

LLMTaxo: Leveraging Large Language Models for Constructing Taxonomy of Factual Claims from Social Media

Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, Chengkai Li

University of Texas at Arlington

{haiqi.zhang, zhengyuan.zhu, zeyu.zhang}@mavs.uta.edu

cli@uta.edu

Abstract

With the rapid expansion of content on social media platforms, analyzing and comprehending online discourse has become increasingly complex. This paper introduces LLMTaxo, a novel framework leveraging large language models for the automated construction of taxonomies of factual claims from social media by generating topics at multiple levels of granularity. The resulting hierarchical structure significantly reduces redundancy and improves information accessibility. We also propose dedicated taxonomy evaluation metrics to enable comprehensive assessment. Evaluations conducted on three diverse datasets demonstrate LLMTaxo’s effectiveness in producing clear, coherent, and comprehensive taxonomies. Among the evaluated models, GPT-4o mini consistently outperforms others across most metrics. The framework’s flexibility and low reliance on manual intervention underscore its potential for broad applicability.

1 Introduction

Misinformation, also known as false or misleading information (Wu et al., 2019), has the potential to sway public perception, cause confusion, and influence people’s decision-making processes (Del Vicario et al., 2016; Muhammed T and Mathew, 2022). Social media platforms, in particular, facilitate the rapid sharing of vast amounts of content, blending accurate information with falsehoods (Allcott et al., 2019). Social media’s global reach and ease of use have transformed how millions of users exchange opinions, news, and factual claims in real-time, making it fertile ground for misinformation (Aïmeur et al., 2023). Factual claims, which are assertions that can be verified as either true or false, are a common vehicle for misinformation (Ni et al., 2024). These claims, whether accurate or not, have a profound societal impact, as the public tends to believe a factual claim is true regardless of

its truthfulness (Moravec et al., 2018; Ognyanova et al., 2020; Xiao et al., 2021; Zhu et al., 2022; Zhang et al., 2024a; Zhu et al., 2025).

The dynamic nature of social media often leads to repetitive or reformulated claims, complicating the identification and validation of factual content (Zhou et al., 2015), making it difficult for non-technical individuals and researchers from different fields to navigate the vast amounts of data (Suarez-Lledo and Alvarez-Galvez, 2021; Hook and Verdeja, 2022; Muhammed T and Mathew, 2022). This challenge calls for scalable and automated tools to systematically organize and analyze factual claims, helping stakeholders—such as researchers and fact-checkers—more effectively navigate the complex information landscape.

One promising approach to addressing this challenge is the use of taxonomies, which provide structured systems for organizing complex information. Taxonomy has been applied in misinformation categorization. For example, Tambini (2017) classified fake news into distinct categories to capture the diverse forms and manifestations of the phenomenon. Kumar and Shah (2018) constructed taxonomies of false information on social media based on various characteristics, such as whether the content is intended to deceive and whether it is opinion-based or fact-based. Zhao and Tsang (2022) proposed a taxonomy of misinformation on social media based on falsity level and evidence type. Yet, these approaches do not construct topic taxonomies, which are particularly valuable for exploring and organizing factual claims. In practice, many fact-checking websites, such as PolitiFact,¹ Snopes,² and Full Fact,³ use topics to organize fact checks. For example, the claim “*Right to Try experimental drug program saved thousands and thousands of lives*” in PolitiFact is given topics such as “Elections” and

¹<https://www.politifact.com/>

²<https://www.snopes.com/>

³<https://fullfact.org/>

“Health Care.” However, these topics are too broad to help users efficiently locate specific claims.

Motivated by the challenges, prior studies, and practical applications discussed above, this paper introduces LLMTaxo, a novel framework utilizing large language models (LLMs) for automatically constructing a topic taxonomy given a collection of factual claims from social media within a specific topic domain. The taxonomy starts with broad categories and branches into increasingly specific subcategories. In this way, factual claims can be categorized into hierarchical categories, allowing for the efficient exploration of information at multiple levels of granularity. For example, broad categories can group COVID-19 vaccine-related claims into overarching topics such as “Public Health,” while more detailed categories can address specific claims about “Vaccine Safety and Effectiveness.”

Our framework, LLMTaxo, organizes factual claims into a coherent hierarchical taxonomy by clustering semantically similar claims, identifying distinct ones, and generating topics at multiple levels of granularity—broad, medium, and detailed. Leveraging the rich background knowledge of LLMs and few-shot learning, LLMTaxo minimizes human involvement and automates the categorization process. It effectively addresses challenges such as semantic variability (the expression of similar ideas in diverse ways) and ensures scalability and adaptability to social media discourse across various datasets and topic domains.

To demonstrate the generalizability of LLMTaxo, we conducted evaluations using carefully designed metrics across three distinct datasets spanning different topic domains and data sources. These include social media posts from *X* (formerly Twitter) and Facebook, covering domains such as COVID-19 vaccines, climate change, and cybersecurity. The evaluation results, showing all models scoring above 3.0 out of 5 across all metrics, with a highest score of 4.9, demonstrate the effectiveness of LLMTaxo in generating clear and comprehensive taxonomies. By identifying distinct factual claims, the framework substantially reduces redundancy. The hierarchical structure of the taxonomy produced by LLMTaxo allows users to explore claims at multiple levels of detail. Moreover, LLMTaxo exhibits consistently strong performance across diverse datasets, demonstrating its adaptability and potential for wide-ranging applications.

In summary, this paper makes several key contributions:

- We introduce the first multi-level taxonomy of factual claims from social media constructed using LLMs. This taxonomy can be integrated into fact-checking workflows and applied across various research domains.

- We are the first to utilize LLMs for generating multi-granularity topics.

- We develop a set of evaluation metrics for comprehensively assessing the structural quality of taxonomies and their claim-topic alignment.

- We evaluate our taxonomies across three diverse datasets. The results demonstrate LLMTaxo’s adaptability to different domains while maintaining accuracy in constructing meaningful taxonomies.

- We release a curated dataset and a public codebase at <https://github.com/idirlab/LLMTaxo>. They are valuable for supporting both future research and reproducibility.

2 Related Work

Taxonomy Construction. Although taxonomy construction has been extensively studied, the definitions of specific problems vary. Generally, taxonomies are hierarchically structured classifications of concepts, terms, and entities that help users organize, retrieve, and navigate information (Carion et al., 2019; Yang, 2012). Generic taxonomy construction tasks typically involve short concept terms or entity names, which are often structured as hypernym-hyponym pairs (Zhang et al., 2018; Huang et al., 2020).

Constructing taxonomies from broader, less formatted content, such as social media posts, differs from traditional taxonomy construction. The inherent variability of such content makes it challenging to establish a precise taxonomy. Several studies have attempted to address this problem. Durham et al. (2023) explored automatic taxonomy generation from disaster-related tweets using topic modeling techniques (Blei et al., 2003; Dumais, 2004). Najem and Hadi (2021) proposed semi-automatic ontology construction of tweets based on semantic feature extraction using WordNet and BabelNet (Miller, 1995; Navigli and Ponzetto, 2010).

The rise of LLMs has significantly advanced many natural language processing tasks, but limited effort has been devoted to taxonomy construction. Chen et al. (2020) employed pretrained language models to construct taxonomic trees, while Chen et al. (2023) compared prompting and fine-tuning approaches for hypernym taxonomy con-

struction. Additionally, [Shah et al. \(2023\)](#) and [Wan et al. \(2024\)](#) introduced end-to-end pipelines that integrate LLMs to generate, refine, and apply labels for user intent analysis in log data. Our work extends this line of research by applying LLMs to generate multi-level taxonomies for factual claims and social media posts. Like prior efforts, we harness LLMs’ generative and semantic capabilities, but we focus on taxonomy construction in open-domain, noisy, and high-variance contexts where entity boundaries and hierarchical relationships are less clear.

Topic Generation. LLMs have emerged as a promising alternative to traditional topic generation approaches. They can generate topics from a given set of documents without requiring predefined labels or training data ([Sarkar et al., 2023](#)). This capability allows for more flexible and context-aware topic extraction. For instance, [Mu et al. \(2024\)](#) introduced a framework that prompts LLMs to generate topics and adhere to human guidelines for refining and merging topics. Recent research has also explored different LLM-based topic modeling techniques. For example, BERTopic ([Grootendorst, 2022](#)) has shown superior performance in terms of diversity and coherence across multiple datasets ([Jung et al., 2024](#)). Yet, to the best of our knowledge, no existing work has generated topics at multiple levels of granularity.

3 Methodology

Analyzing social media data presents numerous challenges, including the overwhelming volume of content with high repetition and the substantial human effort required to explore the data. Our LLMTaxo framework is designed to systematize online content through automated construction of a taxonomy of factual claims. This taxonomy serves as a hierarchical classification system. It enhances the accessibility and navigability of information for users by categorizing claims into broad, medium, and detailed topics.

LLMTaxo initially identifies factual claims from social media posts, and subsequently clusters similar claims to discern and select distinct ones, thereby reducing redundancy. It then leverages LLMs to generate topics for each distinct claim at multiple levels of granularity, ultimately constructing a structured taxonomy. The overview of our framework is shown in Figure 1.

3.1 Taxonomy Construction

Our primary goal is to construct a hierarchical taxonomy. The hierarchical design is informed by the prototype theory ([Geeraerts, 2006](#)), which guides cognitive categorization, i.e., the three-level hierarchy of the taxonomy. Formally, a hierarchical taxonomy $\mathcal{T} = (T_b, T_m, T_d, f_m, f_d)$ contains topics at multiple levels of granularity, ranging from broader to more fine-grained topics. Specifically, the taxonomy uses a three-tiered structure, consisting of broad topics T_b (representing general themes), medium topics T_m (reflecting intermediate distinctions), and detailed topic T_d (highlighting finer aspects). The functions f_m and f_d define the hierarchical relationships between topics, where $f_m : T_m \rightarrow T_b$ maps each medium topic to its corresponding broad topic, and $f_d : T_d \rightarrow T_m$ maps each detailed topic to its respective medium topic.

Note that the same topic label may appear under different parent topics, but these instances represent conceptually distinct topics. For example, the medium topic “Vaccine Safety” under broad topic “Public Opinion” is essentially “Public Opinion about Vaccine Safety,” which is meaningfully different from “Vaccine Safety” under “Government Policies.” To minimize redundancy while preserving clarity, we use concise topic labels for medium and detailed topics, as their inherent meanings are informed by their parent topics.

Given a collection of factual claims $\mathcal{C} = \{c_1, c_2, \dots, c_n\}$, each claim $c \in \mathcal{C}$ is assigned a tuple of topics at three levels—broad, medium, and detailed—with a mapping function ϕ . More formally, $\phi(c) = (t_b, t_m, t_d)$ where $t_b \in T_b$, $t_m \in T_m$, and $t_d \in T_d$. Note that the detailed topic is optional because some claims may be too brief or ambiguous to be categorized distinctly at all three levels.

For example, the right side of Figure 1 illustrates the topic tuple for a claim related to COVID-19 vaccine: “Myocarditis is up TEN times due to the Covid Vaccine...” The broad topic t_b is “Vaccine Safety and Effectiveness,” the medium topic t_m is “Vaccine Side Effects,” and the detailed topic t_d is “Myocarditis Side Effect.”

To automate the process of taxonomy construction given a collection of factual claims \mathcal{C} , we create learning examples that contain both sample claims and their corresponding topics. The topics from the learning examples form a seed taxonomy, which serves as a foundation for LLM-based expansion. To generate topics for each distinct factual

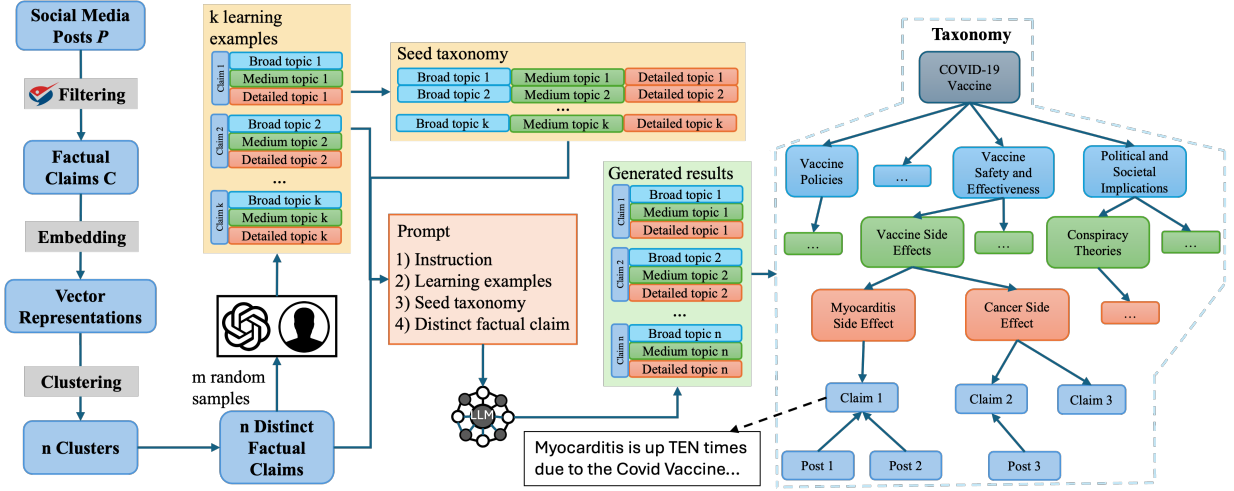


Figure 1: The LLMTaxo framework.

claim in \mathcal{C} , we prompt the LLMs with both the learning examples and the seed taxonomy. After generating topics for all claims, we consolidate them to construct a refined taxonomy. This process is illustrated in Figure 1.

3.1.1 Learning Examples and Seed Taxonomy

While LLMs possess broad general knowledge, they are not pretrained to directly generate hierarchical topics for factual claims at different granularities. Our initial experiments showed that LLMs often generate inconsistent topics for similar claims, resulting in an excessive and unwieldy number of topics that hinder comprehension, as further detailed in Section 4.4. To mitigate this issue, we propose to expand and refine a seed taxonomy \mathcal{S} derived from a set of sample factual claims. This initial taxonomy aids the LLMs by providing a foundation to expand upon. We leverage an LLM with human-in-the-loop supervision to create k learning examples $\mathcal{L} = \{\langle c_1, (t_b^{(1)}, t_m^{(1)}, t_d^{(1)} \rangle), \dots, \langle c_k, (t_b^{(k)}, t_m^{(k)}, t_d^{(k)} \rangle)\}$, representing claims with their respective topics at each level of granularity. The 3-level topic tuples from the k claims form the seed taxonomy, denoted as $\mathcal{S} = \{(t_b^{(1)}, t_m^{(1)}, t_d^{(1)}), \dots, (t_b^{(k)}, t_m^{(k)}, t_d^{(k)})\}$. These learning examples and the seed taxonomy facilitate few-shot in-context learning, helping to stabilize the variation and number of topics generated. The size k of learning examples varies across datasets.

To compile the learning examples and the seed taxonomy, we randomly select a subset $\mathcal{R} \subset \mathcal{C}$ containing m distinct claims for annotation. Recognizing that direct human annotation is labor-intensive and the annotation requires high accuracy and con-

sistency, we employ GPT-3.5 (Brown et al., 2020) to assist human in topic annotation for these claims.

Each claim $c \in \mathcal{R}$ is processed by an LLM prompted to generate topics at three levels of granularity. Below is an example of the prompt used for the COVID-19 vaccine topic domain.

You will be given a tweet related to COVID-19 vaccine. Please generate topics for the tweet from different granularities such as broad topic, medium topic, and detailed topic. Each generated topic should be no more than eight words and you should try to minimize the number of topics generated.

Examples: ...

LLMTaxo adopts a multi-round human-in-the-loop annotation process. In each round, the LLM generates topics for a batch of 10 claims. These topics are then evaluated and refined by annotators, resulting in a finalized annotation of each claim c with topic tuple (t_b, t_m, t_d) , forming a claim-topic pair $\langle c, (t_b, t_m, t_d) \rangle$. The LLM may produce diverse topic labels in the initial run. To ensure consistency and minimize topic variation, we restrict the number of distinct topics and standardize topic labels across similar claims during the refinement process. For example, if the LLM generates “COVID-19 Vaccine Mandates” for claim A and “Vaccine Mandates” for claim B, the annotators are expected to retain a single, generalized topic (e.g., “Vaccine Mandates”) for both claims.

The refined claim-topic pairs are subsequently used as in-context examples to guide the LLM in the next round. This process is iterated until topics are generated for all $c \in \mathcal{R}$. The human-in-the-loop process reduces manual effort and allows for tailored annotations based on specific requirements,

which is adaptable across various topic domains.

Upon annotating all m claims, the annotators review the most frequent broad topics and manually pick k representative claim-topic pairs as the learning examples \mathcal{L} . The topics from these k examples form the seed taxonomy \mathcal{S} . The seed taxonomy and the learning examples guide the LLMs to generalize the task of topic generation across a diverse range of claims while ensuring consistent and structured outputs.

3.1.2 Multi-level Topic Generation

To automate the taxonomy construction process and minimize the human effort, we employed LLMs to generate topics (t_b, t_m, t_d) for each factual claim $c \in \mathcal{C}$. The learning examples \mathcal{L} and the seed taxonomy \mathcal{S} are utilized as part of the prompt for the LLM. Specifically, the prompt consists of $l_i \in \mathcal{L}$ for $i = 1, 2, \dots, k$ (where $l_i = \langle c_i, (t_b^{(i)}, t_m^{(i)}, t_d^{(i)}) \rangle$ is the claim-topic pair of the i -th learning example), the seed taxonomy \mathcal{S} , the instruction and questions that ask the LLM to produce the topic tuple for the claim c . Due to the limited context length of LLMs, each prompt generates the topic tuple $\phi(c) = (t_b, t_m, t_d)$ for only one c . This generation process is iterated until finishing generating topic tuples for all $c \in \mathcal{C}$. The prompt is detailed in Figure 2 in Appendix A.

3.1.3 Topic Consolidation

After the LLM generates the topic tuples for all claims, LLMTaxo consolidates the results to build the taxonomy. Given that (t_b, t_m, t_d) are generated as the topic tuple for a claim, t_b , t_m , and t_d inherently form parent-child relationships. Medium topics that align with the same broad topic are treated as child nodes of that broad topic, and detailed topics are similarly considered child nodes of their respective medium topics. Formally, from the generated topic tuples, LLMTaxo derives the two mapping functions $f_m: T_m \rightarrow T_b$ and $f_d: T_d \rightarrow T_m$. f_m maps each medium topic to a broad topic: $f_m(t_m) = t_b$ if $\exists c \in \mathcal{C}$ such that $\phi(c) = (t_b, t_m, -)$. Similarly, f_d maps each detailed topic to a medium topic: $f_d(t_d) = t_m$ if $\exists c \in \mathcal{C}$ such that $\phi(c) = (-, t_m, t_d)$.

For instance, consider two claims c_1 and c_2 with the generated topic tuples $\phi(c_1) = (\text{Vaccine Safety and Effectiveness}, \text{Vaccine Side Effects}, t_d^{(1)})$ and $\phi(c_2) = (\text{Vaccine Safety and Effectiveness}, \text{Vaccine Injury}, t_d^{(2)})$. The derived mapping function f_m will include $f_m(\text{Vaccine Side Effects}) = \text{Vaccine Safety and Effectiveness}$ and $f_m(\text{Vaccine In-$

jury”) = “Vaccine Safety and Effectiveness.” This establishes that “Vaccine Side Effects” and “Vaccine Injury” are child nodes of “Vaccine Safety and Effectiveness.”

3.2 Claim Detection

Social media contains a wide range of content, including personal opinions, personal experiences, and entertainment. Prior to taxonomy construction, we apply claim detection to identify social media posts that are more likely to contain factual claims, which could potentially carry misinformation. We employ the ClaimBuster (Hassan et al., 2017) model which assigns a score to each sentence or paragraph, indicating the likelihood of it being a check-worthy factual claim. A higher score suggests a greater likelihood of check-worthiness. We set a threshold of 0.5 as it effectively balances precision and recall in identifying check-worthy claims. Posts that score above this threshold are retained for further steps of LLMTaxo. This approach reduces the dataset to a more manageable subset (see Section 4.1 for details). It also ensures that the retained posts are more likely to contain check-worthy factual information relevant to our taxonomy construction.

3.3 Identifying Distinct Claims

Many factual claims on social media are frequently repeated or rephrased. For example, the posts “*BREAKING: Pentagon rescinds COVID-19 vaccine mandate*” and “*The Pentagon officially rescinds COVID-19 vaccine mandate*” convey the same claim but are phrased slightly differently. To reduce redundancy and focus on unique claims, we apply clustering to group identical or nearly identical factual claims. We use HDBSCAN (Campello et al., 2013) due to its ability to handle noise and detect outliers, which is particularly useful given that many posts do not closely resemble others. To capture the semantic meaning of claims, we employ Sentence-BERT (Reimers, 2019) to generate dense vector representations. After clustering, we identify distinct claims by selecting the first post from each returned cluster list while excluding the outlier cluster. A distinct claim is the representation of a cluster of similar factual claims, and the identified distinct claims \mathcal{C} are used for taxonomy construction. The outliers represent infrequently discussed content, whereas we only focus on content that appears multiple times. Repeated exposure to information increases belief in its accuracy (Pen-

nycook et al., 2018), underscoring the significance of identifying widely circulated claims.

4 Experiments

4.1 Datasets

To evaluate our method, we conducted experiments on three social media datasets, each covering a specific topic: COVID-19 Vaccine, Climate Change, and Cybersecurity, denoted as *CV*, *CC* and *CS*, respectively. These datasets were collected from two social media platforms to ensure diversity in the content and structural characteristics of the posts.

COVID-19 Vaccine (CV). We collected tweets related to COVID-19 vaccines using Wildfire (Zhang et al., 2024b). For this dataset, we targeted tweets containing various keyword variations related COVID-19 vaccines, such as “covid19 vaccination,” “covid-19 vaccine,” and “covid vax.” The data collection period spanned from January 1, 2023 to April 25, 2023, resulting in a total of 384,676 tweets. After applying claim detection described in Section 3.2, 232,368 tweets were retained for distinct claim identification.

Climate Change (CC). For the climate change dataset, we utilized CrowdTangle (CrowdTangle, 2024)—a now-discontinued tool—to collect Facebook posts related to climate change. We retrieved posts containing the keyword “climate change” between January 1, 2024, and May 7, 2024, yielding a total of 229,913 posts. After applying claim detection, we retained 89,412 posts.

Cybersecurity (CS). We collected Facebook posts related to cybersecurity using CrowdTangle, with the keyword “cybersecurity.” The collection period also spans from January 1, 2024 to May 7, 2024. Initially, 107,905 posts were gathered, and after claim detection, 38,530 posts were retained.

4.2 Implementation Details

For the HDBSCAN clustering model, we set the minimum cluster size to 3 for the *CV* dataset. The minimum cluster size for *CC* and *CS* datasets was set to 2 because they have fewer posts, and we hoped to avoid the majority of posts being classified as outliers. We also set a maximum cluster size of 3,000 to prevent the formation of overly large clusters. We noticed some clusters share identical posts, the reason for which is that Sentence-BERT may generate slightly different vector embeddings for the same sentence (Reimers, 2019). To avoid duplication, we only keep the same post once in

the distinct claims \mathcal{C} . The final numbers of distinct claims for *CV*, *CC*, and *CS* datasets are 8,103, 14,408, and 5,731, respectively.

For taxonomy construction, we employed three LLMs, Zephyr (Tunstall et al., 2023), GPT-4o mini (OpenAI, 2024), and Gemini 2.0 Flash (DeepMind, 2025) for performance comparison. Zephyr is selected for its competitive performance in language understanding tasks among all 7-billion-parameter LLMs (Chiang et al., 2024), while GPT-4o mini and Gemini 2.0 Flash are chosen for their balance of cost-efficiency and performance. For each dataset, we randomly selected 100 distinct factual claims and annotated them with broad, medium, and detailed topics. From annotated factual claims, we then chose representative samples based on their frequency of occurrence to serve as learning examples for the LLMs. These annotated claims were used to guide the LLMs in generating topics for the distinct claims \mathcal{C} identified through clustering. All experiments were conducted on three A100 GPUs.

4.3 Results

4.3.1 Clustering

The statistics of the clusters are shown in Table 1. We used Silhouette Coefficient (Rousseeuw, 1987) to evaluate the clusters’ quality. Since HDBSCAN was employed, the outlier cluster was excluded during the evaluation. The *CV* dataset achieved the highest Silhouette score of 0.940, reflecting highly cohesive and well-separated clusters. Although the *CC* and *CS* datasets have lower scores of 0.488 and 0.554, these still indicate reasonably good cluster quality. The variation in Silhouette scores across different datasets can be attributed to differences in their characteristics, particularly the higher sparsity and greater length of Facebook posts, which make clustering more challenging. These results suggest that the clustering method performs reasonably well across diverse datasets, even with different clustering configurations.

Datasets	Posts	Clusters	Outliers
<i>CV</i>	232,368	10,995	25,962
<i>CC</i>	89,412	15,794	42,923
<i>CS</i>	38,530	7,398	15,946

Table 1: Cluster statistics for different datasets.

4.3.2 Multi-level Topic Generation

The statistics of the generated topics across datasets are presented in Table 2—the rows labeled “w/” under the method column. It is evident that Gemini and GPT-4o mini outperformed Zephyr in constraining the number of broad and medium topics to a narrower range. Gemini also demonstrated superior ability in limiting the number of detailed topics. However, we observed that some generated topics were associated with only a few factual claims. To enhance the taxonomy’s readability, we consolidated broad topics that appear fewer than 50 times into a new broad category labeled “Other.” For medium and detailed topics, we retained those with occurrences exceeding 4 for medium topics and 3 for detailed topics, respectively. Topics with fewer occurrences were grouped under “Other” topic within their respective parent topics. After merging, the topic statistics of each dataset are shown in Table 3. This approach limits the taxonomy to a more manageable size.

Dataset	Model	Method	Broad Topic	Medium Topic	Detailed Topic
CV	Zephyr	w/o	839	1585	6553
		w/	125 (85.1% ↓)	899 (43.3% ↓)	6060 (7.5% ↓)
	GPT-4o mini	w/o	1028	2301	6399
		w/	12 (98.8% ↓)	41 (98.2% ↓)	2073 (67.6% ↓)
	Gemini	w/o	309	1062	5828
		w/	12 (96.1% ↓)	50 (95.3% ↓)	49 (99.2% ↓)
CC	Zephyr	w/o	1046	3831	12977
		w/	124 (88.1% ↓)	1092 (71.5% ↓)	7414 (42.9% ↓)
	GPT-4o mini	w/o	1668	4998	12638
		w/	8 (99.5% ↓)	274 (94.5% ↓)	8722 (31.0% ↓)
	Gemini	w/o	385	2306	12360
		w/	7 (98.2% ↓)	42 (98.2% ↓)	573 (95.4% ↓)
CS	Zephyr	w/o	377	1684	5340
		w/	126 (66.6% ↓)	656 (61.0% ↓)	3376 (36.8% ↓)
	GPT-4o mini	w/o	688	2184	5335
		w/	12 (98.3% ↓)	111 (94.9% ↓)	4887 (8.4% ↓)
	Gemini	w/o	225	1572	5206
		w/	9 (96.0% ↓)	70 (95.5% ↓)	760 (85.4% ↓)

Table 2: Topic counts for different datasets with and without prompting seed taxonomy.

Dataset	Model	Broad Topic	Medium Topic	Detailed Topic
CV	Zephyr	11	66	114
	GPT-4o mini	8	18	110
	Gemini	9	27	12
CC	Zephyr	8	146	163
	GPT-4o mini	7	46	229
	Gemini	7	22	30
CS	Zephyr	10	48	32
	GPT-4o mini	9	25	61
	Gemini	9	32	26

Table 3: Topic counts after merging less frequent topics.

4.4 Ablation Study

To evaluate the effectiveness of the presence of the seed taxonomy in the prompt, we conducted experiments with removing the seed taxonomy from the prompt. Specifically, we only provide the LLMs with instructions, learning examples, and the target factual claims. The total topic counts with and without seed taxonomy are shown in Table 2. We can see that adding the seed taxonomy to the prompt effectively restricts the taxonomy size, reducing the number of broad topics by up to 99.5%.

5 Evaluation

We assessed LLMTaxo from two perspectives: 1) the quality of the generated taxonomy, and 2) the relevance of the generated topics to the factual claims. We developed a set of evaluation metrics to assess these two aspects. We engaged both human evaluators and GPT-4 (Achiam et al., 2023) for the evaluation. The effectiveness of using LLMs for model performance evaluation has been validated by previous studies (Fu et al., 2024; Liu et al., 2023). Identical instructions and metrics were provided to both human and GPT-4 evaluators. They were instructed to rate each criterion on a scale from 1 (strongly disagree) to 5 (strongly agree), where 5 indicates the best quality and 1 the worst.

To evaluate the quality of taxonomies, we presented taxonomies generated by Zephyr, GPT-4o mini, and Gemini across the three datasets to the evaluators, resulting in a total of 9 taxonomies. For the evaluation of the claim-topic suitability, we randomly selected 102 factual claims along with their corresponding lowest-level topics (i.e., leaf nodes of the taxonomy) generated by the three models from each dataset, resulting a total of 306 claim-topic pairs for evaluation. We exclusively evaluated the leaf node topics with claims, as the broader topics had already been assessed in taxonomy quality evaluation. Note that not every claim has three-level topics, as mentioned in Section 3.1. Therefore, we focused our evaluation on the leaf node topics. To mitigate bias, we shuffled claim-topic pairs from different models before presenting them to the evaluators. Both human evaluators and GPT-4 reviewed these pairs using our predefined metrics. Human evaluators were required to provide rationales for the scores. The evaluation prompt for GPT-4 is detailed in Appendix B.

5.1 Taxonomy Evaluation

To design the taxonomy evaluation metrics, we adopted the Goal Question Metric (GQM) approach (Caldiera and Rombach, 1994) and consulted existing metrics from Kaplan et al. (2022). Due to differences in our tasks, we retained only the *orthogonality* and *completeness* metrics from Kaplan et al. (2022). We further refined these metrics to better align with our objectives and introduced additional metrics tailored to our evaluation needs. Each metric is defined with a clear goal, a guiding question, and specified evaluation criteria. An overview of the metrics is provided below, with comprehensive details available in Appendix C.

Clarity. The goal of this metric is to ensure that each topic label communicates its content effectively to avoid confusion. To assess whether the topic labels are clear, precise, and unambiguous, we evaluate precision, unambiguity, consistency, and accessibility.

Hierarchical coherence. The goal is to ensure that the taxonomy’s structure facilitates easy navigation and understanding by clearly organizing information from the most general to the most specific. To assess whether the taxonomy follows a clear and meaningful hierarchical structure, we evaluate gradational specificity, parent-child coherence, and consistency.

Orthogonality. The goal is to maintain clear boundaries between topics, ensuring that each one captures a unique aspect of the overall topic domain. To assess whether the topics are well-differentiated and free from redundancy, we evaluate their distinctiveness and degree of overlap.

Completeness. The goal of this metric is to cover as many areas of the topic domain as possible to ensure the taxonomy is comprehensive. To assess whether the taxonomy captures a broad and representative set of topics across different facets of the topic domain, we evaluate its domain coverage, depth, as well as balance.

Three human evaluators and GPT-4 were provided with evaluation instructions and the background of the taxonomy construction. Each of them rated the taxonomies based on the evaluation criteria. For each metric, we calculated the average score of the evaluation criteria for each human evaluator. We then computed the mean of these averages across the three human annotators. The evaluation scores are presented in Table 4.

The results affirm the overall efficacy of LLMTaxo, with taxonomies generally receiving high ratings (above 3.0 across all metrics). Notably, GPT-4o mini consistently outperformed Zephyr and Gemini, suggesting its greater suitability for taxonomy construction. Human evaluators tended to give higher scores than GPT-4, particularly for taxonomies generated by Zephyr and Gemini, indicating possibly stricter criteria or different interpretations of taxonomy quality by GPT-4. It is also reasonable to assume that GPT-4 shares greater similarity with GPT-4o mini than with Zephyr or Gemini.

The models generally scored well on clarity and completeness, indicating their effectiveness in producing clear, precise and comprehensive topic labels. GPT-4o mini also scored higher in hierarchical coherence with the highest score of 4.7, suggesting it does well in structuring information from general to specific in a meaningful way. Orthogonality has slightly lower scores compared to others, particularly for Zephyr and Gemini, which indicates some overlap or less distinct boundaries between topics. We observed Gemini was more strongly influenced by the seed taxonomy and tended to generate topics closely aligned with it. This is reflected in the limited number of unique topics produced by Gemini, as shown in Table 3 and Table 2. Such alignment negatively affected both orthogonality and completeness. To assess the agreement level among human evaluators, we computed the inter-rater reliability using Gwet’s AC2 (Gwet, 2008) with quadratic weights, which yielded a score of 0.80.

5.2 Claim-Topic Evaluation

To assess how well the factual claims are aligned with the topics assigned to them from the LLM-generated taxonomy, we performed evaluation on two aspects: *accuracy* and *granularity*. As Section 5.1 already evaluated the taxonomy itself, which includes the relationships among broad, medium, and detailed topics, we only evaluated claims and their leaf node topics. The detailed evaluation metrics are:

Accuracy. This criterion assesses how accurately the leaf node topics reflect the content and context of the corresponding factual claims. This involves determining if the topics are relevant and if they correctly represent the underlying information in the claims without misinterpretation or error.

Dataset	Evaluator	Model	M ₁	M ₂	M ₃	M ₄	Ac	Gr
CV	Human	Zephyr	4.2	4.0	3.3	4.3	4.7	3.6
		GPT-4o mini	4.3	4.2	3.8	4.7	4.4	4.6
		Gemini	4.4	3.9	3.5	3.9	4.4	4.6
	GPT-4	Zephyr	3.5	3.7	3.0	3.7	4.0	4.6
		GPT-4o mini	4.3	4.3	4.0	4.7	4.0	4.4
		Gemini	3.4	3.0	3.0	3.3	4.0	4.4
CC	Human	Zephyr	4.6	4.2	3.9	4.6	3.3	3.9
		GPT-4o mini	4.7	4.7	4.2	4.8	4.4	4.5
		Gemini	4.3	3.9	3.3	4.7	4.4	4.6
	GPT-4	Zephyr	3.5	3.3	4.0	4.0	3.4	3.4
		GPT-4o mini	4.3	4.0	4.5	4.3	3.8	4.5
		Gemini	3.8	3.7	3.5	3.7	3.8	4.4
CS	Human	Zephyr	4.3	3.9	3.7	4.4	3.8	4.0
		GPT-4o mini	4.5	4.0	4.0	4.6	4.7	4.6
		Gemini	4.6	3.6	3.5	3.9	4.8	4.6
	GPT-4	Zephyr	3.3	3.0	3.0	3.3	3.7	4.0
		GPT-4o mini	4.3	4.7	3.5	4.7	4.0	4.9
		Gemini	3.0	3.0	3.5	3.3	4.0	4.9

Table 4: Taxonomy and claim-topic pairs evaluation scores for different models across datasets. In the table, M_1 , M_2 , M_3 , M_4 , Ac, and Gr represent Clarity, Hierarchical Coherence, Orthogonality, Completeness, Accuracy, and Granularity, respectively.

Granularity. This criterion evaluates the specificity of the leaf node topics. This involves determining whether the topics are detailed enough to uniquely categorize and differentiate between factual claims, yet broad enough to maintain practical applicability over multiple claims.

Nine human evaluators participated in the claim-topic evaluation. They were divided into three groups, each consisting of three individuals. Each group assessed the same set of 306 claim-topic pairs, comprising 34 pairs per model for each dataset. After completing the evaluations, we calculated the average scores for each model across the datasets, as presented in Table 4.

Overall, GPT-4o mini and Gemini demonstrated stronger performance in both accuracy and granularity compared to Zephyr, particularly on the CS dataset, where they achieved a granularity score of 4.9. According to the feedback from human evaluators, Zephyr often produced lengthy descriptions for detailed topics, occasionally repeating the content of the social media posts, which contributed to its lower granularity scores. However, Zephyr showed slightly higher accuracy in the CV dataset in the human evaluation, indicating it may perform better with shorter texts such as tweets. In contrast, Facebook posts (CC and CS) likely provided more verbose and detailed content, which may have benefited GPT-4o mini and Gemini.

To measure the inter-rater reliability, we computed Gwet’s AC2 scores for each group. The

average scores of the three groups were 0.67 for accuracy and 0.75 for granularity.

5.3 Error Analysis

Based on the feedback from the human evaluators, errors in the taxonomies can be categorized into three types: 1) overlapping of topics, 2) lack of specificity in topic labels, and 3) generation of noisy data. For criteria receiving lower scores in taxonomy evaluation, evaluators highlighted issues caused by overlapping and similar topic labels, such as “*Vaccine Mandates*” versus “*COVID-19 Vaccine Mandates*” and “*Cybersecurity Levy on Transactions*” versus “*Cybersecurity Levy on Bank Transactions*.” Additionally, there are ambiguous labels, such as “*Lawsuits*” and “*Cybersecurity*,” which lack specificity. In the evaluation of claim-topic pairs, some detailed topics directly mirrored the factual claims. In addition, the LLMs may produce irrelevant outputs, such as detailed topics labeled “*not mentioned in the given post*.” These issues highlight areas where LLMs’ performance can be improved to enhance the accuracy and relevance of the taxonomy, as well as the necessity of post-processing the generated topics.

6 Conclusion

In this study, we introduced LLMTaxo, a novel framework that leverages LLMs to construct taxonomies of factual claims from social media. Through evaluations across three distinct datasets, our approach demonstrated effectiveness in organizing claims into hierarchical structures with broad, medium, and detailed topics. The results highlight the framework’s potential in reducing redundancy, improving information accessibility, and assisting users such as researchers and fact-checkers in navigating online factual claims.

Limitations

The framework’s reliance on LLMs such as Zephyr and GPT-4o mini introduces potential limitations. These models, while powerful, may generate inconsistent or irrelevant topics for certain factual claims, which can impact the overall quality and reliability of the taxonomy. Despite efforts to automate the process, manual intervention remains necessary for refining and annotating learning examples. This introduces subjectivity, which may affect the consistency of the taxonomy.

Scalability poses an additional challenge. While LLMTaxo is designed to be adaptable for different topic domains, its performance may be constrained when applied to larger datasets with a broader range of topics. The computational demands associated with LLM-based topic generation could also limit its feasibility for large-scale applications.

Ethics and Risks

The deployment of LLMTaxo could potentially raise ethical considerations and risks. One key concern is bias in topic generation. Since LLMs are trained on vast amounts of pre-existing data, they may inadvertently reflect biases present in those sources. This can lead to biased topic categorizations, affecting the neutrality and fairness of the taxonomy. Identifying and mitigating such biases is crucial to maintaining objectivity.

Privacy considerations must also be addressed, as social media posts used in the study may contain personal information. Steps have been taken to anonymize and aggregate the data, but ongoing vigilance is required to ensure compliance with ethical guidelines and protect individual privacy.

The framework could also be susceptible to misuse. Malicious actors may attempt to manipulate the taxonomy to frame narratives that align with specific agendas. To counteract this risk, transparency in methodology and responsible use of the framework should be prioritized.

Finally, the taxonomy has the potential to influence public perception of factual claims on social media. Care must be taken to ensure that it presents an accurate, balanced, and comprehensive view of claims, avoiding any unintentional misrepresentation of content.

Acknowledgments

This work is partially supported by the United States National Science Foundation Award 2346261. We acknowledge the Texas Advanced Computing Center (TACC) for providing HPC resources that have contributed to the results reported in this paper. We are grateful to the human evaluators—Jacob Devasier, Abhishek Divakar Goudar, Juan Guajardo Gutierrez, Akshay Kumar Rayapet Madhusudhan, Alankrit Moses, Sai Sandeep Naraparaju, Jeremiah Pitts, Theodora Toutountzi, Mai Tran, Qing Wang, Devin Wingfield—for their thorough and dedicated work in taxonomy quality evaluation.

References

- Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.
- Esma Aïmeur, Sabrine Amri, and Gilles Brassard. 2023. Fake news, disinformation and misinformation in social media: a review. *Social Network Analysis and Mining*, 13(1):30.
- Hunt Allcott, Matthew Gentzkow, and Chuan Yu. 2019. Trends in the diffusion of misinformation on social media. *Research & Politics*, 6(2):2053168019848554.
- David M Blei, Andrew Y Ng, and Michael I Jordan. 2003. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022.
- Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.
- Victor R Basili, Gianluigi Caldiera, and H Dieter Rombach. 1994. The goal question metric approach. *Encyclopedia of software engineering*, pages 528–532.
- Ricardo JGB Campello, Davoud Moulavi, and Jörg Sander. 2013. Density-based clustering based on hierarchical density estimates. In *Pacific-Asia conference on knowledge discovery and data mining*, pages 160–172.
- Belen Carrion, Teresa Onorati, Paloma Díaz, and Vasiliki Triga. 2019. A taxonomy generation tool for semantic visual analysis of large corpus of documents. *Multimedia Tools and Applications*, 78:32919–32937.
- Boqi Chen, Fandi Yi, and Dániel Varró. 2023. Prompting or fine-tuning? a comparative study of large language models for taxonomy construction. In *2023 ACM/IEEE International Conference on Model Driven Engineering Languages and Systems Companion (MODELS-C)*, pages 588–596.
- Catherine Chen, Kevin Lin, and Dan Klein. 2020. Constructing taxonomies from pretrained language models. *arXiv preprint arXiv:2010.12813*.
- Wei-Lin Chiang, Lianmin Zheng, Ying Sheng, Anastasios Nikolas Angelopoulos, Tianle Li, Dacheng Li, Hao Zhang, Banghua Zhu, Michael Jordan, Joseph E. Gonzalez, and Ion Stoica. 2024. Chatbot arena: An open platform for evaluating llms by human preference.
- CrowdTangle. 2024. [CrowdTangle: A Public Insights Tool](#). Accessed: 2024-06-15.

- Google DeepMind. 2025. [Gemini 2.0 flash lite](#). Accessed: 2025-05-26.
- Michela Del Vicario, Alessandro Bessi, Fabiana Zollo, Fabio Petroni, Antonio Scala, Guido Caldarelli, H Eugene Stanley, and Walter Quattrociocchi. 2016. The spreading of misinformation online. *Proceedings of the national academy of Sciences*, 113(3):554–559.
- Susan T Dumais. 2004. Latent semantic analysis. *Annual Review of Information Science and Technology (ARIST)*, 38:189–230.
- James Durham, Sudipta Chowdhury, and Ammar Alzarrad. 2023. Unveiling key themes and establishing a hierarchical taxonomy of disaster-related tweets: A text mining approach for enhanced emergency management planning. *Information*, 14(7):385.
- Jinlan Fu, See-Kiong Ng, Zhengbao Jiang, and Pengfei Liu. 2024. GPTScore: Evaluate as you desire. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 6556–6576.
- Dirk Geeraerts. 2006. Prototype theory. *Cognitive linguistics: Basic readings*, 34:141–165.
- Maarten Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *arXiv preprint arXiv:2203.05794*.
- 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.
- Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by Claim-Buster. In *Proceedings of the 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining (KDD)*, pages 1803–1812.
- Kristina Hook and Ernesto Verdeja. 2022. Social media misinformation and the prevention of political instability and mass atrocities. Accessed: 2025-02-10.
- Jiaxin Huang, Yiqing Xie, Yu Meng, Yunyi Zhang, and Jiawei Han. 2020. Corel: Seed-guided topical taxonomy construction by concept learning and relation transferring. In *Proceedings of the 26th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 1928–1936.
- Hae Sun Jung, Haein Lee, Young Seok Woo, Seo Yeon Baek, and Jang Hyun Kim. 2024. Expansive data, extensive model: Investigating discussion topics around llm through unsupervised machine learning in academic papers and news. *Plos one*, 19(5):e0304680.
- Angelika Kaplan, Thomas Kühn, Sebastian Hahner, Niko Benkler, Jan Keim, Dominik Fuchß, Sophie Corallo, and Robert Heinrich. 2022. Introducing an evaluation method for taxonomies. In *Proceedings of the 26th International Conference on Evaluation and Assessment in Software Engineering*, pages 311–316.
- Srijan Kumar and Neil Shah. 2018. False information on web and social media: A survey. *arXiv preprint arXiv:1804.08559*.
- Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. 2023. G-eval: Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*.
- George A. Miller. 1995. Wordnet: A lexical database for english. Princeton University.
- Patricia Moravec, Randall Minas, and Alan R Dennis. 2018. Fake news on social media: People believe what they want to believe when it makes no sense at all. *Kelley School of Business research paper*, (18-87).
- Yida Mu, Chun Dong, Kalina Bontcheva, and Xingyi Song. 2024. Large language models offer an alternative to the traditional approach of topic modelling. *arXiv preprint arXiv:2403.16248*.
- Sadiq Muhammed T and Saji K Mathew. 2022. The disaster of misinformation: a review of research in social media. *International journal of data science and analytics*, 13(4):271–285.
- Yasir Abdalhamed Najem and Asaad Sabah Hadi. 2021. Semi-automatic ontology learning for twitter messages based on semantic feature extraction. In *New Trends in Information and Communications Technology Applications: 5th International Conference, NTICT 2021, Baghdad, Iraq, November 17–18, 2021, Proceedings 5*, pages 3–16.
- Roberto Navigli and Simone Paolo Ponzetto. 2010. Babelnet: Building a very large multilingual semantic network. In *Proceedings of the 48th annual meeting of the association for computational linguistics*, pages 216–225.
- Jingwei Ni, Minjing Shi, Dominik Stammach, Mrinmaya Sachan, Elliott Ash, and Markus Leippold. 2024. Afacta: Assisting the annotation of factual claim detection with reliable llm annotators. *arXiv preprint arXiv:2402.11073*.
- Katherine Ognyanova, David Lazer, Ronald E Robertson, and Christo Wilson. 2020. Misinformation in action: Fake news exposure is linked to lower trust in media, higher trust in government when your side is in power. *Harvard Kennedy School Misinformation Review*.
- OpenAI. 2024. [Gpt-4o-mini: Optimized mini version of gpt-4](#). Accessed: 2025-02-10.
- Gordon Pennycook, Tyrone D Cannon, and David G Rand. 2018. Prior exposure increases perceived accuracy of fake news. *Journal of experimental psychology: general*, 147(12):1865.

- N Reimers. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*.
- Peter J Rousseeuw. 1987. Silhouettes: a graphical aid to the interpretation and validation of cluster analysis. *Journal of computational and applied mathematics*, 20:53–65.
- Souvika Sarkar, Dongji Feng, and Shubhra Kanti Kar-maker_Santu. 2023. Zero-shot multi-label topic inference with sentence encoders and llms. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 16218–16233.
- Chirag Shah, Ryen W White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Snigdha Sarathi Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Xiaochuan Ni, et al. 2023. Using large language models to generate, validate, and apply user intent taxonomies. *arXiv preprint arXiv:2309.13063*.
- Victor Suarez-Lledo and Javier Alvarez-Galvez. 2021. Prevalence of health misinformation on social media: systematic review. *Journal of medical Internet research*, 23(1):e17187.
- Damian Tambini. 2017. Fake news: public policy responses.
- Lewis Tunstall, Edward Beeching, Nathan Lambert, Nazneen Rajani, Kashif Rasul, Younes Belkada, Shengyi Huang, Leandro von Werra, Cl  mentine Fourier, Nathan Habib, Nathan Sarrazin, Omar Sanseviero, Alexander M. Rush, and Thomas Wolf. 2023. Zephyr: Direct distillation of lm alignment.
- Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W White, Longqi Yang, et al. 2024. Tnt-llm: Text mining at scale with large language models. In *Proceedings of the 30th ACM SIGKDD Conference on Knowledge Discovery and Data Mining*, pages 5836–5847.
- Liang Wu, Fred Morstatter, Kathleen M Carley, and Huan Liu. 2019. Misinformation in social media: definition, manipulation, and detection. *ACM SIGKDD explorations newsletter*, 21(2):80–90.
- Xizhu Xiao, Porismita Borah, and Yan Su. 2021. The dangers of blind trust: Examining the interplay among social media news use, misinformation identification, and news trust on conspiracy beliefs. *Public Understanding of Science*, 30(8):977–992.
- Hui Yang. 2012. Constructing task-specific taxonomies for document collection browsing. In *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1278–1289.
- Chao Zhang, Fangbo Tao, Xiushi Chen, Jiaming Shen, Meng Jiang, Brian Sadler, Michelle Vanni, and Jiawei Han. 2018. Taxogen: Unsupervised topic taxonomy construction by adaptive term embedding and clustering. In *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*, pages 2701–2709.
- Haiqi Zhang, Zhengyuan Zhu, Zeyu Zhang, Jacob Devasier, and Chengkai Li. 2024a. Granular analysis of social media users’ truthfulness stances toward climate change factual claims. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 233–240.
- Zeyu Zhang, Zhengyuan Zhu, Haiqi Zhang, Foram Patel, Josue Caraballo, Patrick Hennecke, and Chengkai Li. 2024b. Wildfire: A twitter social sensing platform for layperson. In *Proceedings of the 17th ACM International Conference on Web Search and Data Mining (WSDM), demonstration description*, pages 1106–1109.
- Xinyan Zhao and Stephanie Tsang. 2022. How people process different types of misinformation on social media: A taxonomy based on falsity level and evidence type. Available at SSRN 4259593.
- Cangqi Zhou, Qianchuan Zhao, and Wenbo Lu. 2015. Impact of repeated exposures on information spreading in social networks. *PloS one*, 10(10):e0140556.
- Zhengyuan Zhu, Zeyu Zhang, Foram Patel, and Chengkai Li. 2022. Detecting stance of tweets toward truthfulness of factual claims. In *Proceedings of the 2022 Computation+Journalism Symposium*.
- Zhengyuan Zhu, Zeyu Zhang, Haiqi Zhang, and Chengkai Li. 2025. RATSD: Retrieval augmented truthfulness stance detection from social media posts toward factual claims. In *Findings of the Association for Computational Linguistics: NAACL 2025*, page 3366–3381.

A Prompt for Topic Generation and Sample Results

There are k learning examples used to guide the LLMs in generating a broad topic, a medium topic, and a detailed topic for each factual claim, as shown in Figure 2. Each prompt example contains the list of topic tuples from the annotated factual claims in the k learning examples (i.e., the seed taxonomy), a factual claim, a question asking the LLMs to generate broad, medium, and detailed topics for the claim, and the answer to the question. In the question, the LLMs are instructed to prioritize generating topics from the existing topics in the seed taxonomy. If none of the existing topics align well with the claim, the LLMs are then directed to generate new topics. This instruction ensures that the LLMs produce a limited number of topics. This prompt is iterated through all the factual claims to

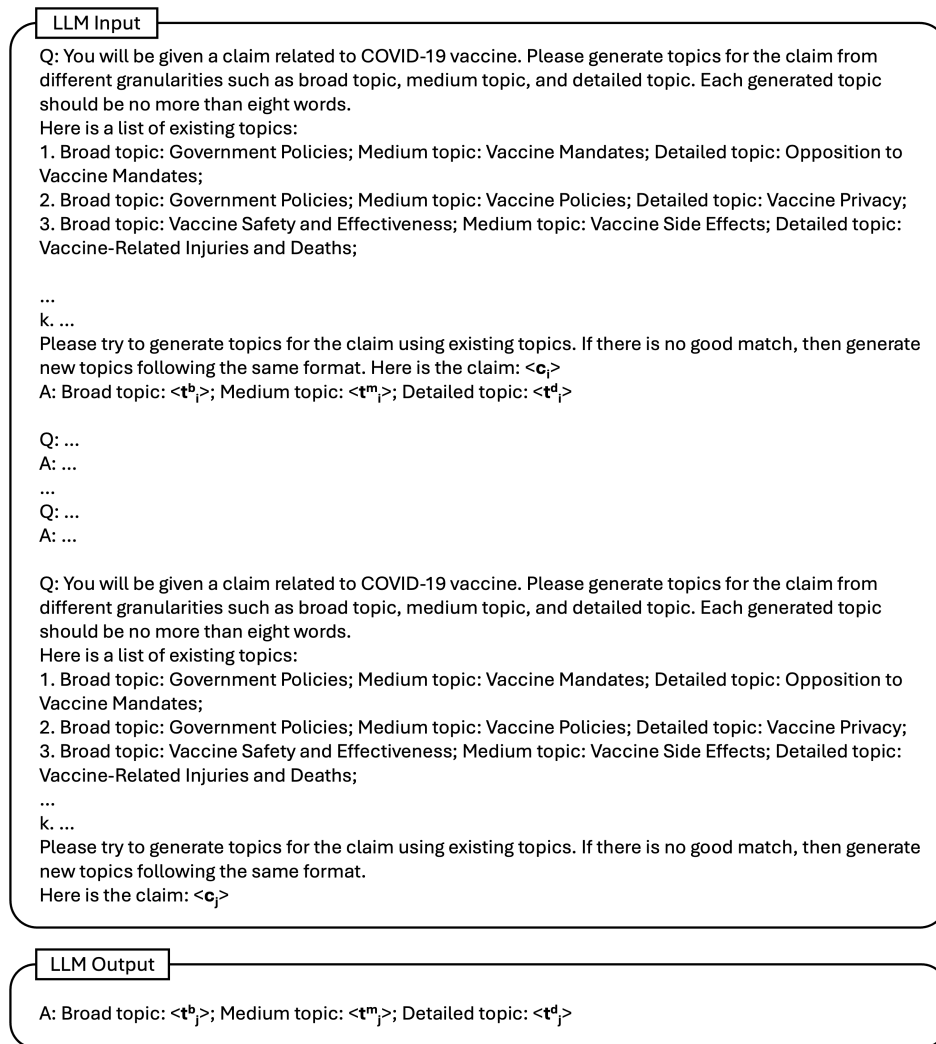


Figure 2: Prompt used to generate topics for each claim.

generate topics for them. A sample of the generated results from the three datasets (CV, CC, and CS) are shown in Table 5.

B GPT-4 Prompt for Evaluation

B.1 Prompt for Evaluating Taxonomy

I used LLMs to construct taxonomy and now I need to evaluate the taxonomy. I created some metrics to evaluate it. The [Taxonomy file name] uploaded contains the taxonomy with three-level topics. Please use the metrics in the [Metrics file name] to evaluate the taxonomy in [Taxonomy file name]. Please read each metric and understand them clearly, and then rate the metrics from 1-5, where 5 is the highest quality and 1 is the lowest. Please also provide judgments for your score and ignore the topic “Other” during evaluation.

B.2 Prompt for Evaluating Claim-Topic Pairs

I used LLMs to generate topics from three levels for factual claims. Now I need to evaluate ONLY

the detailed topics from two aspects: **accuracy** and **granularity**. Here are the two aspects: **Accuracy**: This criterion assesses how accurately the leaf node topics reflect the content and context of the corresponding factual claims. This involves determining if the topics are relevant and if they correctly represent the underlying information without misinterpretation or error. **Granularity**: This criterion evaluates the specificity of the leaf node topics. This involves determining whether the topics are detailed enough to uniquely categorize and differentiate between factual claims, yet broad enough to maintain practical applicability across multiple claims. If there is no detailed topic for a claim then evaluate the medium topic. If there is no medium topic existing, then evaluate broad topic.

Please read the evaluation metrics carefully and evaluate the claim-topic pairs and give one score for accuracy and one score for granularity for each claim-topic pair. The score ranges from 1-5, with 5

Dataset	Claim	Broad Topic	Medium Topic	Detailed Topic
COVID-19 Vaccine	John Stockton boldly suggests ‘thousands’ of pro athletes died after Covid vaccine shot https://t.co/nXbt6Apm2q via @marca	Vaccine Safety and Effectiveness	Vaccine Side Effects	Vaccine-Related Injuries and Deaths
COVID-19 Vaccine	A lot of people in ‘stage 4 cancer’ after #Covid #Vaccine https://t.co/z0YAqGgQrL	Vaccine Safety and Effectiveness	Vaccine Side Effects	Cancer Side Effect
Climate Change	Climate change is an existential threat to humanity. On Earth Day and every day, we remain committed to taking the most aggressive climate action ever.	Activism and Public Awareness	Climate Advocacy	Aggressive Climate Action
Climate Change	Climate change causes Dry spell in Kashmir, The weather in Kashmir is warmer than Delhi and Chandigarh, No snow rain in Kashmir During Chillai Kalan	Environmental Impact	Global Warming	Climate Change Effects in Kashmir
Cybersecurity	CBN Exempts 16 Items from Cybersecurity Levy...including Salary, Loans, Pension, Donations	Policies and Governance	Government Regulations	Cybersecurity Levy Exemptions
Cybersecurity	Streaming giant Roku has recently been targeted by a pair of cyberattacks, and the company confirmed over a half million Roku accounts were compromised.	Threats	Cyberattacks	Roku Account Compromise

Table 5: Factual claims and their topics generated by GPT-4o mini in different datasets.

being the best and 1 being the worst.

«EXAMPLES»

«EXAMPLE 1»

Factual claim: I worked for 18 months to end Biden’s unscientific and unethical military COVID vaccine mandate. Thanks to your phone calls and letters, we gained 92 sponsors on HR 3860. Repeal of the mandate just became a reality with the signing of the NDAA. Now let’s end the other mandates.

broad topic: Government Policies; medium topic: Vaccine Mandates; detailed topic: Opposition to Vaccine Mandates.

Accuracy: 5. Granularity: 5.

«EXAMPLE 2»

Factual claim: Myocarditis is up TEN times due to the Covid Vaccine... Nearly 30 % of young people have measurable cardiac injuries post-vaccine.. The CDC is LYING about this. . .

broad topic: Vaccine Safety and Effectiveness; medium topic: Vaccine Side Effects; detailed topic: Myocarditis Side Effect

Accuracy: 5. Granularity: 5.

«EXAMPLE 3»

Factual claim: Graphen oxide resonates at 26ghz microwaves from a 5G cell towers that’s in the COVID vaccine! You can neutralise the EMF and 5G radiation from mobile devices and detox from heavy metals.

broad topic: Political and Societal Implications; medium topic: Conspiracy Theories

Accuracy: 5. Granularity: 5.

«EXAMPLE 4»

Factual claim: Study published in Dec. 2020 proved COVID Vaccines could cause Strokes, Alzheimer’s, Parkinson’s, Multiple Sclerosis, and Autoimmune Disorder – Is there any wonder why the Five Eyes; Europe have suffered 2 Million Excess Deaths in the past 2 years?

broad topic: Vaccine Safety and Effectiveness; medium topic: Scientific and Medical Discussions; detailed topic: Discussions about Strokes, Alzheimer’s, Parkinson’s, Multiple Sclerosis, and Autoimmune Disorder.

Accuracy: 4. Granularity: 2.

«EXAMPLE 5»

Factual claim: ‘The doctor said that the probable cause of her heart attack was the vaccine, but he was too scared to put that on the report.’ South African politician Jay Naidoo reacts to the South African court being asked to conduct a judicial review of the Covid vaccine.

broad topic: Political and Societal Implications; medium topic: Vaccine Injury; detailed topic: Court Review of Covid Vaccine.

Accuracy: 2. Granularity: 5.

«END EXAMPLES»

Now, please evaluate the topics for the following

claim-topic pairs and only provide the scores for accuracy and granularity separated by a comma. For example. 3, 4.

Claim: {claim}

Broad Topic: {broad_topic}

Medium Topic: {medium_topic}

Detailed Topic: {detailed_topic}

C Evaluation Metrics

This section provides a detailed explanation of the taxonomy evaluation metrics.

Clarity. Assess whether the topic labels are clear, precise, and unambiguous.

Purpose: Ensure that each topic label communicates its content effectively to avoid confusion.

Evaluation Criteria:

- **Precision:** Each topic label uses specific and well-defined terms.
- **Unambiguity:** Topic labels should have only one interpretation, preventing misunderstanding.
- **Consistency:** Use of terminology is consistent across all levels of the taxonomy.
- **Accessibility:** Language is straightforward, avoiding jargon where possible unless it is standard within the covered domain.

Hierarchical Coherence. Assess whether the taxonomy follows a clear and meaningful hierarchical structure.

Purpose: Ensure that the taxonomy's structure facilitates easy navigation and understanding by clearly organizing information from the most general to the most specific.

Evaluation Criteria:

- **Gradational Specificity:** There is a logical progression from broader to more specific categories.
- **Parent-Child Coherence:** Parent-child relationships are well-formed, ensuring that child nodes logically belong to their parent nodes.
- **Consistency:** The hierarchy maintains consistent levels of detail throughout the taxonomy, ensuring that no topics are too broad or too narrow relative to others at the same level.

Orthogonality. Assess whether the topics are well-differentiated without duplication.

Purpose: Maintain distinct boundaries between topics to ensure that each topic captures unique aspects of the domain.

Evaluation Criteria:

- **Distinctiveness:** Topics at each level progressively add meaningful distinctions rather than just rephrasing broader topics.
- **Non-overlap:** For each topic, there is minimal to no overlap in the scope or content with other topics. Note that the topics with different parent topics are always different. For example, the medium topic "Vaccine Safety" under broad topic "Public Opinion" is essentially "Public Opinion about Vaccine Safety" and distinctly different from "Vaccine Safety" under "Government Policies." To minimize redundancy, we use succinct descriptions that are sufficient to convey the distinct meaning of each topic.

Completeness. Assess whether the taxonomy captures a broad and representative set of topics across different aspects of the domain.

Purpose: Cover as many areas of the topic to ensure the taxonomy is comprehensive.

Evaluation Criteria:

- **Domain Coverage:** The taxonomy covers a variety of significant aspects of the domain it represents.
- **Depth:** The taxonomy provides sufficient depth in each branch to capture nuanced distinctions within topics.
- **Balance:** The topics are evenly distributed across the taxonomy. This involves assessing whether some branches are disproportionately detailed while others are underdeveloped, which could lead to an imbalance that might skew the taxonomy's effectiveness and navigability.

Note that the intrinsic evaluation criteria of the metrics cannot completely eliminate overlap due to the inherent characteristics of taxonomy.

D Use of AI Assistants

Some of our code was developed using GitHub Copilot, and the writing was polished using ChatGPT and Grammarly.

E Human Evaluators

The human evaluators involved in the human evaluation are lab members of the research team.