

Researchy Questions: A Dataset of Multi-Perspective, Decompositional Questions for LLM Web Agents

Corby Rosset
Microsoft

Ho-Lam Chung
National Taiwan Univ.

Guanghui Qin
Johns Hopkins Univ.

Ethan C. Chau
Microsoft

Zhuo Feng
Microsoft

Ahmed Awadallah
Microsoft

Jennifer Neville
Microsoft

Nikhil Rao
Microsoft

Abstract

Existing question answering (QA) datasets are no longer challenging to most powerful Large Language Models (LLMs). Traditional QA benchmarks like TriviaQA, NaturalQuestions, ELI5 and HotpotQA mainly study “known unknowns” with clear indications of both what information is missing, and how to find it to answer the question. Hence, good performance on these benchmarks provides a false sense of security. A yet unmet need of the NLP community is a bank of non-factoid, multi-perspective questions involving a great deal of unclear information needs, i.e. “unknown unknowns”. We claim we can find such questions in search engine logs, which is surprising because most question-intent queries are indeed factoid. We present **Researchy Questions**, a dataset of search engine queries tediously filtered to be non-factoid, “decompositional” and multi-perspective. We show that users spend a lot of “effort” on these questions in terms of signals like clicks and session length, and that they are also challenging for GPT-4. We also show that “slow thinking” answering techniques, like decomposition into sub-questions shows benefit over answering directly. We release¹ ~ 100k Researchy Questions, along with the Clueweb22 URLs that were clicked.

1 Introduction

The advent of Large Language Models (LLMs) has ushered in a new era for the field natural language processing, with short- and long- form question-answering (QA) at the forefront of recent achievements (OpenAI et al., 2023). Historically, QA benchmarks served as the crucible for evaluating a model’s grasp of natural language understanding. However, LLMs have nearly perfected many QA datasets, particularly those involving answering short, factual questions like “What is the capital

Dataset	Num.	Topics	Sub-Ques.	Sub-Query
Hotpot QA	300	2.9	3.8	3.6
OpenBook QA	300	3.8	6.3	5.9
Strategy QA	300	3.8	5.3	4.9
Truthful QA	300	3.8	6.4	6.0
Aquamuse	300	3.7	5.4	5.2
Reddit/askh	300	4.9	9.4	8.5
Reddit/asks	300	5.1	9.2	8.8
Reddit/eli5	300	4.5	9.7	9.3
Stack Exchange	300	6.1	8.4	7.6
Wikihow	300	4.8	11.7	11.2
Researchy	96k	3.9	14.3	12.6

Table 1: We asked GPT-4 to decompose questions into a natural hierarchy of subquestions, as well as concrete queries it would issue to a search engine. As expected, HotpotQA requires the fewest sub-questions to answer.

of Brunei?”. The re-emergence of chatbots and “agentic” AI assistants, which represent complex systems based on LLMs as a building block, have led to new opportunities for users to ask deeper, more nuanced questions as shown in Figure 1.

However, the capabilities of powerful AI-assisted QA tools has outpaced the metrics needed to evaluate them. A number of traditional QA benchmarks like Natural Questions (Kwiatkowski et al., 2019), TriviaQA (Joshi et al., 2017), WebQuestions (Berant et al., 2013), SearchQA (Dunn et al., 2017) have been more or less solved by modern LLMs. These datasets consist primarily of factoid questions (from search engine logs, Trivia, Jeopardy!, etc), where answers are typically found within a single sentence or paragraph that almost certainly exist in standard pre-training web corpora (Zhou et al., 2023). For instance, MS Marco (Campos et al., 2016) is up to 55% factoid questions (Bolotova et al., 2022). The obvious drawback to these kinds of QA datasets is that the answers can be memorized by LLMs or solved with simple pattern-matching or keyword-based search.

Multi-hop reasoning tasks including HotpotQA (Yang et al., 2018), HybridQA (Chen et al., 2020), MuSiQue (Trivedi et al., 2022) aim to challenge

¹https://huggingface.co/datasets/corbyrosset/researchy_questions

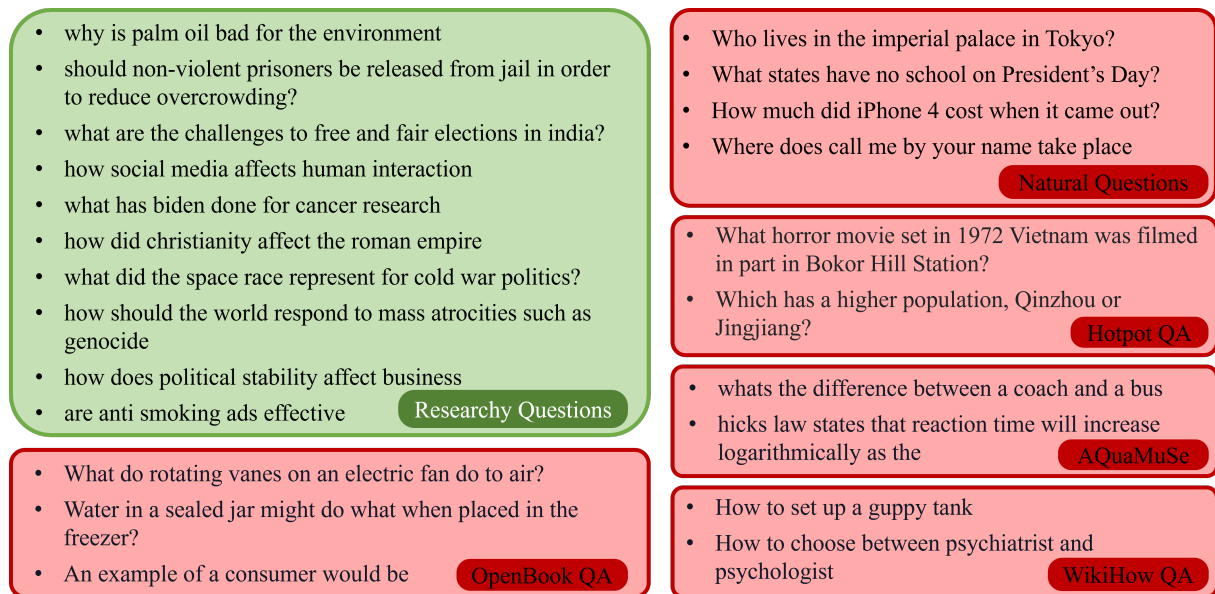


Figure 1: Qualitative comparison of how Researchy Questions differs from other Question Answering datasets. Researchy Questions involve a greater deal of complexity and “unknown unknowns” than other QA datasets.

QA systems to logically bridge information across multiple documents or paragraphs. While these datasets have made strides in increasing question complexity, **the answers are still ultimately factoid, and it’s clear what sub-questions ought to be asked to recall the missing information.** Furthermore, the construction of these datasets (e.g. synthetically generated from paths of wikipedia links) leads to a distribution mismatch w.r.t questions humans ask.

There exist several sources of long-form, non-factoid QA datasets like ELI5 (Fan et al., 2019), Stack Exchange, Yahoo Answers (Zhang et al., 2016), and WikiHowQA (Bolotova-Baranova et al., 2023). While answers to these questions are more involved than factoid questions, ELI5 and WikiHowQA elicit more expository answers than analytical ones. The Oxford Allsouls dataset (Liu et al., 2023b) contains 1k college-level essay prompts which are multi-perspective, but are designed to evaluate persuasive writing skills and do not have associated documents to ground responses. AquaMuSe (Kulkarni et al., 2020) is an excellent attempt at filtering Natural Questions (NQ) for multi-faceted queries, but their method was constrained by relying on the relatively short paragraph-length answers already present in NQ.

The rise of “LLM Agents” e.g. (Wu et al., 2023b) have opened the door for even deeper collaboration between users, LLMs, and tools. In response, more recent datasets have focused on completing challenging tasks in open-ended envi-

ronments with tools like a web browser, file system, database, etc.

In particular, Gaia (Mialon et al., 2023) tests understanding of multi-modal inputs (images and text), as well as complex reasoning across levels of difficulty to solve. AgentBench (Liu et al., 2023c) provides closed environments for an LLM to interact with APIs in various scenarios, including coding (interacting with file systems or databases), games/puzzles, and web browsing/shopping. While these datasets advance the field of metrics for LLM agents, they are small, consisting of only 466 and 1,091 questions respectively that were hand-curated by the authors.

The need for more challenging QA datasets also stems from some alarming trends: while there are hundreds of public LLMs, they are pretrained on only a handful existing corpora (Gao et al., 2020; Raffel et al., 2023), or distilled from one of a few teacher LLMs (Peng et al., 2023). Furthermore, much more data scraped from the internet for training is itself going to be AI-generated content, leading to an echo-chamber effect (Dohmatob et al., 2024; Wu et al., 2023a). **Hence, the convergent evolution (Stayton, 2015) of LLMs increases the risk that they will not recognize that they don’t know something,** e.g. see GPT-4 and Mixtral 8x7b in Table 10. This is especially true when LLMs act in the capacity of LLM-as-a-judge (Zheng et al., 2023a; Yuan et al., 2024), or when faced with very multi-faceted/multi-

Researchy Question: how does public transport help the economy	
Hierarchical Decomposition of the Question	Clicked Clueweb22 URLs
<ol style="list-style-type: none"> What is public transport? <ol style="list-style-type: none"> What are the different types of public transport? How many people use public transport in different regions or countries? What are the direct economic benefits of public transport? <ol style="list-style-type: none"> How public transport reduces costs for users e.g. fuel, parking, maintenance, etc.? How does public transport generate revenue... What are the indirect economic benefits of public transport? <ol style="list-style-type: none"> How does public transport reduce congestion... How does public transport increase access to education, employment, health, ... How does public transport enhance productivity and innovation... How does public transport contribute to environmental and social goals... How do the economic benefits of public transport compare to the costs of providing and maintaining public transport? <ol style="list-style-type: none"> What are the main costs of public transport... How are the costs of public transport financed... How are the benefits and costs of public transport measured and evaluated... 	<ol style="list-style-type: none"> infrastructureusa.org nationalgeographic.org quora.com accessmagazine.org ced.berkeley.edu greenertransportsolutions bts.gov apta.com
<p>Example of a Pivotal Fact from URL: accessmagazine.org</p> <p>“...Even in cities with a high concentration of office space in the [central business district], we estimate that increasing transit ridership by 10 percent will increase office rents by no more than 0.5 percent. For all other cities, we estimate that increasing transit ridership will have no effect on office rents...”</p>	

Table 2: An example Researchy Question, how GPT-4 decomposed it into sub-questions (closed-book), as well the URLs in ClueWeb22 that real users clicked on, and an example of a **Pivotal Fact** from one of those URLs.

perspective questions, where the consequences could be that users “miss the whole picture” or worse, are misled (Zheng et al., 2023b; Liu et al., 2023b). While retrieval-augmentation (Lewis et al., 2021; Borgeaud et al., 2022; Guu et al., 2020) can help complement LLM agents, the risk is merely shifted to whether a sub-system retrieves the *right* information and uses it correctly (Liu et al., 2023a).

We believe the well-studied phenomenon of “**unknown unknowns**” (United States Congress et al., 1981) applies to LLM Agents in scenarios addressing complex questions requiring “slow thinking” (Kahneman, 2011). Simply put, one strategy is to iteratively re-frame or decompose the problem into a set of “**known unknowns**” (which characterize most of the aforementioned QA datasets). For these sub-problems, it should be clearer what information is missing, how to find it, and once found, how the “known known” contributes to the *final answer*. Several techniques such as chain-of-thought question decomposition (Radhakrishnan et al., 2023) and tree-of-thought (Yao et al., 2023a) prompting take a similar approach to plan long-horizon solutions to complex problems. However, those studies still operate over traditional QA benchmarks like HotpotQA, or over simple games like crossword puzzles. Hence, the right benchmark of questions for these advanced decomposition techniques still does not exist for open-domain

web scenarios (Krishna et al., 2021).

We present Researchy Questions to study the dynamics of how LLM agents handle unclear information needs associated with very complex questions. We define a **Researchy Question** as a *non-factoid* question that expects a *long-form answer* (longer than a paragraph!) entailing substantial research or effort to synthesize. A Researchy Question can be instantiated as a complex search task (Aula and Russell, 2008) with unclear information needs that requires analyzing *multiple documents* or pieces of evidence. A Researchy Question does not have a single correct answer, but rather *multiple perspectives* allowing a dense manifold of answers over which varying criteria can determine which is better. In practice, the act of answering a Researchy Question probably involves *decomposition* into sub-questions that aid the retrieval of comprehensive information, reducing the risk of *missing unknown unknowns*. Lastly, a Researchy Question represents a *genuine information need* that real people asked. Figure 1 qualitatively compares other canonical QA datasets.

Researchy Questions is primarily a QA dataset to evaluate question answering systems or LLM Agents with the ultimate goal of achieving ever higher-quality answers using any tools necessary. Notwithstanding, it is also a search/retrieval dataset in the sense that finding and correctly incor-

First	MS Marco	QnA	Non-factoid	Researchy (↓)
how	17.0%	34.3%	29.4%	41.2%
why	1.64%	6.26%	33.4%	22.9%
what	34.9%	12.2%	21.3%	19.1%
is/are/do	5.77%	15.0%	6.50%	9.67%
should	0.11%	0.53%	0.59%	1.85%
can	1.84%	4.31%	1.01%	0.97%
who	3.27%	4.77%	0.90%	0.47%
which	1.78%	2.78%	1.87%	0.46%
when	2.70%	5.03%	0.44%	0.43%
pros	< 0.01%	0.03%	0.45%	0.28%
explain	0.05%	0.06%	0.12%	0.23%
where	3.54%	4.08%	0.37%	0.17%
factors	< 0.01%	0.01%	0.08%	0.15%
will	0.10%	0.69%	0.08%	0.15%
describe	0.05%	0.07%	0.04%	0.09%

Table 3: Top most-prevalent first-words of the three main stages of our question filtering funnel (with MS Marco for comparison).

porating the right evidence is a critical sub-system to meet expectations of trustworthiness and groundedness (Zheng et al., 2023b; Liu et al., 2023b). While we believe that question decomposition is a critical piece of solving Researchy Questions, it is unclear how to define or measure the quality of sub-questions. To aid this endeavor, we reveal what URLs end users found useful, with the hope that good sub-questions will at least lead to the information found in those clicked documents.

We release about 96K Researchy Questions consisting of real users’ queries to a commercial search engine, and additionally:

1. The decomposition of the question into a 2-level hierarchical plan (See Table 2 Left).
2. For each question, the user-aggregated click distribution over URLs in a publicly available web corpus, ClueWeb22.
3. Ordered list of sub-queries corresponding roughly to the sub-questions that can be readily issued to a search engine

In Section 2 we describe how Researchy Questions were obtained and then characterize them in Section 3. In Section 4 we verify that web users expend more effort on Researchy Questions over other queries. In Section 5 we evaluate and compare compositional answering techniques from (Radhakrishnan et al., 2023).

2 Researchy Questions Construction

Researchy Questions are real user queries from search logs. While search logs have a rich diversity of query types and intents (Bolotova et al., 2022; Bu et al., 2010), they mostly contain factoid or navigational queries, which need to be filtered.

2.1 Stage 1: Mining Search Logs

We obtained a set of query-url click pairs from a commercial search engine that were logged between July 2021 and August 2022, which maximally overlaps with the creation of the Clueweb22 snapshot of web documents (Overwijk et al., 2022). This way, we can simply indicate which Researchy Questions clicked on which documents. We began with a large sample of english, non-adult queries that had least one click. We denote these as “General Queries”, which were further filtered.

An important filtering criteria is frequency: we kept queries if they were issued at least 50 times in the logs. This criteria is simple but powerful: it helps denoise the dataset (reduces spelling errors), and also allows us to focus on questions that are not “one-offs”. This helps us glean insights into repeated user behaviors as they interact with the search engine.

In order to select queries which are answer-seeking intent (i.e. actual “questions” as distinct from navigational queries like “facebook login”, shopping intent like “fast running shoes”, or local intent like “best pizza near me”), we used a suite of rules and existing production classifiers:

- Query Language: English
- Adult intent: False
- Number of Distinct Occurrences: ≥ 50
- $3 \leq \text{Number of Query Words} \leq 15$
- Number of Distinct URLs Clicked On: ≥ 2
- Question Intent Classifier: True
- Navigational Intent: False
- Local / Real Estate / Map Intent: False
- Retail/Shopping Intent: False
- Coding/Technical Intent: False
- Health/Medical Intent: False
- Triggered Possible Answer Cards: ≥ 1
- Triggered lots of Ads: False

To explain some of the points above: an Answer Card is a high-precision feature in a search engine where a paragraph containing the answer is shown at the top of the results page, distinct from the “ten blue links”. Because search engines are constantly being updated with new features and triggering rules, all the above statistics are aggregated and normalized over the full year period. For instance, “a lot of Ads” is determined by summing the total number of ads shown for the query over the full year and dividing by the number of times the query was issued, and then picking a threshold above which queries seemed “shopping intent”. The Ads

Query Type	Count	Avg. unique clicked urls	Avg. clicks	Avg. sat. clicks
General Queries	$\geq 1B$	1.88	4.83	2.54
QnA Queries	15.7M	3.99	9.31	5.10
Non-factoid QnA	1.0M	4.20	8.99	4.86
Researchy Queries	100k	6.31	15.85	8.54

Session Type	Count	# Turns	# unique. queries	# sat clicks
General Sessions	$\geq 10B$	2.42	2.11	0.76
QnA Sessions	$\geq 100M$	6.28	5.53	1.15
N.F. QnA Sessions	$\geq 10M$	12.89	11.33	1.91
Researchy Sessions	$\geq 1M$	13.45	11.81	2.46

Table 4: Our query filtering funnel; each row is a subset from the row above it. (Left) Researchy Questions have much more clicks (Sat clicks have longer dwelltime), and require more unique documents; the full distribution is in Figure 2. (Right) Search statistics *per session in which each type of query appeared*, indicating that harder questions appear in longer sessions. These behaviors confirm our filtering yields more complex questions.

requirement also helped catch any shopping intent queries missed by the Retail intent classifier.

We wanted to remove Coding/Technical queries because such questions are often motivated by very specific problems that are often solved by one piece of documentation after lots of clicking around, which is not the behavior we wish to target in this dataset. Health and medical questions are avoided largely because they often overlap too much with those that ought to be addressed by a licensed medical professional. Many shopping / retail queries could be construed as “researchy” e.g. “what are the best headphones”, but we avoid them in this dataset because it is hard to distinguish whether a URL was clicked due to aggressive advertising or a real information need.

After this stage of filtering, we arrived at 15.7M “QnA Queries” which are largely recognizable as natural language questions that deal with open-domain knowledge. This size was manageable enough to efficiently run our own bert-large scale classifiers we use in the next stage of filtering.

2.2 Stage 2: Factoid Classifier

We needed a way to distinguish which QnA Queries were factoid vs. non-factoid; we trained a binary classifier on auto-labeled data for this purpose. The training data was a uniform sample of 200k questions from the 15.7M QnA Queries. Labels for the questions were collected from gpt3 (text-davinci-003) prompted with few-shot examples as shown in Figure 4. The labels were then used to train a bert-large non-factoid question classifier, which was then inferenced on the full set of 15.7M queries. By manual inspection, a threshold was chosen above which we were satisfied that the questions were meaningfully non-factoid. The resulting 1.0M met the non-factoid threshold of 0.75 as shown on the left hand side of Figure 3, which we denote as “Non-factoid QnA Queries”.

2.3 Stage 3: Decompositional Classifier

Not all the resulting non-factoid QnA queries exhibited the “decompositional” information needs. Namely, they often looked expository or “how-to” with generally one correct answer which did not have many perspectives. We trained a second classifier to score the extent to which a question requires asking sub-questions. The specific definition of “requiring sub-questions” is described in a prompt in Figure 5 given to ChatGPT (gpt-35-turbo) to collect labels. We used ChatGPT because we figured this was a relatively intense cognitive task. We inferenced ChatGPT on about 40k outputs from the non-factoid classifier that met the non-factoid threshold of 0.75. We used the labels to train a separate bert-large “decompositional” classifier.

Again, we chose a threshold by manual inspection to indicate which of the 1.0M Non-factoid QnA Queries were also decompositional, which happened to be 0.6 as shown on the right-hand side of Figure 3. Out of the 1.0M queries that met the non-factoid threshold, 146k also met the decompositional threshold. These 146k became the Researchy Questions candidates *before* deduplication. This procedure is not perfect; in Table 6 we show examples of some non-factoid questions that did not meet the decompositional threshold.

2.4 Stage 4: Deduplication

The final step of filtering is deduplication. We take an agglomerative clustering approach (Everitt, 1974), where the only parameter is a distance threshold ϵ below which two queries are considered “duplicate intent”.

We represent the semantic intent of a query by an ANCE-based (Xiong et al., 2020) vector encoder $\vec{q}_i \leftarrow \text{encoder}(q_i)$. We instantiate a metric space defined by $1 - \text{cosine}(\cdot, \cdot)$ of the vector encodings using the faiss implementation of an approximate nearest neighbor (ANN) index (Johnson et al., 2019). For each question in the index, we search

for the nearest neighbors $\{q_j \sim \text{ANN}(q_i) \text{ s.t. } 1.0 - \vec{q}_i \cdot \vec{q}_j < \epsilon\}$. For agglomerative clustering, we define a “group” as a set of queries in which all pairwise distances are within ϵ . We found that about 63% of queries were singletons (did not belong to a group of size greater than one), and the average group size was 3.8. For example, the queries “*what were tanks used for in ww1*”, “*how were the tanks used in ww1*” and “*why were tanks needed in ww1*” were all part of the same group. For all groups of size greater than one, we selected the query that was issued most often in the logs as the representative “head” of the group. After combining the heads of groups and singletons, about 70% of queries remained, yielding 102k Researchy Questions. Although we did our best to de-duplicate question intent, some clusters of topics remain, e.g. a quick keyword count shows about 600 contain the string “ww2”, or about 80 for “supreme court”.

2.5 Stage 5: Final GPT-4 Filtering

As a final quality control step after deduplication, we had all 102k questions labeled by GPT-4 for intrinsic attributes of the question like how **multi-faceted** it is, how **reasoning-intensive** it is likely to be, etc. The full set of eight attributes is defined in Figure 7, along with histograms of those scores in Figure 6 for both Researchy Questions and Natural Questions. All eight are scored on a scale of 1-10. About 3% of 102k questions were removed based on the attributes “**ambiguous**” and “**incomplete**” which indicate defective questions too difficult to answer; some examples are shown in Table 7. Another 2% were removed for being too “**assumptive**”, meaning the question was phrased in a presumptuous way that was likely to bias the answer as shown in Table 8. Another 2% was removed for safety reasons as shown in Table 9, where we deem risk of harm too high by attempting to answer the question. Not all “assumptive” questions are harmful per se. Finally, the remaining 96k queries are what we release.

3 Characterizing Researchy Questions

Across the 96k Researchy Questions (split into 90k train, 6.4k test), there are in aggregate 350k unique documents clicked, 48% of which can be found in the english subset of Clueweb22 Set B (Overwijk et al., 2022); the rest are in Sets A or L. For each question we release, there are on average 4.9 +/- 3.5 clicked documents (See Figure 2

Right), indicating a good diversity of information needs and much higher than the average query overall. Conversely, for each document there is only 1.4 +/- 2.3 associated Researchy Questions (See Figure 2 Left) indicating good query de-duplication.

To get a sense of how intrinsically difficult Researchy Questions are compared to other datasets, we asked GPT-4 how many sub-questions or search engine queries would be necessary to ask/issue in order to fully answer each question. An example decomposition is shown in Table 2, and the aggregate results are shown in Table 1. Clearly, GPT-4 thinks that most factoid QA datasets (top of the table) require the fewest sub-questions to answer, whereas Researchy Questions require the most even among longer-form QA datasets.

We also compared Researchy Questions to another search-log based QA dataset – Natural Questions (Kwiatkowski et al., 2019) – along the 8 quality dimensions described in Section 2.5, such as how reasoning- and knowledge-intense they are. The comparative histograms are shown in Figure 6. Clearly, GPT-4 thinks that Researchy Questions require much more knowledge, reasoning, and are inherently more multi-faceted.

Table 3 shows the distribution of first words across Researchy Questions (as well as intermediate datasets used in the filtering funnel described in the next section). For comparison, MS Marco queries (also from web search logs) are much more factoid – e.g. only 1.64% of them start with “why” (Bajaj et al., 2018).

Lastly, an emergent property of Researchy Questions we observed is that some information found in the clicked URLs is very surprising, e.g. for the question “*should the death penalty be legalized*”, the fact that in the US, “*the cost of enforcing the death penalty cost millions more than life imprisonment*”² would have a large impact on the economic arguments in that answer. We define a **Pivotal Fact** as a piece of information that is so surprising and consequential that, once known, drastically changes how an LLM Agent would answer the question (and the answer wouldn’t be as good without it); but it wouldn’t know about it unless it asked the right sub-question to retrieve it, e.g. “*does the death penalty cost more than life imprisonment*”. Hence, a pivotal fact is the information analogue of a black swan, a critical event that is hard to predict but highly impactful (Taleb, 2008);

²supremecourt.gov/opinions

Method	Samples	Direct Answer		CoT decomp.		Factor decomp.		Decomp.
		Acc.	Score	Acc.	Score	Acc.	Score	Score gain
Hotpot QA	300	0.843	83.4	0.877	83.5	0.837	81.3	+0.1
OpenBook QA	300	0.926	86.1	0.843	83.5	0.750	80.7	-2.6
Strategy QA	300	0.757	80.8	0.810	83.7	0.777	82.6	+2.9
Truthful QA	300	0.703	73.7	0.789	82.4	0.739	81.5	+8.7
Aquamuse	300	0.916	83.0	0.940	84.9	0.926	85.0	+2.0
Reddit/askh	300	0.759	79.8	0.736	77.3	0.732	79.3	-0.5
Reddit/asks	300	0.783	81.1	0.743	79.4	0.796	82.7	+1.6
Reddit/eli5	300	0.883	83.1	0.890	85.0	0.890	86.5	+3.4
StackExchange	300	0.717	78.4	0.599	70.7	0.628	73.4	-5.0
Wikihow QA	300	0.93	82.9	0.937	84.4	0.950	88.2	+5.3
Researchy Questions	1k	N/A	82.7	N/A	84.6	N/A	88.3	+5.6

Table 5: Comparison of various question decomposition techniques to prompt GPT-4 as an answer module. The top half of the table is short-form questions, the bottom half is longform questions.

another example is shown in the bottom of Table 2. Although we don’t have a good way to quantify the prevalence of **Pivotal Facts**, we believe that Researchy Questions is the best dataset to study the dynamics of how LLM Agents search for, and respond to, these unknown unknowns.

4 Agreement with User Search Behavior

More complicated questions ought to require more effort to answer (Kelly et al., 2015). We can approximate the amount of effort users expend in terms of behavior signals like clicks and session length.

In Table 4 (Left) we show aggregate click statistics for each subset of queries. The results show that Researchy Questions, which are both non-factoid and compositional, lead to more in-depth consumption (clicks and sat-clicks) of more diverse information (unique urls), agreeing with prior work (Hassan et al., 2014).

In Table 4 (Right) we show behavioral signals exhibited by users at the session level rather than individual click level. For instance, if a QnA-type query appeared at any time in any session in the date range, the whole session is included in the “QnA Sessions” row. The results clearly show users were twice as engaged answering non-factoid questions than sessions devoted to factoid-y intents, and six times longer than the average session overall.

5 Evaluating Answer Techniques to Researchy Questions

Since there is no one “correct” answer to a Researchy Question, we contend that they ought be evaluated in a relative, side-by-side fashion *a la* Alpaca-Eval style (Li et al., 2023), with e.g. the “closed-book” answer as the reference.

Since Researchy Questions are intended to be answered by decomposing them into sub-questions, we evaluate two decompositional question answering answer techniques – chain-of-thought decomposition, and factored decomposition – against the direct answering baseline. Factored decomposition makes a separate call to an LLM for each sub-question independently, and then a final “recomposition” call to synthesize the main answer (Radhakrishnan et al., 2023).

Table 5 shows the results of a side-by-side auto-eval of the three answering techniques across a range of datasets. The LLM providing the answers was GPT-4, and the LLM-as-a-judge was also GPT-4, prompted separately using the template in Figure 8. Since the gold correct answers exist for many of the datasets we consider, the judge was asked to ascertain “Accuracy” as a binary score of whether the candidate answer was consistent with the gold. The “Score” is on a scale of 1-100 indicating the overall quality. The top half of Table 5 corresponds to short-answer datasets (where Accuracy is more important), while the bottom half is longer answer questions where the overall Score is better suited to evaluate. Out of the long-form answer datasets, Researchy Questions benefits the most from decomposition techniques.

We draw several conclusions from Table 5. Firstly, we largely confirm the results of (Radhakrishnan et al., 2023) that decomposition techniques improve the accuracy on short-form multi-hop datasets over zero-shot direct answering. Secondly, factored decomposition leads to the highest scores for long-form questions, particularly questions involving reasoning about complex processes like Wikihow and Researchy Questions. On the other hand, chain-of-thought decomposition may

be better suited to deduce correct answers to more logic-intensive questions. **We believe the results for Researchy Questions using factored decomposition would be much higher if we incorporated retrieved information.**

6 Related Work

6.1 Human Behavior in Search Sessions

There has been some fundamental research into understanding user behavior in search sessions, from user studies (Kelly et al., 2015) to large scale click log evaluations (Hassan et al., 2014). The latter attempt to ascertain signals that could indicate whether a user is “struggling” or “exploring” in their search session; we use many of the same signals. Namely, they conclude that “exploring” sessions contain more clicks because users want to locate information for multiple facets of a topic. We concur with their results, e.g. non-factoid questions involve more clicks than factoid ones. Similarly, the definition of a “complex search task” is almost congruent to that of a Researchy Question in its information seeking behaviors [Aula and Russell](#). Other studies try to identify complex search tasks and offer recommended sub-tasks ([Hassan Awadallah et al., 2014](#); [Zhang et al., 2021](#)). Much of the intuition on how to filter Researchy Questions stems from these and similar studies on user search behavior.

6.2 Iterative Retrieval-Augmented Generation

Many papers discuss variants of chain-of-thought ([Wei et al., 2023](#)) adapted to decomposing multi-faceted questions such as query refinement prompts ([Amplayo et al., 2022](#)), Decomposed Prompting ([Khot et al., 2023](#)) and ReAct ([Yao et al., 2023b](#)). A step further is to marry a generative LLM with a vector-based retrieval system, e.g. ([Ren et al., 2023](#); [Xiong et al., 2020](#); [Karpukhin et al., 2020](#); [Izacard and Grave, 2021](#)). Such methods are instantiated in many flavors: Self-Ask ([Press et al., 2023](#)) was one of the first chain-of-thought-like decomposed prompting techniques to integrate a search engine, forcing an LLM to iteratively ask and subquestions. IRCot interleaves retrieval with chain-of-thought such that what to retrieve next depends on what was previously retrieved ([Trivedi et al., 2023](#)) Iter-RetGen iteratively retrieves and generates candidate answers to use as inputs to the next phase of retrieval ([Shao et al., 2023](#)), while Beam Retrieval maintains run-

ning hypotheses of the relevant passages at every step ([Zhang et al., 2023](#)).

6.3 Agentic Question Answering

There are several “agentic” frameworks to facilitate the dynamic interaction between tools e.g. retrieval systems and LLMs-as-agents which are adept at iterative retrieval-augmented QA tasks. One is Demonstrate-Search-Predict ([Khattab et al., 2023](#)), another is AutoGen ([Wu et al., 2023b](#)). Some existing Web-based Agents include WebGPT ([Nakano et al., 2022](#)) which imitate how users surf the web and ask queries, others like WebAgent ([Gur et al., 2023](#)) interact with the web programmatically by understanding the raw HTML.

There are also several consumer-facing products for agentic search assistants such as Bing Chat ³, YouPro ⁴ (Research Mode) and SciPhi ⁵. **All of these systems clearly decompose queries into sub-questions and then retrieve/crawl the necessary pages to synthesize the final outputs.** An example of each addressing a Researchy Question is shown in Figures 9, 10, and 11, respectively.

7 Conclusion

For some time now, many QA datasets have been saturated, and many researchers have dismissed search logs as sources of complex questions. We present Researchy Questions, a large dataset to advance the field of complex question answering over multiple documents and perspectives, targeting LLM-assisted web search agents. We detail how these complex queries are mined from search logs, and confirm they involve significantly more effort than other types of search queries. We also provide some preliminary evidence that compositional answering techniques yield stronger results on Researchy Questions than direct answering.

By design, these questions don’t have a gold answer, so regrettably, it difficult to quantify the “headroom” of existing models, but qualitatively (e.g. Table 10) there seems to be substantial room for improvement. There also remains much work surrounding how to measure the quality of sub-questions, and studying how LLM agents find and interact with Pivotal Facts. We hope this dataset can help aid the development of new metrics and unlock new experiences for web users.

³<https://bing.com/chat>

⁴<https://you.com/search>

⁵<https://search.sciphi.ai/research>

Limitations

One of the primary limitations of this study is that while we argue question decomposition is key to solving complex Researchy Questions, we don't propose a good way to measure the quality of a candidate set of sub-questions. Instead, we loosely argue that good sub-questions ought to at least lead to the same set of information found in the documents users clicked on for the question. We admit these two things are not the same, but clicks are a powerful signal. We try to make up for this gap by releasing the hierarchical question/query decomposition that GPT-4 gives closed-book, but we do not evaluate whether that decomposition does indeed lead to the same set of gold documents in a real retrieval system e.g. built from the entirety of ClueWeb22.

We also admit that it would potentially be better to produce the question decomposition by “working backwards” from the documents that were clicked on – i.e. identify key information in the clicked document for a researchy question and then identify what sub-question would lead to that information being retrieved. On the other hand, one could argue that the “forward” direction is what needs to be implemented in practice. While we do not address these points in this study, others in the academic community can explore these topics more with the dataset we release.

Another limitation of this study is that Pivotal Facts are purely an observational phenomenon that we have in no way quantified. Future researchers could create a LLM-prompt to count the number of such statements in the clicked documents.

We regret that this dataset is not multi-lingual. This is due to the large amount of uncertainty and trial-and-error in curating this dataset, which meant frequent manual inspection of data. We believe the same framework can be used to construct a multi-lingual version of Researchy Questions.

Ethics Statement

We obtained approval to release this dataset after thorough IRB proceedings to ensure compliance with privacy, safety, and legal guidelines.

We would like to make a few notes: while it may appear that we tried to remove queries that seemed “controversial”, it is not our intention to act as moral or political regulators to decide if user queries are well-meaning. At the scale of web search, one will notice that users ask a lot of questions for a lot of

reasons, and it is beyond the scope of our job to speculate what the motives are behind a query. It is our job to assess whether the act of attempting to answer a question will lead to a reasonable risk of harm. Also, it is not the case that “GPT-4 is our moral authority” on the topic of safety, it is merely one of many tools we employ in a stack to ensure our internal requirements are being met.

References

- Reinald Kim Amplayo, Kellie Webster, Michael Collins, Dipanjan Das, and Shashi Narayan. 2022. [Query refinement prompts for closed-book long-form question answering](#).
- Anne Aula and Daniel Russell. 2008. Complex and exploratory web search.
- Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. [Ms marco: A human generated machine reading comprehension dataset](#).
- Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. [Semantic parsing on Freebase from question-answer pairs](#). In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.
- Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2022. [A non-factoid question-answering taxonomy](#). In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '22, page 1196–1207, New York, NY, USA. Association for Computing Machinery.
- Valeriia Bolotova-Baranova, Vladislav Blinov, Sofya Filippova, Falk Scholer, and Mark Sanderson. 2023. [WikiHowQA: A comprehensive benchmark for multi-document non-factoid question answering](#). In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5291–5314, Toronto, Canada. Association for Computational Linguistics.
- Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. [Improving language models by retrieving from trillions of tokens](#).
- Fan Bu, Xingwei Zhu, Yu Hao, and Xiaoyan Zhu. 2010. [Function-based question classification for general QA](#). In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*, pages 1119–1128, Cambridge, MA. Association for Computational Linguistics.
- Daniel Fernando Campos, Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, Li Deng, and Bhaskar Mitra. 2016. [Ms marco: A human generated machine reading comprehension dataset](#). *ArXiv*, abs/1611.09268.
- Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020. [HybridQA: A dataset of multi-hop question answering over tabular and textual data](#). In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.
- Elvis Dohmatob, Yunzhen Feng, Pu Yang, Francois Charton, and Julia Kempe. 2024. [A tale of tails: Model collapse as a change of scaling laws](#).
- Matthew Dunn, Levent Sagun, Mike Higgins, V. Ugur Guney, Volkan Cirik, and Kyunghyun Cho. 2017. [Searchqa: A new q&a dataset augmented with context from a search engine](#).
- Brian Everitt. 1974. *Cluster analysis*. Heinemann Educational [for] the Social Science Research Council.
- Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. [ELI5: long form question answering](#). In *Proceedings of the 57th Conference of the Association for Computational Linguistics, ACL 2019, Florence, Italy, July 28- August 2, 2019, Volume 1: Long Papers*, pages 3558–3567. Association for Computational Linguistics.
- Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. [The pile: An 800gb dataset of diverse text for language modeling](#).
- Izzeddin Gur, Hiroki Furuta, Austin Huang, Mustafa Safdari, Yutaka Matsuo, Douglas Eck, and Aleksandra Faust. 2023. [A real-world webagent with planning, long context understanding, and program synthesis](#).
- Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. [Realm: Retrieval-augmented language model pre-training](#).
- Ahmed Hassan, Ryen W. White, Susan T. Dumais, and Yi-Min Wang. 2014. [Struggling or exploring? disambiguating long search sessions](#). In *Proceedings of the 7th ACM International Conference on Web Search and Data Mining, WSDM '14*, page 53–62, New York, NY, USA. Association for Computing Machinery.
- Ahmed Hassan Awadallah, Ryen W. White, Patrick Pantel, Susan T. Dumais, and Yi-Min Wang. 2014. [Supporting complex search tasks](#). In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management, CIKM '14*, page 829–838, New York, NY, USA. Association for Computing Machinery.
- Gautier Izacard and Edouard Grave. 2021. [Leveraging passage retrieval with generative models for open domain question answering](#). In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*,

- pages 874–880, Online. Association for Computational Linguistics.
- Albert Q. Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, Gianna Lengyel, Guillaume Bour, Guillaume Lample, L  lio Renard Lavaud, Lucile Saulnier, Marie-Anne Lachaux, Pierre Stock, Sandeep Subramanian, Sophia Yang, Szymon Antoniak, Teven Le Scao, Th  ophile Gervet, Thibaut Lavril, Thomas Wang, Timoth  e Lacroix, and William El Sayed. 2024. [Mixture of experts](#).
- Jeff Johnson, Matthijs Douze, and Herv   J  gou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547.
- Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. [triviaqa: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension](#). *arXiv e-prints*, page arXiv:1705.03551.
- Daniel Kahneman. 2011. *Thinking, Fast and Slow*. Farrar, Straus and Giroux, New York.
- Vladimir Karpukhin, Barlas O  uz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. [Dense passage retrieval for open-domain question answering](#).
- Diane Kelly, Jaime Arguello, Ashlee Edwards, and Wanching Wu. 2015. [Development and evaluation of search tasks for iir experiments using a cognitive complexity framework](#). In *Proceedings of the 2015 International Conference on The Theory of Information Retrieval, ICTIR ’15*, page 101–110, New York, NY, USA. Association for Computing Machinery.
- Omar Khattab, Keshav Santhanam, Xiang Lisa Li, David Hall, Percy Liang, Christopher Potts, and Matei Zaharia. 2023. [Demonstrate-search-predict: Composing retrieval and language models for knowledge-intensive nlp](#).
- Tushar Khot, Harsh Trivedi, Matthew Finlayson, Yao Fu, Kyle Richardson, Peter Clark, and Ashish Sabharwal. 2023. [Decomposed prompting: A modular approach for solving complex tasks](#).
- Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. [Hurdles to progress in long-form question answering](#).
- Sayali Kulkarni, Sheide Chammas, Wan Zhu, Fei Sha, and Eugene Ie. 2020. [Aquamuse: Automatically generating datasets for query-based multi-document summarization](#).
- 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*.
- Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich K  ttler, Mike Lewis, Wen tau Yih, Tim Rock  tschel, Sebastian Riedel, and Douwe Kiela. 2021. [Retrieval-augmented generation for knowledge-intensive nlp tasks](#).
- Xuechen Li, Tianyi Zhang, Yann Dubois, Rohan Taori, Ishaan Gulrajani, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. AlpacaEval: An automatic evaluator of instruction-following models. https://github.com/tatsu-lab/alpaca_eval.
- Nelson F. Liu, Kevin Lin, John Hewitt, Ashwin Paranjape, Michele Bevilacqua, Fabio Petroni, and Percy Liang. 2023a. [Lost in the middle: How language models use long contexts](#).
- Nelson F. Liu, Tianyi Zhang, and Percy Liang. 2023b. [Evaluating verifiability in generative search engines](#).
- Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. 2023c. [Agent-bench: Evaluating llms as agents](#).
- Gr  goire Mialon, Cl  mentine Fourier, Craig Swift, Thomas Wolf, Yann LeCun, and Thomas Scialom. 2023. [Gaia: a benchmark for general ai assistants](#).
- Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2022. [Webgpt: Browser-assisted question-answering with human feedback](#).
- OpenAI, :, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mo Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madeleine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Sim  n Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian

- Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rameev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O’Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2023. [Gpt-4 technical report](#).
- Arnold Overwijk, Chenyan Xiong, Xiao Liu, Cameron VandenBerg, and Jamie Callan. 2022. [Clueweb22: 10 billion web documents with visual and semantic information](#).
- Baolin Peng, Chunyuan Li, Pengcheng He, Michel Galley, and Jianfeng Gao. 2023. Instruction tuning with gpt-4. *arXiv preprint arXiv:2304.03277*.
- Ofir Press, Muru Zhang, Sewon Min, Ludwig Schmidt, Noah A. Smith, and Mike Lewis. 2023. [Measuring and narrowing the compositionality gap in language models](#).
- Ansh Radhakrishnan, Karina Nguyen, Anna Chen, Carol Chen, Carson Denison, Danny Hernandez, Esin Durmus, Evan Hubinger, Jackson Kernion, Kamilė Lukošiuūtė, Newton Cheng, Nicholas Joseph, Nicholas Schiefer, Oliver Rausch, Sam McCandlish, Sheer El Showk, Tamera Lanham, Tim Maxwell, Venkatesa Chandrasekaran, Zac Hatfield-Dodds, Jared Kaplan, Jan Brauner, Samuel R. Bowman, and Ethan Perez. 2023. [Question decomposition improves the faithfulness of model-generated reasoning](#).
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2023. [Exploring the limits of transfer learning with a unified text-to-text transformer](#).
- Ruiyang Ren, Yingqi Qu, Jing Liu, Wayne Xin Zhao, Qiaoqiao She, Hua Wu, Haifeng Wang, and Ji-Rong Wen. 2023. [Rocketqav2: A joint training method for dense passage retrieval and passage re-ranking](#).
- Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. [Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy](#).
- C. T. Stayton. 2015. [What does convergent evolution mean? the interpretation of convergence and its implications in the search for limits to evolution](#). *Interface Focus*, 5(6):20150039.
- Nassim Nicholas Taleb. 2008. *The Black Swan*. Penguin Books, Harlow, England.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2022. MuSiQue: Multi-hop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*.
- Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. [Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions](#).
- House Committee on Science United States Congress, Subcommittee on Space Science Technology, and Applications. 1981. *NASA Program Management and*

- Procurement Procedures and Practices: Hearings Before the Subcommittee on Space Science and Applications of the Committee on Science and Technology, U.S. House of Representatives, Ninety-seventh Congress, First Session.* U.S. Government Printing Office, Washington, D.C.
- Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc Le, and Denny Zhou. 2023. [Chain-of-thought prompting elicits reasoning in large language models.](#)
- Jiayang Wu, Wensheng Gan, Zefeng Chen, Shicheng Wan, and Hong Lin. 2023a. [Ai-generated content \(aigc\): A survey.](#)
- Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, Ahmed Hassan Awadallah, Ryen W White, Doug Burger, and Chi Wang. 2023b. [Autogen: Enabling next-gen llm applications via multi-agent conversation.](#)
- Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. [Approximate nearest neighbor negative contrastive learning for dense text retrieval.](#)
- Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. [Hotpotqa: A dataset for diverse, explainable multi-hop question answering.](#)
- Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L. Griffiths, Yuan Cao, and Karthik Narasimhan. 2023a. [Tree of thoughts: Deliberate problem solving with large language models.](#)
- Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023b. [React: Synergizing reasoning and acting in language models.](#)
- Weizhe Yuan, Richard Yuanzhe Pang, Kyunghyun Cho, Xian Li, Sainbayar Sukhbaatar, Jing Xu, and Jason Weston. 2024. [Self-rewarding language models.](#)
- Jiahao Zhang, Haiyang Zhang, Dongmei Zhang, Yong Liu, and Shen Huang. 2023. [Beam retrieval: General end-to-end retrieval for multi-hop question answering.](#)
- Xiang Zhang, Junbo Zhao, and Yann LeCun. 2016. [Character-level convolutional networks for text classification.](#)
- Yi Zhang, Sujay Kumar Jauhar, Julia Kiseleva, Ryen White, and Dan Roth. 2021. [Learning to decompose and organize complex tasks.](#) In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2726–2735, Online. Association for Computational Linguistics.
- 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. 2023a. [Judging llm-as-a-judge with mt-bench and chatbot arena.](#)
- Shen Zheng, Jie Huang, and Kevin Chen-Chuan Chang. 2023b. [Why does chatgpt fall short in providing truthful answers?](#)
- Kun Zhou, Yutao Zhu, Zhipeng Chen, Wentong Chen, Wayne Xin Zhao, Xu Chen, Yankai Lin, Ji-Rong Wen, and Jiawei Han. 2023. [Don’t make your llm an evaluation benchmark cheater.](#)

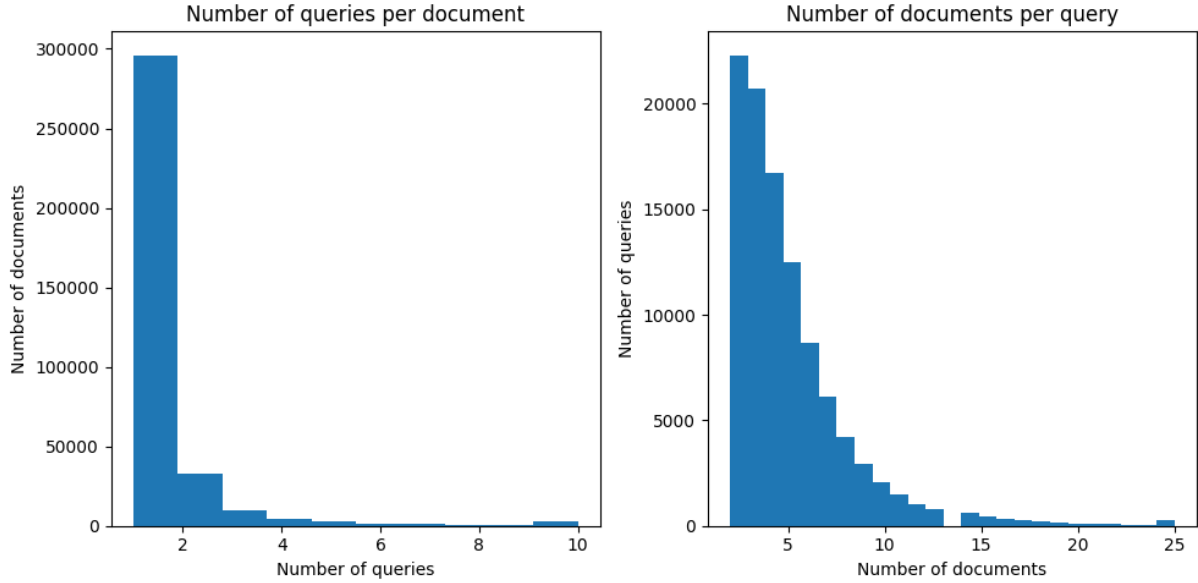


Figure 2: (Right) Histogram of number of documents clicked per question for Researchy Questions which is much higher than for general web search queries. (Left) number of queries associated with each document. The fact that not very many queries are associated with each document validates the effectiveness of our query-deduplication procedure.

A GPT-4-based Filtering Details

We had GPT-4 label certain attributes of each Researchy Question using the prompt given in Figure 7. Particularly, we focus on removing questions that were too **incomplete** or underspecified to meaningfully answer (shown in Table 7), or those that are too **presumptuous** (shown in Table 8). Incomplete queries could arise when users ask queries that e.g. refer to previous queries or subjects in their session.

A.1 Additional Safety Filtering

In addition to the official offensiveness/adult query classifiers used in our commercial search engine, we also had GPT-4 label on a scale of 1-10 “*to what extent the question could be interpreted as being harmful (physically or psychologically to oneself, others, or animals), offensive, overly biased, sexually explicit, or otherwise inappropriate for e.g. someone of the age of 12 to be exposed to.*”. We found roughly 2% of queries had an unsatisfactory harmfulness score and hence were removed from the released dataset.

Table 9 shows some examples of those removed. Real people asked these questions: we acknowledge that some of these may be motivated by genuine information-seeking intent, notwithstanding, we believe there is too high a risk that the act of answering these questions may lead to harm.

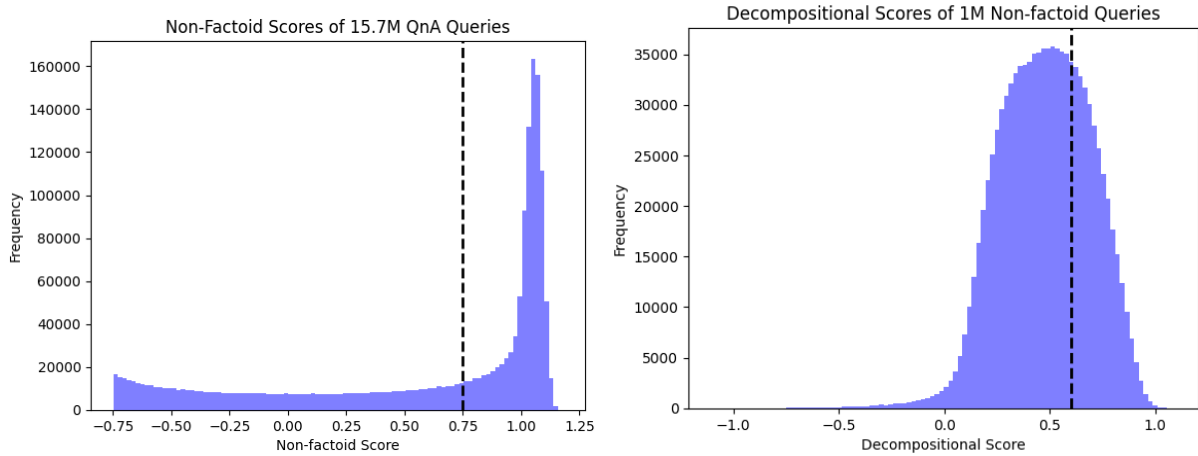


Figure 3: (Left) Non-factoid scores of the 15.7M QnA Queries. The roughly 1M queries whose score exceeded the threshold +0.75 were sent to the Decompositional classifier. Note that because this was a binary classifier, 89% of the non-factoid scores were less than -0.75, which is cut off from the left-hand histogram to make it easier to visualize. (Right) The Decompositional classifier’s scores of the roughly 1M Non-factoid queries. Around 146k queries exceeding the 0.6 threshold line resulted are considered both Non-factoid and Decompositional, and were then de-duplicated to arrive at the final Researchy Questions dataset of around 100k.

Question	Factoid	Decompositional (↓)
Why is it illegal to destroy money	1.02	0.59
Which law firms offer the best opportunities for international work?	1.07	0.58
How should a grad student communicate with faculty	0.80	0.56
How does a short story differ from a novel	1.07	0.54
Why is Beowulf an important literary work	1.10	0.51
Why are e scooters illegal	0.90	0.50
Why did the declaration of independence start	1.12	0.49
How did marsupials evolve	1.02	0.48
Is turbulence dangerous	0.97	0.47
Why do people cut down the Amazon rainforest	1.09	0.45
Why are flags at half mast in Indiana today	1.06	0.45
Why the houses in Rajasthan have thick walls and flat roofs?	1.06	0.43
How do forensic pathologists determine cause and manner of death?	1.02	0.43
What causes changes in the moon’s appearance	1.10	0.41
How did Edgar Allan Poe start his writing career	0.97	0.40
How common was vasectomy in the 1950’s	0.90	0.40
Why are behavioral questions important	0.79	0.38
Is pepper spray harmful	0.93	0.37
How do fertilizers increase productivity	1.05	0.35
How do we get matter and energy from food	1.11	0.32
What type of competition occurs when only a few firms dominate a market?	1.08	0.31
Why can’t waves travel through a vacuum?	1.08	0.30
What caused the Dixie fire	0.84	0.29
Why camels live in desert	1.02	0.29
How was COVID-19 discovered	0.92	0.28
How are elements organized into groups	0.96	0.27
Why did Montezuma give the Spanish gold?	1.01	0.27
Differences between debt and equity financing	0.98	0.27
How are fossils older than 60,000 years dated?	1.08	0.17
What happens when a data collision occurs on an Ethernet bus?	1.03	0.16
How is gold formed in the earth’s crust	0.98	0.12

Table 6: Examples of QnA Queries that were non-factoid (had a non-factoid classifier score above 0.75), but were **not** decompositional (had decompositional classifier score **less** than 0.6). Further down this list, more of the examples have a single correct answer (even if it is somewhat long-form), but there is obviously a lot of gray area, highlighting the challenges of trying to filter large amounts of queries by simple thresholds.

For the following (Question | score | reason) triples, the score indicates how "good" of a non-factoid question they are in the sense that they can lead to interesting and in-depth analysis.

Definition: A good non-factoid question is specific, with potential to amount to a good research report with a clear and refutable thesis, supported by evidence and analysis.

Characteristic formats of good non-factoid questions (not exhaustive):

- Good non-factoid questions will often talk about the relationship between two things, e.g. "Compare and contrast X and Y", "How/why does X affect/impact Y?", "Why X is significant to Y", or "What role does X play in Y?", or "to what extent does X lead to Y?", etc.
- A good non-factoid question can also ask "Why does X happen", "What factors play a role in X?", "How is X significant" or "What is the cause of X", but it should be specific about what kind of analysis is expected.
- Other forms of good non-factoid questions can ask about the pros/cons, benefits/detriments of something, or compare/contrast two things, etc.

Instructions: Rate each question on a scale of 0-10, where 0 is a factoid question and 10 is an excellent non-factoid question and then provide a brief reason for your rating

Q: how tall is abraham lincoln | 0 | factoid

Q: can i change the weather | 2 | personal question

Q: was the civil war fought over slavery | 5 | fair, but could more directly ask about other important facets of the causes of the civil war and their role in the conflict

Q: to what extent was the civil war fought over slavery | 8 | good, will lead to in-depth analysis on the causes of the civil war

Q: what impact do human activities have on the weather | 10 | excellent, many in-depth reports written to answer this question

Q: should LA invest more in railway or highway infrastructure for public transport | 9 | great

Q: what is an example of blackbody radiation? | 0 | asking for an example

Q: could not determine type for | 0 | not a question

Q: what typically signals the end of the olympic games | 2 | factoid, olympic closing ceremony can be looked up easily

Q: Why were Navajo code talkers used during WW2? | 7 | good, could lead to analysis of how culture and language can be used in warfare

Q: When does protein folding begin? | 1 | has a single, known correct answer

Q: what is the cost and necessary materials to build a refinery | 5 | fair, asks about a complex process but will not likely elicit analysis

Q: What is the Navavidha Bhakti? | 0 | asking for a definition

Q: why is technological change bad? | 5 | fair, but could be more specific

Q: analyze how technological changes have historically impacted cultures | 10 | excellent, very specific

Q: who owns phone number 280-626-1435 | 0 | personally identifiable information

Q: What are the main differences between regulations of the NFL and the CFL? | 4 | has potential for in-depth analysis but doesn't explicitly ask for it

Q: Why do planes using rivets & not welded construction? | 7 | good, will require in-depth analysis on aerospace technology

Q: How did the Catholic Pope manage to become more powerful than Kings in old Europe? | 9 | much potential for historical analysis

Q: interesting facts about korea | 0 | not specific

Q: {Question} |

Figure 4: Prompt given to text-davinci-003 to collect labels of whether a question is non-factoid. The current question is substituted at the end. Labels on 1-10 were binarized based to train the non-factoid classifier.

Question: {Question}
 ### Instructions: how clearly does the above question require sub-questions in order to answer it? Think silently about how complex or multi-facted the question is. Imagine you are a person trying to answer the question using a search engine like Google. Would the person likely have to issue multiple queries to arrive at a comprehensive answer? Would they need to exert more effort to fully understand the nuances behind the question, or different perspectives surrounding it? etc. Or, is the person likely to find the best answer with a single search with low effort? Give your rating on 1-100 between <score> and </score> tags, where 1 means the question is trivial or common knowledge, while 100 means it likely involves effort to break the question down into many sub-questions or facets. <score>

Figure 5: Prompt given to gpt-35-turbo to collect labels of how appropriate a question is for “decomposition” into sub-questions. These labels were used to train the Decompositional classifier.

how can climate change	why did the market crash
are perspectives that change when presented with fresh experiences	how did the war for independence end
in what ways did native peoples transform	how dangeous can a persona affect a person personality
how would the law treat the issues confronting a diverse family?	how did president lyndon b. johnson respond to this event?
compare and contrast social contract	what is the difference in how the government reacted in these two situations?
how were others influenced by miyamoto	explain why summers were difficult for europeans.
what effect did the recent looting have on emigration	what is the main reason why some teams never reach stage 4

Table 7: Examples of incomplete or ambiguous queries a.k.a too underspecified to attempt to answer meaningfully (roughly 3% of queries in the last stage of filtering) that were **removed** from the dataset by additional GPT-4 filtering.

how college became a ruthless competition divorced from learning	why is airline customer service so bad
why are police departments afraid of change	why immigration is bad for the us
why did chinese families not desire to have female children?	why do games make people socially disconnected from people
why are catholics democrats	why junior doctors are treated so badly
how casinos have ruined towns	why should zoos be shut down
how is the government controlling us	why rehabilitation does not work in america
why veganism is bad for animals	why most clinical research is not useful

Table 8: Examples of “presumptuous” queries we removed from the dataset (roughly 2% of queries in the last stage of filtering), roughly defined as those where the question is slanted or biased in some way. Note, not all assumptive queries were harmful.

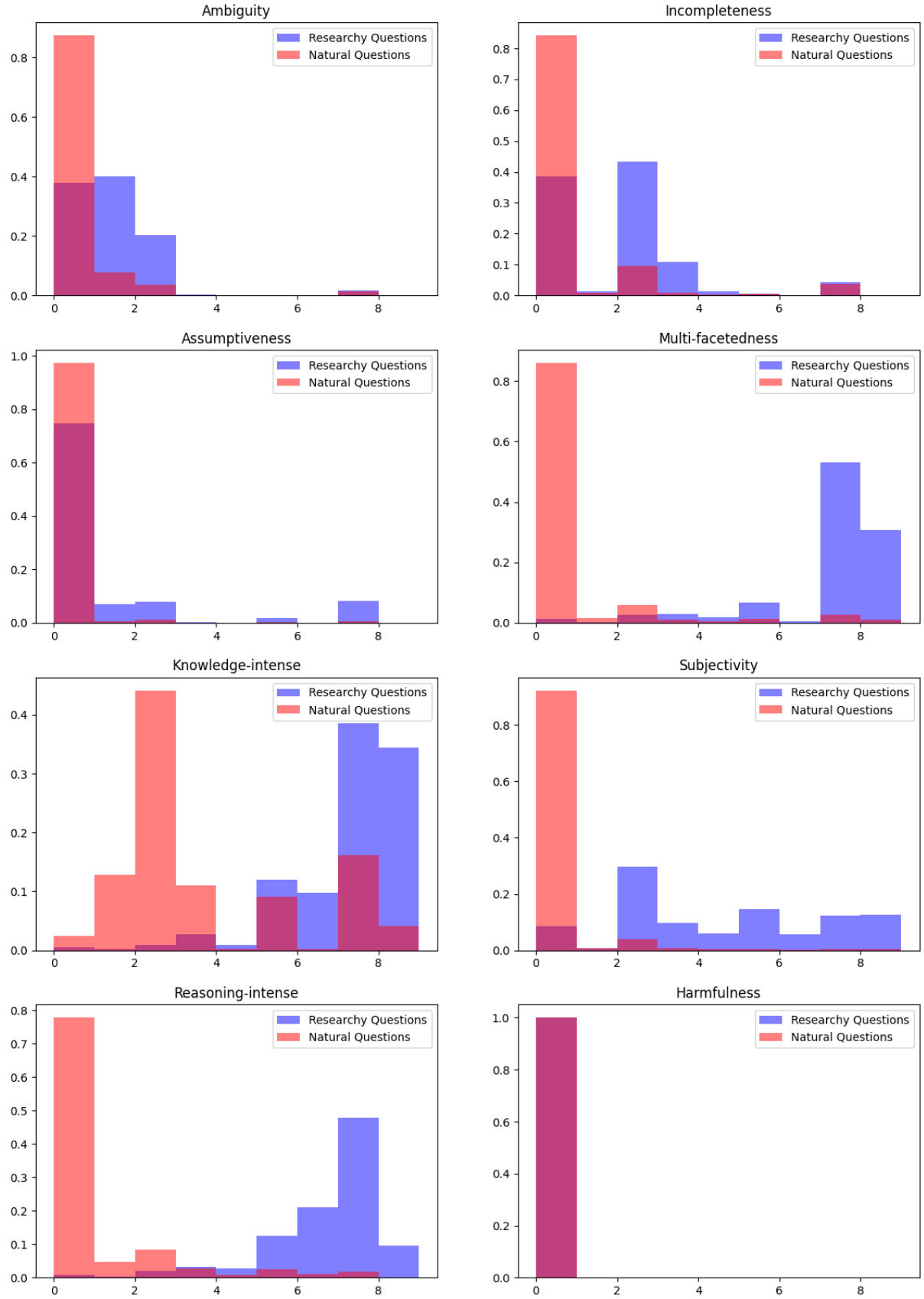


Figure 6: We had GPT-4 label all 102k Research question – and a uniform sample of 2k Natural Questions – along 8 dimensions using the prompt in Figure 7. Each dimension was on a scale of 1-10; we plot their normalized densities. For Research Questions, these are the scores of the final 96k released questions after final filtering described in Section 2.5. Lower scores are better for Ambiguity, Incompleteness, Assumptiveness and Harmfulness. We intended Research Questions to have higher scores for Knowledge-intensiveness, Reasoning-intensiveness and Multi-facetedness.

Given the question: {Question}

Instructions: Please output a python dictionary with fields scoring the question on the following criteria:

1. "**ambiguous**" : Int 0-10 to what extent is the intent of the question ambiguous (has more than one interpretation); 0 means no major ambiguity. Not to be confused with subjectiveness or incompleteness.
2. "**incompleteness**" : Int 0-10 indicating how difficult it is to determine the intent of the question, whether it is missing crucial context or details that ought to be specified in order to answer the question; 0 means the question is answerable and self-contained, 10 means the question is un-answerable because it is incomplete or under-specified.
3. "**assumptive**" : Int 0-10 the degree to which the question has built-in assumptions or biases (that are not offensive, which is point 8 below); 0 means no notable or unreasonable assumptions.
4. "**multi-faceted**" : Int 0-10 the degree to which the question has multiple facets or perspectives that need to be considered in order to answer it; 0 means the question is straightforward and has a single, undisputed answer.
5. "**knowledge-intensive**" : Int 0-10 the degree to which the question would require specialized knowledge (like textbooks, scholarly articles, etc) to provide a thorough and grounded answer; 0 means the answer is common knowledge or can be looked up instantly in common references, 10 means the questions probably entails a lot of work to find and analyze specialized knowledge.
6. "**subjective**" : Int 0-10 the degree to which the question is subjective, meaning an answer(s) exist, but there is no agreed-upon way to determine which one is better; 0 means the question is largely objective i.e. the overwhelming majority of people would agree on the answer if they knew it.
7. "**reasoning-intensive**" : Int 0-10 the degree to which the question requires reasoning to synthesize an answer; 0 means the question can be answered trivially e.g. by looking up a fact, referencing an encyclopedia or database, or using a calculator (once).
8. "**harmful**" : Int 0-10 to what extent the question could be interpreted as being harmful (physically or psychologically to oneself, others, or animals), offensive, overly biased, sexually explicit, or otherwise inappropriate for e.g. someone of the age of 12 to be exposed to.

Note that the above criteria are not mutually exclusive, e.g. a question can be both subjective and knowledge-intensive, for example "is capitalism better than socialism" would be both. Make sure to output only the valid python dictionary without comments or other extraneous output.

Figure 7: Prompt given to GPT-4 to classify attributes of each of the Researchy Questions. The output labels were used to further filter the data for release.

Is it fair to blame younger people for covid spike	How bad is illegal immigration
Why do people pretend to be depressed	Why veganism is bad for animals
what factors affect when deciding to terminate a down syndrome child	Why is the business of sex so popular and lucrative?
how people make money illegally	Was the holocaust real?
Why was gender queer banned	Which countries persecute christians
Why do so many women get abortions	Was the holocaust worse than slavery
Could a small nuclear war reverse global warming	How do holocaust deniers try to gain followers?
Is china a threat to world peace	Why christians believe conspiracy theories
Why do so many incompetent men become leaders	Why did africans sell other africans into slavery
Why is bribery being used to force people to take vaccine	Why were many whites scared away from supporting civil rights
What makes a church a cult	Why did poor southerners fight for slavery
Are minorities more likely to commit crime	Why isn't christianity considered a cult
Why corporal punishment should be legal	Pros and cons of sexism
Should we stratify people based on intelligence or other specific qualities?	Why are we challenged by the arrival of migrants in our country?
Why are we protecting the unvaccinated	Is the death penalty racially motivated
Why is bayonetta so sexualized	Why is racial profiling important in society
Does islam promote violence	Is affirmative action racism
Who is behind mass immigration	Why is the vaccine killing people
Why is the west so intent in bringing in so many illegal migrants	Why might the artist have thought reconstruction was "worse than slavery" for black southerners?

Table 9: Here we show some examples of the roughly 2% of queries that were **removed** from the dataset by additional GPT-4 harmfulness filtering.

Question: {Question}

#####

Reference Answer (assumed to be true): {ReferenceAnswer}

{CandidateAnswers}

#####

Keep in mind the following Guidelines when evaluating the Candidate Answers:

Guidelines:

- **Usefulness:** A useful answer adds value by providing in-depth analysis, actionable steps, and relevant information that helps users make informed decisions, complete tasks, and understand complex problems in a simpler way. It also considers real-world scenarios, cause and effect relationships, and historical context to enhance the user's understanding.
- **Relevance:** A relevant answer directly addresses the question's intent in a sensible and specific manner, without being overly broad, redundant, or distracting. It should be clear, on-topic, and provide helpful information tailored to the question asked.
- **Truthfulness** in an answer means providing accurate and verifiable information, without making false claims, unverifiable statements, or promoting misleading information. It should agree with the Reference Answer. Contradicting the Reference Answer *must* result in lower score.
- **Groundedness** in an answer means providing information supported by reliable sources and following a logical chain of reasoning, instilling confidence in its accuracy. The answer should be based on credible evidence, address uncertainties, and avoid relying on personal opinions, rumors, or hearsay.
- **Thoroughness** involves considering the full scope of a question, addressing multiple perspectives, scenarios, and outcomes, and ensuring all important aspects are covered without omission. It requires analyzing pros and cons, empathizing with various stakeholders, and addressing different interpretations or facets of a question.

Instructions: Above are {NumCandidates} Candidate Answers to the Question: {Question}. Please read them carefully along with the Guidelines for how to evaluate an answer's quality. Then:

1. Briefly **critique** each Candidate Answer with respect to the Guidelines, formatted in **bullet points** between "<CritiqueX>" and "</CritiqueX>" tags where X is the Candidate Answer number. Each bullet is between sub-tags of either <strength>, <weakness>, or <missinginfo>. A <strength> is where the Answer makes a good point that follows the Guidelines and contributes to a helpful response. A <weakness> is where the Answer makes an incorrect, irrelevant, unreasonable or overly broad claim that fails to address the full scope of the Question, or otherwise falls short of the Guidelines. <missinginfo> is when some key event, facts or other material information is omitted but should have included to strengthen the Candidate Answer.
2. Give a **binary correctness** score indicating whether each Candidate Answer generally agrees with the Reference Answer – "<CorrectX>1</CorrectX>" – or contradicts it "<CorrectX>0</CorrectX>", where X is the Candidate Answer number.
3. **Explain** which Candidate Answer is better and why, i.e. how well it adheres to the Guidelines, between "<Explanation>" and "</Explanation>" tags."
4. Finally, **score** each Candidate Answer on 1-100, where 100 is a perfect Answer that aligns with the Guidelines. Indicate the score between "<ScoreX>" and "</ScoreX>" tags where X is the Candidate Answer number.

Make sure to address all {NumCandidates} Candidate Answers. When you are finished, conclude your response with "====".

<Critique1>

Figure 8: Prompt given to GPT-4 to act as a “critique-then-score” judge of candidate answers side-by-side.

Researchy Question: how does public transport help the economy
GPT-4 Turbo Closed-book Answer (2023-07-01-preview)
<p>Public transport provides numerous economic benefits to communities and countries as a whole. Here are some of the key ways in which public transportation supports economic growth and efficiency:</p> <ol style="list-style-type: none"> 1. Reduced Congestion Costs: Public transport can significantly reduce the number of vehicles on the road, which in turn reduces traffic congestion. This leads to a more efficient transportation system and increased productivity. 2. Cost Savings for Individuals: Public transport can be more cost-effective than owning a private vehicle, leading to increased disposable income for individuals. 3. Job Creation: Public transport systems create jobs for operation, maintenance, and construction, contributing to employment opportunities. 4. Increased Property Values: Good public transport connectivity often increases property values, leading to higher tax revenues. 5. Environmental Benefits: Public transport reduces the number of cars on the road, decreasing air pollution and greenhouse gas emissions. 6. Tourism: Efficient public transport systems enhance tourism by making it easier for tourists to navigate and spend in local economies. 7. Business Productivity: Reliable public transport reduces employee tardiness and absenteeism, increasing business productivity. 8. Social Inclusion: Public transport provides mobility for those without access to private vehicles, allowing for greater workforce participation and access to services. 9. Urban Development: Public transport encourages efficient land use and can be a catalyst for urban revitalization and development. 10. Energy Efficiency: Public transport is more energy-efficient per passenger kilometer than private vehicles, contributing to energy security. <p>By providing an affordable and efficient alternative to private vehicles, public transport systems can enhance the overall economic performance of a region and improve the quality of life for its residents.</p>
Mixtral 8x7b Closed-book Answer (https://stablediffusion.fr/mixtral)
<p>Public transportation can have a positive impact on the economy in several ways:</p> <ol style="list-style-type: none"> 1. Reduced traffic congestion: Public transportation helps to reduce traffic congestion by providing an alternative to driving, which can lead to decreased travel times and increased productivity. This can also reduce the need for costly road expansion projects. 2. Increased economic activity: Public transportation can help to stimulate economic activity by making it easier for people to access jobs, shops, and other services. This can lead to increased spending and investment in local communities. 3. Lower transportation costs: Public transportation can provide a more affordable transportation option for many people, especially those who cannot afford to own and maintain a car. This can help to reduce household transportation costs and free up money for other expenses. 4. Improved air quality: Public transportation can help to reduce air pollution by taking cars off the road. This can lead to improved public health and reduced healthcare costs. 5. Increased property values: Properties located near public transportation hubs, such as train stations and bus stops, often have higher property values than those located further away. This can lead to increased tax revenues for local governments. 6. Job creation: Public transportation agencies and companies create jobs for drivers, mechanics, engineers, and other professionals. These jobs can provide stable employment and contribute to the local economy. 7. Energy efficiency: Public transportation is generally more energy-efficient than individual car travel. This can help to reduce dependence on fossil fuels and decrease greenhouse gas emissions. <p>Overall, public transportation can play an important role in supporting a strong and sustainable economy by providing affordable, efficient, and environmentally-friendly transportation options.</p>

Table 10: Closed-book direct answers from GPT-4 and Mixtra 8x7b (Jiang et al., 2024) to the Researchy Question used in Table 2. We find these two answers to be strikingly similar (suggesting convergent evolution of LLMs), and both paint in very broad strokes an answer that isn’t “wrong” but rather, superficial.

Public transportation plays a significant role in boosting the economy in several ways:

1. Job Creation: Investment in public transportation can create millions of jobs. For instance, every \$1 billion invested could create 50,000 jobs [1,2].
2. Economic Returns: Every \$1 invested in public transportation could generate \$5 in economic returns [1].
3. Reduced Traffic Congestion: A well-functioning public transit system reduces traffic congestion, leading to more efficient travel and productivity [3].
4. Economic Growth: Public transportation can shape land use and development patterns, stimulate commerce, and increase property values [4].
5. Supports Vulnerable Populations: It improves equality, particularly for the more vulnerable members of society, by providing them with mobility and access to jobs [1].
6. Environmental Impact: Public transit helps improve air quality and reduce greenhouse gas emissions, contributing to a sustainable future [3].

In summary, public transportation is more than just a way to move people around. It's a vehicle for opportunity, equity, and a better quality of life.

References

- [1] <https://www.weforum.org/agenda/2021/04/here-s-why-cities-should-invest-in-public-transport/>
- [2] <https://infrastructureusa.org/the-economic-impact-of-public-transportation/>
- [3] <https://www.ncesc.com/is-public-transit-better-for-the-economy/>
- [4] <https://www.apta.com/wp-content/uploads/Resources/resources/reportsandpublications/Documents/Economic-Recovery-APTA-White-Paper.pdf>
- [5] <https://profound-tips.com/advice/how-does-public-transportation-help-the-economy/>

Figure 9: A “screenshot” of the output from Bing Chat in precise mode (a.k.a. Microsoft Copilot) for the question “how does transportation help the economy”.

Comprehensive Research Report: The Role of Transportation in Economic Development

Introduction Transportation is a critical factor in the economic development of any region or nation. It serves as the backbone of economic growth, facilitating the movement of goods, services, and labor, which are essential for the efficient functioning of the economy. This report synthesizes key points from various sources to illustrate how transportation contributes to economic prosperity.

Key Contributions of Transportation to the Economy ## Job Creation and Economic Growth Investments in transportation infrastructure, such as public transit, are directly linked to job creation and economic growth. For every \$1 billion invested in transit, an estimated 49,700 to 50,731 jobs can be created, offering a significant return on investment ranging from 4 to 1 to as high as 5 to 1. This investment not only supports American jobs and industry but also generates income for local businesses and their workers. Moreover, spending on transportation programs creates jobs in the short term, particularly in construction, and can stimulate the economy.

Enhancing Business Productivity and Competitiveness Transportation improvements lead to reduced travel times and costs, as well as increased network reliability, which in turn enhances business productivity and competitiveness. By reducing transport costs, businesses can access higher-quality or lower-cost inputs and connect with potential suppliers. This connectivity also allows individuals to access a wider range of jobs, thereby participating more fully in the labor market.

Market Accessibility and Economic Activity An efficient transport system provides market accessibility by linking producers and consumers, enabling transactions to occur. This linkage is crucial for the economy as transportation accounts for a significant portion of the GDP, ranging from 6% to 12%. By offering cost, time, and reliability advantages, transportation enables goods to be transported over longer distances, fostering economic changes, most of which are positive.

Induced Investment and Regeneration Impacts Transportation investments can influence the level or location of private sector investment due to improved connectivity. These investments can also lead to local economic impacts resulting from an improved local image and the attraction of land use development, known as regeneration impacts.

Structural and Developmental Transformations Sustained improvements in transportation, along with parallel advancements in information and production technologies and institutional structures, can lead to structural and developmental transformations within the economy. This demonstrates the dynamic nature of transportation's role in economic development.

Economic Efficiency and Scale An efficient transportation system reduces costs across many economic sectors, while an inefficient one increases these costs. Moreover, more efficient transport costs less, leading to rising productivity and economies of scale, which should cause GDP to rise faster than transport activity.

Influence on Economic Development and Military Strength Transportation activities contribute to a nation's total economic product and strengthen the national or regional economy. A well-developed transportation network not only makes markets more competitive but also adds to a nation's military strength.

Conclusion Transportation plays a pivotal role in the economic development of a country. It is essential for job creation, enhancing business productivity, providing market accessibility, inducing investment, and contributing to structural transformations. Governments recognize the importance of investing in transportation infrastructure as a means to stimulate economic growth and maintain a competitive edge in the global market. As such, transportation is not just a means of moving people and goods; it is a fundamental driver of economic prosperity.

References

- Economic Impact of Public Transit. (2019, April 22). American Public Transportation Association.
- The relationship between transport and the economy. (2017, February 13). Greener Vision.
- Economic Impact Of Public Transportation Investment. (2019, March 27). American Public Transportation Association.
- 3.1 – Transportation and Economic Development | The Geography of Transport Systems. (2017, December 4). Transport Geography.
- Economic Consequences of Transport Improvements. (2022, October 11). ACCESS Magazine.
- Transportation, Jobs, and Economic Growth. (2018, February 14). ACCESS Magazine.
- Transportation economics | Definition, Importance, & Facts. (1998, September 8). Britannica.
- The role of transport in economic development. FutureLearn.

Figure 10: A “screenshot” of the output from YouPro Research engine for the query “how does transportation help the economy”. <https://you.com/search?q=how+does+transportation+help+the+economy>

Query #1 - how does transportation help the economy
Query #2 - How does transportation infrastructure affect economic growth?
Query #3 - What are the long-term effects of transportation investments?
Query #4 - How does transportation contribute to GDP?
Query #5 - What are the environmental impacts of transportation?
Query #6 - How does transportation affect urban planning and development?
Query #7 - What are the challenges in financing transportation infrastructure?
Query #8 - How can transportation policy influence economic productivity?

Final Research Result for how does transportation help the economy

Transportation plays a pivotal role in the economy, contributing to the movement of goods, services, labor, and capital, which are essential for the functioning of a modern society and the growth of an economy [11]. The ability to transport goods and people efficiently is fundamental to economic life, as seen in the early United States where a robust transportation network was instrumental in the country's development. In a macroeconomic sense, transportation activities are a part of the total economic product, and the demand for transportation services contributes to the GDP [2]. The transportation sector not only supports the movement of goods and services but also affects employment, wages, and business income, thereby contributing to the overall economic growth [3].

Investments in transportation infrastructure can lead to job creation and improved connectivity, which in turn, promotes economic productivity and well-being [22]. Furthermore, efficient transportation systems can reduce poverty by increasing access to employment, recreational, and other essential life-shaping opportunities [20]. The economic importance of transportation is also evident in the fact that transportation-related final demand has been increasing since the recession, indicating a strong correlation between transportation and economic activity [8]. Additionally, the World Bank and ITDP report highlights the importance of active mobility in promoting economic productivity and socially inclusive development [21].

Transportation is a critical component of a society, reflecting the aspirations of a society and influencing the location decisions of people and businesses. It is a vital determinant of economic growth, job creation, and improved living conditions, and its efficiency can lead to a green future by reducing congestion and promoting sustainable development [19].

In conclusion, transportation is not only crucial for the movement of goods and services but also plays a significant role in the overall economic growth of a country. It contributes to the GDP, supports employment, and affects wages and business income [2][3]. Investments in transportation infrastructure can lead to job creation and improved connectivity, promoting economic productivity and well-being [22]. Additionally, transportation plays a vital role in reducing poverty by increasing access to essential opportunities [20]. The economic importance of transportation is evident in the strong correlation between transportation and economic activity [8]. Therefore, transportation is a critical component of a society, reflecting its aspirations and influencing the location decisions of people and businesses [19].

References

- [2] <https://www.britannica.com/money/topic/transportation-economics>
- [3] <https://data.bts.gov/stories/s/Transportation-Economic-Trends-Contribution-of-Tra/smrm-36nv/>
- [8] <https://www.bts.gov/sites/bts.dot.gov/files/docs/browse-statistical-products-and-data/transportation-economic-trends/215926/transportation-economic-trends-2017-chapter-2.pdf>
- [11] <https://www.nationalgeographic.org/article/effects-transportation-economy/5th-grade/>
- [19] <https://transportgeography.org/contents/chapter3/transportation-and-society/>
- [20] <https://blogs.iadb.org/transporte/en/how-can-transport-reduce-poverty-and-promote-socially-inclusive-development/>
- [21] <https://www.usatoday.com/story/special/contributor-content/2024/01/10/itdp-and-world-bank-report-highlights-economic-and-environmental-benefits-of-financing-active-mobility/72177281007/>
- [22] <https://www.accessmagazine.org/spring-2011/transportation-jobs-economic-growth/>

Figure 11: A “screenshot” of the output from SciPhi Research engine for the query “how does transportation help the economy”. This research engine also asked and separately answered multiple queries before arriving at the final research result. <https://search.sciphi.ai/research?q=how+does+transportation+help+the+economy>