

Walert: Putting Conversational Search Knowledge into Action by Building and Evaluating a Large Language Model-Powered Chatbot

Sachin Pathiyan Cherumanal
RMIT University
Melbourne, Australia
s3874326@student.rmit.edu.au

Angel Felipe Magnossão de Paula
Universitat Politècnica de València
Valencia, Spain
adepau@doctor.upv.es

Danula Hettiachchi
RMIT University
Melbourne, Australia
danula.hettiachchi@rmit.edu.au

Lin Tian
RMIT University
Melbourne, Australia
lin.tian2@student.rmit.edu.au

Kaixin Ji
RMIT University
Melbourne, Australia
kaixin.ji@student.rmit.edu.au

Johanne R. Trippas
RMIT University
Melbourne, Australia
j.trippas@rmit.edu.au

Damiano Spina
RMIT University
Melbourne, Australia
damiano.spina@rmit.edu.au

Futoon M. Abushaqra
RMIT University
Melbourne, Australia
futoon.abu.shaqra@student.rmit.edu.au

Halil Ali
RMIT University
Melbourne, Australia
halil.ali@rmit.edu.au

Falk Scholer
RMIT University
Melbourne, Australia
falk.scholer@rmit.edu.au

ABSTRACT

Creating and deploying customized applications is crucial for operational success and enriching user experiences in the rapidly evolving modern business world. A prominent facet of modern user experiences is the integration of chatbots or voice assistants. The rapid evolution of Large Language Models (LLMs) has provided a powerful tool to build conversational applications. We present Walert, a customized LLM-based conversational agent able to answer frequently asked questions about computer science degrees and programs at RMIT University. Our demo aims to showcase how conversational information-seeking researchers can effectively communicate the benefits of using best practices to stakeholders interested in developing and deploying LLM-based chatbots. These practices are well-known in our community but often overlooked by practitioners who may not have access to this knowledge. The methodology and resources used in this demo serve as a bridge to facilitate knowledge transfer from experts, address industry professionals' practical needs, and foster a collaborative environment. The data and code of the demo are available at <https://github.com/rmit-ir/walert>.

CCS CONCEPTS

• **Information systems** → **Evaluation of retrieval results**; Search interfaces; • **Human-centered computing** → *Natural language interfaces*.

KEYWORDS

conversational information seeking, large language models, retrieval-augmented generation

ACM Reference Format:

Sachin Pathiyan Cherumanal, Lin Tian, Futoon M. Abushaqra, Angel Felipe Magnossão de Paula, Kaixin Ji, Halil Ali, Danula Hettiachchi, Johanne R. Trippas, Falk Scholer, and Damiano Spina. 2024. Walert: Putting Conversational Search Knowledge into Action by Building and Evaluating a Large Language Model-Powered Chatbot. In *Proceedings of the 2024 ACM SIGIR Conference on Human Information Interaction and Retrieval (CHIIR '24)*, March 10–14, 2024, Sheffield, United Kingdom. ACM, New York, NY, USA, 5 pages. <https://doi.org/10.1145/3627508.3638309>

1 INTRODUCTION

Conversational agents based on Large Language Models (LLMs) such as OpenAI's ChatGPT¹ provide many benefits to stakeholders, reducing the cost of various tasks by saving time and resources, especially for closed-domain questions such as translating responses into different formats, retrieving policy documents, and preparing legal document drafts [5]. However, two major concerns arise in this

Permission to make digital or hard copies of part or all of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for third-party components of this work must be honored. For all other uses, contact the owner/author(s).

CHIIR '24, March 10–14, 2024, Sheffield, United Kingdom

© 2024 Copyright held by the owner/author(s).

ACM ISBN 979-8-4007-0434-5/24/03.

<https://doi.org/10.1145/3627508.3638309>

¹<https://openai.com/>

setting: (i) the risk of giving away sensitive data about the organization [7, 19]; and (ii) the limited access to structured and comprehensible documentation might impede practitioners (e.g., data scientists without a strong background on information retrieval) grasp of the principles, theories, and best practices influencing product quality. Efforts have been taken in this direction of bringing together researchers and experts in conversational information-seeking with industry practitioners to deal with real-world problems – as in the case of Amazon’s Alexa Prize challenges [1, 14, 16].

We aim to combine our diverse research expertise in machine learning, natural language processing, and information retrieval, to bring conversational search knowledge into action and address best practices for building an LLM-powered chatbot. Our goals are to explore the challenges and approaches for implementing a chatbot, compare various methods, and offer best practices in evaluation that stakeholders (not necessarily experts in conversational information seeking) can effectively use to assess and improve the quality of chatbot products.

Using a manually curated Frequently Asked Questions (FAQ) guide from RMIT University’s School of Computing Technologies as a Knowledge Base (KB), we were able to implement a conversational agent, named Walert², that allows potential future students to get answers to questions related to the computer science programs offered at RMIT University.³

In this demo, the primary focus was to characterize the challenges associated with integrating LLMs into the development process of voice-based conversational information-seeking systems based on an existing KB. The process allowed us to address challenges related to (i) handling private/sensitive information by deploying our instance of an open-source LLM; (ii) monitoring the problem of hallucinations (i.e., the introduction of facts that are not true) [9] and generation of inaccurate information by using a human-in-the-loop approach [6] and the inclusion of out of KB questions in our testbed; and (iii) evaluating the effectiveness of intent-based using Natural Language Understanding (NLU) and Retrieval-Augmented Generation (RAG), both at component and end-to-end levels. Our evaluation highlights shortcomings in recent RAG pipeline studies, particularly regarding the lack of ranking evaluation.

2 METHODOLOGY

Two different approaches were used to implement the prototype of our application: Intent-Based (IB) and RAG. The first one, IB, is suitable for structured and predictable interactions (i.e., pre-defined intents), while RAG chatbots retrieve information from a KB and are better for open domain and more dynamic conversations. Figure 1 presents the overall framework for both approaches.

2.1 Data Collection and Testbed

We utilized a manually curated FAQ from RMIT University’s School of Computing Technologies as a KB. The FAQ contains a wide range of common questions that incoming students ask regarding course

²The term “Walert” means “possum” in the native languages of the Woi Wurrung and Boon Wurrung peoples. Possum skin cloaks are essential to the Traditional Owners and Custodians of the land where the authors live and work. Our chatbot, Walert, is named as a tribute to this cultural heritage [4, 15].

³<https://www.rmit.edu.au/partner/hubs/race/news>

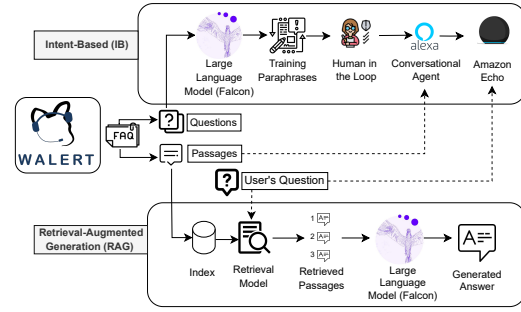


Figure 1: Overall architecture of the two approaches implemented in Walert: IB and RAG.

offerings and academic programs related to computer science. Using the question-answer pairs in the FAQ, we generate a set of questions Q and passages P . The question set $Q = \{q_1, q_2, \dots, q_n\}$ contains existing questions in the FAQ that have known answers (i.e., passages directly extracted from the FAQ), new questions with inferred answers (i.e., manually generated questions that have no direct answers but the answer can be inferred from multiple passages in the FAQ), and questions that do not have an answer in the KB (questions were manually generated by checking that they cannot be answered with the passages in the KB).

To simulate the scenario of different users expressing similar questions differently, we generated multiple semantically equivalent variations for each unique question initially included in the FAQ. The passages set $P = \{p_1, p_2, \dots, p_m\}$ is a corpus of passages extracted from the FAQ, representing our KB. Finally, each question is associated with a gold answer from a set $A = \{a_1, \dots, a_n\}$. Each answer a is obtained from one or more passages. We created a testbed that consists of relevance judgments at the passage level and gold answers for three types of questions:

Questions with Known Answers (Known). These questions have a direct answer in the FAQ. Therefore, the corresponding passage p to the answer a in the FAQ is judged as *Highly Relevant* (label = 2). All questions related to the same topic have the same passage judged as relevant, which is also the gold answer ($p = a$).

Questions with Inferred Answers (Inferred). Questions that do not have a direct answer in the FAQ, but have an answer that can be extracted from the KB, i.e., from one or more passages. Passages that partially contain relevant information to answer the question are judged as *Partially Relevant* (label = 1). The gold answer is manually generated by combining multiple passages.

Out-of-Knowledge Base Questions (Out of KB). These questions cannot be answered with the information available in the KB – even though, these questions are within the domain and likely to be asked. Therefore, there are no relevant passages, and the gold answer consists of communicating to the user that there is no information available to answer that question.

Table 1 shows example passages and answers for each question type. We have 106 questions (including variations) and 120 passages. In our collection, 84 questions have known answers (passages directly extracted from the FAQ), 12 have inferred answers, and 10 do not have an answer in the KB. These three question types in

our testbed provide a comprehensive evaluation that allows us to compare IB against RAG conversational approaches.

In particular, using the questions with known answers, we can assess the system’s ability to correctly respond to questions for which we know a passage in the KB contains the complete answer. By evaluating the effectiveness of the answers for the questions with inferred answers, we can assess the system’s ability to generate answers by combining multiple (partially) relevant passages. Finally, including questions not covered in the FAQ helps assess the chatbot’s ability to identify unanswerable questions – which is a critical step in controlling hallucinations.

2.2 Intent-Based (IB)

The IB approach consists of a conversational model built using Amazon Alexa Skills [16] (upper part of Figure 1). Each question in the FAQ is mapped to an *intent* in the conversational model, i.e., one of the possible actions recognizable by the system.⁴ Intuitively, this approach aims to optimize the correctness of the answers (high precision) but only handles a limited number of questions (low recall), i.e., those present in the FAQ (questions with known answers). Building effective intent recognition models requires multiple instances to train each intent. Since manually creating variations of training utterances is time-consuming, we experimented with using open-source LLMs to automatically create semantically equivalent variations of utterances (i.e., training data augmentation). In our case, these instances would be the semantic variations of the questions from the FAQ. We deployed Falcon-7B [2] in Amazon SageMaker Studio with a 5xlarge-GPU configuration⁵ to generate up to eight question variations for each intent (i.e., a question in the FAQ), using a zero-shot approach and the following prompt: “generate up to eight paraphrases of the following question: QUESTION”. After manually inspecting the variations generated, we established a threshold and selected the top five. These variations were then used to train the conversational model. We also normalized the instances by resolving the acronyms (e.g., replacing CS with Computer Science) and used them along with the original questions for training, making it a total of six training instances per the intent of the conversational model. The answers associated with the original questions in the FAQ were used as responses returned for each question. The Alexa Skill was finally deployed in an Amazon Echo (5th generation) device, which allowed users to interact with the system in an audio-only setting. To enable this, we utilized the Automatic Speech Recognition and NLU features built into Amazon Alexa Skill.

2.3 Retrieval-Augmented Generation (RAG)

In contrast to the precision-oriented IB approach, we sought to investigate a more open-ended methodology known as RAG [11]. Here, we cover two use cases: (i) instances where the questions the user raises vary from what is available in the FAQ, and (ii) scenarios

in which questions can only be answered using abstractive multi-document summarization from multiple passages in the KB. The RAG approach consists of two main stages (bottom part of Figure 1). The initial stage involves retrieving potential passages that may contain the answer, while the second stage involves generating a summary from the top- K retrieved passages.

For the retrieval model, we experimented with two approaches: (i) Okapi BM25 [17] with default parameters ($k_1 = 1.2$; $b = 0.75$) and (ii) dense retrieval using Dense Passage Retrieval (DPR) [10] implementations in the pyserini toolkit [13]. To generate a summary from the top- K retrieved passages, we used the same LLM that was used to generate semantically equivalent variations of the questions for the IB approach, i.e., falcon-7b-instruct (refer Section 2.2) along with the following prompt to generate the summaries:

Generate an answer to be synthesized with text-to-speech for a virtual assistant, the answer should be based on the retrieved documents for the following question. If the retrieved documents are not related to the question, then answer NA.

[QUESTION + LIST OF k PASSAGES]

We experimented with three top-heavy k cutoffs: 1 (which is comparable to IB), 3, and 5 top retrieved passages.

3 QUANTITATIVE EVALUATION

The test collection created from the knowledge base described in Section 2, allows us to apply effectiveness evaluation practices to compare our proposed approaches at both the component and end-to-end levels. Table 2 displays results for IB and RAG approaches using evaluation measures across two dimensions: (i) retrieval effectiveness (Normalized Discounted Cumulative Gain, NDCG [8]) and (ii) natural language generation (BERTScore [20] and ROUGE-1 [12]). ROUGE and BERTScore have been used to quantify hallucinations in LLMs automatically [9]. All the results in Table 2 have been tested for statistical significance using Tukey’s HSD and significance level $\alpha = 0.01$. Below, we discuss the results across two dimensions separately.

Retrieval Effectiveness. When it comes to retrieving a response, the IB approach only retrieves one passage (i.e., the response associated with the recognized intent), whereas RAG approaches retrieve multiple passages for a given question (and the final response is generated using the top- k passages). In our evaluation, we explore top-heavy cutoffs $k=\{1,3,5\}$ to minimize the risk of hallucinations. Table 2 shows that, for the Known questions, IB and RAG using DPR have comparable performance in terms of NDCG@1 and RAG approaches perform better with more aggressive ranking truncation ($k = 1$). For the Inferred questions, RAG approaches outperform IB (which can only return, at most, a passage partially relevant to the question). RAG approaches benefit from more context, and BM25 with $k = 3$ obtains the highest NDCG score. In terms of out-of-KB questions, IB performs substantially better than RAG approaches, being able to identify 80% of the unanswerable questions. RAG-based approaches fail by attempting to generate an answer for most of the questions, which means that it is likely to hallucinate instead of not warning the user about the lack of information in the KB.

End-to-end evaluation. BERTScore and ROUGE-1 scores indicate that IB performs significantly better for the Known questions, whereas RAG approaches are likely to generate better answers for

⁴<https://developer.amazon.com/en-US/docs/alexa/custom-skills/create-intents-utterances-and-slots.html>

⁵We explored different alternatives that would allow us to better understand the process of managing a privacy-aware LLM solution in-house to avoid the potential leaking of sensitive data by using third-party solutions. We found this alternative to be a good fit for our needs, also the most flexible for our future research, e.g., experiments involving fine-tuning.

Table 1: Examples of questions, relevant passages and gold answers in our testbed.

Question Type	Question	Passage(s)	Answer
Questions with Known Answers	Is the transfer from Associate Degree to Bachelors automatic?	No, it is not. You are required to apply when you are closer to the completion of the Associate Degree. (<i>Highly Relevant</i>)	No, it is not. You are required to apply when you are closer to the completion of the Associate Degree.
Questions with Inferred Answers	What does the final year of Computer Science (CS) program include?	(1) [...] Software Engineering (SE) students will do another large in-house project and more SE electives, while CS students will do a slightly smaller project and a few more core [...]. (<i>Partially Relevant</i>) (2) [...] students are required by RMIT rules to do a capstone project in their final year. [...] with an industry partner [...] (<i>Partially Relevant</i>)	It includes a small capstone project with a supervisor that work with an industry partner, as well as a few more core courses and electives.
Out-of-KB Questions	When does the application for program transfer open?	Not available (<i>No Relevant Passages</i>)	I'm sorry, I don't have an answer.

Table 2: Quantitative evaluation of the retrieval phase (NDCG) and generated answers (BERTScore and ROUGE-1), broken down by type of questions. Effectiveness for Out of KB base questions is reported with the percentage of empty rankings / “I don't have an answer” responses. Boldface indicates the best score for each measure and * indicates statistically significant differences against all the other approaches according to Tukey's HSD and significance level $\alpha = 0.01$.

Approach	Retrieval Cutoff k	Known (84 Questions)			Inferred (12 Questions)			Out of KB (10 Questions)		
		NDCG	BERTScore	ROUGE-1	NDCG	BERTScore	ROUGE-1	% Unanswered	BERTScore	ROUGE-1
Intent-Based (IB)		0.643	0.771*	0.671*	0.083	0.332*	0.062	80.00	0.866*	0.813*
RAG (BM25 + Falcon)	1	0.512	0.536	0.179	0.167	0.493	0.185	0.00	0.336	0.045
	3	0.491	0.543	0.209	0.256	0.447	0.106	10.00	0.293	0.054
	5	0.473	0.543	0.209	0.329	0.447	0.106	10.00	0.293	0.054
RAG (DPR + Falcon)	1	0.691	0.545	0.193	0.250	0.513	0.192	10.00	0.340	0.048
	3	0.612	0.564	0.244	0.235	0.476	0.123	20.00	0.311	0.045
	5	0.581	0.564	0.244	0.235	0.476	0.123	20.00	0.311	0.045

the Inferred questions. It is worth noting that the LLM tend to benefit from having less context (i.e., fewer passages in the prompt), achieving higher BERTScore and ROUGE-1 scores for lower cutoffs for both RAG approaches. Results also corroborate that IB performs significantly better than RAG approaches for Out of KB questions.

4 IMPACT AND FUTURE WORK

We built Walert, a conversational agent that answers FAQs about programs of study offered in the School of Computing Technologies at RMIT University. The IB approach, deployed on an Amazon Echo device, was showcased as a demo at the university's Open Day in August 2023 where potential future students learned about the use of LLM-based conversational systems and its risks and limitations. The demo, which was also showcased to visiting high-school students in September 2023, generated university-wide interest and connections, including the IT service team building a university-wide solution.

There are several limitations that we are aiming to address in future work. The current demo relies upon a limited knowledge base (i.e., a manually curated FAQ). We aim to reproduce our methodology with a more extensive set of documents, including brochures and internal web pages related to the delivery of CS programs.

Further research on evaluation measures (beyond BERTScore and ROUGE) is needed to evaluate the validity of generated responses. We aim to explore other evaluation measures, including those for truncated rankings [3] and other dimensions of LLM-based conversational systems [18]. Finally, we plan to deploy RAG approaches to perform online experimentation.

The process of building Walert helped us not only to share complementary knowledge across our group but also to facilitate knowledge translation within the university. We believe the approach and the lessons learned can help other researchers aiming to bridge the gap between experts and practitioners interested in building (and testing) LLM-based conversational information-seeking systems.

ACKNOWLEDGMENTS

Walert was designed and developed in the unceded lands of the Wurundjeri and Boon Wurrung peoples of the eastern Kulin Nation. We pay our respects to their Ancestors and Elders, past, present, and emerging. This research is partially supported by the Australian Research Council (DE200100064, CE200100005) and is undertaken with the assistance of computing resources from RACE (RMIT AWS Cloud Supercomputing). We thank Amina Hossain and Santha Sumanasekara for their valuable contributions.

REFERENCES

- [1] Eugene Agichtein, Yoelle Maarek, and Oleg Rokhlenko. 2022. Alexa Prize TaskBot Challenge. In *Alexa Prize TaskBot Challenge 1 Proceedings*. <https://www.amazon.science/alexa-prize/proceedings/alexa-prize-taskbot-challenge>
- [2] Ebtesam Almazrouei, Hamza Alobeidli, Abdulaziz Alshamsi, Alessandro Cappelli, Ruxandra Cojocaru, Merouane Debbah, Etienne Goffinet, Daniel Heslow, Julien Launay, Quentin Malartic, Badreddine Noune, Baptiste Pannier, and Guilherme Penedo. 2023. Falcon-40B: An Open Large Language Model with State-of-the-art Performance. *Findings of the Association for Computational Linguistics: ACL 2023* (2023), 10755–10773.
- [3] Enrique Amigó, Stefano Mizzaro, and Damiano Spina. 2022. Ranking Interruptus: When Truncated Rankings Are Better and How to Measure That. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Madrid, Spain) (SIGIR '22). Association for Computing Machinery, New York, NY, USA, 588–598. <https://doi.org/10.1145/3477495.3532051>
- [4] Vicki Couzens, Jeph Neale, Hilary Jackman, Grace Leone, and Jessica Clark. 2019. Wurrunggi Biik: Law Of The Land. https://issuu.com/rmitculture/docs/wurrunggi_biik_law_of_the_land Accessed: 15 Dec 2023.
- [5] Aniket Deroy, Kripabandhu Ghosh, and Saptarshi Ghosh. 2023. How Ready are Pre-trained Abstractive Models and LLMs for Legal Case Judgement Summarization?. In *Proceedings of the 3rd Workshop on Artificial Intelligence and Intelligent Assistance for Legal Professionals in the Digital Workplace (LegalAIIA 2023)*. <https://ceur-ws.org/Vol-3423/paper2.pdf>
- [6] Zihan Gao and Jiepu Jiang. 2021. Evaluating Human-AI Hybrid Conversational Systems with Chatbot Message Suggestions. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management (Virtual Event, Queensland, Australia) (CIKM '21)*. Association for Computing Machinery, New York, NY, USA, 534–544. <https://doi.org/10.1145/3459637.3482340>
- [7] Christian Göbel. 2013. The Information Dilemma: How ICT Strengthen or Weaken Authoritarian Rule. *Statsvetenskaplig tidskrift* 115, 2013 (2013), 367–384.
- [8] Kalervo Järvelin and Jaana Kekäläinen. 2002. Cumulated Gain-Based Evaluation of IR Techniques. *ACM Trans. Inf. Syst.* 20, 4 (oct 2002), 422–446. <https://doi.org/10.1145/582415.582418>
- [9] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of Hallucination in Natural Language Generation. *ACM Comput. Surv.* 55, 12, Article 248 (mar 2023), 38 pages. <https://doi.org/10.1145/3571730>
- [10] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781. <https://doi.org/10.18653/v1/2020.emnlp-main.550>
- [11] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems (Vancouver, BC, Canada) (NIPS'20)*. Curran Associates Inc., Red Hook, NY, USA, Article 793, 16 pages.
- [12] Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*. Association for Computational Linguistics, Barcelona, Spain, 74–81. <https://aclanthology.org/W04-1013>
- [13] Jimmy Lin, Xueguang Ma, Sheng-Chieh Lin, Jheng-Hong Yang, Ronak Pradeep, and Rodrigo Nogueira. 2021. Pyserini: A Python Toolkit for Reproducible Information Retrieval Research with Sparse and Dense Representations. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval* (Virtual Event, Canada) (SIGIR '21). Association for Computing Machinery, New York, NY, USA, 2356–2362. <https://doi.org/10.1145/3404835.3463238>
- [14] Yoelle Maarek. 2022. Alexa, Let's Work Together! How Alexa Helps Customers Complete Tasks with Verbal and Visual Guidance in the Alexa Prize TaskBot Challenge. In *Proceedings of the 30th ACM International Conference on Multimedia (Lisboa, Portugal) (MM '22)*. Association for Computing Machinery, New York, NY, USA, 1–2. <https://doi.org/10.1145/3503161.3549912>
- [15] City of Port Phillip. 2016. Boonwurrung Walert (Possum Skin) Cloak. <https://www.portphillip.vic.gov.au/explore-the-city/first-peoples/first-peoples-arts/boonwurrung-walert-possum-skin-cloak> Accessed: 15 Dec 2023.
- [16] Ashwin Ram, Rohit Prasad, Chandra Khatri, Anushree Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrew. 2017. Conversational AI: The science behind the Alexa Prize. (2017). <https://www.amazon.science/publications/conversational-ai-the-science-behind-the-alexa-prize>
- [17] Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at TREC-3. *Nist Special Publication Sp* 109 (1995), 109.
- [18] Tetsuya Sakai. 2023. SWAN: A Generic Framework for Auditing Textual Conversational Systems. *arXiv preprint arXiv:2305.08290* (2023).
- [19] Winson Ye and Qun Li. 2020. Chatbot security and privacy in the age of personal assistants. In *2020 IEEE/ACM Symposium on Edge Computing (SEC)*. IEEE, 388–393. <https://doi.org/10.1109/SEC50012.2020.00057>
- [20] Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. BERTScore: Evaluating Text Generation with BERT. In *Proceedings of the International Conference on Learning Representations (ICLR'20)*. <https://openreview.net/forum?id=SkeHuCVFDr>