Beyond Accurate Answers: Evaluating Open-Domain Question Answering in Enterprise Search

Daniel Xiaodan Zhou
AWS AI Labs
Ann Arbor, MI, United States
danzhou@amazon.com

Lan Liu
AWS AI Labs
Santa Clara, CA, United States
liuall@amazon.com

Anmol Anubhai
AWS AI Labs
Seattle, WA, United States
anubhaia@amazon.com

Maansi Shandilya
AWS AI Labs
Seattle, WA, United States
maansis@amazon.com

Steph Sigalas
Amazon Inc
Seattle, WA, United States
stephsig@amazon.com

William Yang Wang
AWS AI Labs
Santa Clara, CA, United States
wyw@amazon.com

Zhiheng Huang
AWS AI Labs
Santa Clara, CA, United States
zhiheng@amazon.com

ABSTRACT

Open-domain question answering (OpenQA) research has grown rapidly in recent years. However, OpenQA usability evaluation in its real world applications is largely left under studied. In this paper, we evaluated the actual user experience of OpenQA model deployed in a large tech company's production enterprise search portal. From qualitative query log analysis and user interviews, our preliminary findings are: 1) There exists a large number of "contingency answers" that cannot be simply evaluated against their face textual values, due to noisy source passages and ambiguous query intents from short keywords queries. 2) Contingency answers contribute to positive search experience for providing "information scents". 3) Click-through-rate (CTR) is a good user-behavior metric to measure OpenQA result quality, despite the rare existence of "good abandonment". This exploratory study reveals an often neglected gap between existing OpenQA research and its search engine applications that disconnects the offline research effort with online user experience. We call for reformulating OpenQA model objective beyond answer face value and developing new dataset and metrics for better evaluation protocols.

ACM Reference Format:
Daniel Xiaodan Zhou, Lan Liu, Anmol Anubhai, Maansi Shandilya, Steph Sigalas, William Yang Wang, and Zhiheng Huang. 2023. Beyond Accurate Answers: Evaluating Open-Domain Question Answering in Enterprise Search. In *ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '23), March 19–23, 2023, Austin, TX, USA.* ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3576840.3578314

CHIIR '23, March 19–23, 2023, Austin, TX, USA

ACM ISBN 979-8-4007-0035-4/23/03.
https://doi.org/10.1145/3576840.3578314

1 INTRODUCTION

Open-domain question answering (OpenQA) [25] aims to provide answers in response to a user's question based on a large collection of documents. It has been an important and long-standing task in natural language understanding (NLP) and information retrieval (IR). Recent OpenQA systems often follow a two stage framework: given an input question, a *retriever* first selects a small subset of relevant passages that potentially contain the answer; afterwards, a *reader* processes the retrieved contexts and extracts or generates answers [4, 30]. OpenQA applications are typically found in two categories. First, chatbots, dialog systems, and virtual assistants such as *Alexa* use OpenQA to respond to users' natural language questions with factoid or descriptive answers [5, 22]. Second, modern search engines use OpenQA to provide direct answers as a part of the search engine result page (SERP), usually in front of the "10-blue-links" document ranking results [12].

While OpenQA research is fast growing, most recent work focuses on improving answer accuracy from retrieval and reader architecture [10], and there is a lack of work evaluating the actual user experience of OpenQA in real world search engine applications. Inspired by the emerging trends of "Human-Centered AI" [14, 29] and "Data-Centric AI" [16, 24], this paper aims to discuss the open problems surfaced from evaluating the end user experience of OpenQA deployed to a large tech company's production enterprise search portal.

The motivation of this study is twofold. First, we observed that accurate OpenQA answers do not necessarily lead to the best user experience. As shown in Figure 1, for a user query "*can i bring dogs to #office_building*", OpenQA returned a seemingly perfect answer "*no dogs are permitted*". However, users reported that the answer came from an obscure and unauthoritative newsletter from five years ago and may not be trusted any longer. Users could not simply take the answer at its face value, and still needed to search more evidence to verify. This indicates a potential gap with current OpenQA research, where an answer is evaluated purely based on its face textual value using EM/F1 metrics [21]. Second, to complement

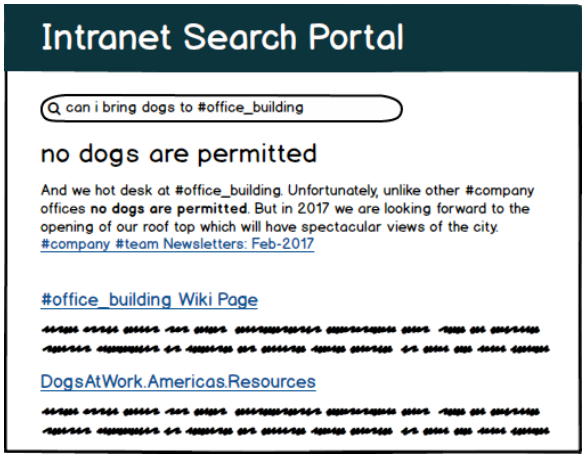

Figure 1: OpenQA answer from a large tech company's production intranet search engine[1].

EM/F1 for OpenQA evaluation in the real world, we investigated online user behavior metrics such as click-through-rate (CTR) that are widely used in search engine evaluation [27]. However, it is important to validate whether CTR could be compromised by "good abandonment" [6, 13], where users abandon search sessions without clicking since their information needs are already met by reading the direct answers from OpenQA on SERP.

Prior work partially addressed our motivating research questions on answers' usefulness [3] and taxonomy [2], the dilemma between trust and convenience [19], and answers' SERP user behavior [28]. This work, however, aims to add on to existing literature by evaluating OpenQA in its real world enterprise search settings. Our scope is limited to providing exploratory and preliminary findings while a more rigorous study is deferred to future work. We believe that these early insights could inspire new research to improve OpenQA design and applications.

2 SYSTEM OVERVIEW

2.1 Extractive OpenQA

We built an OpenQA model following the standard retriever-reader architecture [26]. During offline ingestion, each document is split into passages to build a sparse passage index. During inference, we first perform a BM25 [23] based retrieval to obtain top 100 passages, which are then reranked by a learning-to-rank BERT [8] model (*BERT-PR*) with cross-attention between query and passages. At the last reader stage, another BERT model (*BERT-RC*) is used to extract answer spans from the top 5 passages from *BERT-PR* [26]. *BERT-PR* is trained with MSMARCO [17], and *BERT-RC* is trained with MRQA [9] augmented by a small internally annotated dataset. The training objective follows the typical OpenQA framework to optimize the answer's EM/F1 score based on its face value. The dataset consists of long natural language questions with factoid answers.

[1]For privacy protection, we use a mockup that mimics the actual SERP. We also anonymize all entities.

2.2 Enterprise search application

We collaborated with a big tech company and deployed the OpenQA model in its intranet search portal. The search engine indexed multiple sources from the company's intranet, with a total of 15.7 million documents including wiki pages, developer documentation, blog posts, tech Q&A, educational videos, mailing list archives, HR policies, and many others. On a typical work day, the search engine has an average of 90K to 120K queries from 18K to 24K unique users, where 80% of queries are from users in the software development roles.

For each query, OpenQA model generates answer predictions, and only answers above a certain confidence threshold are displayed to the end users. Answers are placed on the top of SERP before document ranking results, as shown in Figure 1.

3 METHODOLOGY

Driven by the motivation to evaluate the user experience of OpenQA, we designed an exploratory study and adopted a) query log analysis to evaluate answer quality, and b) user interviews to collect qualitative feedback.

3.1 Query log analysis

We dumped all 5.7 million queries from the search engine's live traffic between Jul.1 and Sep.12, 2021, where each query has its associated document ranking and OpenQA results along with user clicks data. The average query length is 3.7 tokens, and 62% of queries have no more than 3 tokens, indicating that real world enterprise search is likely dominated by short keyword queries. To focus on OpenQA study, we removed the queries without answers returned, and a subset of 380K queries, or 7%, remained. The resulting OpenQA queries were categorized into 2 groups based on user clicks pattern: 1) *HAS-CLICK*, including 22.3% queries that received clicks on OpenQA source documents, and 2) *NO-CLICK*, including 77.7% queries that received no OpenQA click.

We randomly sampled 200 query-answer pairs from each of the 2 groups. In order to better represent head queries, we additionally sampled 100 random unique pairs with 5+ occurrences and another 100 most popular pairs. For these 600 query-answer pairs, we carefully conducted qualitative content analysis [11] to gain insights and came up with a spectrum of "answer quality types", to be discussed in *Findings*.

3.2 User interviews

We conducted five semi-structured interviews [15] and in-depth cognitive walk-through sessions with the actual users of the search engine, including three software developers, one manager, and one business analyst. Participants were randomly recruited from the company's UX Research channels, and the interviews were conducted through video conferencing in multiple 30-60 minutes sessions. We recorded all research sessions, coded the qualitative findings and distilled insights. The key user research questions we asked were: 1) How reliable / not reliable do users find OpenQA results? 2) What is users' definition of a "good answer"? 3) What is users' definition of a trustworthy answer? 4) How useful / not useful do users find the search engine when OpenQA is triggered?

5) When unsure about the accuracy of search result, what types of verification processes do users follow?

4 PRELIMINARY FINDINGS

4.1 Answers are more nuanced beyond face textual values

OpenQA models are typically evaluated by EM/F1 score, where the answer quality fully depends on its overlapping with ground truth answers. However, query log analysis indicated that a large number of answers can not be evaluated solely based on their face textual values, but can depend on other factors such as recency or trustworthiness. More concretely, we categorized answers into eight "answer quality types", as shown in Table 1.

Of the eight types, *PERFECT* and *IRRELEVANT* map to the conventional "correct vs incorrect" answers respectively. We define the other six types collectively as **contingency answers** – they do not perfectly answer the queries, but they are not entirely wrong either. To the end users, whether an answer is correct or not depends on other "contingencies" beyond its textual value.

For example, query-answer pair "*can i bring dogs to #office_building → no dogs are permitted*" (labeled as *TO-BE-VERIFIED*) would have been 100% perfect with EM/F1 score. However, the answer came from an outdated newsletter and users did not consider it trustworthy. EM/F1 failed to capture the critical signals on answer's authoritativeness and recency. For another example, "*preferred browser → Chrome*" (labeled as *SITUATIONAL*) was more of a subjective opinion to a particular team than a fact, and was only correct under certain situations, where EM/F1 failed to capture.

We hypothesize that contingency answers exist in enterprise search scenario due to two reasons. First, low quality contents are abundant and OpenQA answers could be extracted from personal opinions or outdated sources, leading to unauthoritative or obsolete answers. Second, search queries are predominantly short keyword queries with broad and ambiguous intents that are hard to expect perfect answers. We'll elaborate these thoughts in *Discussions*.

4.2 Contingency answers contribute to positive search experience

Early evidence from user interviews suggest that contingency answers, while imperfect, still improve search user experience, which is consistent to [28]. First, contingency answers might provide "information scent" [18] to help users quickly navigate to relevant sources. For example, one interviewee shared her experience that, to the query-answer pair "*#ML_model accuracy → >95%*" (labeled as *TO-BE-VERIFIED*), the answer stood out despite of its correctness because it was a "straight number" and appeared more relevant than other search results.

Second, contingency answers can complement "10-blue-links" document ranking (DR) results by surfacing new content. Only 48% of OpenQA answers in our dataset came from the same top-10 DR results in their respective SERP. This is due to that OpenQA model predicts answers directly from local, specific passage context instead of globally relevant documents. One interviewee emphasized that non-overlapping OpenQA answers complemented top DR results and were useful and informative in SERP.

Third, contingency answers can impress users and earn trust, which is an important factor that leads to successful adoption of enterprise search engine [7]. We observed that OpenQA answers could usually solicit the "wow" magical moments from the interviewees – it is a convincing instrument to demonstrate the strength of modern intelligent search and encourage adoption.

4.3 Click-through-rate (CTR) indicates answer quality

To complement accuracy-oriented offline OpenQA evaluation using EM/F1, we investigated CTR online user behavior metrics for the search engine application. We observed that "good abandonment" that *PERFECT* answers result in users abandoning search sessions without click [6, 13] were actually rare (3% in our data samples). Instead, the majority of contingency answers were likely to invite clicks if users still found them useful. On a larger scale, we hypothesize that "good abandonment" is negligible due to the scarcity of *PERFECT* answers in the enterprise search use case.

Furthermore, we found that *IRRELEVANT* answers appeared more often in the *NO-CLICK* group than in the *HAS-CLICK* group (49.5% vs 36.5%), implying that irrelevant answers were less likely to get clicks and vice versa. To sum up, we tentatively suggest that CTR can indicate answer quality, where better answers are more likely to receive higher CTR.

5 DISCUSSIONS AND CONCLUSION

First, it's critical to encourage answers from authoritative and up-to-date sources. Typical OpenQA architecture is optimized to find correct answer spans, assuming all source passages have equally high quality (eg, extracted from Wikipedia articles). However, real world passages can be noisy, outdated, and untrustworthy, and it could lead to catastrophic consequences if users blindly trust obsolete answers when such answers are crucial to high stake tasks [19, 24]. We argue that it's important for OpenQA models to incorporate recency and authoritativeness in addition to pure answer text into its evaluation .

Second, it could be an important OpenQA model objective to predict multiple answers in order to increase answer diversity and recall. Existing OpenQA datasets are dominated by factoid questions with unique specific answers (eg, "*who won Super Bowl 45 → Green Bay Packers*" from SQuAD). However, real world queries often seek relevant information that may be diverse or even conflicting. For example, questions can be about personal opinions (eg, "*preferred browser → Chrome*"), or can be transient in nature (eg, "*#team office hour → 2019/06/19*"). One single answer could be incomplete, misleading, biased, or unfair.

Third, the dichotomies of "answerable vs unanswerable" [1, 20] or "correct vs incorrect" answers, while intuitive, are often oversimplifications to the more nuanced use cases in real world search applications. One might argue that contingency answers from short keywords queries such as "*average cold start time #product → 150ms*" was actually "unanswerable" due to being situational. However, for the search engine use case, contingency answers can still improve users search experience by providing "information scent", and disregarding them as "unanswerable" would be a missed opportunity.

Table 1: OpenQA Answer Quality Types (from 600 question-answer pairs)

Answer Quality Type	Query	Answer + Passage (Abbr.)
PERFECT - Answer completely meet searchers' information need; search activity can stop here.	TIL stands for	Every Thursday, selected team members share a "Today I Learned" (TIL) , which can be an interesting fact, anecdote, or learning that they encountered over the previous week.
TO-BE-VERIFIED - Answers require further validation due to their questionable authoritativeness; or, questions require precise answers that deserve careful verification.	can i bring dogs to *#office_building*	Unfortunately, unlike other *#company* offices no dogs are permitted. But in 2017 we are looking forward to the opening of our roof top ... (*Source*: a 5-year-old newsletter.)
	#ML_model accuracy	As per *#team* analysis *#ML_model* gave an accuracy of >95% from purely translation point and further considering it for matching it is expected to have even higher accuracy. (*Source*: a personal blog post.)
SITUATIONAL - The correctness of answers depend on query context. Eg, asking for *office hour* and the correct answer should be a future date, not in the past.	*#team* office hours	Update [2019/05/27]: ... Office hours for *#team* is set for 2019/06/19.
	preferred browsers	Is there a preferred browser to use with the *#product*? Yes, Chrome is the preferred browser for use with *#product*, but we have had good results with Firefox and other browsers as well.
INCOMPLETE - Answers only partially address information seeking queries. Users typically need to continue search to learn more.	python web scraping	Here are some links for getting started with Scrapy, a Python web scraping framework. Scrapy Tutorial...
WHAT-IS - Query *X* gets translated into *what is X*, but query intent might be much broader (eg, *why X* or *how to do X*).	self review	What is a self review? It is a summary of your key accomplishments, metrics, strengths, and development areas over the previous 12 months. Why do a self review? How is a self review used?
RELEVANT-DOC - Answers are irrelevant, but their source passages or documents are still relevant, and is still a good search result.	Target "release" does not exist in the project	... failed in the pipeline with the error message: Target "release" does not exist in the project "*#project_name*". I checked in the config file this dependency is there. How do I fix this? (*Source*: tech FAQ)
PLAUSIBLE - Answers are seemingly correct but are actually not, the same definition as in SQuAD 2.0 [20].	how to subscribe email list using team email	Our team email-list - team. Subscribe to this list. (Enter your username under the "Subscription Info" on the right of the page, and hit "Subscribe")
	#functions taking long time to build	Our pipeline includes several *#functions* that take a total build time of almost 18 minutes, the longest of which is around 7 minutes.
IRRELEVANT - Neither the answers, passages, nor the documents are relevant. Searchers learn nothing from the result.	create file in sftp server	SFTP/SFTP > Map to Remote... You will enter your connection parameters and a new file will be created named sftp-config.json. (*Source*: personal blog post)

We suggest that OpenQA research take into account "answer quality types" or answer taxonomies [2] to determine the best model behavior for such nuanced query cases.

Fourth, existing OpenQA dataset such as Natural Questions [12] might not generalize well to the real world enterprise search use case, because – *a).* real world queries, as shown from our study, are primarily short keyword queries instead of long natural language queries typically found in OpenQA datasets; *b).* a large number of real world queries tend to be ambiguous, seeking descriptive answers and diverse opinions as opposed to the short factoid answers;

c). real world corpus is far from clean, but rather is very noisy and contains rich metadata such as creation date, content type (eg, HR policies vs personal blogs) and so on, not seen in existing datasets. We raise the open problem to create proper datasets to bridge the gap for the real world enterprise search applications.

Fifth, user-oriented online evaluation of OpenQA on a large scale is limited by the lack of well-established metrics due to the "good abandonment" phenomenon, but initial evidence from this paper suggests that CTR might still be valid. Future work is needed to establish CTR or other user behavior metrics in order to complement EM/F1 accuracy-oriented offline evaluation for OpenQA research.

To conclude, by conducting exploratory query log analysis and user interviews, we revealed the limitations of "model-centric" OpenQA research such as lack of noisy corpus and under representation of short keyword queries, and showed that OpenQA evaluation in the real world should go beyond answer accuracy. We call for reformulating OpenQA model objective beyond answers' face value and developing new dataset and metrics for better evaluation protocols to improve real world OpenQA applications.

ACKNOWLEDGMENTS

We greatly appreciate the anonymous reviewers for their valuable comments and suggestions.

REFERENCES

[1] Akari Asai and Eunsol Choi. 2020. Challenges in Information Seeking QA:Unanswerable Questions and Paragraph Retrieval. (2020). arXiv:2010.11915 http://arxiv.org/abs/2010.11915

[2] Valeriia Bolotova, Vladislav Blinov, Falk Scholer, W. Bruce Croft, and Mark Sanderson. 2022. A Non-Factoid Question-Answering Taxonomy. *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*, 1196–1207. https://doi.org/10.1145/3477495.3531926

[3] 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. *CHIIR 2021 - Proceedings of the 2021 Conference on Human Information Interaction and Retrieval*, 75–84. https://doi.org/10.1145/3406522.3446028

[4] Danqi Chen, Adam Fisch, Jason Weston, and Antoine. Bordes. 2017. Reading Wikipedia to answer opendomain questions. *In Association for Computational Linguistics (ACL)* (2017), 1870–1879.

[5] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen Tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2020. QUAC: Question answering in context. *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing, EMNLP 2018* (2020), 2174–2184. https://doi.org/10.18653/v1/d18-1241 arXiv:1808.07036

[6] Aleksandr Chuklin and Pavel Serdyukov. 2012. Good abandonments in factoid queries. *WWW'12 - Proceedings of the 21st Annual Conference on World Wide Web Companion* (2012), 483–484. https://doi.org/10.1145/2187980.2188088

[7] Paul H. Cleverley and Simon Burnett. 2019. Enterprise search: A state of the art. *Business Information Review* 36, 2 (2019), 60–69. https://doi.org/10.1177/0266382119851880

[8] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina. Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. *In North American Association for Computational Linguistics (NAACL)* (2019).

[9] Adam Fisch, Alon Talmor, Robin Jia, Minjoon Seo, Eunsol Choi, and Danqi Chen. 2019. MRQA 2019 shared task: Evaluating generalization in reading comprehension. *MRQA@EMNLP 2019 - Proceedings of the 2nd Workshop on Machine Reading for Question Answering* (2019), 1–13. https://doi.org/10.18653/v1/d19-5801 arXiv:1910.09753

[10] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Ledell Wu, Sergey Edunov, Danqi Chen, and Wentau. Yih. 2020. Dense passage retrieval for open-domain question answering. *Empirical Methods in Natural Language Processing (EMNLP 2020)* (2020), 1870–1879.

[11] Klaus Krippendorff. 1989. Content Analysis: An Introduction to its methodology. 1 (1989), 403–407.

[12] Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics* 7 (2019), 453–466. https://doi.org/10.1162/tacl_a_00276

[13] Jane Li, Scott Huffman, and Akihito Tokuda. 2009. Good abandonment in mobile and PC internet search. *Proceedings - 32nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR 2009* (2009), 43–50. https://doi.org/10.1145/1571941.1571951

[14] Gina Neff, Manuel Gomez Rodriguez, and Adrian Weller. 2021. NeurIPS 2021 Workshop Proposal : Human Centered AI. November (2021). https://doi.org/10.13140/RG.2.2.33952.51206

[15] Kathryn E Newcomer, Harry P Hatry, and Joseph S Wholey. 2015. Conducting semi-structured interviews. *Handbook of practical program evaluation* 492 (2015), 492.

[16] Andrew Ng. 2021. *MLOps: From Model-centric to Data-centric AI.* Technical Report. https://www.deeplearning.ai/wp-content/uploads/2021/06/MLOps-From-Model-centric-to-Data-centric-AI.pdf

[17] Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. MS MARCO: A human generated MAchine reading COmprehension dataset. *CEUR Workshop Proceedings* 1773 (2016), 1–10. arXiv:1611.09268

[18] Peter Pirolli and Stuart Card. 1999. Information foraging. *Psychological Review* 106, 4 (1999), 643–675. https://doi.org/10.1037/0033-295X.106.4.643

[19] Martin Potthast, Matthias Hagen, and Benno Stein. 2020. The dilemma of the direct answer. *ACM SIGIR Forum* 54, 1 (2020), 1–12. https://doi.org/10.1145/3451964.3451978

[20] Pranav Rajpurkar, Robin Jia, and Percy Liang. 2018. Know what you don't know: Unanswerable questions for SQuAD. *ACL 2018 - 56th Annual Meeting of the Association for Computational Linguistics, Proceedings of the Conference (Long Papers)* 2 (2018), 784–789. https://doi.org/10.18653/v1/p18-2124 arXiv:arXiv:1806.03822v1

[21] Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuad: 100,000+ questions for machine comprehension of text. In *EMNLP 2016 - Conference on Empirical Methods in Natural Language Processing, Proceedings*. 2383–2392. arXiv:arXiv:1606.05250v3

[22] Siva Reddy, Danqi Chen, and Christopher D. Manning. 2019. CoQA: A Conversational Question Answering Challenge. *Transactions of the Association for Computational Linguistics* 7 (2019), 249–266. https://doi.org/10.1162/tacl_a_00266 arXiv:1808.07042

[23] Stephen Robertson and Zaragoza Hugo. 2009. The probabilistic relevance framework: BM25 and beyond. *Foundations and Trends in Information Retrieval* 3, 4 (2009), 333–389. https://doi.org/10.1561/1500000019

[24] Nithya Sambasivan, Shivani Kapania, and Hannah Highfll. 2021. Everyone wants to do the model work, not the data work: Data cascades in high-stakes AI. *Conference on Human Factors in Computing Systems - Proceedings* (2021). https://doi.org/10.1145/3411764.3445518

[25] Ellen M Voorhees. 1999. The TREC-8 question answering track report. *TREC* 99 (1999), 77–82.

[26] Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage BERT: A Globally Normalized BERT Model for Open-domain Question Answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 5877–5881. http://arxiv.org/abs/1908.08167

[27] Ryen W White. 2016. *Interactions with Search Systems.* https://www.amazon.com/Interactions-Search-Systems-Ryen-White/dp/1107034221

[28] Zhijing Wu, Mark Sanderson, B. Barla Cambazoglu, W. Bruce Croft, and Falk Scholer. 2020. Providing Direct Answers in Search Results: A Study of User Behavior. *International Conference on Information and Knowledge Management, Proceedings*, 1635–1644. https://doi.org/10.1145/3340531.3412017

[29] Wei Xu. 2019. Toward Human-Centered AI: A Perspective from Human-Computer Interaction. *Interactions* April 2019 (2019). https://dl.acm.org/doi/10.1145/3328485

[30] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua. 2021. Retrieving and Reading: A Comprehensive Survey on Open-domain Question Answering. (2021), 1–21. arXiv:2101.00774 http://arxiv.org/abs/2101.00774