Effects of Collaboration on the Performance of Interactive Theme Discovery Systems

Alvin Po-Chun Chen* **Rohan Das*** **Dananjay Srinivas***
Alexandra Barry **Maksim Seniw** **Maria Leonor Pacheco**
University of Colorado Boulder
{alvin.chen, rohan.das, dananjay.srinivas}@colorado.edu

Abstract

NLP-assisted solutions to support qualitative data analysis have gained considerable traction. However, no unified evaluation framework exists which can account for the many different settings in which qualitative researchers may employ them. In this paper, we propose a framework to evaluate the way collaboration settings may produce different research outcomes across a variety of interactive systems. Specifically, we study the impact of synchronous vs. asynchronous collaboration using three different NLP-assisted qualitative research tools and present a comprehensive analysis of the differences in the consistency, cohesiveness, and correctness of their outcomes.

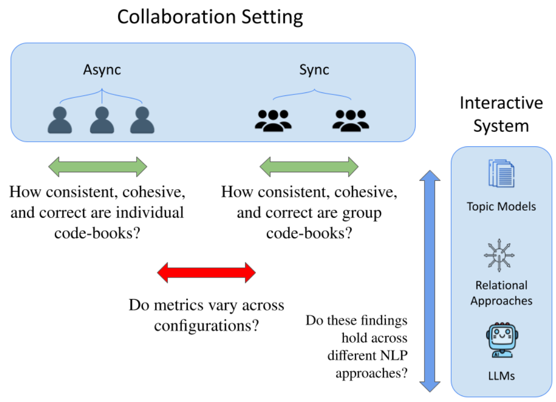


Figure 1: In this study, we measure the quality of coded themes using different interactive systems under different coding configurations.

1 Introduction

Making sense of large textual datasets is a common challenge across academic disciplines and is traditionally addressed through qualitative methods such as Thematic Analysis (Braun and Clarke, 2006) and Grounded Theory (Glaser et al., 1968). These approaches rely on manual *inductive coding*, in which researchers identify abstract themes by closely reading the data. However, as datasets grow in size, manual coding becomes impractical, motivating the use of Natural Language Processing (NLP) techniques to automate parts of the analysis process (Brady, 2019; Hilbert et al., 2019).

In recent years, a range of NLP-based systems have been developed to support qualitative research. These systems assist researchers by uncovering latent semantic structures through topic modeling (Smith et al., 2018; Fang et al., 2023), clustering documents and propagating limited human annotations across datasets (Pacheco et al., 2023; Chew et al., 2023), or offering real-time coding recommendations (Dai et al., 2023; Gao et al., 2024). To maintain researcher agency, such systems typically adopt human-in-the-loop (HitL) strategies that balance automation with manual interpretation.

Previous work typically evaluates HiTL qualitative analysis tools in isolation, focusing on specific technical strengths and weaknesses. For example, by comparing topic coherence with and without human input (Fang et al., 2023) or contrasting machine-assisted and manual code-book generation (Dai et al., 2023). However, qualitative analysis in practice is often collaborative, with teams of researchers jointly coding and interpreting data (Flick, 2014), and supported by tools built on diverse methodologies that are applied to datasets with vastly differing characteristics (Baden et al., 2022). By abstracting away collaboration settings, methodological variation, and dataset diversity, existing evaluations risk misrepresenting how HitL systems operate in real-world settings and how broadly their findings apply. In this work, we seek to answer the following questions: (1) Does the collaboration setting measurably affect the quality of resulting code-books? (2) Do these findings hold across different NLP approaches? (3) How do dataset characteristics influence the outcomes

*Equal Contribution

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 42507–42526
July 2-7, 2026

of NLP-assisted inductive coding tools?

We focus on two common but contrasting collaboration settings: asynchronous coding, where individuals code independently before consolidating results, and synchronous coding, where teams identify themes through live discussion. These settings offer complementary strengths, with asynchronous coding supporting flexibility across time and location, and synchronous coding facilitating shared understanding and efficient coordination. To compare outcomes across these settings, we introduce an evaluation framework that measures consistency between synchronous and asynchronous coding, as well as the cohesiveness and correctness of themes produced within each setting.

We evaluate three NLP-assisted inductive coding tools built on distinct methodological foundations: a human-in-the-loop topic modeling system (Fang et al., 2023), a concept-driven thematic modeling approach (Pacheco et al., 2023), and an LLM-based system for evaluating and propagating code definitions (Chew et al., 2023). To assess generalizability across data characteristics, we test all tools and collaboration settings on two markedly different datasets: a corpus of short social media posts and a collection of advertising texts. We further complement our quantitative analysis with a small-scale user study to capture coder experiences and derive design recommendations.

Our contributions are twofold: (1) we demonstrate that collaboration setting affects the results of NLP-assisted inductive coding tools across diverse methodologies and datasets, and (2) we provide an evaluation strategy that captures multiple dimensions of analysis quality. Together, these findings aim to inform the design and evaluation of language technologies that better align with real-world qualitative research workflows. Our code and experimental data are publicly available[†].

2 Related Work

The overarching goal of the systems we investigate is to partially automate the qualitative coding process, either by inducing topics in an interactive, semi-supervised manner (Fang et al., 2023; Smith et al., 2018), by learning user-defined themes interactively (Pacheco et al., 2023; Gao et al., 2023), or by prompting LLMs with natural language definitions of the observed themes (Chew et al., 2023; Dai et al., 2023). A separate but related line of work exemplified by Gao et al. (2024) uses LLMs to generate label recommendations as users perform the coding process. While this system is explicitly designed for asynchronous collaboration, the systems we study differ in their ability to annotate large portions of the dataset without extensive supervision.

Our research addresses a real-world use case for qualitative researchers using HiTL systems and is informed by the Human-Computer Interaction (HCI) literature (Jiang et al., 2021; Feuston and Brubaker, 2021; Chen et al., 2018). Prior HCI work shows that coders place particular emphasis on identifying and resolving ambiguity. In traditional settings, this is supported by an independent close reading of the data. However, in large-scale, NLP-assisted analysis, coders have limited visibility into where such ambiguities arise. Solutions have been proposed to either visualize codes (Drouhard et al., 2017) or rank document disagreement (Zade et al., 2018) regardless of dataset size. In contrast, our evaluation methods show different qualities of the resulting themes and document assignments by using signals from group overlaps, relationships in the semantic embedding space, and post-hoc evaluations. This highlights areas where the coders diverge both with each other and with the model, providing another perspective on the ambiguity question.

Previous evaluation methods introduced with HiTL systems for qualitative coding have generally been ad hoc, with experiments conducted in various group settings (Choo et al., 2013; Hoque and Carenini, 2016; Smith et al., 2018), on individual participants (Rietz et al., 2020), and through platforms such as MTurk (Zade et al., 2018). Our contribution provides a standardized framework for performing experiments regardless of the collaboration modality, using a suite of metrics for evaluating consistency, cohesiveness, and correctness in experimental results.

3 Interactive Systems

We identify three categories of NLP techniques used in interactive systems for qualitative coding with large datasets: topic models, relational approaches, and LLMs. Other techniques exist, but we focus on these three for their ability to help code large datasets. To maximize coverage across systems, we select a representative system from each category to use in our experiments. In this section, we briefly describe the unique aspects of

[†] https://github.com/blast-cu/interactive-systems

each category and introduce the selected system.

3.1 Topic Models

Tools in this category use some variation of topic modeling to find emerging themes and facilitate document assignment. These systems benefit from the relative speed of the topic model, which allow users to quickly visualize and explore the dataset. Early exploration incorporated visualizations to help users adjust parameters (Chuang and McFarland, 2013), while later works implemented refinement operations that allow users to directly edit topic words and remove documents (Smith et al., 2018). However, topic modeling systems are limited by their lack of malleability and predictability. Refinement operations mostly edit topic words, which can have limited impact in the final results since users cannot directly reassign documents.

We select the HitL query-driven topic model (QDTM) system introduced by Fang et al. (2023). This topic model is initialized by providing input queries (words that represent concepts of interest for the user) which the model uses to generate the initial topics. Users begin by iterating through each topic and naming them based on identified themes. They then use a set of *refinement operations* to edit the topic model. These include: merging and splitting topics based on topic words, adding, removing, or reordering topic words, and removing documents from topics. The next iteration of the model is only produced when the users choose to apply refinements. Each iteration is saved, allowing users to return to prior iterations to test different operations. Once satisfied with the state of the topic model, the user downloads the document distribution for that iteration.

To ensure comparable results, we use the same starting distribution of 13 topics for all our experiments using the same hyperparameters as Fang et al. (2021) ($\alpha = 1.0$, $\beta = 0.5$, $\gamma = 1.5$) and without any input queries. The same initial topic model is provided for all experiments.

3.2 Relational Approaches

Relational approaches combine vector semantics and structured inference to model relationships between high-level concepts. Instead of treating themes as distributions over words (as topic models do), these frameworks define themes as distributions over generalized concepts. This reflects the inductive coding process, where researchers identify patterns and concepts that are then synthesized into more abstract themes. However, their computational complexity grows with the number of dependencies considered, which hinders their ability to quickly adapt during coding sessions. Further, they rely on users to define informative concepts, making them less suited for inexperienced researchers.

We select the relational system introduced by Pacheco et al. (2023), which uses a two-stage relational framework. In the first stage, the system automatically partitions the dataset based on semantic similarity. We use SBERT (Reimers and Gurevych, 2019) to embed each document and partition the dataset using K-Means clustering (Jin and Han, 2010). The users explore each partition to identify themes, assign "good" and "bad" example documents for each theme, and input or correct supporting concepts for each example. In the second stage, the system uses the provided examples and concept relations to map the remaining dataset, only leaving documents unmapped if no theme is a sufficiently good match. Assignments are produced by a structured inference procedure, formulated as an approximate integer linear program, that explicitly models dependencies between concepts and themes. The unmapped documents are repartitioned as in the first stage and users are prompted to review unmapped partitions again. The process is performed iteratively until all documents are mapped or until users are satisfied.

In our experiments, researchers are provided with initial clusters to identify themes. We generate $K = 10$ clusters to form the initial partitions, which are kept consistent across experiments. Within each partition, users name the themes they observe and select positive and negative examples for each. These examples are embedded and used to score semantic similarity with unlabeled documents. In parallel, users annotate each of these examples with its supporting concepts, which are then used to learn concept predictions over the corpus. The framework assumes a soft rule of the form *concept* $\Rightarrow$ *theme* for every concept-theme pair, with rule weights learned from the annotated examples. For instance, when coding the COVID-19 vaccine debate, the rule *anti-vax* $\Rightarrow$ *Natural Immunity is Effective* would receive a high weight if most examples of that theme were annotated with an anti-vax stance. The rules and their learned weights are then combined into a structured inference task that predicts label assignments jointly across the corpus. By drawing on both distributional similarity and learned concept-theme dependencies, the model extrapolates efficiently from a small set of manually labeled examples. Table 3 in Appendix A details all operations available to users during coding.

3.3 Large Language Models

LLMs are ideal candidates for interactive systems, especially for tasks such as qualitative coding where the model can be prompted to produce themes or explanations without ad-hoc training (Kojima et al., 2024). They have been used for theme recommendation (Gao et al., 2023), for code conflict resolution (Gao et al., 2024), and for automated document assignment (Xiao et al., 2023). However, the flexibility of LLM outputs also leads to hallucinations, which are only partially addressed by prompt engineering. Models further suffer from biases in training which are difficult to identify and can impact their ability to produce quality labels or recommendations (Chen et al., 2018). Additionally, their massive size is prohibitive when working with large datasets due to the high cost of inference.

We select the framework introduced by Chew et al. (2023). Here, the human coder first manually codes a representative subset of the data and drafts definitions for each code. The LLM is then prompted to label the data sample with the provided definitions. Agreement is calculated between human and model annotations using GWET's AC_1 (Gwet, 2008). The prompt is then tweaked iteratively to achieve a satisfactory level of agreement, and the best-performing version is used to prompt the model to code the rest of the dataset.

For our experiments, we select Llama 3.2 3B-Instruct as the base LLM for automated labeling. We use the same starting partitions as in the relational approach (Sec. 3.2), and themes are instantiated following the same process, except that users provide richer textual definitions for each theme rather than positive and negative examples and concept annotations. Additional details about our experimental settings can be found in Appendix B.

4 Study Design

To study the effects of collaboration settings on the performance of the three selected systems, we design a protocol that can be used for both synchronous and asynchronous settings. For each system, we conduct three asynchronous experiments with one coder each and two synchronous experiments with three coders each for a total of 30 experiments (15 per dataset). Evaluation metrics are calculated by comparing the resulting code-books within each experimental setting (e.g. the two code-books independently created by the two synchronous groups using the topic model). The rest of this section details the datasets, participant demographics, and experimental protocols.

4.1 Datasets

We perform our experiments using two distinct datasets. The first consists of 85,799 tweets about COVID-19 vaccines posted by users located in the US, uniformly distributed between Jan.-Oct. 2021 (Pacheco et al., 2022). The corpus also contains labels for vaccination stance (e.g. pro-vax, anti-vax) and morality frames (e.g. fairness/cheating and their actor/targets) (Roy et al., 2021), which are used as auxiliary concepts for the relational model. The second dataset consists of 5,471 climate related English language ads from the Facebook Ad Library (Islam et al., 2023). The other dataset focuses on Facebook ads shown in the US between Jan. 2021-Jan. 2022, and contain labels for climate change stance. The two corpora differ substantially in length, complexity, and semantic structure: COVID tweets are short and semantically homogeneous, while climate ads are longer, more linguistically complex, and more formulaic. We characterize these differences quantitatively in Appendix C and discuss their implications in our evaluation (Section 5.4).

4.2 Participants

We recruited a group of 33 researchers in NLP and Computational Social Science, 9 female and 24 male, between the ages of 20 and 45. This group included professors at different levels of seniority, postdoctoral researchers, and graduate and undergraduate students from two different universities. This group covers the range of researchers likely to use interactive coding systems. All participants were either well-versed in qualitative data analysis, or were explicitly trained by senior researchers to perform the task. Due to the large number of experiments in our study, some participants took part in multiple experiments. These participants always performed the asynchronous experiment first to prevent external influence and took part in at most one synchronous experiment.

		Topic Model			Relational			LLM-Based		
		Jaccard	Centroid	Group Avg.	Jaccard	Centroid	Group Avg.	Jaccard	Centroid	Group Avg.
COVID	Sync	**0.56(0.23)**	**0.98(0.05)****	**0.52(0.10)**	**0.36(0.19)**	**0.98(0.01)***	**0.52(0.07)***	0.14(0.08)	0.98(0.03)	0.44(0.03)
	Async	0.30(0.17)	0.96(0.05)**	0.51(0.09)	0.30(0.22)	0.94(0.07)*	0.44(0.10)*	**0.17(0.11)**	0.98(0.02)	**0.45(0.03)**
Climate	Sync	0.37(0.31)*	0.89(0.14)	0.43(0.14)	0.27(0.17)	**0.95(0.05)****	**0.43(0.08)**	0.09(0.07)*	**0.95(0.04)**	**0.37(0.07)**
	Async	**0.58(0.29)***	**0.93(0.12)**	**0.46(0.16)**	**0.33(0.22)**	0.94(0.09)**	0.42(0.08)	**0.13(0.07)***	0.94(0.06)	0.36(0.06)

Table 1: Avg. Consistency between Best Theme Matches *across* Coder Groups (standard deviation in brackets). **Bold** figures highlight higher consistency value between collaboration settings. *Statistically significant using a two-sample unpaired t-test with $p < 0.05$. ** Near statistically significant with $p \approx 0.05$.

4.3 Coding Protocol

At the start of each experiment, participants were provided with a demonstration of all the operations in their respective systems. Every system starts with an initial partition of the data, so participants were instructed to read the first 25 samples in each partition, and manually create/name any themes they identified before freely exploring the rest of the dataset and start performing operations to find more themes.

In the topic model experiments, we suggested that participants merge and split topics based on their identified themes before making fine-grained refinements. They were then asked to refine the topic model based on their identified themes such that every topic corresponds to a unique theme. They kept re-running the model and making refinements until they were satisfied with the results, or until they failed to effect any meaningful changes.

In the relational system experiments, participants were tasked with selecting example documents for each identified theme, as well as determining concept relations for them. Following Pacheco et al. (2023), the supporting concepts considered were vaccination stances and morality frames (e.g., the identified theme "natural immunity" has an "anti-vax" stance, and is tied to the "purity" frame). Once participants were satisfied with their themes and selections, the system automatically coded the rest of the dataset. Unmapped examples were repartitioned and returned to the participants for a second (and last) round of coding.

In the experiments for the LLM-based system, participants produced natural-language definitions for each identified theme and selected a set of good examples for them. We then prompted the LLM with different task-prompt templates to find the best prompt for each set of participant-generated definitions, which was then used to code the rest of the dataset. Details of the templates, as well as the human-model agreement for the best template can be found in Appendix B.

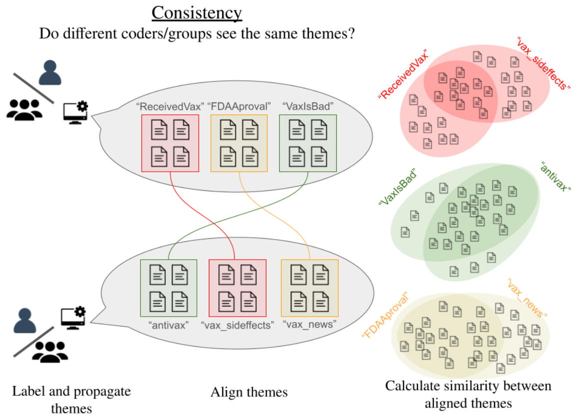


Figure 2: Two sets of coders use a particular HiTL system to find themes. Since the same theme can be named differently by different coders, we find the best match. In this example, the coder 1's theme "`VaxIsBad`" has been matched with coder 2's theme "`antivax`". After aligning, we calculate the similarity between these two themes using methods like Jaccard Similarity or Centroid Distance.

5 Evaluation

We use both descriptive metrics and a user study to provide a comprehensive analysis on the differences when coding in synchronous and asynchronous settings. Our evaluation framework is comprised of three dimensions; **consistency**, **cohesiveness & distinctiveness**, and **correctness**, each of which uses metrics that are well-established in the literature (Ben-David and Ackerman, 2008; Hoyle et al., 2021; Pacheco et al., 2023). The themes identified in each experiment are presented in Appendix D.

5.1 Consistency

Coders risk overgeneralizing or overlooking key themes, leading to unsystematic results (Cornish et al., 2014). We address this by measuring consistency, defined as the extent to which different coders elicit the same themes from the same texts (Fig. 2). In semi-automated coding, consistency is nontrivial to assess: similarly named themes may cover different documents, while differently named themes may overlap substantially. We therefore measure consistency based on document overlap between themes. Specifically, we compute the maximum Jaccard similarity between each theme and all themes produced by another coder, treating this maximum as the theme's best alignment. Consistency is then calculated as the average similarity across all aligned theme pairs. Example alignments can be found in Appendix E.

To account for semantically similar themes with differing document assignments, we also measure semantic consistency using SBERT document embeddings (Reimers and Gurevych, 2019). We compute (1) centroid similarity, based on the cosine similarity between theme centroids, and (2) group average similarity, based on the average pairwise similarity across documents in two themes. As with Jaccard, we report maximum similarities per theme and average them for comparison across settings.

Table 1 reports average maximum Jaccard and embedding similarities across experiments. The high variance in Jaccard similarity across experiments underscores the importance of semantics-based metrics. For example, the topic model results on the COVID dataset show a high difference between collaboration settings: 0.56 average maximum Jaccard similarity in the synchronous experiment compared to 0.30 for the asynchronous one. However, the other two metrics show a much lower difference, suggesting a lower real difference in consistency.

For the COVID dataset, synchronous groups using the topic model and relational systems produced more consistent themes. The LLM-based system showed no statistically significant difference across collaboration settings and has less explanatory power. Results for the Climate dataset were less conclusive, likely due to greater semantic diversity and longer documents which may have encouraged synchronous coders to surface more varied themes through discussion; as shown in the next section, these themes are also more cohesive.

Notably, the LLM-based system afforded the fewest opportunities for user intervention during coding. Unlike the other systems, which supported operations such as splitting and merging topics or defining relations between concepts, the LLM system only allowed users to edit theme definitions. We hypothesize that richer intervention mechanisms better enable coders to leverage the deliberation inherent to synchronous collaboration.

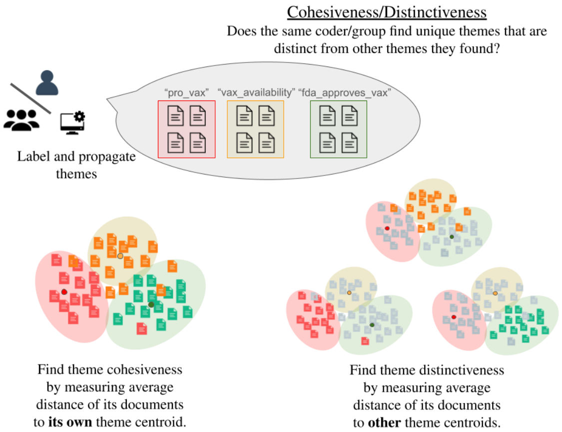


Figure 3: Once a coder has identified themes and they have been propagated the full dataset, we calculate *intra-theme similarity* by measuring the avg. of the pairwise distances between each document within a theme (left). We calculate *inter-theme similarity* by measuring the avg. of pairwise distances between each document in a theme and documents assigned to all other themes (right)

5.2 Cohesiveness and Distinctiveness

Another dimension for determining the systematicity and clarity of coding outcomes is by evaluating the similarities and differences between themes within the same code-book. We propose two metrics to measure this: cohesiveness and distinctiveness. A theme is said to be cohesive if its documents are similar to each other (measured by *intra*-theme similarity) and distinctive if it is dissimilar from documents in other themes within the same code-book (measured by *inter*-theme similarity). Intuitively, the purpose of grouping documents by theme is to create abstract representations of a dataset, where each theme represents a distinct facet of the data. If themes are not cohesive and distinctive, then it becomes hard to tell which theme a given document should belong to and the code-book falls apart.

Figure 3 shows how to evaluate these metrics for a single coder (or coder group). We calculate both the intra-theme similarity and the inter-theme similarity for all the themes in the code-book. *Intra-theme similarity* is calculated by taking the average of pair-wise similarity between all documents of the same theme. *Inter-theme similarity* for a given

			Topic Model		Relational		LLM-based	
			Intra-Theme	Inter-Theme	Intra-Theme	Inter-Theme	Intra-Theme	Inter-Theme
COVID	All	Sync	$0.52(0.10)$	$0.40(0.04)$	$\mathbf{0.51(0.08)}^*$	$0.42(0.05)^*$	$\mathbf{0.44(0.06)}$	$0.40(0.04)$
		Async	$0.52(0.10)$	$0.40(0.04)$	$0.45(0.10)^*$	$\mathbf{0.34(0.11)}^*$	$0.43(0.05)$	$\mathbf{0.39(0.04)}$
	Top 25%	Sync	$0.56(0.11)$	$0.39(0.05)$	$\mathbf{0.70(0.09)}^*$	$0.52(0.07)^*$	$0.63(0.07)$	$0.55(0.05)$
		Async	$0.56(0.11)$	$0.39(0.05)$	$0.64(0.09)^*$	$\mathbf{0.46(0.13)}^*$	$0.63(0.05)$	$\mathbf{0.54(0.05)}$
Climate	All	Sync	$\mathbf{0.57(0.23)}^*$	$0.25(0.08)^*$	$\mathbf{0.44(0.10)}^*$	$0.30(0.07)^*$	$\mathbf{0.39(0.11)}^*$	$0.29(0.05)^*$
		Async	$0.50(0.19)^*$	$\mathbf{0.24(0.07)}^*$	$0.43(0.09)^*$	$\mathbf{0.29(0.07)}^*$	$0.38(0.08)^*$	$\mathbf{0.28(0.05)}^*$
	Top 25%	Sync	$\mathbf{0.72(0.21)}^*$	$0.26(0.09)^*$	$\mathbf{0.66(0.11)}^*$	$0.39(0.10)^*$	$\mathbf{0.63(0.11)}^*$	$0.40(0.09)^*$
		Async	$0.65(0.20)^*$	$0.26(0.08)^*$	$0.65(0.10)^*$	$0.39(0.11)^*$	$0.62(0.12)^*$	$0.40(0.09)^*$

Table 2: Group Avg. Similarity *within* Coder Groups (standard deviation in brackets). Themes are considered to be more **cohesive** if intra-theme similarity is high and more **distinctive** if inter-theme similarity is low. **Bold** figures highlight more cohesive *or* distinctive value between collaboration settings. *Statistically significant using a two-sample unpaired t-test with $p < 0.05$.

theme is calculated by taking the average pair-wise similarity of documents in that theme with documents in all other themes.

A confounding factor in these measures is that all systems provide broad coverage of documents such that even distantly related documents may be assigned to a theme. To more accurately represent the cohesiveness and distinctiveness of the themes in each experiment, we perform the same calculations on a subset comprised of only the top 25% of documents most closely related to each theme. For the relational and LLM-based systems, this top quartile is selected using the distance from the centroid. For the interactive topic model, we use the weights assigned by the model.

Table 2 shows results for both the whole dataset as well as the subset of the documents closest to each theme. Overall, we find that the intra-theme similarities are always higher than inter-theme similarities, which means that themes are at least moderately cohesive and distinctive across the board. For the COVID dataset, we find that themes may be more cohesive but not more distinctive in the synchronous setting especially for the relational system. This may be due to the homogeneity of the dataset, which further explains the uniformity of results in the topic model and LLM experiments.

The results from the Climate dataset present a more compelling finding, especially since all differences are statistically significant, with the synchronous experiments uniformly producing much more cohesive themes. When only the top 25% of documents are considered, the increase in cohesiveness is not correlated with a decrease in distinctiveness, unlike in the COVID case. This is best exemplified by the topic model results, where the synchronous experiment achieving 0.72 on intra-theme similarity, much higher than the 0.65 achieved by the asynchronous experiment despite both experiments producing the same inter-theme similarity. Our results strongly suggest that synchronous collaboration facilitates deeper analysis to uncover clearer themes in the data. This is further supported by the greater distinctiveness in the Climate data experiments: we expect coders to find more distinct themes in a more complex and heterogenous dataset.

5.3 Correctness

Interactive systems allow users to automate large portions of the coding process at the risk of producing inaccurate theme assignments. To estimate how correct the outputs of each system are, we conduct a post-hoc analysis by manually checking a randomly selected sample of $2,400$ document-theme pairs (200 per experimental setting). To ensure the representativeness of our samples, we split the data into quartiles based on document relatedness to each theme and select a uniform sample of documents per theme. As before, relatedness is calculated using the theme weight distribution for the interactive topic model and distance from the centroid for the the other two systems. To assess reliability, each assignment is evaluated by two annotators, with a third weighing in for tie breaks. The human evaluators demonstrate moderate-to-high agreement, with a Krippendorff's α of 0.632 for the COVID tweets dataset and 0.696 for the Climate ads dataset, suggesting that we can trust these estimations.

Figure 4 shows the correctness results for each quartile sample per experiment. For the COVID dataset, we observe that the relational system is the most accurate and shows negligible differences

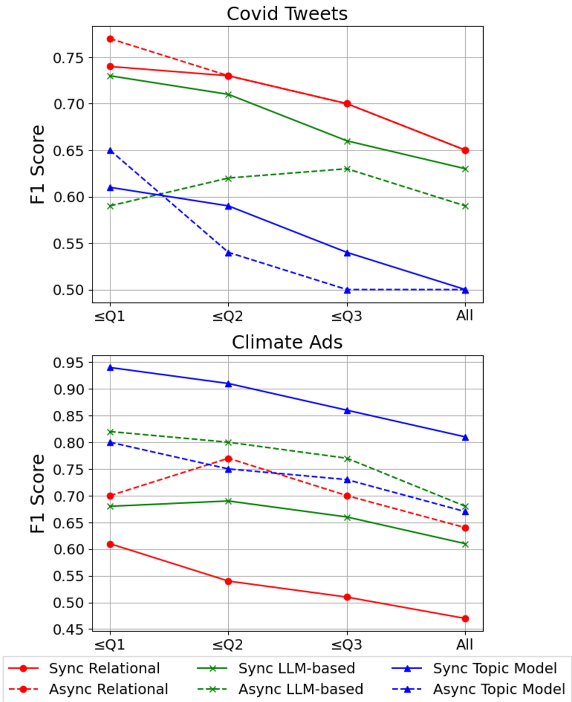


Figure 4: Correctness w.r.t. distance from theme.

across collaboration settings. This an encouraging result, given that this system took the most advantage of synchronous deliberation based on our other metrics of quality. The other two systems showed marked differences, with the topic modeling approach producing more accurate assignments in asynchronous operation, and the LLM producing more accurate assignment in a synchronous paradigm.

On the other hand, the Climate dataset shows almost the opposite result, with the synchronous relational experiment producing the least accurate results and the synchronous topic model significantly outperforming all the other experiments. While there is no strong relationship between collaboration setting and correctness scores, our findings suggest that dataset characteristics may significantly impact the performance of interactive systems. The Climate dataset has almost double the number of words per document compared to the COVID dataset (53.1 vs. 27.5), which may allow the topic model to capture more meaningful lexical statistics. The Climate dataset is also only a fraction of the size of the COVID dataset (5,471 documents vs. 85,799 documents), which means that topic model users see a much greater proportion of the documents and make more accurate refinement decisions.

5.4 Impact of Dataset Differences

To investigate how structural and semantic characteristics of a dataset affect the coding process, we carried out a comprehensive analysis across multiple dimensions, including readability, templatic or formulaic patterns, semantic theme ambiguity, and semantic similarity distribution (details in Appendix C). The higher formulaic content in climate ads (20.1% vs. 3.7%) may facilitate easy pattern recognition, but could also induce coder fatigue through repetitive content. The greater linguistic complexity (Flesch Reading Ease 47.4 vs. 54.9; documents averaging 53.1 vs. 27.5 words) and semantic diversity (mean pairwise similarity 0.26 vs. 0.43) of climate ads likely impose higher cognitive load per coding decision.

Conversely, the semantic homogeneity of COVID tweets, evidenced by higher pairwise similarity (0.43 vs. 0.26) and centroid proximity (0.65 vs. 0.51), gives coders shared semantic reference points that can support more efficient real-time discussion in synchronous settings. The less formulaic nature of COVID discourse may further sustain coder engagement through content variety. These dynamics are consistent with our finding that average consistency scores favor synchronous groups for COVID (Table 1), while for climate ads, asynchronous coders perform on par with or slightly above synchronous groups. Together, these results suggest that coding protocol selection should account for corpus-specific characteristics: semantically diverse, linguistically complex corpora may benefit from asynchronous approaches that allow extended reflection, while homogeneous corpora can be efficiently processed through synchronous collaborative coding.

6 User Study and Recommendations

We conducted semi-structured interviews to understand the participant experiences with the task and tools, with a focus on synchronous versus asynchronous coding (see Appendix F for interview script). Several findings emerged.

Synchronous teamwork eased coding and improved outcomes. Participants reported that working synchronously helped them contextualize data, resolve disagreements quickly, and "break ties" through discussion. These experiences align with our quantitative findings showing higher consistency and cohesiveness in synchronous settings, suggesting the value of systems that explicitly support real-time deliberation.

Asynchronous coders were more sensitive to tool limitations. Because they worked largely in isolation, asynchronous participants focused more on usability issues and tool constraints, highlighting the need to improve support for independent coding workflows.

Limited control reduced user trust, particularly in topic modeling. Participants reported a loss of agency when using the topic modeling system, citing insufficient control over operations and difficulty tracking theme evolution – *'the merge process did not offer the ideal amount of control and made it difficult to keep track of the theme groups."*

Initial topics also sometimes conflated opposing themes due to lexical similarity, frustrating users' attempts to refine results – *"Many Anti-Vax and Pro-Vax standpoints use the same words/phrases in their tweets, which the Fang et al. (2023) model groups together despite the stark difference in message between the two."*

While some appreciated the model's initial theme induction, participants desired greater control and clearer explanations of cluster structure. These findings highlight the need for systems to maximize the degree of control afforded to users.

LLM-based coding was costly and unreliable at scale. Despite strong reasoning capabilities, LLMs proved inefficient for large-scale annotation due to high computational costs and inconsistent classification performance. This opens an opportunity for NLP researchers to make LLMs more reliable inductive reasoners and to come up with prompting strategies that can allow LLMs to reliably classify documents in bulk, especially when working at scale.

NLP researchers should consider ideal collaboration settings and dataset characteristics when designing tools. Our results show that these factors have outsized impact on coding outcomes, and that no single configuration works best across conditions. Interactive systems should therefore balance automation with user control and offer tailored workflows that support both real-time deliberation and independent coding at scale.

7 Conclusion and Future Work

We examined three categories of NLP-assisted qualitative research tools across synchronous and asynchronous collaboration settings, applying them to two structurally distinct English corpora. We designed an evaluation framework that captures theme consistency, cohesiveness, distinctiveness, and correctness, and used it to analyze how coding outcomes vary across methods, settings, and data.

We find that collaboration modality significantly affects output quality, but the direction of the effect depends on both the system and the dataset. On homogeneous corpora, synchronous collaboration produces more consistent and cohesive themes. On heterogeneous corpora, synchronous coding yields markedly more cohesive themes without sacrificing distinctiveness, while consistency scores are comparable across settings. We further observe that systems affording richer user intervention benefit more from synchronous deliberation, whereas systems with limited intervention show little sensitivity to collaboration setting.

Beyond collaboration, we find that dataset characteristics independently shape system performance, as the same system can be the strongest or among the weakest depending on the corpus. Topic modeling, for instance, achieves the highest correctness on Climate ads in the synchronous setting but produces highly divergent code-books across asynchronous coders on COVID. LLM-based solutions show promise but remain costly and inconsistent at scale, and offer fewer intervention points for coders to leverage deliberation.

While this study focuses on collaboration modalities, there are numerous other variables that can affect a tool's efficacy for qualitative coding. We believe that our proposed evaluation framework can be repurposed and expanded to evaluate a wide range of interventions, such as the underlying NLP technology, the interactive interface, the expertise of the coders, and the type of data being annotated. Future work should focus on evaluating other interactive systems to determine whether our findings are representative of relationships between these variables. We hope to inform the development of more robust evaluations of NLP tools for qualitative research in realistic settings.

Limitations

The study presented in this paper has two main limitations.

(1) While we selected three distinct, representative tools to perform our analysis of synchronous vs. asynchronous settings, as well as two datasets with distict characteristics, the list is of course non-exhaustive. A larger study incorporating more tools and datasets could yield additional insights.

(2) While we look at an important variable in qualitative research settings (collaboration modality), there are several other variables that can influence the outcome of NLP-assisted solutions (e.g., choice of tool, expertise and live experience of annotators, type of data being annotated, etc.). In addition to this, we did not explore the many different consolidation strategies that are often used to bring together the perspectives of asynchronous coders. We leave the explorations of these questions for future work.

Ethical Considerations

To the best of our knowledge, no code of ethics was violated during the development of this project. We used publicly available tools and datasets according to their licensing agreements. For our annotation experiments, we followed IRB protocol and did not retain any personally identifiable information.

All information needed to replicate our experiments is presented in the paper. We reported all experimental settings, as well as any pre-processing steps, learning configurations, hyper-parameters, and additional technical details. Due to space constraints, some of this information was relegated to the Appendix. In addition to this, we have made the results of the annotation experiment available to the community, as well as the code to produce all of our reported results. We believe that the results reported in this paper support our claims.

Acknowledgments

This work utilized computing resources at the University of Colorado Boulder, including the CUmulus on-premise cloud service and the Blanca condo computing resource. CUmulus is jointly funded by the National Science Foundation (award OAC-1925766) and the University of Colorado Boulder. Blanca is jointly funded by its computing users and the University of Colorado Boulder.

References

Christian Baden, Christian Pipal, Martijn Schoonvelde, and Mariken A. C. G van der Velden. 2022. Three gaps in computational text analysis methods for social sciences: A research agenda. *Communication Methods and Measures*, 16(1):1–18.

Shai Ben-David and Margareta Ackerman. 2008. Measures of clustering quality: A working set of axioms for clustering. In *Advances in Neural Information Processing Systems*, volume 21. Curran Associates, Inc.

Henry E. Brady. 2019. The challenge of big data and data science. *Annual Review of Political Science*, 22(1):297–323.

Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative Research in Psychology*, 3:77–101.

Nan-Chen Chen, Margaret Drouhard, Rafal Kocielnik, Jina Suh, and Cecilia R. Aragon. 2018. Using machine learning to support qualitative coding in social science: Shifting the focus to ambiguity. *ACM Trans. Interact. Intell. Syst.*, 8(2).

Robert Chew, John Bollenbacher, Michael Wenger, Jessica Speer, and Annice Kim. 2023. Llm-assisted content analysis: Using large language models to support deductive coding. *Preprint*, arXiv:2306.14924.

Jaegul Choo, Changhyun Lee, Chandan K. Reddy, and Haesun Park. 2013. Utopian: User-driven topic modeling based on interactive nonnegative matrix factorization. *IEEE Transactions on Visualization and Computer Graphics*, 19(12):1992–2001.

Jason Chuang and Daniel A. McFarland. 2013. Document exploration with topic modeling : Designing interactive visualizations to support effective analysis workflows.

Flora Cornish, Alex Gillespie, and Tania Zittoun. 2014. *The SAGE Handbook of Qualitative Data Analysis*. SAGE Publications Ltd, London; London.

Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging large language model for thematic analysis. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9993–10001, Singapore. Association for Computational Linguistics.

Margaret Drouhard, Nan-Chen Chen, Jina Suh, Rafal Kocielnik, Vanessa Peña-Araya, Keting Cen, Xiangyi Zheng, and Cecilia R. Aragon. 2017. Aeonium: Visual analytics to support collaborative qualitative coding. In *2017 IEEE Pacific Visualization Symposium (PacificVis)*, pages 220–229.

Zheng Fang, Lama Alqazlan, Du Liu, Yulan He, and Rob Procter. 2023. A user-centered, interactive, human-in-the-loop topic modelling system. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 505–522, Dubrovnik, Croatia. Association for Computational Linguistics.

Zheng Fang, Yulan He, and Rob Procter. 2021. A query-driven topic model. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1764–1777, Online. Association for Computational Linguistics.

Jessica L. Feuston and Jed R. Brubaker. 2021. Putting tools in their place: The role of time and perspective in human-ai collaboration for qualitative analysis. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW2).

Uwe Flick. 2014. The sage handbook of qualitative data analysis.

Jie Gao, Kenny Tsu Wei Choo, Junming Cao, Roy Ka-Wei Lee, and Simon Perrault. 2023. Coaicoder: Examining the effectiveness of ai-assisted human-to-human collaboration in qualitative analysis. *ACM Trans. Comput.-Hum. Interact.*, 31(1).

Jie Gao, Yuchen Guo, Gionnieve Lim, Tianqin Zhang, Zheng Zhang, Toby Jia-Jun Li, and Simon Tangi Perrault. 2024. Collabcoder: A lower-barrier, rigorous workflow for inductive collaborative qualitative analysis with large language models. *Preprint*, arXiv:2304.07366.

Barney G Glaser, Anselm L Strauss, and Elizabeth Strutzel. 1968. The discovery of grounded theory; strategies for qualitative research. *Nursing research*, 17(4):364.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, and 542 others. 2024. The llama 3 herd of models. *Preprint*, arXiv:2407.21783.

Kilem Li Gwet. 2008. Computing inter-rater reliability and its variance in the presence of high agreement. *British Journal of Mathematical and Statistical Psychology*, 61(1):29–48.

Martin Hilbert, George Barnett, Joshua Blumenstock, Noshir Contractor, Jana Diesner, Seth Frey, Sandra González-Bailón, PJ Lamberson, Jennifer Pan, Tai-Quan Peng, Cuihua (Cindy) Shen, Paul E. Smaldino, Wouter van Atteveldt, Annie Waldherr, Jingwen Zhang, and Jonathan J. H. Zhu. 2019. Computational communication sciencel computational communication science: A methodological catalyzer for a maturing discipline. *International Journal of Communication*, 13(0).

Enamul Hoque and Giuseppe Carenini. 2016. Interactive topic modeling for exploring asynchronous online conversations: Design and evaluation of convisit. *ACM Trans. Interact. Intell. Syst.*, 6(1).

Alexander Hoyle, Pranav Goel, Denis Peskov, Andrew Hian-Cheong, Jordan Boyd-Graber, and Philip Resnik. 2021. Is automated topic model evaluation broken? the incoherence of coherence. In *Proceedings of the 35th International Conference on Neural Information Processing Systems*, NIPS '21, Red Hook, NY, USA. Curran Associates Inc.

Tunazzina Islam, Ruqi Zhang, and Dan Goldwasser. 2023. Analysis of climate campaigns on social media using bayesian model averaging. In *Proceedings of the 2023 AAAI/ACM Conference on AI, Ethics, and Society*, AIES '23, page 15–25, New York, NY, USA. Association for Computing Machinery.

Jialun Aaron Jiang, Kandrea Wade, Casey Fiesler, and Jed R. Brubaker. 2021. Supporting serendipity: Opportunities and challenges for human-ai collaboration in qualitative analysis. *Proc. ACM Hum.-Comput. Interact.*, 5(CSCW1).

Xin Jin and Jiawei Han. 2010. *K-Means Clustering*, pages 563–564. Springer US, Boston, MA.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2024. Large language models are zero-shot reasoners. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Julia Mendelsohn, Ceren Budak, and David Jurgens. 2021. Modeling framing in immigration discourse on social media. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2219–2263, Online. Association for Computational Linguistics.

Maria Leonor Pacheco, Tunazzina Islam, Monal Mahajan, Andrey Shor, Ming Yin, Lyle Ungar, and Dan Goldwasser. 2022. A holistic framework for analyzing the COVID-19 vaccine debate. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 5821–5839, Seattle, United States. Association for Computational Linguistics.

Maria Leonor Pacheco, Tunazzina Islam, Lyle Ungar, Ming Yin, and Dan Goldwasser. 2023. Interactive concept learning for uncovering latent themes in large text collections. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 5059–5080, Toronto, Canada. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Tim Rietz, Peyman Toreini, and Alexander Maedche. 2020. Cody: An interactive machine learning system for qualitative coding. In *Adjunct Proceedings* *of the 33rd Annual ACM Symposium on User Interface Software and Technology*, UIST '20 Adjunct, page 90–92, New York, NY, USA. Association for Computing Machinery.

Shamik Roy, Maria Leonor Pacheco, and Dan Goldwasser. 2021. Identifying morality frames in political tweets using relational learning. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 9939–9958, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Alison Smith, Varun Kumar, Jordan Boyd-Graber, Kevin Seppi, and Leah Findlater. 2018. Closing the loop: User-centered design and evaluation of a human-in-the-loop topic modeling system. In *Proceedings of the 23rd International Conference on Intelligent User Interfaces*, IUI '18, page 293–304, New York, NY, USA. Association for Computing Machinery.

Laurens van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of Machine Learning Research*, 9(86):2579–2605.

Ziang Xiao, Xingdi Yuan, Q. Vera Liao, Rania Abdelghani, and Pierre-Yves Oudeyer. 2023. Supporting qualitative analysis with large language models: Combining codebook with gpt-3 for deductive coding. In *Companion Proceedings of the 28th International Conference on Intelligent User Interfaces*, IUI '23 Companion, page 75–78, New York, NY, USA. Association for Computing Machinery.

Himanshu Zade, Margaret Drouhard, Bonnie Chinh, Lu Gan, and Cecilia Aragon. 2018. Conceptualizing disagreement in qualitative coding. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, page 1–11, New York, NY, USA. Association for Computing Machinery.

A Relational Approach Operational Details

Table 3 details all operations available to participants using the Pacheco et al. (2023) system.

Operations	Description
Finding Partitions	Experts can find partitions in the space of unassigned instances. We currently support the K-means (Jin and Han, 2010) and Hierarchical Density-Based Clustering (Mendelsohn et al., 2021) algorithms.
Text-based Queries	Experts can type any query in natural language and find instances that are close to the query in the embedding space.
Finding Similar Instances	Experts have the ability to select each instance and find other examples that are close in the embedding space.
Listing Themes and Instances	Experts can browse the current list of themes and their mapped instances. Instances are ranked in order of "goodness", corresponding to the similarity in the embedding space to the theme representation. They can be listed from closest to most distant, or from most distant to closest.
Visualizing Local Explanations	Experts can visualize aggregated statistics and explanations for each of the themes. To obtain these explanations, we aggregate all instances that have been identified as being associated with a theme. Explanations include wordclouds, frequent entities and their sentiments, and graphs of concept distributions.
Visualizing Global Explanations	Experts can visualize aggregated statistics and explanations for the global state of the system. To do this, we aggregate all instances in the database. Explanations include theme distribution, coverage statistics, and t-sne plots (van der Maaten and Hinton, 2008).

(a) Exploratory Operations

Operations	Description
Adding, Editing and Removing Themes	Experts can create, edit, and remove themes. The only requirement for creating a new theme is to give it a unique name. Similarly, themes can be edited or removed at any point. If any instances are assigned to a theme being removed, they will be moved to the space of unassigned instances.
Adding and Removing Examples	Experts can assign "good" and "bad" examples to existing themes. Good examples are instances that characterize the named theme. Bad examples are instances that could have similar wording to a good example, but that have different meaning. Experts can add examples in two ways: they can mark mapped instances as "good" or "bad", or they can directly contribute example phrases.
Adding or Correcting Concepts	We allow users to upload additional observed or predicted concepts for each textual instance. For instances and phrases added as "good" and "bad" examples, we allow users to add or edit the values of these concepts. The intuition behind this operation is to collect additional information for learning to map instances to themes.

(b) Intervention Operations

Table 3: Interactive Operations for the Pacheco et al. (2023) System

B LLM-based Experimental Configuration and Prompt Details

We used the Llama 3.2 3B-Instruct (Grattafiori et al., 2024) model with a batch size of 32 for all generation tasks. In total, we ran 10 jobs at an average compute time of 24 hours per job using an A100 40GB VRAM GPU.

COVID dataset		
code-book	Gwet's AC_1	# Unlabeled Docs
Sync 1	0.42	$5,548\ (6.5\%)$
Sync 2	0.49	$9,766\ (11.4\%)$
Async 1	0.61	$611\ (0.7\%)$
Async 2	0.62	$2,506\ (2.9\%)$
Async 3	0.46	$13,816\ (16.1\%)$
Climate dataset		
code-book	Gwet's AC_1	# Unlabeled Docs
Sync 1	0.38	$291\ (5.3\%)$
Sync 2	0.33	$372\ (6.8\%)$
Async 1	0.38	$150\ (2.7\%)$
Async 2	0.39	$277\ (5.1\%)$
Async 3	0.25	$120\ (2.2\%)$

Table 4: Results for the selected prompt for each coding session using the LLM-based system. Gwet's AC_1 is used to select the best prompt for running the full dataset. The number of unlabeled documents represent documents where the LLM produced a label not created by human annotators after running the full dataset (percentage of the dataset unlabeled).

```
To code this tweet, do the following:
- First, read the codebook and the
tweet.
- Next, decide which code is most
applicable and explain your reasoning
for the coding decision.
- Finally, generate json with your
code and your reason for the coding
decision.  The response MUST be
formatted as JSON.
Codes:
-
<codes>
-
Codebook:
-
<codebook>
-
Tweet:
-
<tweet>
-
JSON Output:
-
"code" :  "",
"reason" :  ""
-
```

Figure 5: Prompt Template 1 for LLM-based experiments.

B.1 LLM Prompts

To choose the best prompt for labeling the full dataset, we run preliminary experiments with each code-book to calculate human-model annotator agreement. The model is prompted to label all of the documents already labeled by the human and agreement is calculated using Gwet's AC_1, as shown in Table 4. Figure 5 is the prompt template adapted from Chew et al. (2023) while figures 6, 7, and 8 are additional templates created based on the first template.

C Dataset Analysis

The two corpora exhibit substantial structural differences as seen in Table 5. The climate dataset comprises 5,471 documents with a mean length of 53.1 words, while the COVID dataset contains 85,799 documents averaging 27.5 words. This approximately two-fold difference in document length is statistically significant (Mann-Whitney U, p<0.001, effect size r=-0.353). Climate advertisements also demonstrate greater variability in length (CV=1.47) compared to the more constrained COVID tweets (CV=0.50). Lexical diversity metrics reveal nuanced differences. While type-token ratios are comparable (0.060 vs. 0.051), the Measure of Textual Lexical Diversity (MTLD) indicates higher lexical sophistication in COVID tweets (119.6 vs. 95.6).

```
To code this tweet, do the following:

First, read the codebook and the
tweet.  Next, decide which code is
most applicable based on the tweet's
content and explain your reasoning
for the coding decision.  Finally,
generate a JSON object with the
selected code and provide a brief
explanation for your coding decision.
The response MUST be formatted as
JSON.
Codebook:
Themes:  <"theme":  "definition">
Tweet:  < "text":  "<text>" >
JSON Output:  < "code":  "", "reason":
"" >
```

Figure 6: Prompt Template 2 for LLM-based experiments.

Metric	Covid	Climate
Number of Documents	85,799	5,471
Words per Document	27.50	53.10
Sentences per Document	2.50	3.40
Type-Token Ratio	0.051	0.060
Measure of Textual Lexical Diversity (threshold = 0.72)	119.65	95.58

Table 5: Corpus-level statistics for the Covid and Climate datasets.

Readability analyses indicate that climate advertisements present greater cognitive demands. The Flesch Reading Ease score for climate ads (M=47.4) falls within the “difficult” range, whereas COVID tweets score higher (M=54.9), indicating moderately easier comprehension (p<0.001, r=0.211). Correspondingly, climate ads require higher grade-level reading ability (Flesch-Kincaid Grade: 10.7 vs. 9.2; Gunning Fog Index: 13.2 vs. 11.2). See Table 6.

We also performed an analysis over a sample of 5000 documents from both datasets to identify the extent of formulaic or template centric content. DBSCAN clustering on TF-IDF similarity matrices (threshold=0.8) identified 137 template clusters in climate ads, encompassing 20.1% of documents, compared to only 31 clusters (3.7%) in COVID tweets. The proportion of high-similarity docu-

```
To generate code for this tweet,
provide a step-by-step explanation
of how to approach the task.

First, analyze the tweet's content
and identify key concepts, such as
the type of object or class being
described, any specific behaviors or
requirements, and relevant keywords.
Next, evaluate the codebook options
and determine which one is most
applicable.  Explain your reasoning
for your decision, including any
similarities between the tweet and
the code definitions, or any specific
requirements mentioned in the tweet
that align with a particular code.
Finally, generate a JSON object
with the selected code and provide
additional context, including:
- A clear explanation of how you
arrived at your chosen code
- Any relevant notes or comments
about the code's functionality and
requirements
- A brief comparison to other codes in
the book, if applicable
The response MUST be formatted as
JSON.
Codebook:  <codebook>
Tweet:  <tweet>
JSON Output:  < "code":  "",
"reasoning":  "", "context":  ""
>
```

Figure 7: Prompt Template 3 for LLM-based experiments.

```
To analyze this tweet and select a
relevant theme, follow these steps:

First, read the tweet and identify
key concepts, such as emotions,
objects, or ideas mentioned in the
text.
Next, evaluate the theme options
and determine which one is most
applicable.  Explain your reasoning
for your decision, including any
connections you see between the
tweet's content and the theme
definitions.
Then, generate a JSON object with the
selected theme and provide additional
insight into your analysis.  Include:
A clear explanation of how you
arrived at your chosen theme
Any specific characteristics or
keywords from the tweet that support
your decision
A brief comparison to other themes,
if applicable
The response MUST be formatted as
JSON.
Themes:  <Codebook>
Tweet:  <tweet>
JSON Output:  <"theme":  "",
"insight":  "">
```

Figure 8: Prompt Template 4 for LLM-based experiments.

ment pairs (>=0.8 cosine similarity) is an order of magnitude greater in climate ads (0.059% vs. 0.006%). See Table 7.

Characteristic n-grams reflect domain-specific discourse patterns. Climate ads prominently feature phrases such as “oil natural gas”, “clean energy jobs”, and “fight climate change”, while COVID tweets center on vaccine-related language (“covid 19 vaccine”, “getting covid vaccine”). These patterns suggest that climate advertising employs standardized messaging templates, whereas COVID discourse largely centers around vaccines. See Table 8.

Gaussian Mixture Model analysis with SBERT embeddings (Reimers and Gurevych, 2019) reveals that both corpora exhibit high thematic clarity, with mean maximum cluster probabilities exceeding 0.98. However, COVID tweets demonstrate marginally higher semantic entropy (M=0.042 vs. 0.023), indicating slightly greater theme ambiguity. The proportion of documents spanning multiple clusters (probability difference < 0.2 between top-2 clusters) is higher for COVID tweets (0.66% vs. 0.28%). Silhouette scores are low for both corpora (0.022 vs. 0.018), suggesting that while documents cluster clearly into dominant themes, inter-cluster boundaries are not sharply defined. See Table 9.

Mean pairwise cosine similarity in the SBERT embedding space is substantially higher for COVID tweets (0.429 vs. 0.261), indicating greater semantic homogeneity. The inter-quartile range confirms this pattern: COVID tweets exhibit similarity scores between 0.365 – 0.518, while climate ads range from 0.167 – 0.348. Notably, 62.6% of climate ad pairs fall below 0.3 similarity, compared to only 14.1% of COVID tweet pairs. Document-to-centroid similarity further corroborates this finding (0.654 vs. 0.510), demonstrating that COVID tweets cluster more tightly around their corpus centroid. See Table 10.

Metric	Covid	Climate
Flesch Reading Ease Score	54.85	47.38
Flesch Kincaid Grade	9.18	10.68
Gunning Fog Index	11.20	13.22
Automated Readability Index	12.34	11.85

Table 6: Measures of Linguistic Complexity of the Covid and Climate datasets.

Metric	Covid	Climate
Template Clusters	31	137
Percentage of Documents in Template Clusters	3.70	20.10
Percentage of High Similarity Document Pairs	0.006	0.059
Average Cluster Size	5.96	7.34
Largest Cluster Size	31	75

Table 7: Template patterns across Covid and Climate datasets.

Metric	Covid	Climate
Average Pairwise Similarity between documents	0.43	0.26
25th Percentile Pairwise Similarity Score	0.37	0.17
75th Percentile Pairwise Similarity Score	0.52	0.35
Average Similarity of documents with the centroid	0.65	0.51
Percentage of Pairs with Similarity > 0.7	0.70	0.20
Percentage of Pairs with Similarity > 0.3	14.14	62.55

Table 10: Semantic Similarity Distribution for Covid and Climate datasets.

D Themes by Code-book

Tables 11 through 16 list the final themes from every experiment, organized by dataset and model. Each table covers one (dataset, model) combination and reports themes across the three asynchronous and two synchronous sessions. Theme counts are summarized in Table 17, and themes are reported verbatim as coders named them.

Covid	Climate
'covid 19 vaccine', 'getting covid vaccine', 'covid vaccine https', 'got covid vaccine', 'covid 19 vaccines'	*'https bit ly', 'oil natural gas', 'clean energy jobs', 'fight climate change', 'oil gas industry'*

Table 8: Top n-gram phrases in Covid and Climate datasets.

Metric	Covid	Climate
Average Entropy across Documents	0.042	0.023
Average Confidence in Primary Cluster Assignment	0.989	0.994
Percentage of Documents Where Top 2 Clusters are within 0.2 Probability	0.660	0.280
Silhouette Score	0.022	0.018

Table 9: Semantic theme ambiguity for Covid and Climate datasets.

E Example Heatmaps of Synchronous Code-book Jaccard Similarities

Heatmaps of the Jaccard similarity of themes from the synchronous experiments can be seen in figures 9, 10, and 11. Each asynchronous experiment has three code-books so heatmaps between two code-books would not accurately reflect the Jaccard metrics reported in the paper.

Async 1	Async 2	Async 3	Sync 1	Sync 2
1. Pro Trump and against Biden/Obama 2. Dispelling misinformation regarding death 3. Vaccine availability 4. 2nd dose side effects 5. Comparisons to other virus vaccines 6. Anger over mask mandates 7. Reminders of the 2nd vaccine 8. Origin of the vaccine 9. Discussing vaccines in schools 10. Positive outcomes of vaccine 11. Concerns regarding Johnson and Johnson vaccine 12. CVS job openings 13. Information aimed towards minority communities 14. Against anti-vaxxers 15. Calling out liars 16. Free food for vaccinated people 17. Vaccine card mandates 18. Vaccine hesistancy 19. Vaccine appointment info	1. VaccineRolloutNews 2. AntiTrump 3. VaccineSideEffects 4. VaccineTestScience 5. HopefulPeopleGetVaccinated 6. WhereToGetVaccine 7. FDANews&Discourse 8. VaccinePassportNews 9. FakeNews 10. VaccineAdvocacy 11. AntiRepublican 12. AntiDemocrat 13. PandemicNews 14. VaccineHesitancy 15. POCVaccineAwareness 16. FightingDisinformation 17. VaccineChildRisk 18. VaccineProblemsNews	1. Anti-Biden and Pro-Trump Sentiment 2. Comparing the COVID Vaccine to other Deadly Disease Vaccines (polio, measles) 3. Anger towards COVID Deniers / COVID Deniers Passing Away 4. Immediate After-Effects from COVID Vaccine (soreness, nausea) 5. Family Members Receiving the COVID Vaccine 6. Vaccine Eligibility Updates 7. Proof of Vaccination (vaccine card/passport) 8. Adverse Childhood Experience's being caused by COVID 9. COVID Vaccine causing Blood Clots 10. Unknown Long-Term Effects of COVID Vaccine 11. Become a Pharmacy Technician Alerts 12. COVID FDA Approval Status 13. Anger Towards Republicans Lying about COVID 14. Giving Info about COVID and the Vaccine to Minorities in America 15. Walgreens Vaccine Appointment Updates	1. trump supporters saying biden is wrongfully taking credit for vaccine 2. side effects of vaccine 3. announcing vaccine eligibility 4. wholesome reactions to getting the vaccine 5. inefficacy of vaccine 6. corelation between trauma and covid 7. halting vaccination due to adverse effects 8. proof of vaccine 9. appeal to get vaccinated 10. covid related deaths 11. advertising for covid related jobs 12. freebies with proof of vaccine 13. concerns about effects of vaccines 14. disproportionate impact and equitable recover across racial lines 15. political conspiracy / nicki minaj controversy 16. vaccine appointments	1. criticism of governments' response to vaccine 2. family covid vaccine experience 3. immediate personal side effects 4. state vaccine eligibility 5. non-covid vaccine comparision 6. institutional adverse childhood experiences 7. vaccine certification 8. anti-vaccine conspiracy outcomes 9. pharmacy hiring 10. FDA vaccine approval 11. long term side effects 12. community vaccine equity 13. vaccine side effects / Nicki Minaj 14. covid appointment

Table 11: Final codebooks for the COVID Tweets dataset using the Topic Model approach.

Async 1	Async 2	Async 3	Sync 1	Sync 2
1. Receiving first dose of covid vaccine 2. Criticising political figures 3. Promoting covid related news articles 4. Vaccine accessibility and distribution 5. Information about covid vaccine availability and appointments 6. Encouraging people to get the covid vaccine 7. The vaccine does not work because you can still get covid 8. Skepticism over the covid vaccine 9. Experiences of vaccine side effects 10. Praising frontline healthcare workers 11. Saying that you can still get covid if you have the vaccine 12. Receiving second vaccine dose 13. Encouraging getting the vaccine 14. Criticising The President 15. FDA Approval of covid vaccine 16. Providing resources to those who are not involved in the covid debate 17. Praising Government Leadership 18. Against covid disbelievers 19. Stating that the vaccine does not cause covid	1. GotVaccinated 2. AntiRepublican 3. WhereToGetVaccine 4. AdvocateForVaccine 5. VaccineSymptomReport 6. VaccineEfficacyDenial 7. VaccineRefusal 8. VaccineKills 9. AntiDemocrat 10. BlameFauci	1. ReceivedFirstDoseOfCovidVaccine 2. RepublicansDownplayingCovid 3. FdaApprovesCovidVaccine 4. GettingTheCovidVaccine 5. NewsArticlesAboutCovidVaccineProgress 6. ProCovidVaccine 7. MentionsAMayor 8. NewsAgenciesReportingOnTheCovidVaccine 9. AntiCovidVaccine 10. CovidVaccineAftereffects 11. GetYourCovidVaccine 12. BidenAdministrationEnforcingCovidVaccine 13. CovidVaccineOnlyReducesTheEffectsOfCovid 14. HealthcareWorkers 15. AntiBiden 16. NegativeViewOnTrump 17. RepublicansResponsibleForCovidDeaths 18. IsraelOutbreak	1. VaxSymptoms 2. GovGoodPolicies 3. VaxAppointmentInfo 4. VaxApprovalInfo 5. VaxDoesntWork 6. UnjustifiedFearOfVax 7. IGotTheVax 8. GovBadPolicies 9. VaxLessensSymptoms	1. PostVaxSymptoms 2. ReasonsForUSLaggingOnVaccines 3. VaxDistributionIssueDueToLocalPolicy 4. VaxAvailabilityInfo 5. #IGotMyVaccine 6. VaxDoesMoreHarmThanGood 7. VaxLessensSymptoms 8. FDAapproval

Table 12: Final codebooks for the COVID Tweets dataset using the Relational approach.

Async 1	**Async 2**	**Async 3**	**Sync 1**	**Sync 2**
1. Receiving first dose of covid vaccine	1. GotVaccinated	1. received_vax	1. GetAVaccine	1. vaccine_get_a_vaccine
2. Praising Government Leadership	2. BlameFauci	2. negative_discourse_around_politicians	2. CovidLiberal	2. vaccine_politics
3. Providing resources to those who are not involved in the covid debate	3. VaccineRefusal	3. vax_appointment_availability	3. ScrewYouGovernment	3. vaccine_rollout
4. Vaccine accessibility and distribution	4. WhereToGetVaccine	4. vax_approved	4. CovidDebunking	4. vaccine_efficacy
5. Praising frontline healthcare workers	5. VaccineEfficacyDenial	5. vax_reduces_fatality	5. GoodNewsAboutVaccines	5. vaccine_first_dose
6. Promoting covid related news articles	6. AdvocateForVaccine	6. positive_discourse_around_politicians	6. VaccineAppointmentAvailable	6. vaccine_symptoms
7. Information about covid vaccine availability and appointments	7. VaccineSymptomReport	7. news_about_vax	7. GotFirstDose	7. online_references
8. Stating that the vaccine does not cause covid	8. AntiDemocrat	8. anti_vax	8. JudgingUnvaccinated	
9. Skepticism over the covid vaccine	9. AntiRepublican	9. encourage_getting_vax	9. ThankYouGovernment	
10. Criticising The President	10. VaccineKills	10. side_effects_of_vaccine	10. VaccineSymptomsNegative	
11. Experiences of vaccine side effects		11. vax_is_ineffective_or_harmful	11. ThankYouDoctors	
12. Receiving second vaccine dose		12. getting_attention_of_politician	12. CovidConservative	
13. Encouraging getting the vaccine		13. sharing_stories_related_to_covid	13. VaccineDoesntPreventCovidNegative	
14. Saying that you can still get covid if you have the vaccine			14. VaccineConspiracy	
15. The vaccine does not work because you can still get covid			15. VaccineSymptomsPositive	

Table 13: Final codebooks for the COVID Tweets dataset using the LLM-based approach.

Async 1	**Async 2**	**Async 3**	**Sync 1**	**Sync 2**
1. AntiGunBan	1. pro-2A	1. AntiGunBan	1. AntiGunBan	1. gun rights
2. AntiWoke	2. donations for climate change	2. BidenPolicies	2. EnvironmentalConservation	2. climate action
3. ProtectEnvironment	3. solar energy	3. SaveEnvironment	3. SchoolBoardFinances/GasInfrastructure	3. renewables
4. CleanEnergyWorks	4. advocating for city level politics	4. AntiLiberalElectionCandidates	4. AntiGridDeregulation/PoliticalActionAds	4. inflation
5. ActionAgainstDemSpending	5. political activism to oppose climate change	5. RisingEnergyCost	5. CleanEnergyGoals	5. anti-woke
6. IndigenousAndBlackAdvocacy	6. criticism of liberal tax policy	6. ProRenewableEnergy	6. BellbrookSugarcreekConservativeElections	6. organized cooking oil theft
7. PassBidenAct	7. indigenous climate justice	7. PoliticalLeadershipTraining	7. Unexplainable	7. indigenous activism
8. AntiOilPolitician	8. promoting legislation against climate change	8. BioFuelTheft	8. ConservativePoliticsEvents	8. pro oil and natural gas
9. AntiPropertyTax	9. pros of wind and solar energy	9. ProUSOil	9. TPUSAAd	9. TPUSA
10. ProUSOilIndustry	10. criticism for disfavoring US ONG production	10. MinorityGroupWomenAdvocacy	10. DomesticOilAndGasProduction	10. conservative school board
11. ProtectRivers	11. kudos for confronting ONG emissions	11. DonationsForProRiverPoliticians	11. ClimateChange	
12. TPUSAAd	12. access to water and promoting renewable energy	12. SupportForBuildBackBetterAct	12. Pipeline	
13. EVCharging	13. discussing problems with school board	13. WaterPollutionAndContamination	13. BuildBackBetterAct	
14. OilTheft	14. promoting news and oped program	14. ProCleanEnergyPolitician	14. AntiGridDeregulation	
		15. TPUSALiveAd	15. LiberalTaxHikes	
		16. AntiLiberalSpending	16. NativeAmericanCulturalPreservation	
		17. AntiPropertyTax	17. OilTheft/ClimateLeader	
		18. EnergyInfrastructureProjects	18. NaturalGasIncentive/CleanEnergyPetition	
			19. CleanEnergyProjects	
			20. ColoradoAirQuality/SupportingBlackWomen	

Table 14: Final codebooks for the Climate Ads dataset using the Topic Model approach.

Async 1	**Async 2**	**Async 3**	**Sync 1**	**Sync 2**
1. AdvocateForGreenEnergy 2. GreenPowerExpensive 3. ClimateActionDonation 4. GreenEnergyWorks 5. AdvocateForGreenEnergyPoliticalAction 6. AntiTax 7. GreenEnergySavesMoney 8. StopOilPollution 9. DemocratOutreach 10. OilCreatesJobs 11. NuclearAdvocacy 12. AntiWoke	1. AgencyToStopClimateChange 2. CallsToInvestInRenewableEnergy 3. NeedQuickResponseForClimateCrisis 4. PoliticalActivismForRenewableEnergy 5. DonateToSupportClimateChangeActivism 6. ChallengesFacingRenewableEnergy 7. PoliticalActivismForOilAndGas 8. SupportForRenewableEnergy 9. ProtectFromHarmsCausedByFossilFuels 10. InroadsMadeByOilAndGas 11. InroadsMadeByRenewableEnergy 12. EconomicAdvantagesOfRenewableEnergy 13. SupportForBuildBackBetterAct 14. SupportForOilAndGas 15. EconomicAdvantagesOfOilAndGas 16. InroadsMadeByNuclearEnergy 17. EconomicAdvantagesOfNuclearEnergy	1. LegislatorsMustActAgainstClimateChange 2. ClimateChangeDonations 3. OilAndGasExtractionRisks 4. NuclearEnergyBenefits 5. TackleClimateChange 6. ReduceEmissions 7. VirginiaConservativeCampaign 8. OilAndGasJobs 9. SouthwestGasInvestment 10. RenewableEnergyInfrastructureGrowth 11. NuclearEnergyEducation 12. OilAndGasTax 13. BuildBetterAct 14. RestoreVirginia	1. DonationMatchingForClimate 2. AbandonedWells 3. GreenEnergyIsCostEffective 4. ClimateEmergency 5. NuclearRevenueSupportsJobs 6. HazardsOfWells 7. VirginiaConservatives 8. DemandClimateActionFromLeaders 9. PublicLandExploitation 10. GreenEnergyCreatesJobs 11. OilAndGasSupportJobs 12. AntiFossilFuelInfrastructure 13. CleanOilAndGas 14. GreenEnergyCapacity 15. CutMethane 16. CommerceCityDemocrats 17. RestoreVirginia 18. ThankRepForBuildBackBetter 19. BellbrookTaxRevolt 20. BellbrookWokeRadicals 21. GreenEnergySupportsCommunities 22. NuclearEnergyProvidesTaxRevenue 23. TellRepToNotRaiseOilGasTaxes 24. OilAndGasProvidesWages 25. CelebrateBuildBackBetter	1. ClimateChangeMovements 2. ClimateChangeDonations 3. OilAndGasTaxesBad 4. LawmakersAndOilAndGas 5. RenewableEnergyPromotionAndGrowth 6. CleanEnergyPromotion 7. RenewableEnergyEconomicBenefits 8. PoliticalTraining 9. CallToActionAgainstOilAndGasIndustries 10. CarbonEmissionReduction 11. OilAndGasBoostEconomy 12. PropertyTaxCampaign 13. BuildBackBetterActCampaign 14. RightWingCampaigns 15. NuclearPowerOtherBenefits 16. NuclearPowerEconomicBenefits

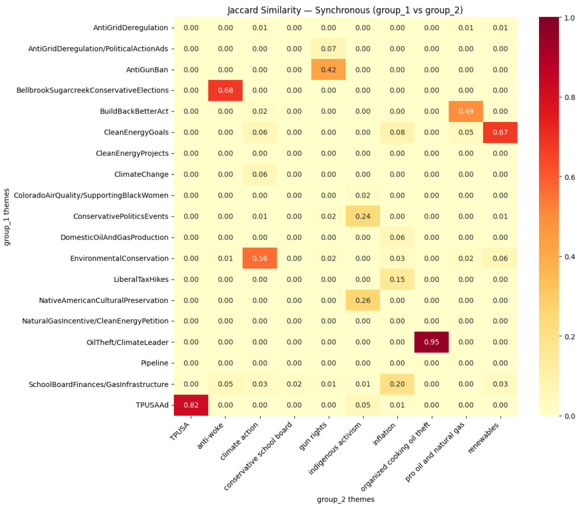
Table 15: Final codebooks for the Climate Ads dataset using the Relational approach.

Async 1	**Async 2**	**Async 3**	**Sync 1**	**Sync 2**
1. AntiWoke 2. GreenEnergySavesMoney 3. DemocratOutreach 4. AdvocateForGreenEnergy 5. GreenPowerExpensive 6. GreenEnergyWorks 7. StopOilPollution 8. AdvocateForGreenEnergyPoliticalAction 9. ClimateActionDonation 10. NuclearAdvocacy 11. AntiTax 12. OilCreatesJobs	1. AgencyToStopClimateChange 2. PoliticalActivismForRenewableEnergy 3. EconomicAdvantagesOfRenewableEnergy 4. ChallengesFacingRenewableEnergy 5. InroadsMadeByRenewableEnergy 6. CallsToInvestInRenewableEnergy 7. ProtectFromHarmsCausedByFossilFuels 8. NeedQuickResponseForClimateCrisis 9. DonateToSupportClimateChangeActivism 10. EconomicAdvantagesOfOilAndGas 11. PoliticalActivismForOilAndGas 12. InroadsMadeByOilAndGas 13. SupportForBuildBackBetterAct 14. EconomicAdvantagesOfNuclearEnergy 15. InroadsMadeByNuclearEnergy 16. SupportForOilAndGas 17. SupportForRenewableEnergy	1. NuclearEnergyBenefits 2. BuildBetterAct 3. ReduceEmissions 4. NuclearEnergyEducation 5. TackleClimateChange 6. OilAndGasExtractionRisks 7. VirginiaConservativeCampaign 8. RenewableEnergyInfrastructureGrowth 9. OilAndGasTax 10. OilAndGasJobs 11. SouthwestGasInvestment 12. ClimateChangeDonations 13. LegislatorsMustActAgainstClimateChange 14. RestoreVirginia	1. VirginiaConservatives 2. CommerceCityDemocrats 3. BellbrookWokeRadicals 4. BellbrookTaxRevolt 5. RestoreVirginia 6. GreenEnergyCreatesJobs 7. GreenEnergyIsCostEffective 8. GreenEnergyCapacity 9. GreenEnergySupportsCommunities 10. AbandonedWells 11. AntiFossilFuelInfrastructure 12. HazardsOfWells 13. PublicLandExploitation 14. CutMethane 15. ClimateEmergency 16. DemandClimateActionFromLeaders 17. DonationMatchingForClimate 18. OilAndGasSupportJobs 19. OilAndGasProvidesWages 20. TellRepToNotRaiseOilGasTaxes 21. CleanOilAndGas 22. NuclearRevenueSupportsJobs 23. NuclearEnergyProvidesTaxRevenue 24. CelebrateBuildBackBetter 25. ThankRepForBuildBackBetter	1. BuildBackBetterActCampaign 2. CallToActionAgainstOilAndGasIndustries 3. CarbonEmissionReduction 4. CleanEnergyPromotion 5. ClimateChangeDonations 6. ClimateChangeMovements 7. LawmakersAndOilAndGas 8. NuclearPowerEconomicBenefits 9. NuclearPowerOtherBenefits 10. OilAndGasBoostEconomy 11. OilAndGasTaxesBad 12. PoliticalTraining 13. PropertyTaxCampaign 14. RenewableEnergyEconomicBenefits 15. RenewableEnergyPromotionAndGrowth 16. RightWingCampaigns

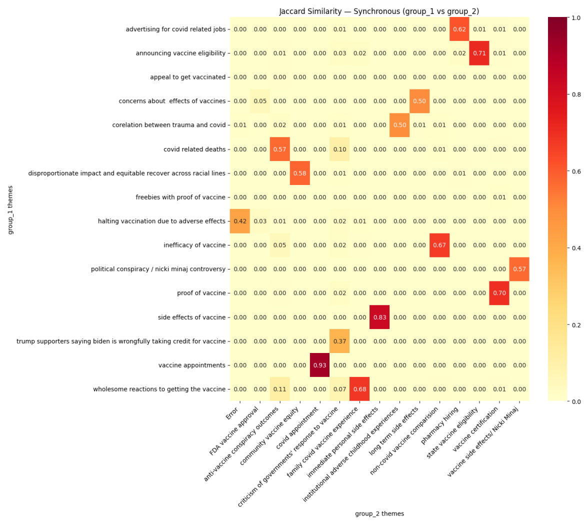
Table 16: Final codebooks for the Climate Ads dataset using the LLM-based approach.

Dataset	Model	Async 1	Async 2	Async 3	Sync 1	Sync 2	Total
COVID Tweets	Topic Model	19	18	15	16	14	82
	Relational	19	10	18	9	8	64
	LLM-based	15	10	13	15	7	60
Climate Ads	Topic Model	14	14	18	20	10	76
	Relational	12	17	14	25	16	84
	LLM-based	12	17	14	25	16	84

Table 17: Number of themes identified per experiment, broken down by dataset, model, and collaboration setting.



(a) Climate ads.



(b) COVID tweets.

Figure 9: Jaccard similarity heatmaps using the topic model approach.

F Semi-Structured Interview

F.1 Interviewing

We administered interviews after annotation sessions. Asynchronous annotators were asked questions individually about their experience, whereas synchronous annotator groups were asked questions with their fellow annotators.

F.2 Script

1. Have you worked on annotation projects before? Did these annotation projects use qualitative coding strategies (ex: grounded theory)? How experienced are you as an annotator?

2. How was your experience on the COVID-19 vaccine annotation session we conducted on Sunday? Particularly, we are interested in your thoughts and feelings over the session.

3. You annotated in a group, working together as a team. Did you find this setup to be beneficial? What were some of the limitations you faced, both individually and and as a group, when working synchronously?

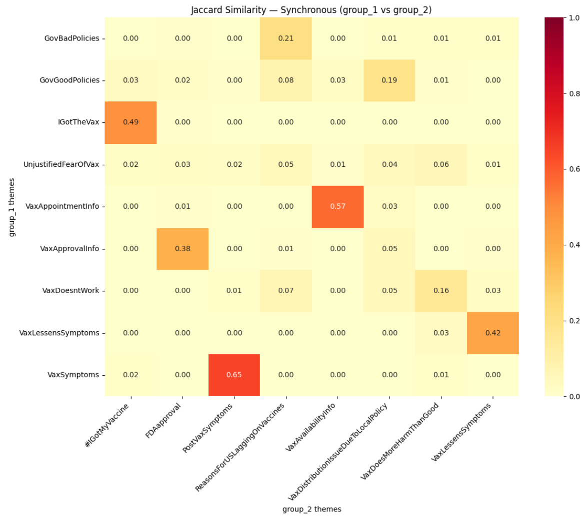
4. On a similar line, what would you consider to be the pros and cons if you were to annotate alone?

The last questions would be flipped based on if we are posing it to synchronous or asynchronous annotators.

G Use of Pre-Existing Artifacts

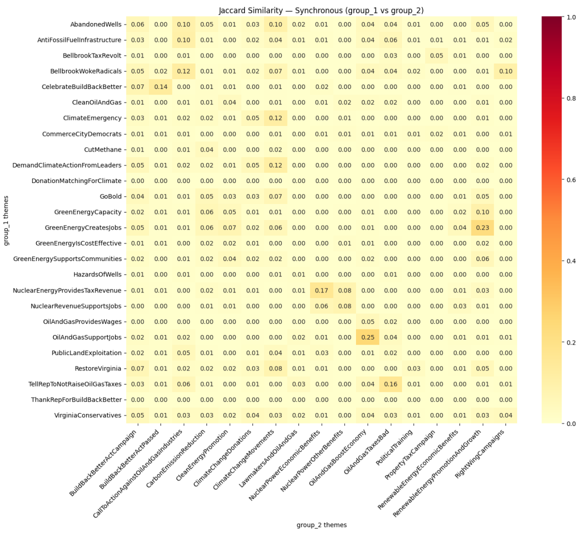
All pre-existing artifacts utilized in this study, including datasets, software libraries, models, and computational tools, are publicly available under open-source and open-access licenses. This academic work adheres to all intended use guidelines and terms of service for the respective resources.

(a) Climate ads.

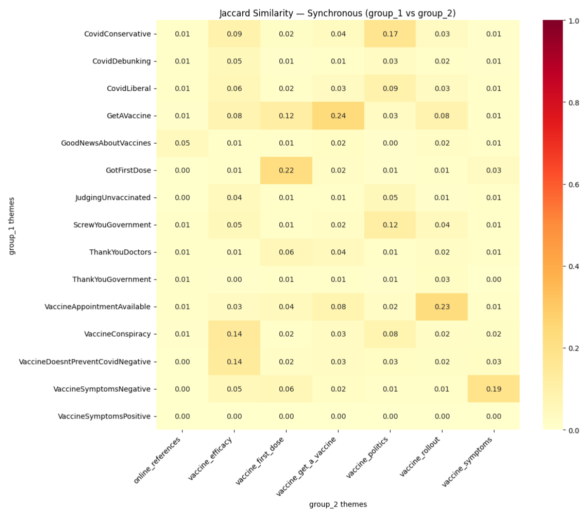


(b) COVID tweets.

Figure 10: Jaccard similarity heatmaps using the relational approach.



(a) Climate ads.



(b) COVID tweets.

Figure 11: Jaccard similarity heatmaps using the LLM-based approach.