Analyzing Human Questioning Behavior and Causal Curiosity through Natural Queries

Roberto Ceraolo[1,*] **Dmitrii Kharlapenko**[2,*] **Ahmad Khan**[2] **Amélie Reymond**[3]
Rada Mihalcea[4] **Bernhard Schölkopf**[5] **Mrinmaya Sachan**[2] **Zhijing Jin**[2,5]
[1]EPFL [2]ETH Zürich [3]UW [4]University of Michigan [5]MPI for Intelligent Systems
ceraolo.rc@gmail.com dkharlapenko@ethz.ch jinzhi@ethz.ch

Abstract

The recent development of Large Language Models (LLMs) has changed our role in interacting with them. Instead of primarily testing these models with questions we already know the answers to, we now use them to explore questions where the answers are unknown to us. This shift, which hasn't been fully addressed in existing datasets, highlights the growing need to understand naturally occurring human questions—those that are more complex, open-ended, and reflective of real-world needs. To this end, we present *NatQuest*, a collection of 13,500 naturally occurring questions from three diverse sources: human-to-search-engine queries, human-to-human interactions, and human-to-LLM conversations. Our comprehensive collection enables a rich understanding of human curiosity across various domains and contexts. Our analysis reveals a significant presence of causal questions (up to 42%) within the dataset, for which we develop an iterative prompt improvement framework to identify all causal queries and examine their unique linguistic properties, cognitive complexity, and source distribution. We also lay the groundwork to explore LLM as a router for these questions and provide six efficient classification models to identify causal questions at scale for future work. [1]

1 Introduction

The rapid advancement of Large Language Models (LLMs) (Brown et al., 2020; Chowdhery et al., 2023) is reshaping our interactions with these technologies (Burns et al., 2024). Instead of primarily acting as "testers"—posing questions to which we *already* know the answers—we are increasingly becoming "inquirers," asking questions that reflect *genuine* curiosity. As contrasted in Figure 1, existing NLP datasets often feature test questions such as "If John hits Tim, will Tim be angry?", but many recent queries to LLMs explore complex topics, such as "What are the causes of economic growth?". This shift highlights that while test questions are typically easier and designed to evaluate models, natural inquiries are often more *challenging*, *open-ended*, and aligned with *real-world needs*, driven by *pure curiosity* (Coenen et al., 2019; Rothe et al., 2018; Gottlieb et al., 2013).

This transition, along with the widespread use of LLMs as chatbots, necessitates a deeper understanding of how humans naturally pose questions (Ouyang et al., 2023). From a computational social science perspective (Biester et al., 2024; Porter et al., 2016), it is intriguing to analyze the linguistic characteristics of these questions, the underlying human needs they express, and the user personality traits they reveal. Additionally, gaining better insight into the topics covered by these questions can aid in developing domain-specific methods to improve them. However, existing questions in popular NLP datasets, as illustrated in Figure 1, are predominantly *test* questions designed for tasks such as reading comprehension (Rajpurkar et al., 2016; Tandon et al., 2019), commonsense reasoning (Roemmele et al., 2011; Sap et al., 2019; Bondarenko et al., 2022), and formal causal inference (Jin et al., 2024, 2023). While datasets like WildChat (Zhao et al., 2024) include natural queries from human-to-LLM interactions, they unfortunately lack diversity and coverage of other facets of human curiosity.

To bridge this gap, we present *NatQuest*, a comprehensive dataset of 13,500 naturally occurring questions derived from three diverse channels: human-to-search-engine queries (H-to-SE) from Google (Kwiatkowski et al., 2019) and Bing (Nguyen et al., 2016), human-to-human interactions (H-to-H) from Quora (Iyer et al., 2017), and human-to-

Equal contribution. Research done while Roberto was a research intern at ETH.

[1]Our code and data are at https://github.com/roberto-ceraolo/natquest.

Dataset Type	Example Question	Data Nature
Reading Comprehension	[Given a passage] In the story, why did Tim quit?	Testing LLMs against existing knowledge of humans
Commonsense Causality	If John hits Tim, will Tim be angry?	Testing LLMs against existing knowledge of humans
Formal Causal Inference	If A correlates with B, can we say A causes B?	Testing LLMs against existing knowledge of humans
Our Dataset	**What are causes of economic growth?** (and many other open-domain natural questions)	**Studying human natural queries & paving the way for human-AI collaboration to answer them**

Figure 1: Our *NatQuest* dataset propose a different paradigm to ask open-ended natural human inquiries, in contrast to previous tasks (e.g., reading comprehension, commonsense causality, and formal causal inference) which craft questions restricted to ones that humans already understand well and purposefully test LLMs with.

LLM interactions (H-to-LLM) from ShareGPT and WildChat (Zhao et al., 2024). This multi-faceted approach ensures a rich tapestry of questions that reflect genuine human curiosity across various domains and contexts.

Our resulting dataset encompasses a wide variety of topics, and different cognitive complexities, from simple factual queries (e.g., "How high is Mountain Everest?") to complex causal investigations (e.g., "What are the causes of economic growth?"). We find that natural inquiries are 38% more open-ended than existing curated datasets, and also more equally cover a comprehensive range of six levels of cognitive complexity (Bloom et al., 1964). Further, we also find that people use various media to post their questions based on different needs, where H-to-SE queries are usually for knowledge and information needs, and H-to-LLM questions cover more needs for problem-solving and leisure.

Furthermore, we identify an important phenomenon, causal inquiries (Pearl, 2009a; Peters et al., 2017), within the natural questions. Our data analysis reveals that up to 42% of the questions are related to causality. These questions are particularly intriguing as they not only concern existing knowledge but often relate to future actions, predictions, and decision-making, rendering them especially meaningful for humans (Pearl, 2019; Sloman and Lagnado, 2015). To systematically study these queries, we develop an iterative prompt improvement method for identifying and categorizing causal questions. Based on our categorization, we analyze the distinct linguistic properties, cognitive complexity, and source distributions of causal questions. Furthermore, we lay the groundwork for future research by investigating the performance of current LLMs on our question set and building efficient causal question classifiers as routers to differentiate between casual and non-casual questions, directing them to an enhanced reasoning pipeline, as the literature demonstrates the efficiency of routing techniques to enhance performance (Chen et al., 2023).

In summary, our contributions are as follows:

1. We present *NatQuest*, a latest collection of 13,500 natural human queries across three diverse sources.
2. We identify the differences between natural questions in our *NatQuest* dataset and existing curated NLP datasets in terms of open-endedness, cognitive complexities, user needs, and knowledge domains.
3. We further explore an important phenomenon in natural questions—causal inquiries—and analyze its distinct linguistic properties, cognitive complexity, and distribution across different sources.
4. To identify causal questions across *NatQuest*, we present an iterative prompt improvement framework combined with a limited set of human expert labels to scale up causal question labeling in our large dataset.
5. We provide preliminary studies on LLM response analysis and efficient causal question classification using 6 non-neural-network or small language models.

2 Exploring Human Natural Queries

2.1 Natural Question Sources

Table 1: Our *NatQuest* equally covers questions from the three source types: human-to-search-engine queries (H-to-SE), human-to-human interactions (H-to-H), and human-to-LLM interactions (H-to-LLM).

Nature of Data	Dataset	# Samples
H-to-SE (33.3%)	MSMarco (2016)	2,250
	NaturalQuestions (2019)	2,250
H-to-H (33.3%)	Quora Question Pairs	4,500
H-to-LLM (33.3%)	ShareGPT	2,250
	WildChat (2024)	2,250

To compose our *NatQuest* dataset, we identify three different channels where people ask questions online: people enter queries on search engines (H-to-SE), post questions to other people on question platforms (H-to-H), and chat with LLMs hosted on web interfaces such as ChatGPT (H-to-LLM). To represent H-to-SE data, we obtain search queries on Google from the NaturalQuestions dataset (Kwiatkowski et al., 2019), and on Bing Search from the MSMarco dataset (Nguyen et al., 2016). For the H-to-H data, we adopt the Quora Question Pairs dataset (Iyer et al., 2017),[2] which compiles actual question data from the widely-used question-answering website Quora. Lastly, we also incorporate several sources of H-to-LLM queries, from ShareGPT,[3] a collection of users' voluntarily-shared queries to ChatGPT, and the WildChat collection of user-LLM conversations through their chat interface powered by ChatGPT and GPT-4 APIs (Zhao et al., 2024).

To preprocess the data, we filter out empty queries, non-English questions, and invalid characters. For LLM conversations, we select the first question and cut off follow-up questions. For long questions with more than 30 words, we generate shorter versions using GPT3.5-turbo-0125 to condense the main idea of the questions. As shown in Table 1, we sample 4,500 questions from each of the three channels, and if a channel has multiple sources, we also distribute the questions evenly across the sources. This results in a total of 13,500 questions for our *NatQuest* dataset, which can be accessed at `https://huggingface.co/datasets/causal-nlp/NatQuest`.

2.2 Data Statistics

Overall. We describe the statistics of our dataset in Table 2. Overall, our dataset has 13,500 total samples, each with a length of 11.34 words on average. The entire vocabulary size of our data is 25.7K unique words, with a type-token ratio (TTR) of 0.168.

Topic Coverage. To explore the topics covered in our dataset, we perform K-means clustering (Hartigan and Wong, 1979) on the embedding space of all the queries, using the text embedding model *text-embedding-3-small* from OpenAI. Figure 2 uses t-SNE (Van der Maaten and Hinton, 2008) to visualize the five main clusters, covering daily life questions; computer-related inquiries; sports, medicine and science; prompt questions; and story generation.

[2] `https://quoradata.quora.com/First-Quora-Dataset-Release-Question-Pairs`

[3] `https://huggingface.co/datasets/anon8231489123/ShareGPT_Vicuna_unfiltered`

Table 2: Overall statistics of our *NatQuest* dataset, including the number of samples (# Samples), average number of words per sample (# Words/Sample), vocabulary size (Vocab) by the number of unique words, and Type-Token Ratio (TTR).

	# Samples	# Words/Sample	Vocab	TTR
NatQuest	13,500	11.34	25,709	0.168

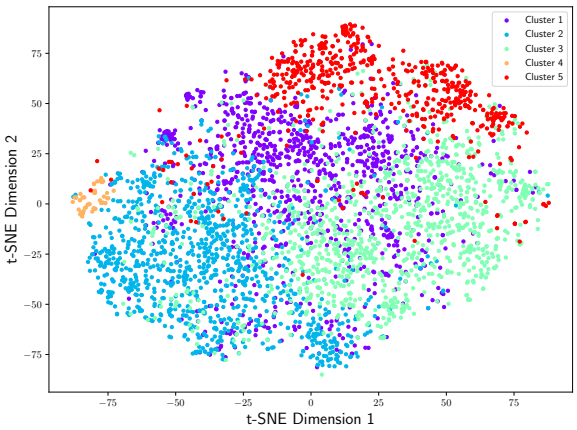

Figure 2: T-SNE visualisation of the main topic clusters in our *NatQuest* dataset. Cluster 1: Daily life. Cluster 2: Computer-related. Cluster 3: Sports, medicine, and science. Cluster 4: Prompt Questions. Cluster 5: Stories and fictional characters. See more detailed information about the clusters in Appendix A.1.

2.3 How Do Natural Inquiries Differ from Curated Test Questions?

Our dataset enables a rich set of explorations on human natural questioning behavior. Specifically, we will explore two research questions: what are the properties distinguishing natural inquiries from curated test questions (in this section), and how does human behavior vary across different platforms (in Section 2.4)?

To compare natural questions in *NatQuest* with non-natural ones, we collect a comparison set consisting of six curated test sets. We include a diverse range of curated test sets covering reading comprehension tasks (Rajpurkar et al., 2018), multiple-choice science questions (Clark et al., 2018), truthful question answering (Lin et al., 2022), grade school math word problems (Cobbe et al., 2021), and questions requiring complex reasoning (Wang et al., 2024)

and domain experts (Rein et al., 2023). We sample 500 questions from each of them, to build a representative sample of non-natural questions.

Methods. We compare natural inquiries and curated tests in terms of cognitive complexity and open-endedness. For cognitive complexity, we adopt the six levels of cognitive abilities in Bloom's Taxonomy (Anderson and Krathwohl, 2001): starting with the simplest skill, remembering, and advancing to understanding, applying, and all the way to creating. For open-endedness, we evaluate whether the question permits semantically different answers (i.e., open-ended) or only has one unique answer (i.e., not open-ended). Both evaluations are implemented using LLMs (GPT-4o-mini with prompts in Appendix B), following the use of LLMs as judge (Zheng et al., 2023). We further verify the quality of the LLM annotations by calculating its agreement with human annotation on a small set of 150 samples, where we obtain an F1 score of 79%, which is relatively reasonable.

Results. Comparing the non-natural questions and our *NatQuest* in Table 3, we find that natural questions tend to be much more open-ended, with 68% questions allowing semantically different answers. Moreover, natural questions often require a more even distribution of all the six levels of cognitive skills, with each skill's portion closer to 13%, and a higher portion of advanced skills such as evaluating and creating.

2.4 Does Human Question Behavior Vary across Platforms?

We are further interested in inferring human behavior from the questions, and how it varies across the three channels of inquiries: H-to-SE, H-to-H, and H-to-LLMs. We start our analysis by inferring user needs and intent from the questions, and compare them across the three platforms. We then analyze the differences in cognitive complexity and knowledge domains of the questions across the platforms.

2.4.1 Inferred User Needs

Inspired by existing frameworks of human needs (Maslow, 1943) and information seeking behavior (Wilson, 1999; Kuhlthau, 2005), we introduce five types of needs relevant for *NatQuest*: "Knowledge and Information," "Problem-Solving and Practical Skills," "Personal Well-being," "Professional and Social Development," and "Leisure and Creativity." To label the needs following our taxonomy, we annotate a small set, and then use LLMs to iteratively improve the prompt, before scaling up LLM labels to the entire set, with details in Appendix B.

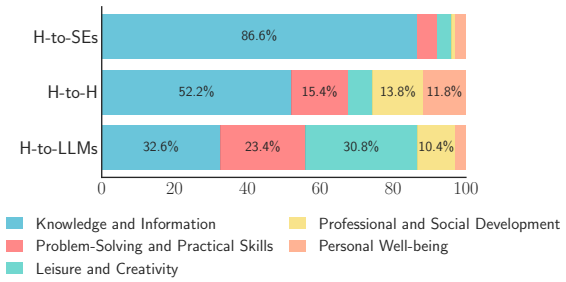

Figure 3: User needs across sources.

We plot in Figure 3 the distribution of user needs inferred from questions across the three platforms. Notably, we find that the basic need of "Knowledge and Information" is dominant in H-to-SE queries, but much less in others, e.g., as low as 32.6% in H-to-LLMs interactions. In contrast, LLMs are used more often to address the needs of "Leisure and Creativity" (30.8%) and "Problem-Solving and Practical Skills" (23.4%). These results show a shift in user expectations from AI systems in contrast to search engines from simple factual queries to more creative and interactive problem-solving. Future work can conduct additional longitudinal studies to investigate whether the advancement of LLMs has a causal effect to change user's choice of the medium to ask questions.

2.4.2 Cognitive Complexity

We show the three platforms' distribution across the six levels of cognitive complexity in Bloom's Taxonomy (Anderson and Krathwohl, 2001) in Table 4. We find that in H-to-SEs, most questions (76.49%) fall under the Remembering category, indicating that users primarily seek factual information retrieval from search engines, which resonates with the dominant "Knowledge and Information" need in Section 2.4.1. In contrast, among H-to-H interactions, the majority of questions are "Evaluating," – requiring subjective judgments and nuanced understanding, which explains our dataset's open-endedness illustrated earlier in Section 2.3. In H-to-LLMs, the largest category is Creating (38.20%), suggesting that users frequently request generative and novel content from LLMs.

Table 3: Comparison of non-natural questions in curated tests and our *NatQuest* in terms of open-endedness and cognitive complexity.

Category	Non-Natural Questions	*NatQuest*
Open-Endedness		
Open-Ended	30%	68%
Cognitive Complexity		
Remembering (least complex)	30.51%	36.82%
Understanding	7.47%	13.47%
Applying	36.41%	13.54%
Analyzing	13.90%	8.8%
Evaluating	11.54%	13.74%
Creating (most complex)	0.17%	13.62%

Table 4: Distribution of cognitive complexity and domain classifications for different sources (H-to-SEs, H-to-H, H-to-LLMs).

Category	H-to-SE	H-to-H	H-to-LLMs
Cognitive Complexity			
Remembering	76.49%	19.42%	14.52%
Understanding	13.82%	15.27%	11.32%
Applying	4.07%	18.73%	17.81%
Analyzing	3.40%	13.00%	10.05%
Evaluating	1.87%	31.27%	8.09%
Creating	0.36%	2.31%	38.20%
Knowledge Domain			
Arts and Culture	13.89%	5.30%	12.54%
Computer Science	6.80%	17.87%	45.11%
Everyday Life & Personal Choices	9.63%	21.57%	11.91%
Health and Medicine	27.99%	10.94%	2.71%
Historical Events & Hypothetical Scenarios	10.04%	3.22%	3.44%
Natural & Formal Sciences	13.39%	7.91%	4.84%
Psychology & Behavior	0.71%	11.47%	3.82%
Society, Economy & Business	17.55%	21.73%	15.63%

2.4.3 Knowledge Domain

We use LLMs to identify the main knowledge categories in *NatQuest* in Table 4. The result suggests that humans predominantly use LLMs for computer science-related questions (45.11%), while search engines receive more inquiries related to health and medicine (27.99%). On the other hand, humans ask each other a more balanced variety of questions, with "Society, Economy, Business" (21.73%) and "Everyday Life and Personal Choices" (21.57%) being the most frequent. This indicates that users (currently) rely on LLMs for technical queries, on search engines for health-related information, and on human interactions for broader, context-rich discussions.

3 Identifying Causal Inquiries

As motivated in the introduction, an interesting phenomenon in human queries is their causality-seeking behavior. In this section, we will introduce how we identify causal inquiries among all the natural questions, their distinct features, and correlations with the other behaviors.

3.1 Formal Definition of Causality

Rooted in philosophy (Beebee et al., 2009; Russell, 2004; Kant, 1781), causality has evolved into a rigorous statistical field (Fisher and Ford, 1927; Rubin, 1980; Spirtes et al., 1993; Pearl, 2009b). To reason about the causal relationships among variables, we introduce a set of formal terminologies from a coarse-grained level (through *causal graphs*) (Spirtes et al., 2000; Zhang and Hyvärinen, 2009) to a fine-grained level (through *structural causal models (SCMs)*) (Pearl, 2009b; Peters et al., 2017). We give an overview of example questions in Table 5, with details of our taxonomy of the four causal types below.

On a coarse-grained level, we can represent all the causal relations among variables via a causal graph $\mathcal{G} := (\boldsymbol{V}, \boldsymbol{E})$. The causal graph $\mathcal{G}$ consists of a set of n variables $\boldsymbol{V} := (X_1, \ldots, X_n)$, and the set of edges $\boldsymbol{E}$ which consist of each directed causal relation e_{ij} if X_i directly causes X_j, also noted as $X_i \to X_j$.

The first two types of causal questions are with regard to the causal graph. **Type 1** asks about **variables $\boldsymbol{V}$** in the causal graph. For example, "what are the causes of fire" enquires the cause variables that lead to fire, namely fuel, heat, and an oxidizer. **Type 2** causal questions ask about the **existence of directed edges** among variables, such as whether a lack of experience (variable X_1) leads to the application failure (variable X_2), which queries the existence of the $e_{12} : X_1 \to X_2$ edge.

Moving to a fine-grained level, we introduce the other two types of causal questions using the SCM

Table 5: Example questions in the four causality types.

Causal Question Type with Examples
Type 1. About Variables
What are the causes of fire?
What nutrition do athletes need?
Type 2. About Relations
Does smoking interfere with the drug effect?
Was my application rejected due to my lack of work experience?
Type 3. About Average Effects
How much do COVID vaccines decrease hospitalization risk?
Are small or big classrooms better for kids?
Type 4. About Mechanisms
Had I not done a PhD, would my life be different?
How do scientists prepare rockets for missions to the Moon?

framework (Pearl, 2009b; Bareinboim et al., 2022). If a directed edge $e_{ij} : X_i \to X_j$ exists, then **Type 3** questions ask about a more quantitative relation in the form of causal effects, such as "On average, how much do COVID vaccines decrease hospitalization risk?". Here, the query concerns a **quantification of the *average* change** in the effect (i.e., hospitalization risk) given the cause (i.e., vaccination). Most questions of Type 3 are formulated as average treatment effect (ATE), which is $\mathbb{E}[Y|\,\mathrm{do}(X)]$, i.e. evaluating the expected changes in the effect variable Y when replacing the cause variable X with a constant, keeping all the rest unchanged.

Lastly, **Type 4** questions can ask about the **functions/mechanisms** among the variables, or anything that can only be answered based on the functions/mechanisms. Such examples include counterfactual questions in the Ladder of Causation (Pearl, 2009a) such as "Had I not done a PhD, would my life be different?", or Partial Differential Equations (PDEs) as in dynamic systems. An SCM is usually presented as $\mathcal{S} := \{X_i := f_i(\mathbf{Pa}_i, U_i)\}_{i=1}^n$, where $\mathbf{Pa}_i \subseteq \boldsymbol{V} \setminus \{X_i\}$ are the direct causes, or *causal parents*, of X_i, f_i are deterministic functions, and U_i are exogenous variables which capture possible stochasticity of the process, and the uncertainty due to unmeasured parts of the system.

3.2 Classifying Causal Questions

Human Annotation of Causal Questions Given the formal formulation of causality, we engaged two expert annotators with extensive knowledge in both causality and NLP to annotate a sample of $N = 500$ data points. They were briefed on the nature of the data and the potential presence of sensitive content. Additionally, the annotators represent diverse genders and cultural backgrounds. We develop the following iterative improvement process: Using an annotation guideline based on our definitions in Section 3.1, annotators independently label 500 data points. Then, we check the inter-annotator agreement rate, achieving a Cohen's κ (Cohen, 1960) of 0.66, indicating moderate consensus among the human labelers (McHugh, 2012; Landis and Koch, 1977; Viera et al., 2005). The annotators then further refine the labels by analyzing the disagreement cases and either: (1) In cases where the initial textual guideline does not communicate the mathematical formulation clearly, we improve the guidelines, and the annotators agree on a clear classification choice. (2) In cases where the two annotators interpret the data sample differently, they have a discussion and agree on one correct label for each question. The agreed-upon label then becomes the ground truth. We identify 238 out of 500 annotated samples as causal questions.

Scaled Labeling by Iterative Prompt Improvement Using the human annotation set, we develop an iterative prompt improvement process on the entire dataset. We start the process by first providing our definitions of causal questions. After this initial annotation by GPT-4 on the sample set, we calculate the classification performance of the LLM with regard to our ground-truth labels, and inspect the error cases manually. Then, we improve the prompt by the canonical prompt engineering techniques including in-context learning (Brown et al., 2020) and chain-of-thought prompting (Wei et al., 2022)), and by clarifying the decision principles around the error cases. We repeat this iterative improvement process until reaching a high classification performance.[4] The final performance of GPT-4 reaches 88.7% F1 scores with regard to the ground-truth causal question labels in the 500 sample set, 89.4% accuracy, 92.9% precision, and 84.9% recall. On a held-out set of 100 additional annotated samples not belonging to the previous set, we observe a good performance of 88% F1 score.

3.3 How Do Causal Questions Differ from Non-Causal Ones?

Based on our scaled labels of causal questions, we investigate two research questions: how do causal questions differ from non-causal questions (in this section), and how are causal questions distributed

[4]We report in Appendix B.8 some examples of early-iteration vs. late-iteration prompts.

(Section 3.4)?

3.3.1 Linguistic Differences

Method. We first explore the differences between causal and non-causal queries by linguistic features. First, we calculate the frequency of different question words in causal and non-causal questions. We further follow Girju and Moldovan (2002) to identify the presence of morphological causatives (verbs ending in *-en* or *-ify*), lexico syntactic patterns (presence of causative verb phrases such as *lead to, induce, result in*), and matching rules by Bondarenko et al. (2022).

Table 6: Distribution of question words in our dataset.

Question Words	Causal	Non-Causal	Overall
How	**2,077**	636	2,713
Why	**635**	12	647
What	1,383	**2,324**	3,707
Who	144	**723**	867
Where	124	**465**	589

Table 7: Distribution of causal indicators in questions, including morphological causative, lexico-syntactic pattern, and causative rules.

	Causal	Non-Causal	Overall
Morphological	**1.65%**	0.56%	1.02%
Lexico-Syntactic	**11.52%**	6.83%	8.81%
Causative Rules	**15.21%**	0.40%	6.65%

Results. As shown in Table 6, causal questions are more frequently led by question words such as "Why" and "How," whereas non-causal questions often start with "What," "Who" and "Where." In terms of other linguistic indicators in Table 7, causal questions have more morphological causatives and lexico-syntactic patterns, as well as matches with more causative rules. However, we also acknowledge that the, three linguistic methods capture a limited set of causal questions, which indicate the linguistic diversity and the richness of our *NatQuest* dataset.

3.3.2 Cognitive Complexity

Another distinction of causal questions is that it requires higher-level cognitive capabilities. We present the cognitive skill distribution across causal vs. non-causal questions in Figure 4, which shows that non-causal questions rely more on lower cognitive requirements like Remembering, whereas causal questions tend to require higher cognitive skills like Applying, Analyzing, and Evaluating.

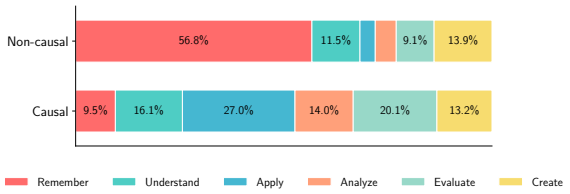

Figure 4: Cognitive skill distribution in causal vs. non-causal questions.

3.4 How Are Causal Questions Distributed?

Connecting causal queries with our previous findings, we analyze how causal questions are distributed across natural vs. curated questions, and across the three different platforms.

Causality in Natural vs. Curated Questions Comparing the percentage of causal questions across curated test sets and our *NatQuest*, we find that our natural questions almost double the percentage of causal queries (42%) than that of non-natural questions (23%) in Table 8. This again highlights a contrast between natural queries made by humans and crafted tests. Our findings in reinforce previous work's insight on the divergence between user queries and NLP benchmarks (Ouyang et al., 2023).

Table 8: Distribution of causal questions across natural and non-natural questions, as well as the three platforms.

	Causal	Non-Causal
Non-Natural	23%	77%
NatQuest	42%	58%
H-to-SE	22%	78%
H-to-H	59%	41%
H-to-LLMs	46%	54%

Causal Queries across Platforms Identifying causality in questions across the three platforms in Table 8, we find that H-to-H has the largest proportion of causal questions at 59%, followed by H-to-LLMs at 46%, and finally, H-to-SE at 22%. This pattern suggests that when seeking to understand causes and effects, internet users view LLMs as more aligned with human responses compared to search engines. In the future, there might be an increasing transition towards conversational systems for information retrieval and understanding of the world (Zhou and Li, 2024).

4 Preparations for Future Research

4.1 Evaluating LLM Responses to Natural Questions

While our study presents an extensive investigation of human questioning behavior, a natural next step is to evaluate the question-answering behavior of LLMs on these natural questions. To this end, we provide some preliminary work in evaluating the performance of the latest LLM, GPT-4o, based on three usability criteria commonly used in user satisfaction surveys (ISO, 2018): effectiveness, efficiency, and satisfaction. On a scale of 1–5, we annotate the answer quality of 50 random causal questions, and find that GPT-4o obtains an average effectiveness of 3.83, efficiency of 2.88, and satisfaction of 3.85. We find it struggles when asked to foresee the causal effects in the future, but does well in answering causal questions that require knowledge lookup. Its answers are often overly verbose, explaining its limited score in efficiency. More details are in Appendix A.5.

4.2 Building Efficient Causal Question Routers

As shown earlier in the study, causal questions has its own unique nature, and occupy a non-trivial percentage of 42% of natural questions. Moreover, our preliminary study above shows that the LLM performance on causal questions is quite limited. Given this motivation, we imagine an emerging future direction is to classify causal questions beforehand, and potentially route them to some specific, reasoning-enhanced solution pipeline.

To efficiently identify causal questions, we explore a set of 7 smaller, more efficient models in Table 9, trained or fine-tuned on *NatQuest* to classify causal questions from non-causal ones, aiming to understand this task's tradeoff between model size and accuracy as such classification models have proven effective for router construction (Ong et al., 2024; Ding et al., 2024).

We see that the largest model, FLAN-T5-XL (LoRA), performs best. However, a much smaller model, FLAN-T5-Small, also provides a good compromise with a small drop in accuracy but significant computation savings. Details on our experiments and further analysis can be found in Appendix A.6. This work provides a starting point for future work to classify causal questions, and potentially build dedicated causal reasoning modules.

Table 9: Performance of efficient models on causal question identification. For each model, we report its number of parameters (# Params), accuracy (Acc.), and the F1, precision (P.), and recall (R.) of the causal class.

	# Params	F1	Acc.	P.	R.
Rule-based classifier	–	58.7	23.3	62.8	55.2
TF-IDF + XGBoost	–	72.3	79.0	78.1	67.4
FLAN-T5-Small	80M	84.2	86.5	83.3	85.1
FLAN-T5-Base	250M	85.7	87.9	85.5	86.0
FLAN-T5-Large (LoRA)	780M	85.8	88.0	85.0	87.0
Phi-1.5 (LoRA)	1.5B	85.3	87.3	83.4	**87.2**
FLAN-T5-XL (LoRA)	2.85B	**87.7**	**89.2**	**91.1**	85.4

5 Related Work

Socio-Linguistic Analysis of Human Behavior Analysis Psychologists and linguists have long been interested in the study of questions, as they provide valuable insights into human cognition, emotion, and social dynamics. Traditionally, this task has been approached through methods such as manual content analysis (Graesser and Person, 1994), discourse analysis (Sinclair and Coulthard, 2013), and corpus linguistics (Biber, 2012). More recently, there has been a shift towards computational methods and large-scale data analysis, as evidenced by the development of tools like Linguistic Inquiry and Word Count (LIWC) (Tausczik and Pennebaker, 2010) and the application of natural language processing techniques (Boyd and Schwartz, 2021). Pennebaker et al. (2003) examine how natural language use, including questioning behavior, reflects personality traits and social processes. Jackson et al. (2022) explore how analyzing language, including questions, can advance psychological science by providing insights into cognitive and emotional processes. To the best of our knowledge, existing research falls short in inferring insights about human curiosity and information-seeking behavior from naturally occurring questions asked online.

Causal Question Datasets There is a growing research interest in causal reasoning of LLMs (Zhang et al., 2023; Kiciman et al., 2023; Zecevic et al., 2023; Jin et al., 2023, 2024). However, existing literature focuses on "test" questions, lacking coverage of natural "inquiries" and a comprehensive collection of natural causal questions. Indeed, while some previous studies focus on a specific type of causality (Jin et al., 2023; Tandon et al., 2019; Gusev and Tikhonov, 2022), others only use linguistic heuristics to label a question as "Causal"

(Lal et al., 2021a; Verberne et al., 2006, 2008; Lal et al., 2021b), and most of them primarily include artificially-generated data (Bondarenko et al., 2022; Mostafazadeh et al., 2020; Roemmele et al., 2011), rarely including sources of natural questions. Moreover, since most datasets were collected before the recent success of LLMs, none of them includes a new source of natural questions – causal questions directly asked to LLMs (Ouyang et al., 2023). Focusing on natural inquiries, we can perform several analyses of human behavior based on the language traces left online.

6 Conclusion

In conclusion, this study identifies the gap between previously curated NLP datasets and natural human questions. With a collection of 13.5K questions in *NatQuest*, we analyze the distinct features of natural questions and further explore the properties of causal inquiries. Our work presents an up-to-date question set in the era of LLMs, and paves the way for future model improvements on causal reasoning.

Limitations

Selection bias *NatQuest* is the first effort towards building a representative sample of causal questions humans ask online, to study human curiosity. Clearly, the data selection process is not devoid of bias. By *bias* here we imply that there might be a difference between the set of questions we gathered (*NatQuest*) and the full set we are trying to make an inference about (human curiosity as a whole) (Blodgett et al., 2020). Before analyzing the bias for the full set of questions humans ask, we wonder about a subset: how representative is *NatQuest* of the questions internet users ask? Each of the sources we used was gathered by researchers independently, so they underwent different filtering procedures, which might affect the distribution of questions. The following are potential sources of bias for each of the components of *NatQuest*:

1. ShareGPT: questions that yielded a denial from ChatGPT were excluded (e.g. "*I'm sorry but*")
2. NaturalQuestions: among Google queries, a subset of questions was selected with specific syntactical patterns, yielding a Wikipedia page among the top 5 Google results.
3. MSMARCO: only questions on Bing looking for a specific answer, and for which human judges could generate an answer based on some retrieved text passages were included.
4. WildChat: since the conversations were collected on Huggingface Spaces, most likely the average user was a developer, or someone involved in the AI community - and not the general internet population at large
5. Quora Question Pairs: the authors declare they used sanitation methods such as removal of questions with long question details

The above filtering procedures constitute a limitation of this work. In the future, the inclusion of more varied data sources can help reduce the bias coming from any single source.

Ideally, we would like to generalize such insights to humans, not just internet users. Can we consider the questions asked online as a representative sample of the full set of human questions? If an individual has a question, there might be several reasons why she does not ask it online and hence does not leave a trace. Those reasons act as confounders and generate a bias in the questions found online. For instance, users concerned about privacy might decide to avoid asking private questions on search engines or to LLMs. Future work could include more data sources or involve conducting surveys that aim to identify and categorize questions not typically found online, thus potentially decreasing the influence of these biases. Also, having filtered out all non-English queries, the insights mainly apply to the countries where people tend to use the Internet in English. Future work could focus on building a multilingual-*NatQuest* to improve its coverage.

Limitations of the Classifiers While our causal question classifiers show promise, they have several limitations. As *NatQuest* is the first dataset with causal question labels, external validation was not possible, highlighting the novelty of our work but also the need for additional labeled datasets in this domain. We focused primarily on binary classification without analyzing performance across different causal question subcategories (e.g., Types 1-4 as defined in Section 3.1). Additionally, we did not explore the impact of this classification on downstream tasks such as question answering or causal inference. Although we compared our models with a rule-based classifier, more comprehen-

sive comparisons with other methods adapted for causal question identification could provide further insights. Finally, *NatQuest* may contain inherent biases due to its online sources, potentially influencing classifier performance and generalizability. Addressing these limitations in future work could involve creating additional labeled datasets, developing fine-grained classification models, exploring practical applications, and establishing standardized benchmarks for causal question identification tasks.

Ethical Considerations

Data License *NatQuest* comprises publicly available sources. We carefully review the licenses of each source used to build *NatQuest*. The shareGPT dataset has a *Apache-2* license, WildChat has a *AI2 ImpACT License – Low Risk Artifacts*, MSMARCO a *Non-commercial research purposes only*, NaturalQuestions a *Creative Commons Share-Alike 3.0* and the license for Quora follows their terms of service. All the above allow the re-publishing of data for non-commercial purposes, hence we release the full dataset, complemented by our annotations.

Risk of Misuse We do not see direct potential for misuse or harm due to *NatQuest*. The intended use is to spur research about human curiosity and information-seeking behavior. All the data sources were already available before this work. The insights about human behavior should be taken into consideration with care, considering their possible limitations as described in Section 6.

Also, since some of the sources used to build *NatQuest* may contain toxic and/or NSFW samples, we advise users that *NatQuest* might necessitate filtering and pre-processing to ensure the safety and appropriateness of the dataset for downstream applications.

Author Contributions

The *idea and design* of the project originated in discussions among Zhijing, Dmitrii, Roberto, and refined with Mrinmaya. In the initial exploration stage, Dmitrii started collecting the data and studying the questions in detail to find meaningful categorizations. The *design of the taxonomies* was developed in discussions among Roberto and Zhijing. The *data collection* (including prompt engineering, manual annotation, experimentation with GPT and labelling of *NatQuest*) was led by Roberto, with the important contribution of Ahmad and Amélie. The *fine-tuning* of the language models was led by Dmitrii. The *cleaning and compilation* of the code and data was done by Roberto. All co-authors contributed to writing the paper, especially Roberto, Zhijing and Ahmad, with significant feedback from Mrinmaya, Rada and Bernhard.

Acknowledgment

This material is based in part upon work supported by the German Federal Ministry of Education and Research (BMBF): Tübingen AI Center, FKZ: 01IS18039B; by the Machine Learning Cluster of Excellence, EXC number 2064/1 – Project number 390727645; by a National Science Foundation award (#2306372); by a Swiss National Science Foundation award (#201009) and a Responsible AI grant by the Haslerstiftung. The usage of OpenAI credits are largely supported by the Tübingen AI Center. Zhijing Jin is supported by PhD fellowships from the Future of Life Institute and Open Philanthropy, as well as travel support from ELISE (GA #951847) for the ELLIS program.

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 edition*. Addison Wesley Longman, Inc. 4, 19

Mousumi Banerjee, Michelle Capozzoli, Laura McSweeney, and Debajyoti Sinha. 1999. Beyond kappa: A review of interrater agreement measures. *Canadian journal of statistics*, 27(1):3–23. 19

Elias Bareinboim, Juan D. Correa, Duligur Ibeling, and Thomas Icard. 2022. On pearl's hierarchy and the foundations of causal inference. In Hector Geffner, Rina Dechter, and Joseph Y. Halpern, editors, *Probabilistic and Causal Inference: The Works of Judea Pearl*, volume 36 of *ACM Books*, pages 507–556. ACM. 6

Helen Beebee, Christopher Hitchcock, Peter Charles Menzies, and Peter Menzies. 2009. *The Oxford handbook of causation*. Oxford Handbooks. 5

Douglas Biber. 2012. Register as a predictor of linguistic variation. *Corpus linguistics and linguistic theory*, 8(1):9–37. 8

Laura Biester, Ryan Boyd, Zhijing Jin, Veronica Perez Rosas, Steven Wilson, James Pennebaker, and Rada Mihalcea. 2024. Inferring human behavior from language. *Nature Human Behavior*. 1

Su Lin Blodgett, Solon Barocas, Hal Daumé III, and Hanna M. Wallach. 2020. Language (technology) is

power: A critical survey of "bias" in NLP. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, ACL 2020, Online, July 5-10, 2020*, pages 5454–5476. Association for Computational Linguistics. 9

Benjamin Samuel Bloom, Max D Engelhart, Edward J Furst, Walker H Hill, and David R Krathwohl. 1964. *Taxonomy of educational objectives*, volume 2. Longmans, Green New York. 2

Alexander Bondarenko, Magdalena Wolska, Stefan Heindorf, Lukas Blübaum, Axel-Cyrille Ngonga Ngomo, Benno Stein, Pavel Braslavski, Matthias Hagen, and Martin Potthast. 2022. CausalQA: A benchmark for causal question answering. In *COLING*, pages 3296–3308. 1, 7, 9, 17

Ryan L. Boyd and H. Andrew Schwartz. 2021. Natural language analysis and the psychology of verbal behavior: The past, present, and future states of the field. *Journal of Language and Social Psychology*, 40(1):21–41. PMID: 34413563. 8

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel M. Ziegler, Jeffrey Wu, Clemens Winter, Christopher Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. *arXiv preprints*. 1, 6

Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, Bowen Baker, Leo Gao, Leopold Aschenbrenner, Yining Chen, Adrien Ecoffet, Manas Joglekar, Jan Leike, Ilya Sutskever, and Jeffrey Wu. 2024. Weak-to-strong generalization: Eliciting strong capabilities with weak supervision. In *Forty-first International Conference on Machine Learning, ICML 2024, Vienna, Austria, July 21-27, 2024*. OpenReview.net. 1

Jonathan P. Chang, Caleb Chiam, Liye Fu, Andrew Z. Wang, Justine Zhang, and Cristian Danescu-Niculescu-Mizil. 2020. Convokit: A toolkit for the analysis of conversations. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue, SIGdial 2020, 1st virtual meeting, July 1-3, 2020*, pages 57–60. Association for Computational Linguistics. 15

Lingjiao Chen, Matei Zaharia, and James Zou. 2023. Frugalgpt: How to use large language models while reducing cost and improving performance. *arXiv preprint arXiv:2305.05176*. 2

Tianqi Chen and Carlos Guestrin. 2016. XGBoost: A scalable tree boosting system. In *Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Francisco, CA, USA, August 13-17, 2016*, pages 785–794. ACM. 15

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, Kensen Shi, Sasha Tsvyashchenko, Joshua Maynez, Abhishek Rao, Parker Barnes, Yi Tay, Noam Shazeer, Vinodkumar Prabhakaran, Emily Reif, Nan Du, Ben Hutchinson, Reiner Pope, James Bradbury, Jacob Austin, Michael Isard, Guy Gur-Ari, Pengcheng Yin, Toju Duke, Anselm Levskaya, Sanjay Ghemawat, Sunipa Dev, Henryk Michalewski, Xavier Garcia, Vedant Misra, Kevin Robinson, Liam Fedus, Denny Zhou, Daphne Ippolito, David Luan, Hyeontaek Lim, Barret Zoph, Alexander Spiridonov, Ryan Sepassi, David Dohan, Shivani Agrawal, Mark Omernick, Andrew M. Dai, Thanumalayan Sankaranarayana Pillai, Marie Pellat, Aitor Lewkowycz, Erica Moreira, Rewon Child, Oleksandr Polozov, Katherine Lee, Zongwei Zhou, Xuezhi Wang, Brennan Saeta, Mark Diaz, Orhan Firat, Michele Catasta, Jason Wei, Kathy Meier-Hellstern, Douglas Eck, Jeff Dean, Slav Petrov, and Noah Fiedel. 2023. Palm: Scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24:240:1–240:113. 1

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Y. Zhao, Yanping Huang, Andrew M. Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling instruction-finetuned language models. *CoRR*, abs/2210.11416. 15

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge. 3

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. 2021. Training verifiers to solve math word problems. 3

Anna Coenen, Jonathan D Nelson, and Todd M Gureckis. 2019. Asking the right questions about the psychology of human inquiry: Nine open challenges. *Psychonomic Bulletin & Review*, 26(5):1548–1587. 1

Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46. 6

Cristian Danescu-Niculescu-Mizil, Moritz Sudhof, Dan Jurafsky, Jure Leskovec, and Christopher Potts. 2013. A computational approach to politeness with application to social factors. In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics, ACL 2013, 4-9 August 2013, Sofia, Bulgaria, Volume 1: Long Papers*, pages 250–259. The Association for Computer Linguistics. 15

Dujian Ding, Ankur Mallick, Chi Wang, Robert Sim, Subhabrata Mukherjee, Victor Ruhle, Laks VS Lak-

shmanan, and Ahmed Hassan Awadallah. 2024. Hybrid llm: Cost-efficient and quality-aware query routing. *arXiv preprint arXiv:2404.14618*. 8

Ronald A. Fisher and E. B. Ford. 1927. The spread of a gene in natural conditions in a colony of the moth panaxia dominula l. *Heredity*, 11:143–174. Early work by Fisher on the application of randomization in agricultural experiments. 5

Roxana Girju and Dan Moldovan. 2002. Mining answers for causation questions. In *Proc. The AAAI Spring Symposium on Mining Answers from Texts and Knowledge Bases*, pages 67–82. 7, 17

Jacqueline Gottlieb, Pierre-Yves Oudeyer, Manuel Lopes, and Adrien Baranes. 2013. Information-seeking, curiosity, and attention: computational and neural mechanisms. *Trends in cognitive sciences*, 17(11):585–593. 1

Arthur C Graesser and Natalie K Person. 1994. Question asking during tutoring. *American educational research journal*, 31(1):104–137. 8

Ilya Gusev and Alexey Tikhonov. 2022. HeadlineCause: A dataset of news headlines for detecting causalities. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 6153–6161, Marseille, France. European Language Resources Association. 8

J. A. Hartigan and M. A. Wong. 1979. Algorithm as 136: A k-means clustering algorithm. *Journal of the Royal Statistical Society. Series C (Applied Statistics)*, 28(1):100–108. 3

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2022. Lora: Low-rank adaptation of large language models. In *The Tenth International Conference on Learning Representations, ICLR 2022, Virtual Event, April 25-29, 2022*. OpenReview.net. 15

ISO. 2018. Ergonomics of human-system interaction - part 11: Usability: Definitions and concepts (iso 9241-11:2018). 8, 15

Shankar Iyer, Nikhil Dandekar, and Kornél Csernai. 2017. Quora question pairs dataset. 1, 3

Joshua Conrad Jackson, Joseph Watts, Johann-Mattis List, Curtis Puryear, Ryan Drabble, and Kristen A Lindquist. 2022. From text to thought: How analyzing language can advance psychological science. *Perspectives on Psychological Science*, 17(3):805–826. 8

Zhijing Jin, Yuen Chen, Felix Leeb, Luigi Gresele, Ojasv Kamal, Zhiheng Lyu, Kevin Blin, Fernando Gonzalez Adauto, Max Kleiman-Weiner, Mrinmaya Sachan, and Bernhard Schölkopf. 2023. CLadder: A benchmark to assess causal reasoning capabilities of language models. In *Advances in Neural Information Processing Systems 36: Annual Conference on Neural Information Processing Systems 2023, NeurIPS 2023, New Orleans, LA, USA, December 10 - 16, 2023*. 1, 8

Zhijing Jin, Jiarui Liu, Zhiheng Lyu, Spencer Poff, Mrinmaya Sachan, Rada Mihalcea, Mona T. Diab, and Bernhard Schölkopf. 2024. Can large language models infer causation from correlation? In *The Twelfth International Conference on Learning Representations, ICLR 2024*. OpenReview.net. 1, 8

Ankur Joshi, Saket Kale, Satish Chandel, and D Kumar Pal. 2015. Likert scale: Explored and explained. *British journal of applied science & technology*, 7(4):396–403. 15

Immanuel Kant. 1781. *Critique of Pure Reason*. Cambridge University Press. 5

Emre Kiciman, Robert Ness, Amit Sharma, and Chenhao Tan. 2023. Causal reasoning and large language models: Opening a new frontier for causality. *CoRR*, abs/2305.00050. 8

Carol Collier Kuhlthau. 2005. Information search process. *Hong Kong, China*, 7(2005):226. 4

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*. 1, 2, 3

Yash Kumar Lal, Nathanael Chambers, Raymond Mooney, and Niranjan Balasubramanian. 2021a. TellMeWhy: A dataset for answering why-questions in narratives. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 596–610, Online. Association for Computational Linguistics. 9

Yash Kumar Lal, Nathanael Chambers, Raymond J. Mooney, and Niranjan Balasubramanian. 2021b. Tellmewhy: A dataset for answering why-questions in narratives. In *ACL/IJCNLP (Findings)*, volume ACL/IJCNLP 2021 of *Findings of ACL*, pages 596–610. Association for Computational Linguistics. 9

J Richard Landis and Gary G Koch. 1977. An application of hierarchical kappa-type statistics in the assessment of majority agreement among multiple observers. *Biometrics*, pages 363–374. 6

Yuanzhi Li, Sébastien Bubeck, Ronen Eldan, Allie Del Giorno, Suriya Gunasekar, and Yin Tat Lee. 2023. Textbooks are all you need II: phi-1.5 technical report. *CoRR*, abs/2309.05463. 15

Stephanie Lin, Jacob Hilton, and Owain Evans. 2022. Truthfulqa: Measuring how models mimic human falsehoods. 3

AH Maslow. 1943. A theory of human motivation. *Psychological Review google schola*, 2:21–28. 4

Mary L McHugh. 2012. Interrater reliability: the kappa statistic. *Biochemia medica*, 22(3):276–282. 6

Nasrin Mostafazadeh, Aditya Kalyanpur, Lori Moon, David Buchanan, Lauren Berkowitz, Or Biran, and Jennifer Chu-Carroll. 2020. GLUCOSE: GeneraLized and

COntextualized story explanations. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4569–4586, Online. Association for Computational Linguistics. 9

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated machine reading comprehension dataset. In *Proceedings of the Workshop on Cognitive Computation: Integrating neural and symbolic approaches 2016 co-located with the 30th Annual Conference on Neural Information Processing Systems (NIPS 2016), Barcelona, Spain, December 9, 2016*, volume 1773 of *CEUR Workshop Proceedings*. CEUR-WS.org. 1, 2, 3

Isaac Ong, Amjad Almahairi, Vincent Wu, Wei-Lin Chiang, Tianhao Wu, Joseph E Gonzalez, M Waleed Kadous, and Ion Stoica. 2024. Routellm: Learning to route llms with preference data. *arXiv preprint arXiv:2406.18665*. 8

OpenAI. 2024. Hello gpt4o. 15

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, EMNLP 2023, Singapore, December 6-10, 2023*, pages 2375–2393. Association for Computational Linguistics. 1, 7, 9

Judea Pearl. 2009a. Causal inference in statistics: An overview. 2, 6

Judea Pearl. 2009b. *Causality: Models, reasoning and inference (2nd ed.)*. Cambridge University Press. 5, 6

Judea Pearl. 2019. The seven tools of causal inference, with reflections on machine learning. *Communications of the ACM*, 62(3):54–60. 2

James W Pennebaker, Matthias R Mehl, and Kate G Niederhoffer. 2003. Psychological aspects of natural language use: Our words, our selves. *Annual review of psychology*, 54(1):547–577. 8

Jonas Peters, Dominik Janzing, and Bernhard Schölkopf. 2017. *Elements of causal inference: Foundations and learning algorithms*. The MIT Press. 2, 5

Shanette C. Porter, Michelle Rheinschmidt-Same, and Jennifer A. Richeson. 2016. Inferring identity from language: Linguistic intergroup bias informs social categorization. *Psychological Science*, 27(1):94–102. 1

Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, Melbourne, Australia. Association for Computational Linguistics. 3

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing, EMNLP 2016, Austin, Texas, USA, November 1-4, 2016*, pages 2383–2392. The Association for Computational Linguistics. 1

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R. Bowman. 2023. GPQA: A graduate-level google-proof q&a benchmark. *CoRR*, abs/2311.12022. 4

Melissa Roemmele, Cosmin Adrian Bejan, and Andrew S. Gordon. 2011. Choice of plausible alternatives: An evaluation of commonsense causal reasoning. In *Logical Formalizations of Commonsense Reasoning, Papers from the 2011 AAAI Spring Symposium, Technical Report SS-11-06, Stanford, California, USA, March 21-23, 2011*. AAAI. 1, 9

Anselm Rothe, Brenden M Lake, and Todd M Gureckis. 2018. Do people ask good questions? *Computational Brain & Behavior*, 1:69–89. 1

Donald B. Rubin. 1980. Randomization analysis of experimental data: The fisher randomization test comment. *Journal of the American Statistical Association*, 75(371):591–593. 5

Bertrand Russell. 2004. *History of western philosophy*. Routledge. 5

Maarten Sap, Ronan Le Bras, Emily Allaway, Chandra Bhagavatula, Nicholas Lourie, Hannah Rashkin, Brendan Roof, Noah A. Smith, and Yejin Choi. 2019. ATOMIC: an atlas of machine commonsense for if-then reasoning. In *The Thirty-Third AAAI Conference on Artificial Intelligence, AAAI 2019, The Thirty-First Innovative Applications of Artificial Intelligence Conference, IAAI 2019, The Ninth AAAI Symposium on Educational Advances in Artificial Intelligence, EAAI 2019, Honolulu, Hawaii, USA, January 27 - February 1, 2019*, pages 3027–3035. AAAI Press. 1

John Sinclair and Malcolm Coulthard. 2013. Towards an analysis of discourse. In *Advances in spoken discourse analysis*, pages 1–34. Routledge. 8

Steven A Sloman and David Lagnado. 2015. Causality in thought. *Annual review of psychology*, 66(1):223–247. 2

Karen Sparck Jones. 1972. A statistical interpretation of term specificity and its application in retrieval. *Journal of documentation*, 28(1):11–21. 15

Peter Spirtes, Clark Glymour, and Richard Scheines. 1993. Causation, prediction, and search. 5

Peter Spirtes, Clark Glymour, and Richard Scheines. 2000. *Causation, Prediction, and Search, Second Edition*. Adaptive computation and machine learning. MIT Press. 5

Niket Tandon, Bhavana Dalvi, Keisuke Sakaguchi, Peter Clark, and Antoine Bosselut. 2019. WIQA: A dataset for "what if..." reasoning over procedural text. In *Proceedings of the 2019 Conference on Empirical Methods*

in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), pages 6076–6085, Hong Kong, China. Association for Computational Linguistics. 1, 8

Yla R Tausczik and James W Pennebaker. 2010. The psychological meaning of words: Liwc and computerized text analysis methods. *Journal of language and social psychology*, 29(1):24–54. 8

Carsten Ullrich and Erica Melis. 2009. Pedagogically founded courseware generation based on htn-planning. *Expert Systems with Applications*, 36(5):9319–9332. 19

Laurens Van der Maaten and Geoffrey Hinton. 2008. Visualizing data using t-sne. *Journal of machine learning research*, 9(11). 3

Suzan Verberne, Lou Boves, Peter-Arno Coppen, and Nelleke Oostdijk. 2006. Discourse-based answering of why-questions. *Trait. Autom. des Langues*, 47(2):21–41. 9

Suzan Verberne, Lou Boves, Nelleke Oostdijk, and Peter-Arno Coppen. 2008. Using syntactic information for improving why-question answering. In *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, pages 953–960, Manchester, UK. Coling 2008 Organizing Committee. 9

Anthony J Viera, Joanne M Garrett, et al. 2005. Understanding interobserver agreement: the kappa statistic. *Fam med*, 37(5):360–363. 6

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, Tianle Li, Max Ku, Kai Wang, Alex Zhuang, Rongqi Fan, Xiang Yue, and Wenhu Chen. 2024. Mmlu-pro: A more robust and challenging multi-task language understanding benchmark. *CoRR*, abs/2406.01574. 3

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, brian ichter, Fei Xia, Ed H. Chi, Quoc V Le, and Denny Zhou. 2022. Chain of thought prompting elicits reasoning in large language models. In *Advances in Neural Information Processing Systems*. 6

Tom D Wilson. 1999. Models in information behaviour research. *Journal of documentation*, 55(3):249–270. 4

Matej Zecevic, Moritz Willig, Devendra Singh Dhami, and Kristian Kersting. 2023. Causal parrots: Large language models may talk causality but are not causal. *CoRR*, abs/2308.13067. 8

Cheng Zhang, Stefan Bauer, Paul Bennett, Jiangfeng Gao, Wenbo Gong, Agrin Hilmkil, Joel Jennings, Chao Ma, Tom Minka, Nick Pawlowski, et al. 2023. Understanding causality with large language models: Feasibility and opportunities. *arXiv preprint arXiv:2304.05524*. 8

Kun Zhang and Aapo Hyvärinen. 2009. Causality discovery with additive disturbances: An information-theoretical perspective. In *Machine Learning and Knowledge Discovery in Databases: European Conference, ECML PKDD 2009, Bled, Slovenia, September 7-11, 2009, Proceedings, Part II 20*, pages 570–585. Springer. 5

Wenting Zhao, Xiang Ren, Jack Hessel, Claire Cardie, Yejin Choi, and Yuntian Deng. 2024. (inthe)wildchat: 570k chatGPT interaction logs in the wild. In *The Twelfth International Conference on Learning Representations*. 1, 2, 3

Lianmin Zheng, Wei-Lin Chiang, Ying Sheng, Siyuan Zhuang, Zhanghao Wu, Yonghao Zhuang, Zi Lin, Zhuohan Li, Dacheng Li, Eric P. Xing, Hao Zhang, Joseph E. Gonzalez, and Ion Stoica. 2023. Judging LLM-as-a-judge with MT-Bench and Chatbot arena. *CoRR*, abs/2306.05685. 4

Tao Zhou and Songtao Li. 2024. Understanding user switch of information seeking: From search engines to generative ai. *Journal of Librarianship and Information Science*, page 09610006241244800. 7

A Additional Analyses

Table 10: Inter-rater reliability measures for different aspects of question evaluation.

Aspect	Cohen's Kappa
Cognitive complexity	0.6926 (weighted)
Users' needs	0.58
Open-endedness	0.63
Domain	0.73
Subjectivity	0.68

A.1 Cluster Details

In Appendix A.1 we show some question examples from each cluster in *NatQuest*.

A.2 Data Statistics for Causal and Non-Causal Questions

In Table 13 we show the overall statistics for causal and non-causal questions in *NatQuest*.

A.3 Subjectivity

Questions are also classified based on whether they are subjective (i.e., involving personalized, culture-specific opinions, etc.) or objective. We find that SE receives mostly objective questions, and users typically don't raise too many subjective, opinion-seeking questions to them. H-to-H online forums receive 53% subjective questions. LLMs receive 29% subjective questions, although the majority are still objective, factual ones. There is still a way to go for LLMs to gain more trust and to let people ask more personally related questions.

A.4 Politeness

We look at politeness using the library ConvoKit (Chang et al., 2020) and follow the rules to compute politeness introduced by Danescu-Niculescu-Mizil et al. (2013). We build a composite *politeness score* by adding 1 for the presence of each positive linguistic indicator of politeness, and subtracting 1 for negative ones. We refer to Danescu-Niculescu-Mizil et al. (2013) for more details about positive and negative markers. In line with intuition, we find that internet users are mostly polite with other humans (Quora). Interestingly, interactions with LLMs follow, and finally SEs.

A.5 GPT-4o evaluation on *NatQuest*

In this section, we evaluate the performance of GPT-4o on *NatQuest*.

NatQuest does not contain ground truth answers due to the open-ended nature of many of the questions (see Appendix A.3). This is also the case for several of our data channels, in particular, both H-to-LLMs and H-to-H datasets don't have answers. Among H-to-SEs datasets, NaturalQuestions have a Wikipedia passage containing the answer, and MSMarco has human-written answers.

Nevertheless, we can evaluate answers based on a metric of 3 usability criteria used in user satisfaction surveys (ISO, 2018): *effectiveness* (does the answer complete the goal specific to this user?), *efficiency* (does the answer provide just the right amount of information or is it too vague / overly detailed?) and *satisfaction* (is the answer friendly and pleasant to read, leaving the user satisfied?).

We assess the answer quality on 50 randomly sampled causal questions from *NatQuest* for GPT-4o (OpenAI, 2024). Two human annotators score the answers on a scale from 0 to 5 (similar to a Likert scale (Joshi et al., 2015)) on our metric defined above. We obtain a mean effectiveness of 3.83 (std 0.97), efficiency 2.88 (std 0.78) and satisfaction 3.85 (std 0.63). We find that GPT-4o struggles when asked to foresee the future or questions related to personal decisions, but performs well on causal questions requiring knowledge lookup. The answers are also overly verbose (reflected in the lower efficiency score) and struggle with empathy and friendliness as the answers almost always just provide a list of factors to consider without engaging with the user. Tackling these weaknesses can be a large avenue for future work to achieve AI agents better equipped to get closer to the types of answers that humans seek.

A.6 Evaluation of Different Models on Causal Question Classification

Here we train or fine-tune various efficient models for causal question classification.

A Selection of Efficient Models We first train a baseline model that does not involve deep learning, namely XGBoost (Chen and Guestrin, 2016) using TF-IDF for vectorization (Sparck Jones, 1972), and we fine-tune 5 language models, with a number of parameters ranging from 80M to 2.85B (Chung et al., 2022; Li et al., 2023). We use full-weight supervised fine-tuning for smaller models and LoRA (Hu et al., 2022) for the largest ones to save on computing power. For models running on 3090s RTX, the hardware has 24GB of memory, and the A100 has 40 GB.

Table 11: Example questions for each topic cluster.

Cluster	Example Questions
Daily life	• How can I marry a millionaire? • What should I say when someone is expressing concern for my health? • How can I cope with feeling constantly numb and unmotivated? • Can you suggest anchor thoughts for optimizing my mindset, especially in the morning? • Should I follow up with him to ask about his feelings, even if it may result in rejection?
Computer-Related	• Why are the signals generated by my code incorrect? • Can you create a 3D wireframe grid in JavaScript on a canvas without frameworks or libraries, allowing for snapping lines and adding points? • How can I explain to the brand/marketing team that iOS app deployment may take up to 24 hours, causing updates to not be instantly available to all customers? • What is the use of static keyword in Java? • Can I use the rules and the design (not logo) of monopoly in my mobile game?
Sports, medicine and science	• Where does the fertilization of the egg occur? • What type of image is formed by a 4 cm object placed 40 cm away from a convex mirror with a radius of curvature of 20 cm? • A decrease in the normal amount of urine is called oliguria. • List of high school players drafted in NBA. • Which hormone is most responsible for signaling satiety as well as reducing food intake during a meal?
Prompt Questions	• Generate detailed image prompts for the AI "Midjourney" based on given concepts, varying in description, environment, composition, atmosphere, and style. • Generate detailed image prompts for the AI "Midjourney" in Chinese ink style, depicting a poet alone in the snow, smelling plum blossoms from a wine glass. • Can you generate imaginative prompts for Midjourney's AI program to inspire unique and interesting images? Start with something beyond human imagination. • Generate detailed image prompts for the AI "Midjourney" featuring a concept, environment, composition, mood, and style. Follow the specified structure and guidelines.
Stories and Fictional characters	• Where does the new Beauty and the Beast take place? • How will Game of Thrones end? • The Legend of Zelda: Why does Link have pointy ears? • Short summary of Miss Peregrine's Home for Peculiar Children book. • Creatively name and describe fruits and vegetables from an alien world

Table 12: Causal and non-causal questions in each topic cluster.

Cluster	# Questions	Causal (%)	Non-Causal (%)
Daily life	3,609	56.08	43.92
Computer-Related	3,806	59.88	40.12
Sports, medicine and science	3,875	22.99	77.01
Prompt Questions	216	32.41	67.59
Stories and Fictional characters	1,994	21.92	78.08

Table 13: Overall statistics of our *NatQuest* dataset.

	Overall	Causal	Non-Causal
# Samples	13,500	5,701	7,799
# Words/Sample	11.34	12.56	10.45
Vocab Size	25,709	14,526	17,367
Type-Token Ratio	0.168	0.202	0.213

Table 9 highlights the tradeoff between the accuracy and size of the model. Future practitioners who may want to use *causal classification* have several valid options, depending on the balance they need between resource usage and accuracy. For example, FLAN-T5-XL (LoRA) performed best but required significantly more resources than all other models. Smaller FLAN-T5 models and Phi-1.5 represent a compromise, giving up 2-3% but saving on computations. Finally, using our baseline can also be an option for use cases with very limited computing power, being extremely fast and lightweight and still achieving a 72% F1 score on the task. We also check the performance of a linguistics-based

classifier, which integrates both lexical rules based on Bondarenko et al. (2022) and lexico-syntactic patterns along with sets of causal and morphological connectives in Girju and Moldovan (2002). This classifier achieves an accuracy of 23% and a precision of 63% in identifying causal questions. The main reason for its limited performance is that our natural questions are often more formulated in an informal manner, contrasting with more standard grammar used in the other datasets. For example, the question word is implicit in the sample "*win a grand slam without losing a set?*". These results suggest that effective identification of causal questions requires a deeper, semantic understanding of the questions as well as background knowledge, in particular, to be able to capture the implicit and handle the ambiguity intrinsic to many causative constructions (Girju and Moldovan, 2002).

B Prompts

All prompts have a similar structure: each possible category is defined in detail, then some examples are provided.

B.1 Summarisation

Needs

Below you'll find a question that a human asked on {source}. Reformulate the question more concisely, while retaining the original key idea. You can skip the details. Maximum length: 30 words.
Question: {question}
Shorter question:

B.2 Causality

Causality

The following is a question that a human asked online. Classify the question in one of the following two categories:
Category 1: Causal. This category includes questions that suggest a cause-and-effect relationship broadly speaking, requiring the use of cause-and-effect knowledge or reasoning to provide an answer. A cause is a preceding thing, event, or person that contributes to the occurrence of a later thing or event, to the extent that without the preceding one, the later one would not have occurred. A causal question can have different mechanistic natures. It can be: 1. Given the cause, predict the effect: seeking to understand the impact or outcome of a specific cause, which might involve predictions about the future or hypothetical scenarios.
2. Given the effect, predict the cause: asking "Why" something occurs (e.g. Why do apples fall?), probing the cause of a certain effect, asking about the reasons behind something, or the actions needed to achieve a specific result or goal, "How to" do something, explicitly or implicitly (e.g. Why does far right politics rise nowadays? How to earn a million dollars? How to learn a language in 30 days?). This also includes the cases in which the effect is not explicit: any request with a purpose, looking for ways to fulfill it. It means finding the action (cause) to best accomplish a certain goal (effect), and the latter can also be implicit. If someone asks for a restaurant recommendation, what she's looking for is the best cause for a certain effect which can be, e.g., eating healthy. If asking for a vegan recipe, she's looking for the recipe that causes the best possible meal. Questions asking for "the best" way to do something, fall into this category. Asking for the meaning of something that has a cause, like a song, a book, a movie, is also causal, because the meaning is part of the causes that led to the creation of the work. A coding task which asks for a very specific effect to be reached, is probing for the cause (code) to obtain that objective.
3. Given variables, judge their causal relation: questioning the causal link among the given entities (e.g. Does smoking cause cancer? Did my job application get rejected because I lack experience?)
Be careful: causality might also be implicit! Some examples of implicit causal questions:
- the best way to do something
- how to achieve an effect
- what's the effect of an action (which can be in the present, future or past)
- something that comes as a consequence of a condition (e.g. how much does an engineer earn, what is it like to be a flight

attendant)
- when a certain condition is true, does something happen?
- where can I go to obtain a certain effect? - who was the main cause of a certain event, author, inventor, founder?
- given an hypothetical imaginary condition, what would be the effect?
- what's the feeling of someone after a certain action?
- what's the code to obtain a certain result?
- when a meaning is asked, is it because an effect was caused by a condition (what's the meaning of <effect>)?
- the role, the use, the goal of an entity, an object, is its effect

Category 2: Non-causal. This category encompasses questions that do not imply in any way a cause-effect relationship.
Let's think step by step. Question: {question}

B.3 Causality Types

Causality types

You are tasked with classifying causal questions into one of four types based on the following taxonomy. Carefully read each type's definition and examples before proceeding to classify the given question. Provide a brief explanation for your classification.

Causal Question Types:
Type 1: About Variables
Definition:
- Questions that ask about the variables in the causal graph. E.g., they inquire about the causes or contributing factors of a phenomenon.
Characteristics:
- Seek to identify variables that play a role in causing an effect in a specific case or scenario.
- Often start with "Why did X ..." or "Where does X come from ...", "What will happen if. . . "
Examples:
"Why did company X close in Europe?"
"What will happen if Y gets elected?"
Type 2: About Relations
Definition:
Questions that ask about the existence of directed edges (causal relationships) among variables. They inquire whether one variable directly affects another.
Characteristics:
- Investigate if a causal link exists between specific variables.
- Often phrased as "Does X cause Y?" or "Is there a relationship between X and Y?"
Examples:
- "Does smoking interfere with the drug effect?"
- "Was my application rejected due to my lack of work experience?"
- "Does stress lead to heart disease?"
Type 3: About Average Effects
Definition:
Questions that ask about the quantification of the average change in an effect given a cause. They often involve measuring the magnitude of an effect.
Characteristics:
- Focus on the extent or degree to which one variable affects another on average.
- Often include phrases like "How much," "To what extent," or "What is the effect size of..."
Examples:
- "How much do COVID vaccines decrease hospitalization risk?"
- "Are small or big classrooms better for kids?"
- "What is the average improvement in test scores due to tutoring?"
Type 4: About Mechanisms Definition:
Questions that ask about the functions or mechanisms among variables, or require understanding the underlying processes. This includes counterfactual questions.
Characteristics:
- These are questions about a causal relation in a phenomenon which holds always, or most of the time.
- May involve hypothetical or counterfactual scenarios.
- Often start with "How does..." or "What would happen if...", "Why does X happen?",

"How can someone achieve Y", "What are the causes of".
Examples:
- "Had I not done a PhD, would my life be different?"
- "How do scientists prepare rockets for missions to the Moon?"
- "What is the biochemical process by which insulin regulates blood sugar?"
- "What makes a good doctor?"
- "Why are some people X?"
- "How can someone achieve Y?"
- "What are the causes of Z?"
—
Instructions:

1. Read the Question Carefully: Understand what the question is asking.
2. Think Step by Step: Analyze the question to identify its underlying causal nature.
3. Match with Definitions and Characteristics: Compare the question with the definitions above.
4. Provide Classification: Assign the question to one of the four types.
5. Explain Your Reasoning: Briefly justify why the question fits that type.
—
Example Classification:
- Question: "Does regular exercise improve mental health?"
- Analysis:
- The question is asking about the existence of a causal relationship between regular exercise and mental health.
- It inquires whether one variable (exercise) affects another (mental health).
- Classification: Type 2: About Relations
- Explanation: The question seeks to determine if there is a causal link between two variables.
—
To recap, remember the following: Variables = when the question wants to know the "content" of a node of the causal graph, e.g. a cause or an effect, or another node like a mediator, in a specific case. It is asking for an element of a graph in a one-time phenomenon, like why did X happen, how did Y happen. Edge = question about the existence or not of a causal relationship. Avg effect = question about quantification of a causal effect. Mechanism = question about how a phenomenon works in general - so once again about the content of the nodes / structure of the graph but not of a specific case, but for something that is always in the same way, governs how things happen. It is asking for what normally, usually happens - what happens if I do X, why does X happen, How does X happen, why would X happen, where can X happen, how can X be achieved.
Now, please classify the following question:
Causal Question: {QUESTION}

B.4 Cognitive Complexity

We follow a commonly used taxonomy for evaluating the kinds of intellectual skills needed to answer a question by Anderson and Krathwohl (2001), using GPT4o-mini to classify questions into the different required skills. We follow a procedure similar to Section 3.2 to validate the efficacy of our prompt. We further evaluated our results against the group truth, achieving a 0.64 Cohen's kappa score (Banerjee et al., 1999; Ullrich and Melis, 2009), which is interpreted as "moderate agreement."

Cognitive Complexity

Your task is to classify given statements or questions according to Anderson and Krathwohl's Taxonomy of the Cognitive Domain. Use the following six categories and their descriptions:
Remembering: Recognizing or recalling knowledge from memory. Remembering is when memory is used to produce or retrieve definitions, facts, or lists, or to recite previously learned information. Factual questions, that do not require reasoning fall into this category.
Understanding: Constructing meaning from different types of functions be they written or graphic messages or activities like interpreting, exemplifying, classifying, summarizing, inferring, comparing, or explaining. Questions asking for the meaning or explanation of a concept fall into this category.

Applying: Carrying out or using a procedure through executing, or implementing. Applying relates to or refers to situations where learned material is used, applied in a concrete situation, is used to present or show something. Questions that require the application of some theory or rule. For example, requiring some calculation, formula, light reasoning, applied to something in the real world. They do not entail a creative effort, but instead applying some rule or principle. Asking to generate a code with a specific goal (e.g. a cmd code that does yyy) is "apply" whereas asking to build a website requires a creative effort. How to do something, how to make something, how to solve something, how to apply some principle etc.
Analyzing: Breaking materials or concepts into parts, determining how the parts relate to one another, or how the parts relate to an overall structure or purpose. Mental actions included in this function are differentiating, organizing, and attributing, as well as being able to distinguish between the components or parts. When one is analyzing, he/she can illustrate this mental function by creating spreadsheets, surveys, charts, or diagrams, or graphic representations. Questions requiring deeper, more complex considerations on a certain thing. e.g. considering several aspects of something, considering pros and cons etc. Explaining why something is the way it is, by providing evidence or logical reasoning.
Evaluating: Making judgments based on criteria and standards through checking and critiquing. Critiques, recommendations, and reports are some of the products that can be created to demonstrate the processes of evaluation. Evaluating comes before creating as it is often a necessary part of the precursory behavior before one creates something. Questions asking to make judgements, suggestions, recommendations. Also making an hypothesis about something uncertain. Judging whether something is better than something else, or the best.
Creating: Putting elements together to form a coherent or functional whole; reorganizing elements into a new pattern or structure through generating, planning, or producing. Creating requires users to put parts together in a new way, or synthesize parts into something new and different creating a new form or product. This process is the most difficult mental function in the taxonomy. Questions asking for generation tasks that require a creative effort fall into this category.

Examples
1. Q: What does the term 'photosynthesis' mean?
Classification: Understanding
Explanation: The question asks for the meaning of a term, which falls under the 'Understanding' category.
2. Q: Calculate the area of a circle with a radius of 5 meters.
Classification: Applying
Explanation: The question requires applying a formula to calculate the area of a circle, which falls under the 'Applying' category.
3. Q: Compare and contrast the advantages and disadvantages of renewable energy sources.
Classification: Evaluating
4. Q: Design a new logo for a tech startup company.
Classification: Creating
5. Q: Explain the causes of World War II.
Classification: Analyzing
6. Q: What category does the word 'dog' belong to?
Classification: Remembering
7. Q: Is surfing easier to learn than snowboarding?
Classification: Evaluating

Please classify the following question:

{question}

B.5 Domain

To classify *NatQuest* into Knowledge Domains, we use an iterative category generation procedure. We begin with an initial categorization, classify 1000 points using GPT-3.5-turbo-0125 including

an “Other” category, and finally manually inspect the “Other” category to refine the categories. This procedure is repeated until a satisfactory categorization is achieved. Based on the categories, we adopt the iterative prompting improvement to optimize the prompt according to the human annotated set.

Domain

Below you’ll find a question. Classify it in one of the following categories:
1. Natural and Formal Sciences: This category encompasses questions related to the physical world and its phenomena, including, but not limited to, the study of life and organisms (Biology), the properties and behavior of matter and energy (Physics), and the composition, structure, properties, and reactions of substances (Chemistry); also formal sciences belong to this category, such as Mathematics and Logic. Questions in this category seek to understand natural laws, the environment, and the universe at large.
2. Society, Economy, Business: Questions in this category explore the organization and functioning of human societies, including their economic and financial systems. Topics may cover Economics, Social Sciences, Cultures and their evolution, Political Science and Law. Questions regarding business, sales, companies’ choices and governance fall into this category.
3. Health and Medicine: This category focuses on questions related to human health, diseases, and the medical treatments used to prevent or cure them. It covers a wide range of topics from the biological mechanisms behind diseases, the effectiveness of different treatments and medications, to strategies for disease prevention and health promotion. It comprises anything related or connected to human health.
4. Computer Science and Technology: Questions in this category deal with the theoretical foundations of information and computation, along with practical techniques for the implementation and application of these foundations. Topics include, but are not limited to, theoretical computer science, coding and optimization, hardware and software technology and innovation in a broad sense. This category includes the development, capabilities, and implications of computing technologies.
6. Psychology and Behavior: This category includes questions about the mental processes and behaviors of humans. Topics range from understanding why people engage in certain behaviors, like procrastination, to the effects of social factors, and the developmental aspects of human psychology, such as language acquisition in children. The focus is on understanding the workings of the human mind and behavior in various contexts, also in personal lives.
7. Historical Events and Hypothetical Scenarios: This category covers questions about significant past events and their impact on the world, as well as hypothetical questions that explore alternative historical outcomes or future possibilities. Topics might include the effects of major wars on global politics, the potential consequences of significant historical events occurring differently, and projections about future human endeavors, such as space colonization. This category seeks to understand the past and speculate on possible futures or alternative historical happenings.
8. Everyday Life and Personal Choices: Questions in this category pertain to practical aspects of daily living and personal decision-making. Topics can range from career advice, cooking tips, and financial management strategies to advice on maintaining relationships and organizing daily activities. This category aims to provide insights and guidance on making informed choices in various aspects of personal and everyday life. Actionable tips fall into this category.
9. Arts and Culture: This category includes topics in culture across various mediums such as music, television, film, art, games, and social media, sports, celebrities.

Assign one of the above categories to the given question.
Question: {question}

B.6 Users' needs

For user needs, LLMs reach an F1 score of 0.80 with corresponding human annotations.

Needs

Analyze the following question and identify the primary user need category it falls into. Consider the broad categories of user needs as defined below:
Knowledge and Information: Seeking factual information, understanding concepts, or exploring ideas. Questions falling in this category are looking for knowledge for its own sake, as far as we can infer from the question. We do not see an underlying need of the user except for curiosity.
Problem-Solving and Practical Skills: Troubleshooting issues, learning new skills, managing daily life, and handling technology. Anything that is actionable, how to do something or solve a problem.
Personal Well-being: Improving mental and physical health, managing finances, and seeking support.
Professional and Social Development: Advancing career, job searching, and improving social interactions. Any information request about work, school, academia (but that is not actionable, in that case it's 2. Problem-Solving and Practical Skills).
Leisure and Creativity: Finding recreational activities, pursuing hobbies, and seeking creative inspiration.
Categorize this question based on the user's primary need asked in the question, choosing among the categories above.
Examples
1. Q: "What is the chemical symbol for gold?"
A: Knowledge and Information
2. Q: "How do I fix a leaky faucet?"
A: Problem-Solving and Practical Skills
3. Q: "What does the paracetamol do to the body?"
A: Personal Well-being
4. Q: "Create a story about a magical kingdom."
A: Leisure and Creativity
5. Q: "How do I improve my resume?"
A: Professional and Social Development

Question: "question"

B.7 Open-Endedness

Open-Endedness

You are tasked with classifying answerable questions based on the uniqueness of their answers. This classification helps understand the nature of the question and the potential diversity of valid responses.
For answerable questions: 1. Unique Answer: There is a single, specific correct answer that is widely accepted based on current human knowledge. Questions about fictional characters are most likely "Unique answer" because most of the times we can assume the answer is in the book / movie
2. Multiple Valid Answers: There are several plausible, valid answers that could be considered correct depending on perspective, context, or interpretation. Multiple valid answers means either (1) there is a subjective judgment involved (the answer varies depending on who answers), or (2) it is a creative task that can be solved in several different ways, or (3) we as humans do not have access to a unique correct answer. Trivial different wordings of the same concept do not qualify as "Multiple answers". e.g. "What's the meaning of bucolic?"

Ambiguous questions are not necessarily "multiple answers" just because they are ambiguous (i.e. since different people could understand the query differently, they might give different answers). In such cases we should assume a meaning for the question, and reason about the possible ways to answer it.
For each given, classify it as either: 1. Unique Answer
2. Multiple Valid Answers

If the question is looking for a list of items, but the list is unique and well defined, it should be classified as "Unique Answer".
If the question is looking for a number which can be found or computed, it should be classified as "Unique Answer".
If the question is asking how something can be achieved, and there is only one way to achieve it, it should be classified as "Unique Answer", and "Multiple Valid Answers" otherwise.
Examples
1. Q: "What is the chemical symbol for gold?"
Classification: Unique Answer
Explanation: There is a single, universally accepted answer in chemistry: Au.

2. Q: "What is the best programming language for web development?"
Classification: Multiple Valid Answers
Explanation: There are several programming languages suitable for web development, and the "best" can depend on project requirements, developer preference, and other factors.

3. Q: "Who was the first president of the United States?"
Classification: Unique Answer
Explanation: There is a single, historically accepted answer: George Washington.

4. Q: "Tell me some short bedtime stories"
Classification: Multiple Valid Answers
Explanation: There are various short stories that can be told at bedtime, and the choice can vary based on cultural background, personal preference, and other factors.

5. Q: "What is the meaning of life?"
Classification: Multiple Valid Answers
Explanation: This question has multiple valid answers based on different philosophical, religious, and personal perspectives.

6 Q: "What does it mean when an economy is in a recession?"
Classification: Unique Answer
Explanation: There is a specific definition of a recession in economics, making this a question with a unique answer.

7 Q: "Name the three primary colors"
Classification: Unique Answer
Explanation: There are three primary colors in the RGB color model: red, green, and blue.

8 Q: "What criteria should I consider when buying a new laptop?"
Classification: Multiple Valid Answers
Explanation: The criteria for buying a laptop can vary based on individual needs, preferences, and budget constraints.

9. Q: "Who is the best neurosurgeon in New York?"
Classification: Multiple Valid Answers
Explanation: The best neurosurgeon can vary based on specialization, patient reviews, and other factors.
Please classify the following question:

{question}

B.8 Prompt Iteration

Here we show the first and the last prompts of the iteration procedure for Causality and Open-Endedness.

B.8.1 Causality

With first prompt we obtained a weighted F1 of 0.798, with the sixth iteration we obtained 0.894.

First Prompt Causality

The following is a question that a human asked on website. Classify the question in one of the following two categories:
Category 1: Causal. This category includes questions that suggest a cause-and-effect relationship broadly speaking, requiring the use of cause-and-effect knowledge or reasoning to provide an answer.
A causal question can have different mechanistic natures. It can be: 1. Given the cause, predict the effect: seeking to understand the impact or outcome of a specific cause, which might involve predictions about the future or hypothetical scenarios (e.g. What

if I do a PhD? Should I learn how to swim? Will renewable energy sources become the primary means of power? What would the world look like if the Internet had never been invented?);
2. Given the effect, predict the cause: asking "Why" something occurs (e.g. Why do apples fall?), probing the cause of a certain effect, asking about the reasons behind something, or the actions needed to achieve a specific result or goal, "How to" do something, explicitly or implicitly (e.g. Why does far right politics rise nowadays? How to earn a million dollars? How to learn a language in 30 days?). This also includes the cases in which the effect is not explicit: any request with a purpose, looking for ways to fulfill it. It means finding the action (cause) to best accomplish a certain goal (effect), and the latter can also be implicit. If someone asks for a restaurant recommendation, what she's looking for is the best cause for a certain effect which can be, e.g., eating healthy. If asking for a vegan recipe, she's looking for the recipe that causes the best possible meal. Questions asking for "the best" way to do something, fall into this category;
3. Given variables, judge their causal relation: questioning the causal link among the given entities (e.g. Does smoking cause cancer? Did my job application get rejected because I lack experience?)
Categorical 2: Non-causal. This category encompasses questions that do not imply in any way a cause-effect relationship. For example a non-causal question can be asking:
To translate, rewrite, paraphrase a text
To generate a story
To play a game
To provide the solution for a mathematical expression, or a riddle requiring mathematical reasoning
To provide information about something (softwares, websites, materials, events, restaurants) or use such information to make a comparison, without much reasoning. This is non-causal because there is not a specific purpose of the user, but they are only looking for information.

Examples: Question: What would the world look like if the Internet had never been invented? Category: <Causal>
Question: I'd like to play a game of Go with you through text. Let's start with a standard 19x19 board. I'll take black. Place my first stone at D4. Category: <Non-causal>
Question: How can I earn a million dollars fast? Category: <Causal>
Question: How should I spend my last month in Argentina before leaving the country for a long time? Category: <Causal>
Question: Translate "hiking" in Italian Category: <Non-causal>
Question: Will renewable energy sources become the primary means of power? Category: <Causal>
Question: What's the derivative of the logarithm? Category: <Non-causal>
Question: What are some high-protein food options for snacks? Category: <Non-causal>
Question: Does smoking cause cancer? Category: <Causal>
Question: What's the best vegan recipe with broccoli? Category: <Causal>
Question: Write a python script to efficiently sort an array. Category: <Causal>
Question: What's more efficient, Python or C++? Category: <Non-causal>
Question: Tell me the names of all bookshops in Zurich Category: <Non-causal>
Question: Best chair for a home office Category: <Causal>
Answer ONLY with the category in the following format: <Category>, e.g. <Causal>, <Non-causal>.
Question: question Category:

Last Prompt Causality (6th iteration)

The following is a question that a human asked on {website}. Classify the question in one of the following two categories:
Category 1: Causal. This category includes questions that suggest a cause-and-effect relationship broadly speaking, requiring the use of cause-and-effect knowledge or reasoning to provide an answer.
A cause is a preceding thing, event, or

person that contributes to the occurrence of a later thing or event, to the extent that without the preceding one, the later one would not have occurred.
A causal question can have different mechanistic natures. It can be: 1. Given the cause, predict the effect: seeking to understand the impact or outcome of a specific cause, which might involve predictions about the future or hypothetical scenarios.
2. Given the effect, predict the cause: asking "Why" something occurs (e.g. Why do apples fall?), probing the cause of a certain effect, asking about the reasons behind something, or the actions needed to achieve a specific result or goal, "How to" do something, explicitly or implicitly (e.g. Why does far right politics rise nowadays? How to earn a million dollars? How to learn a language in 30 days?).
This also includes the cases in which the effect is not explicit: any request with a purpose, looking for ways to fulfill it. It means finding the action (cause) to best accomplish a certain goal (effect), and the latter can also be implicit. If someone asks for a restaurant recommendation, what she's looking for is the best cause for a certain effect which can be, e.g., eating healthy. If asking for a vegan recipe, she's looking for the recipe that causes the best possible meal. Questions asking for "the best" way to do something, fall into this category.
Asking for the meaning of something that has a cause, like a song, a book, a movie, is also causal, because the meaning is part of the causes that led to the creation of the work. A coding task which asks for a very specific effect to be reached, is probing for the cause (code) to obtain that objective.
3. Given variables, judge their causal relation: questioning the causal link among the given entities (e.g. Does smoking cause cancer? Did my job application get rejected because I lack experience?)
Be careful: causality might also be implicit! Some examples of implicit causal questions:
- the best way to do something
- how to achieve an effect
- what's the effect of an action (which can be in the present, future or past)
- something that comes as a consequence of a condition (e.g. how much does an engineer earn, what is it like to be a flight attendant)
- when a certain condition is true, does something happen?
- where can I go to obtain a certain effect?
- who was the main cause of a certain event, author, inventor, founder?
- given an hypothetical imaginary condition, what would be the effect?
- what's the feeling of someone after a certain action?
- what's the code to obtain a certain result?
- when a meaning is asked, is it because an effect was caused by a condition (what's the meaning of <effect>)?
- the role, the use, the goal of an entity, an object, is its effect

Category 2: Non-causal. This category encompasses questions that do not imply in any way a cause-effect relationship.
Let's think step by step. Answer in the following format: Reasoning: [Reasoning]
Category: [Casual / Non-casual]
Always write "Category" before providing the final answer.
Question: {question}

B.8.2 Open-Endedness

With the first prompt we obtained a weighted F1 of 0.713, with the fourth iteration we obtained 0.795.

First Prompt Open-Endedness

You are tasked with classifying answerable questions based on the uniqueness of their answers. This classification helps understand the nature of the question and the potential diversity of valid responses.
For answerable questions: 1. Unique Answer: There is a single, specific correct answer that is widely accepted based on current human knowledge. Questions about fictional characters are most likely "Unique answer" because most of the time we can assume the answer is in the book / movie 2.

Multiple Valid Answers: There are several plausible, valid answers that could be considered correct depending on perspective, context, or interpretation. Multiple valid answers means either (1) there is a subjective judgment involved (the answer varies depending on who answers), or (2) it is a creative task that can be solved in several different ways, or (3) we as humans do not have access to a unique correct answer. Trivial different wordings of the same concept do not qualify as "Multiple answers". e.g. "What's the meaning of bucolic?"
Ambiguous questions are not necessarily "multiple answers" just because they are ambiguous (i.e. since different people could understand the query differently, they might give different answers). In such cases we should assume a meaning for the question, and reason about the possible ways to answer it.
For each given hypothetically answerable question, classify it as either: 1. Unique Answer 2. Multiple Valid Answers
Examples
1. Q: "What is the chemical symbol for gold?" Classification: Unique Answer Explanation: There is a single, universally accepted answer in chemistry: Au.
2. Q: "What is the best programming language for web development?" Classification: Multiple Valid Answers Explanation: There are several programming languages suitable for web development, and the "best" can depend on project requirements, developer preference, and other factors.
3. Q: "Who was the first president of the United States?" Classification: Unique Answer Explanation: There is a single, historically accepted answer: George Washington.
4. Q: "What is the most effective way to reduce stress?" Classification: Multiple Valid Answers Explanation: There are various effective stress reduction techniques, and the most effective method can vary from person to person.
Please classify the following question:
question

Last Prompt Open-Endedness (4th iteration)

You are tasked with classifying answerable questions based on the uniqueness of their answers. This classification helps understand the nature of the question and the potential diversity of valid responses.
For answerable questions: 1. Unique Answer: There is a single, specific correct answer that is widely accepted based on current human knowledge. Questions about fictional characters are most likely "Unique answer" because most of the times we can assume the answer is in the book / movie 2. Multiple Valid Answers: There are several plausible, valid answers that could be considered correct depending on perspective, context, or interpretation. Multiple valid answers means either (1) there is a subjective judgment involved (the answer varies depending on who answers), or (2) it is a creative task that can be solved in several different ways, or (3) we as humans do not have access to a unique correct answer. Trivial different wordings of the same concept do not qualify as "Multiple answers". e.g. "What's the meaning of bucolic?"
Ambiguous questions are not necessarily "multiple answers" just because they are ambiguous (i.e. since different people could understand the query differently, they might give different answers). In such cases we should assume a meaning for the question, and reason about the possible ways to answer it.
For each given hypothetically answerable question, classify it as either: 1. Unique Answer 2. Multiple Valid Answers
If the question is looking for a list of items, but the list is unique and well-defined, it should be classified as "Unique Answer". If the question is looking for a number which can be found or computed, it should be classified as "Unique Answer". If the question is asking how something can be achieved, and there is only one way to achieve it, it should be classified as "Unique Answer", and "Multiple Valid Answers" otherwise.
Examples
1. Q: "What is the chemical symbol for gold?" Classification: Unique Answer Ex-

planation: There is a single, universally accepted answer in chemistry: Au.
2. Q: "What is the best programming language for web development?" Classification: Multiple Valid Answers Explanation: There are several programming languages suitable for web development, and the "best" can depend on project requirements, developer preference, and other factors.
3. Q: "Who was the first president of the United States?" Classification: Unique Answer Explanation: There is a single, historically accepted answer: George Washington.
4. Q: "Tell me some short bedtime stories" Classification: Multiple Valid Answers Explanation: There are various short stories that can be told at bedtime, and the choice can vary based on cultural background, personal preference, and other factors.
5. Q: "What is the meaning of life?" Classification: Multiple Valid Answers Explanation: This question has multiple valid answers based on different philosophical, religious, and personal perspectives.
6 Q: "What does it mean when an economy is in a recession?" Classification: Unique Answer Explanation: There is a specific definition of a recession in economics, making this a question with a unique answer.
7 Q: "Name the three primary colors" Classification: Unique Answer Explanation: There are three primary colors in the RGB color model: red, green, and blue.
8 Q: "What criteria should I consider when buying a new laptop?" Classification: Multiple Valid Answers Explanation: The criteria for buying a laptop can vary based on individual needs, preferences, and budget constraints.
9. Q: "Who is the best neurosurgeon in New York?" Classification: Multiple Valid Answers Explanation: The best neurosurgeon can vary based on specialization, patient reviews, and other factors.
Please classify the following question:
question

C Distributions

We report here the full distributions of labels in *NatQuest*.

Table 14: Distribution of causal question types in the dataset.

Causal Type	Count	Percentage
About Mechanisms	2,942	51.58%
About Variables	2,073	36.36%
About Relations	442	7.75%
About Average Effects	244	4.28%

D Experimental Details

D.1 OpenAI Experiments and Labeling

The total amount spent on the OpenAI API was around 200 dollars. When possible, the BatchAPI was used, to save on expenses. Several models were used across this work, depending on the availability at the moment of each analysis, and

Hyperparameters:

- *seed*: 42
- *temperature*: 1
- *max tokens*: 1000

We used the OpenAI API endpoint for prompt engineering, testing, and gathering answers, and the BatchAPI to label the full dataset.

D.2 Trainings

Learning Rate

The learning rate was always set to 2×10^{-4}.

FLAN-T5 LoRA Configuration

Listing 1: FLAN-T5 LoRA Configuration

```
lora_config = LoraConfig(
    r=32,
    lora_alpha=32,
    target_modules=["q", "v"],
    lora_dropout=0.05,
    bias="none",
    task_type=TaskType.SEQ_2_SEQ_LM
)
```

PHI LoRA Configuration

Listing 2: PHI LoRA Configuration

```
lora_config = LoraConfig(
    r=32,
    lora_alpha=16,
```

Table 15: Distribution of cognitive complexity across overall dataset, non-causal questions, and causal questions. Significant increases in causal questions are highlighted in bold.

Cognitive Complexity	Overall	Non-Causal	Causal
Remembering	36.82%	56.75%	9.54%
Understanding	13.47%	11.54%	16.11%
Applying	13.54%	3.67%	**27.04%**
Analyzing	8.82%	5.04%	13.98%
Evaluating	13.74%	9.09%	**20.11%**
Creating	13.62%	13.90%	13.23%

Table 16: Distribution of cognitive complexity across different data sources. Notable values are highlighted in bold.

Cognitive Complexity	Overall	H-to-SEs	H-to-H	H-to-LLMs
Remembering	36.82%	**76.49%**	19.42%	14.52%
Understanding	13.47%	13.82%	15.27%	11.32%
Applying	13.54%	4.07%	18.73%	17.81%
Analyzing	8.82%	3.40%	13.00%	10.05%
Evaluating	13.74%	1.87%	**31.27%**	8.09%
Creating	13.62%	0.36%	2.31%	**38.20%**

```
    target_modules="all-linear",
    lora_dropout=0.05,
    bias="none",
    task_type=TaskType.CAUSAL_LM
)
```

Batch Size and Accumulation Details

- **PHI:** Batch size = 4, Accumulation = 1
- **XL:** Batch size = 1, Accumulation = 4
- **Large:** Batch size = 1, Accumulation = 4
- **Base:** Batch size = 1, Accumulation = 4
- **Small:** Batch size = 4, Accumulation = 1

Additional plots

We report here additional plots on the performance of the fine-tuned classifiers.

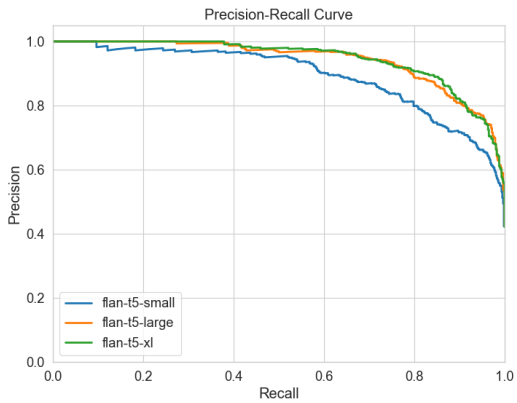

Figure 5: Combined Precision - Recall curve for the FLAN models.

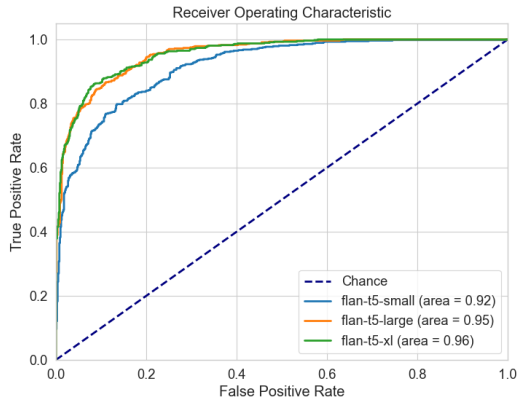

Figure 6: Combined ROC curve for the FLAN models.

E More examples

Here we show some additional examples from *NatQuest*.

E.1 Domain Classification

Table 20 shows more examples classified by domain.

Table 17: Comparison of domain classes, cognitive complexity, open-endedness, and user needs between Non-Natural Questions and CausalQuest datasets. Significant differences are highlighted in bold.

Category	Non-Natural Questions	*NatQuest*
Domain Class		
Natural and Formal Sciences	40.68%	7.74%
Society, Economy, Business	28.64%	18.79%
Everyday Life and Personal Choices	7.74%	**16.00%**
Computer Science	6.64%	**25.83%**
Historical Events and Hypothetical Scenarios	5.90%	4.48%
Arts and Culture	3.83%	**9.41%**
Health and Medicine	3.77%	**10.91%**
Psychology and Behavior	2.80%	**6.83%**
Bloom Taxonomy		
Applying	**36.41%**	13.54%
Remembering	30.51%	36.82%
Analyzing	13.90%	8.82%
Evaluating	11.54%	13.74%
Understanding	7.47%	13.47%
Creating	0.17%	**13.62%**
Open-Endedness		
Unique Answer	**70.43%**	32.30%
Multiple Valid Answers	29.57%	**67.70%**
User Needs		
Knowledge and Information	**73.70%**	57.09%
Problem-Solving and Practical Skills	24.63%	14.77%
Professional and Social Development	0.70%	**8.42%**
Personal Well-being	0.60%	**5.88%**
Leisure and Creativity	0.37%	**13.84%**

Table 18: Distribution of causal and non-causal questions across different data sources. Notable values are highlighted in bold.

Source	Non-Causal	Causal
Overall	57.77%	42.23%
H-to-SEs	**78.07%**	21.93%
H-to-H	41.29%	**58.71%**
H-to-LLMs	53.96%	46.04%

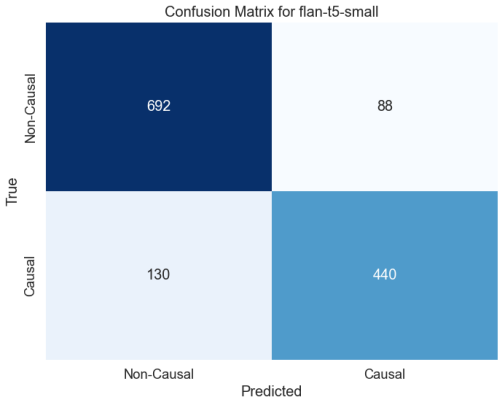

Figure 7: Confusion matrix for FLAN-T5-Small.

Table 19: GPT models used for various tasks.

Task	GPT Model
Summarisation	gpt-3.5-turbo-0125
Causality classification	gpt-4-turbo-2024-04-09
All other classifications	gpt-4o-mini-2024-07-18
Sect 5.1	gpt-4o-2024-05-13

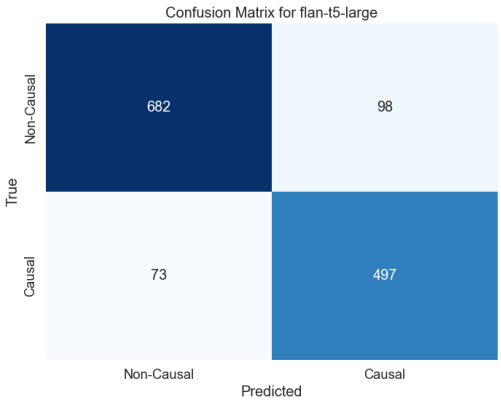

Figure 8: Confusion matrix for FLAN-T5-Large.

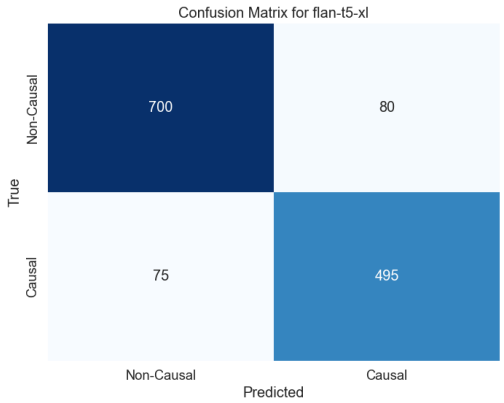

Figure 9: Confusion matrix for FLAN-T5-XL.

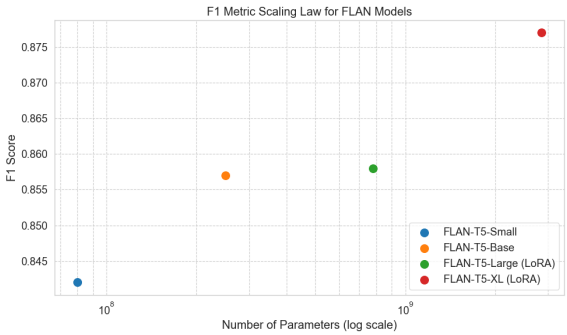

Figure 10: F1 score scaling law, FLAN models.

E.2 Subjectivity

Table 21 shows more examples classified by subjectivity.

F Annotators' instructions

We report here the instructions given to annotators when manually labeling the data.

The annotators were given the exact same prompts as GPT. They were given the first iteration of the prompt engineering since the iterations were done by computing accuracy with respect to annotators' labels. Annotators were advised about the sources of the data considering that they could imply the potential presence of toxic / NSFW examples.

F.1 Annotation Instructions for Causality / Domain, Subjectivity / Cognitive Complexity / Needs / Open-Endedness

DISCLAIMER - Sensitive / Offensive content The questions were sourced from social media platforms, search engines, and conversational AI. This means that questions might contain offensive or sensitive content. By proceeding with the annotation task you are acknowledging this possibility and that you are comfortable reading and classifying the questions.

Task Description In the file {FILE_NAME}.xlsx there are {N} questions that were asked either on Quora, CharGPT, Bing, or Google. The task is to classify each question depending on its {FEATURE}. The categories are explained below, as they were explained to GPT. The reason we are doing this is to understand whether the classification that we made via GPT is reliable / humans would agree with them.

{PROMPTS}

F.2 Annotation Instructions for Answer Grading

In file {FILENAME} there are pairs of questions and answers. the questions were asked either on Quora, CharGPT, Bing, or Google. Judge the answers given the following rubric:

- Effectiveness: Does the answer complete the goal specific to this user? How far is it with respect to the ideal perfect answer? 0: useless, 5: perfect
- Efficiency: Is the answer going to the point, without useless waste of time of the user? 0: a lot of useless information provided, 5: the answer provides just the right amount of information
- Satisfaction: Is the answer leaving the user satisfied, being empathetic, friendly, and pleasant to read? 0: annoying, 1: very pleasant

BE AWARE: because of the sources of the data, the prompts can contain toxic, disturbing, or NSFW topics. If you do not feel comfortable with this, feel free to skip the question or stop the annotation.

Table 20: Examples of various classifications of questions based on their Domain class.

Domain Class	Question
Domain Class: Everyday Life and Personal Choices	
Everyday Life and Personal Choices	How can I go about finding an overseas T-shirt manufacturer for my clothing business?
Everyday Life and Personal Choices	How do I care for raw denim?
Domain Class: Computer Science	
Computer Science	Why was the question "What does Jimmy Wales think of Wikipedia" merged with "What does Jimmy Wales think of Wikipedia Redefined"?
Computer Science	Where can I get the design of CSR implementation on Block RAM?
Domain Class: Psychology and Behavior	
Psychology and Behavior	What does it mean when a girl says 'You're so affectionate.'?
Psychology and Behavior	What do dreams signify?
Domain Class: Health and Medicine	
Health and Medicine	Is it healthy to eat seedless fruits?
Health and Medicine	Why do you have swollen lymph nodes and what is the medicine for it?
Domain Class: Natural and Formal Sciences	
Natural and Formal Sciences	How is AC converted into DC?
Natural and Formal Sciences	What are the different parametric study done on the cable stayed bridge for dynamic loading?
Domain Class: Society, Economy Business	
Society, Economy Business	Why are people now more interested in pop music rather that the good old classic rock?
Society, Economy, Business	Is it true that when purchasing a vehicle, if the cost is under 7000 you do not need to have full coverage insurance?
Domain Class: Historical Events and Hypothetical Scenarios	
Historical Events and Hypothetical Scenarios	What if Swami Vivekanand didn't get his guru?
Historical Events and Hypothetical Scenarios	Will Facebook shut down by the end of 2016?
Domain Class: Other	
Other	What are some of the worst examples of plagiarism you've witnessed?
Other	Is Sanskrit considered a religious language?

Table 21: Examples of questions classified by their Subjectivity.

Is Subjective	Question
Subjectivity: False	
False	What is the origin of ES-IS and IS-IS? Is there any connection to the OSI model?
False	Why does Hilary Clinton cough so much?
Subjectivity: True	
True	I got a BigData internship and I want to get an internship in a bigger company next year. What skills should I spend my time on this summer?
True	I am in college and it seems like every guy ignores me. Is there something wrong with me?