Thematic-LM: A LLM-based Multi-agent System for Large-scale Thematic Analysis

Tingrui Qiao*
tqia361@aucklanduni.ac.nz
University of Auckland
Auckland, New Zealand

Caroline Walker
caroline.walker@auckland.ac.nz
University of Auckland
Auckland, New Zealand

Chris Cunningham
C.W.Cunningham@massey.ac.nz
Massey University
Wellington, New Zealand

Yun Sing Koh
y.koh@auckland.ac.nz
University of Auckland
Auckland, New Zealand

Abstract

Thematic analysis (TA) is a widely used qualitative method for identifying underlying meanings within unstructured text. However, TA requires manual processes, which become increasingly labour-intensive and time-consuming as datasets grow. While large language models (LLMs) have been introduced to assist with TA on small-scale datasets, three key limitations hinder their effectiveness. First, current approaches often depend on interactions between an LLM agent and a human coder, a process that becomes challenging with larger datasets. Second, with feedback from the human coder, the LLM tends to mirror the human coder, which provides a narrower viewpoint of the data. Third, existing methods follow a sequential process, where codes are generated for individual samples without recalling previous codes and associated data, reducing the ability to analyse data holistically. To address these limitations, we propose Thematic-LM, an LLM-based multi-agent system for large-scale computational thematic analysis. Thematic-LM assigns specialised tasks to each agent, such as coding, aggregating codes, and maintaining and updating the codebook. We assign coder agents different identity perspectives to simulate the subjective nature of TA, fostering a more diverse interpretation of the data. We applied Thematic-LM to the Dreaddit dataset and the Reddit climate change dataset to analyse themes related to social media stress and online opinions on climate change. We evaluate the resulting themes based on trustworthiness principles in qualitative research. Our study reveals insights such as assigning different identities to coder agents promotes divergence in codes and themes.

CCS Concepts

• **Applied computing** → **Sociology**; • **Computing methodologies** → **Multi-agent systems**; *Information extraction.*

*Corresponding author

WWW '25, April 28-May 2, 2025, Sydney, NSW, Australia

ACM ISBN 979-8-4007-1274-6/2025/04
https://doi.org/10.1145/3696410.3714595

Keywords

Computational Social Science, Thematic Analysis, Large Language Model, Multi-agent System

ACM Reference Format:
Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. 2025. Thematic-LM: A LLM-based Multi-agent System for Large-scale Thematic Analysis. In *Proceedings of the ACM Web Conference 2025 (WWW '25), April 28-May 2, 2025, Sydney, NSW, Australia.* ACM, New York, NY, USA, 10 pages. https://doi.org/10.1145/3696410.3714595

1 Introduction

The growing availability of unstructured text data, particularly from social media, presents both an opportunity and a challenge for researchers [19]. While such data can hold valuable insights, analysing them effectively requires robust methods to extract meaning from vast volumes of information. Computational approaches, such as topic modelling [1] and sentiment analysis [8], can handle large datasets but often produce surface-level results that fail to uncover the underlying, context-specific meanings within the data. In contrast, qualitative methods, such as thematic analysis (TA) [5], are designed to explore nuanced interpretations by focusing on the subjective experiences and contexts that shape the data. As illustrated in Fig. 1, thematic analysis involves systematically identifying patterns of meaning within text data [5]. It begins with coding, where key concepts or ideas relevant to the research question are labelled. These codes are then reviewed and grouped into broader themes, which capture underlying patterns and contextual meanings within the data. Despite its effectiveness, TA is labour-intensive and time-consuming, requiring manual processes [6] such as familiarization, coding, theme development, and interpretation of themes, which become increasingly burdensome as the dataset grows. For larger datasets, a team of trained coders is often required to work collaboratively to manage the volume while maintaining the reliability and credibility of the results [44], which is both expensive and logistically challenging.

Recent advances in large language models (LLMs) have opened new possibilities for automating thematic analysis, as LLMs have demonstrated impressive capabilities in processing unstructured data by learning from vast corpora of texts [2, 24, 46]. Researchers have begun applying LLMs as single agents to assist in the thematic analysis process [12, 13, 16]. However, existing approaches have

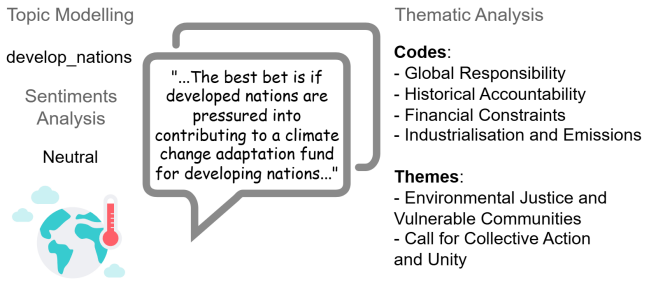

Figure 1: Differences between topic modelling, sentiment analysis and thematic analysis, illustrated through an example of climate change-related post from social media.

three major limitations. First, current computational thematic analysis methods require interaction with a human coder. The human coder needs to be familiar with the entire dataset, oversee the LLMs' outputs, and provide feedback to the LLMs, which is infeasible given a large dataset. Second, due to the iterative feedback process, the LLMs often tend to imitate the human coder's perspective, producing results that mirror the coder's viewpoint [12], limiting the diversity of viewpoints and resulting in a narrower analysis [12]. Third, current approaches to thematic analysis are sequential; LLMs do not revisit previously coded data to update codes as new information arises, undermining a key principle of thematic analysis: ensuring that codes accurately reflect consistent meanings across the dataset. Apart from these limitations, evaluating the themes generated by LLMs presents further challenges due to the volume of data and the subjective nature of thematic analysis. While qualitative evaluation by humans becomes impractical at scale, automatic evaluation metrics, such as inter-rater reliability [33], often assess the similarity between the LLMs' output and manual results and overlook the fact that thematic analysis can reflect multiple valid perspectives within the data. Consequently, lower inter-rater reliability may not necessarily indicate lower-quality thematic analysis.

To address these limitations, we introduce an LLM-based multi-agent system for large-scale computational TA, which we term Thematic-LM. Thematic-LM assigns distinct components of TA as specialised tasks to individual LLM agents, fully automating the process. To simulate the process of refining codes in response to new data, we implement an adaptive codebook that stores prior codes and their corresponding quotes. A reviewer agent retrieves similar codes and quotes from the codebook, compares them to the new data, and updates the codes accordingly. We allocate coder agents with different identity perspectives to generate different views on the codes and themes. Additionally, we analyse the quality of themes based on the computational adaptation of the principles of trustworthiness in qualitative research [25, 27, 38], including credibility, dependability, confirmability, and transferability. Our main contributions are as follows:

- We propose **Thematic-LM**, an LLM-based multi-agent system for large-scale computational thematic analysis. To the best of our knowledge, Thematic-LM is the first to employ multiple LLM agents for qualitative analysis.
- We encourage different perspectives on themes by assigning different identities to the LLM agents, prompting them to reflect on their identities while performing the analysis.
- We apply Thematic-LM to the Dreaddit [47] and the Reddit climate change dataset [1], uncovering underlying themes regarding social media stress and opinions on climate change.

2 Related Work

Social Media Data Analytic. With over a billion people using social media, enormous amounts of unstructured data are generated through daily interactions on these platforms [19]. Various machine learning techniques have been employed to extract insights to handle the scale of such data. Topic modelling approaches [4, 11, 56] are applied to uncover abstract topics or clusters of similar content from large datasets. Sentiment analysis [8, 37, 52] focuses on determining the emotional tone or attitude behind a piece of text, such as classifying whether social media posts reflect positive, negative, or neutral sentiment. Other classification approaches [43, 45] are typically employed to categorize social media posts into predefined categories, such as news, entertainment and sports. However, these methods tend to produce high-level categorizations and descriptive outputs, offering surface-level insights into the data, which do not capture deeper, contextual meanings or allow for nuanced interpretations of complex social media interactions. Our Thematic-LM automates TA through a multi-agent system with multiple coders, enabling deeper exploration of underlying meanings and perspectives from large-scale datasets.

Computational Thematic Analysis. Several studies have explored the use of LLMs for automating TA. De Paoli [13] and Drápal et al. [16] applied LLMs to relatively small datasets by guiding the models through structured, step-by-step coding instructions. Drápal et al. [16] found that LLM performance closely aligns with human coders when iterative feedback is provided. Similarly, Dai et al. [12] proposed a feedback loop where expert input helps refine the LLM's output. While these approaches demonstrate promise, their reliance on human intervention and focus on small datasets limit their scalability. In contrast, Thematic-LM assigns a team of LLM agents to handle different components of TA, fostering a broader perspective by simulating independent coders. Each coder agent is given a unique identity, encouraging analysis from diverse viewpoints. Additionally, Thematic-LM employs an adaptive codebook that revisits and updates previously coded data, ensuring scalability and adaptability to large datasets.

LLM-based Multi-agent System. Recent research has shown that collaboration between multiple LLM agents can enhance inter-consistency [51], improve factuality and reasoning [17], and encourage divergent thinking [29]. Motivated by the benefits, various LLM-based multi-agent systems have been developed [10, 28, 50]. Multi-agents are often employed for problem-solving or simulation. For example, Hong et al. [21] uses specialised LLM agents as a software engineering team for developing applications collaboratively. Chan et al. [9] proposed ChatEval, which uses multi-agent debate to evaluate the quality of LLM outputs. For research on simulation, Zhang et al. [53] explored simulating collaborative intelligence in human society by assigning LLM agents various personal traits and

[1] https://www.kaggle.com/datasets/pavellexyr/the-reddit-climate-change-dataset

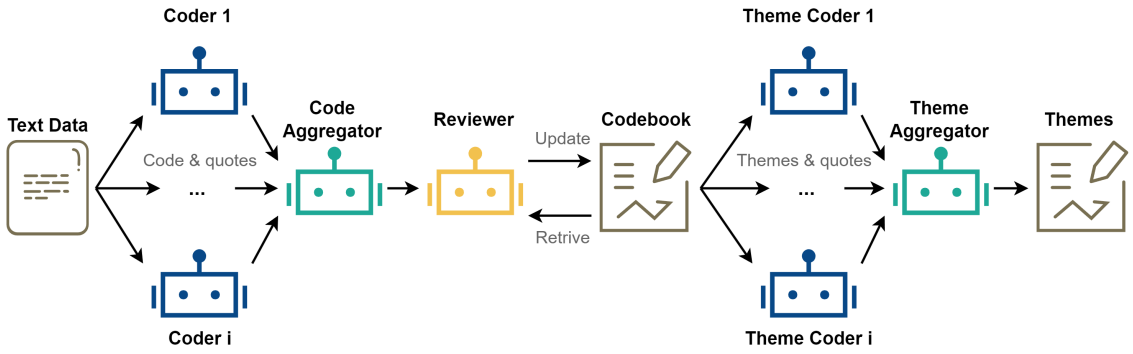

Figure 2: Thematic-LM consists of coder, aggregator, reviewer, and theme coder agents, organized into two stages: coding and theme development. In the coding stage, multiple coder agents independently analyse text data and output codes and quotes to the code aggregator. The code aggregator refines and organizes the codes before sending them to the reviewer. The reviewer maintains an adaptive codebook, ensuring codes are consistent with prior coded data. In the theme development stage, theme coder agents use the codebook to identify themes, which are then refined by the theme aggregator to produce the final themes.

thinking styles. Similarly, Park et al. [40] established a community of 25 agents in a sandbox environment simulating a small town, while Kovač et al. [26] constructed a school environment with LLM agents to explore developmental psychology. Moreover, Zhao et al. [54] examine the competition between LLM agents by simulating a virtual town with restaurant agents competing over customer agents. Thematic-LM focuses on TA as a problem-solving task and simulates coders with different identities to encourage a broader viewpoint regarding the data.

3 LLM-based Multi-agent System for Thematic Analysis

We adopt the inductive thematic analysis (TA) approach outlined by Braun and Clarke [5]. In TA, coding identifies items of analytic interest from the data and assigns short-phrase labels, while themes are built by synthesizing and refining insights from these codes. In traditional team-based TA, each coder often works independently to generate codes, followed by regular meetings to compare and consolidate codes, reducing redundancy or overlap and ensuring cohesion in the analysis [44]. Inductive TA is data-driven, which develops themes from the data rather than with a predefined codebook of themes [6]. In contrast to the conventional predefined codebook approach, we implement an adaptive codebook that continuously updates codes throughout the coding process, accommodating new data and insights. Building on previous work of computational TA [12, 13, 16], Thematic-LM performs TA in two stages: coding and theme development. In the coding stage, the codebook is finalized as codes are generated and refined, while the theme development stage focuses on synthesizing themes from the codebook. We provide details of the system in Section 3.1 and the coders' identity perspectives in Section 3.2.

3.1 Multi-agent System

As illustrated in Fig. 2, our multi-agent system consists of three types of LLM agents: coder, aggregator, and reviewer. Each agent has a specialised role in the TA process, contributing to coding and theme development stages to fully automate these tasks. The agents are implemented with conversational agents from AutoGen [50].

Coder Agents are responsible for coding in the first stage and identifying themes in the second stage. In the coding stage, the coders are instructed to write one to three codes for each piece of data to capture concepts or ideas with the most analytical interest. For each code, the coder extracts a representative quote from the data as evidence. The resulting codes, quotes and corresponding quote IDs are passed to the code aggregator agent. During the theme development stage, the theme coders are given a complete version of the codebook from the coding stage. The codebook is compressed with LLMLingua [22, 23] to reduce token costs. The coder agents then analyse the codes and associated quotes holistically to identify overarching themes that reflect deeper insights into the data. These themes, along with theme descriptions and the most relevant quotes, are then passed to the theme aggregator.

Aggregator Agents refine and organize the outputs from the coder agents into structured formats suitable for the next stage. During the coding stage, the code aggregator merges codes with similar meanings, retaining differences where necessary, and organizes the codes, quotes, and quote IDs into JSON format, which the reviewer agent uses to update the codebook. Similarly, in the theme development stage, the theme aggregator refines and organizes the identified themes and associated quotes, merging similar themes and outputting the final themes in JSON format.

Reviewer Agent operates exclusively during the coding stage, maintaining and updating the codebook. This codebook stores previous codes, their corresponding quotes, and quote IDs in JSON format. Each entry in the codebook is a code, and its associated quotes are nested below each code along with their quote IDs. Codes are represented both as texts and as embeddings, generated using a Sentence Transformer model [42]. The reviewer agent processes new codes and quotes from the aggregator and retrieves the top-k similar codes and quotes from the codebook by computing the cosine similarity between their code embeddings. The reviewer compares the new codes and quotes with existing codes and quotes to determine whether these codes can be updated and whether

similar existing codes can be merged. After making these decisions, the reviewer updates the codebook to save new codes and quotes and merge similar codes. The reviewing and updating process is crucial in TA, as it plays a central role in ensuring the codes remain dynamic, interpretative, and responsive to the data. Once finalized, the codebook is passed to the theme development stage.

Evaluation We evaluate the quality of themes based on the principles of trustworthiness in qualitative research [25, 27, 38]. Existing metrics used in computational TA, such as inter-rater reliability [33], assume that there is one "correct" set of themes to match against for measuring the level of accuracy. We propose that trustworthiness principles, which emphasize meaningful, coherent, and data-grounded analysis, provide a more robust framework for evaluating themes. As shown in Fig. 3, we adopt the trustworthiness principles for evaluating the computational TA approaches: (1) *Credibility and Confirmability*: Credibility evaluates whether the themes accurately represent the data, while confirmability assesses whether the themes are data-driven rather than driven by biases. We measure credibility and confirmability at the same time by retrieving the associated data through quote IDs and assigning an LLM-as-a-judge [55] evaluator agent to determine whether the themes are consistent with the data. We compute the percentage of quoted data that is consistent with the corresponding themes. The inconsistency with the data can be caused by hallucinations or internal biases within the LLM models. (2) *Dependability*: asses whether the same process can be repeated by a separate researcher and reveal similar findings. The dependability of the computational approach can be measured by repeating the process and measuring the inter-rater reliability of the resulting themes. We measure the inter-rater reliability in themes by conducting the TA several times and computing the average pairwise ROGUE scores [30], which measures the amount of overlap between the themes. For each pair of theme sets A and B, we first calculate the ROUGE-1 and ROUGE-2 scores by using set A as the reference set:

$$\begin{aligned} \text{ROUGE-1}_{A\to B} &= \frac{\text{Number of overlapping unigrams in } B}{\text{Total number of unigrams in } A} \\ \text{ROUGE-2}_{A\to B} &= \frac{\text{Number of overlapping bigrams in } B}{\text{Total number of bigrams in } A} \end{aligned} \tag{1}$$

We calculate the ROUGE-1 and ROUGE-2 scores by using sets B as the reference set and compute the average of the ROUGE-1 and ROUGE-2 scores for the pair of sets:

$$\begin{aligned} \text{ROUGE-1} &= \frac{1}{2}\left(\text{ROUGE-1}_{A\to B} + \text{ROUGE-1}_{B\to A}\right) \\ \text{ROUGE-2} &= \frac{1}{2}\left(\text{ROUGE-2}_{A\to B} + \text{ROUGE-2}_{B\to A}\right) \\ \text{ROUGE} &= \frac{1}{2}\left(\text{ROUGE-1} + \text{ROUGE-2}\right) \end{aligned} \tag{2}$$

(3) *Transferability*: assesses whether the identified themes and codes can be meaningfully applied to other contexts or datasets with similar characteristics. Since it might be challenging to find or construct a dataset with similar characteristics from other contexts, we evaluated whether the identified themes from one subset of a dataset can generalize to another subset from the same dataset. We split the dataset into a training and validation set, where we perform TA separately and measure whether the themes from the training set can transfer to the themes in the test set by computing the overlap between themes via pairwise ROGUE scores shown in Eq. (2). While this approach does not address the application of themes to datasets from different sources, it provides insight into how well the themes represent unseen data with similar characteristics.

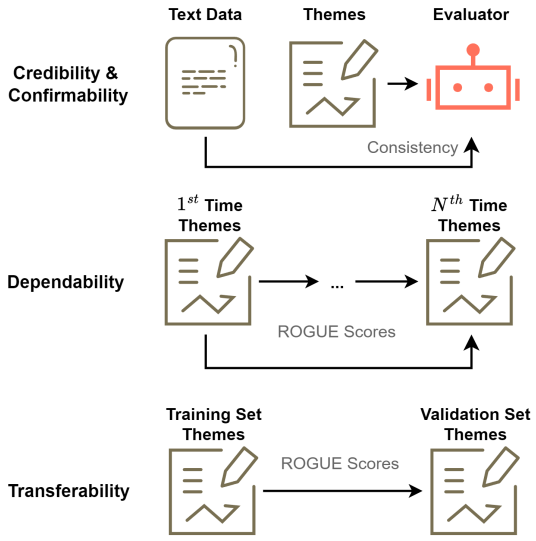

Figure 3: Evaluation framework: we employ an evaluator agent to check the consistency between the themes and associated data to assess the credibility and confirmability of the themes. To assess their dependability, we repeat the thematic analysis (TA) process N times and examine whether the themes remain stable by computing the overlap between themes. For transferability, we perform TA independently on two split data sets and compute the overlap between themes.

3.2 Coder Identities

TA inherently embraces subjectivity, recognizing that researchers bring their own perspectives, assumptions, and interpretations to the data [6, 18]. The identification of themes is guided by the coder's insights and understanding, which plays an active role in deciding what is meaningful in the data. Consequently, the same data may yield different themes depending on who is conducting the analysis. Coders may interpret the same information in diverse ways, especially when they come from varied social, cultural, or professional backgrounds. This variability does not undermine the reliability of the analysis but instead highlights the subjectivity that enriches qualitative research. The subjective nature of thematic analysis allows it to delve deeply into human experiences, emotions, and meanings while providing the contextual understanding needed to explore nuanced social and cultural issues [18, 38].

In previous work on computational thematic analysis, the LLM's outputs are aligned with a human coder through iterative feedback. In contrast, Thematic-LM simulates coders with varied backgrounds to foster diverse perspectives in data interpretation. In profiling the coder agents, we draw from existing literature on different viewpoints and opinions related to the subject matter, assigning distinct identities to the system message of each agent. These agents

Table 1: Themes and description of the themes produced by Thematic-LM on the Dreaddit dataset, compared with topic modelling and sentiment analysis results.

Theme	Description	Related Topic	Sentiment Score
Navigating Emotional and Mental Health Challenges	The experiences of individuals dealing with anxiety, mental health concerns, and the impact of emotional stressors on daily life.	therapist_to_therapy	0.00
Impact of Economic Stress and Resource Scarcity	The financial struggles and efforts to manage limited resources and economic instability.	homeless_the_to_have	0.10
Familial Dynamics and Responsibilities	The complex relationships within families, the burden of roles, and the tension between obligations and emotional needs.	food_you_for	0.15
Coping with Academic and Professional Pressures	The stress associated with academic performance and the pressures to succeed in professional life.	job_and_have	0.20
Interpersonal Conflict and Relationship Stress	The conflicts in personal relationships, whether with romantic partners, friends, or colleagues, the emotional and mental strain caused by unresolved tensions or disagreements.	sex_that_sexual	0.15
Seeking Validation and Emotional Support	The frequent attempts by individuals to gain reassurance or validation from others, often by sharing their struggles and emotions openly in the hope of receiving empathy or encouragement.	you_are_support	0.35
The Dichotomy of Online and Offline Identities	The differences between one's real and virtual personas and the impact of social media on identity and interactions.	he_him_me	0.25

are instructed to interpret the data through the lens of their assigned identities, reflecting on how someone with such a background might perceive and analyse the information. This approach allows us to explore the diversity of perspectives that may emerge from the data and offers a way to measure the divergence between coders' interpretations due to different backgrounds.

4 Experiments and Analysis

We conduct a series of experiments to assess Thematic-LM's effectiveness in performing thematic analysis on social media datasets. At the time of writing, no other multi-agent systems were designed for qualitative analysis, such as thematic analysis. Therefore, our experiments focus primarily on Thematic-LM. Specifically, we aim to answer the following questions:

- What insights can Thematic-LM uncover from the Dreaddit and the Reddit climate change dataset? (Section 4.1)
- Does Thematic-LM produce higher quality themes compared to a single LLM agent? (Section 4.2)
- How do the different coder identities influence the codes and themes? (Section 4.3)

Experimental Setup. We use GPT-4o [2] to serve as the LLM agents. The *temperature* and *top_p* are set at the default value of one. JSON mode needs to be enabled for the agents to ensure the consistency of the output format. For each code and theme, the agents save up to 20 of the most relevant quotes associated with the concept. The reviewer retrieves the top 10 most similar codes for each new code. To measure the dependability, we conducted the TA three times and calculated the average pairwise ROGUE scores between the resulting sets of themes. To measure transferability, we split the dataset into a 50% training set and a 50% validation set. We employ Thematic-LM to analyse the Dreaddit [47] and the Reddit climate change dataset. Dreaddit contains over 190k posts from subreddits related to abuse, anxiety, financial issues, PTSD and relationship problems. The Reddit climate change dataset consists of 4.6 million Reddit posts and comments related to climate change.

[2]https://platform.openai.com/docs/models/gpt-4o

4.1 Thematic Analysis of Social Media Data

We assign two coder agents and two theme coder agents for TA on Dreaddit and the Reddit climate change dataset. The agents are provided with the instructions only, with no personal identities given to the agent. To compare the TA results, we perform topic modelling and sentiment analysis on the datasets. We employ BERTopic [20] to categorize the posts into topics and RoBERTa [31] from TweetNLP [7] for sentiment analysis. We set the number of neighbours, number of components for UMAP [35] and minimum cluster size for HDBSCAN [34] in BERTopic as 15, 10 and 10, respectively. We use cosine similarity as the distance metric in HDBSCAN. For each theme, we select the most relevant topic by looking at the majority of the topics of the data points associated with the theme through the quote IDs. Similarly, we computed the average sentiment of the data points. Sentiment scores of zero, one and two denote negative, neutral and positive, respectively.

Dreaddit Datset. As shown in Table 1, Thematic-LM produced seven themes from the Dreaddit dataset. The sentiment analysis returns mostly negative labels for the data associated with the themes. We observe that the themes identified by Thematic-LM on the Dreaddit dataset highlight a broader and more meaningful understanding of the data compared to the related topics generated by topic modelling. For instance, the theme "Navigating

Table 2: Themes and description of the themes produced by Thematic-LM on the Reddit climate change dataset, compared with topic modelling and sentiment analysis results.

Theme	Description	Related Topic	Sentiment Score
Emotional Burden of Climate Change Awareness	The complex emotions individuals face, such as anxiety, guilt, frustration, and helplessness, stemming from the overwhelming nature of climate change and a perceived lack of control over its outcomes.	climate_change_years	0.20
Generational and Cultural Disconnection	The perceived gaps in understanding and values across different generations and cultures are often exacerbated by rapid societal and technological changes.	denier_climate_scope_change	0.30
Call for Collective Action and Unity	Emphasizing the necessity for collective action and unity in addressing societal challenges, including political divisions and significant issues like climate change and economic inequality.	jobs_people_us	1.30
Critique and Skepticism of Political and Economic Systems	A critical examination of current political and economic systems, highlighting concerns about inequality, inefficiency, and the shortcomings of existing policies.	bank_companies_billion	0.25
Personal and Community Resilience	Personal and communal efforts to adapt to the impacts of climate change, emphasizing the importance of strengthening social connections and local initiatives	people_change	0.85
Role of Technology in Climate Solutions	The potential of technology and innovation to mitigate climate change effects, including renewable energy advancements, carbon capture technologies, and sustainable agriculture practices.	energy_nuclear_power	0.90
Impact on Biodiversity and Ecosystems	Individuals express worries about endangered species, habitat destruction, and the overall health of the planet's ecosystems.	meat_animals	0.10
Climate Migration and Displacement	The challenges faced by communities and individuals who are forced to relocate due to climate change impacts such as rising sea levels, extreme weather events, and resource scarcity.	housing_cities_city	0.20

Table 3: Comparison of the theme quality scores on the Dreaddit dataset.

Method	Credibility & Confirmability	Dependability	Transferability
Single	0.63	0.45	0.41
Single (Codebook)	0.75	0.61	0.67
System (1 Coder)	0.92	**0.81**	0.86
System (2 Coders)	**0.94**	0.78	**0.87**

Table 4: Comparison of the theme quality scores on the Reddit climate change dataset.

Method	Credibility & Confirmability	Dependability	Transferability
Single	0.66	0.56	0.73
Single (Codebook)	0.74	0.69	0.78
System (1 Coder)	0.96	0.84	**0.90**
System (2 Coders)	**0.98**	**0.86**	0.89

Emotional and Mental Health Challenges" captures the users' struggles with anxiety and emotional stressors. In contrast, the related topic "therapist_to_therapy" provides a more fragmented association of words, missing the depth in the narrative. Similarly, "Impact of Economic Stress and Resource Scarcity" encapsulates the daily struggles of managing limited resources, while the related topic "homeless_the_to_have" only loosely connects words around homelessness and possession, failing to capture the specific challenges individuals face in their economic lives. This comparison illustrates how thematic analysis delves into the underlying meanings and human experiences, offering a much more insightful and comprehensive picture than topic modelling, which often yields superficial groupings of co-occurring terms.

The themes reflect various levels of human needs, resonating with Maslow's Hierarchy of Needs [36]. "Navigating Emotional and Mental Health Challenges" and "Impact of Economic Stress and Resource Scarcity" correspond to Maslow's foundational physiological and safety needs, as they involve mental well-being and financial stability. "Familial Dynamics and Responsibilities" and "Interpersonal Conflict and Relationship Stress" align with belongingness and love needs, highlighting the importance of relationships and emotional bonds in individuals' lives. Meanwhile, "Coping with Academic and Professional Pressures" and "Seeking Validation and Emotional Support" relate to esteem needs, where individuals seek recognition, achievement, and emotional validation. Finally, "The Dichotomy of Online and Offline Identities" reflects the higher-order need for self-actualization as individuals navigate personal identity and the complexities of presenting themselves in digital and real-world environments.

The Reddit Climate Change Dataset. As shown in Table 2, the eight themes identified by Thematic-LM present a more nuanced and interconnected understanding of climate change discourse, emphasizing emotional and social dimensions rather than solely categorizing discussions by co-occurring words. The emotion captured by the themes generally aligns with the sentiment score. The "Emotional Burden of Climate Change Awareness" theme highlights the psychological distress, anxiety, and feelings of helplessness that

individuals face, reflecting the concept of eco-anxiety [41]. This emotional struggle is intertwined with the "Generational and Cultural Disconnection", which points to the gaps in understanding and values that can arise between different generations, further complicating collective responses to climate change. The theme "Call for Collective Action and Unity" underscores the necessity for collaboration in addressing climate-related challenges, emphasizing a shared responsibility that can help foster social cohesion. This is complemented by the theme "Personal and Community Resilience," which showcases the importance of local initiatives and social connections as individuals and communities adapt to the changing environment. Meanwhile, the "Critique and Skepticism of Political and Economic Systems" reflects a growing awareness of systemic failures and the need for significant reforms to ensure effective climate action. The theme "Role of Technology in Climate Solutions" highlights the potential for innovation and advancements in technology to mitigate the impacts of climate change, showcasing a hopeful perspective amid the challenges. Finally, the theme "Impact on Biodiversity and Ecosystems" serves as a reminder of the broader ecological implications of climate change, emphasizing the interconnectedness of human actions and environmental health. Together, these themes illustrate a complex multifacet of emotional, social, and systemic factors shaping climate change discourse from social media.

4.2 Quality of Themes

To investigate the benefits brought by the adaptive codebook and the multi-agent system, we compare the quality of themes between a single LLM agent, a single LLM agent with an adaptive codebook, Thematic-LM with one coder for coding and another for theme development and Thematic-LM with two coders for coding and theme development, respectively. The single LLM agent is instructed to first label the data sequentially with codes, define themes from the codes, and save the data IDs of the most relevant codes for each theme. The single LLM agent approach is similar to Drápal et al. [16] and De Paoli [13]. The single LLM agent with the adaptive codebook adds steps for retrieving similar codes for comparison and saving codes and quotes into the codebook. After the coding stage, the single LLM agent takes the codebook as input and defines themes from the codebook. The coder agents are given instructions without assigning any identities.

As shown in Tables 3 and 4, we observe that the introduction of an adaptive codebook improves the quality of the themes of the LLM agent, and multi-agents perform better than single agents. The credibility & confirmability score of the single agent with the codebook is improved due to the LLM agent having access to retrieve past quotes, which provides a chance to reflect on past codes and quotes whenever similar data arrive. The similar codes and associated quotes together give a more holistic view and improve the contextual understanding of the data, which makes the coding less affected by randomness brought by a single data sample. This is shown by improvements in both dependability and transferability. The multi-agent systems have higher credibility & confirmability, dependability and transferability scores. In the multi-agent system, the distribution of specialised tasks has made the tasks simpler and shorter for each agent, improving factuality and reducing hallucination brought by doing complex tasks. This has led to more stable and transferable themes across different runs.

Table 5: Examples of themes not captured in Thematic-LM with no assigned coder identities but emerged in Thematic-LM where coders are assigned different identity perspectives.

Theme	Description
Economic Impact of Climate Policies	Concerns about the economic consequences of aggressive climate regulations, particularly their impact on industries and job markets.
Environmental Stewardship	The deep responsibility to protect and maintain the natural environment, viewing humans as caretakers of the earth.
Scepticism of Climate Science	Questioning about the extent of human influence on global warming and discuss whether climate change is part of the natural cycle.
Environmental Justice and Vulnerable Communities	Highlights the disproportionate impact of climate change on marginalized communities, advocating for policies that address environmental justice and protect vulnerable populations.

4.3 Divergence of Pespectives

We aim to investigate whether assigning different identities can broaden the views of the TA and the effects of assigning identities to coders. We conducted the experiments on the Reddit climate change dataset, as climate change is a polarizing issue due to the intersection of social, economic, political and cultural values, which lead to divergent opinions [14, 15]. The subjectivity of TA and the scale of the dataset might lead to some views being underrepresented in the resulting themes. We assign five coders with different identity perspectives: (1) *Human-Driven Climate Change Agent*: This agent adopts the widely accepted scientific view that human activities are the primary drivers of climate change [15]. It focuses on the role of industrialization, fossil fuel emissions, deforestation, and other anthropogenic activities in accelerating global warming. (2) *Natural Climate Change Agent*: This agent approaches climate change from the viewpoint that it is a natural phenomenon, part of Earth's long-term climatic cycles. It reflects the arguments that climate fluctuations have occurred over millennia due to factors like solar radiation, volcanic activity, and ocean currents, suggesting that current climate shifts may not be solely due to human activities [3, 32]. (3) *Progressive View Agent*: The progressive agent is given the progressive perspective rooted in environmental justice, equity, and sustainability, advocating for systemic changes that address not only environmental issues but also social inequalities exacerbated by climate impacts [15]. The agent emphasizes green technologies, grassroots activism, and policies that ensure vulnerable communities are not disproportionately affected. (4) *Conservative View Agent*: This agent reflects the conservative perspective on climate change, focusing on gradual, market-driven solutions rather than large-scale regulatory interventions [15]. It prioritizes economic stability, energy independence, and limited government involvement in climate policies. From this viewpoint,

Figure 4: Comparison of codes and themes generated by five coders with no identities given, with the same identities of "human-driven climate change" given, and with five different identities. The five different identities are agents which are instructed to believe in "human-driven climate change" and "climate change as a natural cycle" and instructed to act with "progressive view", "conservative view" and "Indigenous view", denoted as H, N, P, C, and I respectively. The pairwise ROGUE scores measure the differences between codes and themes from different agents.

climate action should not jeopardize economic growth, jobs, or individual freedoms. (5) *Indigenous View Agent*: The Indigenous agent operates from the perspective that climate change is deeply intertwined with human relationships with nature and the environment [48, 49]. It emphasizes traditional ecological knowledge, the interconnectedness of all living beings, and the sacred responsibility to care for the land. This agent highlights climate change's cultural, spiritual, and community-based dimensions.

To measure the effects of identity perspectives on the coders, we measure the inter-rater reliability via pairwise ROGUE scores (Eq. (2)) among the coder agents during the coding and theme development stage within Thematic-LM. For example, during the coding stage, we compare codes between each pair of agents by calculating the ROGUE scores. The final scores for each system are the average scores from the coding and theme development stages. We calculate the ROGUE scores as the average between ROGUE-1 and ROGUE-2 scores. We compare the differences between assigning no identities to coder agents, assigning the same identities to coder agents and assigning different identities to coder agents. For the first system, we assign five coders for coding and theme development without any identities. Second, we assign the five coders the same identities of the "human-driven climate change" view to measure the effect of having the same identity perspectives. Third, we assign the five different identities to the coders to measure the divergence of perspectives. As shown in Fig. 4, the agents with different identities produced divergent codes and themes with overall lower ROGUE scores than agents with no identities assigned, while the agents with the same identities produced more similar codes than agents with no identities assigned. For agents with no identities assigned, there are some variations in the codes and themes, as the ROGUE scores indicate there are about 58% to 77% of overlap words and word pairs. While the codes and themes of the agents with different identities diverge from each other, there is a relatively higher overlap of codes and themes for agents with related views. For example, there is some overlap between codes from human-driven climate change and progressive views, such as codes related to collective action to reverse the effect of pollution on climate change.

Although in the Thematic-LM with five different coders, the codes and themes diverge from each other, we have instructed the code and theme aggregator to retain the different codes and themes to maintain the different perspectives. As a result, the resulting themes are more diverse than agents with no agent identities assigned. For Thematic-LM with five different agent identities, 15 themes are identified from the Reddit climate change dataset. As shown in Table 5, we illustrate examples of themes that are not captured in Thematic-LM with no given coder identities but have emerged in Thematic-LM with different coder identities. We observe that with different identity perspectives, the agents might highlight unaddressed issues by considering different viewpoints. For example, themes such as "Economic Impact of Climate Policies" and "Scepticism of Climate Science" reflect concerns and beliefs that differ from those captured in a more homogenized analysis without different identities.

5 Conclusion

We presented Thematic-LM, the first LLM-based multi-agent system for large-scale thematic analysis. Thematic-LM addresses key challenges in the computational thematic analysis of large-scale datasets by distributing the tasks among specialised agents, maintaining an adaptive codebook and assigning different identity perspectives. We employ Thematic-LM to analyse the Dreaddit and Reddit climate change datasets. Our work lays a foundation for conducting qualitative research with LLM agents. Future work could investigate combining other qualitative methods and incorporating the imaging modality into the analysis.

Acknowledgments

Tingrui Qiao is supported by the University of Auckland Doctoral Scholarship. This work is supported by the Our Voices study, funded by an Endeavour grant (UOAX1912) by the Ministry of Business, Innovation and Employment (2019-2025). The views reported in this paper are those of the authors and do not necessarily represent the views of the Our Voices Investigators. Funders had no role in this article's design, analysis or writing.

References

[1] Aly Abdelrazek, Yomna Eid, Eman Gawish, Walaa Medhat, and Ahmed Hassan. 2023. Topic modeling algorithms and applications: A survey. *Information Systems* 112 (2023), 102131.

[2] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774* (2023).

[3] Richard P Allan, Mathew Barlow, Michael P Byrne, Annalisa Cherchi, Hervé Douville, Hayley J Fowler, Thian Y Gan, Angeline G Pendergrass, Daniel Rosenfeld, Abigail LS Swann, et al. 2020. Advances in understanding large-scale responses of the water cycle to climate change. *Annals of the New York Academy of Sciences* 1472, 1 (2020), 49–75.

[4] Stuart J Blair, Yaxin Bi, and Maurice D Mulvenna. 2020. Aggregated topic models for increasing social media topic coherence. *Applied Intelligence* 50 (2020), 138–156.

[5] Virginia Braun and Victoria Clarke. 2006. Using thematic analysis in psychology. *Qualitative research in psychology* 3, 2 (2006), 77–101.

[6] Virginia Braun and Victoria Clarke. 2012. *Thematic analysis.* American Psychological Association.

[7] Jose Camacho-collados, Kiamehr Rezaee, Talayeh Riahi, Asahi Ushio, Daniel Loureiro, Dimosthenis Antypas, Joanne Boisson, Luis Espinosa Anke, Fangyu Liu, and Eugenio Martínez Cámara. 2022. TweetNLP: Cutting-Edge Natural Language Processing for Social Media. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Wanxiang Che and Ekaterina Shutova (Eds.). Association for Computational Linguistics, Abu Dhabi, UAE, 38–49. https://doi.org/10.18653/v1/2022.emnlp-demos.5

[8] Koyel Chakraborty, Siddhartha Bhattacharyya, and Rajib Bag. 2020. A survey of sentiment analysis from social media data. *IEEE Transactions on Computational Social Systems* 7, 2 (2020), 450–464.

[9] Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. 2023. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201* (2023).

[10] Weize Chen, Yusheng Su, Jingwei Zuo, Cheng Yang, Chenfei Yuan, Chi-Min Chan, Heyang Yu, Yaxi Lu, Yi-Hsin Hung, Chen Qian, Yujia Qin, Xin Cong, Ruobing Xie, Zhiyuan Liu, Maosong Sun, and Jie Zhou. 2024. AgentVerse: Facilitating Multi-Agent Collaboration and Exploring Emergent Behaviors. In *The Twelfth International Conference on Learning Representations.* https://openreview.net/forum?id=EHg5GDnyq1

[11] Stephan A Curiskis, Barry Drake, Thomas R Osborn, and Paul J Kennedy. 2020. An evaluation of document clustering and topic modelling in two online social networks: Twitter and Reddit. *Information Processing & Management* 57, 2 (2020), 102034.

[12] Shih-Chieh Dai, Aiping Xiong, and Lun-Wei Ku. 2023. LLM-in-the-loop: Leveraging Large Language Model for Thematic Analysis. In *The 2023 Conference on Empirical Methods in Natural Language Processing.* https://openreview.net/forum?id=NuMemgzPYT

[13] Stefano De Paoli. 2023. Performing an inductive thematic analysis of semi-structured interviews with a large language model: an exploration and provocation on the limits of the approach. *Social Science Computer Review* (2023), 08944393231220483.

[14] Antoine Dechezleprêtre, Adrien Fabre, Tobias Kruse, Bluebery Planterose, Ana Sanchez Chico, and Stefanie Stantcheva. 2022. *Fighting climate change: International attitudes toward climate policies.* Technical Report. National Bureau of Economic Research.

[15] Thomas Dietz. 2020. Political events and public views on climate change. *Climatic Change* 161, 1 (2020), 1–8.

[16] Jakub Drápal, Hannes Westermann, Jaromir Savelka, et al. 2023. Using Large Language Models to Support Thematic Analysis in Empirical Legal Studies.. In *JURIX*. 197–206.

[17] Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. 2023. Improving factuality and reasoning in language models through multiagent debate. *arXiv preprint arXiv:2305.14325* (2023).

[18] Linda Finlay. 2021. Thematic analysis:: the 'good', the 'bad'and the 'ugly'. *European Journal for Qualitative Research in Psychotherapy* 11 (2021), 103–116.

[19] Norjihan Abdul Ghani, Suraya Hamid, Ibrahim Abaker Targio Hashem, and Ejaz Ahmed. 2019. Social media big data analytics: A survey. *Computers in Human behavior* 101 (2019), 417–428.

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

[21] Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, Chenyu Ran, Lingfeng Xiao, Chenglin Wu, and Jürgen Schmidhuber. 2024. MetaGPT: Meta Programming for A Multi-Agent Collaborative Framework. In *The Twelfth International Conference on Learning Representations.* https://openreview.net/forum?id=VtmBAGCN7o

[22] Huiqiang Jiang, Qianhui Wu, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2023. LLMLingua: Compressing Prompts for Accelerated Inference of Large Language Models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, Houda Bouamor, Juan Pino, and Kalika Bali (Eds.). Association for Computational Linguistics, Singapore, 13358–13376. https://doi.org/10.18653/v1/2023.emnlp-main.825

[23] Huiqiang Jiang, Qianhui Wu, Xufang Luo, Dongsheng Li, Chin-Yew Lin, Yuqing Yang, and Lili Qiu. 2024. LongLLMLingua: Accelerating and Enhancing LLMs in Long Context Scenarios via Prompt Compression. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, Lun-Wei Ku, Andre Martins, and Vivek Srikumar (Eds.). Association for Computational Linguistics, Bangkok, Thailand, 1658–1677. https://doi.org/10.18653/v1/2024.acl-long.91

[24] Mikhail V Koroteev. 2021. BERT: a review of applications in natural language processing and understanding. *arXiv preprint arXiv:2103.11943* (2021).

[25] Irene Korstjens and Albine Moser. 2018. Series: Practical guidance to qualitative research. Part 4: Trustworthiness and publishing. *European Journal of General Practice* 24, 1 (2018), 120–124.

[26] Grgur Kovač, Rémy Portelas, Peter Ford Dominey, and Pierre-Yves Oudeyer. 2023. The socialai school: Insights from developmental psychology towards artificial socio-cultural agents. *arXiv preprint arXiv:2307.07871* (2023).

[27] Helvi Kyngäs, Maria Kääriäinen, and Satu Elo. 2020. The trustworthiness of content analysis. *The application of content analysis in nursing science research* (2020), 41–48.

[28] Guohao Li, Hasan Abed Al Kader Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. 2023. CAMEL: Communicative Agents for "Mind" Exploration of Large Language Model Society. In *Thirty-seventh Conference on Neural Information Processing Systems.* https://openreview.net/forum?id=3IyL2XWDkG

[29] Tian Liang, Zhiwei He, Wenxiang Jiao, Xing Wang, Yan Wang, Rui Wang, Yujiu Yang, Zhaopeng Tu, and Shuming Shi. 2023. Encouraging divergent thinking in large language models through multi-agent debate. *arXiv preprint arXiv:2305.19118* (2023).

[30] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out.* Association for Computational Linguistics, Barcelona, Spain, 74–81. https://aclanthology.org/W04-1013

[31] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692 [cs.CL] https://arxiv.org/abs/1907.11692

[32] H Damon Matthews, Andrew J Weaver, Katrin J Meissner, NP Gillett, and M Eby. 2004. Natural and anthropogenic climate change: incorporating historical land cover change, vegetation dynamics and the global carbon cycle. *Climate Dynamics* 22 (2004), 461–479.

[33] Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica* 22, 3 (2012), 276–282.

[34] Leland McInnes, John Healy, Steve Astels, et al. 2017. hdbscan: Hierarchical density based clustering. *J. Open Source Softw.* 2, 11 (2017), 205.

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

[36] Saul McLeod. 2007. Maslow's hierarchy of needs. *Simply psychology* 1, 1-18 (2007).

[37] Federico Neri, Carlo Aliprandi, Federico Capeci, and Montserrat Cuadros. 2012. Sentiment analysis on social media. In *2012 IEEE/ACM international conference on advances in social networks analysis and mining.* IEEE, 919–926.

[38] Lorelli S Nowell, Jill M Norris, Deborah E White, and Nancy J Moules. 2017. Thematic analysis: Striving to meet the trustworthiness criteria. *International journal of qualitative methods* 16, 1 (2017), 1609406917733847.

[39] Bo Pang, Tingrui Qiao, Caroline Walker, Chris Cunningham, and Yun Sing Koh. 2025. LIBRA: Measuring Bias of Large Language Model from a Local Context. In *Advances in Information Retrieval: 47th European Conference on Information Retrieval, ECIR 2025, Lucca, Italy, April 06–10, 2025, Proceedings* (Lucca, Italy). Springer-Verlag, Berlin, Heidelberg.

[40] Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. 2023. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology.* 1–22.

[41] Panu Pihkala. 2020. Anxiety and the ecological crisis: An analysis of eco-anxiety and climate anxiety. *Sustainability* 12, 19 (2020), 7836.

[42] N Reimers. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. *arXiv preprint arXiv:1908.10084* (2019).

[43] Timo Reuter and Philipp Cimiano. 2012. Event-based classification of social media streams. In *Proceedings of the 2nd ACM International Conference on Multimedia Retrieval.* 1–8.

[44] K Andrew R Richards and Michael A Hemphill. 2018. A practical guide to collaborative qualitative data analysis. *Journal of Teaching in Physical education* 37, 2 (2018), 225–231.

[45] Lei Tang and Huan Liu. 2011. Leveraging social media networks for classification. *Data mining and knowledge discovery* 23 (2011), 447–478.
[46] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).
[47] Elsbeth Turcan and Kathleen McKeown. 2019. Dreaddit: A reddit dataset for stress analysis in social media. *arXiv preprint arXiv:1911.00133* (2019).
[48] Kyle Whyte. 2017. Indigenous climate change studies: Indigenizing futures, decolonizing the Anthropocene. *English Language Notes* 55, 1 (2017), 153–162.
[49] Kyle Whyte. 2020. Too late for indigenous climate justice: Ecological and relational tipping points. *Wiley Interdisciplinary Reviews: Climate Change* 11, 1 (2020), e603.
[50] Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Shaokun Zhang, Erkang Zhu, Beibin Li, Li Jiang, Xiaoyun Zhang, and Chi Wang. 2023. Autogen: Enabling next-gen llm applications via multi-agent conversation framework. *arXiv preprint arXiv:2308.08155* (2023).
[51] Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. 2023. Examining Inter-Consistency of Large Language Models Collaboration: An In-depth Analysis via Debate. In *The 2023 Conference on Empirical Methods in Natural Language Processing*. https://openreview.net/forum?id=XEwQ1fDbDN
[52] Lin Yue, Weitong Chen, Xue Li, Wanli Zuo, and Minghao Yin. 2019. A survey of sentiment analysis in social media. *Knowledge and Information Systems* 60 (2019), 617–663.
[53] Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. 2023. Exploring Collaboration Mechanisms for LLM Agents: A Social Psychology View. *CoRR* abs/2310.02124 (2023). https://doi.org/10.48550/ARXIV.2310.02124
[54] Qinlin Zhao, Jindong Wang, Yixuan Zhang, Yiqiao Jin, Kaijie Zhu, Hao Chen, and Xing Xie. 2024. CompeteAI: Understanding the Competition Dynamics of Large Language Model-based Agents. In *Forty-first International Conference on Machine Learning*.
[55] Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric Xing, et al. 2023. Judging llm-as-a-judge with mt-bench and chatbot arena. *Advances in Neural Information Processing Systems* 36 (2023), 46595–46623.
[56] Xujuan Zhou, Xiaohui Tao, Md Mostafijur Rahman, and Ji Zhang. 2017. Coupling topic modelling in opinion mining for social media analysis. In *Proceedings of the international conference on web intelligence*. 533–540.

A Limitations

Evaluation of Qualitative Results. While our evaluation framework provides a foundation for the computational evaluation of qualitative results, there are a few limitations. The credibility and confirmability measured by the evaluator agents might produce biased results due to self-enhancement bias [55], as LLMs may be unaware of their own potential mistakes. Furthermore, transferability is constrained by the availability of suitable datasets for evaluation, as assessing generalizability beyond the original dataset remains challenging. While splitting a dataset into subsets provides an internal measure of transferability, it does not fully capture the ability of themes to generalize across different contexts or domains.
Fidelity of Identity Perspectives. We have employed different identity perspectives in the profiling of coder agents to improve the diversity of themes in Thematic-LM. However, results from the agents with assigned perspectives should not be interpreted as representative of the population with these viewpoints. While Thematic-LM demonstrates the ability to simulate diverse coder perspectives, its fidelity is inherently constrained by the underlying LLM's training data and its representation of those perspectives. This limitation is particularly significant for underrepresented or marginalized viewpoints, such as Indigenous perspectives, where the risk of oversimplification or misrepresentation is heightened. Existing research [39] has found that LLMs encounter knowledge boundaries when handling mixed use of English and Indigenous languages, which may further limit their ability to capture these views authentically. To address these limitations, future research may focus on the evaluation of fidelity of identity perspectives, such as incorporating community-driven feedback mechanisms and validation processes, where individuals from relevant communities review and refine outputs.

B Main Prompts

The prompt for the coder with no identity given is shown below:
"You are a coder in thematic analysis of social media data. When given a social media post, write 1-3 codes for the post. The code should capture concepts or ideas with the most analytical interests. For each code, extract a quote from the post corresponding to the code. The quote needs to be an extract from a sentence. Output the codes and quotes in the following format..."
The prompt for the aggregator is shown below:
"You are an aggregator coder in the thematic analysis of social media data. Your job is to take the codes and corresponding quotes from other coders, merge the similar codes and retain the different ones. Store the quotes under the merged codes, and keep the top {K} most relevant quotes. Output the codes and quotes in JSON format. Don't output anything else. Quote_id is the same as data_id. Example..."
The prompt for the reviewer is shown below:
"You are a review coder in the thematic analysis of social media data. Your job is to review the previously coded data with new codes, merge similar codes, and give them more representative codes. You will be given two items. The first contains new codes and quotes; the second contains similar codes and corresponding quotes to each new code. Decide if there are previously similar coded data with the same meaning that can be merged with the new codes. Update the new code according to the previous code if needed. If the previous codes are all different or there are no similar codes, leave the merge_codes empty in the output. Output the updated codes and quotes in JSON format..."
The prompt for the theme coder is shown below:
"You are a coder in the thematic analysis of social media data. Your job is to develop themes from codes and their corresponding quotes from the data. When given the codebook in JSON with codes and quotes, identify themes which reflect deeper meanings of the data. For each theme, write one sentence to describe what the theme talks about. Keep top {K} most relevant quotes; each theme has no more than ten quotes. Output the themes, description and related quotes in the following JSON format..."

C Additional Experiment

Table 6: Comparison of the theme quality scores on the Dreaddit dataset using LLaMA-3-8B.

Method	Credibility & Confirmability	Dependability	Transferability
Single	0.44	0.37	0.35
Single (Codebook)	0.68	0.52	0.50
System (1 Coder)	0.71	0.57	0.56
System (2 Coders)	0.74	0.61	0.63
System (5 Coders)	**0.80**	**0.69**	**0.74**

We observe that while the local, smaller LLaMA-3-8B [46] achieves lower results than GPT-4o, the improvement from the codebook and multi-agent system are consistent with previous results.