Credible Without Credit: Domain Experts Assess Generative Language Models

Denis Peskoff
Princeton University
Office of Population Research
dp2896@princeton.edu

Brandon M. Stewart
Princeton University
Sociology and Office of Population Research
bms4@princeton.edu

Abstract

Language models have recently broken into the public consciousness with the release of the wildly popular ChatGPT. Commentators have argued that language models could replace search engines, make college essays obsolete, or even write academic research papers. All of these tasks rely on accuracy of specialized information which can be difficult to assess for non-experts. Using 10 domain experts across science and culture, we provide an initial assessment of the coherence, conciseness, accuracy, and sourcing of two language models across 100 expert-written questions. While we find the results are consistently cohesive and concise, we find that they are mixed in their accuracy. These results raise questions of the role language models should play in general-purpose and expert knowledge seeking.

1 Do Experts Agree with ChatGPT?

Since its release in late November 2022, ChatGPT has gained over 100 million users in just two months and been the subject of breathless coverage news coverage which claims it threatens to "replace search engines" (Loten, 2022; Grant and Metz, 2022), kill the college essay (Marche, 2022), and automate the writing of scientific research (Stokel-Walker, 2023). These tasks are distinct from the kind usually evaluated in NLP because they all rely on expert-level knowledge. In this paper, we survey 10 experts to obtain subjective assessments of how two recent language models engage with questions in diverse domains.

Our efforts build on prior work to evaluate the capabilities of language models. Language models are now regularly subjected to extensive benchmarks which cover a variety of standard NLP tasks (Wang et al., 2019; Brown et al., 2020; Ribeiro et al., 2020; Srivastava et al., 2022). Recent efforts engage in domain-specific tasks such as taking the bar or medical licensing exams (Katz et al., 2023; Kung et al., 2023) and making political arguments (Palmer and Spirling, 2023). Liu et al. (2023), released on arXiv while this paper was under review, evaluates the ability of generative search engines to answer a range of general knowledge queries. We complement these efforts by having experts craft their own information-seeking questions and evaluate the generated responses.

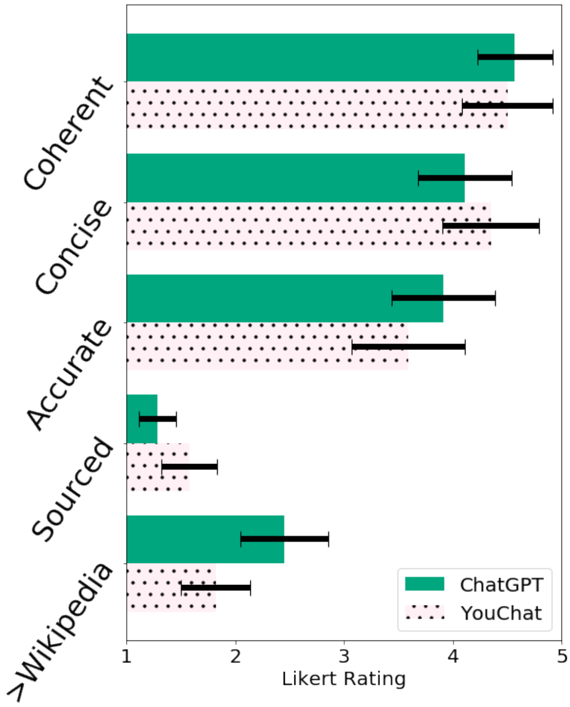


Figure 1: Average ratings by domain experts of language model generated answers to 100 questions across 10 domains (higher values indicate better performance of the language model, 95% confidence intervals computed by expert-blocked bootstrap). Language models score high on cohesion and conciseness, modest on accuracy, and poor on sourcing of their answers. Both models are sometimes ranked ahead of Wikipedia.

In the next section, we briefly discuss the role of expertise in language models and our goals in evaluating it. We then describe our methodology

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 2: Short Papers, pages 427–438
July 9-14, 2023

(Section 3). We find ChatGPT and YouChat to be cohesive and coherent (Section 4.1), generally accurate with some misses (Section 4.2), and ubiquitously lacking in sources (Section 4.3). A majority of our experts recommend these models for general purpose questions but not for professional settings (Section 4.4). We conclude with implications and contrast with the contemporaneous findings in Liu et al. (2023) (Section 5).

2 "Expertise" in Language Models

Individuals and companies are increasingly looking to language models as a source of expert question answering. For example, Metzler et al. (2021) lays out a vision for search that involves language models providing answers to user-generated questions. Unfortunately, the challenge for many language models is that they are trained to generate language, not to have correct answers. As Shah and Bender write, "to the extent that [language models] sometimes get the right answer to…questions [it] is only because they happen to synthesize relevant strings out of what was in their training data. No reasoning is involved" (Shah and Bender, 2022, pg. 222). This has led Narayanan and Kapoor (2022) to characterize ChatGPT as a "bullshit generator"—plausible, but not accurate (Frankfurt, 2005). While language models might incidentally produce accurate answers to simple and uncontested queries (e.g., "what is the capital of France?"), we might be understandably skeptical that it will produce correct answers to more nuanced questions. Generated language reflects its training data and—to the extent the training data is publicly known—it is more reflective of the web than expert speech (Bender et al., 2021). By using experts evaluating material in their domain of choice, we provide an initial assessment of expertise provided by these models. Ultimately what constitutes sufficient accuracy for broader use depends on the use case.

3 Methodology

We evaluate two recently-released language models: OpenAI's ChatGPT and You.com's YouChat (Google's Bard and many other options weren't released at the time of initial submission). OpenAI's ChatGPT is the wildly popular evolution of the GPT-3 model (Brown et al., 2020) and YouChat is built specifically for search. Both systems have a free and public option (at the time of writing) which makes them generally accessible.

We survey 10 experts across a range of arbitrarily-chosen disciplines from quantum information to ballroom dance (see a complete list in the appendix). We recruited experts from our personal networks aiming to cover a wide-range of different types of knowledge (with the understanding we cannot be exhaustive or representative). The majority hold a doctorate or medical degree.

We asked each expert to fill out an online survey with their own description of their area of expertise, two Wikipedia pages pertinent to it, and five common questions and five niche questions from their domain (see Table 1 for examples).[1] In a second wave of the survey, we provide answers generated from these questions using ChatGPT and YouChat and ask them to rank the answers on a 5-point Likert-type item for coherence, conciseness, accuracy, sourcing, and quality of content relative to Wikipedia (Likert, 1932). We ask for open-ended feedback on answers and alternate which system the experts evaluate first. Questions are designed to allow experts to focus on their own area of expertise while providing an opportunity to distinguish between different levels of knowledge-specialization. The survey took one hour on average. Six experts were surveyed in January and four in May of 2023.

The survey design elicits subjective expert judgment of system performance. We evaluate coherence, conciseness, and accuracy as important properties in information-seeking (Cambazoglu et al., 2021). Comparing assessments to Wikipedia provides a difficult-to-beat baseline with which many people are already familiar. We also ask whether the language model provides a source for its information. Evaluating the source of the information in the response is important not only for the purposes of giving credit, but also as a mechanism for accountability (Bender et al., 2021; Liu et al., 2023). After all the questions, we directly ask whether the expert would recommend the tool for general purpose and professional use, and if the style of the content is obviously automated (Dou et al., 2022). We make our data (including the full context for all quotes we use here) publicly available to help support future work.[2]

[1]Prompts for the questions request different conditions such as length, uncertainty, and requesting a source. See the supplement for the survey instrument.

[2]https://denis.ai/bot_thoughts/

Area	Example Common Question	Example Niche Question
Family Medicine	Does everyone get cancer cells in their body?	Are there some parts of cognitive function that improve with age after age 40?
Radiology	Can imaging studies help diagnose traumatic injuries?	What is better to detect and diagnose necrotizing enterocolitis in infants, x rays or ultrasound?
Toxicology	What are the most common laboratory findings in a patient with methanol poisoning?	In what scenario should methylene blue be administered for methemoglobinemia?
Geo-archaelogy	What formed parallel and perpendicular straight water channels in Mesoamerica?	Could geological faulting create parallel channels (streams or canals) that cross at right angles?
Neuro-science	How do neurons communicate in the brain?	Are DREADDs a "good" method for manipulating neural activity?
Quantum Information	What speedups can quantum computers achieve over classical computers?	Are superconducting qubits the best platforms for realizing quantum computers when compared to trapped ion or semiconductor or neutral-atom qubits?
Linguistics	Are there drawbacks to being bilingual?	Is any aspect of syntax universal across all languages?
Digital Humanities	Which was the first novel?	What is the fate of formalism today in literary criticism?
Ballroom	What are some of the most popular ballroom or Latin American partner dances?	Should a competitive dancer seek the opinions of a wide variety of dance judges?
Opera	Which is the best version of "Tannhäuser"?	Why did Rossini stop composing operas at age 37?

Table 1: A common and niche question, selected from one of 10, written by the experts from our 10 domains.

4 Results Across Disciplines

We discuss two running examples: one about Maria Callas, a famous New York-born opera singer with an international career often associated with Greece, and one about quantum information. Our opera expert asks, "Was Maria Callas American?". ChatGPT answers *"No, Maria Callas was not American. Maria Callas, whose full name was Maria Anna Cecilia Sofia Kalogeropoulos, was born on December 2, 1923, in New York City, United States. However, she was of Greek descent. Her parents were Greek immigrants, and she later acquired Greek citizenship. Maria Callas is renowned as one of the greatest opera singers of the 20th century and was particularly known for her performances in the bel canto repertoire."*

As a second example, a quantum expert asks "Can quantum computers be realized in the physical world?". YouChat answers, *"Yes, quantum computers can be realized in the physical world. Researchers have been able to build prototype systems that demonstrate the principles of quantum computing, and there are a number of research groups and companies working on building more advanced versions. Additionally, several companies have announced plans to build full-scale quantum computers in the next few years."*

4.1 Answers Are Credible

Answers are judged as coherent (avg 4.5), and concise (avg 4.2) by our experts. Responses generally restate the question, provide relevant information, are grammatically correct, and are formal in tone. The responses were quite stable on regeneration.

Although results are relatively concise, they do differ in length. ChatGPT's answer to the question about Maria Callas is four sentences including a final sentence about her career that is completely unrelated (YouChat's is 3). For the question on quantum information we gave above, ChatGPT provided a three paragraph answer which our expert described as “"a well constructed and nuanced answer that synthesizes information from multiple perspectives"” while YouChat used three sentences.

4.2 Uneven Accuracy

While responses are fairly uniform in coherence, they are uneven in terms of accuracy (with 111 of the 200 responses marked as one of the two most accurate categories and 38 marked in the two lowest accuracy categories). Surprisingly, niche questions were only slightly less accurate than common ones (-.16). Examining the comments suggest that the rankings reflect fairly different standards for what counts as accurate (expert ratings are included in parentheses below where a 1 is completely inaccurate and a 5 is completely accurate). We urge caution in interpreting the averages.

On the question about Maria Callas, ChatGPT asserts "No" while clarifying that she was born in New York (1) while YouChat answers "Yes" (2). Both comment on her additional Greek citizenship. Our expert gave both quantum information answers top marks for accuracy (ChatGPT:5, YouChat:4).

Seven experts gave at least one answer the lowest accuracy score suggesting it is completely wrong. For example in a toxicology answer, ChatGPT gave “a list of causes of *anion gap* acidosis instead of ***NON-anion*** gap acidosis” (1). Similarly YouChat answered the wrong question from our ballroom dance expert by confusing “the Viennese Waltz with the Waltz. The answer describes an entirely different dance from the dance the question is about” (1). Many other answers though were quite accurate. Our geoarcheologist expressed a common sentiment that the responses are “basic but generally correct” (4). Other answers were “excellent, nuanced” (5, toxicologist). The fairly uniform coherence makes it difficult for a non-expert to discern the correct information from the noise.

The answers also varied in their ability to capture uncertainty in the field overall. Our neuroscientist noted that ChatGPT “accurately captured the controversy surrounding use of DREADDs” (5) but that YouChat “was unable to capture the longstanding controversy” (4). The toxicologist noted that ChatGPT offered a “definitive answer to something that is not totally agreed upon” on the subject of dialysis for lithium poisoning (3). By contrast, our linguist observed on a niche question that “the response to the query about complex predicates is appropriately waffly” (5).

We close this section by noting that even for experts, assessing accuracy can be complicated. Our linguist notes “I would say that the response is invalid, but there are linguists who would agree with it and YouChat does flag the fact its controversial” (2) and the geoarcheologist cited overclaiming, writing that YouChat “takes too strong a position that the evidence does not back up” (1). Such cases are difficult to adjudicate—what counts as sufficient evidence?—but the difficulty is inevitable with complex questions.

4.3 Sourcing is Almost Completely Absent

Our clearest finding is that most answers by the language models do not provide any source for their information. Only 11 out of 100 ChatGPT answers and 19 of 100 YouChat answers were scored more than the lowest value for sourcing. Neither system provides a source for Maria Callas’ biographical information nor concrete examples of physically-realized quantum computers.

When sources are provided, they are often vague, irrelevant, or misleading. The neuroscientist remarked on the first problem writing, “the references are vague; it can cite the names of scientific journals and books but not specific articles or book chapters”. When the models provide a source we found that it was often a (only tangentially relevant) Wikipedia article (Figure 2 provides an anomalous example). These are sometimes loosely related by keywords, but still irrelevant such as a reference to Wikipedia’s article on post-traumatic epilepsy for a question about using imaging to diagnose traumatic injuries. In a question on quantum information, a relevant Stephen Hawking paper was recommended, but an unrelated link was provided.

Perhaps the most serious concern is where an authoritative source is invoked, but inaccurately. When asked “What should a radiologist recommend to a patient after the radiologist incidentally detects a thyroid nodule on a chest CT scan done for another reason?” ChatGPT claims, *“The American Thyroid Association recommends that patients with a thyroid nodule larger than 1 cm or with suspicious features on imaging should undergo a fine-needle aspiration (FNA) biopsy.”* But, “neither the ACR not ATA recommend that patients with a thyroid nodule larger than 1 cm should categorically undergo fine-needle aspiration”! This echoes previous findings in the domain of medicine, where work evaluating previous generations of voice as-

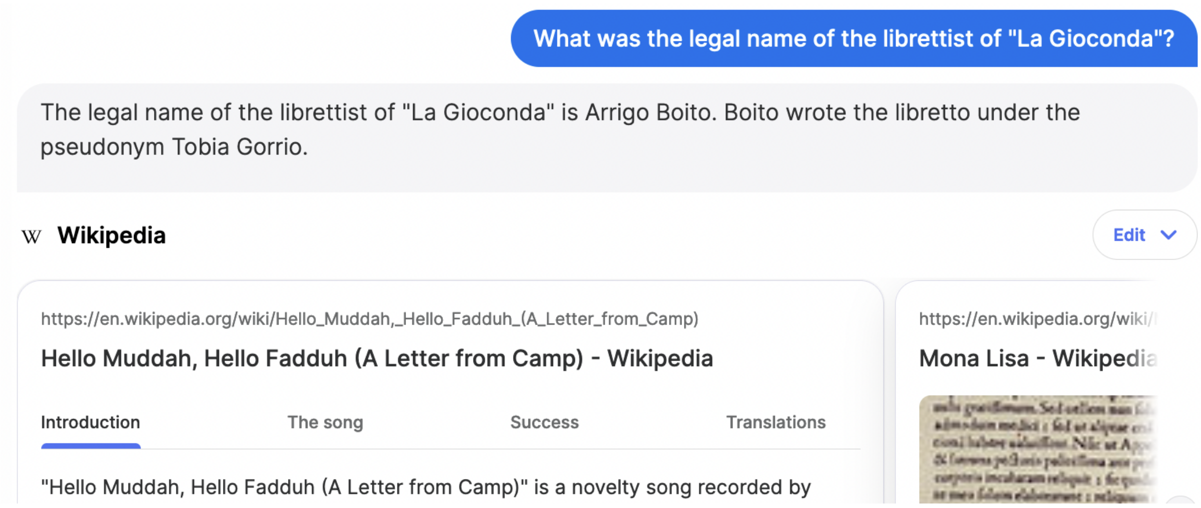


Figure 2: The Wikipedia pages for *Hello Muddah, Hello Fadduh (A Letter from Camp)* and *Mona Lisa* are unsurprisingly not the correct sources for the legal name of a librettist, despite the YouChat interface's suggestion.

sistants has shown that they provided inaccurate medical answers that could have proven fatal (Bickmore et al., 2018). Our neuroscientist asked a niche question where YouChat identified a specific journal article, but it appears to be made up (neither we, nor she, were able to find it) although she did judge the answer as completely accurate.

4.4 Mixed Recommendations for Use

Only 3 of the 10 experts would recommend using ChatGPT and 0 of the 10 would recommend YouChat in professional setting (rating of 4 or higher, where 5 is "full confidence"). However, the majority would endorse both systems for general purposes questions about their domain (70% rating of 4 or higher)—more than would endorse Wikipedia for the same (60% rating of 4 or higher). The family physician summarized a common theme, "once again Wikipedia has extensive articles on life expectancy extension but nowhere near as concise as this" and the linguist wrote on YouChat's answer, "this is an excellent concise response, although wiki provides more information (as usual)."

5 Discussion

Language models were coherent, but undersourced and not always accurate. They were generally not endorsed for professional use, but were seen as valuable by some experts as a source of knowledge for people out of the domain. Providing sourcing citations will be an important step in building confidence. Citations are sufficiently inconsistent when they appear to merit verifying important results.

Our findings are reinforced by the contemporaneous work of Liu et al. (2023) which provides a more systematic audit of four generative search engines (including YouChat, but not ChatGPT) on a diverse series of queries (including common google searches and questions on Reddit) using 34 prescreened MTurk annotators. They also find that these search engines are "credible without credit"—having high fluency and perceived utility, but insufficient sourcing. They find that about half of the responses are fully supported by citations and three fourths of the citations given didn't actually support the sentence. One of their main findings is a negative correlation between citation recall/precision and fluency/perceived utility. Sourcing is so absent in our study that we observe no meaningful correlation with other variables and accuracy has positive correlation with cohesion and conciseness. Further work could investigate if these discrepancies are due to differences in the systems evaluated, the kinds of questions asked, or the judgments of experts vs. annotators. This difference aside, their findings resonate with ours that credibility without credit should make us cautious in looking to language models as a source of expertise.

Limitations

Our study has three important limitations. First, our study is small in scope. By their nature, experts are difficult to recruit and consequently the domains we can cover are limited. The small sample also suggests that the quantitative measures may not be stable in a larger or more representative sample.

Second, our observation process was somewhat artificial. We generated replies for our experts and did not to do any prompt tuning. This reflects the way the expert chose to ask the question, but does not capture the ceiling of performance that would be possible in a conversation. As the Family Medicine expert noted about our question comparing Wikipedia to ChatGPT, "for more detail one could spend more time with Wikipedia and to the organization themselves, but chat provides an immediate general summary and the opportunity to drill down further with ongoing questions and conversation.I have used chat GTP to do medical and biological research In a matter of minutes which would have taken me hours previously". A more extensive study on information seeking behaviors would be of interest and Liu et al. (2023) is a useful step in that direction.

Third, the responses across experts are not necessarily comparable. We allowed experts to choose their own questions and provide their own interpretations of the key measures like coherence or conciseness. Comparability of scales across contexts is a long-standing problem in survey research (King and Wand, 2007) and we highlight some of the concerns around the accuracy question above. Nevertheless, we felt that asking a set of closed-ended questions would help to provide some aggregate judgment, adding some systematic data to the anecdotes shared in public forums. While we caution about drawing any binding conclusions from our preliminary work, we felt that given the fast-evolving nature of the technology, a quick assessment was merited. Our findings are broadly supported using different questions and methodology in Liu et al. (2023).

One important aspect that is out of scope in our analysis is differential accuracy by question asker. Latanya Sweeney's classic study of racial discrimination in online ads (Sweeney, 2013) points to the possibility that *how* a question is asked or where it is asked *from* could result in inaccurate or harmful answers for marginalized communities (see also Noble, 2018; Benjamin, 2019). We have also focused exclusively on English language questions and answers, but differences in easily-available training data across languages can produce substantial differences in the information offered. For example, Yang and Roberts (2021) shows that embeddings trained on Baidu Baike—an online Chinese encyclopedia—encode substantially different associations with sensitive historical events and people than Chinese Language Wikipedia (which is regularly blocked in China). There is much more to understand about the degree to which large language models can mimic expertise.

Ethics Statement

Work was approved by Princeton University's IRB under Proposal 15346. No deception was used in the experiment and we screened language model responses for any sensitive content before passing them to the experts (although we did not encounter any). Participants were not compensated for participation and gave consent to be identified. All appropriate IRB protocols in providing instructions and gathering consent were followed.

Acknowledgements

Research reported in this publication was supported by The Eunice Kennedy Shriver National Institute of Child Health & Human Development of the National Institutes of Health under Award Number P2CHD047879. Additionally, this material is based upon work supported by the National Science Foundation under Grant # 2127309 to the Computing Research Association for the CIFellows 2021 Project. This work would not have been possible without all the experts generously donating their busy time. We are extremely grateful to them.

References

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 610–623.

Ruha Benjamin. 2019. *Race after technology: Abolitionist tools for the new jim code*. Polity, Cambridge.

Timothy W Bickmore, Ha Trinh, Stefan Olafsson, Teresa K O'Leary, Reza Asadi, Nathaniel M Rickles, and Ricardo Cruz. 2018. Patient and consumer safety risks when using conversational assistants for medical information: an observational study of siri, alexa, and google assistant. *Journal of medical Internet research*, 20(9):e11510.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens

Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

B Barla Cambazoglu, Valeriia Baranova, Falk Scholer, Mark Sanderson, Leila Tavakoli, and Bruce Croft. 2021. Quantifying human-perceived answer utility in non-factoid question answering. In *Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, pages 75–84.

Yao Dou, Maxwell Forbes, Rik Koncel-Kedziorski, Noah A. Smith, and Yejin Choi. 2022. Is GPT-3 text indistinguishable from human text? scarecrow: A framework for scrutinizing machine text. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 7250–7274, Dublin, Ireland. Association for Computational Linguistics.

Harry G Frankfurt. 2005. *On bullshit*. Princeton University Press.

Nico Grant and Cate Metz. 2022. A new chat bot is a 'code red' for google's search business. *The New York Times*.

Daniel Martin Katz, Michael James Bommarito, Shang Gao, and Pablo Arredondo. 2023. Gpt-4 passes the bar exam. *Available at SSRN 4389233*.

Gary King and Jonathan Wand. 2007. Comparing incomparable survey responses: Evaluating and selecting anchoring vignettes. *Political Analysis*, 15(1):46–66.

Tiffany H Kung, Morgan Cheatham, Arielle Medenilla, Czarina Sillos, Lorie De Leon, Camille Elepaño, Maria Madriaga, Rimel Aggabao, Giezel Diaz-Candido, James Maningo, et al. 2023. Performance of chatgpt on usmle: Potential for ai-assisted medical education using large language models. *PLoS digital health*, 2(2):e0000198.

Rensis Likert. 1932. A technique for the measurement of attitudes. *Archives of psychology*, 22(140):5–55.

Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848*.

Andrew Loten. 2022. Chatty ai and protein-predicting algorithm defined the year in emerging tech. *The Wall Street Journal*.

Stephen Marche. 2022. The college essay is dead. *The Atlantic*.

Donald Metzler, Yi Tay, Dara Bahri, and Marc Najork. 2021. Rethinking search: making domain experts out of dilettantes. In *ACM SIGIR Forum*, volume 55, pages 1–27. ACM New York, NY, USA.

Arvind Narayanan and Sayash Kapoor. 2022. Chatgpt is a bullshit generator. but it can still be amazingly useful. *AI Snake Oil*.

Safiya Umoja Noble. 2018. Algorithms of oppression. In *Algorithms of Oppression*. New York University Press.

Alexis Palmer and Arthur Spirling. 2023. Large language models can argue in convincing and novel ways about politics: Evidence from experiments and human judgement. Technical report, Working paper), Technical report.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Chirag Shah and Emily M Bender. 2022. Situating search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval*, pages 221–232.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, and et. al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Chris Stokel-Walker. 2023. Chatgpt listed as author on research papers: many scientists disapprove. *Nature*, 613(7945):620–621.

Latanya Sweeney. 2013. Discrimination in online ad delivery. *Communications of the ACM*, 56(5):44–54.

Alex Wang, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2019. GLUE: A multi-task benchmark and analysis platform for natural language understanding. In *7th International Conference on Learning Representations, ICLR 2019, New Orleans, LA, USA, May 6-9, 2019*. OpenReview.net.

Eddie Yang and Margaret E Roberts. 2021. Censorship of online encyclopedias: Implications for nlp models. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, pages 537–548.

A Appendix

A.1 Study Participants

The experts who wrote and evaluated these questions are enumerated in Table 2.

A.2 Reproducibility

We provided three reproductions of the first questions for evaluation. An example is provided in Table 3.

A.3 Data Examples

We provide example ChatGPT and YouChat answers for the opera and quantum information examples in Table 4.

A.4 Data Collection

These questions were sourced through the form shown in Figures 3 and evaluated through the form in Figure 4.

A.5 Additional Data

We release our data publicly at
`https://denis.ai/bot_thoughts/`.

Speciality	Name
Family Medicine	Dr. D. Andrew MacFarlan
Radiology	Dr. Dionysios Koroulakis
Toxicology	Dr. David Schaffer
Geoarchaeology	Dr. Timothy Beach
Neuroscience	Elizabeth Tawa
Quantum Info	Dr. Andrew Guo
Linguistics	Dr. Adele Goldberg
Digital Humanities	Dr. Ryan Heuser
Ballroom	Dan Calloway
Opera	Fred Plotkin

Table 2: Participants and their specialties.

REGENERATION OF QUESTION

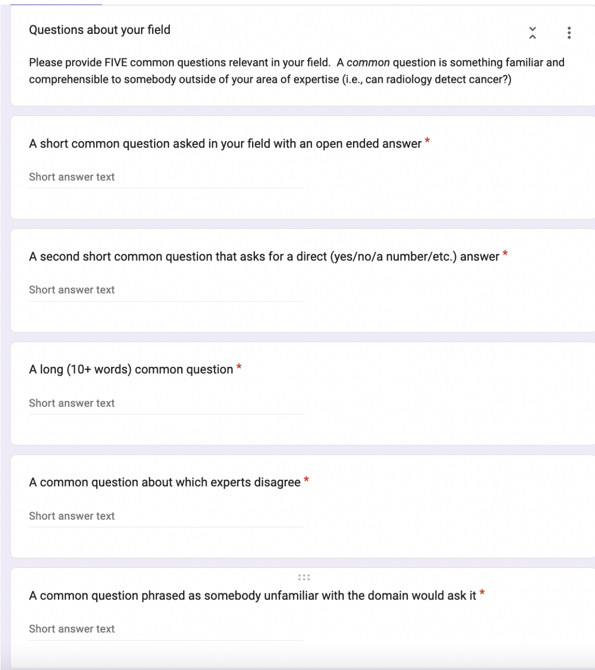
Human: "Can imaging studies help diagnose traumatic injuries?"

ChatGPT: *"Yes, imaging studies such as X-rays, CT scans, and MRI can be used to help diagnose traumatic injuries. They can be used to identify broken bones, internal bleeding, and other injuries that may not be visible on the surface of the body. Imaging studies can also be used to monitor the healing process and determine the severity of an injury."*

ChatGPT: *"Imaging studies such as X-rays, CT scans, and MRIs can help diagnose traumatic injuries by providing detailed images of the affected area. These images can reveal bone fractures, dislocations, and injuries to internal organs and soft tissue. They can also be used to monitor the healing process and detect complications. In some cases, a combination of imaging studies may be needed to fully evaluate the extent of an injury."*

ChatGPT: *"Yes, imaging studies such as X-rays, CT scans, and MRI scans can help diagnose traumatic injuries. These imaging techniques can provide detailed pictures of the bones, soft tissue, and internal organs, which can help physicians identify and diagnose injuries such as broken bones, head injuries, internal bleeding, and other conditions. However, it's important to note that imaging studies are not always necessary for the diagnosis of traumatic injuries and the decision to use imaging should be based on the individual patient's history and clinical presentation."*

Table 3: Regeneration of a question by ChatGPT.


Questions about your field

Please provide FIVE common questions relevant in your field. A *common* question is something familiar and comprehensible to somebody outside of your area of expertise (i.e., can radiology detect cancer?)

A short common question asked in your field with an open ended answer *

Short answer text

A second short common question that asks for a direct (yes/no/a number/etc.) answer *

Short answer text

A long (10+ words) common question *

Short answer text

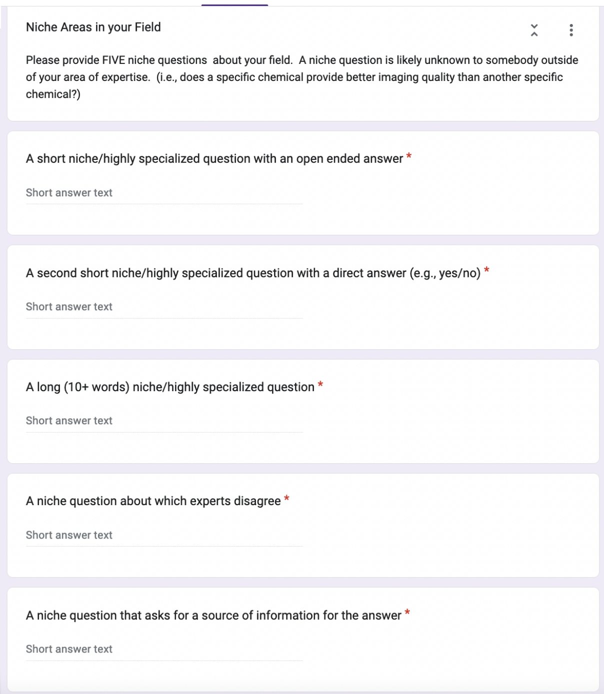
A common question about which experts disagree *

Short answer text

A common question phrased as somebody unfamiliar with the domain would ask it *

Short answer text

(a) Prompts for common questions.


Niche Areas in your Field

Please provide FIVE niche questions about your field. A niche question is likely unknown to somebody outside of your area of expertise. (i.e., does a specific chemical provide better imaging quality than another specific chemical?)

A short niche/highly specialized question with an open ended answer *

Short answer text

A second short niche/highly specialized question with a direct answer (e.g., yes/no) *

Short answer text

A long (10+ words) niche/highly specialized question *

Short answer text

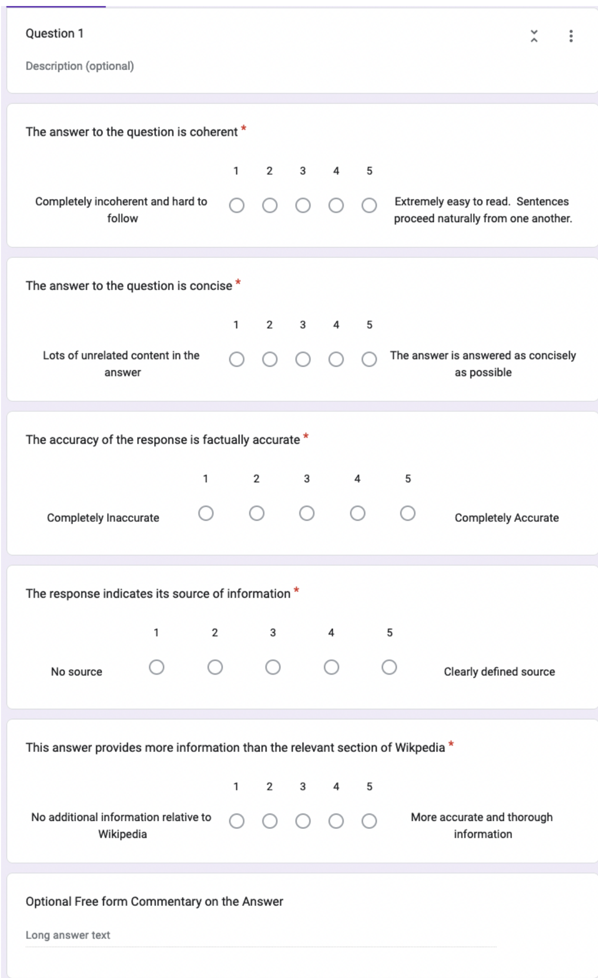
A niche question about which experts disagree *

Short answer text

A niche question that asks for a source of information for the answer *

Short answer text

(b) Prompts for niche questions.


Question 1

Description (optional)

The answer to the question is coherent *

1 2 3 4 5

Completely incoherent and hard to follow — Extremely easy to read. Sentences proceed naturally from one another.

The answer to the question is concise *

1 2 3 4 5

Lots of unrelated content in the answer — The answer is answered as concisely as possible

The accuracy of the response is factually accurate *

1 2 3 4 5

Completely Inaccurate — Completely Accurate

The response indicates its source of information *

1 2 3 4 5

No source — Clearly defined source

This answer provides more information than the relevant section of Wikpedia *

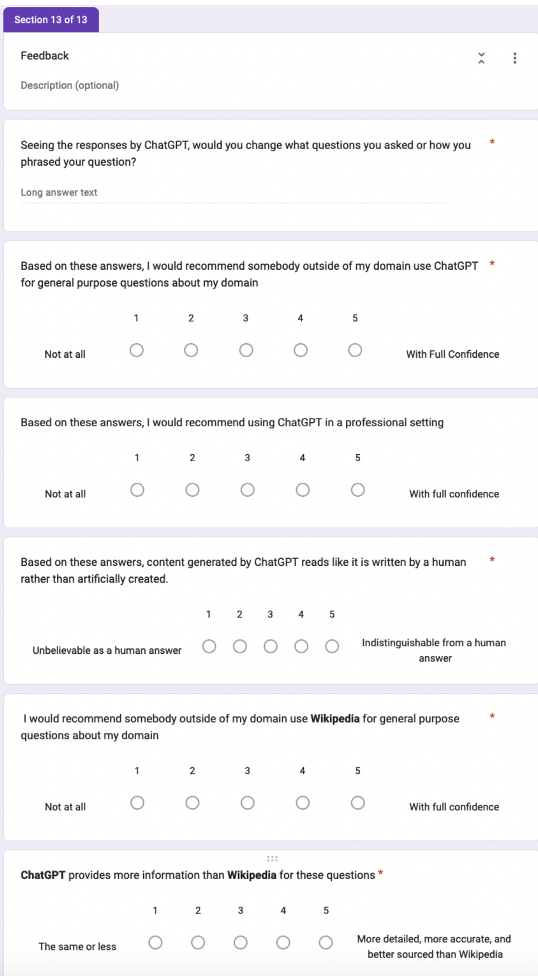
1 2 3 4 5

No additional information relative to Wikipedia — More accurate and thorough information

Optional Free form Commentary on the Answer

Long answer text

(a) Feedback is provided along a Likert scale for each question.


Section 13 of 13

Feedback

Description (optional)

Seeing the responses by ChatGPT, would you change what questions you asked or how you phrased your question? *

Long answer text

Based on these answers, I would recommend somebody outside of my domain use ChatGPT for general purpose questions about my domain *

1 2 3 4 5

Not at all — With Full Confidence

Based on these answers, I would recommend using ChatGPT in a professional setting

1 2 3 4 5

Not at all — With full confidence

Based on these answers, content generated by ChatGPT reads like it is written by a human rather than artificially created. *

1 2 3 4 5

Unbelievable as a human answer — Indistinguishable from a human answer

I would recommend somebody outside of my domain use **Wikipedia** for general purpose questions about my domain *

1 2 3 4 5

Not at all — With full confidence

ChatGPT provides more information than **Wikipedia** for these questions *

1 2 3 4 5

The same or less — More detailed, more accurate, and better sourced than Wikipedia

(b) A summary judgement is made after all questions.

Human: "Was Maria Callas American?"

ChatGPT: *"No, Maria Callas was not American. Maria Callas, whose full name was Maria Anna Cecilia Sofia Kalogeropoulos, was born on December 2, 1923, in New York City, United States. However, she was of Greek descent. Her parents were Greek immigrants, and she later acquired Greek citizenship. Maria Callas is renowned as one of the greatest opera singers of the 20th century and was particularly known for her performances in the bel canto repertoire."*

YouChat: *"Yes, Maria Callas was American-born. She was born in Manhattan, New York City, to Greek immigrant parents. However, she later established her career in Italy and became known as a famous Greek soprano."*

Human: "Can quantum computers be realized in the physical world?"

ChatGPT: *"Quantum computers are a theoretical concept that has not yet been fully realized in the physical world. However, significant progress has been made in recent years in the development of quantum computing hardware, and several companies and research groups have built prototype quantum computers that are capable of running small-scale quantum algorithms. These prototypes are often referred to as "Noisy Intermediate-Scale Quantum" (NISQ) computers, as they are not yet capable of achieving the full potential of quantum computing, but they are able to perform some quantum computations that are beyond the capabilities of classical computers.*
There are several different physical implementations of quantum computing being researched, including superconducting qubits, trapped ions, topological qubits and so on. Each of these approaches has its own set of challenges and limitations, and it is not yet clear which approach will ultimately be the most successful in building large-scale, practical quantum computers.
It's worth to note that, currently, the most advanced quantum computers are still far from being able to perform any practical computations, and it's still uncertain when, or even if, fully-functional, large-scale, fault-tolerant quantum computers will be built."

YouChat: *"Yes, quantum computers can be realized in the physical world. Researchers have been able to build prototype systems that demonstrate the principles of quantum computing, and there are a number of research groups and companies working on building more advanced versions. Additionally, several companies have announced plans to build full-scale quantum computers in the next few years."*

Table 4: The full ChatGPT and YouChat responses to the two in-paper examples.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
LIMITATIONS

☑ A2. Did you discuss any potential risks of your work?
ETHICS

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract, 1

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☒ Did you use or create scientific artifacts?

Left blank.

☐ B1. Did you cite the creators of artifacts you used?
No response.

☐ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
No response.

☐ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
No response.

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
No response.

☐ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
No response.

☐ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
No response.

C ☒ Did you run computational experiments?

Left blank.

☐ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
No response.

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

□ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
No response.

□ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
No response.

□ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
No response.

D ☑ Did you use human annotators (e.g., crowdworkers) or research with human participants?
Ethics

☑ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
Appendix, Section 2

☑ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
Ethics / IRB

☑ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
Ethics/ IRB

☑ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
Ethics

□ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
Not applicable. Participants are identified directly. Appendix A