LLM Benchmark–User Need Misalignment for Climate Change

Anonymous ACL submission

Abstract

Climate change is a major socio-scientific issue with significant implications for public decision-making and policy discussions. As large language models (LLMs) increasingly serve as an interface for accessing climate knowledge, whether existing benchmarks reflect user needs is critical for evaluating LLM performance in real-world settings. We propose a Proactive Knowledge Behaviors Framework that captures the different human–human and human–AI knowledge seeking and knowledge provision behaviors. We further develop a fine-grained Topic–Intent–Form taxonomy and apply it to analyze climate-related data representing different knowledge behaviors. Our results reveal a substantial mismatch between current benchmarks and real-world user needs, while knowledge interaction patterns between humans and LLMs closely resemble those in human–human interactions. These findings provide actionable guidance for benchmark design, RAG system development, and LLM training.

1 Introduction

Climate change is a critical socio-scientific challenge whose impacts extend beyond climate science to domains such as food systems, public health, and economic development (IPCC, 2021, 2022a,b). As climate risks increasingly affect societies, public demand for climate-related knowledge continues to grow (Calzada et al., 2024). At the same time, large language models (LLMs) are rapidly becoming common interfaces through which people seek information and produce written content (Chatterji et al., 2025). This raises concerns about the reliability of AI-generated climate information, particularly when non-expert users may over-trust and further disseminate inaccurate or misleading outputs (Pan et al., 2023; Hao et al., 2024; Lu et al., 2025). These concerns highlight the importance of evaluation benchmarks that reflect real-world user needs.

However, it remains unclear whether existing benchmarks used to evaluate LLM knowledge of climate change truly reflect the questions that users ask when consulting LLMs about climate change. In particular, it is uncertain whether these benchmarks accurately capture the diversity of topics, user intents, and expected answer forms that arise in real-world interactions.

To address this question, we first perform a systematic comparison between datasets representing real-world needs and existing benchmarks. Specifically, we identify climate-change–related queries from real user–LLM interaction datasets, including *WildChat* (Zhao et al., 2024), *LMSYS-Chat* (Zheng et al., 2024), and *ClimateQ&A* (Calzada et al., 2024). We also design an LLM-based topic modeling approach and construct taxonomies of topics, intents, and answer forms to annotate the data. Our results reveal a significant misalignment. Existing benchmarks focus on a narrow subset of climate knowledge and limited question types, whereas real user needs cover a broader range of topics (e.g., policy, transition and action), require higher-level procedural and metacognitive support (e.g., advice and actionable writing), and often demand more structured output formats (e.g., explanatory paragraphs or itemized lists).

A natural solution is to construct benchmarks directly from user queries. However, such queries are typically open-ended, diverse, and context-dependent, which makes it difficult to obtain reference answers at scale. This raises an important question: are there other readily accessible sources of high-quality knowledge that closely match these real-world knowledge needs? Here, we propose a *Proactive Knowledge Behaviors Framework* (see Figure 1 and Section 3) to conceptualize human *knowledge seeking* (i.e., asking) and *knowledge provision* (i.e., guiding, informing) behaviors. The framework serves to guide the empirical discovery and identification of candidate knowledge sources.

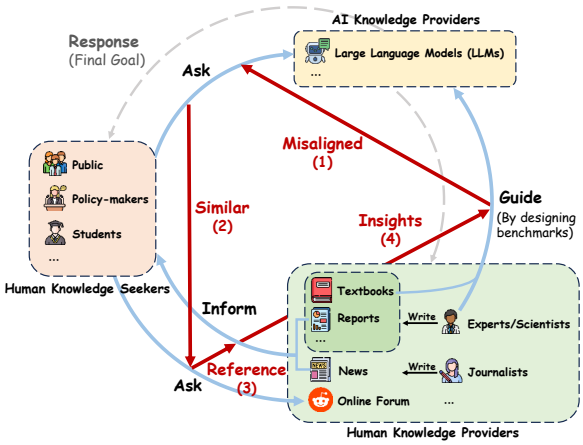

Figure 1: Our Proactive Knowledge Behaviors Framework. The proactive knowledge behaviors between three key actors are shown as blue arrows and the red arrows reflect our analytical logic.

Our analysis guided by the framework reveals that *human–human* and *human–AI* knowledge interactions are fundamentally aligned on knowledge needs. Specifically, by comparing the topic–intent–form distributions of questions asked when humans seek knowledge from other humans with those queries posed to LLMs, we observe a strong similarity. This alignment provides the premise for directly leveraging data from *human–human* knowledge interactions as a reference to support *human–AI* knowledge interactions. Furthermore, we find that within human-to-human knowledge provision, the topic distribution of the comprehensive and authoritative *IPCC AR6* reports (IPCC, 2021, 2022a,b) are highly consistent with human needs. This suggests that it can serve as a high-quality minimal knowledge base when constructing benchmarks (through sub-sampling), or when building retrieval-augmented generation (RAG) retrieval corpora.

In summary, in this work we make the following contributions:

- We identify a misalignment between existing climate change LLM benchmarks and real-world knowledge needs.
- We provide reference distributions of topic–intent–form and corpora for climate-change LLM evaluation and development.
- We introduce reusable taxonomies for climate-change topics, user intents, and expected answer forms.

2 Related Work

In the climate change domain, most previous NLP resources are designed for well-defined traditional tasks such as stance detection (Vaid et al., 2022; Morio and Manning, 2023), fact verification (Diggelmann et al., 2020), claims identification (Stammbach et al., 2023) and information retrieval (Schimanski et al., 2024), which are significantly different from open-ended human-LLM interactions. Recent work has begun to explore LLM-based climate question answering and communication, including literacy evaluation (Bulian et al., 2024), QA benchmarks (Manivannan et al., 2025), grounded dialogue systems (Vaghefi et al., 2023; Mullappilly et al., 2023; Hsu et al., 2024), multi-source data integration (Kuznetsov et al., 2025), and personalized recommendations (Nguyen et al., 2024). Nevertheless, these efforts remain largely science-oriented, leaving open questions about how to better support real-world, user-facing climate-related LLM interactions.

3 Framework and Data

Directly searching for solutions from real user queries is costly. When it is not feasible to start from the target scenario, a practical strategy is to draw inspiration from a related and mature pattern. Our *Proactive Knowledge Behaviors Framework* (Figure 1) is built on this idea. Specifically, we leverage mature human–human knowledge interaction patterns to analogy-inspire emerging human–LLM knowledge interactions. Within this framework, we identify three types of actors: human knowledge seekers (e.g., the public, policy makers, and students), human knowledge providers (e.g., experts, journalists, scientists, and contributors on online Q&A platforms such as Reddit), and AI knowledge providers (i.e., large language models in this work). We focus on proactive knowledge behaviors in both human–human and human–AI interactions. These behaviors fall into two categories: *knowledge seeking* and *knowledge provision*.

For knowledge seeking, we examine two types of behaviors. The first is human knowledge seekers **asking** AI knowledge providers. We collect the corresponding data as **Human-to-AI Queries**. The data come from three sources: *WildChat* (Zhao et al., 2024) and *LMSYS-Chat-1M* (Zheng et al., 2024), two public LLM conversation logs, and *ClimateQ&A* (Calzada et al., 2024), a dataset of climate-related questions posed by the French pub-

Category	Dataset	Task/Format	Brief Description	Count
Human-to-AI Queries	WildChat	LLM–User Conversations (Queries)	Large-scale multilingual LLM–user logs; diverse IPs/languages/regions/domains.	1,706
	LMSYS-Chat-1M	LLM–User Conversations (Queries)	Large-scale multilingual LLM–user logs; diverse IPs/languages/regions/domains.	1,331
	ClimateQ&A	LLM–User Conversations (Queries)	Authentic climate-related questions from French public to a climate-specialized LLM.	3,033
Human-to-Human Questions	Reddit (4 subreddits)	Questions (Forum Posts)	Questions from *r/climatechange*, *r/climate*, *r/climate_science*, *r/GlobalClimateChange*; collected from top/new/hot, deduplicated, interrogatives only.	1,023
Human-to-AI Guidance Knowledge	ClimaQA-Gold	QA Benchmark (Expert-annotated)	Graduate-level climate science QA; expert annotated for high reliability/authority.	540
	ClimaQA-Silver	QA Benchmark (Synthetic)	Large-scale synthetic graduate-level climate science QA for broad coverage in evaluation/fine-tuning.	2,543
Human-to-Human Knowledge Provision	SciDCC	News Articles	Climate-related news articles from *Science Daily*; media/journalistic narratives of climate issues.	7,476
	IPCC *AR6* (WG I/II/III)	Scientific Reports (Paragraph-level)	Authoritative synthesis for policymakers/public. WG I: Physical Basis; WG II: Impacts/Adaptation/Vulnerability; WG III: Mitigation. Paragraphs treated as instances.	18,889

Table 1: An overview of the eight core climate-related datasets used in this study. "Count" refers to the number of valid samples after cleaning and filtering. See three "Auxiliary Corpora" in Appendix B.1.

lic to an LLM-based system. For *WildChat* and *LMSYS-Chat-1M*, we use an LLM to filter out conversations unrelated to climate change. From each conversation, we retain only the first-turn user query to capture the user's original intention.

The second type is human knowledge seekers **asking** human knowledge providers. We define the corresponding data as **Human-to-Human Questions**, which are collected from question-style posts in four major climate-related subreddits on Reddit (Reddit Inc., 2025).[1]

For knowledge provision, we also consider two forms of behaviors. The first refers to human knowledge providers **guiding** AI knowledge providers. This refers to guiding LLM development through benchmark design. We collect the corresponding data as **Human-to-AI Guidance Knowledge** using the *ClimaQA* dataset (Manivannan et al., 2025), one of the most comprehensive benchmarks for evaluating LLM climate knowledge. We analyze its two subsets separately: *ClimaQA-Gold* and *ClimaQA-Silver*. Both subsets were generated by LLMs based on textbooks and include multiple-choice, cloze, and free-form questions. In our analysis, we use only the question components.

The second form is human knowledge providers proactively **providing** knowledge to human knowledge seekers. We collect the data as **Human-to-Human Knowledge Provision**. The data come from: *SciDCC* (Mishra and Mittal, 2021), a climate news dataset, and *IPCC AR6 (WG I/II/III)* (IPCC, 2021, 2022a,b), a highly authoritative and comprehensive assessment report on climate change.

In addition, we collect several auxiliary corpora to support topic modeling. These include *Climate-FEVER* (Diggelmann et al., 2020), *Environmental Claims* (Stammbach et al., 2023), and *ClimSight* (CliDyn Team, 2025). Details of all eleven datasets are summarized in Table 1, with examples provided in Appendix B.8.

We focus exclusively on proactive behaviors and exclude passive behaviors such as responses, as proactive behaviors determine whether and how knowledge flows emerge and therefore reveal the underlying motivations and information needs in knowledge interactions. Using this framework, we first identify the misalignment between evaluation and demand in human–AI knowledge interactions. We then examine whether knowledge needs in human–AI interactions align with those in human–human interactions. Based on this premise, we discover high-quality data from human–human knowledge interactions and translate them into insights for LLM benchmark construction and LLM development, with the ultimate goal of improving LLM responses on climate change topics.

4 Methods

This section describes our methodology for analyzing data. We first construct a topic taxonomy with the help of an LLM. We then further classify each data point along two dimensions—user intent and expected answer form—again leveraging an LLM. Finally, we represent queries as weighted vectors over topics, intents, and forms to enable quantitative analyses.

[1]The Reddit data used in this study was extracted on 2025-08-04 20:14 AEST (UTC+10:00) via the official Reddit API.

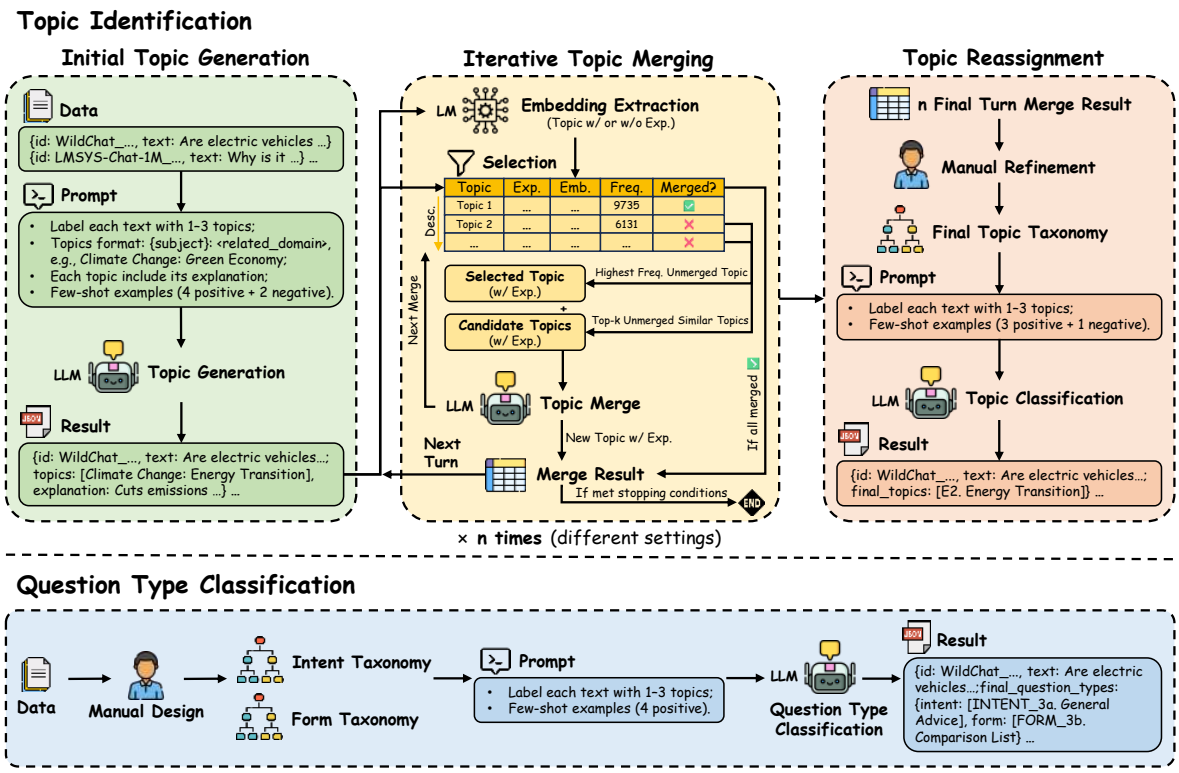

Figure 2: Pipelines for data annotation in Topic Identification and Question Type Classification.

4.1 Topic Identification

Initial Topic Generation. Traditional topic models such as LDA (Blei et al., 2003) and BERTopic (Grootendorst, 2022), which rely on lexical co-occurrence or static embeddings, often struggle to accurately capture semantic relations in complex or multi-faceted text. Inspired by TopicGPT and HICode (Pham et al., 2024; Zhong et al., 2025), we instead leverage the stronger semantic reasoning capabilities of LLMs to generate free-form topics. For each data sample, we apply a 6-shot in-context learning prompt (see Appendix B.2) to instruct the LLM to generate one to three topic labels together with brief explanations. Samples judged by the model as irrelevant to climate change are removed. Details are provided in Appendix B.3.

Iterative Topic Merging. Free-form topic generation often yields redundant or semantically overlapping labels, necessitating topic merging. We sort topics by descending frequency and compute LM-based embeddings. At each step, the most frequent unmerged topic is used as an anchor to retrieve its top-k most similar topics by cosine similarity. An LLM then merges only semantically equivalent or granularity-reconcilable topics into a new topic and description (Appendix B.2). One turn of merging processes all topics once, and we iteratively run multiple turns of merging until no merging occurs. Finally, the topic set is reduced to 1.58%–6.73% of its original size.

To improve coverage and robustness, we repeat the merging process under multiple settings (Appendix B.5): varying the use of explanations, embedding LMs, and merging LLMs. These variations mitigate missed merges due to embedding limitations and account for the inherently subjective nature of taxonomy construction, reducing bias introduced by any single configuration.

Topic Reassignment. Although a single round of merging already produces a high-quality topic list, we further conduct manual refinement across multiple merging results to construct a comprehensive and well-defined final taxonomy. For each merge setting, we take the union of the top-5 most frequent topics across the eleven datasets and supplement essential top-10 topics (Appendix B.6). This process standardizes topic granularity and merges residual semantic overlaps. The final taxonomy adopts a two-level hierarchy (Figure 12 in Appendix A), consisting of five coarse-grained categories (i.e., *Climate Science Foundations & Method*, *Ecological Impacts*, *Human Systems & Socioeconomic Impacts*, *Adaptation Strategies*, and

Mitigation Mechanisms, twenty-five fine-grained topics (e.g., *Biodiversity Loss*, *Agriculture & Food Security*, and *Energy Transition*), and an *Others* category. After establishing the taxonomy, we reassign topics to samples using an LLM with 4-shot in-context learning, allowing each sample to be associated with up to three relevant topics to capture multi-topic or coupled cases.

4.2 Question Type Classification

We analyze queries' question type through a taxonomy defined by two dimensions: **user intent** and **expected answer form**, capturing not only what users ask but also what they expect in response. Unlike topic modeling, both dimensions admit relatively clear ground-truth labels, enabling a fixed taxonomy design (Figure 13 in Appendix A). Although our analysis focuses on climate change, the taxonomy is intentionally designed to be broadly applicable. To ensure a comprehensive coverage and robustness to evolving LLM capabilities, our design builds on prior work (Ouyang et al., 2023; Wang et al., 2024; Wan et al., 2024; Shah et al., 2025) and incorporates empirical observations from our data as well as newly introduced commercial LLM functionalities (OpenAI, 2025a).

User intent and answer form are each organized into eight major categories, comprising twenty-nine types in total, along with nine "Others" types. For instance, intent includes *Fact Lookup* and *General Advice*, while form includes *Concise Paragraph* and *Item List*. Each intent type is further associated with a **knowledge-type label** derived from an extended Bloom's taxonomy (Anderson and Krathwohl, 2001), including *Factual*, *Conceptual*, *Procedural*, and *Metacognitive* knowledge (Appendix A). This mapping enables us to infer both the question category and the cognitive competencies required of the LLM. Finally, we use an LLM to label each instance with between 1 and 3 ranked intent and form labels. Details regarding the human verification for LLM annotation can be found in Appendix C.6.

4.3 Method for Analysis

For quantitative analysis, we retained the 42,261 samples classified under the 25 finalized topics. Each sample was represented as weighted vectors along three dimensions—Topic, Intent, and Form—where higher-ranked labels received greater weights normalized to a sum of one. Specifically, if a sample in a given dimension has K ordered labels (sorted by relevance), their weights follow the ratio: $w_1 : w_2 : \cdots : w_K = K : (K-1) : \cdots : 1$. All subsequent analyses were based on these weighted representations. To compare distributions across datasets and groups, we computed cosine similarities between normalized fixed-length weight vectors (25 dimensions for Topics and 38 for Intents and Forms), quantifying the alignment and divergence of knowledge behaviors across data sources.

5 Results

5.1 Benchmark-User Need Misalignment

To evaluate **whether current LLM benchmarks for climate change reflect real-world knowledge needs**, we compare Human-to-AI Queries and Human-to-AI Guidance Knowledge. Recall that the former includes *WildChat*, *LMSYS-Chat-1M*, and *ClimateQ&A* while the latter includes *ClimaQA-Gold* and *ClimaQA-Silver*. The three Human-to-AI query datasets are merged as the **Real-World Group**, while the two Human-to-AI Guidance Knowledge datasets are merged as the **LLM Benchmark Group**, with each dataset weighted proportionally to its size rather than applying uniform weights. Figure 3(a) shows pairwise cosine similarities of topic distributions among the five datasets, while Figure 3(c) shows cosine similarities between each dataset and the two groups. These results confirm the validity of grouping: the three Human-to-AI query datasets exhibit highly consistent topic distributions (with similarities 0.94–0.98), indicating strong real-world representativeness across distinct data sources, while the two Human-to-AI Guidance Knowledge datasets show nearly identical internal distributions.

We now examine the divergence of the two groups in terms of **Topic**. Figure 3(b) and 3(d) show that the topic distribution of the Real-World Group is more dispersed, with non-trivial proportions spanning many topics; in contrast, the LLM Benchmark Group is highly concentrated in *A1. Atmospheric Science & Climate Processes*, where its share exceeds that of the Real-World Group by approximately 63%. Meanwhile, the Real-World Group shows substantially greater attention to *E. Mitigation Mechanisms*, particularly *E1. Climate Policy, Governance & Finance Mechanism* and *E2. Energy Transition*, while the LLM Benchmark Group pays almost no attention to these categories.

For comparison regarding **User Intent**, which is shown in Figure 4, all datasets contain non-

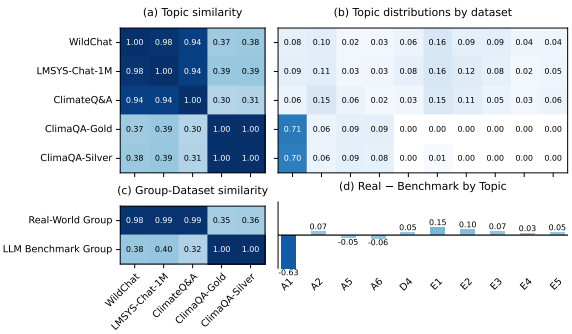

Figure 3: Topic comparison between Human-to-AI Queries and Human-to-AI Guidance Knowledge. (a) Pairwise topic-distribution similarities across the five datasets; (b) Probabilities of the 10 most diverging topics, i.e., those topics with the highest absolute probability differences under the two groups; (c) Topic-distribution similarities between each dataset and each group; (d) Probability differences of the most diverging topics. The interpretations and presentations of Figures 4 and 5 follow the same logic.

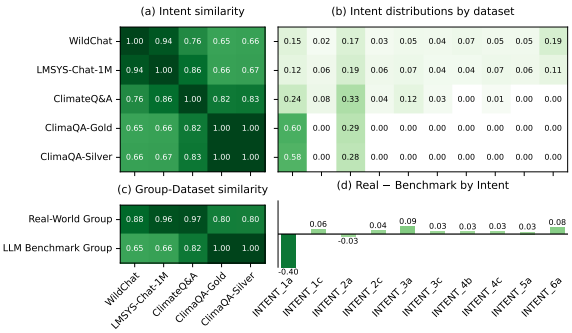

Figure 4: User intents comparison between Human-to-AI Queries and Human-to-AI Guidance Knowledge.

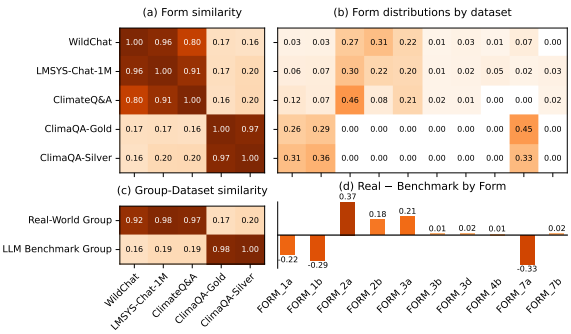

Figure 5: Expected answer forms comparison between Human-to-AI Queries and Human-to-AI Guidance Knowledge.

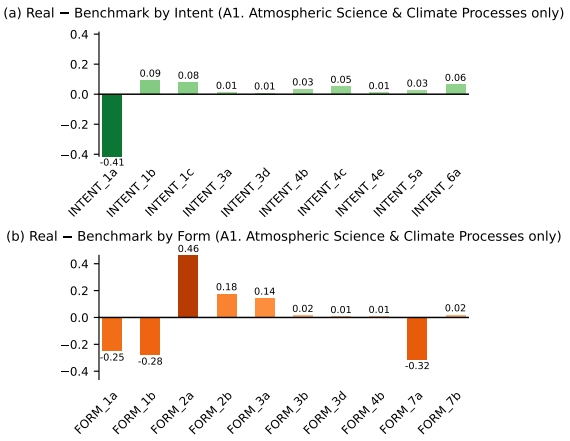

Figure 6: Question-type differences within topic *A1. Atmospheric Science&Climate Processes*: Real-World queries to LLM vs LLM QA benchmarks. (a) Top-10 intent differences; (b) Top-10 answer form differences.

negligible proportions (≥12%) of *INTENT_1a. Fact Lookup* and *INTENT_2a. Reasoning / Causal Analysis*. The Benchmark Group allocates as much as 60% to *INTENT_1a. Fact Lookup*, exceeding the Real-World Group by roughly 40% on average. In contrast, Figure 4(b) indicates that real-user intents are more diverse, and Figure 4(d) shows that humans more frequently seek *INTENT_3a. General Advice* and *INTENT_6a. Operational Writing*.

For **Expected Answer Form**, shown in Figure 5, within-group consistency remains high, while between-group differences are substantial. The Benchmark Group focuses on *FORM_1a. Concise Value(s) / Entity(ies)*, *FORM_1b. Brief Statement*, and *FORM_7a. Multiple Choice*. In contrast, the Real-World Group exhibits richer diversity but primarily concentrates on *FORM_2a. Concise Paragraph*, *FORM_2b. Detailed Multi-paragraph*, and *FORM_3a. Item List*, with between-group gaps reaching up to 37% (Figure 5(d)).

Because the Benchmark Group is highly concentrated in topic *A1*, we further examine whether the observed differences in intent and form are driven solely by topic imbalance. Restricting both groups to topic *A1* (Figure 6), we find that the differences in intent and form remain robust and align with the patterns observed in the full data (cf. Figure 4(d) and Figure 5(d)).

Taken together, these results show that current LLM climate evaluations exhibit a systematic misalignment with real user needs across the *Topic–Intent–Form* space. Benchmarks concentrate narrowly on climate-science fundamentals, prioritize *Factual* and limited *Conceptual knowledge* via fact-lookup and reasoning-analysis query types (see the intent–knowledge mapping in Figure 13 in Appendix A), and remain restricted to relatively simple answer forms. In contrast, users' actual knowl-

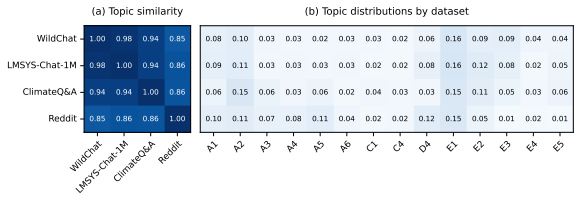

Figure 7: Topic comparison between real-world queries to LLMs and questions to humans. (a) Pairwise topic-distribution similarity across four datasets; (b) Topic distributions for the union of Top-10 topics (by share) across the four datasets. The interpretations and presentations of Figures 8 and 9 follow the same logic.

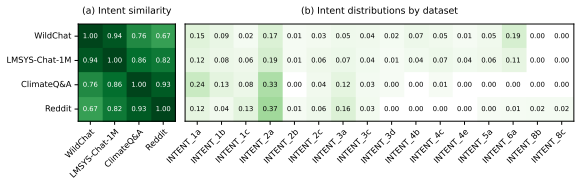

Figure 8: User intent comparison between real-world queries to LLMs and real-world questions to humans.

edge needs span a much broader set of topics (such as policy, transition, and industry action), require higher-level *Procedural* and *Metacognitive knowledge* to support advisory and operational-writing intents, and involve more complex answer forms such as paragraphs and itemized lists.

5.2 Similarity in Knowledge Needs

To verify whether **human–human knowledge interactions can serve as a reference for climate-related LLM evaluation**, we compare Human-to-AI Queries with Human-to-Human Questions. Recall that the latter consists of questions posted on Reddit. At the **Topic** level, as shown in Figure 7, the four datasets exhibit high distributional similarity (minimum 85%), suggesting that users do not systematically avoid or prefer particular topics when seeking answers from different knowledge providers. *E1. Climate Policy, Governance & Finance Mechanism* and *A2. Greenhouse Gas & Biogeochemical Cycles* are the most prominent topics (with minimum shares of roughly 15% and 10%, peaking at 16% and 15%). *A1. Atmospheric Science & Climate Processes*, *D4. Public Awareness, Communication & Community Engagement*, and *E2. Energy Transition* also generally maintain nontrivial proportions.

At the **User Intent** level (Figure 8), overall patterns are similar, yet differences become more noticeable: *INTENT_1a. Fact Lookup* is markedly

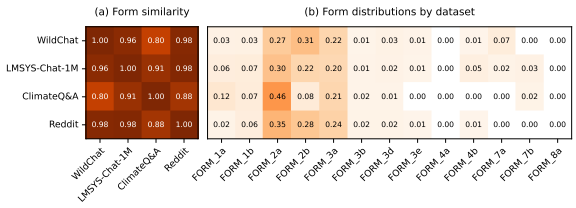

Figure 9: Expected answer form comparison between real-world queries to LLMs and real-world questions to humans.

higher in *ClimateQ&A* (around 24%), consistent with the intuition that a domain-specialized LLM is frequently used for factual lookup. Meanwhile, *INTENT_2a. Reasoning / Causal Analysis* is higher in both *ClimateQ&A* and *Reddit*–the two datasets representing domain-specific climate contexts–about 33% and 37%. In contrast, the intent distributions of *WildChat* and *LMSYS-Chat-1M* are more dispersed; relative to the other datasets, these two open-domain conversational datasets contain a higher proportion of *INTENT_6a. Operational Writing* (i.e., writing/creation tasks).

At the **Form** level (Figure 9), the four datasets are highly similar, with a general preference for *FORM_2a. Concise Paragraph*, *FORM_2b. Detailed Multi-paragraph*, and *FORM_3a. Item List*.

In summary, users largely focus on the same topics when addressing LLMs versus humans, and consistently prefer well-explained, clearly structured text or itemized forms. However, intent varies by context: domain-specific LLMs and forums emphasize analysis-oriented and fact-based questions (corresponding to *Conceptual* and *Factual knowledge*), while using general-purpose LLMs for climate-related writing remains an important use case (corresponding to *Procedural knowledge*). Recognizing these differences helps tailor evaluation strategies to specific application scenarios.

5.3 Knowledge Sources Align with Needs

To **identify high-quality knowledge sources in human–human knowledge interactions that align with real-world knowledge needs**, we analyze the **Topic** distributions between Human-to-Human Knowledge Provision—represented by *IPCC AR6* and *SciDCC*—and both Human-to-Human Questions and Human-to-AI Queries. Overall, as shown in Figure 10, although *IPCC AR6* is commonly perceived as highly specialized and policy-oriented, and has raised concerns about whether it can meet the public's knowledge needs,

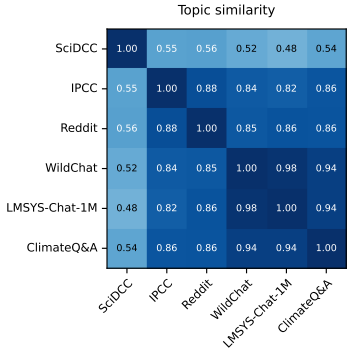

Figure 10: Topic comparison between inform-human knowledge and both real-world questions to human and queries to LLMs.

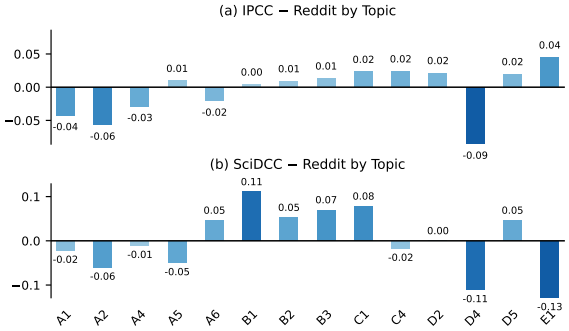

Figure 11: Topic differences: inform-human knowledge vs real-world questions to human. (a) IPCC minus Reddit topic distributions; (b) SciDCC minus Reddit topic distributions. Only shows the union of Top-10 absolute differences.

its topic distribution is in fact highly aligned with the human climate knowledge demands (with similarities of 88%, 84%, 82%, and 86% to Reddit, WildChat, LMSYS-Chat-1M, and ClimateQ&A, respectively). In contrast, the topic similarity of *SciDCC* appears to be more moderate.

To further leverage these corpora, we compute the distribution differences between *IPCC/SciDCC* and Human needs (Figure 11). Considering the strong consistency in knowledge demands between Human-to-AI queries and Human-to-Human questions, we simplify the analysis by using Reddit as a proxy for Human needs. We observe the first major contrast: *SciDCC* is more strongly oriented toward *B. Ecological Impacts*, particularly *B1. Biodiversity Loss*. We further identify a shared gap: compared with *Reddit*, both *IPCC* and *SciDCC* devote substantially less attention to *D4. Public Awareness, Communication & Community Engagement*. This suggests that although these issues are salient to users, they remain underrepresented in knowledge-provision sources.

5.4 Insights

Based on the findings in Results 5.1, 5.2, and 5.3, we derive the following implications to help design climate-change–related LLM evaluations and improve LLM knowledge provision

First, Benchmark Design. Benchmark questions should be designed according to the observed distributions of topic, user intent and expected answer form, so as to better reflect real-world user needs. In addition, *IPCC AR6* closely aligns with real-world user topic demands, making it an ideal source for (down-)sampling data when constructing benchmark datasets.

Second, Retrieval-Augmented Generation (RAG) Systems. Similarly, *IPCC AR6* serves as an excellent retrieval corpus. Its strong alignment with real-user topic distributions and its rich content can reduce redundant data collection and improve retrieval relevance, making it suitable as a minimal knowledge base. This finding also helps explain why some previous studies (Vaghefi et al., 2023) achieved better performance when using IPCC reports for RAG systems. However, for the topic *D4: Public Awareness, Communication & Community Engagement*, additional external sources are still required to supplement the retrieval corpus.

Finally, Training and Fine-tuning. The distributional differences we identified in topics and question types can directly inform the composition of training datasets for climate-specific LLMs. These insights provide guidance for data selection and balancing, thereby improving the alignment between models and real-world knowledge needs.

6 Conclusion

We investigate whether climate change benchmarks for LLMs reflect real-world knowledge needs. By comparing benchmarks with real user–LLM interactions, we find a systematic misalignment across the *Topic–Intent–Form* space: existing benchmarks emphasize science and fact-lookup, whereas real users seek broader, application-oriented knowledge and structured explanations. We further show that human–human and human–AI knowledge-seeking patterns are highly similar, and identify that *IPCC AR6* closely matches real-world topic distributions. This offers guidance for more realistic climate-related LLM evaluation, development and RAG system construction.

Limitations

This work focuses on climate change, therefore, its methodology and some of its conclusions may have limited direct generalizability to other domains. In addition, certain data categories are constrained by the availability of high-quality datasets. For example, human–human queries are sourced solely from *Reddit*, primarily because existing public climate-related datasets from pllatform like Twitter (now X) tend to exhibit substantial noise in both content and format. Similarly, the Human-to-AI Guidance Knowledge category relies exclusively on *ClimaQA*, as it is currently one of the few mainstream and relatively well-curated LLM-oriented QA datasets in the climate change domain.

Although we consider the Human-to-AI Queries to be of sufficient scale, and datasets such as *WildChat* and *LMSYS-Chat-1M* cover multilingual and geographically diverse users, with distributions across Topic-Intent-Form showing strong consistency across sources, the data may still exhibit potential biases. In particular, users who consent to sharing their interactions with LLMs may constitute a self-selected population, and English speakers remain dominant in the data. This may result in the underrepresentation or omission of certain populations and their needs.

Furthermore, the construction of the topic taxonomy in this work involves a degree of manual decision-making, which may introduce subjectivity. Although we show in Appendix C.3 that directly using the topic list obtained after Topic Merge yields results similar to those after Topic Reassignment in the overall analysis, the taxonomy design itself may still influence the final conclusions.

Finally, the fine-grained topic and question taxonomy increases the complexity of manual annotation and validation, and places higher demands on annotator expertise, which in turn raises the difficulty of thoroughly validating the results.

References

Lorin W. Anderson and David R. Krathwohl. 2001. *A taxonomy for learning, teaching, and assessing : a revision of Bloom's taxonomy of educational objectives*, complete ed. edition. Longman, New York.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent dirichlet allocation. *J. Mach. Learn. Res.*, 3(null):993–1022.

Jannis Bulian, Mike S. Schäfer, Afra Amini, Heidi Lam, Massimiliano Ciaramita, Ben Gaiarin, Michelle Chen Hübscher, Christian Buck, Niels G. Mede, Markus Leippold, and Nadine Strauß. 2024. Assessing large language models on climate information. In *Proceedings of the 41st International Conference on Machine Learning*, ICML'24. JMLR.org.

Natalia De La Calzada, Théo Alves Da Costa, Annabelle Blangero, and Nicolas Chesneau. 2024. Climateq&a: Bridging the gap between climate scientists and the general public. *CoRR*, abs/2403.14709.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. 2025. How people use chatgpt. Working Paper 34255, National Bureau of Economic Research.

CliDyn Team. 2025. Climsight-gpt-4.1-mini-climate-assessments. `https://huggingface.co/datasets/CliDyn/ClimSight-GPT4.1-Mini-Climate-Assesments`. Available on Hugging Face. Accessed: 2025-11-13.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*.

Maarten R. Grootendorst. 2022. Bertopic: Neural topic modeling with a class-based tf-idf procedure. *ArXiv*, abs/2203.05794.

Guozhi Hao, Jun Wu, Qianqian Pan, and Rosario Morello. 2024. Quantifying the uncertainty of llm hallucination spreading in complex adaptive social networks. *Scientific Reports*, 14.

Angel Hsu, Mason Laney, Ji Zhang, Diego Manya, and Linda Farczadi. 2024. Evaluating ChatNetZero, an LLM-chatbot to demystify climate pledges. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 82–92, Bangkok, Thailand. Association for Computational Linguistics.

IPCC. 2021. *Climate Change 2021: The Physical Science Basis. Contribution of Working Group I to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

IPCC. 2022a. *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

IPCC. 2022b. *Climate Change 2022: Mitigation of Climate Change. Contribution of Working Group III to the Sixth Assessment Report of the Intergovernmental Panel on Climate Change*. Cambridge University Press, Cambridge, United Kingdom and New York, NY, USA.

Ivan Kuznetsov, Antonia Jost, Dmitrii Pantiukhin, Boris Shapkin, Thomas Jung, and Nikolay Koldunov. 2025. Transforming climate services with llms and multi-source data integration. *npj Climate Action*, 4:97.

Zhuoran Lu, Gionnieve Lim, and Ming Yin. 2025. Understanding the effects of large language model (llm)-driven adversarial social influences in online information spread. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, CHI EA '25, New York, NY, USA. Association for Computing Machinery.

Veeramakali Vignesh Manivannan, Yasaman Jafari, Srikar Eranky, Spencer Ho, Rose Yu, Duncan Watson-Parris, Yian Ma, Leon Bergen, and Taylor Berg-Kirkpatrick. 2025. ClimaQA: An automated evaluation framework for climate question answering models. In *The Thirteenth International Conference on Learning Representations*.

Prakamya Mishra and Rohan Mittal. 2021. Neuralnere: Neural named entity relationship extraction for end-to-end climate change knowledge graph construction. In *ICML 2021 Workshop on Tackling Climate Change with Machine Learning*.

Gaku Morio and Christopher D Manning. 2023. An nlp benchmark dataset for assessing corporate climate policy engagement. In *Advances in Neural Information Processing Systems*, volume 36, pages 39678–39702. Curran Associates, Inc.

Sahal Shaji Mullappilly, Abdelrahman Shaker, Omkar Thawakar, Hisham Cholakkal, Rao Muhammad Anwer, Salman Khan, and Fahad Khan. 2023. Arabic mini-ClimateGPT : A climate change and sustainability tailored Arabic LLM. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 14126–14136, Singapore. Association for Computational Linguistics.

Vincent Nguyen, Sarvnaz Karimi, Willow Hallgren, Ashley Harkin, and Mahesh Prakash. 2024. My climate advisor: An application of NLP in climate adaptation for agriculture. In *Proceedings of the 1st Workshop on Natural Language Processing Meets Climate Change (ClimateNLP 2024)*, pages 27–45, Bangkok, Thailand. Association for Computational Linguistics.

OpenAI. 2025a. Introducing chatgpt agent: Bridging research and action. https://openai.com/index/introducing-chatgpt-agent/. Accessed: 2025-11-16.

OpenAI. 2025b. Introducing gpt-4.1 in the api. https://openai.com/index/gpt-4-1/. Accessed: 2026-2-25.

OpenAI. 2025c. Introducing gpt-5. https://openai.com/index/introducing-gpt-5/. Accessed: 2026-2-25.

Siru Ouyang, Shuohang Wang, Yang Liu, Ming Zhong, Yizhu Jiao, Dan Iter, Reid Pryzant, Chenguang Zhu, Heng Ji, and Jiawei Han. 2023. The shifted and the overlooked: A task-oriented investigation of user-GPT interactions. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 2375–2393, Singapore. Association for Computational Linguistics.

Yikang Pan, Liangming Pan, Wenhu Chen, Preslav Nakov, Min-Yen Kan, and William Wang. 2023. On the risk of misinformation pollution with large language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 1389–1403, Singapore. Association for Computational Linguistics.

Chau Minh Pham, Alexander Hoyle, Simeng Sun, Philip Resnik, and Mohit Iyyer. 2024. TopicGPT: A prompt-based topic modeling framework. 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 2956–2984, Mexico City, Mexico. Association for Computational Linguistics.

Reddit Inc. 2025. Reddit api documentation. https://www.reddit.com/dev/api/. Accessed: 2026-2-25.

Tobias Schimanski, Jingwei Ni, Roberto Spacey Martín, Nicola Ranger, and Markus Leippold. 2024. ClimRetrieve: A benchmarking dataset for information retrieval from corporate climate disclosures. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 17509–17524, Miami, Florida, USA. Association for Computational Linguistics.

Chirag Shah, Ryen White, Reid Andersen, Georg Buscher, Scott Counts, Sarkar Das, Ali Montazer, Sathish Manivannan, Jennifer Neville, Nagu Rangan, Tara Safavi, Siddharth Suri, Mengting Wan, Leijie Wang, and Longqi Yang. 2025. Using large language models to generate, validate, and apply user intent taxonomies. *ACM Trans. Web*, 19(3).

Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 1051–1066, Toronto, Canada. Association for Computational Linguistics.

Sentence Transformers. 2020. all-minilm-l6-v2. https://huggingface.co/sentence-transformers/all-MiniLM-L6-v2. Accessed: 2025-11-13.

Saeid Ashraf Vaghefi, Dominik Stammbach, Veruska Muccione, Julia Bingler, Jingwei Ni, Mathias Kraus, Simon Allen, Chiara Colesanti-Senni, Tobias Wekhof, Tobias Schimanski, Glen Gostlow, Tingyu Yu, Qian Wang, Nicolas Webersinke, Christian Huggel, and Markus Leippold. 2023. ChatClimate: Grounding conversational AI in climate science. *Communications Earth and Environment*, 4(1):480.

Roopal Vaid, Kartikey Pant, and Manish Shrivastava. 2022. Towards fine-grained classification of climate change related social media text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop*, pages 434–443, Dublin, Ireland. Association for Computational Linguistics.

Mengting Wan, Tara Safavi, Sujay Kumar Jauhar, Yujin Kim, Scott Counts, Jennifer Neville, Siddharth Suri, Chirag Shah, Ryen W. White, Longqi Yang, Reid Andersen, Georg Buscher, Dhruv Joshi, and Nagu Rangan. 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*, KDD '24, page 5836–5847, New York, NY, USA. Association for Computing Machinery.

Jiayin Wang, Fengran Mo, Weizhi Ma, Peijie Sun, Min Zhang, and Jian-Yun Nie. 2024. A user-centric multi-intent benchmark for evaluating large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3588–3612, Miami, Florida, USA. Association for Computational Linguistics.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. Qwen3 technical report. *CoRR*, abs/2505.09388.

Yanzhao Zhang, Mingxin Li, Dingkun Long, Xin Zhang, Huan Lin, Baosong Yang, Pengjun Xie, An Yang, Dayiheng Liu, Junyang Lin, Fei Huang, and Jingren Zhou. 2025. Qwen3 embedding: Advancing text embedding and reranking through foundation models. *CoRR*, abs/2506.05176.

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. Wildchat: 1m chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*.

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Tianle Li, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zhuohan Li, Zi Lin, Eric Xing, Joseph E. Gonzalez, Ion Stoica, and Hao Zhang. 2024. LMSYS-chat-1m: A large-scale real-world LLM conversation dataset. In *The Twelfth International Conference on Learning Representations*.

Mian Zhong, Pristina Wang, and Anjalie Field. 2025. HICode: Hierarchical inductive coding with LLMs. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 31060–31078, Suzhou, China. Association for Computational Linguistics.

A Taxonomies

Figures 12 and 13 present the final taxonomies developed and used for data annotation in our work. Specifically, Figure 12 illustrates the climate change *topic taxonomy*, while Figure 13 shows the *question type taxonomy*.

The topic taxonomy is organized as a two-level hierarchical structure. It contains five primary categories—*A. Climate Science Foundations & Method*, *B. Ecological Impacts*, *C. Human Systems & Socioeconomic Impacts*, *D. Adaptation Strategies*, and *E. Mitigation Mechanisms*—comprising a total of 25 fine-grained subtopics. In addition, we introduce an auxiliary category *F. Others* to capture a portion of climate-related data not covered by the main taxonomy, as well as irrelevant samples. Data labeled as *Others* are excluded from the final analysis.

The question type taxonomy consists of two complementary two-level classification schemes: one describing *user intent* and the other specifying the *expected answer form*. Each scheme includes eight core categories and 29 fine-grained subtypes. Additionally, both taxonomies contain a global *Others* category and an extra *others* subtype within each core category to account for edge cases.

Furthermore, we incorporate the knowledge dimension framework from Bloom's taxonomy (Anderson and Krathwohl, 2001), including *Factual knowledge* (basic facts and terminology), *Conceptual knowledge* (relationships among concepts, principles, and theories), *Procedural knowledge* (methods and processes for performing tasks), and *Metacognitive knowledge* (awareness and regulation of one's own cognition, including strategies for learning). Since the taxonomy was originally developed to describe human learning processes, we adapt and operationalize these categories to better reflect knowledge usage in LLMs. Specifically, we interpret *Factual knowledge* as the recall of basic facts and definitions, *Conceptual knowledge* as the ability to capture relationships among concepts (e.g., causal or relational reasoning), *Procedural knowledge* as knowledge of how to execute tasks or processes, and *Metacognitive knowledge* as the capacity for planning, strategy selection, and decision-making during problem solving. This extension is intended to provide an operational categorization of knowledge needed by users in LLM outputs rather than to model human cognition directly.

These labels indicate the primary and most representative knowledge requirement for addressing each intent type. For example, *1a. Fact Lookup* primarily rely on *Factual* knowledge (F), whereas *2a. Reasoning / Causal Analysis* typically depends on *Conceptual* knowledge (C). It is important to note that we annotate only the most distinctive and discriminative knowledge dimension for each category. Additional supporting knowledge requirements and the detailed mapping rationale are documented in Appendix B.9.

B Method and Data Details

B.1 More details of the dataset

Auxiliary Corpora The **Auxiliary Corpora** included three datasets: *Climate-FEVER* (Diggelmann et al., 2020), a fact verification dataset; *Environmental Claims* (Stammbach et al., 2023), which focuses on corporate climate claims; and *ClimSight* (CliDyn Team, 2025), a climate question–answering dataset. More details in Table 2.

Data Consent We use a combination of publicly available datasets and data collected via official APIs, and adhere to their respective licenses and terms of use. All datasets used in this work are obtained from publicly released sources. We rely on the original data collection procedures and consent processes as described by their authors (in accordance with their respective licenses, e.g., Apache-2.0 License, BSD 3-Clause License,), as specified in their official releases.), and use these datasets strictly for research purposes. Reddit data is collected using the official Reddit API in compliance with the platform's terms of service. The data consists only of publicly available posts. We do not collect private or deleted content. The IPCC AR6 reports (Working Groups I, II, and III) are publicly available documents intended for open dissemination and are used for research purposes only.

Sensitive or Offensive Content The study uses publicly available datasets (e.g., *WildChat*, *LMSYS-Chat*, *ClimateQ&A*). These datasets consist of user queries and may contain naturally occurring personally identifiable information or offensive content. We rely on the original dataset providers' collection and filtering procedures and do not introduce additional sensitive data. This work does not involve direct interaction with human subjects, and no attempt is made to identify individuals represented in the data.

B.2 LLM Prompts

The prompt below omits the example and taxonomy. In the following prompt, {subject}, {n}, and {m} correspond to *Climate Change*, 4, and 20, respectively.

1. Topic Identification: Initial Topic Generation

```
You are an expert research curator
    specialising in "{subject}". Your
    task is to identify the main topics
    touched on by the text below, and for
     each topic provide a brief (<= {m}
    words) explanation.
The texts you need to identify will come
    from various sources, including
    reports, real-world user queries,
    online Q&A, AI-generated questions,
    social media posts, and more. At the
    same time, these texts may appear in
    any language (including English,
    Chinese, Russian, French, etc.). You
    need to detect the language, but
    always produce the output in English.

GUIDELINES:

1. Think of a "topic" as a high-level
    theme-but neither too abstract (e.g.
    "science") nor too fine-grained (e.g.
     "NO2 emissions from Euro-6 diesel
    cars in Canberra on 12 Jan 2021").
    Some examples of granularity guide
    fot generated topics: Climate Change:
     Extreme Weather; Climate Change:
    Climate Modeling; Climate Change:
    Urban Planning; Climate Change:
    Economic Impacts.
2. Express each topic in the form "{
    subject}: <related_domain>" where `<
    related_domain>` is:
- A short, self-contained descriptor (no
    more than n={n} words), such as "
    entrepreneurship", "agriculture", "
    politics", or "natural ecology";
- A concrete subdomain within {subject},
    or an interdisciplinary area related
    to {subject}.
3. Return between 1 and 3 topics. If the
    content is no plausible connection to
     {subject}, return exactly:[  {"topic
    ":"Irrelevant Data","explanation":"
    None" }], not"Climate Change:
    Irrelevant Data".
4. Output a pure array of objects in
    English, each with:
-"topic": a string like"Climate Change:
    Economic Phenomena";
-"explanation": a string that explains the
     meaning of the topic in the context
    of {subject}, using no more than m =
    {m} words.

[Examples (4 positive and 2 negative).]
```

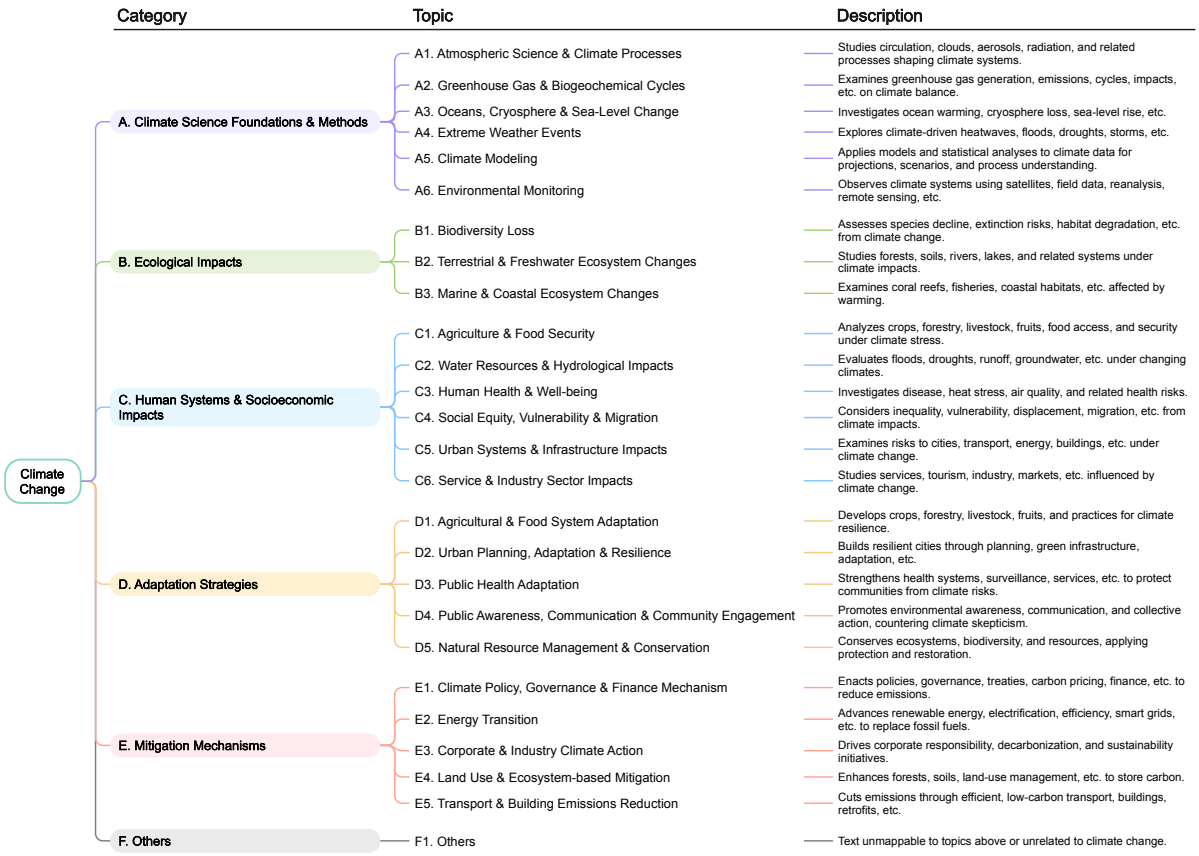

Figure 12: Final Topic Taxonomy of Climate Change.

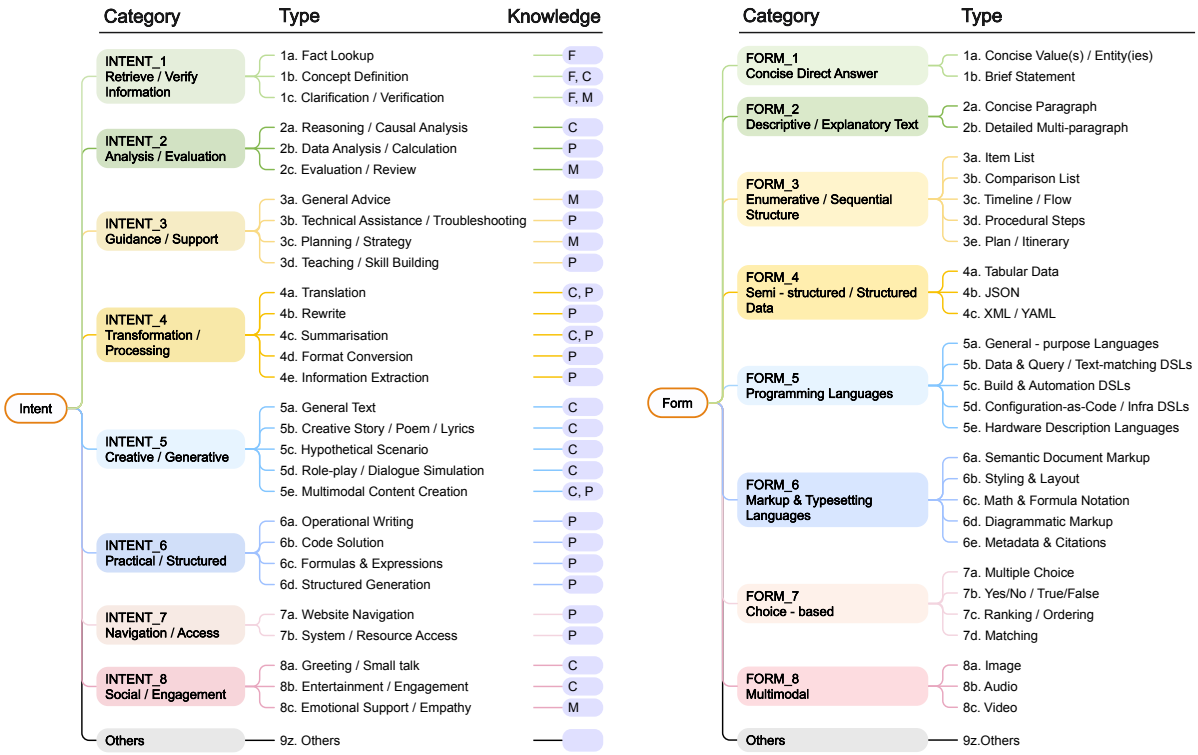

Figure 13: Taxonomy of question types, including user intents the expected answer forms. The Intent Taxonomy also contains a "Knowledge" label indicating the type of knowledge required by an LLM for each intent: **F** for Factual, **C** for Conceptual, **P** for Procedural, and **M** for Metacognitive. Each category also has an *z. Others* type, which is omitted from the figure.

Category	Dataset	Task/Format	Brief Description	Count
Auxiliary Corpora	Climate-FEVER	Fact Verification	Internet climate-related claims for factuality assessment.	1,452
	Environmental Claims	Claim Detection (Corporate)	Company environmental statements from annual/sustainability reports; corporate perspective.	1,270
	ClimSight	QA Dataset	A question-answering dataset related to climate change.	2,998

Table 2: An overview of the 3 Auxiliary Corpora used in this study.

2. Topic Identification: Iterative Topic Merging

```
You are a meticulous topic curator. Each
    topic is represented as a JSON object
     with two keys:
-"topic": the topic label
-"explanation": a short description of
    what that topic covers

The system will pass you a JSON array
    where the first element is the parent
     topic, and the remaining elements
    are the candidate topics to consider.
     **Each object also includes an"id"
    field which is the 1-based row index
    (as a string). Use these"id" values
    for selection:**
[
  {"id":"1","topic":"Climate Change:
    ParentLabel","explanation":"Parent
    explanation here" },
  {"id":"2","topic":"Candidate A","
    explanation":"Explanation A" },
  {"id":"3","topic":"Candidate B","
    explanation":"Explanation B" },
  ...
]

Your job is to merge those candidates
    whose both topic and explanation
    clearly match the parent concept.

Return a JSON object with exactly three
    keys:

{
 "merged_ids": [ /* array of candidate IDs
     (strings like"2","3") to merge */ ],
 "parent_topic":     "...",    /* either
    unchanged or a clearer single topic
    label */,
 "parent_explanation":"..."    /* a
    concise explanation for the chosen
    parent_topic */
}

Rules:
1. ONLY merge when the candidate's topic
    label is a near-synonym of the parent
    's label (same concept, same
    granularity) and the candidate's
    explanation bidirectionally entails
    the parent's explanation (each would
    still be true if swapped). Do not
    merge if the candidate is a sibling
    area.
2. If unsure, do not merge! Just returning
     an empty"merged_ids" over a risky
    merge; only merge high-confidence
    near-duplicates.
3. Maintain the same semantic level (e.g.,
     do not merge into a broader or
    narrower topic).
4. Do not include the parent itself in"
    merged_ids".
5. You must only select items from the
    provided list and return their"id"
    values in"merged_ids" (do not invent
    new IDs).
6. If no candidates should merge, return
    an empty array for"merged_ids" and
    repeat the original"parent_topic" and
    "parent_explanation".
7. You may revise the"parent_topic" to
    improve clarity or specificity, but
    it must follow the required topic
    naming format below.

Topic Naming Format (Mandatory for
    parent_topic and all candidates):
1. Each topic must be expressed in the
    form:"{subject}: <related_domain>"
    where <related_domain> must be:
- A short, self-contained descriptor (no
    more than n={n} words), such as "
    entrepreneurship", "agriculture", "
    politics", or "natural ecology";
- A concrete subdomain within {subject},
    or an interdisciplinary area related
    to {subject}.
- Avoid vague or overly general terms like
     "issues", "overview", or "aspects".
2. Each topic's explanation explains the
    general meaning of the topic in the
    context of {subject}, using no more
    than m = {m} words.

[A complete example of topic merging.]
```

3. Topic Identification: Topic Reassignment

```
You are an expert in climate change
    science.
Classify the input text into at most 3
    core topics based on the taxonomy
    below.

**Decision Rules:**
1. Identify only **core, directly relevant
    ** topics (do not assign marginal or
    weakly related ones).
```

```
2. Return up to 3 topics; fewer is
     acceptable.
3. If the text **does not clearly map to
     any topic or unrelated to climate
     change**, return Others only.
   - When classifying as Others, return
     only one item:
   [
     {"topic":"F1. Others" }
   ]
4. Do not create new topics. Use only the
     taxonomy provided.

Output Format (STRICT)
Return ONLY a JSON array of 1-3 objects,
     each with a single field:
[
  {"topic":"A5. Climate Modeling" },
  {"topic":"E1. Climate Policy, Governance
      & Finance Mechanism" }
]

[The complete, two-level topic taxonomy (
     including explanations for each topic
     ).]
[Examples (3 positive and 1 negative).]
```

4. Question Type Classification

```
You are a query classification expert.
Given a user query Q, classify it along
     two independent axes:
AXIS A - INTENT (What the user wants to
     achieve)
AXIS B - FORM (The most appropriate form
     of the answer)
The texts you need to identify will come
     from various sources, including real-
     world user queries, online Q&A, AI-
     generated questions, social media
     posts, and more. At the same time,
     these texts may appear in any
     language (including English, Chinese,
      Russian, French, etc.). You need to
     detect the language, but always
     produce the output in English.

==============================
Classification rules
==============================

1. For both INTENT and FORM, assign only
     one main category if possible.
2. If the query reasonably fits multiple
     categories (e.g., user expects both
     evaluation and advice), assign all
     applicable categories but no more
     than 3.
3. When ambiguous in FORM (e.g., concept
     explanation may be concise or
     detailed), assign multiple categories
      as needed but no more than 3.
4. If multiple categories are assigned for
      either INTENT or FORM, list them
     together in an array in the output,
     in priority order (most relevant
     first).
5. If no category fits, assign the
     appropriate Others category (marked
     with "z").
6. Output must in English and format must
     be:
   {
     "intent": ["INTENT_xx. xxx", "
     INTENT_yy. yyy", ...],
     "form": ["FORM_xx. xxx ", "FORM_yy.
     yyy", ...]
   }
7. Do not output any text outside the JSON
      structure.

[The complete, two-level intent and form
     taxonomy (including brief
     explanations for each intent/form and
      linguistic clues (e.g., common query
      patterns) for each intent).]
[Examples (4 positive).]
```

B.3 LLM usage and configurations.

For **Filtering Data** from *WildChat* and *LMSYS-Chat-1M*, we used the `Qwen-3 30B` model (Yang et al., 2025) (temperature=0, reasoning mode disabled). For **Initial Topic Generation**, we employed `GPT-4.1-mini` (OpenAI, 2025b) (temperature=0.2). For **Iterative Topic Merging**, we used `GPT-4.1-mini` (temperature=0.2) or `GPT-5-mini` (minimal reasoning effort) depends on merge settings. Both **Topic Reassignment** and **Question Type Classification** were performed using `GPT-5-mini` (OpenAI, 2025c) (low reasoning effort).

B.4 Additional Details on Topic Modeling.

Initial Topic Generation. To ensure consistency in the generated topics, we impose a constrained naming format and granularity requirement within the prompt. Each topic follows the structure *{subject}: <related-domain>*, where *<related-domain>* denotes a specific area closely associated with the research subject (in our case, climate change). Empirically, this naming convention helps align the generated topics more closely with the semantic scope of the target subject.

To further control topic granularity, we impose two additional constraints: the *related-domain* component is limited to at most four words, and the accompanying explanation is restricted to at most twenty words. These restrictions prevent the model from generating overly fine-grained topics that capture idiosyncratic details of individual samples rather than generalizable thematic concepts.

Iterative Topic Merging. In the iterative merging stage, we first sort the initial topic list by frequency in descending order. In each iteration, the most frequent topic that has not yet been merged is selected as the anchor topic. This strategy is motivated by the intuition that high-frequency topics tend to be more representative of the corpus and therefore serve as effective anchors for consolidating semantically related topics.

For each anchor topic, we retrieve the top-k most similar candidate topics as potential merge targets, where $k = \min(10, \max(1, \lfloor \textit{Topic List Size}/10 \rfloor))$. Empirically, this value provides a practical balance between coverage and quality: a smaller k reduces the efficiency of the merging process, while a larger k tends to degrade LLM performance during the semantic comparison stage.

Although embedding similarity provides a useful signal for identifying potential candidates, it may still be affected by biases or coverage limitations in the embedding model's training data. As a result, some topics with similar vector representations may correspond to conceptually unrelated themes. For example, topics such as "Drought Risk" and "Urban Planning" may appear close in the embedding space despite describing different concepts.

After each round of merging, the remaining topics are carried forward to the next iteration. The process continues until one of the following stopping conditions is satisfied: (1) **Inactivity**, where no merges occur during the current iteration, indicating that the process has converged; or (2) **Spread-based criteria**, where the semantic similarity among the remaining topics is sufficiently low (mean similarity < 0.3 or maximum similarity < 0.5), suggesting that the remaining topics are already sufficiently diverse and further merging is unlikely to be beneficial.

In practice, the merging process typically terminates due to the inactivity condition. The topics produced in the *Iterative Topic Merging* stage retain the same naming format as those generated in the *Initial Topic Generation* stage. The pseudocode of the topic merging algorithm is presented in Algorithm 1, and the experiment analyzing the topic list obtained after topic merging is presented in Appendix C.3.

Topic Reassignment. In the Topic Reassignment stage, we use the data labeled under the S5 configuration of the topic merging experiments (see Appendix B.5) as the basis for relabeling. From this dataset, we remove 4,645 samples that were labeled solely as *"Irrelevant Data"* in order to avoid unnecessary computational cost during reassignment. The S5 configuration is selected because, based on a qualitative inspection of the topic merging trajectories, it exhibits the most coherent and well-structured merging behavior among the six experimental settings (see the visualization in Appendix C.5).

After filtering out the irrelevant samples, a total of 46,346 samples are reassigned to topics. Among these, 42,261 samples correspond to core-topic categories after excluding 4,085 samples labeled as *"F. Others"*.

B.5 Experimental Configurations for Six Topic Merging Runs

Table 3 summarizes the six experimental configurations of topic merging pipeline. Each configuration differs in (i) the textual input selected for embedding extraction (topic only vs. topic name + explanation), (ii) the embedding model (all-MiniLM-L6-v2 or Qwen3-Embedding-4B) employed (Transformers, 2020; Zhang et al., 2025), and (iii) the LLM for executing the topic merging step. The last column reports the final number of merged topics obtained from the initial pool of 10,730 topics.

B.6 Manual Topic Reassignment Details

To construct a comprehensive taxonomy, we collected topic candidates from each of the six experimental configurations. The selected lists are shown below. Table 4 maps the selected topics from six experimental configurations (S1-S6) to the final taxonomy, grouped by higher-level categories for clarity.

Setting 1. Amount=23

Climate Change: Agriculture, Climate Change: Air Pollution, Climate Change: Atmospheric Science, Climate Change: Biodiversity Conservation, Climate Change: Climate Modeling, Climate Change: Corporate Sustainability, Climate Change: Cryosphere Dynamics, Climate Change: Environmental Impact, Climate Change: Extreme Weather Events, Climate Change: Health Impacts, Climate Change: Hydrological Change, Climate Change: Mitigation Strategies, Climate Change: Policy & Governance, Climate Change: Public Perception,

Algorithm 1: Iterative Topic Merging

Input: Initial topic set $\mathcal{T} = \{(t_i, e_i, c_i, \mathbf{v}_i)\}_{i=1}^{N}$, where t_i is the topic label, e_i its explanation, c_i its frequency, and $\mathbf{v}_i$ its embedding; maximum batch size B; stopping thresholds θ_{mean} and $\theta_{\max}$.

Output: A refined topic set $\mathcal{T}^*$ and a hierarchical merge tree $\mathcal{M}$.

Collapse topics with identical normalized (t_i, e_i) by summing counts and recomputing embeddings
repeat
  Sort $\mathcal{T}$ in descending order of c_i
  Mark all topics as unmerged
  Initialize $\mathcal{T}_{\text{new}} \leftarrow \varnothing$ and $\mathcal{M}_{\text{round}} \leftarrow \varnothing$
  Set `any_merge` $\leftarrow$ `false`
  foreach *unmerged topic* $p \in \mathcal{T}$ *in sorted order* **do**
    if p *is locked* **then**
      Move p to $\mathcal{T}_{\text{new}}$ and record a self-link in $\mathcal{M}_{\text{round}}$
      mark p as merged and **continue**
    Let $\mathcal{U}$ be the set of unmerged, unlocked topics excluding p
    if $\mathcal{U} = \varnothing$ **then**
      Move p to $\mathcal{T}_{\text{new}}$ and record a self-link in $\mathcal{M}_{\text{round}}$
      mark p as merged and **continue**
    Compute cosine similarities between p and all topics in $\mathcal{U}$
    Select the top-k most similar candidates, where

$$k = \min\big(B, \max(1, \lfloor |\mathcal{T}|/10 \rfloor)\big).$$

    Query the LLM with p and the selected candidates
    The LLM returns a subset $\mathcal{C}$ of candidates to merge with p, together with an updated parent topic label and explanation $(t^\star, e^\star)$
    Let $\mathcal{S} = \{p\} \cup \mathcal{C}$
    if $|\mathcal{S}| > 1$ **then**
      `any_merge` $\leftarrow$ `true`
    Construct a new parent topic from $\mathcal{S}$:

$$c^\star = \sum_{x \in \mathcal{S}} c_x, \qquad \mathbf{v}^\star = \text{Normalize}\left(\frac{\sum_{x \in \mathcal{S}} c_x \mathbf{v}_x}{\sum_{x \in \mathcal{S}} c_x}\right).$$

    Append $(t^\star, e^\star, c^\star, \mathbf{v}^\star)$ to $\mathcal{T}_{\text{new}}$
    Record the parent–children relation in $\mathcal{M}_{\text{round}}$
    Mark all topics in $\mathcal{S}$ as merged
  Add self-links for any uncovered topics, if needed
  Deduplicate $\mathcal{T}_{\text{new}}$ by identical normalized (t, e), summing counts and recomputing embeddings
  Update the global merge tree $\mathcal{M}$ using the deduplicated parent topics
  Set $\mathcal{T} \leftarrow \mathcal{T}_{\text{new}}$
  Compute the mean and maximum pairwise cosine similarities among topics in $\mathcal{T}$
until *no merge occurs in the current round* ***or*** *the mean inter-topic similarity is below* θ_{mean} ***or*** *the maximum inter-topic similarity is below* θ_{max}
return $\mathcal{T}, \mathcal{M}$

Setting ID	Text for embedding	Embedding model	Merging model	# Merged topics
S1	Topic only	all-MiniLM-L6-v2	gpt-4.1-mini	722
S2	Topic only	all-MiniLM-L6-v2	gpt-5-mini	286
S3	Topic name + explanation	all-MiniLM-L6-v2	gpt-4.1-mini	540
S4	Topic name + explanation	all-MiniLM-L6-v2	gpt-5-mini	187
S5	Topic name + explanation	Qwen3-Embedding-4B	gpt-4.1-mini	411
S6	Topic name + explanation	Qwen3-Embedding-4B	gpt-5-mini	170

Table 3: Six experimental configurations of the topic merging pipeline.

Climate Change: Regional Climate Impacts, Climate Change: Renewable Energy, Climate Change: Economic Impacts, Climate Change: Greenhouse Gas Emissions, Climate Change: Social Vulnerability, Climate Change: Societal Transformation, Climate Change: Adaptation Strategies, Climate Change: Energy Transition, Climate Change: Biogeochemical Cycles.

Setting 2. Amount=22

Climate Change: Agriculture, Climate Change: Atmospheric Science, Climate Change: Biological Responses, Climate Change: Climate Modeling, Climate Change: Climate Variability, Climate Change: Cloud Processes, Climate Change: Community Engagement, Climate Change: Energy Systems, Climate Change: Extreme Weather, Climate Change: Fossil Fuels, Climate Change: Geophysical Events, Climate Change: Governance, Climate Change: Inventory Methodologies, Climate Change: Mitigation Strategies, Climate Change: Public Health, Climate Change: Radiative Forcing, Climate Change: Regional Policy, Climate Change: Economic Assessment, Climate Change: Economic Structures, Climate Change: Tourism Impacts, Climate Change: Transportation, Climate Change: Urban Planning.

Setting 3. Amount=25

Climate Change: Agricultural Biotechnology, Climate Change: Agriculture, Climate Change: Atmospheric Processes, Climate Change: Biodiversity Loss, Climate Change: Climate Modeling, Climate Change: Corporate Sustainability, Climate Change: Cryosphere Changes, Climate Change: ENSO Dynamics, Climate Change: Energy Transition, Climate Change: Environmental Monitoring, Climate Change: Extreme Weather, Climate Change: Oceanography, Climate Change: Policy & Governance, Climate Change: Public Awareness, Climate Change: Temperature Trends, Climate Change: Tourism Impacts, Climate Change: Water Resources, Climate Change: Adaptation Strategies, Climate Change: Climate Finance, Climate Change: Community Engagement, Climate Change: Evolutionary Biology, Climate Change: Financial Risk, Climate Change: Human Health and Well-being, Climate Change: Sustainable Development, Climate Change: Urban Planning.

Setting 4. Amount=31

Climate Change: Agriculture, Climate Change: Biodiversity Conservation, Climate Change: Carbon Cycle, Climate Change: Climate Modeling, Climate Change: Climate Science, Climate Change: Cryosphere Changes, Climate Change: Emissions Mitigation, Climate Change: Expert Judgment, Climate Change: Extreme Weather, Climate Change: Fossil Fuel Extraction, Climate Change: Hydrological Impacts, Climate Change: Industry Sustainability, Climate Change: Marine Ecology, Climate Change: Policy & Governance, Climate Change: Sectoral Impacts, Climate Change: Social Equity, Climate Change: Data & Observation, Climate Change: Human-Environment Interaction, Climate Change: Natural Disturbances, Climate Change: Public Health, Climate Change: Electrification Impacts, Climate Change: Sustainable Development, Climate Change: Transportation Emissions, Climate Change: Indoor Air Quality, Climate Change: Radiative Forcing, Climate Change: Socioeconomic Policy, Climate Change: Urban Heat Mitigation, Climate Change: ESG Investing, Climate Change: Environmental Pollution, Climate Change: Net Zero Targets.

Setting 5. Amount=34

Climate Change: Aerosol Radiative Effects, Climate Change: Agriculture, Climate Change: Agriculture & Pollution, Climate Change: Atmospheric Processes, Climate Change: Biodiversity Loss, Climate Change: Carbon Cycle, Climate Change: Circular Economy, Climate Change: Climate Teleconnections, Climate Change: Corporate Responsibility, Climate Change: Ecosystem Impacts, Climate Change: Energy Demand Management, Climate Change: Energy Efficiency, Climate Change: Energy Sector, Climate Change: Environmental Monitoring, Climate Change: Extreme Weather, Climate Change: Geoengineering, Climate Change: Human Well-being, Climate Change: Hydrological Changes, Climate Change: Impact Modeling, Climate Change: Industrial Emissions, Climate Change: Internal Variability, Climate Change: Marine Ecosystems, Climate Change: Mitigation Pathways, Climate Change: Polar Regions, Climate Change: Policy & Governance, Climate Change: Public Engagement, Climate Change: Radiative Forcing, Climate Change: Risk Management, Climate Change: Scientific Literature, Climate Change: Sectoral Analysis, Climate Change: Sustainable Products, Climate Change: Temperature Trends, Climate Change: Tourism Impacts, Climate Change: Urban Adaptation.

Setting 6. Amount=21

Climate Change: Carbon Accounting, Climate Change: Climate Monitoring, Climate Change: Ecosystem Services, Climate Change: Emergency Preparedness, Climate Change: Energy Systems, Climate Change: Food Systems, Climate Change: Governance, Climate Change: Human Health, Climate Change: Hydrology & Flooding, Climate Change: Land Use, Climate Change: Physical Mechanisms, Climate Change: Planetary Geology, Climate Change: Pollution Management, Climate Change: Renewable Energy, Climate Change: Climate Finance, Climate Change: Human Habitability, Climate Change: Infrastructure Development, Climate Change: Misinformation, Climate Change: Ozone Depletion, Climate Change: Paleoclimate, Climate Change: Urbanization.

B.7 Framework Analysis Perspectives

Under the *Proactive Knowledge Behaviors Framework*, we identify three meaningful comparative perspectives based on the knowledge behaviors among the three actors.

The first perspective is termed *multi-actor behavioral comparison on one target*, which examines behavioral differences among different actors (e.g., different human knowledge providers) when interacting with the same target (e.g., an AI knowledge provider).

The second perspective is termed *intra-actor behavioral comparison*, which focuses on how the same actor exhibits different behaviors when interacting with different targets.

The third perspective is termed *inter-actor behavioral comparison*, which compares behavioral differences across different types of actors.

Based on these three comparative perspectives, with moderate extensions, it becomes possible to systematically operationalize the analyses corresponding to Steps 1–4 in Table 5. We exclude comparisons that are sequential in nature and lack informational value, such as comparing *human ask human* vs. *human guide AI*, or *human inform human* vs. *human ask AI*. By leveraging these three analytical perspectives, this approach also enables the framework to be systematically extended beyond the context of climate change to other socio-scientific research domains.

B.8 Dataset and Annotation Examples

Here we present examples from the eight core datasets used in this study, along with their annotated labels for topic and question type.

B.8.1 WildChat

"text": "Water management is a critical adaptation strategy for climate change. Which of the following does not support this position A. Technology should focus on use of marine water resources B. Finding alternative water stores C. Implementing legislation to ensure the 'fair' distribution of water D. Increasing water conservation in periods of water surplus",
"Final_Topics": ["C2. Water Resources & Hydrological Impacts", "D5. Natural Resource Management & Conservation"],
"Final_Question_Types":
"Intent": ["INTENT_1a. Fact Lookup"],
"Form": ["FORM_7a. Multiple Choice"]

Category	Topic	Mapping to Taxonomy
Climate Science Foundations & Methods	Atmospheric Science & Climate Processes	Atmospheric Science (S1, S2); Cloud Processes (S2); Radiative Forcing (S2, S4, S5); Atmospheric Processes (S3, S5); Climate Science (S4); Internal Variability (S5)
	Greenhouse Gas & Biogeochemical Cycles	Biogeochemical Cycles (S1); Greenhouse Gas Emissions (S1); Fossil Fuels / Fossil Fuel Extraction (S2, S4); Carbon Cycle (S4, S5); Carbon Accounting (S6)
	Oceans, Cryosphere & Sea-Level Change	Cryosphere Dynamics (S1); Oceanography (S3); Cryosphere Changes (S3, S4); Marine Ecology/Ecosystems (S4, S5); Polar Regions (S5)
	Extreme Weather Events	Extreme Weather Events (S1); Extreme Weather (S2, S3, S4, S5)
	Climate Modeling	Climate Modeling (S1, S2, S3, S4); Impact Modeling (S6)
	Environmental Monitoring	Environmental Monitoring (S3, S5); Data & Observation (S4); Climate Monitoring (S6)
Ecological Impacts	Biodiversity Loss	Biodiversity Conservation (S1, S4); Biodiversity Loss (S3, S5); Evolutionary Biology (S3)
	Terrestrial & Freshwater Ecosystem Changes	Hydrological Change/Changes (S1, S5); Hydrological Impacts (S4); Ecosystem Impacts (S5); Hydrology & Flooding (S6)
	Marine & Coastal Ecosystem Changes	Environmental Impact (S1); Oceanography (S3); Marine Ecology/Ecosystems (S4, S5); Ecosystem Impacts (S5)
Human Systems & Socioeconomic Impacts	Agriculture & Food Security	Agriculture (S1, S2, S3, S4, S5); Agricultural Biotechnology (S3); Food Systems (S6)
	Water Resources & Hydrological Impacts	Hydrological Change/Changes (S1, S5); Water Resources (S3); Hydrological Impacts (S4); Hydrology & Flooding (S6)
	Human Health & Well-being	Health Impacts (S1); Public Health (S2, S4); Human Health and Well-being (S3); Human Well-being (S5); Human Health (S6); Human Habitability (S6)
	Social Equity, Vulnerability & Migration	Social Vulnerability (S1); Societal Transformation (S1); Social Equity (S4); Human-Environment Interaction (S4)
	Urban Systems & Infrastructure Impacts	Urban Planning (S2, S3); Urban Heat Mitigation (S4); Infrastructure Development (S6); Urbanization (S6)
	Service & Industry Sector Impacts	Economic Impacts (S1); Tourism Impacts (S2, S3, S5); Economic Assessment (S2); Economic Structures (S2); Sectoral Impacts (S4); Industry Sustainability (S4); Electrification Impacts (S4); Sectoral Analysis (S5)
Adaptation Strategies	Agricultural & Food System Adaptation	Adaptation Strategies (S1, S3); Food Systems (S6)
	Urban Planning, Adaptation & Resilience	Adaptation Strategies (S1, S3); Urban Planning (S2,S3); Urban Heat Mitigation (S4); Urban Adaptation (S5); Infrastructure Development (S6)
	Public Health Adaptation	Health Impacts (S1); Public Health (S2, S4); Human Health and Well-being (S3); Human Well-being (S5); Human Health (S6)
	Public Awareness, Communication & Community Engagement	Public Perception (S1); Community Engagement (S2, S3); Public Awareness (S3); Public Engagement (S5); Misinformation (S6)
	Natural Resource Management & Conservation	Adaptation Strategies (S1, S3); Natural Disturbances (S4); Ecosystem Services (S6); Land Use (S6); Pollution Management (S6)
Mitigation Mechanisms	Climate Policy, Governance & Finance Mechanism	Policy & Governance (S1, S3, S4, S5); Governance (S2, S6); Regional Policy (S2); Climate Finance (S3, S6); Financial Risk (S3); Socioeconomic Policy (S4); ESG Investing (S4)
	Energy Transition	Energy Transition (S1, S3); Renewable Energy (S1, S6); Energy Systems (S2, S6); Electrification Impacts (S4); Energy Sector (S5); Energy Efficiency (S5)
	Corporate & Industry Climate Action	Corporate Sustainability (S1, S3); Sustainable Development (S3, S4); Industry Sustainability (S4); Corporate Responsibility (S5); Sustainable Products (S5); Circular Economy (S5)
	Land Use & Ecosystem-based Mitigation	Carbon Cycle (S4, S5); Natural Disturbances (S4); Land Use (S6); Ecosystem Services (S6)
	Transport & Building Emissions Reduction	Transportation (S2, S4); Transportation Emissions (S4); Electrification Impacts (S4); Infrastructure Development (S6)

Table 4: Primary mapping from collected topics to the final taxonomy.

"text": "*(Translated from Chinese into English)* 2. The attached file scrippsm.txt is a list of 180 numbers representing atmospheric carbon dioxide concentration, measured in ppv (parts per volume), i.e., the volume of CO2 contained in one million volumes of air. The data are records from the Mauna Loa volcano from January 1996 to December 2010, and this portion of the data was obtained from the Scripps study. (a) Perform a least-squares fit to the CO2 data using the model $f(t) = c_1 + c_2 t + c_3 \cos(2\pi t) + c_4 \sin(2\pi t)$, where t is measured in years. Report the optimal fitting coefficients Ci and the RMSE fitting error. Plot the continuous curve from January 1989 to the end of this year, together with the 180 data points. (b) Use your model to predict the CO2 concentrations in May 2004, September 2004, May 2005, and September 2005. These months tend to contain the annual maximum and minimum of the CO2 cycle. The actual recorded values are 380.63, 374.06, 382.45, and 376.73 ppv, respectively. Report the errors at these four points. (c) Add the extra term c5cos(4*pi*t) and redo (a) and (b). Compare the new RMSE and the model errors at the four points. (d) Repeat (c) using the additional term $c_6 t^2$. Which term provides the greatest improvement in the model, (c) or (d)? (e) Add both terms from (c) and (d), and redo (a) and (b). Prepare a table summarizing your results for each part of this problem, and provide an explanation for the outcomes as much as possible.", "Final_Topics": ["A5. Climate Modeling", "A2. Greenhouse Gas & Biogeochemical Cycles"], "Final_Question_Types": "Intent": ["INTENT_2b. Data Analysis / Calculation"], "Form": ["FORM_2b. Detailed Multi-paragraph", "FORM_4a. Tabular Data"]

"text": "*(Translated from Russian into English)* Greening the economy and economizing ecology: list the methods and examples of using these methods in the interaction between society and nature.", "Final_Topics": ["E1. Climate Policy, Governance & Finance Mechanism"], "Final_Question_Types": "Intent": ["INTENT_1a. Fact Lookup", "INTENT_3a. General Advice"], "Form": ["FORM_3a. Item List"]

B.8.2 LMSYS-Chat-1M

"text": "Describe the evolution of succulent plants with reference to past climate and climate change.",
"Final_Topics": ["B2. Terrestrial & Freshwater Ecosystem Changes"],
"Final_Question_Types":
"Intent": ["INTENT_2a. Reasoning / Causal Analysis", "INTENT_1b. Concept Definition"],
"Form": ["FORM_2b. Detailed Multi-paragraph"]

"text": "peux tu résume l'article M̈ethods to Optimize Carbon Footprint of Buildings in Regenerative Architectural Design with the Use of Machine Learning, publié par Convolutional Neural Network, and Parametric Design NAME_1 en 2020?",
"Final_Topics": ["E5. Transport & Building Emissions Reduction"],
"Final_Question_Types":
"Intent": ["INTENT_4c. Summarisation"],
"Form": ["FORM_2a. Concise Paragraph"]

"text": "Task: Analyzing Perspectives on Climate Change Background: Climate change is a complex and multifaceted issue that has been the subject of intense scientific study and public debate. The following paragraph provides some information on the causes and impacts of climate change: Paragraph: Climate change is caused by human activities such as burning fossil fuels, deforestation, and industrial processes, which release greenhouse gases into the atmosphere. These greenhouse gases trap heat from the sun, causing the Earth's temperature to rise and leading to a range of impacts, including more frequent and severe heatwaves, droughts, floods, and storms. Instructions: 1. Read the paragraph above. 2. Create a claim that either expresses skepticism or belief about climate change, using the information presented in the paragraph. 3. For each claim, classify it as either representing the views of a climate skeptic or a climate believer. Use the following format for your claims: Claim format: Claim statement. Classification: Skepticism or Not-skepticism. You can create as many claims as you like, but make sure you have at least one claim from each class.",

"Final_Topics": ["A2. Greenhouse Gas & Biogeochemical Cycles", "A4. Extreme Weather Events"],
"Final_Question_Types":
"Intent": ["INTENT_5a. General Text"],
"Form": ["FORM_4b. JSON"]

B.8.3 ClimateQ&A

"text": "can nuclear be a sustainable energy despite the difficulty of managing radioactive waste?",
"Final_Topics": ["E2. Energy Transition"],
"Final_Question_Types":
"Intent": ["INTENT_2a. Reasoning / Causal Analysis"],
"Form": ["FORM_2a. Concise Paragraph"]

"text": "what is the problem with biowaste management",
"Final_Topics": ["A2. Greenhouse Gas & Biogeochemical Cycles", "E1. Climate Policy, Governance & Finance Mechanism"],
"Final_Question_Types":
"Intent": ["INTENT_2a. Reasoning / Causal Analysis",
"INTENT_1b. Concept Definition"], "Form": ["FORM_2a. Concise Paragraph", "FORM_3a. Item List"]

"text": "What is the Global Warming Potential of R-454C?",
"Final_Topics": ["A2. Greenhouse Gas & Biogeochemical Cycles"],
"Final_Question_Types":
"Intent": ["INTENT_1a. Fact Lookup"],
"Form": ["FORM_1a. Concise Value(s) / Entity(ies)"]

B.8.4 ClimaQA-Gold

"text": "How did the CLAW Hypothesis proposed by Charlson et al. in 1987 suggest a mechanism to stabilize Earth's temperature?",
"Final_Topics": ["A1. Atmospheric Science & Climate Processes", "A2. Greenhouse Gas & Biogeochemical Cycles"],
"Final_Question_Types":
"Intent": ["INTENT_2a. Reasoning / Causal Analysis", "INTENT_1b. Concept Definition"],
"Form": ["FORM_1b. Brief Statement"]

"text": "Biogenic <MASK> contribute to the production of extremely low volatile organic compounds with atmospheric implications.",
"Final_Topics": ["A1. Atmospheric Science & Climate Processes"],
"Final_Question_Types":
"Intent": ["INTENT_1a. Fact Lookup"],
"Form": ["FORM_1a. Concise Value(s) / Entity(ies)"]

"text": "How does aerosol alter the radiative balance in the atmosphere in a hypothetical scenario where there is a significant increase in anthropogenic aerosol emissions?",
"Final_Topics": ["A1. Atmospheric Science & Climate Processes"],
"Final_Question_Types":
"Intent": ["INTENT_1a. Fact Lookup"],
"Form": ["FORM_7a. Multiple Choice"]

B.8.5 Clima-Silver

"text": "How would the overall number of cloud droplets be affected if the maximum supersaturation in a cloud formation process was lower than usual?",
"Final_Topics": ["A1. Atmospheric Science & Climate Processes"],
"Final_Question_Types":
"Intent": ["INTENT_2a. Reasoning / Causal Analysis"],
"Form": ["FORM_1b. Brief Statement"]

"text": "Disjunctive <MASK> allows for the assessment of probabilities of exceeding specific thresholds, enabling quantitative evaluation of the risk of inaction in environmental management scenarios.",
"Final_Topics": ["A5. Climate Modeling"],
"Final_Question_Types":
"Intent": ["INTENT_1a. Fact Lookup"],
"Form": ["FORM_1a. Concise Value(s) / Entity(ies)"]

"text": "How does the net greenhouse effect of Arctic low clouds contribute to the Earth's energy budget?",
"Final_Topics": ["A1. Atmospheric Science & Climate Processes"],
"Final_Question_Types":

"Intent": ["INTENT_1a. Fact Lookup"],
"Form": ["FORM_7a. Multiple Choice"]

B.8.6 Reddit

"text": "Former climate change denier here. This has been the hottest winter I've experienced (in Texas). Is this a result of climate change or a coincidence? Not being sarcastic. I've noticed that, at least in the part of Texas I'm in right now, it's been in the 80s for most of winter. It got below freezing today, but is expecting to go back into the 80s later this week. I know there's a difference between weather and climate, but I've never seen anything this erratic and weird. Y'all know more than me, so I'm interested in your thoughts.",
"Final_Topics": ["A4. Extreme Weather Events"],
"Final_Question_Types":
"Intent": ["INTENT_2a. Reasoning / Causal Analysis", "INTENT_1c. Clarification / Verification"],
"Form": ["FORM_2a. Concise Paragraph", "FORM_3a. Item List"]

"text": "When do you think climate change will become so undeniable that even the most stubborn deniers will no longer be able to ignore its impact? The question is asking when climate change will get so bad that even the biggest skeptics can't deny it anymore. It points to a tipping point where extreme weather, higher temperatures, and obvious signs of damage will make denial impossible. It reflects frustration with the ongoing doubt and the hope that clear proof will finally push everyone to take it seriously and act.",
"Final_Topics": ["A4. Extreme Weather Events", "D4. Public Awareness, Communication & Community Engagement"],
"Final_Question_Types":
"Intent": ["INTENT_2a. Reasoning / Causal Analysis"],
"Form": ["FORM_2b. Detailed Multi-paragraph", "FORM_2a. Concise Paragraph"]

"text": "Rising Temperatures, Rapid Aging: How Ambient Heat Affects DNA in Seniors Global warming is driving more frequent extreme heat events, posing severe health risks to senior citizens. By 2050, over 100 million Americans could be affected, with rising hospital admissions and cardiovascular issues. A recent study links heat exposure to epigenetic aging, with research showing DNA methylation changes in heart tissue and immune systems. What are your thoughts on this study?
Read more here : [Rising Temperatures and Rapid Aging](https://medtigo.com/news/rising-temperatures-rapid-aging-how-ambient-heat-affects-dna-in-seniors/)",
"Final_Topics": ["C3. Human Health & Well-being", "A4. Extreme Weather Events"],
"Final_Question_Types":
"Intent": ["INTENT_2c. Evaluation / Review"],
"Form": ["FORM_2a. Concise Paragraph"]

B.8.7 SciDCC

"text": "Climate. Warmer springs mean more offspring for prothonotary warblers. Climate change contributes to gradually warming Aprils in southern Illinois, and at least one migratory bird species, the prothonotary warbler, is taking advantage of the heat. A new study analyzing 20 years of data found that the warblers start their egg-laying in southern Illinois significantly earlier in warmer springs. This increases the chances that the birds can raise two broods of offspring during the nesting season, researchers found. They report their findings in the journal. Warmer springs in temperate regions of the planet can create a mismatch between when food is available to breeding birds and when their energy demands are highest – when they are feeding nestlings, said Illinois Natural History Survey avian ecologist Jeffrey Hoover, who conducted the study with INHS principal scientist Wendy Schelsky. If climate change diminishes insect populations at critical moments in the birds' nesting season, food shortages may cause some chicks and even adults to die. Other long-distance migratory birds suffer ill consequences from warmer springs on their breeding grounds. But this mismatch seems not to occur in prothonotary warblers, the researchers found. In the Cache River watershed, where the study was conducted, the warblers nest in swamps and forested wetlands, which are abundant sources of insects throughout the spring and summer. The warblers eat caterpillars, flies, midges, spi-

ders, mayflies and dragonflies. As long as it does not dry up, the wetland habitat provides a steady supply of food to sustain the birds. We studied populations of prothonotary warblers nesting in nest boxes in these areas from 1994–2013, Schelsky said. The researchers captured and color-banded the birds, studying 2,017 nesting female warblers in all. They visited the nests every three to five days from mid-April to early August, keeping track of when each bird laid its first egg and the birds' overall reproductive output for each nesting season. They also compared local April temperature trends with those of the Annual Global Land Temperature Anomaly data compiled by NOAA. They found that local conditions reflected global temperature changes. When local temperatures were warmer in April, the birds tended to lay their first eggs earlier in the spring and produced a greater number of offspring over the course of the breeding season. As the warming trend continues, conditions may change in ways that harm the birds' reproductive capacity. Warmer temperatures could cause local wetlands to dry up during the nesting season, cutting off the steady supply of insects the birds rely on to raise their young and increasing the exposure of nests to predators.",
"Final_Topics": ["B2. Terrestrial & Freshwater Ecosystem Changes", "B1. Biodiversity Loss"],
"Final_Question_Types":
"Intent": ["INTENT_4e. Information Extraction"],
"Form": ["FORM_4b. JSON"]

B.8.8 IPCC AR6

"text": "CMIP5 models with a model top within the stratosphere seriously underestimate the amplitude of the variability of the wintertime NAM expression in the stratosphere, in contrast to CMIP5 models which extend well above the stratopause (Lee and Black, 2015). However, even in the latter models, the stratospheric NAM events and their downward influence on the troposphere are insufficiently persistent (Charlton-Perez et al., 2013; Lee and Black, 2015). Increased vertical resolution does not show any significant added value in reproducing the structure and magnitude of the tropospheric NAM (Lee and Black, 2013) nor in the NAO predictability assessed in a seasonal prediction context with a multi-model approach (Butler et al., 2016). On the other hand, there is mounting evidence that a correct representation of the Quasi Biennial Oscillation, extratropical stratospheric dynamics (the polar vortex and sudden stratospheric warmings), and related troposphere-stratosphere coupling, as well as their interplay with ENSO, are important for NAO/NAM timing (Scaife et al., 2016; Karpechko et al., 2017; Domeisen, 2019; Domeisen et al., 2019), in spite of underestimated troposphere–stratosphere coupling found in models compared to observations (O'Reilly et al., 2019b).",
"Final_Topics": ["A5. Climate Modeling", "A1. Atmospheric Science & Climate Processes"],
"Final_Question_Types":
"Intent": ["INTENT_9z. Others"],
"Form": ["FORM_9z. Others"]

B.9 LLM Knowledge Requirements by User Intent

We introduce how each user intent maps to the knowledge dimensions: F=Factual, C=Conceptual, P=Procedural, M=Metacognitive. We mark $\star$ for primary (indispensable) knowledge and $\circ$ for auxiliary (helpful but not core). See Table 6, 7, 8, 9, 10, 11, 12, 13.

C Additional Results

C.1 Overall Distribution

The complete Topic, Intent, and Form similarities and distributions for all 11 datasets are shown in Figures 14, 15, 16, 17, 18, and 19. The Intent labels for *SciDCC*, *Climate-FEVER*, and *Environmental Claims* are fixed according to their original tasks. For *ClimaQA-Gold*, *ClimaQA-Silver*, *SciDCC*, *Climate-FEVER*, and *Environmental Claims*, the Form labels are fixed based on the datasets' original answer types. For *IPCC*, both the intent and form are assigned to *9z. Others*.

C.2 Cross-analysis

We further conduct a cross-analysis of the *Topic*, *Intent*, and *Form* of questions posed by humans (including both queries to LLMs and questions directed to other humans). Specifically, we examine Topic×Intent, Topic×Form, and Intent×Form (see Figure 20, 21 22). Overall, although Topic×Intent and Topic×Form show relatively high similarity

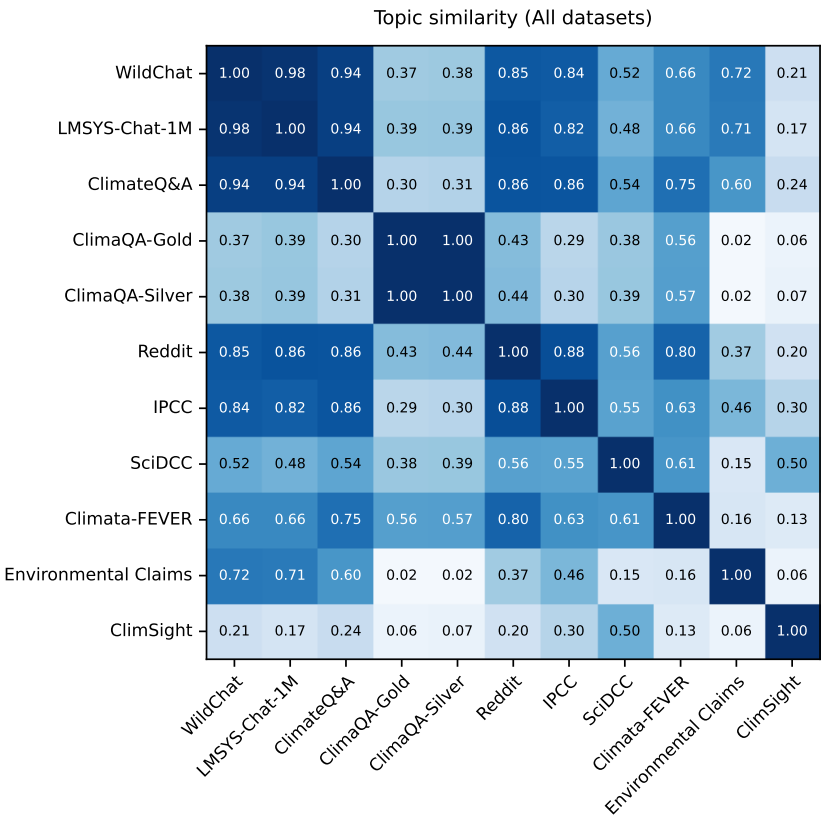

Figure 14: Topic Similarity for All 11 Datasets.

across datasets, their distributions are diffuse, with no strongly dominant pairings. By contrast, the results for Intent×Form align more closely with intuition: on Reddit, "*INTENT_2a. Reasoning / Causal Analysis* × *FORM_2a/2b*" accounts for higher shares (about 15% and 14%); in ClimateQ&A, "*INTENT_2a* × *FORM_2a*" reaches as high as 21%.

C.3 Analysis with Merged Topic Lists

For completeness, we also examined the topic list obtained directly after the topic merging step (Figure 23,fig:mt2, 25, 26, 27, 28). An analysis based on this merged topic list yields results that are broadly consistent with those obtained after applying our taxonomy and performing topic reassignment, this reflects the rationality and effectiveness of our Topic Merging algorithm design. However, to ensure the highest possible accuracy and alignment between topics and the designed classification scheme, we still conducted the topic reassignment procedure before performing the final analysis.

C.4 More Visualization

We built a local, interactive data-visualization web application (Figure 29) that supports comprehensive, in-depth, and cross-dimensional analysis across seven analytical dimensions. The web provides visualizations through tables, heatmaps, bar charts, and differential bar charts. It allows analysis by individual datasets or by groups (i.e., datasets under a specific knowledge behavior). Users can choose whether to include "Others" data, apply any of three different data-weighting methods, and conduct analysis at either the Category level or the more granular Topic/Type level.

Specifically, by visualizing the results under three different weighting methods, we can still arrive at conclusions that are broadly consistent with the main analysis of this paper. The three weighting methods are defined as follows:

1. **Label-count weighting**: each label is assigned an equal weight of 1, and the total is normalized by the sum of all labels.

2. **Per-sample weighting**: for each sample with K labels, the weight is evenly distributed, assigning $\frac{1}{K}$ to each label.

3. **Ranked weighting** (Used for): labels are weighted according to their order using a triangular scheme, where higher-ranked labels re-

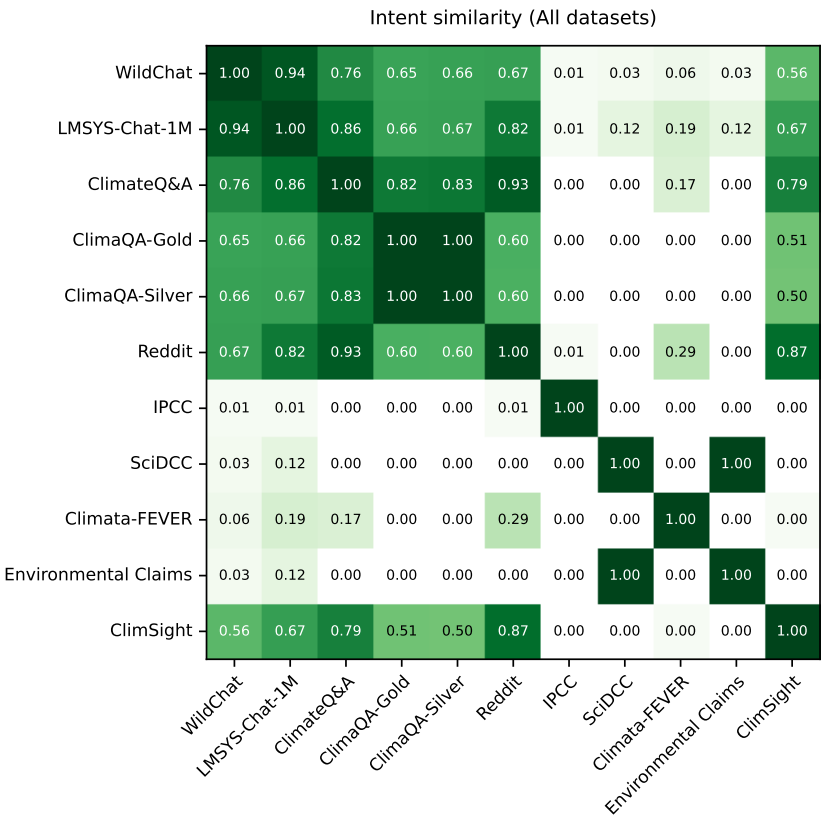

Figure 15: Intent Similarity for All 11 Datasets.

ceive greater weights. This weighting method is used in the main body of the paper.

C.5 Topic Merging Tree

We store the Topic Merging process using a tree structure and visualize it in an HTML page (as shown in Figure 30) to make it easier to inspect the merging workflow and its quality, and to support experimental design. Each topic can be clicked to reveal its id, corresponding data volume, explanation, and randomly sampled examples. Localized visualization via search is also supported.

C.6 Human Verification

We randomly sampled 150 data instances from the full dataset, including 15 instances from each of the eight core datasets and 10 instances from each of the three auxiliary datasets. We mapped the original labels to high-level taxonomy categories (Topic, Intent, and Form) for evaluation. Figure 31 shows that the sampled validation subset closely matches the full dataset in terms of Topic, Intent, and Form distributions, supporting its representativeness. Six annotators (four PhD-level, one Master's-level, and one undergraduate) each labeled 50 samples. The instructions for the annotators are as follows:

> Annotate the data with Topic and Question_Type (Intent + Form).
>
> - Each data topic/intent/form can be annotated with 1–3 labels, meaning a single data instance may ultimately have 3–9 labels.
> - When multiple labels exist, please try to rank them by relevance as much as possible, with the most relevant ones placed first.
> - Some of the intents or forms in the dataset have already been filled in, because those items don't require intent or form labeling.
> - Within the Topic categories, some type labels under C. Human Systems & Socioeconomic Impacts and D. Adaptation Strategies appear similar. However, Category C emphasizes impacts, while Category D emphasizes responses and adaptation.

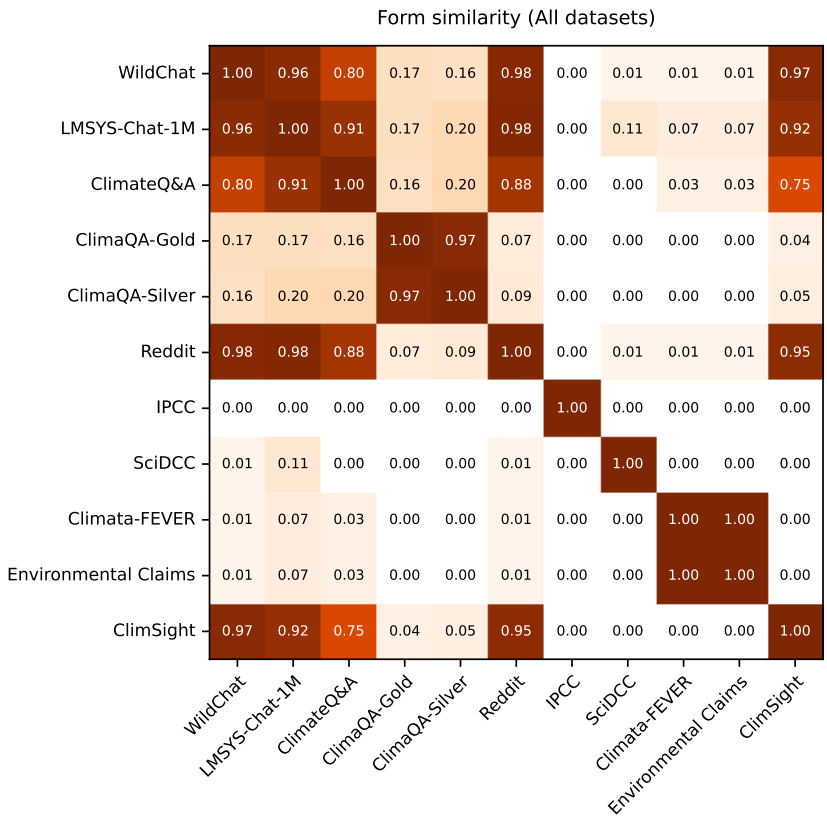

Figure 16: Form Similarity for All 11 Datasets.

- In addition, distinguishing between D. Adaptation Strategies and E. Mitigation Mechanisms can also be difficult. The goal of adaptation is to cope with climate change impacts that have already occurred or are unavoidable — climate change is already happening, and adaptation aims to make social, economic, and ecological systems more capable of resisting or withstanding these shocks. Mitigation, on the other hand, aims to reduce the magnitude of climate change itself — that is, to lower greenhouse gas emissions or enhance carbon sinks so that global warming becomes less severe.

Files:

- Taxonomy_Figure.pdf contains two figures of the taxonomy. In it, the topic taxonomy provides several descriptions, and the Topic and Type columns are the dimensions we need to annotate (e.g., A5. Climate Modeling, INTENT_5a. General Text, FORM_2a. Concise Paragraph).
- The Label_List folder provides the complete sets of labels.
- The Prompt folder contains descriptions, explanations, clues and examples for determining Topic and Question Type.
- Examples.jsonl provides several fully annotated examples.

Model–human agreement was measured using Jaccard similarity and Micro-F1, with 95% confidence intervals (CI) estimated via 1,000-round bootstrap resampling. As shown in Table 14, the LLM exhibits consistently high agreement with human across all categories (Jaccard scores of 0.706, 0.743, and 0.783 on topic, intent and form, respectively). The relatively narrow CI widths (approximately 0.10–0.12) suggest that the sample size is sufficient to support stable estimation of model–human agreement under the current evaluation setting. We further evaluated human inter-annotator agreement using pairwise Cohen kappa. The average pairwise Cohen kappa scores for Topic, Intent, and Form are 0.594, 0.656, and 0.735, respectively, indicating moderate to substantial agreement.

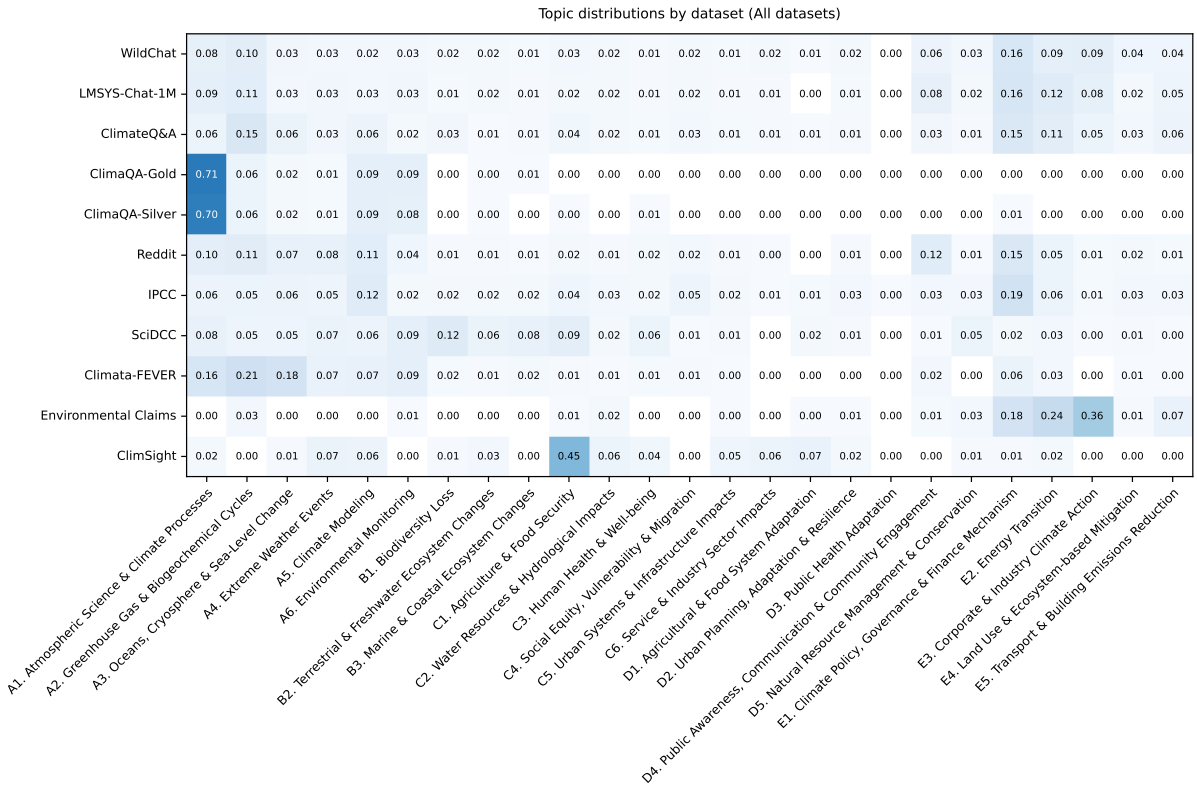

Figure 17: Topic Distribution for All 11 Datasets.

Step	Actor(s) Involved	Analytical Perspective	Behavioral Comparison
(1) Misalignment	Human Knowledge Seeker + Human Knowledge Provider	Multi-actor behavioral on one target comparison	Seeker Ask LLM vs. Provider Guide LLM
(2) Similarity	Human Knowledge Seeker	Intra-actor behavioral comparison	Ask Human vs. Ask LLM
(3) Reference	Human Knowledge Seeker ↔ Human Knowledge Provider	Inter-actor behavioral comparison	Seeker Ask Human vs. Provider Inform Human
(4) Insights	Human Knowledge Provider	Intra-actor behavioral comparison	Inform Human vs. Guide LLM

Table 5: Mapping of analysis steps to actors, analytical perspectives, and behaviors under the *Proactive Knowledge Behaviors Framework*.

Topic	F	C	P	M	Reason
1a-Fact lookup	⋆			○	Retrieve specific facts/data → atomic verifiable information; occasional uncertainty disclosure.
1b-Concept definition	⋆	⋆			Definitions/characteristics → boundaries and attributes of a concept.
1c-Clarification / verification	⋆	○		⋆	Verify truth/resolve ambiguity → fact checking plus criteria selection/uncertainty handling.

Table 6: Knowledge Requirements for INTENT_1 (Information retrieval).

Topic	F	C	P	M	Reason
2a-Reasoning / causal analysis		⋆	○	○	Explain causes/effects/comparisons → conceptual relations frame the analysis; procedural steps and strategy are supportive.
2b-Data analysis / calculation		○	⋆	○	Perform numerical/statistical computation → execution of methods and pipelines is core; method choice is supportive.
2c-Evaluation / review		○	○	⋆	Assess and rate → establishing criteria, weighting evidence, and judgment are metacognitive.

Table 7: Knowledge Requirements for INTENT_2 (Analysis / evaluation).

Topic	F	C	P	M	Reason
3a-General advice		○	○	⋆	Non-technical recommendations → context-sensitive trade-offs and strategy selection.
3b-Technical assistance / troubleshooting		○	⋆	○	Practical help resolving issues → reproducible diagnose–test–verify procedure.
3c-Planning / strategy		○	○	⋆	Plans/roadmaps/schedules → goal decomposition, prioritization, and monitoring.
3d-Teaching / skill building		○	⋆	○	Tutorials/step-by-step guidance → executable sequences are central.

Table 8: Knowledge Requirements for INTENT_3 (Guidance / support).

Topic	F	C	P	M	Reason
4a-Translation		⋆	⋆		Convert between languages → semantic/pragmatic mapping (C) executed via translation procedures (P).
4b-Rewrite		○	⋆	○	Paraphrase/change tone → controlled transformation process; style/genre knowledge assists.
4c-Summarisation		⋆	⋆		Condense to key points → discourse structure grasp (C) + filtering/compression steps (P).
4d-Format conversion			⋆		Transform file/data formats → structural mapping procedure.
4e-Information extraction	○	○	⋆		Extract entities/facts → extraction rules/pipeline; schema awareness can help.

Table 9: Knowledge Requirements for INTENT_4 (Text transformation).

Topic	F	C	P	M	Reason
5a-General text generation		⋆	○	○	Open-ended writing → genre/discourse patterns guide content organization.
5b-Creative story / poem / lyrics		⋆	○	○	Narrative/poetic forms and conventions drive creation.
5c-Hypothetical scenario		⋆	○	○	Counterfactual "what if" → conceptual modeling of assumptions/relations.
5d-Role-play / dialogue simulation		⋆	○	○	Personas and pragmatics (register, intent) are conceptual; execution is secondary.
5e-Multimodal creation		⋆	⋆	○	Prompts for images/audio/video → cross-modal semantics (C) plus construction procedures (P).

Table 10: Knowledge Requirements for INTENT_5 (Creative / generative).

Topic	F	C	P	M	Reason
6a-Operational writing		○	⋆	○	Emails/reports/ads → templates, drafting workflow, and deliverable constraints.
6b-Code solution	○	○	⋆	○	Write/fix code → syntax, APIs, test–debug procedures; specific constants as facts.
6c-Formulas & expressions	○	○	⋆		LaTeX/Excel/regex → expression/syntax construction procedures.
6d-Structured generation		○	⋆	○	Tables/schedules/checklists → schema-constrained generation and formatting steps.

Table 11: Knowledge Requirements for INTENT_6 (Practical / structured outputs).

Topic	F	C	P	M	Reason
7a-Website navigation		○	⋆	○	Open URLs/pages → path/step execution; choose shortest/robust route.
7b-System / resource access		○	⋆	○	Open files/systems → operation sequences with basic safety/permission strategy.

Table 12: Knowledge Requirements for INTENT_7 (Navigation / access).

Topic	F	C	P	M	Reason
8a-Greeting / small talk		⋆	○	○	Casual conversation → pragmatics/etiquette schemas; light execution/adjustment.
8b-Entertainment / engagement		⋆	○	○	Jokes/games → genre/game conventions; rule execution secondary.
8c-Emotional support / empathy		○	○	⋆	Supportive/empathetic responses → boundary setting, validation, escalation judgment.

Table 13: Knowledge Requirements for INTENT_8 (Social / engagement).

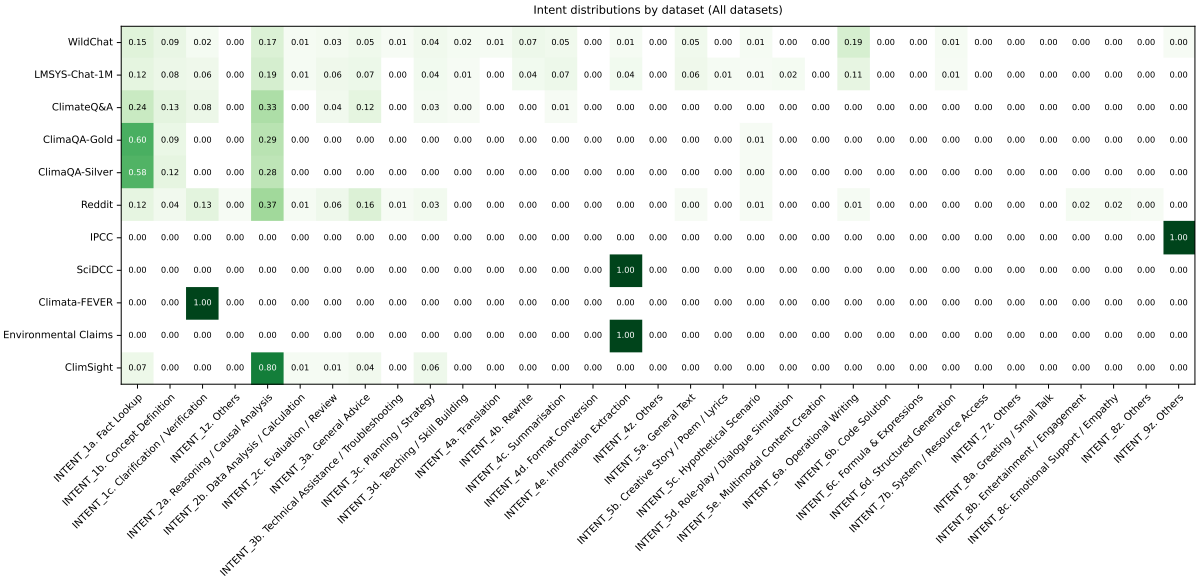

Figure 18: Intent Distribution for All 11 Datasets.

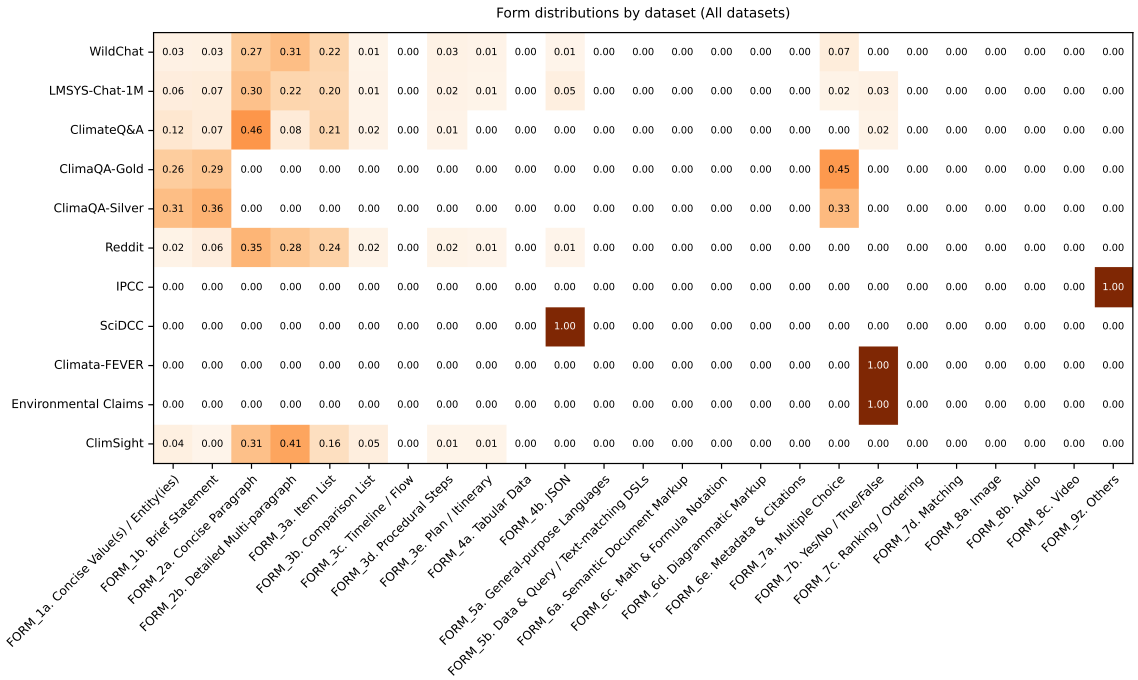

Figure 19: Form Distribution for All 11 Datasets.

		1–50			51–100			101–150			Overall					
Annoator	**Metric**	Topic	Intent	Form	Topic	Intent	Form	Topic	Intent	Form	Topic	CI	Intent	CI	Form	CI
A	Jaccard	0.710	0.737	0.737	0.780	0.855	0.810	0.713	0.687	0.727	0.734	[0.671, 0.789]	0.759	[0.701, 0.813]	0.757	[0.696, 0.814]
A	Micro-F1	0.739	0.739	0.713	0.787	0.864	0.821	0.735	0.712	0.707	0.755	[0.699, 0.806]	0.772	[0.716, 0.822]	0.745	[0.684, 0.804]
B	Jaccard	0.737	0.727	0.807	0.655	0.717	0.830	0.643	0.737	0.790	0.678	[0.620, 0.743]	0.727	[0.669, 0.786]	0.809	[0.756, 0.859]
B	Micro-F1	0.754	0.739	0.796	0.688	0.719	0.844	0.632	0.737	0.811	0.691	[0.634, 0.752]	0.732	[0.674, 0.789]	0.817	[0.764, 0.864]
Agr(A,B)	Cohen's κ	0.596	0.646	0.661	0.630	0.701	0.797	0.555	0.632	0.748	0.605	–	0.663	–	0.745	–

Table 14: Segment-wise and overall performance (mean with 95% confidence intervals).

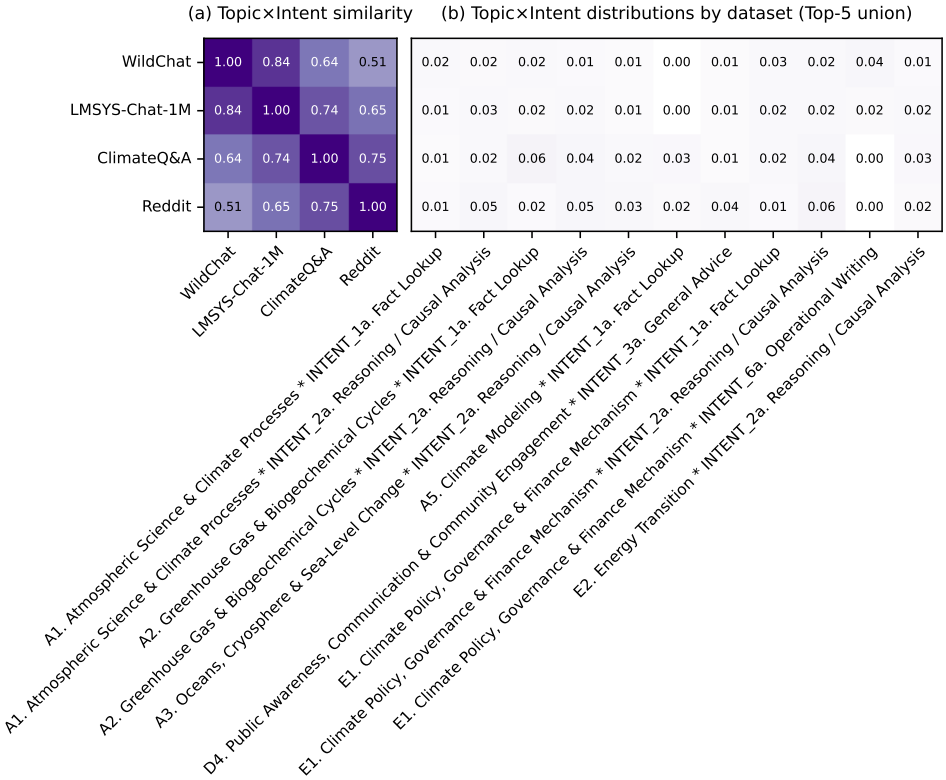

Figure 20: Topic×Intent cross-analysis.

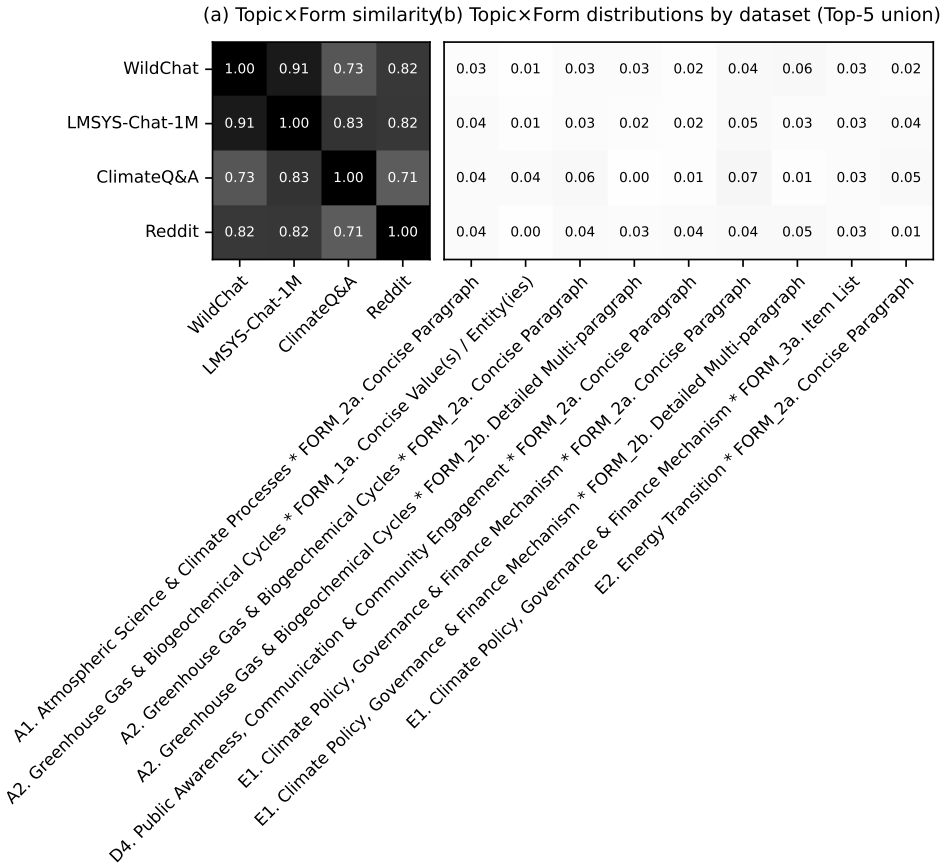

Figure 21: Topic×Form cross-analysis.

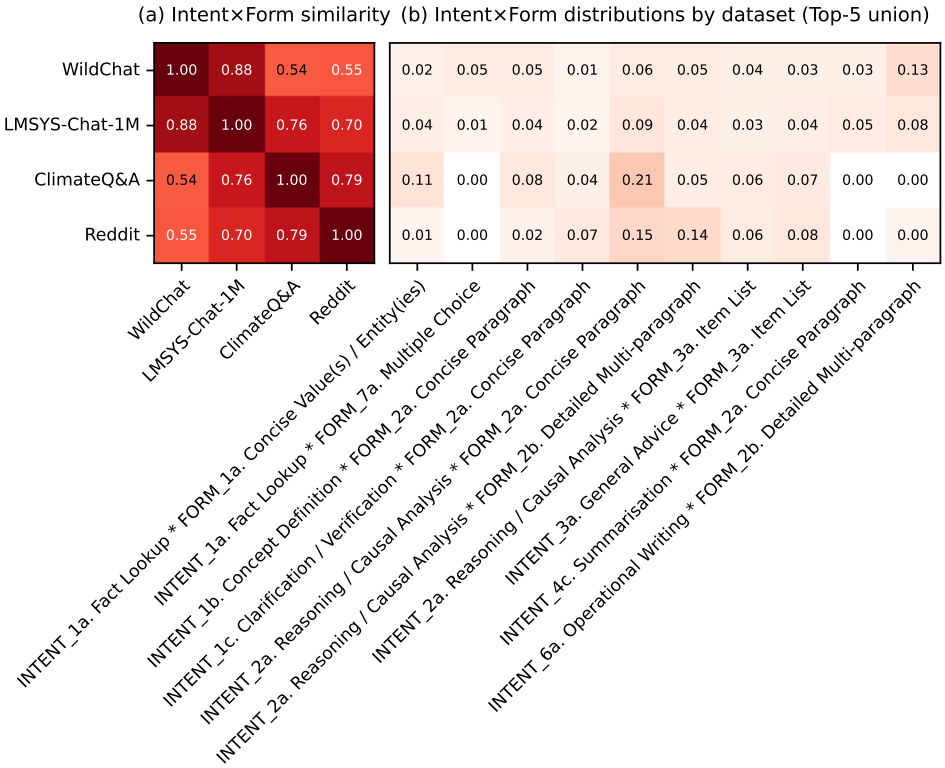

Figure 22: Intent×Form cross-analysis.

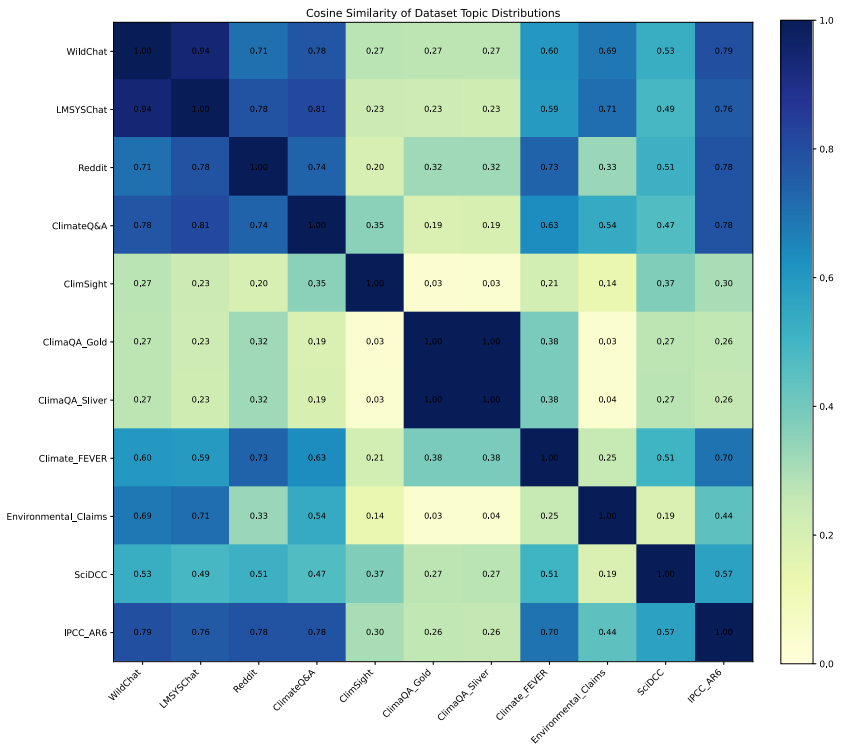

Figure 23: Topic Similarity for All 11 Datasets under S1 Merge Setting.

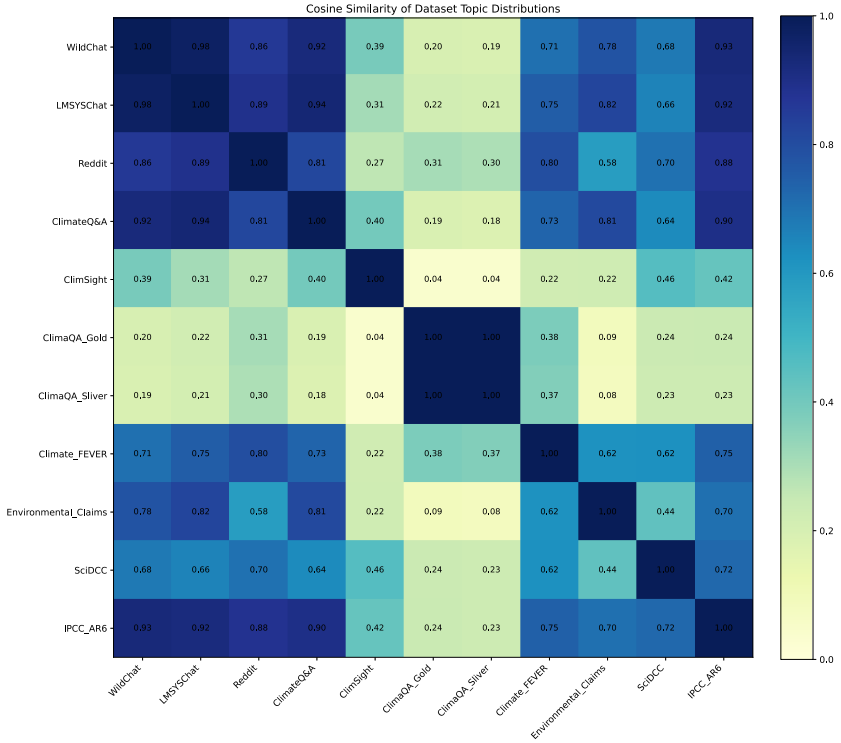

Figure 24: Topic Similarity for All 11 Datasets under S2 Merge Setting.

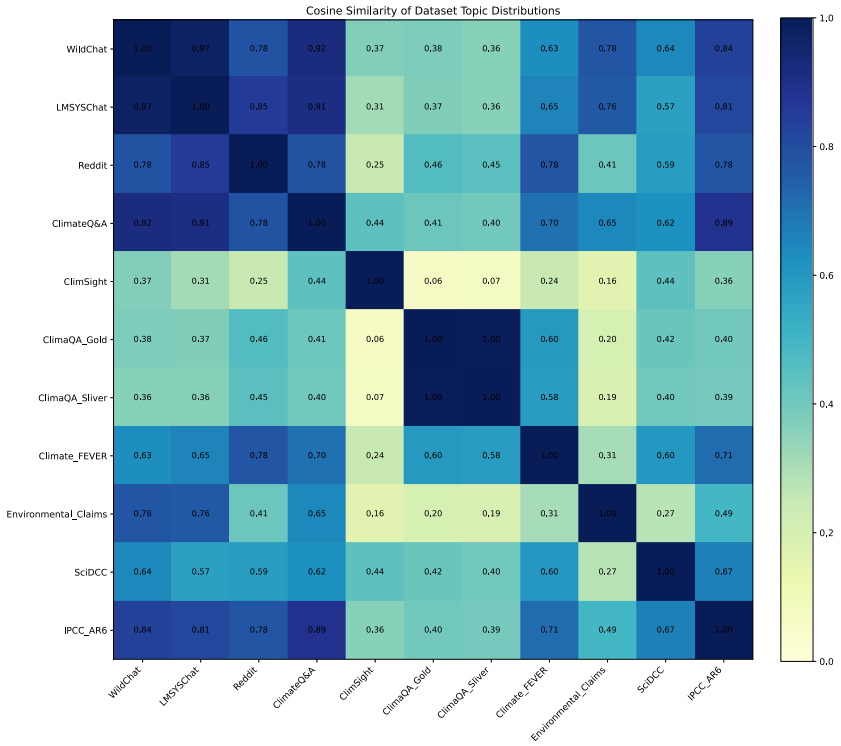

Figure 25: Topic Similarity for All 11 Datasets under S3 Merge Setting.

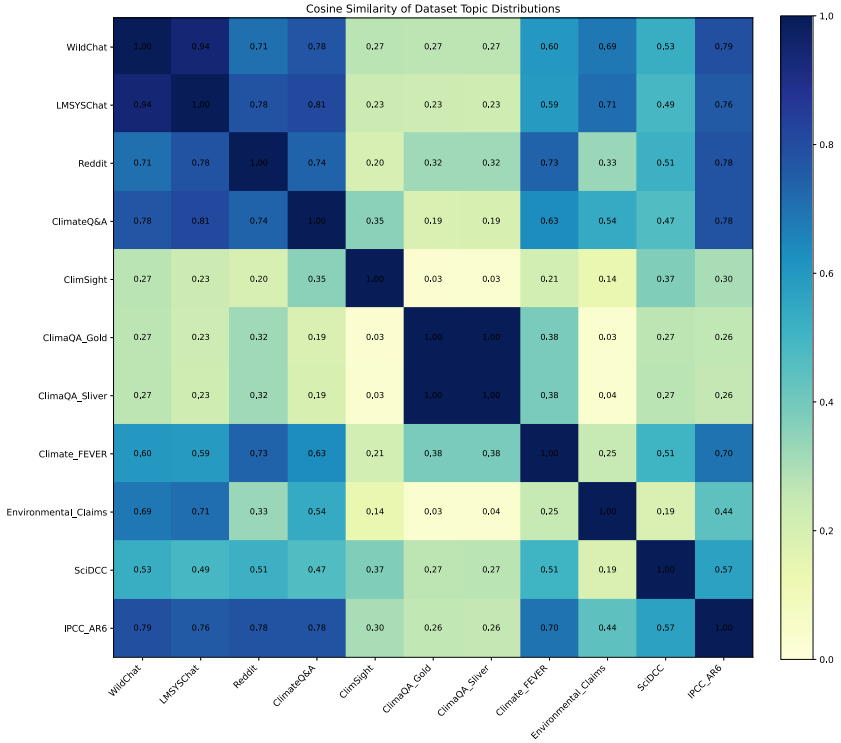

Figure 26: Topic Similarity for All 11 Datasets under S4 Merge Setting.

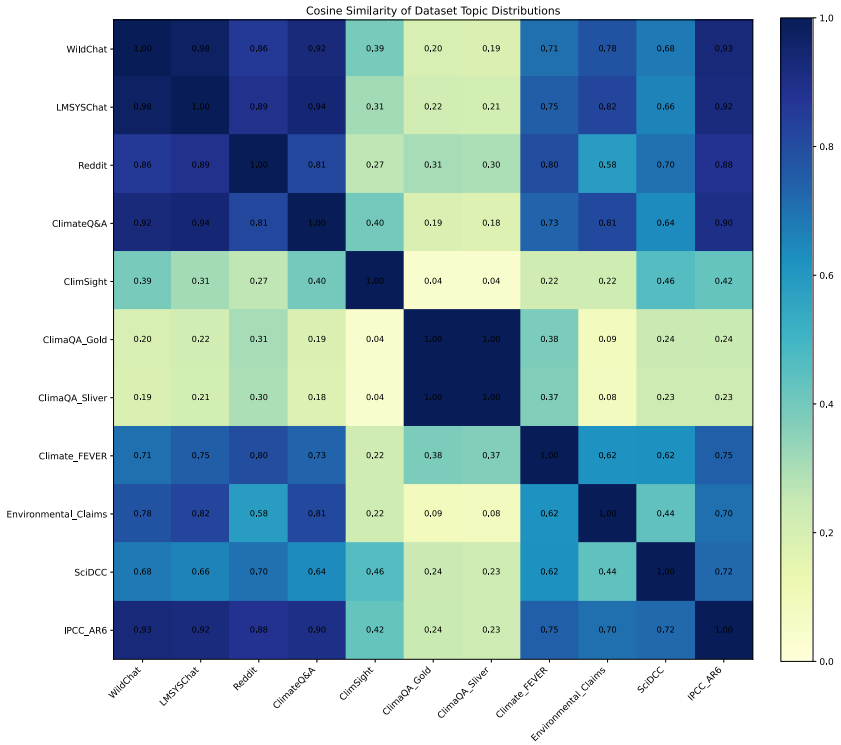

Figure 27: Topic Similarity for All 11 Datasets under S5 Merge Setting.

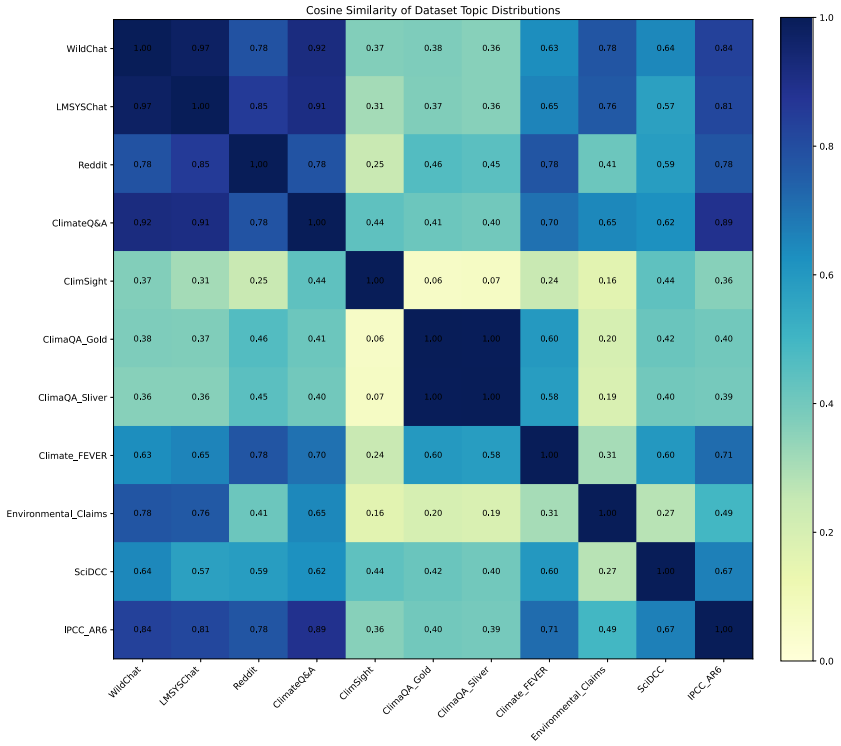

Figure 28: Topic Similarity for All 11 Datasets under S6 Merge Setting.

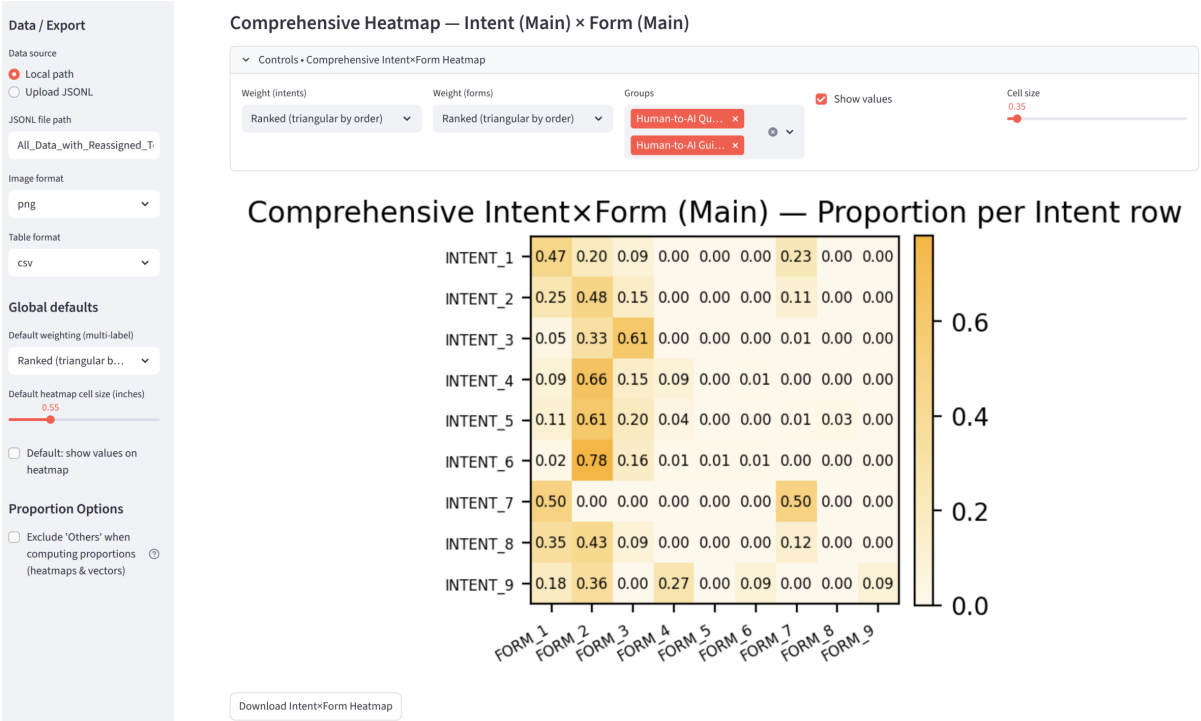

Figure 29: Interactive data visualization enabling flexible data selection, sample weighting, and both multidimensional and cross-dimensional analysis. Shown here is a partial cross-analysis between Intent and Form.

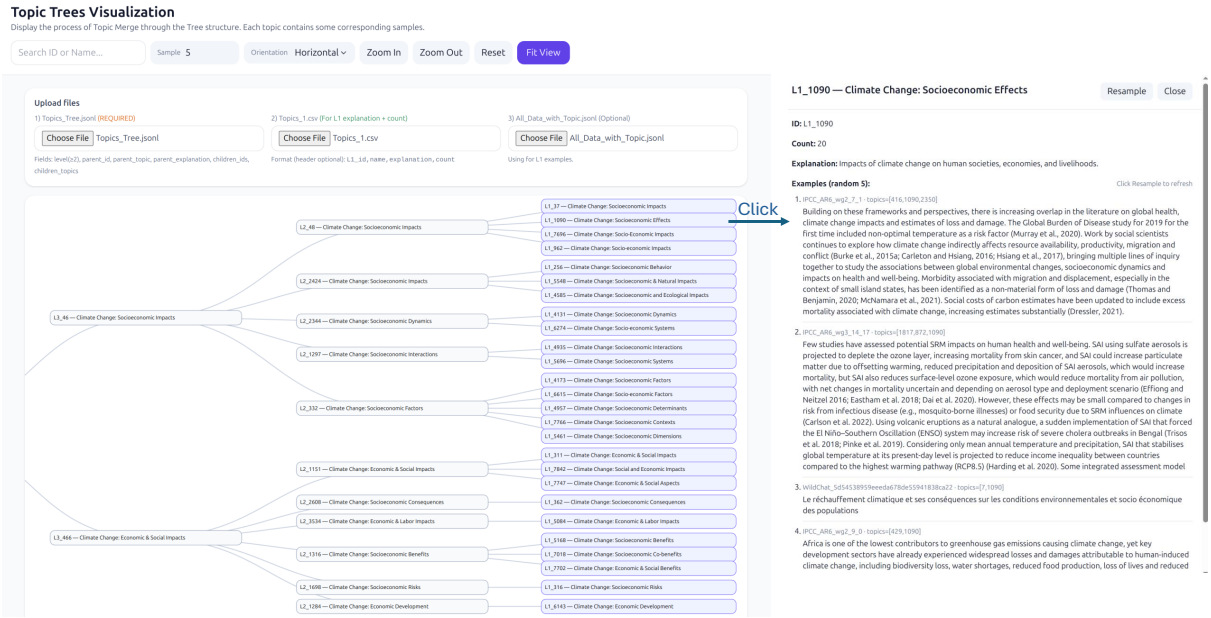

Figure 30: Visualization of the topic-merging process, with interactive access to each topic's explanation, sample count, and randomly sampled instances. Leaf nodes are displayed in purple, while other nodes are shown in gray.

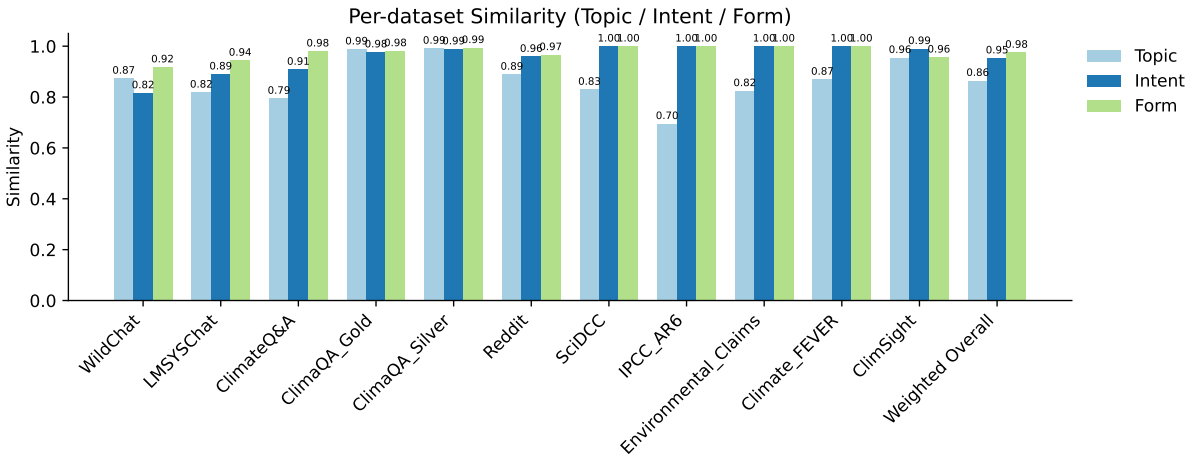

Figure 31: Similarity between the sampled validation subset and the full dataset.