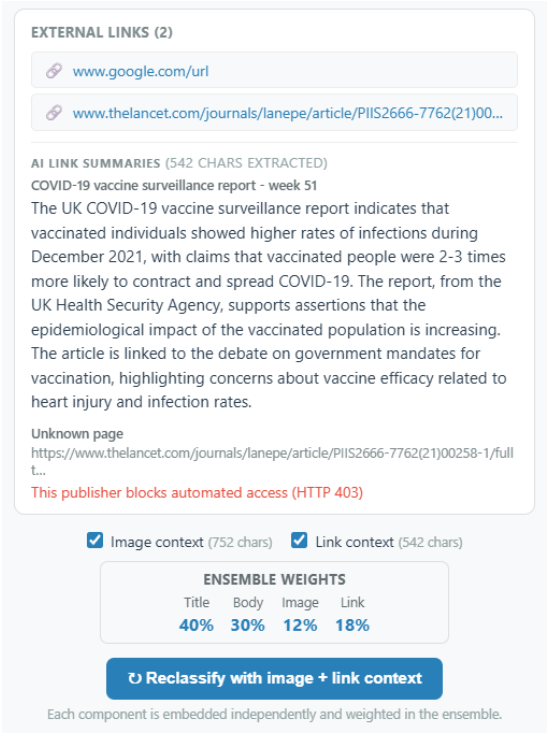


(b) Link-only post: the linked FEC page is summarised and re-embedded as part of the post (46/34/20).

Figure 4.17: Single-modality enrichment: only one of the image or link endpoints contributes extra context, and the base weights are renormalised over the three active components.



(a) Image + link post with both contexts resolved successfully: all four components feed the ensemble at the full 40/30/12/18 weighting.



(b) Image + link post where one linked publisher blocks automated access: the HTTP 403 is reported per-link without dropping the rest of the link context.

Figure 4.18: Full multimodal enrichment: both image and link context are available, so the classifier runs with all four components active.

4.6.5 Guard Rails and Edge Cases [METHODOLOGY]

The extension distinguishes between *soft warnings*, which are contextual notes attached to a classification that the model still produces, and *hard refusals*, which suppress the classification entirely because the input is outside the regime the model was trained for.

Figure 4.19 summarises the full decision logic that the backend applies to every post before a result is surfaced to the extension. The paragraphs that follow describe each decision point in prose.

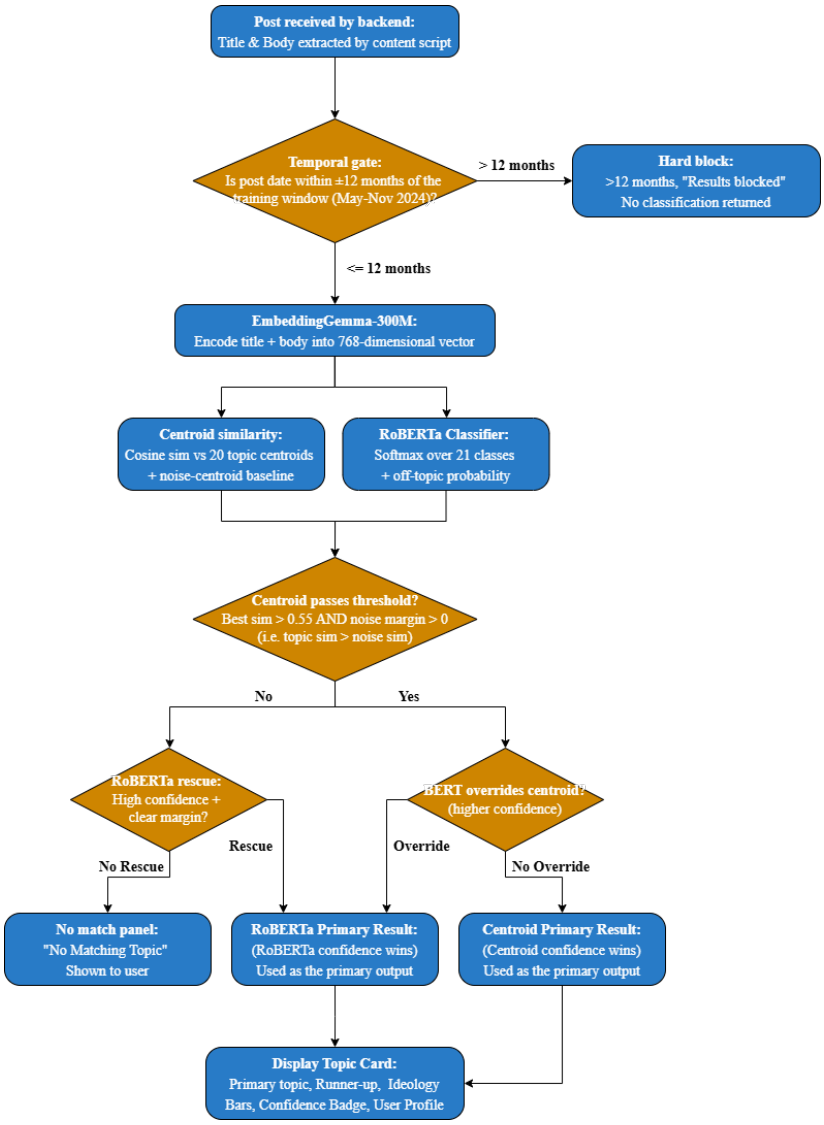


Figure 4.19: Classification decision flowchart for the Chrome extension backend. Each post passes through a temporal gate, then two classifiers run in parallel (centroid cosine similarity against 20 topic centroids, and RoBERTa softmax over 21 classes), and a threshold check picks the primary result or falls back to the no-match panel.

Temporal relevance: Every post carries the date extracted from Reddit's metadata. The backend compares that date to the training window (May-November 2024) and returns one of three states: posts inside the window trigger no warning, posts up to twelve months outside the window trigger a *soft* temporal warning that is shown alongside the normal classification card, saying that topic assignments may not reflect the discourse of that period, and posts more than twelve months outside the window trigger a *hard* block where the classifier is short-circuited before any centroid or RoBERTa result can be surfaced, the card is replaced by the no-match panel, and the warning banner wording changes to "Results blocked". The RoBERTa rescue path described below is explicitly suppressed when the hard temporal block fires, so a confident BERT prediction cannot override the

date gate. The rationale is that the topic scheme is time-stamped to the election period and extrapolating it to content from, say, a year and a half later would be misleading rather than just uncertain.

Content-length warnings: The backend counts words after normalisation and raises a tiered warning, all of which are soft and leave the classification card in place. "extremely short" (fewer than 3 words, "rough estimate only"), "very short" (fewer than 15 words, "results may be unreliable"), and "limited" (fewer than 30 words, "include more post content or comments"). If there is no body at all (an image or link post) or the body was deleted but the title survives, the popup injects an equivalent warning explaining that classification is running on the title alone.

Confidence warnings: When the classification is confident enough to surface but sits near the noise baseline, the backend attaches a borderline-match warning. This is either a low-topic-similarity note (best similarity under 65%) or a borderline-match note (best similarity in the 65-70% range with a noise margin below 0.02). These are soft warnings, the topic card is still shown, but the reader is told not to treat the match as definitive.

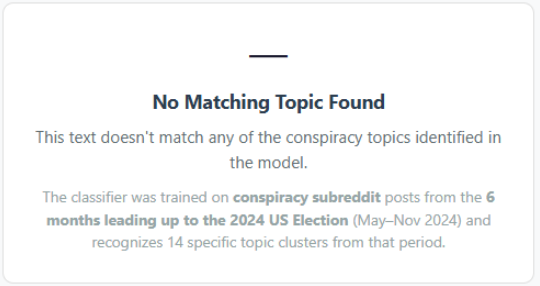
No-match state: A proper no-match, where the topic card is replaced by a "No Matching Topic" panel, is reserved for inputs that are genuinely far from every cluster. It is triggered when the best centroid similarity falls below an absolute floor of 0.55, when the noise centroid is more similar than any real topic (noise margin below zero), when BERT is strongly off-topic and the centroid signal is weak, or when the hard temporal block fires. No-match is not a silent failure mode, as the panel explicitly states that the text does not match any of the fourteen topics from the training window, so the reader knows the extension chose not to force the post into its nearest neighbour. The RoBERTa classifier can also *rescue* a borderline no-match when its top probability and margin are both high enough, so the two classifiers act as a mutual sanity check rather than as independent voters, the one exception is the hard temporal block, which suppresses the rescue regardless of BERT's confidence. When the centroid does surface a match but RoBERTa disagrees, RoBERTa is allowed to override the centroid only if its softmax confidence exceeds 0.60, its margin to the second-place class exceeds 0.20, *and* the centroid's own top-1-vs-top-2 similarity gap is below 0.08. The last condition exists because a wide centroid gap means the cosine pipeline has a confidently separated winner among its topics, which is exactly the case where the trained head is the more error-prone classifier (RoBERTa is known to be poorly calibrated on small, imbalanced classes, and several of the topics, such as 9/11 Truth at 116 posts, have far fewer training examples than the dominant Vaccine and Trump/Biden topics). When the centroid is itself uncertain (top-1 and top-2 within 8 per-

centage points), RoBERTa's supervised signal genuinely adds information and is allowed to override.

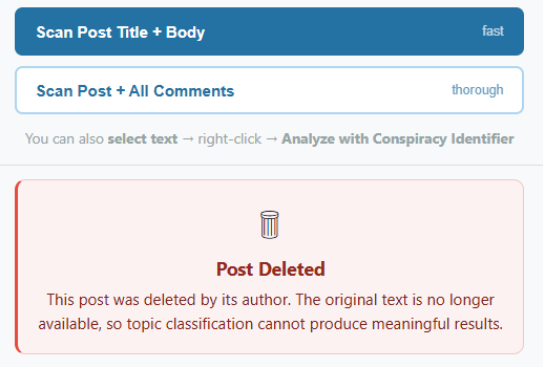
Deleted posts and empty inputs: Posts whose body has been deleted or removed are detected in the content script before any backend call and shown as a distinct "Post Unavailable" state rather than passed in as a near-empty string, since there is nothing meaningful left to classify. The detector covers both Reddit layouts, handling the legacy API markers `[deleted]` and `[removed]` (with and without the "by user" / "by moderators" suffixes), and the longer prose placeholders rendered by the newer shreddit Web Components for author deletions, moderator removals, admin and spam-filter removals, and third-party mirrors that echo the same sentence in a slightly different form. The backend duplicates the same regex family as a second line of defence in case the frontend check is bypassed. If the post author is still known (for example, the body was deleted but the username survived), the user-profile block is still attached underneath the unavailable-post card. Completely empty inputs (nothing survives normalisation) are caught by the backend with an HTTP 400 and rendered as a concise error message in the popup.

Finally, the "Analyze with Conspiracy Identifier" context-menu entry makes the topic model available on any page, not just Reddit, and right-clicking a text selection on an arbitrary website sends that text to `/analyze` and renders a compact, read-only version of the same topic card in a small overlay pinned to the current page. This is primarily meant for quick checks against off-Reddit articles, news headlines, or forum posts that the reader wants to situate against the conspiracy topic scheme without leaving the page.

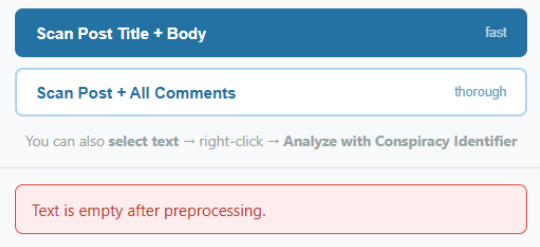
(a) Hard temporal block (18 months after the training window, "Results blocked") co-occurring with a limited-text content warning on the same post.



(b) No-match state: the post is not close enough to any of the 14 topics from the training window, so no classification is surfaced.



(c) Unavailable post: the extension detects a deletion or moderator-removal marker (any of the layouts recognised by the expanded regex family) and refuses to classify rather than embedding an empty placeholder.



(d) Empty text after cleaning: nothing survives the normalisation filter, so the backend returns an early error instead of classifying.

Figure 4.20: Guard-rail states. Each represents a case where the pipeline deliberately yields no classification rather than forcing the post into the nearest topic.

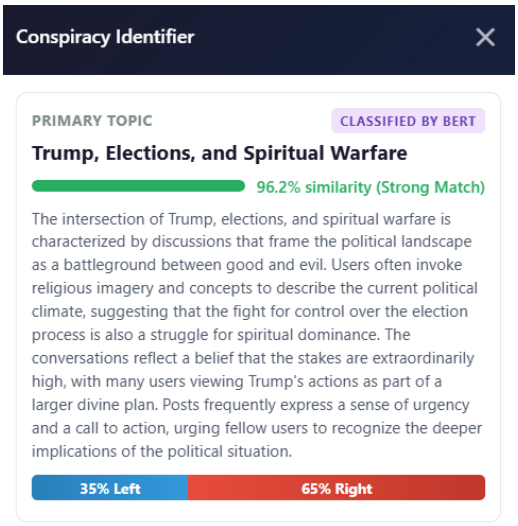


Figure 4.21: Compact overlay rendered by the context-menu action. The card mirrors the popup's primary-topic layout with the confidence bar, description, and ideology category, but omits the transparency panels for readability.

4.7 Scalability and Computational Considerations

The pipeline was developed on a single consumer laptop (Intel i7, 16 GB RAM, NVIDIA RTX 3060 6 GB VRAM) and runs end-to-end in under two hours on the full corpus of 55,154 users and 13,194 HVU documents. The most expensive stage in the pipeline is the EmbeddingGemma-300M embedding step, which takes roughly 40 minutes on the RTX 3060 for the full document set. Every other stage, from the overlap computation through to the network construction and polarisation metrics, completes in minutes or seconds because the operations are either simple set intersections, vector arithmetic, or sparse-graph traversals that scale linearly with the number of users or edges.

The design is modular by construction, meaning each of the six stages reads its inputs from the shared `pipeline_results/` directory and writes its outputs back to the same location, with no in-memory coupling between stages. This means that a single stage can be re-run after a code change without repeating the entire pipeline, and the `verify_outputs.py` integrity checker catches stale or missing artefacts before they propagate downstream. For a researcher who wants to extend the corpus, the main constraint is the embedding step, which grows linearly with the number of documents, and the HDBSCAN clustering, which grows roughly as $O(n \log n)$ for the default leaf-cluster selection used here. Neither is a bottleneck at the current scale, and both would remain tractable on the same hardware for corpora several times larger.

Extending the pipeline to new Reddit data is straightforward. Adding more conspiracy or political subreddits requires only updating the subreddit classification lists and re-running the pipeline from Stage 1, since the overlap logic, ideology computation, and

downstream stages all work off the subreddit labels rather than hard-coded subreddit names. The data format itself is the standard Reddit submission and comment schema exported by tools like Pushshift or the Reddit API, so no format adaptation is needed for additional subreddits.

Adapting the pipeline to a different social media platform would require more substantial changes. The ideology inference depends on the concept of named communities with known political leanings (subreddits in the Reddit case), and a platform without that structure (e.g. Twitter/X, where ideology would need to be inferred from follow graphs or hashtag usage) would need a different Stage 2 entirely. The topic clustering and Chrome extension stages, on the other hand, operate on plain text and are platform-agnostic, so they could be reused with minimal modification as long as the upstream stages provide the same output format (user ideology vectors and a corpus of labelled posts).

Chapter 5

Results

This chapter presents the results for each of the seven research questions defined in Section 1.3, which were derived from the three high-level objectives proposed by the thesis supervisor:

- **Identification** (supervisor RQ1 - *what conspiracy theories were most discussed?*): addressed by the BERTopic topic discovery reported in Chapter 4 and elaborated in the analysis populations and per-topic composition of Section 5.1.
- **Ideological structure** (supervisor RQ2 - *how do conspiracy narratives differ across left vs. right communities?*): addressed by RQ1-RQ6 (Sections 5.1-5.6), which test ideological composition, engagement intensity, audience entropy, predictive power of ideology, bridging-score differences, and subreddit ecosystem sorting.
- **Classification utility** (supervisor RQ3 - *can a reliable classifier be built?*): addressed by the Chrome extension methodology in Section 4.6 and evaluated quantitatively in Section 5.8.

Statistical methods: Because the discussion below leans on a handful of standard statistical tools, each one is introduced here in layman's terms before the RQ-specific results are presented.

- *p-value.*
  - **Definition:** the probability of seeing an observation at least as extreme as the one we got, *assuming the null hypothesis is true* (i.e. assuming there is no real effect). Small p-values mean the observed pattern would be unlikely by chance alone. The conventional cut-off is $\alpha = 0.05$, we call a result *statistically significant* when $p < 0.05$. A p-value says nothing about how *large* or *important* an effect is, only about how surprising it would be under the null.

- **Formula:** if T is the test statistic and t_{obs} its observed value, $p = \Pr(|T| \geq |t_{\text{obs}}| \mid H_0)$.
- **Worked Example (coin flipped 100 times, H_0: coin is fair).** Under H_0, $X \sim \text{Binomial}(100, 0.5)$: mean 50, standard deviation 5. The two-sided p-value for an observed X is $\Pr(|X - 50| \geq |X_{\text{obs}} - 50|)$:
  - $X = 50$: $p = 1.00$, perfectly consistent with fairness.
  - $X = 55$: $p \approx 0.37$, well above 0.05, not significant.
  - $X = 58$: $p \approx 0.13$, still above 0.05, not significant.
  - $X = 60$: $p \approx 0.057$, just above 0.05, not quite significant.
  - $X = 61$: $p \approx 0.035$, below 0.05, significant, reject fairness.
  - $X = 40$: $p \approx 0.057$, symmetric to $X = 60$ (same distance from 50), a two-sided test treats tail-heavy and head-heavy outcomes identically.
  - $X = 39$: $p \approx 0.035$, symmetric to $X = 61$, significant.

  So the $\alpha = 0.05$ cutoff is on the *p-value*, not on the head count directly, inverting it gives the equivalent head-count cutoff $X \leq 39$ or $X \geq 61$.

- *Benjamini-Hochberg (BH) correction [8].*
  - **Definition:** when we run many tests at once, the chance of at least one false positive grows quickly. With $m = 14$ independent tests at $\alpha = 0.05$, the probability of *at least one* false positive is $1 - 0.95^{14} \approx 0.51$. BH controls the expected *false discovery rate* (FDR) across the family of tests.
  - **Formula:** sort the m raw p-values ascending as $p_{(1)} \leq p_{(2)} \leq \cdots \leq p_{(m)}$, find the largest k such that $p_{(k)} \leq \frac{k}{m}\alpha$, reject all hypotheses with $p \leq p_{(k)}$. Unless stated otherwise, every p-value reported below is BH-adjusted.
  - **Worked Example ($m = 14$ topics, $\alpha = 0.05$).** The BH thresholds are $p_{(k)} \leq k \cdot 0.05/14$:
    - $k = 1$: threshold ≈ 0.0036.
    - $k = 2$: threshold ≈ 0.0071.
    - $k = 3$: threshold ≈ 0.0107.
    - $k = 14$: threshold $= 0.05$ (the raw cut-off only applies to the very last one).

    Case A: sorted p-values $\{0.004, 0.020, 0.040, \ldots\}$. The first fails ($0.004 > 0.0036$), the second fails ($0.020 > 0.0071$), the third fails ($0.040 > 0.0107$): *none* survive, even though each looks individually significant at $\alpha = 0.05$. **Case B:** sorted p-values $\{0.001, 0.003, 0.040, \ldots\}$. Now $0.001 < 0.0036$ and

$0.003 < 0.0071$, so both survive, $0.040 > 0.0107$ so the third does not. BH protects us from over-claiming without discarding genuinely strong signals.

- *Chi-squared test of independence [39].*
  - **Definition:** a test for whether two categorical variables occur together more often than chance would predict. It compares observed counts against the counts we would expect if the two variables were independent, and asks whether the gap is too large to be luck.
  - **Formula:** for an $r \times c$ contingency table with observed counts O_{ij} and expected counts $E_{ij} = (\text{row}_i \cdot \text{col}_j)/n$,

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{c} \frac{(O_{ij} - E_{ij})^2}{E_{ij}},$$

    compared against a χ^2 distribution with $(r-1)(c-1)$ degrees of freedom. Large $\chi^2 \Rightarrow$ small p-value $\Rightarrow$ the variables are probably not independent.
  - **Worked Example (topic with** 100 **posters, baseline** 37% **left,** 63% **right, so** $E = (37, 63)$**, df** $= 1$**).**
    * Observed $O = (37, 63)$: $\chi^2 = 0$, $p = 1.00$. Matches baseline exactly, no evidence of skew.
    * Observed $O = (45, 55)$: $\chi^2 = (45-37)^2/37 + (55-63)^2/63 \approx 1.73 + 1.02 = 2.75$, $p \approx 0.097$. Mild skew, not significant.
    * Observed $O = (50, 50)$: $\chi^2 \approx 4.57 + 2.68 = 7.25$, $p \approx 0.007$. Significant left-lean.
    * Observed $O = (60, 40)$: $\chi^2 \approx 14.3 + 8.4 = 22.7$, $p < 10^{-5}$. Extreme left-lean.
    * Observed $O = (20, 80)$: $\chi^2 \approx 7.81 + 4.59 = 12.4$, $p \approx 0.0004$. Significant right-lean (chi-squared is two-sided, any direction of deviation inflates it).

- *Cramér's V.*
  - **Definition:** an effect-size companion to chi-squared. Chi-squared tells us whether a deviation is statistically surprising, V tells us whether it is *large enough to matter*. Conventional cut-offs: $V < 0.1$ negligible, 0.1-0.3 small, 0.3-0.5 medium, > 0.5 large.
  - **Formula:** $V = \sqrt{\frac{\chi^2}{n \cdot \min(r-1, c-1)}}$, bounded in $[0, 1]$, where n is the total sample size.
  - **Worked Example (**2×2 **table, baseline** 37/63**,** $\min(r-1, c-1) = 1$**).**

* $n = 100$ users, observed 45/55: $\chi^2 \approx 2.75$, $p \approx 0.097$, $V = \sqrt{2.75/100} \approx 0.166$. Not significant, but the effect size is already "small"-a bigger sample would probably push it over $p = 0.05$.
* $n = 10{,}000$ users, observed 3,800/6,200 vs expected 3,700/6,300: $\chi^2 \approx 4.3$, $p \approx 0.038$ ("significant"), but $V = \sqrt{4.3/10{,}000} \approx 0.021$ (negligible). The huge sample made a tiny deviation detectable, the p-value alone would overstate the finding.
* $n = 100$ users, observed 60/40: $\chi^2 \approx 22.7$, $V \approx 0.476$: medium-to-large effect and strongly significant-the genuine case where both pieces agree.

- *Mann-Whitney U test [29].*
  - **Definition:** a non-parametric test for "do these two groups come from different distributions?", based on ranks rather than raw values. Small U means one group consistently sits on one end of the ranking, which is unlikely if both groups share a distribution. We prefer it to the t-test here because per-topic engagement is heavily right-skewed (most users post once or twice, a handful post hundreds of times), violating the t-test's normality assumption.
  - **Formula:** pool all $n_1 + n_2$ observations, rank them ascending, and let R_1 be the sum of ranks in group 1. Then

$$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1, \qquad U = \min(U_1, U_2).$$

  - **Worked Example (two groups of 5 post counts each, ranks 1-10).**
    * Left $\{1,1,2,3,150\}$ vs Right $\{1,2,2,4,5\}$. The outlier 150 pulls the left *mean* to 31.4 vs 2.8, but after ranking the outlier becomes just rank 10. Rank sums are similar: Mann-Whitney is *not* fooled and returns a high p-value.
    * Left $\{1,2,3,4,5\}$ vs Right $\{6,7,8,9,10\}$. Left gets ranks 1-5 ($R_1 = 15$), right gets 6-10: $U_1 = 5 \cdot 5 + 15 - 15 = 25$, $U_2 = 0$, $U = 0$. Perfect separation-as small a U as possible, the test returns $p \approx 0.008$.
    * Left $\{3,4,5,6,7\}$ vs Right $\{3,4,5,6,7\}$. Identical groups give $R_1 = R_2$, $U_1 = U_2 = 12.5$, $U = 12.5$ (the maximum for $n_1 = n_2 = 5$). $p = 1.00$, no evidence of difference.

- *Cohen's d [13].*
  - **Definition:** effect-size companion to the t-test and Mann-Whitney. It is the difference between the two group means, expressed in units of pooled standard

deviation, so it is comparable across studies and scales. Conventional cut-offs: $|d| < 0.2$ negligible, 0.2-0.5 small, 0.5-0.8 medium, > 0.8 large. We report d alongside the p-value so statistical significance and practical significance can be judged separately.

  - **Formula:** $d = \dfrac{\bar{x}_1 - \bar{x}_2}{s_p}, \quad s_p = \sqrt{\dfrac{(n_1-1)s_1^2 + (n_2-1)s_2^2}{n_1+n_2-2}}$.
  - **Worked Example (pooled SD fixed at $s_p = 10$).**
    * $\bar{x}_1 = 5, \bar{x}_2 = 5$: $d = 0$, identical groups.
    * $\bar{x}_1 = 5, \bar{x}_2 = 6$: $d = -0.1$, negligible-this is the kind of difference a huge sample can flag as "significant" while still being uninteresting.
    * $\bar{x}_1 = 5, \bar{x}_2 = 8$: $d = -0.3$, small.
    * $\bar{x}_1 = 5, \bar{x}_2 = 12$: $d = -0.7$, medium.
    * $\bar{x}_1 = 5, \bar{x}_2 = 15$: $d = -1.0$, large.
    * $\bar{x}_1 = 15, \bar{x}_2 = 5$: $d = +1.0$-sign just flips, indicating group 1 is higher.

- *Odds ratio (OR).*
  - **Definition:** the *odds* of an outcome are odds $= p/(1-p)$, where p is the probability of the outcome. The odds ratio compares two groups on this odds scale. We use it in RQ4 because it is the natural output of logistic regression.
  - **Formula:** $\text{OR} = \dfrac{p_1/(1-p_1)}{p_2/(1-p_2)}$.
  - **Worked Example (right-leaning vs left-leaning engagement with a topic).**
    * $p_1 = p_2 = 0.25$: OR $= 1.00$, no difference.
    * $p_1 = 0.30$, $p_2 = 0.20$: odds 0.43 vs 0.25, OR ≈ 1.72-right-leaning users have $\approx 72\%$ higher odds of engaging.
    * $p_1 = 0.40$, $p_2 = 0.20$: odds 0.67 vs 0.25, OR ≈ 2.67-right-leaning odds are $\approx 2.7\times$ higher.
    * $p_1 = 0.20$, $p_2 = 0.40$: OR ≈ 0.375-inverse of the above, right-leaning odds are $\approx 37.5\%$ of left-leaning, i.e. the topic skews left.
    * $p_1 = 0.50$, $p_2 = 0.50$: OR $= 1.00$, again no difference even though engagement is high.

  Note: the odds ratio is *not* the same as the relative risk p_1/p_2. For small p_1, p_2 they are close, but as probabilities approach 1 the odds blow up and the two diverge.

- *Independent-samples t-test.*

- **Definition:** a parametric test for whether the means of two groups differ, assuming each group's values are approximately normally distributed with comparable variances. It is the classical counterpart to Mann-Whitney U, it has more power when the normality assumption holds but can be misled by outliers and heavy tails. We use Mann-Whitney for per-topic engagement counts (which are heavily right-skewed) and a Welch t-test for the bridging-score comparisons in RQ5 (which are roughly symmetric and bounded in $[0,1]$).
- **Formula:** $t = \dfrac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$ (Welch variant, unequal variances), compared against a t-distribution with Welch-Satterthwaite degrees of freedom.
- **Worked Example (two groups, $n_1 = n_2 = 50$, $s_1 = s_2 = 1$).**
  - $\bar{x}_1 = 0.50$, $\bar{x}_2 = 0.50$: $t = 0$, $p = 1.00$. Identical means.
  - $\bar{x}_1 = 0.52$, $\bar{x}_2 = 0.50$: $t \approx 0.14$, $p \approx 0.89$. Tiny difference, not significant.
  - $\bar{x}_1 = 0.60$, $\bar{x}_2 = 0.50$: $t \approx 0.71$, $p \approx 0.48$. Still not significant at this sample size.
  - $\bar{x}_1 = 0.90$, $\bar{x}_2 = 0.50$: $t \approx 2.83$, $p \approx 0.006$. Strongly significant, though the effect is driven as much by having $s = 1$ and $n = 50$ as by the raw 0.4 gap.

• *Shannon entropy (normalised) [45].*

- **Definition:** a summary of how spread out a discrete distribution is. It is zero when all mass concentrates on one category and maximal when every category is equally likely. We use it in RQ3 to describe each topic's audience shape (left vs. right) and in RQ4 as a per-user *political diversity* feature (how evenly a user's activity spreads across the six ideology tiers). The *normalised* form divides by $\log_2 k$ (where k is the number of categories) to keep the output in $[0,1]$ regardless of k, so audience-level entropies and user-level entropies are directly comparable.
- **Formula:** for a distribution $p = (p_1, \ldots, p_k)$ with $\sum_i p_i = 1$, $H(p) = -\sum_{i=1}^{k} p_i \log_2 p_i$, and the normalised form is $\tilde{H}(p) = H(p)/\log_2 k$.
- **Worked Example ($k = 2$, left/right audience).**
  - $p = (1.00, 0.00)$: $\tilde{H} = 0$, fully one-sided audience.
  - $p = (0.85, 0.15)$: $H \approx 0.610$, $\tilde{H} \approx 0.610$. Matches the lowest-entropy topic in RQ3.
  - $p = (0.70, 0.30)$: $H \approx 0.881$, $\tilde{H} \approx 0.881$.
  - $p = (0.50, 0.50)$: $H = 1$, $\tilde{H} = 1$, maximally mixed.

- *Spearman rank correlation (ρ) [48].*
  - **Definition:** a non-parametric measure of monotonic association between two variables, computed as the Pearson correlation of their *ranks* rather than their raw values. Like Mann-Whitney, it is robust to outliers and to non-linear (but still monotonic) relationships. Values lie in $[-1,1]$, with 0 meaning no monotonic relationship, positive values meaning the two variables tend to rank together, and negative values meaning they rank oppositely. Conventional cut-offs: $|\rho| < 0.1$ negligible, 0.1-0.3 weak, 0.3-0.5 moderate, > 0.5 strong.
  - **Formula:** if r_i^x and r_i^y are the ranks of x_i and y_i, $\rho = 1 - \frac{6\sum_i d_i^2}{n(n^2-1)}$, where $d_i = r_i^x - r_i^y$ (simplified form, assuming no ties).
  - **Worked Example ($n = 50$ users, score vs. engagement).**
    * $\rho = 0.01$: no monotonic relationship, scatter is essentially random.
    * $\rho = 0.15$: weak positive relationship, the RQ5 value for right-leaning score-engagement.
    * $\rho = 0.26$: weak-to-moderate, the RQ5 value for left-leaning score-engagement.
    * $\rho = 0.70$: strong positive, users with high scores almost always post more.
    * $\rho = -0.30$: moderate negative, the two would rank in opposite directions.
- *Topic-level null.*
  - **Definition:** not a single test but a way of reading a *family* of tests run across the topic set. When every per-topic test is run at $\alpha = 0.05$ and none survives BH correction, we say the family of comparisons is a *topic-level null*: there is no topic at which the corrected evidence supports rejecting the null hypothesis, and the family as a whole is consistent with no effect. It is a statement about the family of tests, not a claim that each individual raw p is large, some may look significant on their own, but none survives the multiple-comparison correction.
  - **How we use it:** RQ1 (composition), RQ2 (intensity), and RQ4 (odds ratios) each produce a family of 14 main-topic and up to 20 subtopic tests. When none survives BH, the corresponding RQ is reported as a topic-level null even though individual topics may have χ^2, d, or OR point estimates slightly away from parity. This is the only honest reading when the family-wide FDR is controlled, and it is also why the RQ2 discussion separates "direction of the point estimates" from "population-level claim".

In short, p-values (with BH correction) tell us whether a pattern is real, while Cramér's V, Cohen's d, and odds ratios tell us whether it is large enough to matter. Both are needed because a large dataset can make even trivial differences statistically significant, so reporting only p-values would overstate the findings.

The main analysis population is **1,106 users** who have both a clear ideological lean and at least one post in a non-noise topic cluster (412 left-leaning, 694 right-leaning, giving the 37.3% / 62.7% baseline used as the reference split throughout). The wider population for the bridging-score analysis in Section 5.5 includes **6,299 users**, every account in the ideology lookup table with a computable score, regardless of whether they posted in a named topic. Each RQ in this chapter corresponds to one analysis script in the project repository, the concrete mapping is kept in the repository's `README` rather than reproduced here, to keep the chapter focused on findings rather than file paths.

The 14 main topics and 9 subtopics used throughout this chapter are defined, described, and sized in Tables 4.5 and 4.7 of Chapter 4 (Section 4.4.4 and 4.4.5). Topic 0 (outlier bucket) is excluded from every analysis here. All user counts cited below refer to unique posters with a classified ideology lean, and the outlier-excluded figures are the ones that feed into the RQ1-RQ4 tests.

5.1 RQ1: Ideological Composition of Conspiracy Topics

Question: Which ideological groups dominate each conspiracy topic?

Chi-squared tests were run for each of the 14 main topics and 20 subtopic-level categories, comparing the observed left/right split to the population baseline (37.3% left, 62.7% right).

Result: No topic shows a statistically significant deviation in ideological composition after BH correction, not at the main topic level (0/14 significant, max Cramér's $V = 0.162$) and not at the subtopic level either (0/19 tested, max Cramér's $V = 0.170$, one of the 20 subtopic categories is dropped from the RQ1/RQ2 tests because it has too few ideology-classified users to support them). All effect sizes are negligible to small.

The highest over-representation is 1.14× (left-leaning in Russia/Ukraine) at the main level and 1.22× (left-leaning in Russian Influence) at the subtopic level. These are descriptive patterns that do not reach statistical significance. Figure 5.1 shows the left/right composition per topic, and Figure 5.2 shows the over-representation heatmap.

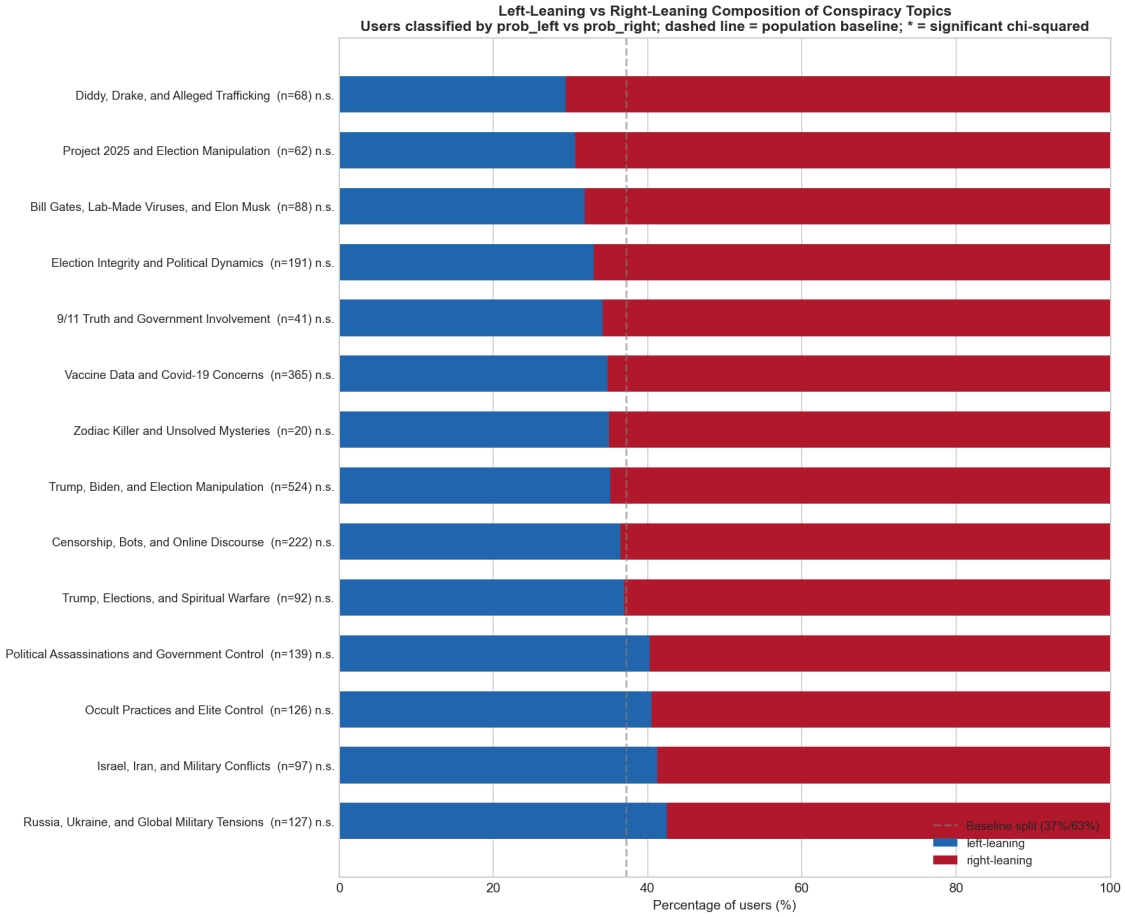


Figure 5.1: Left vs. right-leaning user composition per conspiracy topic (main topics).

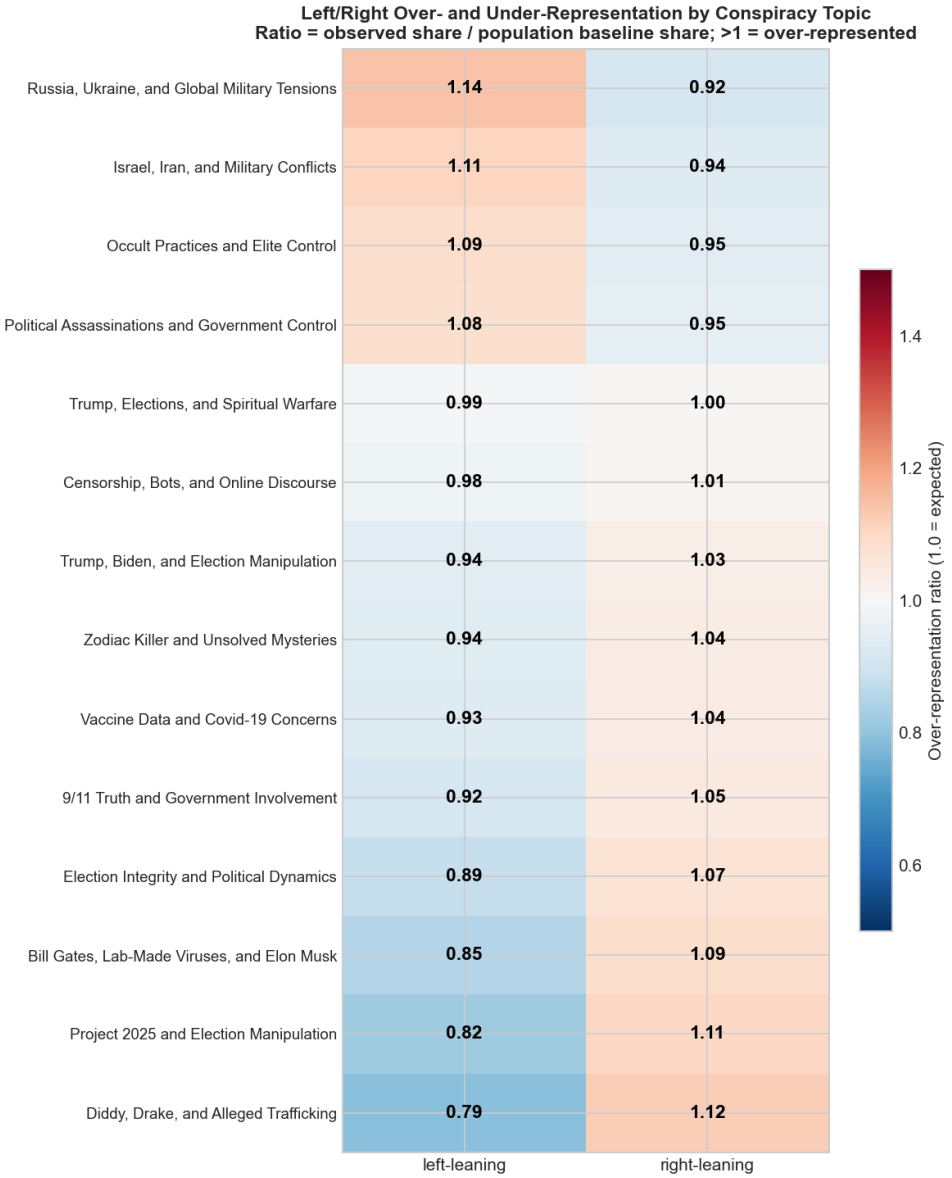


Figure 5.2: Over-representation heatmap: ratio of observed to expected ideological composition per topic.

Figure 5.3 shows the same data at the six-tier granularity level, breaking each topic's audience into strong/lean/weak left and right. The pattern is consistent across all topics, with no tier dominating any individual topic.

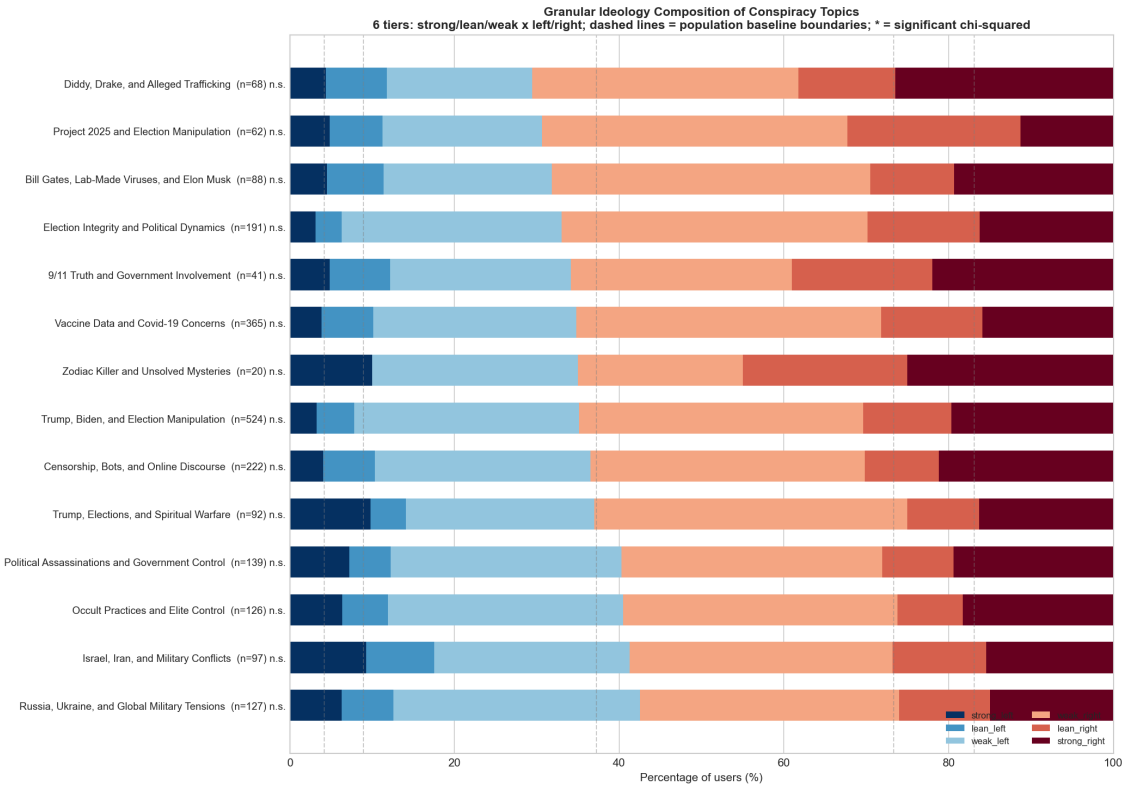


Figure 5.3: Six-tier ideology composition per topic. Dashed lines mark population baselines. No topic reaches significance.

Table 5.1 lists the exact χ^2, raw p, BH-adjusted p, and Cramér's V for all 14 main topics so that the null result can be audited directly.

Table 5.1: Full chi-squared results for RQ1 (ideological composition per topic), including raw and BH-adjusted p-values. All adjusted p-values exceed 0.05, confirming no statistically significant ideological skew at the topic level.

#	Topic	χ^2	p (raw)	p (BH)	Cramér's V
1	Trump/Biden/Election	1.024	0.312	0.643	0.044
2	Vaccine/Covid-19	0.943	0.332	0.643	0.051
3	Election Integrity	1.488	0.223	0.643	0.088
4	Bill Gates/Lab Viruses	1.111	0.292	0.643	0.112
5	Censorship/Bots	0.056	0.814	0.899	0.016
6	Russia/Ukraine	1.508	0.219	0.643	0.109
7	Israel/Iran	0.659	0.417	0.643	0.082
8	Political Assassinations	0.548	0.459	0.643	0.063
9	Occult/Elite Control	0.561	0.454	0.643	0.067
10	Trump/Spiritual Warfare	0.003	0.953	0.953	0.006
11	Diddy/Drake/Trafficking	1.788	0.181	0.643	0.162
12	Zodiac Killer	0.043	0.835	0.899	0.047
13	9/11 Truth	0.169	0.681	0.867	0.064
14	Project 2025	1.158	0.282	0.643	0.137

Discussion: Two things are significant in the results. First, the *direction* of the small deviations shows right-leaning users are slightly over-represented in most narratives and left-leaning users are slightly over-represented in geopolitics-adjacent ones (Russia/Ukraine, Russian Influence), but in every case the gap is within 14-22% of the baseline rate, which is far below what could support a "this is a right-wing topic" framing. Second, the subtopic-level analysis was specifically included to test whether a coarse 14-topic scheme was hiding stronger sorting inside each cluster, which is not the case. The max Cramér's V rises only from 0.162 (14 main topics) to 0.170 (19 subtopics tested), which is still in the negligible-to-small regime, so the null pattern is not an artefact of topic granularity. The practical implication for later RQs is that topic identity by itself carries very little ideological signal, which sets up the RQ4 finding that *how broadly* a user participates matters more than *which side* they are on. More broadly, this result pushes back against the popular framing that specific conspiracy theories "belong" to one end of the political spectrum. In this corpus, during a heated election cycle, the audience of every single conspiracy narrative is drawn from both sides in roughly the same proportions as the overall population. If a platform or a media-literacy programme were to target conspiracy content by the assumed ideology of its audience, this finding suggests they would be targeting the wrong axis entirely.

5.2 RQ2: Engagement Intensity by Ideology

Question: Do left-leaning and right-leaning users differ in how intensely they engage with each topic?

Mann-Whitney U tests were used to compare per-topic engagement intensity between left and right-leaning users.

Result: No topic shows a significant difference in engagement intensity after BH correction (0/14 main topics, 0/19 subtopics tested, the same small-N exclusion applies as in RQ1). The largest Cohen's d is 0.24 (small) at the main level and 0.48 (small) at the subtopic level. Figure 5.4 shows the effect sizes per topic.

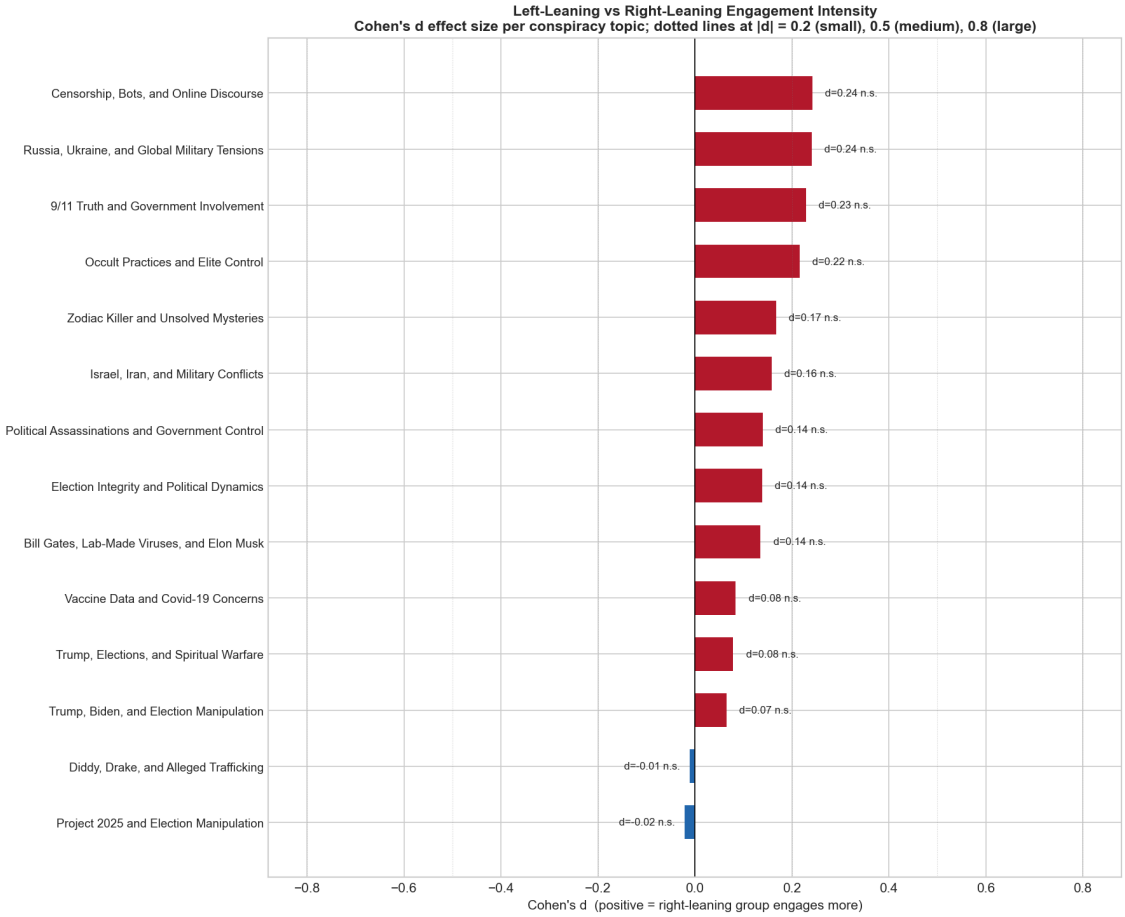


Figure 5.4: Cohen's d effect sizes for left vs. right engagement intensity per topic. All values fall inside the "negligible to small" band ($|d| < 0.5$).

Figure 5.5 shows the mean engagement intensity side by side for each topic. The figure requires careful reading, as for several topics the two bars look very different in height, yet all Cohen's d values stay inside the small band. The reason is that the bars plot the group *mean*, which is easily inflated by a handful of heavy-posting users, especially when one group is very small. For example, Zodiac Killer ($n_L = 7$) and 9/11 Truth ($n_L = 14$) have so few left-leaning posters that one user posting a lot can triple the group mean, whereas Bill Gates/Musk and Censorship/Bots have the same issue for right-leaning users. Cohen's d controls for this by scaling the mean difference by the pooled standard deviation, so if a group's mean is high because a few outliers dragged it up, the standard deviation is also high, and d stays small. The bars therefore look more dramatic than the effect sizes warrant, and no topic survives BH correction.

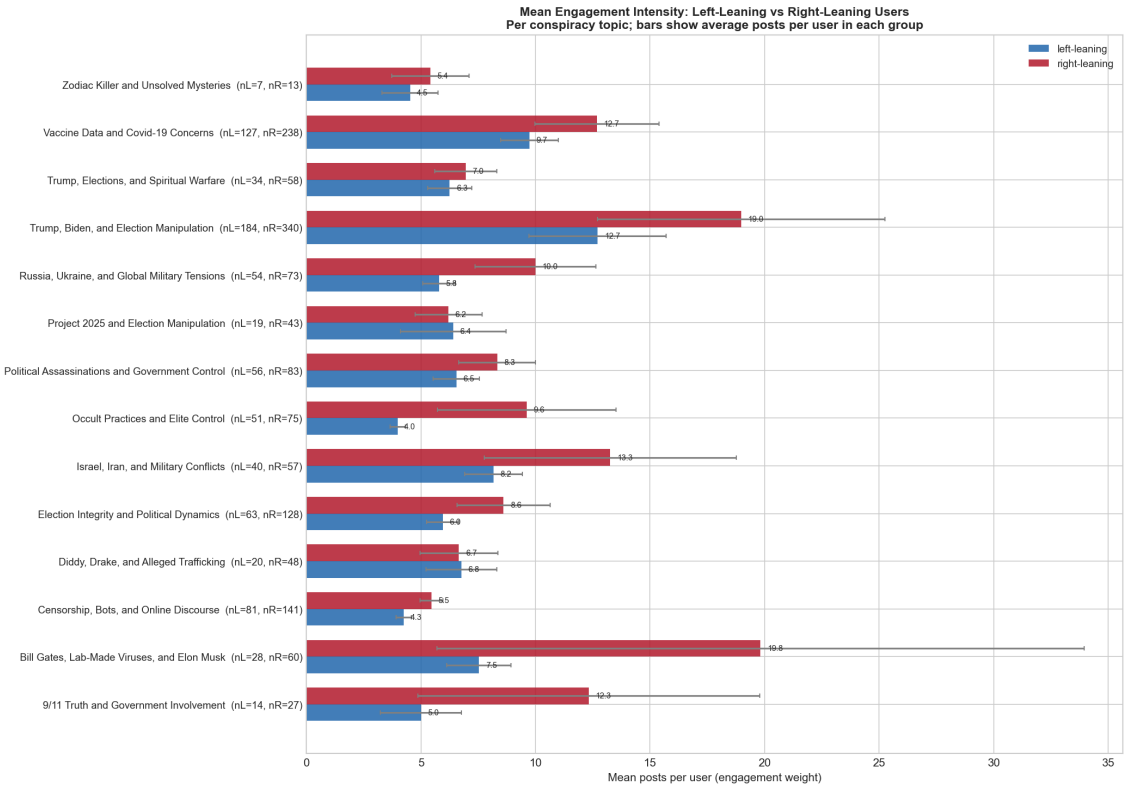


Figure 5.5: Mean engagement intensity (posts per user) by ideology for each topic. Bars can appear visually large because group means are sensitive to heavy-posting outliers, especially in small groups. Error bars show the standard error of the mean. Cohen's d (Figure 5.4) is the correct effect-size measure, with all values falling in the negligible-to-small band ($|d| \leq 0.24$) and none surviving BH correction.

Figure 5.6 expresses the same intensity comparison on a different scale. For each topic, it plots the *right-leaning engagement-share gap*, defined as the share of a topic's total posts produced by right-leaning users minus their share of that topic's users. A value of 0 means right-leaning users post at the same per-user rate as everyone else, positive values mean they over-contribute (red bar), negative values mean they under-contribute (which would render the bar blue). Every topic in the corpus sits on the positive side of zero, so every bar is red. This means that right-leaning users consistently produce a slightly larger slice of each topic's engagement than their headcount alone would predict. The magnitudes are small (typically a few percentage points), which is why the associated Cohen's d values remain in the negligible-to-small band and why no test survives BH correction, but the direction is uniform and worth acknowledging rather than hiding behind the null label.

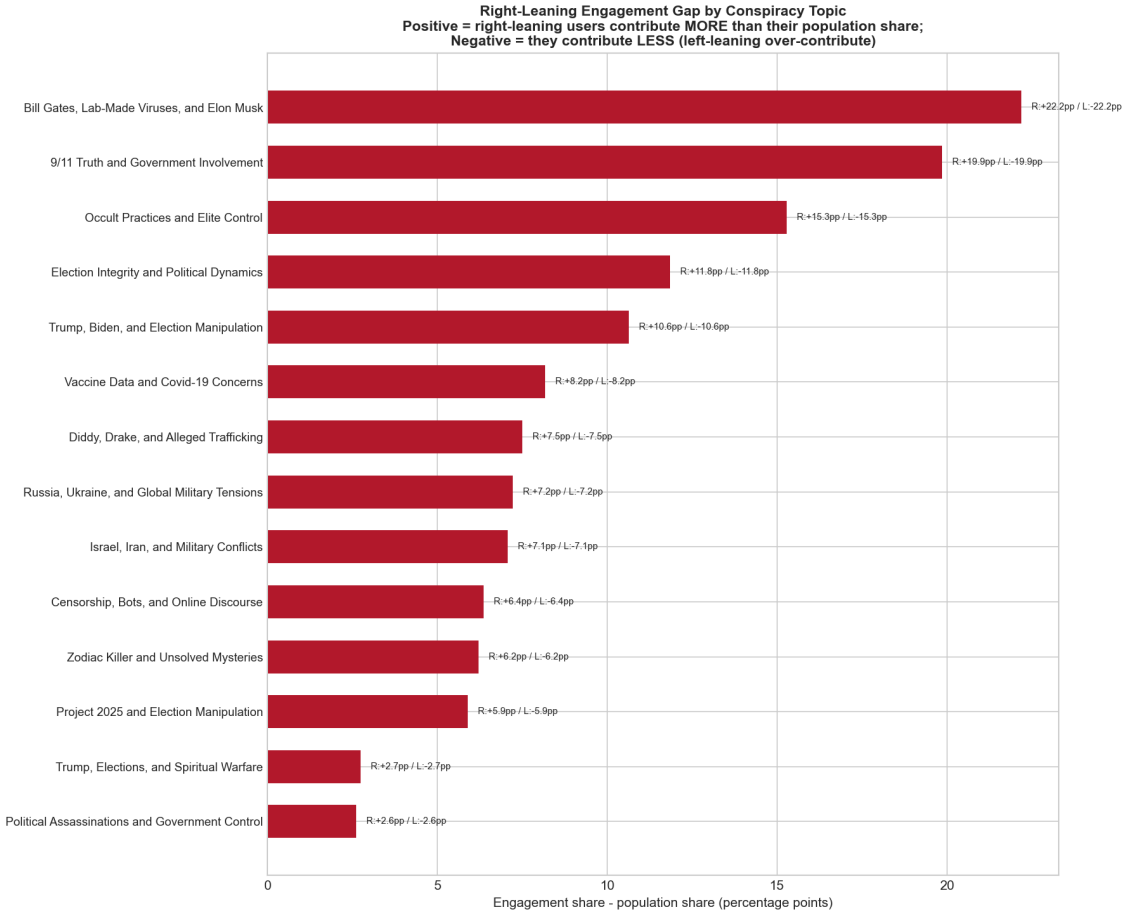


Figure 5.6: Right-leaning engagement-share gap per topic (engagement share − population share, in percentage points). Bars would be blue for any topic where right-leaning users under-contribute, all bars are red because the gap is positive for every topic, though the magnitudes are small.

Breaking the analysis down to the six-tier granularity (Figure 5.7) confirms the same pattern. Strong, lean, and weak tiers on both sides contribute to each topic roughly in proportion to their population share. The figure is restricted to the six busiest topics because finer-grained tiers require at least three users per cell to produce a reliable mean and many smaller topics fail that threshold for at least one tier, extending the panel to all twenty categories would fill much of the plot with hatched “unreliable” bars.

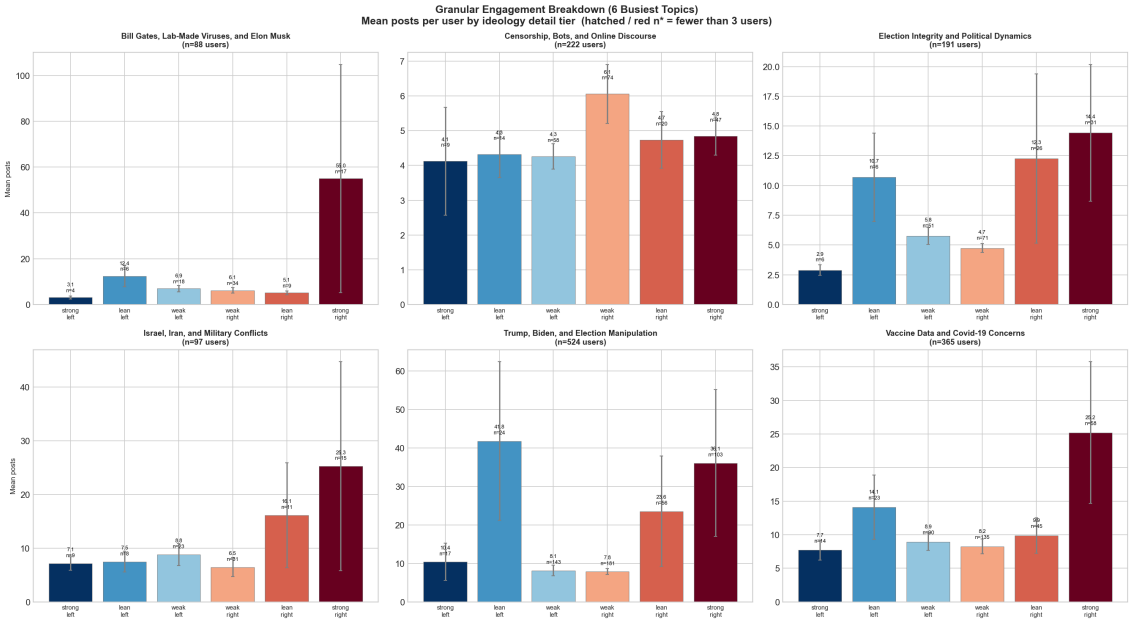


Figure 5.7: Six-tier engagement intensity for the six busiest topics (topics with enough users in every tier to support a reliable per-tier mean). As in Figure 5.5, right-side tiers (weak/lean/strong right) tend to sit marginally above their left-side counterparts, but no tier dominates any individual topic beyond its baseline share.

Discussion: The contrast between *composition* (RQ1) and *intensity* (RQ2) is the main thing to notice here. Under a strict echo-chamber model one would expect both membership and per-user activity to sort by ideology, but the data show only a faint, non-significant tilt in membership and a similarly faint, non-significant tilt in intensity. The intensity tilt has a consistent direction, right-leaning users post slightly more per-user in nearly every topic, visible both in the side-by-side means of Figure 5.5 and in the uniformly-positive engagement-share gap of Figure 5.6, but the magnitude stays inside the "small or smaller" band ($|d| \leq 0.24$ at the main level) and no test survives BH correction. Calling the pattern "null" should therefore be understood in the formal sense, as the effect sizes are too small and the corrected p-values too large to claim a reliable population-level difference. The one place where the effect size creeps up to "small" ($d = 0.48$ at the subtopic level) is in a narrative where the sub-population is already small (Russian Influence), so the larger d mostly reflects thinner per-group samples rather than a genuine behavioural gap, this is exactly what BH correction is designed to filter out, and the corrected p remains above 0.05. Taken together, RQ1 and RQ2 argue that ideology is not the variable that discriminates conspiracy topic engagement in this corpus. The uniform direction of the intensity tilt is still worth flagging for practical purposes, since it means that aggregate engagement metrics (total posts, total comments) in conspiracy subreddits will look slightly right-heavy even when the per-user rate difference is negligible. A naive reading of raw volume counts without per-capita adjustment would overstate the

role of ideology in driving conspiracy discussion.

5.3 RQ3: Ideological Diversity of Conspiracy Topics

Question: How ideologically diverse is each conspiracy topic's audience?

Where RQ1 asks whether any topic deviates significantly from the baseline composition, RQ3 looks at the *shape* of each topic's audience directly: how concentrated or spread out is the ideological mix? The natural summary is the normalised Shannon entropy defined in the statistical-methods block, computed over the binary left/right audience distribution of each topic, bounded in $[0, 1]$, with 0 meaning every user in the audience shares the dominant lean and 1 meaning the audience is a perfectly even left/right split.

A brief note on scope relative to RQ4: the entropy in this section is an *audience-level* quantity, one number per topic, computed over the two-class (left, right) distribution of that topic's posters. The entropy used later in RQ4 is a *user-level* quantity, one number per user, computed over that user's six-tier ideology probability vector (strong/lean/weak $\times$ left/right) to summarise how politically omnivorous the user is. The two share the name "entropy" but act on completely different distributions and serve different roles, they are not double-counting the same signal.

Result: Every main topic has high audience entropy (range 0.61–0.93, mean ≈ 0.86), meaning no topic is dominated by one lean to the exclusion of the other. The lowest-entropy topic (*Bill Gates, Lab-Made Viruses, and Elon Musk*, $H = 0.61$, 85% right) is still far from a pure one-sided audience, the highest-entropy topics (*Political Assassinations* and *Trump / Spiritual Warfare*, both $H \approx 0.93$) are essentially indistinguishable from the population split. Figure 5.8 shows the per-topic values. The colour gradient is the entropy itself (darker blue = higher entropy = more balanced audience), and the italic annotation on each bar reports the dominant side's share as a descriptive cue, not a classification label.

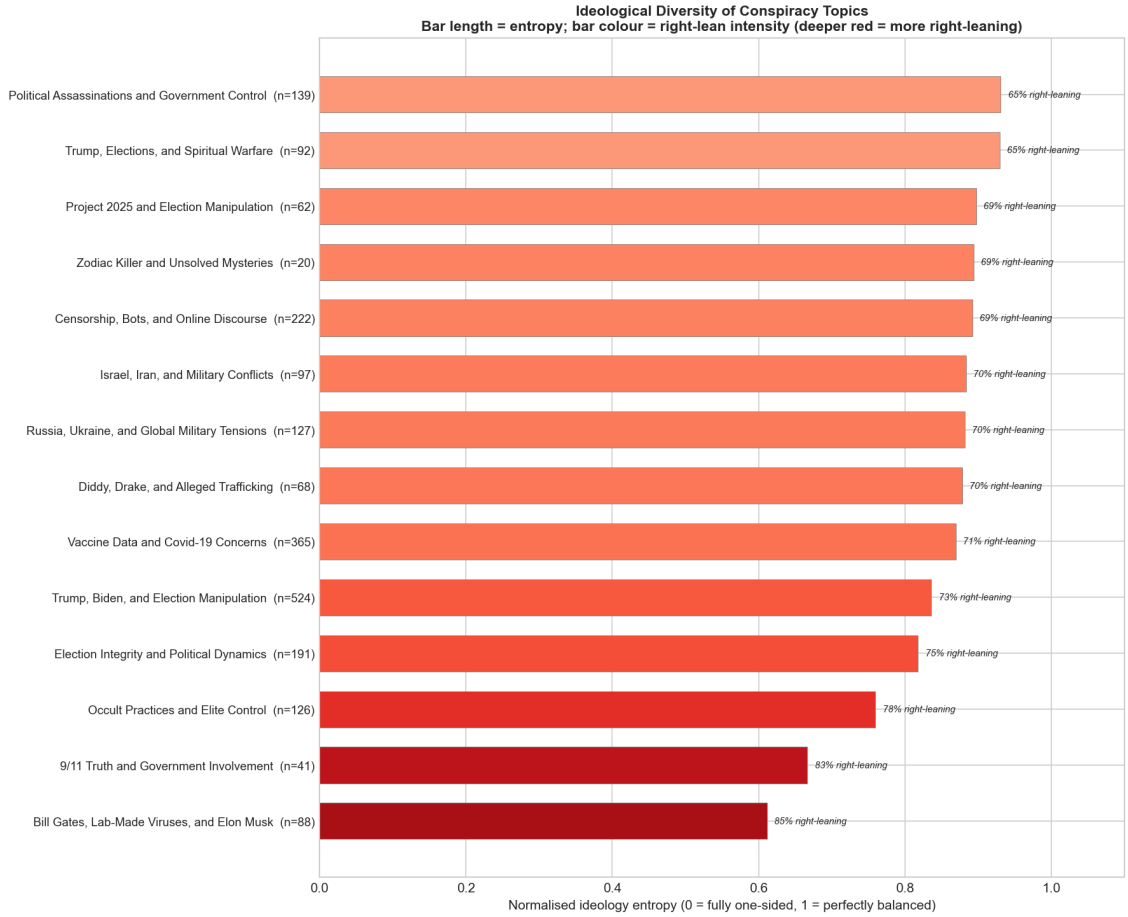


Figure 5.8: Normalised audience-level Shannon entropy per topic (0 = fully one-sided, 1 = perfectly balanced). **Bar length** encodes entropy and **bar colour** encodes right-lean intensity, with deeper red indicating a higher right-lean percentage, consistent with the thesis's red/right convention. The italic annotation on each bar shows the dominant side's percentage share.

Discussion: The entropy values should be read strictly as descriptive summaries of each topic's audience shape, not as an inferential classification. Every topic in the corpus has a meaningfully mixed audience, which is consistent with the RQ1 null. If no topic deviates significantly from the population baseline after BH correction, then no topic's audience can collapse toward a single lean, and the floor on entropy is correspondingly high. The entropy values also matter for RQ4, but through a different construct. A topic with high audience entropy is, by construction, hard to predict from ideology *direction* alone, which is part of why political *user-level diversity* (computed over each user's six-tier vector, not the audience two-class vector used here) turns out to outperform political direction as a predictor in the logistic regression.

5.4 RQ4: Can Ideology Predict Conspiracy Preferences?

Question: Can user ideology predict which conspiracy topics they engage with?

RQ4 moves from describing the audience (RQ1-RQ3) to asking whether a user's ideology *predicts* which topics they post in. Two kinds of models are fitted per topic:

- an **odds-ratio** test comparing right-leaning vs. left-leaning engagement probability (Figure 5.9), which is a direction-only predictor, and
- a **logistic regression** that jointly includes ideology direction (left/right probability), political diversity (the user-level six-tier entropy), and a dominance score, so that direction and diversity can compete as predictors for the same outcome (Figure 5.10 and Table 5.2).

Both are reported per topic, and both are BH-corrected across the topic family.

Result: 0/14 main topics and 0/20 subtopics reach significance after BH correction. At the main level the odds ratios range from 0.78 (Russia/Ukraine, left-skewed) to 1.46 (Diddy/Drake, right-skewed). At the subtopic level the range widens to 0.70–1.46 but still contains OR $= 1$ (no difference) within every confidence interval. All ratios are close to parity.

A note on the figure: the bars in Figure 5.9 plot $\log_2(\text{OR})$, not the raw odds ratio, because $\log_2$ places the "no difference" point at 0 and makes opposite effects symmetric (a topic with OR $= 2$ and one with OR $= 0.5$ are drawn as equal-magnitude bars on opposite sides of zero). Mapping back: $\log_2(0.78) \approx -0.36$ and $\log_2(1.46) \approx +0.55$, which is why the visible bar lengths span roughly -0.4 to $+0.55$ even though the raw OR range is 0.78 to 1.46. The bar text labels on the left-hand side report the raw OR value and its 95% CI for each topic, so both scales are readable directly from the figure.

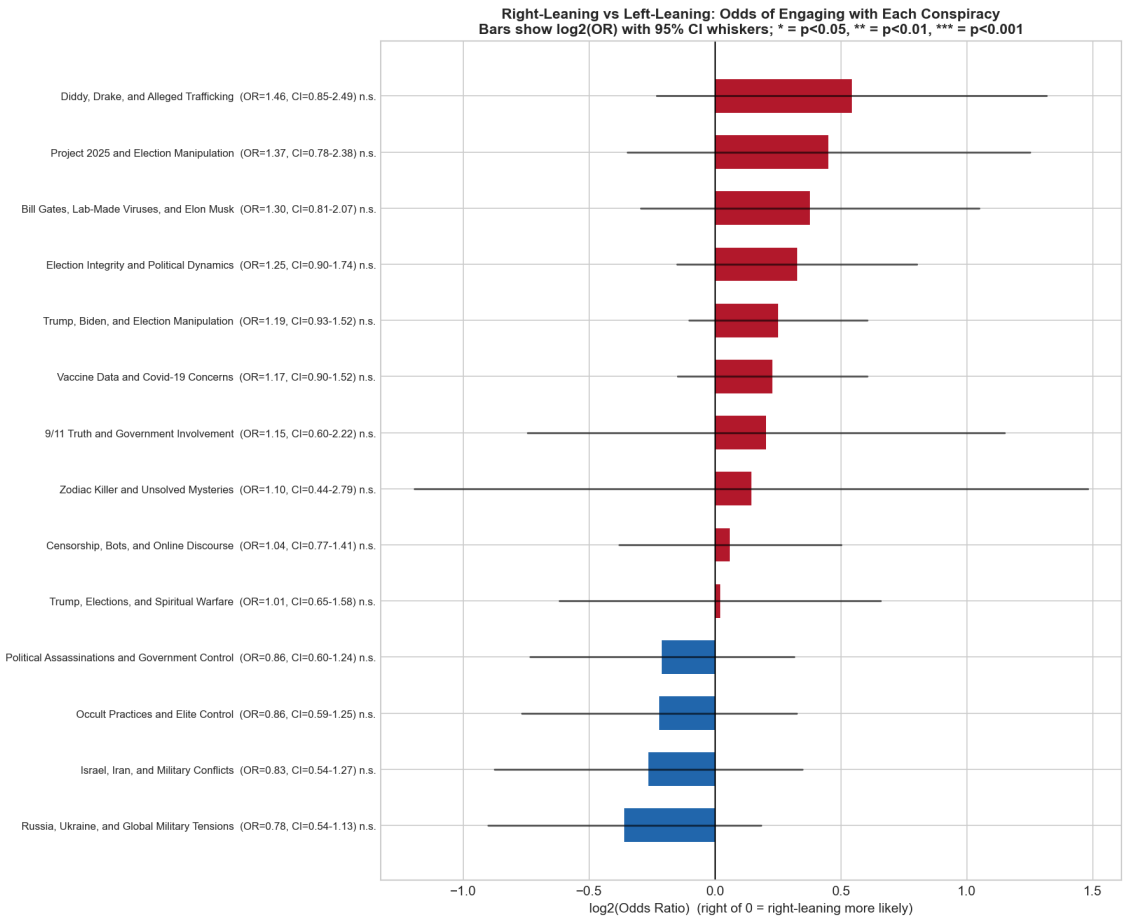


Figure 5.9: Right-leaning-vs-left-leaning odds of engaging with each main topic. Bars are drawn on a $\log_2(\text{OR})$ axis for left/right symmetry, where the 0 line corresponds to $\text{OR} = 1$ (no ideological preference), red bars indicate $\text{OR} > 1$ (right-leaning over-index) and blue bars indicate $\text{OR} < 1$ (left-leaning over-index). Each bar's text label shows the raw OR and 95% confidence interval. All confidence intervals cross 1 and no topic survives BH correction (all marked n.s.).

The more substantive finding sits in the logistic regression. When direction, diversity and dominance all compete for the same binary outcome ("did this user engage with topic T?"), the coefficient with the largest absolute size, i.e. the one that moves the log-odds the most, is not the same across topics. Table 5.2 counts, per topic, which predictor wins that competition.

Table 5.2: Number of topics where each predictor has the largest |coefficient| in the per-topic logistic regression. "Entropy" is the user-level six-tier diversity entropy, "Left / Right probability" are the six-tier probabilities summed within side, "Dominance" is 1 minus that entropy.

Predictor	Main Topics	Subtopics
Entropy (user-level diversity)	**6/14**	**7/20**
Left probability	3/14	5/20
Right probability	3/14	3/20
Dominance	1/14	1/20

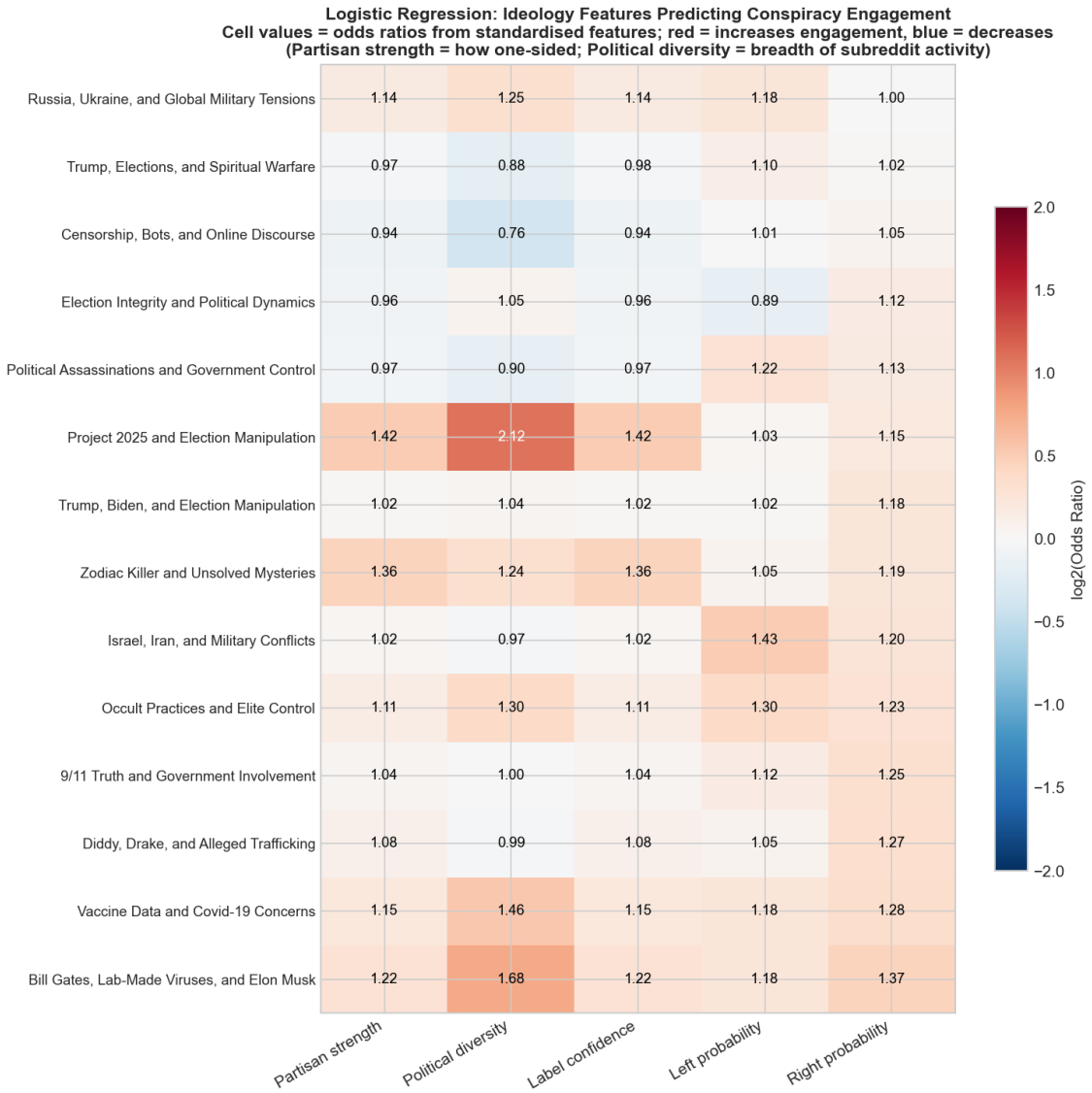


Figure 5.10: Logistic-regression coefficient heatmap: each row is a topic, each column is a predictor (direction vs. user-level political diversity), cell colour is the standardised coefficient. Diversity wins the most rows.

Key takeaway: under BH correction, *no* per-topic predictor is individually significant, so the reported diversity-over-direction pattern is a ranking of effect-sizes within a

topic-level null, not a claim of statistical significance. That caveat aside, the ranking itself is consistent: in the majority of topics, the variable that moves the log-odds of engagement most is *how broadly* a user participates across the six ideology tiers, not *which side* they sit on.

Figure 5.11 complements the odds-ratio view by plotting the raw engagement rate per topic, i.e. the percentage of left-leaning users who post in that topic versus the percentage of right-leaning users who post in it, ignoring diversity entirely. The blue and red bars sit very close to each other in almost every topic, which is the same null picture as Figure 5.9 on a different scale: once you strip out baseline population shares, the per-user probability of engaging with a given topic is roughly the same on both sides.

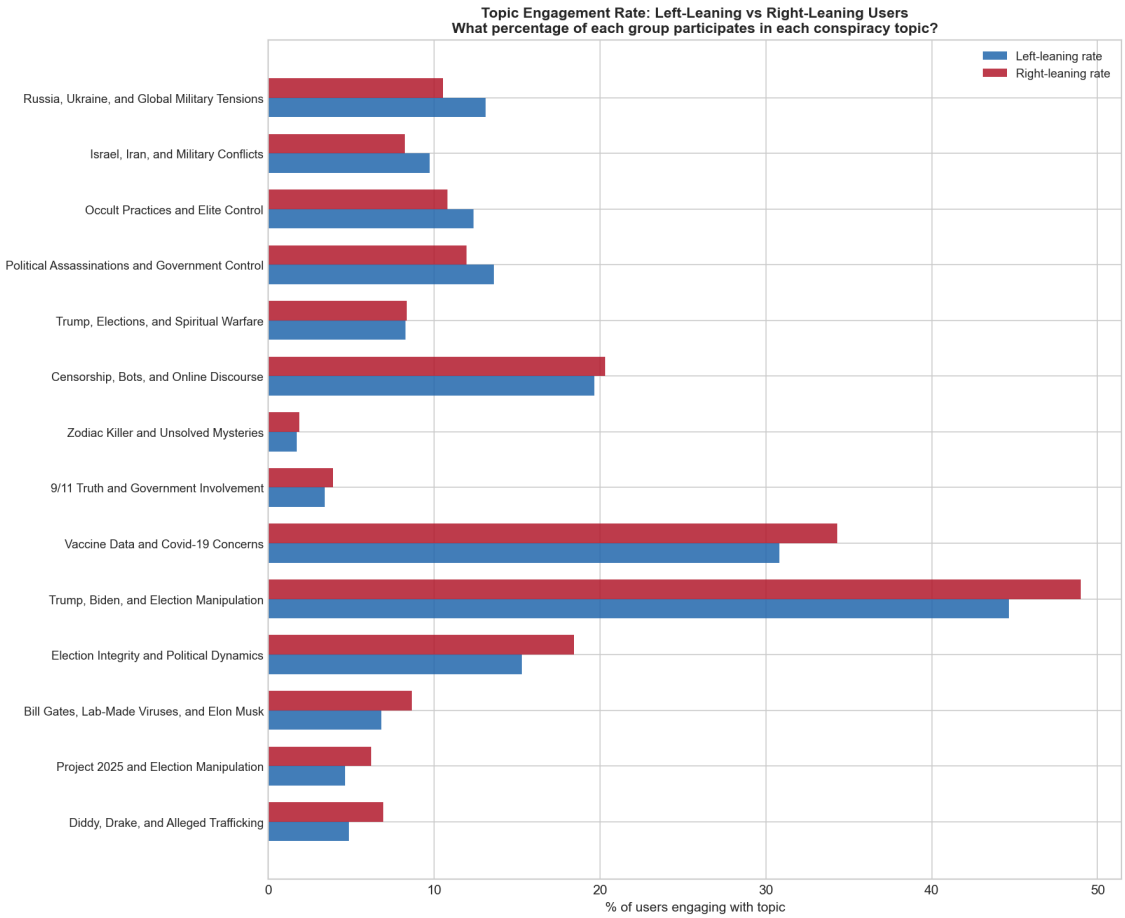


Figure 5.11: Topic engagement rate: proportion of each side's users who post in each topic. Rates track closely across every topic, echoing the odds-ratio null in Figure 5.9.

Discussion: Three things are worth stating plainly.

First, **significance**. Neither the odds-ratio family nor the logistic-regression family contains a single topic that survives BH at $\alpha = 0.05$. The raw odds ratios are close to parity and all 95% CIs cross 1, so reading any individual topic as "right-coded" or "left-coded" is unsupported by the data.

Second, **the direction-vs-diversity ranking**. Inside the null, the relative ordering of predictors is still informative as a description of the point estimates: in 6/14 main topics (and 7/20 subtopics), the coefficient with the largest magnitude is user-level political-diversity entropy, more than either left or right probability on their own. Read as an effect-size ranking rather than a significance claim, this says that *how omnivorous* a user's political footprint is, rather than *which side* it leans toward, is the stronger candidate for a predictor of conspiracy engagement. Diversity here is strictly user-level and six-tiered, it is not the same number as the two-class audience entropy in RQ3.

Third, **why diversity might out-predict direction**. Two readings are compatible with the data. A compositional one: politically omnivorous users simply spend more aggregate time in political subreddits, so mechanically they are more likely to turn up in conspiracy subreddits as well, a selection effect rather than a belief-driven one. A content one: conspiracy narratives appeal disproportionately to users whose political identity is not cleanly tribal, consistent with the long-standing framing of conspiracy thinking as a cross-cutting style rather than a partisan one. The cross-sectional design cannot tell these apart, but both readings share the practical implication that "which side" is not the right axis to sort conspiracy engagement along. If a researcher or a platform wanted to identify users at higher risk of deep conspiracy engagement, the user's political diversity score would be a more informative feature than their left-right lean, which is a non-obvious result that cuts against the common assumption that conspiracy thinking is a partisan phenomenon.

5.5 RQ5: User Bridging Scores and Ideology

Question: Do composite bridging scores differ by ideology?

Result: All four composite scores show negligible differences between left and right-leaning users (Table 5.3). Bridging capacity does not depend on ideology (Figure 5.12).

Table 5.3: Composite bridging scores by ideology ($N = 6{,}299$). All differences are negligible.

Score	Left Mean	Right Mean	Cohen's d	Sig.
Weighted	0.537	0.537	0.006	n.s.
Geometric	0.488	0.489	−0.008	n.s.
Harmonic	0.465	0.469	−0.037	n.s.
Ensemble	0.491	0.492	−0.020	n.s.

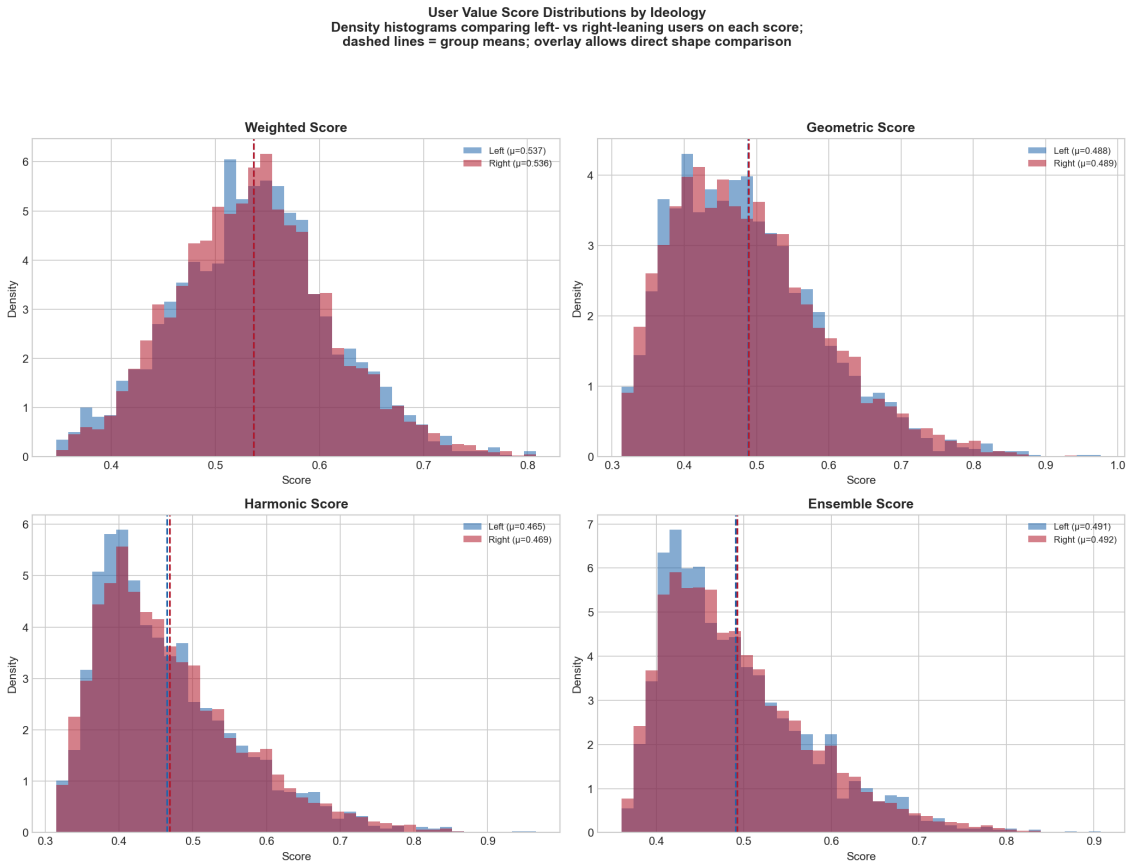


Figure 5.12: Composite bridging score distributions by ideology. The distributions are nearly identical.

Figure 5.13 breaks down the three component sub-scores (political activity, conspiracy activity, and balance) by ideology. The raw means sit within a few thousandths of a point of each other, but because the sample is large ($N = 6{,}299$) the BH-adjusted tests do flag two of the three components as significant, those being conspiracy activity and balance. Political activity is indistinguishable across sides ($d = +0.054$, $p_{\text{adj}} = 0.08$, n.s.), conspiracy activity is marginally higher for right-leaning users ($d = -0.069$, $p_{\text{adj}} = 0.006$, significant), and balance is also slightly higher for right-leaning users ($d = -0.102$, $p_{\text{adj}} < 0.001$, significant). These are the textbook case from the methods block where a large N makes a truly tiny gap detectable, all three effect sizes are comfortably inside the negligible band ($|d| < 0.2$), so the practical reading is still that ideology does not shape how users split their time between political and conspiracy domains, the direction (slight right-side edge on conspiracy intensity and balance) is consistent with the mild right-ward tilt already seen in RQ2.

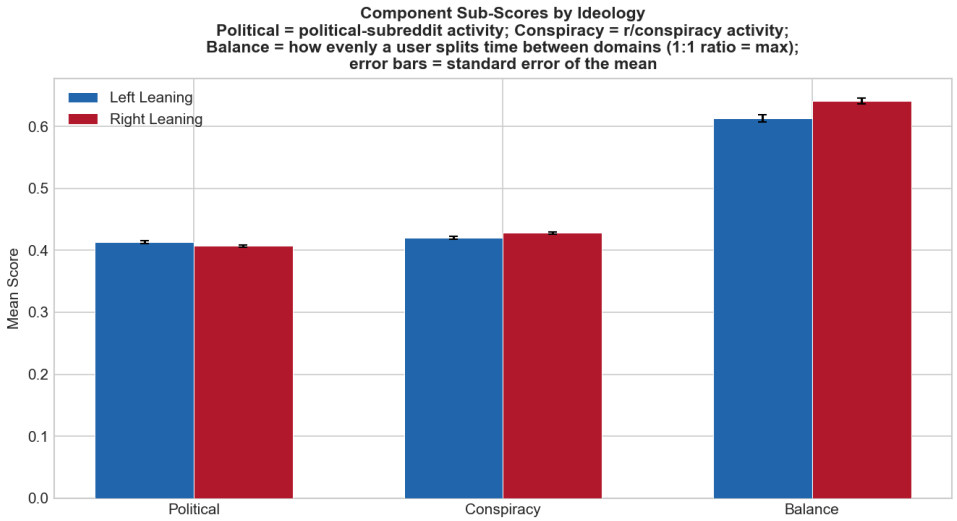


Figure 5.13: Component sub-scores (political, conspiracy, balance) by ideology. All three components are nearly equal.

Figure 5.14 shows the correlation matrix among all scoring components. The three composite scores (weighted, geometric, harmonic) are highly correlated with each other ($r > 0.9$), which validates that the ensemble average is a stable summary and that the choice of aggregation function does not change the ranking of users in any meaningful way. The more telling number is the moderate correlation between political and conspiracy sub-scores ($r \approx 0.5$), which says that users who are active in one domain tend to be active in the other as well. This cross-domain co-activity is not obvious in advance, since one could reasonably expect conspiracy engagement to be independent of political engagement, or even inversely related if conspiracy communities attract users who have disengaged from mainstream politics. Instead, the positive correlation suggests that conspiracy participation on Reddit is part of a broader pattern of high platform activity rather than a niche that draws in otherwise-inactive accounts.

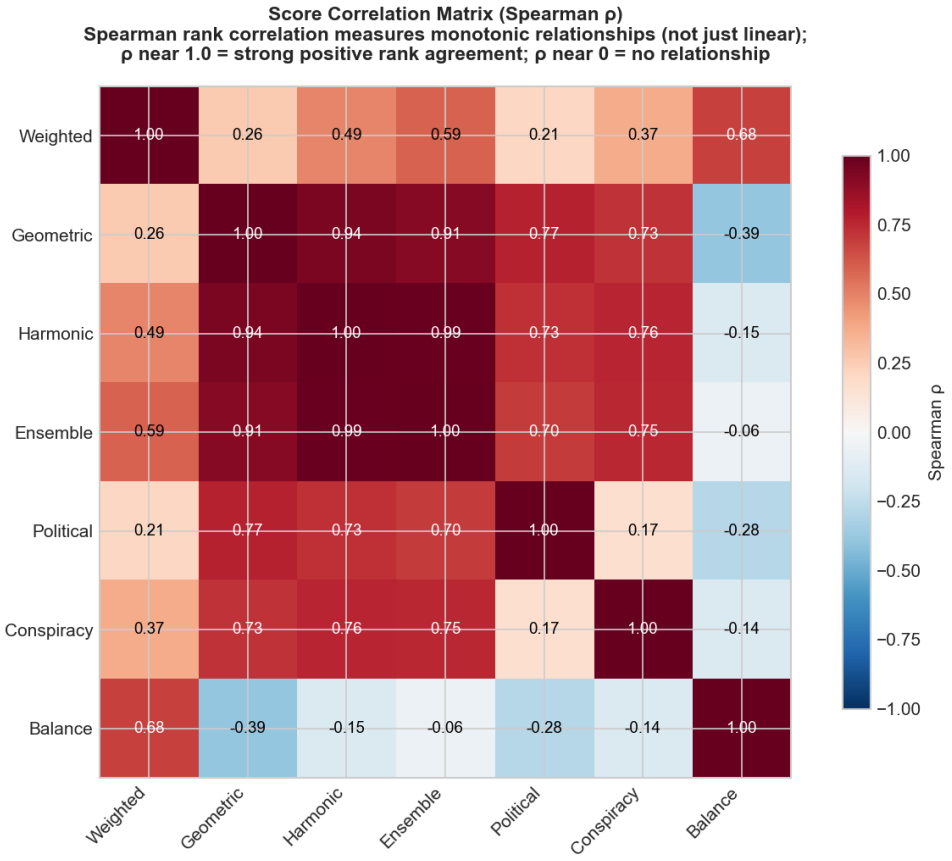


Figure 5.14: Correlation matrix of scoring components. The three composite scores are highly intercorrelated, political and conspiracy sub-scores show moderate positive correlation.

Figure 5.15 plots ensemble score against total conspiracy engagement (post count). Higher-scoring users are, by construction, more active, but the spread at each score level shows considerable variation. This spread matters because it confirms that the ensemble score is not simply a proxy for volume. Two users can have the same total post count but very different ensemble scores if one is heavily lopsided toward conspiracy while the other balances both domains, so the score captures the *shape* of a user's cross-domain activity, not just the amount.

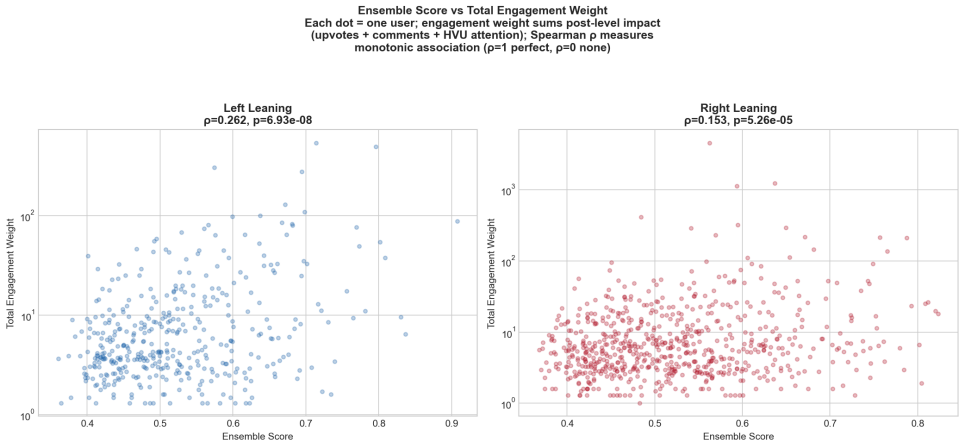


Figure 5.15: Ensemble score vs. total conspiracy engagement. The positive trend confirms the score captures activity level, but the spread at each level shows it also reflects balance.

That said, the correlations between scores and engagement are roughly $2\times$ stronger for left-leaning users ($\rho = 0.262$ vs. 0.153 for score-engagement, $\rho = 0.315$ vs. 0.152 for score-breadth). When left-leaning users do bridge both political and conspiracy domains, they tend to do so with more intensity and across more topics. Figure 5.16 visualises this relationship, as higher-scoring users engage with more topics, but the slope is steeper for left-leaning users.

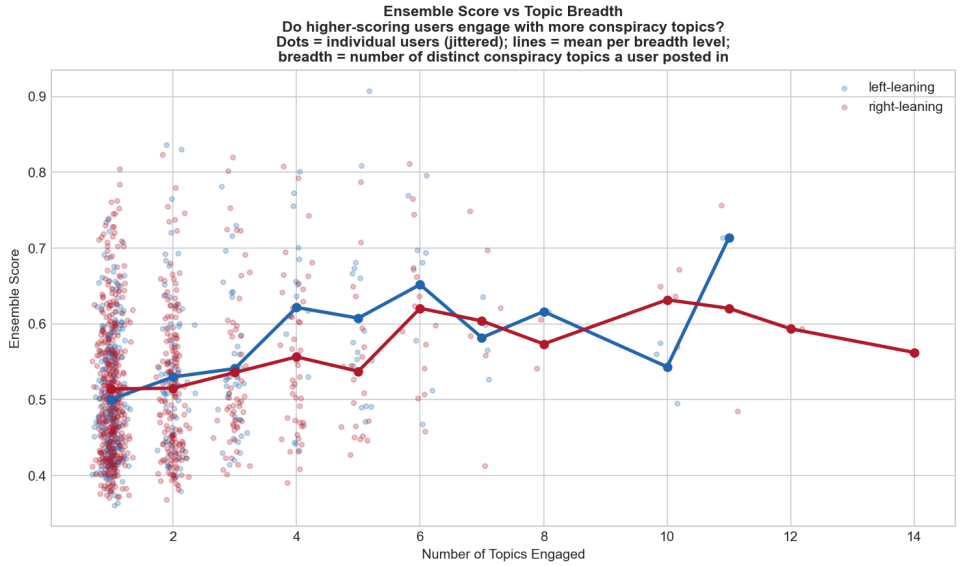


Figure 5.16: Ensemble score vs. topic breadth by ideology. Left-leaning users (blue) show a stronger score-breadth correlation.

The top 10% of HVU show no significant ideological skew (chi-squared $= 0.68$, $p = 0.41$). Figure 5.17 shows that the ideology split in the top decile is almost identical to the overall population. This is the strongest version of the bridging-score null, because it tests the users with the most cross-domain activity, the ones most likely to show an

ideological skew if one existed, and still finds none. If high-intensity conspiracy bridging were a predominantly right-leaning or left-leaning behaviour, it would show up here, and it does not.

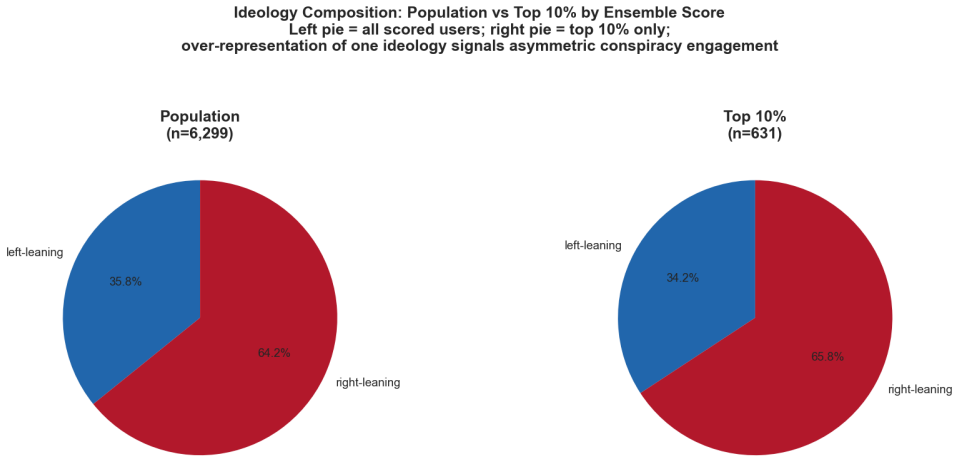


Figure 5.17: Ideology composition: full population (left) vs. top 10% by ensemble score (right). The distributions are nearly identical.

Discussion: The bridging-score null is noteworthy because it replicates the topic-level null at the aggregate user level. Neither the content people post about nor the structural role they play (as measured by the ensemble score) separates cleanly by lean. The interesting asymmetry lies in the slopes, where the score-engagement and score-breadth correlations are roughly twice as strong for left-leaning users as for right-leaning users ($\rho \approx 0.26$ vs. 0.15, $\rho \approx 0.32$ vs. 0.15). This says that among users who do bridge politics and conspiracy, left-leaning bridgers tend to be more intense and more topically diverse when they do it, even though the average bridging score is the same on both sides. The practical consequence is that raw ideology-level comparisons can hide meaningful within-group structure, a median comparison would have reported "no effect" and missed this shape entirely. The top-decile chi-squared finally confirms that the most active bridging accounts come from both sides in the population ratio, which rules out the simplest alternative explanation, that bridging is dominated by one political camp.

5.6 RQ6: Network Ideology Mixing and Null Model

Question: Do r/conspiracy users interact with ideologically similar or different peers, and is any observed pattern attributable to the graph structure rather than genuine ideological sorting?

RQ1-RQ5 ask whether ideology predicts which topics users engage with or how much they engage. RQ6 shifts to the network level and asks whether ideology shapes *who*

interacts with whom: specifically, whether the r/conspiracy co-activity network exhibits ideological homophily (users preferentially interacting with ideologically similar peers), and whether that pattern goes beyond what the base-rate ideological composition of the network would produce by chance.

Methodology: Each user's ideology participation vector $\mathbf{P}_u = (p_{u,L}, p_{u,C}, p_{u,R})$ (left, centre, right activity proportions, normalised to sum to 1) was loaded from the ideology-vector file. For each edge (u, v) in the user-user co-activity network (4,368 users, 193,876 edges), the *cosine distance* $d(u,v) = 1 - \cos(\mathbf{P}_u, \mathbf{P}_v)$ was computed as the ideological distance between the two endpoints: $d = 0$ means identical ideology profiles, $d = 1$ means maximally opposed. The *neighbour diversity* $\mathrm{ND}(u)$ of each user is the mean cosine distance to all their direct neighbours.

To establish a null baseline, ideology vectors were randomly permuted across users 500 times while the graph structure was held fixed, and the mean edge distance and mean $\mathrm{ND}(u)$ were recomputed in each permutation. The observed value is compared to this null distribution to obtain an empirical percentile and two-sided p-value.

Result: The network shows statistically significant ideological homophily.

- **Observed mean edge distance:** 0.4425, compared to a null mean of 0.4685 ± 0.0070. The observed value falls at the 0th percentile of the 500-shuffle null distribution ($p < 0.001$), meaning no shuffled network produced a mean edge distance as low as the one observed. Users in r/conspiracy interact with ideologically more similar peers than would be expected by chance alone.

- **Observed mean neighbour diversity** $\mathrm{ND}(u)$**:** 0.4362, compared to a null mean of 0.4686 ± 0.0091 (observed at the 0th percentile, $p < 0.001$). The average r/conspiracy user has less ideologically diverse immediate neighbours than chance would predict.

Figures 5.18 and 5.19 show the null distributions alongside the observed values.

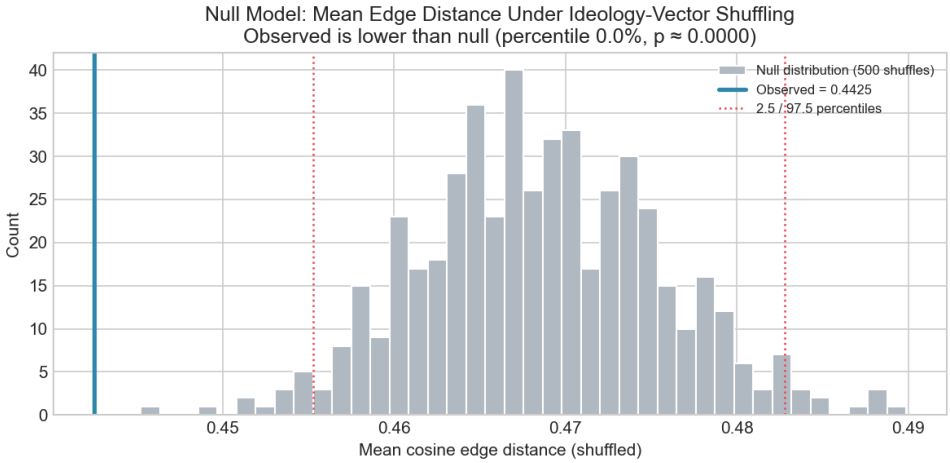


Figure 5.18: Null-model distribution of mean edge cosine distance across 500 ideology-vector shuffles (grey histogram). The blue vertical line marks the observed value (0.4425), which falls well below the null distribution, confirming homophily.

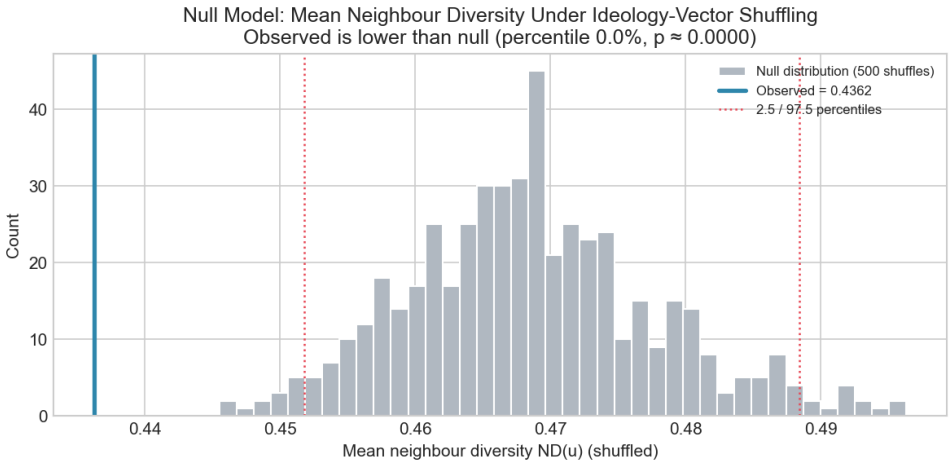


Figure 5.19: Null-model distribution of mean neighbour diversity $\mathrm{ND}(u)$ across 500 shuffles. The observed mean (0.4362) is again below the null, consistent with ideological homophily.

Ideology mix and neighbour diversity: The Spearman correlation between a user's normalised entropy $H(u)$ (how ideologically mixed their own political activity is) and their neighbour diversity $\mathrm{ND}(u)$ is $\rho = -0.369$ ($p < 0.001$). The *negative* sign is counter-intuitive at first but has a geometric explanation, being that a maximally mixed user has a participation vector close to $(0.33, 0.33, 0.33)$. Two such vectors are geometrically close in cosine space (small cosine distance), so mixed users who interact with other mixed users produce *low* ND values. The negative correlation therefore reflects the geometry of the ideology vector space rather than a behavioural tendency of mixed users to seek out ideologically similar peers. Figure 5.20 shows the scatterplot.

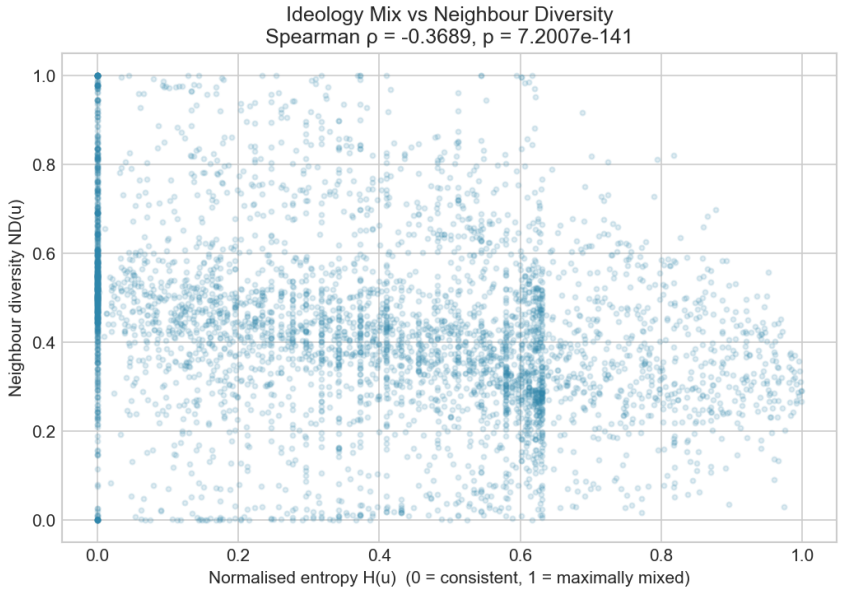


Figure 5.20: Spearman correlation between user ideology entropy $H(u)$ and neighbour diversity $\mathrm{ND}(u)$ ($\rho = -0.369$, $p < 0.001$). The negative slope reflects the geometry of the ideology vector space rather than an absence of mixing among consistent-ideology users.

Structural roles of ideologically mixed users: Table 5.4 compares three structural-position metrics (weighted degree, approximate betweenness centrality [18], and PageRank [36]) between the top quartile of $H(u)$ (most mixed, $n = 1{,}118$) and the bottom quartile (most consistent, $n = 1{,}095$). None of the three comparisons reaches significance (all $p > 0.05$, all $|d| < 0.05$, negligible effect sizes). Figure 5.21 shows the boxplots.

Table 5.4: Structural position of ideologically mixed (high-H) vs consistent (low-H) users. None of the metrics differs significantly.

Metric	**High-H mean**	**Low-H mean**	**Cohen's d**	p	**Sig.**
Weighted degree	720.6	671.9	0.026	0.119	n.s.
Betweenness centrality	0.0004	0.0003	0.044	0.193	n.s.
PageRank	0.0002	0.0002	0.031	0.058	n.s.

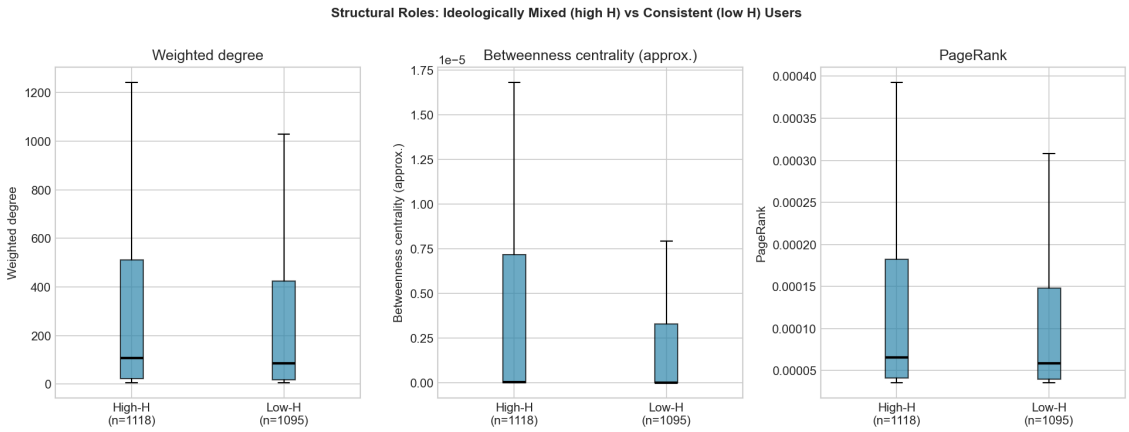


Figure 5.21: Structural position metrics for high-entropy (most ideologically mixed) vs low-entropy (most ideologically consistent) users. Distributions are comparable across all three metrics.

Discussion: The homophily result is the most unambiguous finding in this chapter. The network is more ideologically assortative than chance, and the effect is large enough that zero of 500 random permutations produced a mean edge distance as low as the one observed. Conspiracy subreddit interaction structures are not ideologically random, users preferentially interact with others who share their political footprint.

At first glance this seems to contradict the topic-level nulls of RQ1-RQ5, since ideology does not predict *what* people talk about but does predict *who* they talk with. The apparent paradox dissolves once the two findings are recognised as operating on different layers of behaviour. Content choice (which conspiracy topic a user engages with) and interaction-partner selection (which other users they end up co-commenting alongside) are separate decisions governed by separate mechanisms. Think of each conspiracy topic as a public square that anyone can enter regardless of political background, and the RQ1-RQ5 results confirm that entry is indeed open to all sides in roughly equal proportions. But within that square, people naturally gravitate toward conversations with others who share their political profile, and the RQ6 null model confirms that this gravitational pull is stronger than chance alone would produce. The content is shared, the social context is not.

This dual-layer pattern is not unique to conspiracy communities. The broader homophily literature documents that assortative ties routinely form not because individuals actively screen for similar partners but because shared social contexts, such as common workplaces, neighbourhoods, or voluntary associations, place similar people in physical or digital proximity and thereby increase the likelihood that they interact [32]. On Reddit, the analogous "social context" is the set of political subreddits a user frequents. Three platform-specific mechanisms can translate that shared context into the network-level homophily we observe, without requiring any user to consciously seek out ideologically

similar peers.

The first is *temporal co-arrival*. Users who frequent the same political subreddits are exposed to the same external news triggers at roughly the same time, because political communities amplify breaking stories within hours. When a news event (a court ruling, a vaccine mandate, a leaked document) sends a wave of users from left-leaning political subreddits into r/conspiracy, those users land in the same comment threads not because they sought each other out but because they arrived during the same narrow window. A right-leaning wave triggered by a different framing of the same event may arrive hours later or cluster in a different set of threads. The co-activity network records these temporal coincidences as edges, and because the arrival waves are ideologically correlated, the resulting edges are ideologically assortative even though no user chose a discussion partner based on ideology. This is the digital equivalent of the "shared foci" mechanism that McPherson et al. [32] identify in offline social networks, where people who frequent the same physical spaces end up with similar ties not by choice but by exposure.

The second is *thread-level framing selection*. Selective exposure theory predicts that individuals preferentially seek out information consistent with their prior beliefs [49], and on Reddit this preference operates at the thread level rather than at the topic level. A single conspiracy topic (e.g. "Vaccines") can spawn dozens of threads within a subreddit, and those threads often differ sharply in how they frame the narrative, with one thread presenting vaccine scepticism through a libertarian, anti-government lens while another frames it through a progressive, anti-corporate lens. Both threads fall under the same BERTopic cluster, so the topic-level analysis of RQ1 correctly sees no ideological skew in the Vaccines topic as a whole. But at the thread level, users self-select into the framing that resonates with their existing worldview, and this thread-level sorting is exactly what the co-activity network captures, because two users who comment in the same thread share an edge while two users who comment in different threads of the same topic do not. The topic is ideology-neutral, but the conversational entry points within it are not.

The third is *reply-chain dynamics*. Reddit's nested comment structure means that a user who replies to a comment is far more likely to receive further replies from users who also engaged with that same parent comment. If the parent comment expresses a view that attracts engagement from one ideological camp, the reply chain beneath it will be disproportionately populated by users from that camp, whether they are agreeing or pushing back. Over thousands of such micro-interactions, the network accumulates a homophily signal that emerges from the architecture of threaded conversation rather than from any user-level decision to filter by ideology. Cinelli et al. [12] document a closely related platform-architecture effect across Facebook, Twitter, Reddit, and Gab, showing that each platform's interaction design shapes the degree of echo-chamber formation independently of the content being discussed, and that Reddit's structure produces measurable

ideological segregation in user interaction patterns even in the absence of algorithmic curation.

These three mechanisms are not mutually exclusive and likely compound each other. A temporally correlated arrival wave funnels ideologically similar users into the same threads, selective exposure to compatible framings keeps them in those threads rather than in adjacent ones, and reply-chain dynamics concentrate their interactions into shared edges. The net effect is a network that looks ideologically sorted even though no individual user had to make an ideologically motivated choice about which conspiracy narrative to engage with. This is why the topic-level nulls of RQ1-RQ5 and the network-level homophily of RQ6 are not in tension. They measure different stages of the same process, and the ideological sorting enters at the interaction stage, not at the content-selection stage.

Separately, ideologically mixed users hold no structural advantage in the network. They are not significantly more central, more influential (by PageRank), or higher-degree than ideologically consistent users. This rules out the hypothesis that cross-cutting users act as structural bridges that moderate the ideological sorting described above.

The implications of this distinction for platform design and content moderation are discussed in the practical implications paragraph at the end of this chapter.

5.7 RQ7: Subreddit Ecosystem Comparison

Question: Does the ideological composition differ across conspiracy subreddit ecosystems?

RQ1-RQ5 hold topic and user variables constant and let subreddit identity drop out of the picture. RQ7 reverses that framing and asks whether the *subreddit* itself explains any of the ideological variation in the corpus. Four quantities are compared across the four conspiracy subreddits, those being volume, ideological composition (both the hard four-class breakdown and the finer eight-class tiers), HVU ensemble scores, and topic share. Each is reported separately because they answer distinct questions. Volume tells us how much a community contributes to the corpus, composition tells us who is in the community, HVU scores tell us how much high-intensity bridging activity each community hosts, and topic share tells us whether communities with different audiences also pick different narratives.

Headline result: A chi-squared test of ideology $\times$ subreddit gives $\chi^2 = 8.58$, $df = 3$, $p = 0.035$, the **only BH-surviving significance test in the entire study**. Table 5.5 and Figure 5.23 show the underlying breakdown, and the rest of this section unpacks the shape of that difference.

Figure 5.22 shows the raw post volume per conspiracy subreddit. r/conspiracy contributes 76.2% of the corpus on its own, r/conspiracy_commons a further 16.6%, and the

remaining two subreddits together account for just over 7%. Any subreddit-level effect has to be read against this volume asymmetry. It is easy for a small community to look extreme on a proportions chart when its absolute counts are in the hundreds rather than tens of thousands, and the chi-squared test above does use the raw counts, so the verdict is not driven by the small-subreddit proportions alone.

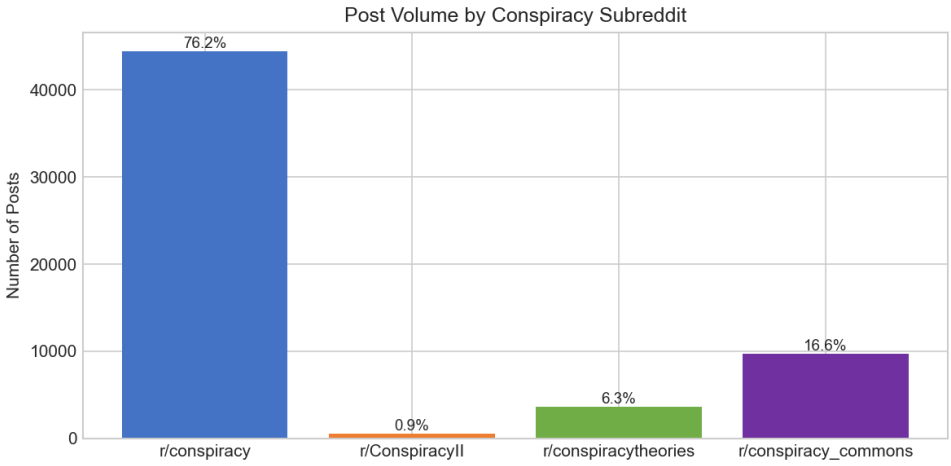


Figure 5.22: Post volume per conspiracy subreddit. r/conspiracy accounts for 76.2% of the corpus, r/conspiracy_commons for 16.6%, the other two for the remainder.

Table 5.5: Ideology composition per conspiracy subreddit (four-class aggregation, $\chi^2 = 8.58$, $df = 3$, $p = 0.035$ on the raw counts).

Subreddit	Users	Left %	Centre %	Right %	R-L Gap
r/conspiracy	6,398	22.2%	34.5%	41.1%	+18.9 pp
r/conspiracy_commons	913	27.1%	21.4%	48.8%	+21.8 pp
r/conspiracytheories	682	27.1%	31.7%	37.4%	+10.3 pp
r/ConspiracyII	110	27.3%	22.7%	44.5%	+17.3 pp
Rows sum to <100% because a small "balanced" residual ($\leq$ 5.5 pp per row) is omitted for clarity.					

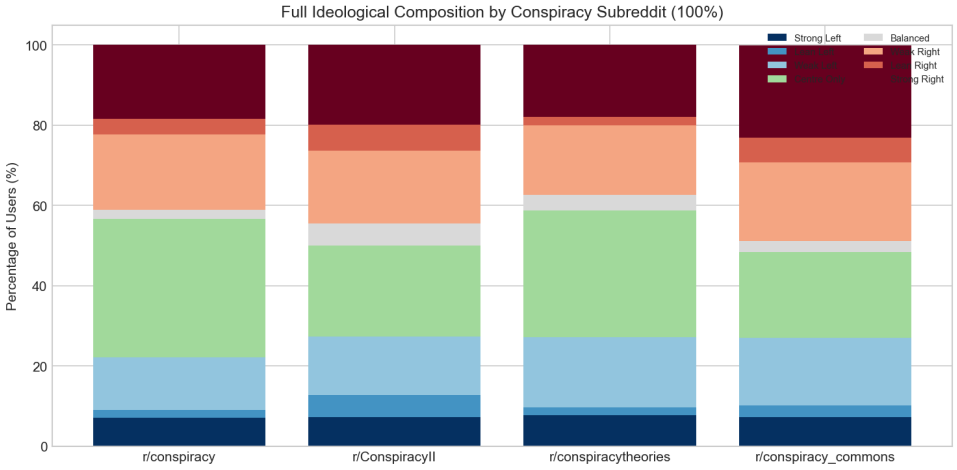


Figure 5.23: Ideological composition per conspiracy subreddit. r/conspiracy_commons is the most right-skewed and the furthest from baseline, r/conspiracytheories the least right-skewed.

The four-class view above is the one used for the significance test, but the eight-tier breakdown sharpens the picture. r/conspiracy_commons carries the highest *strong-right* share of the four (23.0%), the highest lean-right share (6.2%), and a centre-only share that is roughly 13 pp lower than r/conspiracy's (21.4% vs. 34.5%). r/conspiracytheories, which is both the smallest and closest-to-baseline subreddit, has the same 18.0% strong-right share as r/conspiracy but a noticeably higher weak-left share (17.4% vs. 13.2%), which is what pulls its R−L gap down to +10.3 pp. In short, the difference between subreddits is not uniformly "more right", it is a specific reshaping of the left/centre/right tiers, and the shape is consistent with subreddit-choice behaviour rather than with a difference in what topics are being discussed.

Figure 5.24 shows the distribution of HVU ensemble scores per subreddit. The medians are similar across all four communities, but the means diverge modestly. Specifically, r/conspiracy_commons carries the highest mean ensemble score (0.5232), r/conspiracytheories the lowest (0.4963), with r/conspiracy and r/ConspiracyII sitting in between (0.5066 and 0.5111, respectively). The same subreddit also over-indexes on the global top-decile of bridging users, 21.8% of its HVU fall in the global top 10% by ensemble score (vs. 10% expected), against 15.8% in r/conspiracy, 18.8% in r/ConspiracyII, and 11.1% in r/conspiracytheories. The most right-skewed community is therefore also the one that attracts the most intense cross-domain bridging users, which is a useful caveat to the "composition, not content" reading, where users engage matters for *how* intensely they engage, not just for which ideology they lean toward.

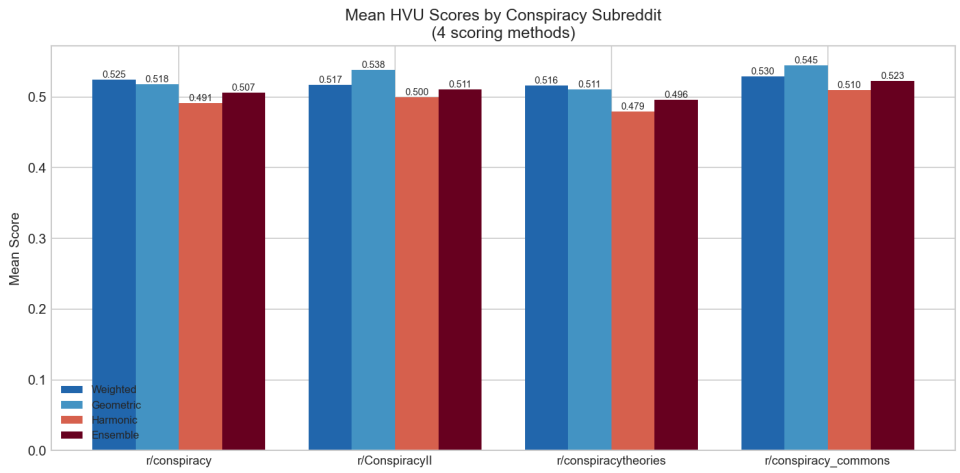


Figure 5.24: HVU ensemble score distributions per conspiracy subreddit. Medians are comparable across communities, but r/conspiracy_commons has the highest mean (0.523) and the largest share of users in the global top decile (21.8%).

Figure 5.25 turns to the content side. At an aggregate level, the topical profiles across the four subreddits are broadly similar, with Trump/Biden, Vaccines, and the outlier bucket dominating everywhere. The smaller subreddits each carry topic-level over-representations that are worth naming because they are exactly the kind of narrative-level signal that would be invisible at the aggregate view. r/ConspiracyII runs Diddy/Drake-trafficking content at 1.5× the corpus rate and Russia/Ukraine at 1.4×. r/conspiracytheories over-indexes on Project 2025 at 3.9× the corpus rate and Trump/spiritual-warfare at 2.1×. r/conspiracy_commons over-indexes on the Bill Gates/Lab-Virus/Musk cluster at 1.7× and on 9/11 Truth at 1.2×. None of these multipliers are large enough to flip the dominant-topic ordering inside any subreddit, but they do explain why subreddit-specific narrative niches persist inside an otherwise broadly similar topic mix.

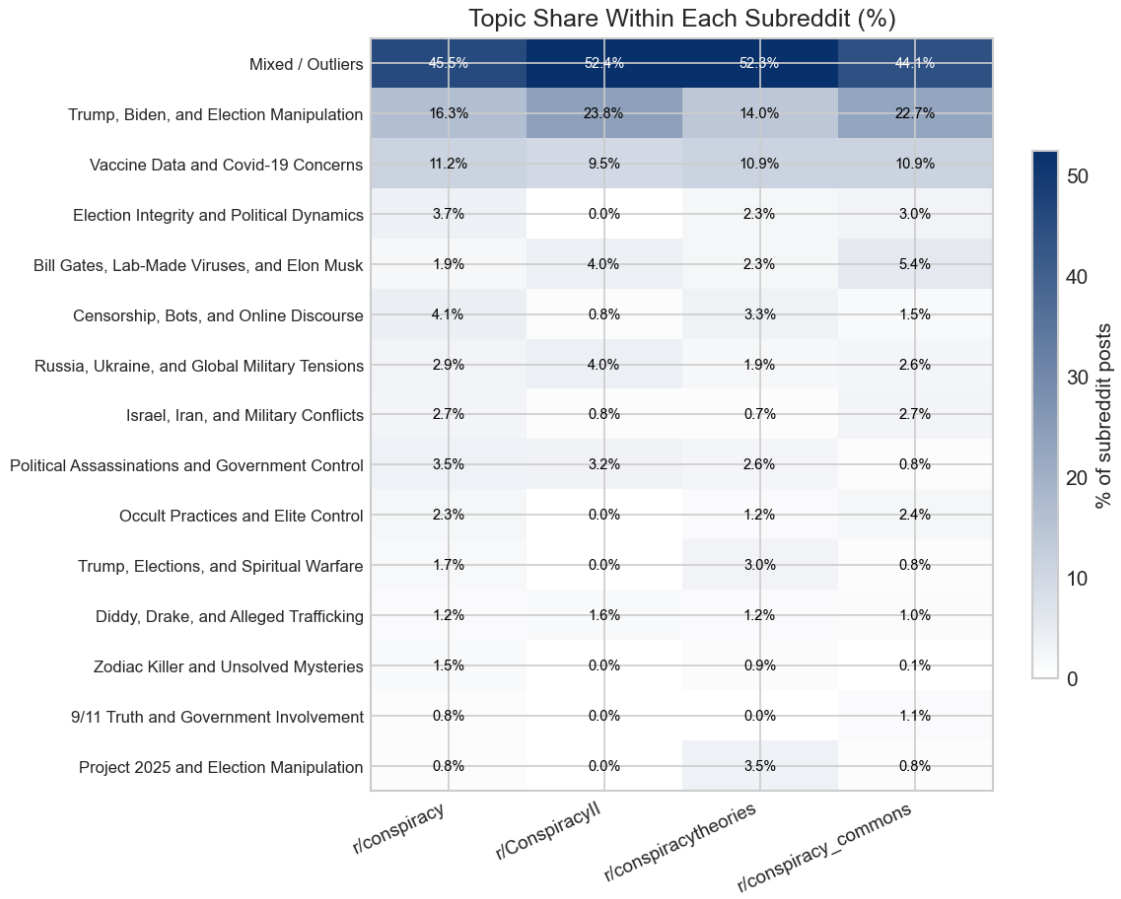


Figure 5.25: Topic share within each conspiracy subreddit. The aggregate profiles are similar across the four communities, but each smaller subreddit has one or two topics it over-indexes on (see text for the specific multipliers).

Discussion: RQ7 is the only place in the chapter where a BH-corrected significance test survives, and the shape of the finding is specific enough to be worth stressing rather than flattened into a headline claim. Three observations hold the finding together. First, the R−L gap ranges from +10.3 pp in r/conspiracytheories, the subreddit closest to the baseline split, to +21.8 pp in r/conspiracy_commons, the most right-skewed of the four, so the subreddit effect is not "every conspiracy subreddit looks the same", it is a gradient across four specific communities. Second, the topic-level picture does *not* line up with the ideological picture. If narrative content drove ideological sorting, one would expect r/conspiracy_commons to specialise in right-coded topics and r/conspiracytheories in left-coded ones, which is not what the topic heatmap shows, the over-representations that do exist (Project 2025 in r/conspiracytheories, Bill Gates/Musk in r/conspiracy_commons) do not track a clean partisan axis. Third, the HVU-score and top-decile patterns point to an intensity story *within* the community effect, meaning the most right-skewed subreddit also carries the highest concentration of global top-decile bridging users, which means its ideological skew coexists with higher-intensity cross-domain activity, not lower-quality

participation. The remaining explanation is community-level rather than content-level, moderation style, welcoming or filtering norms, and the subreddit-specific demographics of the people who show up and stay. This also clarifies why the earlier RQs returned nulls, since ideological sorting in this corpus is happening at the *subreddit-choice* layer, which the topic-level and user-level analyses of RQ1-RQ5 are not designed to detect. It is the one finding that survives correction, and the only one that should be carried forward into the conclusion as a positive claim.

5.8 RQ8: Classifier Evaluation

Question: How accurately can the pipeline classify a short-form Reddit post into one of the conspiracy topic categories?

The classification pipeline (Section 4.6) has two components, a *centroid similarity* method (unsupervised, distance-based, no labelled training data required) and a *fine-tuned RoBERTa head* (supervised, probability-based). Each is evaluated separately below with full disclosure of the data splits, augmentation pipeline, and hyperparameter choices so that the reported metrics can be assessed critically.

Data preparation and split

The labelled corpus used for the RoBERTa head is the set of HVU posts assigned to named topics by BERTopic (i.e. the non-noise fraction). The BERTopic assignment serves as the ground-truth label, which is an important caveat returned to in the Limitations subsection. For the final (subtopic) model, the 3 parent topics (Vaccines, Bill Gates/Musk, Russia/Ukraine) are replaced by their 9 subtopic labels, giving 20 real topic classes plus one off-topic class (21 total).

The full labelled set is split once, before any augmentation, using a **stratified 85 / 15 train/validation split** with a fixed random seed (42) for reproducibility. Stratification ensures that every class's proportions are preserved in both halves. The validation set is never touched during augmentation or training. It is used only for the final evaluation reported here.

Table 5.6: Pre-augmentation data sizes for the three training runs. "Original train" is the 85% split before any synthetic samples are added. "Augmented train" is the size after augmentation, which is the value stored in `classifier_metadata.json`. The val set is identical across runs within the same label scheme.

Model version	Classes	Orig. train	Aug. train	Val
Baseline (no augmentation)	15	11,506	11,506	2,031
Augmented, flat topics	15	11,506	24,600	2,031
Augmented, subtopics (final)	21	≈6,280	14,976	1,108

The subtopic dataset is smaller than the flat-topic dataset because the subtopic labelling is applied only within the 3 re-split parent topics, while posts from the other 11 unchanged topics carry their original flat-topic labels. The 3 parent posts are redistributed among 9 finer classes, so the total labelled pool for the subtopic run is smaller ($\approx$7,388 original samples: 6,280 train + 1,108 val). The per-class counts for some subtopics are consequently very thin, in particular EcoHealth with only $n = 65$ posts.

Data augmentation

Augmentation is applied to the training half only, while the val set is kept clean. Two operations are performed:

1. **Class upsampling.** Any class whose original count falls below the 75th-percentile class count is upsampled to that target by generating augmented copies until the target is reached. This addresses the severe imbalance between large topics (Trump/Election: $>$2,000 posts) and small ones (Zodiac, 9/11: $<$120 posts).

2. **One synthetic copy per original sample.** Regardless of class size, one augmented variant of every training sample is added to the training set. This doubles, or more than doubles, the training set for all classes.

Each augmented variant is produced by applying *one* of four text-perturbation strategies, chosen uniformly at random:

- **Random word deletion:** 10-20% of words are dropped.
- **Random adjacent word swap:** 2-4 adjacent word pairs are swapped.
- **Title-only:** the post body is dropped and only the title is kept, simulating the short-form input the extension will receive in practice.
- **Sentence shuffle:** sentences in the post body are reordered randomly while the title is preserved.

These are dependency-free operations that do not require synonym databases or language models. Their combined effect is to teach the model that topic identity should not depend on word order or the presence of every sentence, which is a useful inductive bias for short-form social-media text.

Training configuration and regularisation

All three model versions share the same backbone (`roberta-base`), the same maximum sequence length (512 tokens), and the same regularisation choices:

- **Label smoothing** [34] ($\varepsilon = 0.10$): the target distribution is smoothed away from hard 0/1 labels. This prevents the model from assigning near-zero probability to correct-but-uncertain classes and improves calibration.
- **Dropout** ($p = 0.20$): applied to the final classification head to reduce co-adaptation of features.
- **Early stopping** (patience = 3 evaluation rounds): training halts when the macro F1 on the validation set has not improved for 3 consecutive evaluations. The final subtopic model stopped at epoch 10 of the configured 20.
- **Temperature scaling** [24] (post-hoc calibration): after training, a scalar temperature T is fitted on the validation set logits to minimise calibration error. The optimal temperature for the final model is $T = 0.952$, slightly below 1, which slightly sharpens the softmax distribution. Calibrated probabilities are what the Chrome extension displays as the BERT confidence score.
- **Class-weighted loss**: per-class weights (inversely proportional to class frequency) are applied during training to further counteract imbalance.

Results

Table 5.7 shows the validation-set metrics for all three model versions. All figures are read directly from the `metrics` block in `classifier_metadata.json` for each model, with no post-hoc adjustments.

Table 5.7: RoBERTa classifier metrics on the held-out validation set. Accuracy = fraction of val samples correctly classified. F1 macro = unweighted mean of per-class F1, penalising poor performance on small classes equally to large ones. F1 weighted = class-frequency-weighted mean F1. Early-stop epoch = the epoch at which validation macro F1 peaked.

Model	Cl.	Val N	Acc.	F1 mac.	F1 wt.	Stop ep.
Baseline (no augmentation)	15	2,031	0.608	0.608	0.606	10/10
Augmented, flat topics	15	2,031	0.689	0.639	0.684	12/20
Augmented, subtopics (final)	21	1,108	**0.769**	**0.716**	**0.770**	10/20

The final model achieves **76.9% accuracy** and **macro F1 = 0.716** on the 1,108-sample validation set. The gap between macro F1 (0.716) and weighted F1 (0.770) reflects class-

size imbalance, where the model does better on large topics (Trump/Biden, Vaccines) and worse on tail topics with few training examples (Zodiac Killer, EcoHealth subtopic). A per-class F1 breakdown was printed to the training console but was not saved to disk, so the macro/weighted gap is the only available proxy for this imbalance.

The three incremental improvements are:

1. **Augmentation alone** (baseline $\rightarrow$ flat-augmented): accuracy $+8.1$ pp, macro F1 $+3.1$ pp. The upsampling step is the dominant contributor, as without it small classes are effectively invisible to the model.

2. **Finer-grained subtopic labels** (flat-augmented $\rightarrow$ subtopic-final): accuracy $+8.0$ pp, macro F1 $+7.7$ pp. Splitting the three heterogeneous parents gave the model cleaner, more coherent class boundaries. The macro F1 gain is larger than the accuracy gain because the subtopics are more separable from each other than the parent topics were, which improves per-class recall uniformly.

3. **Early stopping** prevented degradation beyond epoch 10 in the final run. Training for all 20 epochs would have overfitted given the small per-class sample sizes for the tail subtopics.

Centroid pipeline

The centroid method is unsupervised and assigns a post to the topic whose centroid has the highest cosine similarity above an absolute threshold of 0.55, with a noise-margin guard. Because the centroids are computed from the BERTopic assignments that also serve as the ground-truth labels for the RoBERTa head, evaluating the centroid method against those same labels would be circular. Instead, two indirect characterisations are reported:

- **BERTopic cluster coherence.** The BERTopic run produces a silhouette score of 0.562 for the Vaccine sub-clusters and 0.420-0.626 across the three sub-clustered parents (Table 4.7), indicating that the geometric structure the centroids encode is reasonably clean, though the Russia/Ukraine four-way split is the weakest ($s = 0.420$).

- **Agreement with RoBERTa.** Across the HVU corpus, the centroid method and the RoBERTa head agree on the top-1 topic assignment in approximately two-thirds of cases, with the exact fraction depending on the confidence floor applied. Disagreements are concentrated on borderline posts where the centroid similarity is close to the noise baseline or where the gap between the top-two topics is small, which is exactly the regime the RoBERTa rescue path handles.

Limitations

1. **No held-out test set.** The reported metrics come from the validation split, not from a third partition that was never seen during training or hyperparameter selection. The validation set was held out before augmentation and its labels were not used to tune any hyperparameter directly, so the metrics are a reasonable estimate of generalisation, but they are not a guarantee. With a corpus of this size, reserving a test set would have further reduced per-class training signal for the small tail topics, which is why the two-split design was chosen.

2. **BERTopic labels as ground truth.** The classifier learns to reproduce BERTopic's cluster assignments, not a human-curated taxonomy. A post that genuinely straddles two conspiracy narratives will be labelled by whichever cluster centroid it falls nearest, and the classifier will learn that potentially arbitrary boundary. The macro F1 therefore measures agreement with BERTopic, not agreement with an independent human annotation.

3. **In-domain evaluation.** Both the training data and the validation data come from the same HVU Reddit corpus collected in the same six-month window. The model's ability to generalise to other platforms, other time windows, or longer-form text has not been tested.

4. **Per-class breakdown unavailable.** The full classification report (per-class precision, recall, F1) was printed to the training console but was not saved to disk. The macro vs. weighted F1 gap is the only available proxy for per-class imbalance.

5.9 Summary of Findings

Table 5.8 gathers the one-line verdicts for each research question on a single sheet. "Sig. tests / total" refers to the number of per-topic tests that survive Benjamini-Hochberg correction at $\alpha = 0.05$. The effect-size columns report the most extreme per-topic point estimate observed at the main-topic level and at the subtopic level, whichever is larger.

Table 5.8: Effect-size summary across research questions. Every per-topic family is a BH-corrected topic-level null except RQ7, which is the sole surviving significance test in the chapter.

RQ	Metric	Main topics	Subtopics
RQ1	Composition χ^2 (sig./total)	0/14	0/19
	Max Cramér's V	0.162 (small)	0.170 (small)
RQ2	Intensity Mann-Whitney (sig./total)	0/14	0/19
	Max |Cohen's d|	0.24 (small)	0.48 (small)
RQ3	Audience entropy range	0.61–0.93	—
	Topics with one-sided audience ($H < 0.5$)	0/14	—
RQ4	Odds-ratio tests (sig./total)	0/14	0/20
	OR range	0.78–1.46	0.70–1.46
	Best-ranked logistic predictor	entropy (6/14)	entropy (7/20)
RQ5	Composite score tests (sig./total)	0/4 composite scores	
	Max composite $\|d\|$	0.037 (negligible)	
	Sub-score tests (sig./total)	2/3 (conspiracy, balance)	
	Max sub-score $\|d\|$	0.102 (negligible)	
RQ6	Mean edge cosine distance (observed vs null)	0.4425 vs null 0.4685, 0th pct., $p < 0.001$ **homophily**	
	Spearman $H(u)$ vs ND(u)	$\rho = -0.369$, $p < 0.001$	
	Structural role difference	n.s. for degree, betweenness, PageRank	
RQ7	Subreddit $\times$ ideology χ^2	$\chi^2 = 8.58$, $p = 0.035$ **significant**	
	R$-$L gap range	$+10.3$ pp to $+21.8$ pp	
RQ8	RoBERTa validation accuracy	76.9%	
	RoBERTa macro F1 / weighted F1	0.716 / 0.770	
	Centroid–BERT top-1 agreement	$\approx$2/3 of corpus posts	

Read top-to-bottom, the table tells a four-part story.

RQ1-RQ5: a topic-level null with a mild right-ward tilt. Every per-topic family returns 0 BH-surviving tests across composition, intensity, and odds-ratio comparisons, and every effect size stays inside the negligible-to-small band. Inside that null, two residual patterns are worth naming so that the chapter is not read as claiming the point estimates are exactly flat. First, right-leaning users post slightly more per-user in almost every topic (uniform positive engagement-share gap, max $|d| = 0.24$ at the main level), and they carry slightly higher conspiracy-activity and balance sub-scores at the population level (both $p_{\text{adj}} < 0.01$, both $|d| < 0.12$), which is the textbook large-N case of a real but practically negligible directional bias. Second, when the logistic regression in RQ4 lets direction and user-level diversity compete for the same outcome, diversity wins the most topics ($6/14$ at the main level, $7/20$ at the subtopic level), which reframes the question from which side engages to how broadly distributed a user's political footprint is.

RQ6: the network is ideologically assortative beyond chance. The null model (500 shuffles of ideology vectors, graph structure held fixed) confirms that the observed mean cosine edge distance (0.4425) is lower than the entire null distribution (0th percentile, $p <$ 0.001). Users in r/conspiracy interact preferentially with ideologically similar peers. This

structural homophily does not conflict with the topic-level nulls of RQ1-RQ5. Content choice and interaction partner selection are separate layers of behaviour. Ideologically mixed users hold no privileged structural position in the network.

RQ7: the one surviving topic-level significance is about where, not what. The only BH-corrected test that rejects its null is the ideology $\times$ subreddit chi-squared ($\chi^2 = 8.58$, $p = 0.035$), with the R$-$L gap running from $+10.3$ pp in r/conspiracytheories to $+21.8$ pp in r/conspiracy_commons. The same right-skewed subreddit carries the highest concentration of global top-decile bridging users (21.8% vs. 10% expected). The topic-share heatmap does not line up with the ideology gradient, so the finding is about community choice rather than narrative content.

RQ8: the classifier reaches a practical accuracy level. The final RoBERTa model achieves 76.9% accuracy and macro F1 = 0.716 on the held-out validation set. These scores are sufficient for the Chrome extension use-case, where the centroid method handles the majority of clear-cut cases and the RoBERTa head resolves borderline ones.

Taken together, the chapter argues that **conspiracy engagement on Reddit is ideology-agnostic at the topic level but shows structural homophily at the network level and ideological stratification at the subreddit-choice level**. Which narratives people engage with separates only weakly by lean. Who they interact with and which community they settle into both show ideological sorting.

Practical implications: These findings carry several consequences for how conspiracy engagement is understood and addressed outside an academic context.

For **content moderation**, the topic-level null (RQ1-RQ5) means that flagging conspiracy narratives by the assumed ideology of their audience is not supported by this data. Every conspiracy topic in the corpus draws from both political sides in roughly the same proportions, so a moderation strategy that treats certain narratives as inherently "right-wing" or "left-wing" would misclassify a large share of its participants. A more defensible approach would target specific narrative features (misinformation claims, incitement, coordinated inauthentic behaviour) regardless of the ideological profile of the community discussing them.

For **media literacy**, the RQ4 finding that political diversity out-predicts political direction as a correlate of conspiracy engagement suggests that the relevant educational intervention is not "your side is more susceptible" but rather "broad, unfocused political consumption correlates with conspiracy exposure". Programmes that help users recognise the structural features of conspiracy reasoning, rather than tying susceptibility to a partisan identity, are better aligned with what the data shows.

For **platform design**, the coexistence of topic-level ideology-agnosticism and network-level homophily (the RQ6 paradox discussed above) implies that recommendation algo-

rithms optimised for engagement can reinforce ideological clustering in the social graph even when the underlying content preferences are cross-partisan. If the algorithm surfaces content based on who a user has interacted with before, and those interaction partners are ideologically similar (as the null model confirms), then the recommendation loop tightens the social bubble without ever restricting the range of topics a user sees. The RQ7 subreddit-choice finding reinforces this point at the community level, where the ideological sorting happens at the "which subreddit do I join" decision rather than at the "which topic do I engage with" decision, so platform features that shape community discovery and onboarding have more leverage over ideological sorting than features that shape within-community content ranking.

These are the observed claims carried forward into Chapter 7.

Chapter 6

Related Work

This chapter reviews the existing literature relevant to this thesis, organised into four areas: conspiracy theory research (Section 6.1), ideology inference from social media (Section 6.2), topic modelling of online content (Section 6.3), and polarisation measurement (Section 6.4).

6.1 Conspiracy Theory Research

Conspiracy theory research has traditionally come from psychology and political science. Early work treated conspiracy belief as a personality trait or cognitive tendency [11, 21], looking at how it correlates with things like distrust of authority and need for cognitive closure. Douglas et al. [16] give a good overview of the field, identifying epistemic, existential, and social reasons why people believe in conspiracies. Van Prooijen et al. [52] found a link between illusory pattern perception and both conspiracy and supernatural beliefs, pointing to a general cognitive mechanism rather than something specific to politics.

More recently, researchers have started using computational methods to study conspiracy communities at scale. Bessi et al. [9] looked at how conspiracy vs. science content spreads on Facebook and found that conspiracy communities tend to form echo chambers with their own diffusion patterns. Vosoughi et al. [54] showed that false news spreads faster and reaches more people than true news on Twitter, a pattern that applies to conspiracy content as well. Zollo et al. [56] found that attempts to debunk conspiracy content are not very effective within conspiracy communities.

On Reddit specifically, Samory and Mitra [44] studied how conspiracy theories are discussed, finding common rhetorical strategies across different conspiracies. Klein et al. [25] looked at the linguistic precursors that lead users into r/conspiracy, and Ribeiro et al. [43] audited a similar radicalisation pathway on YouTube, tracing how users migrate from mainstream to fringe communities over time. Most of this work, however, uses keyword-based classification or manual labelling rather than the unsupervised topic discovery approach used in this thesis.

6.2 Ideological Inference from Social Media

Figuring out someone's political ideology from their online behaviour is a well-studied problem. Approaches range from text-based classification [40] to network methods that look at who follows or interacts with whom. Barberá [5] built a Bayesian ideal-point estimation method using Twitter follow networks and showed that social media behaviour can recover ideological positions that are comparable to what you get from surveys. Gentzkow and Shapiro [20] measured ideological segregation online and found that online exposure is actually more ideologically diverse than people usually assume.

On Reddit, prior work has classified subreddits on a left-right spectrum using manual labelling, user overlap analysis, and embedding methods [47, 55]. The subreddit classifications used in this thesis build on that work. Behavioural ideology inference has also been used across platforms, Lai et al. [26] estimate the ideology of political YouTube videos by treating Reddit posts that link to those videos as behavioural traces and applying correspondence analysis to the resulting user-video matrix, a methodology that parallels the participation-based inference used here but applied to a cross-platform setting.

One thing that sets this thesis apart is the use of *continuous ideology vectors* based on participation patterns rather than simple discrete labels. This is in line with recent arguments in the literature that political identity is more nuanced and multi-dimensional than a left/right binary [40].

6.3 Topic Modeling of Online Discourse

Topic modelling has been widely used on social media data. Traditional approaches like LDA [10] have been applied to find topics in Twitter, Reddit, and news datasets. More recently, neural topic models like BERTopic [23], Top2Vec [3], and others based on pre-trained language models [15] have shown better results on short, noisy text. These newer methods use sentence embeddings [42] to capture meaning that word-counting approaches miss.

BERTopic, the framework used here, combines UMAP [31] for dimensionality reduction with HDBSCAN [30] for density-based clustering, and it does not need you to specify how many topics to look for. UMAP was chosen over the older t-SNE [51] projection because it preserves more of the global document structure and scales better to the tens of thousands of embeddings produced by the HVU corpus, which matters when centroids computed in the reduced space have to be stable across repeated runs. It has been used in several social media studies and handles the vocabulary diversity and short posts typical of user-generated content well. This thesis extends the standard BERTopic pipeline with engagement weighting and topic validation to deal with the uneven quality

of posts in conspiracy communities.

6.4 Polarisation Measurement

Measuring political polarisation online is an active area of research. Common approaches include looking at partisan sorting in social networks [14], measuring ideological homophily in follower or interaction graphs [2], and computing opinion divergence on specific issues [19]. Conover et al. [14] found that retweet networks on Twitter are highly polarised while mention networks are more cross-partisan, showing that different types of interaction reveal different polarisation patterns.

Paschalides et al. [37] proposed the POLAR framework for modelling polarisation in news media, which influenced the design of the Chrome extension in this thesis. A more direct precedent is Check-It [38], a browser plugin from the same advisor and supervisor that combines flag-lists, fact-check similarity, a user blacklist, and a logistic regression model to detect fake news on the web, while running entirely on the user's machine for GDPR compliance. The Conspiracy Identifier extension in this thesis follows the same locally-running, multi-signal design philosophy, swapping keyword lookups and logistic regression for semantic centroids and a RoBERTa head. Liotatis extended POLAR into a mobile app for on-demand polarisation detection, work that was done under the same supervisor and advisor as this thesis.

The eight polarisation metrics used here (defined in Section 4.5.2, including homophily, cross-engagement, subreddit cross-participation, narrative clustering, top-3 topic concentration, ideological bridge score, ideology imbalance, and an overall aggregate) are based on established network analysis concepts [32,35] but adapted to the Reddit-specific structure of the dataset.

6.5 Positioning of This Work

This thesis differs from previous work in a few key ways.

- It combines *behavioural ideology inference* with *unsupervised topic discovery*, avoiding the limitations of both self-report ideology measures and keyword-based conspiracy classification.
- It uses proper statistical testing with multiple-comparison correction and reports null results honestly-something often missing from studies that only look at descriptive patterns.
- It covers *four* conspiracy subreddits (not just r/conspiracy), giving a broader view of the ecosystem.

- It includes a *sub-clustering* step that uncovers ideological patterns that are invisible at a coarser level, showing why multi-resolution analysis matters.
- It delivers an *applied tool* (Chrome extension) that connects the offline analysis to real-time content classification.

Chapter 7

Conclusion

7.1 Discussion

This section situates the findings against the three research objectives outlined in the initial thesis document and discusses what the results mean in the context of prior work.

Objective 1: Identification of conspiracy topics

The first objective enquired about which conspiracy theories related to Trump, Biden, and Harris were most actively discussed in polarised Reddit communities during the 2024 election cycle. The unsupervised topic discovery pipeline (described in Chapter 4) identified 14 main conspiracy narratives. The most prominent by volume, *Trump, Biden, and Election Manipulation* ($N = 628$ unique posters), directly maps to the candidate focus. It covers electoral fraud claims, Trump shooting coverage, and accusations of manipulation by both campaigns. A second election-adjacent topic, *Election Integrity and Political Dynamics* ($N = 230$), captures narrower ballot-handling and debate-framing narratives. Together these two topics account for roughly a third of the clustered corpus and confirm that candidate-related conspiracy content was the dominant narrative thread in r/conspiracy and related subreddits during the election window.

Several additional topics carry candidate-adjacent content even without being named as such. For example, *Bill Gates, Lab-Made Viruses, and Elon Musk* references figures closely associated with the Trump campaign's rhetoric, and *Project 2025 and Election Manipulation* directly invokes a policy document that became a major point of contention between the two campaigns. The remaining topics (Vaccines, Russia/Ukraine, 9/11, Zodiac, Occult, Diddy/Drake, etc.) represent the broader conspiracy landscape that co-exists with election-specific narratives in the conspiracy subreddits, which is consistent with prior research showing that conspiracy communities are not mono-thematic [44]. The scope of the analysis naturally expanded to cover all discovered topics rather than filtering to candidate-specific ones only, because the research questions concern the relation-

ship between ideology and conspiracy engagement across the full topic landscape, and because no specific candidate-named topics were identified as being dominated by one single ideology. Additionally, restricting to candidate-named topics would discard a large portion of the corpus without evidence that doing so would change the findings.

Objective 2: Network analysis and ideological differences

The second objective was about how conspiracy narratives differ in structure and prominence across left and right communities. The results present a layered picture. At the *topic-composition and engagement* layer (RQ1-RQ5), no significant left-right sorting was found after Benjamini-Hochberg correction across any of the 14 main topics or 20 subtopic categories. This is consistent with the cross-cutting nature of conspiracy thinking identified in prior work [11, 16], which argues that conspiratorial thinking is a cognitive style not well predicted by partisan identity. The one structural variable that does predict conspiracy engagement is user-level *ideological diversity* (how broadly a user spreads activity across ideology tiers), which outperforms simple left-right direction as a predictor in 6 of 14 topics in logistic regression (RQ4). This aligns with the hypothesis that politically omnivorous users are structurally more likely to encounter conspiracy content regardless of which side they lean toward. At the *network interaction* layer (RQ6), the null model analysis reveals significant ideological homophily. Specifically, the observed mean cosine edge distance between co-active user pairs (0.4425) falls at the 0th percentile of 500 ideology-vector shuffles ($p < 0.001$), meaning r/conspiracy users interact preferentially with ideologically similar peers. This structural finding sits alongside the topic-level null without contradiction. Content choice (what topics to post about) and interaction partner selection (who to engage with in discussion threads) are distinct behaviour patterns, and ideology appears to influence the latter but not the former. At the *community* layer (RQ7), the ideology $\times$ subreddit chi-squared test ($\chi^2 = 8.58$, $p = 0.035$) is the only test to survive BH correction, with r/conspiracy_commons carrying the largest right-left gap (+21.8 pp) and the highest concentration of top-decile bridging users. This suggests that ideological sorting in the conspiracy ecosystem happens at the platform architecture level (which subreddit a user gravitates toward) rather than at the narrative level (which topics they post about within a subreddit).

Objective 3: Classification and utility

The fine-tuned RoBERTa classifier achieves 76.9% accuracy and a macro F1 of 0.716 on the 21-class (20 topics + off-topic) held-out validation set. The classifier is integrated into the Chrome extension alongside the centroid similarity pipeline. The centroid method handles clear-cut cases efficiently, while the RoBERTa head resolves borderline assign-

ments and provides the off-topic probability used to suppress false classifications. Together they provide a reliable classifier for short-form online discussions. The limits of the evaluation (in-domain data, BERTopic labels as ground truth, no separate test set) are documented in RQ8.

Overall interpretation

The thesis argues that the ideology-conspiracy link is real but operates at a different level of abstraction than simple left-right sorting would suggest. Ideology shapes which communities users inhabit and who they talk to, but once inside the conspiracy space, left and right users engage with essentially the same narratives. The most informative predictor is not *which side* a user is on but *how broadly* they participate across political communities, suggesting that political curiosity or omnivory, rather than partisan commitment, is the underlying driver of heavy conspiracy engagement.

7.2 Summary

This thesis set out to investigate whether political ideology predicts conspiracy theory engagement on Reddit during the 2024 U.S. presidential election cycle. To do this, a modular pipeline was built that infers user ideology from behavioural signals (political subreddit participation), discovers conspiracy topics through unsupervised clustering (BERTopic with EmbeddingGemma-300M), builds multi-layer networks for structural analysis, and tests the ideology-conspiracy link using proper statistical methods. The central empirical contribution is a set of findings at three distinct levels of analysis.

At the *topic and user level*, the main result is a null. Across 14 main topics and 20 subtopic-level categories, no statistically significant relationship was found between ideology and topic composition, engagement intensity, or predictive odds ratios after Benjamini-Hochberg correction. All effect sizes are negligible to small. Left and right-leaning users engage with the same conspiracy topics at similar rates.

At the *network level*, a null model based on 500 ideology-vector shuffles confirms significant ideological homophily in the user co-activity network. The observed mean cosine edge distance (0.4425) falls at the 0th percentile of the null distribution ($p < 0.001$). Users interact preferentially with ideologically similar peers even though they engage with the same topics. Ideologically mixed users do not occupy structurally privileged positions in the network.

At the *community level*, subreddit choice is linked to ideology ($\chi^2 = 8.58$, $p = 0.035$), with r/conspiracy_commons showing the largest right-left gap (+21.8 pp). This points to community culture and moderation norms as the mechanism through which ideology shapes conspiracy participation, rather than the actual content of the conspiracies.

Two further findings are worth highlighting. First, political *diversity* (how broadly someone participates across ideological subreddits) is a better predictor of conspiracy engagement than which *side* they are on, suggesting that curiosity or information-seeking behaviour matters more than partisan commitment. Second, sub-clustering revealed a left-leaning over-representation in the *Russian Influence* subtopic that was hidden at the main topic level, demonstrating the value of multi-granularity analysis.

7.3 Limitations

There are several limitations to keep in mind.

- **Ideology inference is behavioural, not attitudinal:** Users are classified based on where they post, which is a proxy for political belief but not a direct measure. A user who posts in r/Conservative to argue *against* conservative positions would be misclassified.
- **Temporal scope:** The dataset covers a single six-month period (May-November 2024). The ideology-conspiracy relationship might look different outside of an election season or under different political conditions.
- **Platform specificity:** Reddit has a particular structure (pseudonymous, community-based through subreddits) that may not generalise to platforms like Twitter/X, Facebook, Telegram, etc.
- **Noise rate:** The BERTopic clustering excludes almost half of all posts (the noise bucket), which means some meaningful but uncommon conspiracy narratives may have been missed. Any replication with a lower noise rate or a different clustering method could surface additional topics.
- **English-only analysis:** The study only covers English-language content and uses U.S.-centric political classifications.
- **Centre users excluded from the main analysis:** The 60.1% of users classified as centre are left out of left-vs-right comparisons. This reduces the sample size and might hide patterns specific to centre users.

7.4 Ethical Considerations

Studying user behaviour on Reddit raises several ethical questions worth discussing.

- **User privacy:** All data used in this study was publicly posted on Reddit, and no attempt was made to de-anonymise users beyond their Reddit usernames. Individual

users are not named or profiled in the thesis, all analysis is reported at the aggregate or group level.

- **Ideology inference without consent:** Users did not consent to having their political ideology inferred from their posting behaviour. This is a common approach in computational social science, but it is important to recognise that the inferred labels are approximations based on behavioural patterns, not ground truth about someone's beliefs. The ideology vectors should not be treated as definitive statements about any individual's political identity.

- **Labelling communities:** Classifying subreddits as "left", "right", or "centre" involves judgment calls that simplify a complex spectrum. The thesis uses and relies on established classification lists from prior research rather than manually classifiying subreddits, but any such classification will miss nuance.

- **Responsible use of findings:** The null result should not be read as "conspiracy content is harmless". It is specifically about whether left-right lean predicts *which* conspiracy topics users participate in, not about the truth or harm of those conspiracies.

- **Chrome extension:** The Conspiracy Identifier extension classifies posts, not people. It is designed as a transparency tool to help users understand what type of content they are reading, not as a surveillance mechanism. The extension runs locally and does not transmit user data.

- **Surfacing user activity and ideology in the extension:** When the author of a Reddit post is a known HVU, the extension displays their six-tier ideology label, activity counts, and bridging percentile alongside the post. Attaching a political label to an individual pseudonymous account is sensitive, and this trade-off is accepted only because the extension is a private, local research tool. The lookup table never leaves the installing machine, nothing is transmitted externally, and the extension is not distributed through the Chrome Web Store. Any future public deployment would have to drop the per-user profile surface or move to an aggregate-only view.

7.5 Future Work

Several directions could extend this work.

- **Stance detection for ideology inference:** The current pipeline infers ideology purely from *where* a user posts, not *how* they post. A user who is active in r/Conservative

to argue against conservative positions would be misclassified as right-leaning. Incorporating stance detection, classifying each post or comment as supportive or critical of the subreddit's dominant viewpoint, would produce more accurate ideology vectors and could reveal users whose actual beliefs diverge from their posting communities.

- **Cross-platform expansion:** The subreddit-as-community-hub structure that makes Reddit well-suited for behavioural ideology inference also exists on platforms like Lemmy and Discord, which the pipeline could be retargeted to with minimal changes. Running it on structurally different platforms (Twitter/X, Telegram) would additionally test whether the null result generalises beyond Reddit's specific ecosystem.

- **Comment-level analysis:** This thesis focuses on posts, despite there being an option to include comments when using the Chrome extension. No comments were used when building the topic model or the classification model, which means that a large portion of user engagement with conspiracy content is not captured. Specifically, the number of comments in our dataset is an order of magnitude higher than the number of posts, and many users engage only through commenting rather than posting. Looking at comments too would capture finer-grained engagement and allow for sentiment analysis of how users *react to* conspiracy content, not just what they post about.

- **Longitudinal tracking:** Running the same analysis across multiple election cycles would show whether the null result holds up over time or is specific to this period.

- **Causal inference:** This analysis shows associations, not cause and effect. To understand whether ideology influences conspiracy engagement, or conspiracy engagement influences ideology, future work could use instrumental variables or natural experiments.

- **Better noise handling:** Techniques like guided topic modelling or hierarchical clustering could reduce the 45.3% noise rate and recover topics from the currently unclustered posts.

- **Multilingual and non-U.S. contexts:** As noted in the limitations above, the current analysis is English-only and U.S.-centric. Extending the pipeline to other languages and political systems (e.g., European left-right spectrums, or multi-party systems) would test how culturally dependent the findings are.

- **Chrome extension improvements:** The Conspiracy Identifier could be extended to work on other platforms, detect trending topics in real time, and give users explanations of why a post was classified a certain way. A concrete next step is the

in-browser deployment discussed in Section 4.6.1. Porting the embedding model to `transformers.js` over WebGPU would remove the dependency on the local FastAPI backend and make the extension installable as a single artefact, at the cost of re-validating the in-browser embeddings against the Python centroids.

Bibliography

[1] Academic Torrents. Reddit political subreddit data. https://academictorrents.com/details/ba051999301b109eab37d16f027b3f49ade2de13, 2024. Accessed: 2025-09-15.

[2] Lada A. Adamic and Natalie Glance. The political blogosphere and the 2004 U.S. election: Divided they blog. *Proceedings of the 3rd International Workshop on Link Discovery*, pages 36–43, 2005.

[3] Dimo Angelov. Top2Vec: Distributed representations of topics. *arXiv preprint arXiv:2008.09470*, 2020.

[4] Arctic Shift. Arctic shift download tool. https://arctic-shift.photon-reddit.com/download-tool, 2024. Accessed: 2025-09-15.

[5] Pablo Barberá. Birds of the same feather tweet together: Bayesian ideal point estimation using Twitter data. In *Political Analysis*, volume 23, pages 76–91, 2015.

[6] Mathieu Bastian, Sebastien Heymann, and Mathieu Jacomy. *Gephi: An Open Source Software for Exploring and Manipulating Networks*. 2009.

[7] Jason Baumgartner, Savvas Zannettou, Brian Keegan, Megan Squire, and Jeremy Blackburn. The Pushshift Reddit dataset. In *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, volume 14, pages 830–839, 2020.

[8] Yoav Benjamini and Yosef Hochberg. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society: Series B*, 57(1):289–300, 1995.

[9] Alessandro Bessi, Mauro Coletto, George Alexandru Davidescu, Antonio Scala, Guido Caldarelli, and Walter Quattrociocchi. Science vs conspiracy: Collective narratives in the age of misinformation. *PLoS ONE*, 10(2):e0118093, 2015.

[10] David M. Blei, Andrew Y. Ng, and Michael I. Jordan. Latent Dirichlet allocation. *Journal of Machine Learning Research*, 3:993–1022, 2003.

[11] Robert Brotherton, Christopher C. French, and Alan D. Pickering. Measuring belief in conspiracy theories: The generic conspiracist beliefs scale. *Frontiers in Psychology*, 4:279, 2013.

[12] Matteo Cinelli, Gianmarco De Francisci Morales, Alessandro Galeazzi, Walter Quattrociocchi, and Michele Starnini. The echo chamber effect on social media. *Proceedings of the National Academy of Sciences*, 118(9):e2023301118, 2021.

[13] Jacob Cohen. Statistical power analysis for the behavioral sciences. 1988.

[14] Michael D. Conover, Jacob Ratkiewicz, Matthew Francisco, Bruno Gonçalves, Filippo Menczer, and Alessandro Flaminini. Political polarization on Twitter. *Proceedings of the International AAAI Conference on Web and Social Media (ICWSM)*, 5(1):89–96, 2011.

[15] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL)*, pages 4171–4186, 2019.

[16] Karen M. Douglas, Joseph E. Uscinski, Robbie M. Sutton, Aleksandra Cichocka, Turkay Nefes, Chee Siang Ang, and Farzin Deravi. Understanding conspiracy theories. *Political Psychology*, 40(S1):3–35, 2019.

[17] Adam M. Enders, Steven M. Smallpage, and Robert N. Lupton. Are all 'birthers' conspiracy theorists? On the relationship between conspiratorial thinking and political orientations. *British Journal of Political Science*, 50(3):849–866, 2020.

[18] Linton C. Freeman. A set of measures of centrality based on betweenness. *Sociometry*, 40(1):35–41, 1977.

[19] Kiran Garimella, Gianmarco De Francisci Morales, Aristides Gionis, and Michael Mathioudakis. Quantifying controversy on social media. *ACM Transactions on Social Computing*, 1(1):1–27, 2018.

[20] Matthew Gentzkow and Jesse M. Shapiro. Ideological segregation online and offline. *The Quarterly Journal of Economics*, 126(4):1799–1839, 2011.

[21] Ted Goertzel. Belief in conspiracy theories. *Political Psychology*, 15(4):731–742, 1994.

[22] Google. EmbeddingGemma-300M: Sentence embedding model. `https://huggingface.co/google/embeddinggemma-300m`, 2024. Accessed: 2025-09-15.

[23] Maarten Grootendorst. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. *arXiv preprint arXiv:2203.05794*, 2022.

[24] Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. On calibration of modern neural networks. In *Proceedings of the 34th International Conference on Machine Learning (ICML)*, pages 1321–1330, 2017.

[25] Colin Klein, Peter Clutton, and Adam G. Dunn. Pathways to conspiracy: The social and linguistic precursors of involvement in Reddit's conspiracy theory forum. In *PLoS ONE*, volume 14, page e0225098, 2019.

[26] Angela Lai, Megan A. Brown, James Bisbee, Joshua A. Tucker, Jonathan Nagler, and Richard Bonneau. Estimating the ideology of political YouTube videos. *Political Analysis*, 32(3):345–360, 2024.

[27] Daniel D. Lee and H. Sebastian Seung. Learning the parts of objects by non-negative matrix factorization. In *Nature*, volume 401, pages 788–791, 1999.

[28] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. RoBERTa: A robustly optimized BERT pretraining approach. In *arXiv preprint arXiv:1907.11692*, 2019.

[29] Henry B. Mann and Donald R. Whitney. On a test of whether one of two random variables is stochastically larger than the other. *The Annals of Mathematical Statistics*, 18(1):50–60, 1947.

[30] Leland McInnes, John Healy, and Steve Astels. hdbscan: Hierarchical density based clustering. In *Journal of Open Source Software*, volume 2, page 205, 2017.

[31] Leland McInnes, John Healy, and James Melville. UMAP: Uniform manifold approximation and projection for dimension reduction. *arXiv preprint arXiv:1802.03426*, 2018.

[32] Miller McPherson, Lynn Smith-Lovin, and James M. Cook. Birds of a feather: Homophily in social networks. *Annual Review of Sociology*, 27:415–444, 2001.

[33] Joanne M. Miller, Kyle L. Saunders, and Christina E. Farhart. Conspiracy endorsement as motivated reasoning: The moderating roles of political knowledge and trust. *American Journal of Political Science*, 60(4):824–844, 2016.

[34] Rafael Müller, Simon Kornblith, and Geoffrey Hinton. When does label smoothing help? In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 32, 2019.

[35] Mark E.J. Newman. Modularity and community structure in networks. *Proceedings of the National Academy of Sciences*, 103(23):8577–8582, 2006.

[36] Lawrence Page, Sergey Brin, Rajeev Motwani, and Terry Winograd. The PageRank citation ranking: Bringing order to the web. Technical Report 1999-66, Stanford InfoLab, 1998.

[37] Demetris Paschalides, Chrysovalantis Christodoulou, Rafael Andreou, George Pallis, and Marios D. Dikaiakos. POLAR: A holistic framework for the modelling of polarization in news media. In *Proceedings of the 12th ACM Conference on Web Science*, 2020.

[38] Demetris Paschalides, Chrysovalantis Christodoulou, Kalia Orphanou, Rafael Andreou, Alexandros Kornilakis, George Pallis, Marios D. Dikaiakos, and Evangelos Markatos. Check-It: A plugin for detecting fake news on the web. *Online Social Networks and Media*, 25:100156, 2021.

[39] Karl Pearson. On the criterion that a given system of deviations from the probable in the case of a correlated system of variables is such that it can be reasonably supposed to have arisen from random sampling. *Philosophical Magazine Series 5*, 50(302):157–175, 1900.

[40] Daniel Preotiuc-Pietro, Ye Liu, Daniel Hopkins, and Lyle Ungar. Beyond binary labels: Political ideology prediction of Twitter users. *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 729–740, 2017.

[41] Juan Ramos. Using tf-idf to determine word relevance in document queries. *Proceedings of the First Instructional Conference on Machine Learning*, 242(1):29–48, 2003.

[42] Nils Reimers and Iryna Gurevych. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 3982–3992, 2019.

[43] Manoel Horta Ribeiro, Raphael Ottoni, Robert West, Virgílio A.F. Almeida, and Wagner Meira. Auditing radicalization pathways on YouTube. In *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency (FAccT)*, pages 131–141, 2020.

[44] Mattia Samory and Tanushree Mitra. 'The Government Doesn't Want You To Know': How conspiracy theories are discussed on Reddit. In *Proceedings of the ACM on Human-Computer Interaction*, volume 2, pages 1–24, 2018.

[45] Claude E. Shannon. A mathematical theory of communication. *Bell System Technical Journal*, 27(3):379–423, 1948.

[46] Ahmed Soliman et al. Reddit political subreddit classification, 2024.

[47] Ahmed Soliman, Jan Hafer, and Florian Lemmerich. A characterization of political communities on Reddit. In *Proceedings of the 30th ACM Conference on Hypertext and Social Media*, pages 259–263, 2019.

[48] Charles Spearman. The proof and measurement of association between two things. *The American Journal of Psychology*, 15(1):72–101, 1904.

[49] Natalie Jomini Stroud. Polarization and partisan selective exposure. *Journal of Communication*, 60(3):556–576, 2010.

[50] Joseph E. Uscinski and Joseph M. Parent. American conspiracy theories. 2014.

[51] Laurens van der Maaten and Geoffrey Hinton. Visualizing data using t-SNE. *Journal of Machine Learning Research*, 9:2579–2605, 2008.

[52] Jan-Willem van Prooijen, Karen M. Douglas, and Clara De Inocencio. Connecting the dots: Illusory pattern perception predicts belief in conspiracies and the supernatural. *European Journal of Social Psychology*, 48(3):320–335, 2018.

[53] Jan-Willem van Prooijen, André P. M. Krouwel, and Thomas V. Pollet. Political extremism predicts belief in conspiracy theories. *Social Psychological and Personality Science*, 6(5):570–578, 2015.

[54] Soroush Vosoughi, Deb Roy, and Sinan Aral. The spread of true and false news online. *Science*, 359(6380):1146–1151, 2018.

[55] Isaac Waller and Ashton Anderson. Quantifying social organization and political polarization in online platforms. In *Nature*, volume 600, pages 264–268, 2021.

[56] Fabiana Zollo, Alessandro Bessi, Michela Del Vicario, Antonio Scala, Guido Caldarelli, Louis Shekhtman, Shlomo Havlin, and Walter Quattrociocchi. Debunking in a world of tribes. *PLoS ONE*, 12(7):e0181821, 2017.