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

Anonymous ACL submission

Abstract

With the vast 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 taxonomy of factual claims from social media by generating topics from multi-level granularities. This approach aids stakeholders in more effectively navigating the social media landscapes. We implement this framework with different models across three distinct datasets and introduce specially designed taxonomy evaluation metrics for a comprehensive assessment. With the evaluations from both human evaluators and GPT-4, the results indicate that LLMTaxo effectively categorizes factual claims from social media, and reveals that certain models perform better on specific datasets.

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).

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). Therefore, there is an urgent need for scalable, automated tools to organize and analyze factual claims in a systematic way, enabling stakeholders such as researchers, fact-checkers, and policymakers to better navigate this information landscape. For example, Tambini (2017) categorized fake news into different types to describe a range of fake news phenomena. Kumar and Shah (2018) constructed taxonomies for false information on social media platforms based on different characteristics, such as with the intention to deceive or not and opinion-based or fact-based. Zhao and Tsang (2022) proposed a taxonomy of misinformation on social media based on falsity level and evidence type. In this paper, we present a novel framework, LLMTaxo, for the automated construction of a taxonomy of factual claims on social media by using large language models (LLMs) to generate multi-level topics for each claim.

A taxonomy for organizing information assets starts with broad categories and branches into increasingly specific subcategories. In this way, factual claims can be categorized into meaningful categories, allowing for the identification of distinct claims, reduction of redundancy, and 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. Such structured organization enhances fact-checking and research workflows, and offers a clearer understanding of the themes and patterns in factual claims.

Our framework, LLMTaxo, clusters semantically similar factual claims, identifies distinct ones, and generates topics at multi-granularities, thus organizing claims into a hierarchical taxonomy with broad, medium, and detailed topics. The LLMs have rich background knowledge that is not only applicable now but also in the future while the online content changes over time. By leveraging the advanced capabilities of LLMs and few-shot learning, the framework minimizes human involvement and automates the time-intensive task of categorizing claims while maintaining adaptability to the evolving nature of social media discourse.

We addressed multiple challenges presented in developing LLMTaxo. One key challenge lies in the semantic variability of claims, where similar ideas are conveyed differently. LLMTaxo addresses this by clustering and identifying distinct claims. Another challenge is the scalability and applicability of the framework across diverse datasets and domains. To tackle this, we incorporate few-short in-context learning for LLMs, enabling them to be easily adapted to large datasets and various domains. To demonstrate the generalizability of LLMTaxo, we evaluated it using carefully designed metrics across three distinct datasets from different topics and data sources, including posts from *X* (formerly Twitter) and Facebook, covering topics such as COVID-19 vaccines, climate change, and cybersecurity.

The evaluation results demonstrate the effectiveness of LLMTaxo in enhancing our understanding of the social media landscape. By identifying distinct factual claims, the framework significantly reduces redundancy. The hierarchical taxonomy produced by LLMTaxo enables a multi-level analysis of claims, allowing for exploration at varying levels of detail. Furthermore, the framework exhibits strong performance across diverse datasets, showcasing its adaptability and potential for broad application. The datasets and codebase are available at https://anonymous.4open.science/r/LLMTaxo-2595.

This paper presents several key contributions:

• We introduce the first taxonomy of factual claims on social media constructed using LLMs. This resource can be directly integrated into fact-checking workflow and various research fields.

• We are the first to utilize LLMs for generating topics at multiple granularities.

• We develop a set of evaluation metrics for taxonomy and claim-topic pairs to comprehensively assess the quality of taxonomy.

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

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 (Carrion et al., 2019; Yang, 2012). Generic taxonomy construction tasks typically involve short concept terms or entity names, often represented 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 construction. Additionally, Shah et al. (2023); 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.

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 from multi-granularities.

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 the chaotic nature of social media through the automated construction of a taxonomy of factual claims. This taxonomy serves as a hierarchical classification system, enhancing the accessibility and navigability of information for users by categorizing claims into broad, medium, and detailed topics. The LLMTaxo framework initially identifies factual claims from social media posts, 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. This structured taxonomy provides an organized way to manage and explore factual claims from social media platforms. The overview of our framework is shown in Figure 1. In this section, we first explain how we utilize LLMs to develop a taxonomy. We then describe the techniques employed for detecting factual claims and the methods used to identify and select distinct claims.

3.1 Taxonomy Construction

The primary goal of this study is to construct a hierarchical taxonomy. The hierarchical design is informed by the prototype theory (Geeraerts, 2006), which guides cognitive categorization. Formally, a hierarchical taxonomy $\mathcal{T} = (T_b, T_m, T_d, f_m, f_d)$, that organizes a collection of factual claims $\mathcal{C} = \{c_1, c_2, ..., c_n\}$ into topics at multiple levels of granularity, ranging from broader to more fine-grained. Here, 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. Specifically, the taxonomy was constructed using a three-tiered structure consisting of broad topic $t_b \in T_b$ (representing the general theme), medium topic $t_m \in T_m$ (reflecting intermediate distinctions), and detailed topic $t_d \in T_d$ (highlighting finer aspects) for each claim. Each claim $c \in C$ is assigned topics at three levels: $\phi(c) = (t_b, t_m, t_d)$, where ϕ is a function that maps each claim c to a tuple containing its broad, medium, and detailed topics. For example, consider a factual claim c with $\phi(c) = (t_b, t_m, t_d)$ 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*, as shown on the right side of Figure 1. The hierarchical structure ensures that the taxonomy captures both high-level themes and more nuanced distinctions across factual claims. To automate the taxonomy construction process, we create learning examples that contain both factual claims and their corresponding topics. The topics from the learning examples form a seed taxonomy, which serves as a foundation for LLM-based expansion. We prompt the LLMs with both the learning examples and the seed taxonomy to generate topics for each distinct factual claim. After generating topics for all claims, we consolidate them to construct a refined taxonomy, as shown in Figure 1.

3.1.1 Learning Examples and Seed Taxonomy

While LLMs possess broad general knowledge, they are not inherently equipped to generate hierarchical topics for factual claims across different granularities specific to our needs. Our initial experiments indicated that LLMs often produce inconsistent topics for similar claims, leading to an unwieldy number of categories that hinder comprehension, as further detailed in Section 4.4. To mitigate this issue, we propose to incorporate 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 manually create k learning examples $\mathcal{L} = \{(c_1, t_{b_1}, t_{m_1}, t_{d_1}), ..., (c_k, t_{b_k}, t_{m_k}, t_{d_k})\}$, representing claims with their respective topics at each level of granularity. The topics from the k claims form the seed taxonomy, denoted as $\mathcal{S} = \{(t_{b_1}, t_{m_1}, t_{d_1}), ..., (t_{b_k}, t_{m_k}, t_{d_k})\}$. 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.

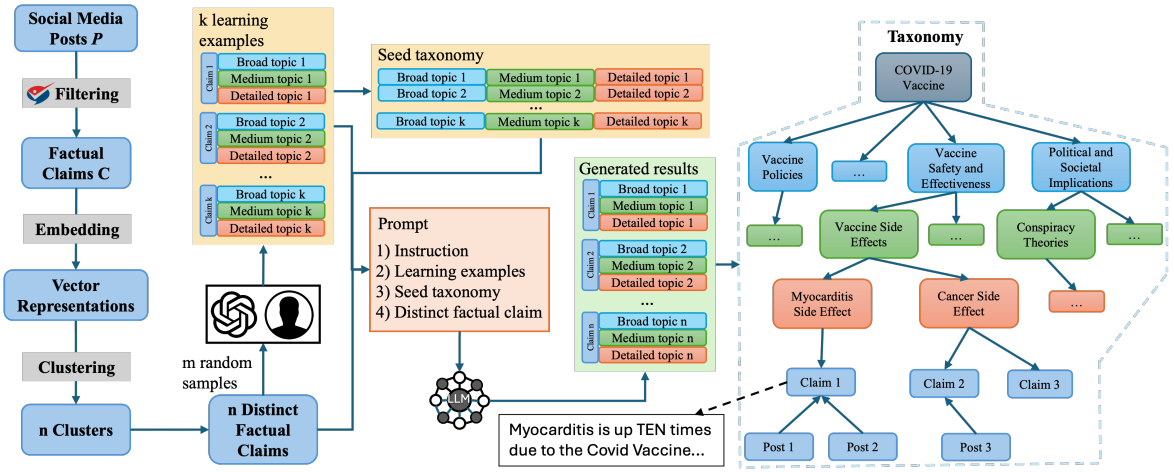

Figure 1: The LLMTaxo framework.

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

Each claim $c_i \in \mathcal{R}$ is processed by the LLMs, which is prompted to generate topics at three levels of granularity. After evaluating the LLM-generated topics, we refine and annotate each claim c_i with $(t_{b_i}, t_{m_i}, t_{d_i})$. The LLM-guided process reduces manual effort and allows for tailored annotations based on specific requirements, which is adaptable across various domains. Below is an example of the prompt used for the COVID-19 vaccine dataset:

```
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.
```

To achieve consistency and minimize topic variation, we restrict the number of topics and assign identical topics to claims within the same category. Note that no all factual claims have three-level topics, as some may be too brief or ambiguous to be categorized distinctly into broad, medium, and detailed topics. In such cases, the claim may align with only one or two topic levels. Upon annotating all m claims, we reviewed the most frequently occurring broad topics and established representative sets. The topics from the final set of k factual claims form the seed taxonomy $\mathcal{S}$. The k claim-topic pairs are used as learning examples $\mathcal{L}$. 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 (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 LLMs. Specifically, the prompt consists of $l_i \in \mathcal{L}$, the seed taxonomy $\mathcal{S}$, the instruction and questions that ask LLMs to produce the broad, medium, and detailed topics for c_i, and the answer to the question (i.e., corresponding $(t_{b_i}, t_{m_i}, t_{d_i})$ of c_i). After the LLMs learn from the k examples, it is provided with a new claim c_j and asked to generate new topics $(t_{b_j}, t_{m_j}, t_{d_j})$ for c_j. Due to the limited context length of the LLMs, one prompt generates $(t_{b_j}, t_{m_j}, t_{d_j})$ for only one c_j. This generation process is iterated until finishing generating $(t_{b_j}, t_{m_j}, t_{d_j})$ for all $c_j \in \mathcal{C}$. All the LLM-generated $(t_{b_j}, t_{m_j}, t_{d_j})$ form T_b, T_m, and T_d. The prompt is detailed in Figure 2 in Appendix A.

3.1.3 Topics Consolidation

After the LLMs generate the topics for each claim, we consolidate the results to build the taxonomy. 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, we employ two functions: $f_m : T_m \rightarrow T_b$, which maps each medium topic to a broad topic, and $f_d : T_d \rightarrow T_m$, which maps each detailed topic to a medium topic. Therefore, every medium topic is

associated with a broad topic: $\forall t_m \in T_m, \quad \exists t_b \in T_b$ such that $f_m(t_m) = t_b$; similarly, every detailed topic is associated with a medium topic: $\forall t_d \in T_d, \quad \exists t_m \in T_m$ such that $f_d(t_d) = t_m$.

For instance, consider two claims c_1 and c_2 with $\phi(c_1)$=(*Vaccine Safety and Effectiveness*, *Vaccine aSide Effects*, $t_d^{(1)}$) and $\phi(c_2)$=("Vaccine Safety and Effectiveness," "Vaccine Injury," $t_d^{(2)}$). By applying f_m, we get f_m("Vaccine Side Effects")="Vaccine Safety and Effectiveness" and f_m("Vaccine Injury")="Vaccine Safety and Effectiveness." This establishes that "Vaccine Side Effects" and "Vaccine Injury" are child nodes of "Vaccine Safety and Effectiveness." In this way, we establish a hierarchical structure of topics. This structure ensures that related topics are appropriately grouped, forming the final taxonomy $\mathcal{T}$.

3.2 Claim Detection

Social media data 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 employed a pioneering model, ClaimBuster (Hassan et al., 2017), for detecting check-worthy factual claims. This model 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 scored above this threshold were retained for further steps of LLMTaxo. This approach reduced 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 platforms 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 factual claim but are phrased slightly differently. To reduce redundancy and focus on unique claims, we applied clustering to group identical or nearly identical factual claims. We used 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 and should form individual clusters. To capture the semantic meaning of each claim, we employed Sentence-BERT (Reimers, 2019) to generate dense vector representations. After clustering, we identified distinct claims by selecting the first post from each cluster while excluding the outlier cluster. Each cluster represents a distinct factual claim, and the identified claims $\mathcal{C}$ form the foundation 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 (Pennycook 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 the COVID-19 vaccine using Wildfire (Zhang et al., 2024b). For this dataset, we targeted tweets containing various keyword variations related to the COVID-19 vaccine, 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 aforementioned 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 as of August 2024—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 two LLMs, Zephyr (Tunstall et al., 2023) and GPT-4o mini (OpenAI, 2024), 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 is chosen for its 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. To minimize the variability in the output, we set the temperature parameter of both models to 0.001. 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. To evaluate the clusters, we used Silhouette Coefficient (Rousseeuw, 1987). 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 of different datasets can be attributed to differences in dataset characteristics, particularly the greater sparsity in Facebook posts, and the Facebook posts are usually lengthier, making it more challenging for clustering. These results suggest that the method performs reasonably well across diverse datasets, even with different clustering configurations.

Datasets	Posts	Clusters	Outliers
CV	232,368	10,995	25,962
CC	89,412	15,794	42,923
CS	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 each dataset are presented in Table 3—the rows labeled "w/" under the method column. It is evident that GPT-4o mini demonstrates a more effective ability than Zephyr in limiting the number of broad and medium topics to a narrower range. 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 2. This approach limits the taxonomy to a more manageable size.

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

Table 2: 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 LLM with instruction, learning examples, and the target factual claim. The total topic counts with and without seed taxonomy are shown in Table 3. 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%.

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**% ↓)
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% ↓)
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% ↓)

Table 3: Topic counts for different datasets with and without prompting seed taxonomy.

5 Evaluation

We assessed LLMTaxo from two aspects: 1) the quality of the taxonomy, and 2) the appropriateness of the generated topics for the factual claims. We engaged both human evaluators and GPT-4 (Achiam et al., 2023), applying two sets of evaluation metrics that we have developed to evaluate each aspect. 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 human and GPT-4 evaluators. They were instructed to rate each criterion on a scale from 1 to 5, where 5 indicates the highest quality and 1 the lowest.

To evaluate the quality of taxonomies, we presented taxonomies generated by Zephyr and GPT-4o mini across the three datasets to the evaluators. For the evaluation of the claim-topic suitability, we randomly selected 50 factual claims along with their corresponding lowest-level topics (i.e., leaf nodes of the taxonomy) generated by both models from each dataset, resulting in a total of 100 claim-topic pairs per dataset. We exclusively evaluated the leaf node topics with claims, as the broader taxonomy quality had already been assessed in taxonomy quality evaluation, making further evaluation of higher-level topics unnecessary. Note that not all claims involve three-level topics, as mentioned in Section 3.1.1. Therefore, we focused our evaluation on the leaf node topics. To mitigate bias, we shuffled the claim-topic pairs from both models before presenting them to the evaluators. Both human evaluators and GPT-4 reviewed these pairs using our predefined metrics. Evaluators were also required to provide rationales for their scores. The evaluation prompt is detailed in Appendix C.

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 B.

Clarity. The goal 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 distinct boundaries between topics to ensure that each topic captures unique aspects of the domain. To assess whether the topics are well-differentiated without duplication, we evaluate non-overlap.

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

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 3.3 across all metrics). Notably, GPT-4o mini consistently outperformed Zephyr, suggesting its greater suitability for taxonomy construction. Human evaluators tended to give higher scores than GPT-4, particularly for Zephyr, indicating possibly stricter criteria or different interpretations of taxonomy

quality by GPT-4. It is also conceivable that GPT-4 shares more similarities with GPT-4o mini than with Zephyr. The models generally scored well on clarity and completeness, indicating their effectiveness in producing clear, precise and comprehensive topic labels. GPT-4o mini also scores highly in hierarchical coherence, 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, which may indicate some overlap or less distinct boundaries between topics.

Dataset	Evaluator	Model	M_1	M_2	M_3	M_4	Ac	Gr
CV	Human	Zephyr	4.2	4.0	3.4	4.3	4.3	4.1
		GPT-4o mini	4.3	4.2	3.8	4.7	3.8	3.8
	GPT-4	Zephyr	3.5	3.7	3.0	3.7	3.9	4.6
		GPT-4o mini	4.3	4.3	4.0	4.7	4.0	4.4
CC	Human	Zephyr	4.6	4.2	3.9	4.6	3.4	4.6
		GPT-4o mini	4.7	4.7	4.2	4.8	4.0	4.0
	GPT-4	Zephyr	3.5	3.3	4.0	4.0	3.3	3.4
		GPT-4o mini	4.3	4.0	4.5	4.3	3.7	4.1
CS	Human	Zephyr	4.3	3.9	3.7	4.4	3.9	4.0
		GPT-4o mini	4.5	4.0	4.0	4.6	4.4	4.2
	GPT-4	Zephyr	3.3	3.0	3.0	3.3	3.8	4.3
		GPT-4o mini	4.3	4.7	3.5	4.7	4.1	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.

5.2 Claim-Topic Evaluation

To assess how well the LLM-generated topics align with their corresponding factual claims, we performed evaluation on two aspects: *accuracy* and *granularity*. As we had already evaluated the taxonomy, which includes the relationship among broad, medium, and detailed topics in Section 5.1, we only evaluate claims and their leaf node topics. The detailed illustration of the evaluation metrics is:

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.

After human evaluators rated each claim-topic pair, we calculated the average scores for each dataset separately for Zephyr and GPT-4o mini, as shown in Table 4. The evaluations highlight the distinct strengths of Zephyr and GPT-4o mini. Zephyr demonstrated consistent granularity across datasets but showed variable accuracy, with its strongest performance in the *CV* dataset. GPT-4o mini exhibited strong granularity, particularly in the *CS* dataset where it scored a high of 4.9. Although it generally scored lower than Zephyr's in the *CV* dataset, it was competitive in the *CC* and *CS* datasets. The brevity and limited context of tweets in the *CV* dataset may have posed challenges for achieving high accuracy. In contrast, Facebook posts (*CC* and *CS*) likely provided more verbose and detailed content, which might have helped in achieving higher scores for GPT-4o mini.

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 score in taxonomy evaluation, evaluators highlighted issues caused by overlapping and similar topic labels, such as "*Vaccine Mandates*" versus "*COVID-19 Vaccine Mandates*" or "*Cybersecurity Levy on Transactions*" versus "*Cybersecurity Levy on Bank Transactions*" Additionally, there are ambiguous labels, such as "*Lawsuits*" and "*Cybersecurity*," which lack specificity. For the claim-topic pairs evaluation, some detailed topics directly replicate the factual claims. In addition, the LLMs may generate irrelevant data, producing detailed topics such as "not mentioned in the given post." These points highlight areas where LLM performance can be refined to improve taxonomy accuracy and relevance.

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 researchers, fact-checkers, and policymakers in navigating online factual claims.

Limitation

Semantic variability presents another challenge, as similar claims may be expressed in different ways. Although the framework reduces the large portion of redundancy through clustering, subtle differences in wording can lead to semantically similar claims being categorized separately, potentially affecting the coherence of the generated taxonomy.

The framework's reliance on LLMs such as Zephyr and GPT-4o mini also 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 across different 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.

References

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*.

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 Basili1 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. Springer.

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. IEEE.

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.

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. Association for Computational Linguistics.

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*.

Naeemul Hassan, Fatma Arslan, Chengkai Li, and Mark Tremayne. 2017. Toward automated fact-checking: Detecting check-worthy factual claims by claimbuster. In *Proceedings of the 23rd ACM SIGKDD international conference on knowledge discovery and data mining*, pages 1803–1812.

Kristina Hook and Ernesto Verdeja. 2022. Social media misinformation and the prevention of political instability and mass atrocities. *online], https://www. stimson. org/2022/social-media-misinformation-and-the-prevention-of-politicalinstability-and-mass-atrocities*.

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. https://wordnet.princeton.edu/. 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. Springer.

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 Stammbach, 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. https://openai.com/index/gpt-4o-mini-advancing-cost-efficient-intelligence/. 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 Karmaker_Santu. 2023. Zero-shot multi-label topic inference with sentence encoders and llms. Association for Computational Linguistics.

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, Clémentine Fourrier, 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, Xiusi 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*, 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"*.

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 a factual claim, a list of topic sets from the k annotated factual claims (i.e., seed taxonomy), 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. 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 generate topics for them. A sample of the generated results from the three datasets are shown in Table 5.

B Evaluation Metrics

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

LLM Input

Q: You will be given a claim related to COVID-19 vaccine. Please generate topics for the claim from different granularities such as broad topic, medium topic, and detailed topic. Each generated topic should be no more than eight words.
Here is a list of existing topics:
1. Broad topic: Government Policies; Medium topic: Vaccine Mandates; Detailed topic: Opposition to Vaccine Mandates;
2. Broad topic: Government Policies; Medium topic: Vaccine Policies; Detailed topic: Vaccine Privacy;
3. Broad topic: Vaccine Safety and Effectiveness; Medium topic: Vaccine Side Effects; Detailed topic: Vaccine-Related Injuries and Deaths;

...
k. ...
Please try to generate topics for the claim using existing topics. If there is no good match, then generate new topics following the same format. Here is the claim: <$\mathbf{c}_i$>
A: Broad topic: <$\mathbf{t}^{\mathbf{b}}_i$>; Medium topic: <$\mathbf{t}^{\mathbf{m}}_i$>; Detailed topic: <$\mathbf{t}^{\mathbf{d}}_i$>

Q: ...
A: ...
...
Q: ...
A: ...

Q: You will be given a claim related to COVID-19 vaccine. Please generate topics for the claim from different granularities such as broad topic, medium topic, and detailed topic. Each generated topic should be no more than eight words.
Here is a list of existing topics:
1. Broad topic: Government Policies; Medium topic: Vaccine Mandates; Detailed topic: Opposition to Vaccine Mandates;
2. Broad topic: Government Policies; Medium topic: Vaccine Policies; Detailed topic: Vaccine Privacy;
3. Broad topic: Vaccine Safety and Effectiveness; Medium topic: Vaccine Side Effects; Detailed topic: Vaccine-Related Injuries and Deaths;
...
k. ...
Please try to generate topics for the claim using existing topics. If there is no good match, then generate new topics following the same format.
Here is the claim: <$\mathbf{c}_j$>

LLM Output

A: Broad topic: <$\mathbf{t}^{\mathbf{b}}_j$>; Medium topic: <$\mathbf{t}^{\mathbf{m}}_j$>; Detailed topic: <$\mathbf{t}^{\mathbf{d}}_j$>

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

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

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.

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.

C GPT-4 Prompt for Evaluation

C.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.

C.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 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}

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.