Report on the 1st Workshop on Generative Information Retrieval (Gen-IR 2023) at SIGIR 2023

Bénédict, G.; Zhang, R.; Metzler, D.; Yates, A.; Deffayet, R.; Hager, P.; Jullien, S.

DOI
10.1145/3642979.3642995

Publication date
2023

Document Version
Final published version

Published in
SIGIR Forum

Report on the 1st Workshop on Generative Information Retrieval (Gen-IR 2023) at SIGIR 2023

Gabriel Bénédict
IRLab & RTL NL
University of Amsterdam
The Netherlands
g.benedict@uva.nl

Ruqing Zhang
ICT
Chinese Academy of Sciences
China
zhangruqing@ict.ac.cn

Donald Metzler
Google Research
USA
metzler@google.com

Andrew Yates
IRLab
University of Amsterdam
The Netherlands
a.yates@uva.nl

Romain Deffayet
IRLab & Naver Labs Europe
University of Amsterdam & Naver
The Netherlands & France
r.e.deffayet@uva.nl

Philipp Hager
Mercury Machine Learning Lab
University of Amsterdam
The Netherlands
p.k.hager@uva.nl

Sami Jullien
AI for Retail Lab
University of Amsterdam
The Netherlands
s.jullien@uva.nl

Abstract

The first edition of the workshop on Generative Information Retrieval (Gen-IR 2023) took place in July 2023 in a hybrid fashion, co-located with the ACM SIGIR Conference 2023 in Taipei (SIGIR 2023). The aim was to bring information retrieval researchers together around the topic of generative AI that gathered attention in 2022 and 2023 with large language models and diffusion models. Given the novelty of the topic, the workshop was focused around multi-sided discussions, namely panels and poster sessions of the accepted proceedings papers. Two main research outcomes are the proceedings of the workshop[1] and the potential research directions discussed in this report.

Date: 27 July 2023.

Website: https://coda.io/@sigir/gen-ir.

[1] https://coda.io/@sigir/gen-ir/accepted-papers-17

1 Introduction

Generative information retrieval (Gen-IR) is an emerging field derived from autoregressive language models, where tokens are predicted one at a time, given an input and previously predicted tokens. If the tokens to be predicted are words, then Gen-IR reduces to grounded answer generation. The predicted tokens can also be document identifiers (e.g., titles, document IDs, etc.). In that case, workshop participants agreed on the term generative document retrieval (oftentimes referred to as DSI[2]). The predicted tokens can also be nodes and edges of a graph representation. In that case, we could talk of knowledge graph generation. More rarely, predicted tokens can be an item to recommend to a user (e.g. a movie to a streaming platform user). We refer to this as generative recommendation. We formally list these research directions here:

- Grounded Answer Generation
- Generative Document Retrieval
- Generative Recommendation
- Summarization and Document Rewriting
- Generative Knowledge Graphs

This is our best attempt at an exhaustive list of what Gen-IR means. The terminology above is, however, ever-evolving. We maintain a list of generative retrieval papers[3] in hopes of encouraging further work in this direction. Gen-IR seems promising because it can be considered an end-to-end paradigm. A single autoregressive system predicts the next token (word or document ID) to perform all retrieval tasks: retrieval, reranking, query reformulation, natural language answer, conversation, etc.

Evidence for the interest Gen-IR can be found in our list of papers above, and in the four related events this year:

(i) Neurosymbolic Generative Models[4] [NeSy-GeMs workshop @ ICLR 2023]
(ii) Retrieval-based Language Models and Applications[5] [Retrieval-LM Tutorial @ ACL 2023]
(iii) Recommendation with Generative Models[6] [RGM Workshop @ CIKM 2023]
(iv) Personalized Generative AI[7] [PGAI Workshop @ CIKM 2023]

This paper is an event report of our own Gen-IR event: the 1st Generative Information Retrieval Workshop (Gen-IR 2023), held in conjunction with SIGIR 2023. The workshop had a poster session with accepted papers and several panel discussions. We report on how we organized the workshop (Section 2), provide a descriptive account on what happened at the workshop (Section 3), and report on what we learned (Section 4). This final section is based on transcripts of the panel discussions. For each, we extract a number of hot takes: interesting points raised by panelists or the audience. We then discuss these hot takes, distill them into concrete research questions, and provide additional thoughts on each.

[2] Tay et al. [2022] pioneered that technique with their differentiable search index (DSI) model and inspired follow up papers.
[3] https://github.com/gabriben/awesome-generative-information-retrieval
[4] https://nesygems.github.io/
[5] https://acl2023-retrieval-lm.github.io/
[6] https://rgm-cikm23.github.io/
[7] https://sites.google.com/view/pgai2023/home

2 Workshop Overview

This section is purely descriptive and gives an account of who reviewed papers and how we divided panel sessions into topics. We first explain what we mean with these topics.

2.1 Topics

During the workshop, we subdivided discussions into three topics:

1. *Model Training* refers to elements related to model, such as architecture, training regime and reinforcement learning methods
2. *Model Behavior* refers to how the model reacts to different input data (inputs, outputs, hallucination, answer generation, etc.).
3. *Broader Issues* deals with evaluation, Human Computer Interactions, ecosystem concerns, societal issues, etc.

Each topic can contain any of the research directions above (recommendation, document retrieval, etc.).

2.2 Format

The workshop was a full-day hybrid workshop held in Taiwan's on the 27th of July 2023. The day was organized as follows:

Time	Activity
Morning	
09.00 – 09.30	Opening
09:30 – 10:30	Panel Discussions (Model Training)
10:30 – 11:00	Coffee break
11:00 – 12:30	Poster Session, shared with the ReneuIR[8] and REML[9] workshops
12:30 – 13:30	Lunch
Afternoon	
13:30 – 14:30	Panel Discussions (Broader Issues)
14:30 – 14:45	Coffee Break
14:45 – 15:45	Panel Discussion (Model Behavior)
15:45 – 16:00	Coffee Break
16:00 – 16:45	Roundtable Discussion
16:45 – 17:00	Closing

[8] https://reneuir.org/
[9] https://reml-workshop.github.io/

2.3 Program Committees

Gen-IR owes its existence to the dedication of eighteen researchers who generously volunteered their time to review the submissions. We extend our heartfelt gratitude to each member for their commitment to the workshop. A list of the program committee members is displayed below.

- Arian Askari (Leiden Institute of Advanced Computer Science, Leiden University)
- Xiao Wang (University of Glasgow)
- Xinyu Ma (Baidu)
- Andrew Yates (University of Amsterdam)
- Vinh Q. Tran (Google)
- James Thorne (KAIST)
- Jiangui Chen (Institute of Computing Technology, Chinese Academy of Sciences)
- Yubao Tang (Institute of Computing Technology, Chinese Academy of Sciences)
- Ronak Pradeep (David R. Cheriton School of Computer Science, University of Waterloo)
- Shengyao Zhuang (The University of Queensland)
- Hainan Zhang (Beihang University)
- Qingyao Ai (Tsinghua University)
- Zhicheng Dou (Renmin University of China)
- Yujia Zhou (Renmin University of China)
- Nicola De Cao (University of Amsterdam)
- Roi Cohen (Tel Aviv University)
- Sheng-Chieh Lin (University of Waterloo)
- Hyunii Lee (KAIST)

3 Workshop Program

In this segment, we present an overview of the Gen-IR workshop, encompassing details about its participants, accepted papers, poster sessions, panels, and roundtable discussions.

3.1 In Numbers

Being the first workshop on generative information retrieval, Gen-IR 2023 has garnered significant interest within the information retrieval community. We are delighted to share that the workshop had a total of 231 registrants, comprising half of attendees who are joining virtually and half of enthusiastic individuals participating in person. This turnout underscores the growing curiosity and engagement surrounding the evolving landscape of generative information retrieval.

3.2 Papers

We have 15 accepted papers. Each submission was reviewed on EasyChair[10] in a double blind fashion by at least three reviewers from the list in Section 2.3. Reviewers could rate the paper with reject / weak-reject / weak-accept / accept. We did not allow for a borderline (i.e. neutral) stance.

[10] https://easychair.org/

Papers that had mixed reviews were decided upon by the organizers; acting as meta reviewers. There was no preset quota on the number of papers to accept / reject. Code and reproducibility efforts were encouraged in the call for paper. All accepted papers are hosted in a non-archival fashion on our website[11] and were presented in a poster session. In the following, we categorize them by topic and present their titles and authors along with a summary of their contributions.

3.2.1 Model Training

Of these 15 accepted submissions, four papers focus on the training strategies of generative models for retrieval problems.

How Does Generative Retrieval Scale to Millions of Passages?
by *Ronak Pradeep, Kai Hui, Jai Gupta, Adam Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler and Vinh Tran* [Pradeep et al., 2023]

Although many different approaches have been proposed to improve the effectiveness of generative retrieval, they have only been evaluated on document corpora on the order of 100k in size. Therefore, in this work, the authors conduct the first empirical study of generative retrieval techniques across various corpus scales, ultimately scaling up to the entire MS MARCO passage ranking task with a corpus of 8.8M passages and evaluating model sizes up to 11B parameters. Experimental results demonstrate that the use of synthetic queries as a document representation strategy is the only approach that remained effective, and highlight the importance of accounting for the compute cost of techniques and keeping the parameter count fixed. The findings will help the research community better understand the current challenges faced when applying generative retrieval models to larger corpora and inspire new research in this direction.

Generative Retrieval as Dense Retrieval
by *Thong Nguyen and Andrew Yates* [Nguyen and Yates, 2023]

Similar to Pradeep et al. [2023], this work also argues that the new generative retrieval paradigm faces challenges with updating the index and scaling to large collections. Specifically, the authors analyze two prominent variants of generative retrieval and demonstrated that the generative retrieval process can be decomposed into dot products between query and document vectors, similar to dense retrieval. This analysis leads the authors to propose a hybrid generative-dense retrieval model with a new atomic document representation/identifier, Tied-Atomic, and a contrastive loss with BM25 negatives sampled from top-k ranked documents for training. Experimental results on two datasets, NQ320k [Kwiatkowski et al., 2019] and the full MS MARCO show that this approach does not reduce retrieval effectiveness while enabling the model to scale to large collections.

Bridging the Gap Between Indexing and Retrieval for Differentiable Search Index with Query Generation

[11] https://coda.io/@sigir/gen-ir/accepted-papers-17

by *Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon and Daxin Jiang* [Zhuang et al., 2023]

The authors identify and tackle an important issue of current DSI models: the data distribution mismatch that occurs between the DSI indexing and retrieval processes during training. Therefore, a simple yet effective indexing framework for DSI, called DSI-QG, is proposed to address this fundamental problem. When indexing, DSI-QG represents documents with a number of potentially relevant queries generated by a query generation model and re-ranked and filtered by a cross-encoder ranker. The presence of these queries at indexing allows the DSI models to connect a document identifier to a set of queries. Experimental results on both mono-lingual and cross-lingual passage retrieval tasks show the effectiveness of DSI-QG.

On Exploring the Reasoning Capability of Large Language Models with Knowledge Graphs
by *Pei-Chi Lo, Yi-Hang Tsai, Ee-Peng Lim and San-Yih Hwang* [Lo et al., 2023]

This paper examines the capacity of Large Language Models (LLMs) to reason with knowledge graphs using their internal knowledge graph, i.e., the knowledge graph they learned during pre-training. Two research questions are formulated: (1) To what extent can LLMs accurately recall information from KG? and (2) To what extent can LLMs infer knowledge graph relations from context? To address these two questions, the authors employ LLMs to perform four distinct knowledge graph reasoning tasks and distinguish content and ontology hallucination that occur in knowledge reasoning tasks. Experimental results demonstrate that LLMs can retrieve knowledge graph information from memory and infer knowledge graph relations from given context.

3.2.2 Model Behaviors

Of these 15 accepted submissions, six papers focus on the model behaviors of generative models for different input data and tasks.

Query Expansion by Prompting Large Language Models
by *Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang and Michael Bendersky* [Jagerman et al., 2023]

Query expansion is a widely used technique to improve the recall of search systems. The authors propose an approach to query expansion that leverages the generative abilities of Large Language Models (LLMs). In contrast to traditional PRF-based query expansion, LLMs are not restricted to the initial retrieved set of documents and may be able to generate expansion terms not covered by traditional methods. The proposed method is simple: the authors prompt a large language model and provide it a query, then use the model's output to expand the original query with new terms that help during document retrieval. Experimental results show that Chain-of-Thought prompts are especially promising for query expansion, since they instruct the model to generate verbose explanations that can cover a wide variety of new keywords.

Generative Query Reformulation for Effective Adhoc Search
by *Xiao Wang, Sean MacAvaney, Craig Macdonald and Iadh Ounis* [Wang et al., 2023]

Similar to query expansion, automatic query reformulation is also a popular paradigm used in information retrieval (IR) for improving effectiveness. In light of recent advancements in generative language models, this paper targets to study the capacity of such models to perform query reformulation and how they compare with long-standing query reformulation methods that use pseudo-relevance feedback. In particular, the authors investigate two representative query reformulation frameworks, GenQR and GenPRF. GenQR directly reformulates the user's input query, while GenPRF provides additional context for the query by making use of pseudo-relevance feedback information. Experimental results demonstrate the effectiveness of the generated query reformulations in comparison to several existing baselines.

Tackling Query-Focused Summarization as A Knowledge-Intensive Task: A Pilot Study
by *Weijia Zhang, Svitlana Vakulenko, Thilina Rajapakse, Yumo Xu and Evangelos Kanoulas* [Zhang et al., 2023]

Query-focused summarization (QFS) task aims to generate a summary from a set of topic-related documents to answer a given query. However, such relevant documents should be annotated manually and thus are not readily available in realistic scenarios. Therefore, the authors tackle the QFS task as a knowledge-intensive (KI) task, where relevant documents should be retrieved from a large-scale knowledge corpus given a query. To this end, the authors first build a new dataset based on the existing DUC datasets, and then benchmark the dataset with two families of models: QFS models and retrieval-enhanced models. Experimental results demonstrate that QFS models perform significantly worse on the KI setting compared to the original QFS task and the KI setting is much more challenging and offers substantial room for improvement.

QontSum: On Contrasting Salient Content for Query-focused Summarization
by *Sajad Sotudeh and Nazli Goharian* [Sotudeh and Goharian, 2023]

Similar to Zhang et al. [2023], this work also focuses on the query-focused summarization (QFS) task. In this study, the authors propose Qontsum, a novel approach for QFS that leverages contrastive learning to help the model attend to the most relevant regions of the input document. The approach is evaluates on a couple of benchmark datasets for QFS and experimental results demonstrate that it surpasses or achieves the performance of existing approaches on relevant benchmark datasets with considerably reduced computational cost through enhancements in the fine-tuning stage. Moreover, the authors conduct automatic performance comparisons, human assessment, and error analyses, to gain insights into the model's strengths, limitations, and potential future research directions.

PALR: Personalization Aware LLMs for Recommendation
by *Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang and Yanbin Lu* [Yang et al., 2023]

Personalized recommendation has emerged as a crucial factor in meeting users' expectations for customized experiences. The authors argue the challenges of directly leveraging parametric knowledge saved in general-purpose LLMs to generate recommended items, e.g., knowledge gaps between LLMs and items that need to be recommended, the possibility of LLMs to generating hallucinatory results and the limitations regarding input token length and efficiency. Therefore, they propose a novel framework, named PALR (Personalization Aware LLMs for Recommendation), aimed at integrating user history behaviors (such as clicks, purchases, ratings, etc.) with LLMs to generate user preferred items. Evaluation on two public datasets demonstrates the strong potential of an LLM for recommendation in comparison to the state-of-the-art approaches.

Generative Sequential Recommendation with GPTRec
by *Aleksandr V. Petrov and Craig Macdonald* [Petrov and Macdonald, 2023]

Sequential recommendation is an important recommendation task that aims to predict the next item in a sequence. The authors argue that existing Transformer-based recommendation models have several limitations, such as the vocabulary of item ids may be many times larger than in language models, and the classical Top-K recommendation approach may not be optimal for complex recommendation objectives. Therefore, inspired by recent progress in generative language models, the authors propose the GPTRec sequential recommendation model based on the GPT-2 architecture, to address large vocabulary issues by splitting item ids into sub-id tokens using a novel SVD Tokenisation algorithm. Besides, a novel Next-K recommendation strategy is proposed to produce complex interdependent recommendation lists.

3.2.3 Broader Issues

Of these 15 accepted submissions, five papers focus on the broader issues, e.g., evaluation and human cognition and interactions, of generative retrieval models.

On the Robustness of Generative Retrieval Models: An Out-of-Distribution Perspective
by *Yuan Liu, Ruqing Zhang, Jiafeng Guo, Wei Chen and Xueqi Cheng* [Liu et al., 2023]

So far, much effort has been devoted to developing effective generative retrieval models. There has been less attention paid to the robustness perspective. When a new retrieval paradigm enters into the real-world application, it is also critical to measure the out-of-distribution (OOD) generalization. This paper first defines OOD robustness from three perspectives in retrieval problems: 1) The query variations; 2) The unforeseen query types; and 3) The unforeseen tasks. Based on this taxonomy, the authors design corresponding experiments and conduct empirical studies to analyze the robustness of several representative generative retrieval models against dense retrieval models. The results reveal generative retrieval models' overall poor performance in OOD robustness, and they have different generalizability performance in different OOD scenarios.

Query Understanding in the Age of Large Language Models
by *Avishek Anand, Venktesh V, Abhijit Anand and Vinay Setty* [Askari et al., 2023]

Querying, conversing, and controlling search and information-seeking interfaces using natural language are fast becoming ubiquitous with the rise and adoption of large-language models (LLM). In this position paper, the authors describe a generic framework for interactive query-rewriting using LLMs, which reformulates ambiguous queries into intepretable, faithful and scrutable natural language queries. The authors point that their vision of improving interactive query understanding using LLMs has far-reaching implications for – how users interact with search engines, how data is collected from user interactions, the quality of data and feedback, and its impact on the learning ecosystem. The authors also envision that using LLM-based rewrites would also bootstrap many subareas of IR like question answering, entity search, temporal IR, and medical IR.

Generating Synthetic Documents for Cross-Encoder Re-Rankers: A Comparative Study of ChatGPT and Human Experts
by *Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas and Suzan Verberne* [Askari et al., 2023]

This paper introduces a new dataset, ChatGPT-RetrievalQA, which contains 24,322 queries, 26,882 responses generated by ChatGPT, and 58,546 human-generated responses. Specifically, the novel ChatGPT-RetrievalQA dataset is presented to address two research questions: (1) How does the effectiveness of cross-encoder re-rankers fine-tuned on ChatGPT-generated responses compare to those fine-tuned on human-generated responses in both supervised and zero- shot settings? and (2) How does the effectiveness of using ChatGPT for generating relevant documents differ between specific and general domains? Then,the authors fine-tune a range of cross-encoder re-rankers on either human-generated or ChatGPT-generated data, evaluating their performance on the proposed dataset.

GPT-4 Synthetic Data Improves Generalizability For Contract Clause Retrieval
by *Shang Gao, Divyanshu Murli, Javed Qadrud-Din and Martin Gajek* [Gao et al., 2023]

In this work, the authors compare the performance of dual encoder ranker models and cross encoder reranker models trained entirely on synthetic data from GPT-4 against the same models trained on human expert annotated data. The authors test performance on clause retrieval using the CUAD and Applica AI Contract Discovery datasets. The results suggest that synthetic data can replace human data in situations where inference may require generalization or involve data drift—namely, models trained on synthetically generated data from GPT-4 can better retrieve a wide range of potentially unseen contract clause types compared to models trained on a limited set of human annotated contract clause types. The authors expect that, while synthetic data may still be lower quality than human expert labeled data, the higher quantity and flexibility of synthetic data expose the resulting models to more diversity.

Retrievability Bias Estimation Using Synthetically Generated Queries
by *Amin Abolghasemi, Suzan Verberne, Arian Askari and Leif Azzopardi* [Abolghasemi et al., 2023]

This paper aims to evaluate the retrievability bias in pre-trained language models (PLM) based ranking models using both human-generated queries and synthetically generated queries. Two research questions are formulated: (1) What is the retrievability bias of BM25 [Robertson et al., 1996] (as one of the least biased traditional lexical ranking models) in comparison to that of PLM-based rankers? and (2) Can we use synthetic query generation to create query sets for the estimation of retrievability bias? Experimental results show the promise of using synthetic queries generated with transformer-based generative models in estimating retrievability bias. Besides, synthetically generated queries might cause less bias and suggest that training with these queries can be considered as a future research direction for reducing retrievability bias in ranking models.

3.3 Poster Session

In light of the acceptance of our 15 papers, a dynamic poster session was organized to facilitate the dissemination of their findings. 9 poster presenters took the stage to personally unveil their research endeavors. Meanwhile, using Jitsi[12] calls, 6 poster presenters attended virtually. A QR code next to the poster could be scanned by a researcher wondering close to the poster. Attendees were encouraged to bring headphones before the event to talk in isolation to the online poster presenters.

We integrated a joint poster session with two other SIGIR workshops, namely Reaching Efficiency in Neural Information Retrieval (ReNeuIR @ SIGIR 2023) and Retrieval Enhanced Machine Learning (REML @ SIGIR 2023).

3.4 Panels and Roundtable Discussions

A significant part of the workshop was devoted to discussions in the form of panels and a roundtable. The panels covered different aspects of generative answer retrieval and generative document retrieval with the goal of starting discussions on their societal implications and on how these approaches can be improved in the future.

The topic of the first panel was *model training*, with most of the discussion focusing on how generative document retrieval models perform and could possibly be improved. The topic of the second panel was *broader issues*. The discussion revolved around the long-term impact of generative answers, such as long-term effects on the information ecosystem (e.g., will people still write Wikipedia articles?) and when and if it can be reasonable to accept answers from a model that may be incorrect. The topic of the third panel was *model behavior*, with the panelists discussing a variety of topics like the right unit of retrieval and what is missing from current LLMs (e.g., fact-checking, high-level reasoning). At the roundtable, the remaining participants had a brief informal discussion about how learning to rank relates to generative IR.

[12] https://jitsi.org/

4 Workshop Outcomes

In this section, we outline learnings from the workshop, namely from the panel discussions and the workshop organization process.

4.1 Potential Research Directions

We identify potential research directions from our three panel discussions on *model training*, *model behavior*, and *broader issues* (see definitions in Section 2.1). We performed live human transcription of the panel discussions. From the transcript, we identified thoughts that deserved a highlight; we name them hot takes here. We then discussed these singled-out hot takes during five sessions of a neural information retrieval reading group at the IRLab[13] at the university of Amsterdam. For each panel, we list hot takes and additional thoughts that were either gathered in the reading group or during the panel discussion itself. All notes below are based on oral discussions.

4.1.1 Model Training

Panelists. Jiafeng Guo, Vinh Q. Tran, Minjoon Seo and Rajhans Samdani

Discussion summary. The panelists debated on the potential of Gen-IR and the challenges it poses, particularly with the use of LLMs. They highlighted the current limitations of LLMs in the IR setting: slow adoption, sometimes limited performance, evaluation when generating via the autoregressive process, etc. They also highlighted the limitations of current models such as dense passage retrievers (DPRs). They are not much-more parameter efficient, if not worse. They concluded that better models might be found by combining insights from traditional IR and LLMs. For instance, mixture of LLMs, different encoders, beam search or embeddings specialized for retrieval. Finally, they added that good, simple open-source resources would help in speeding up the adoption of DSI models.

Hot takes.

1. **DPRs are actually larger than most LLMs** With a model size of ∼20B parameters, DPR is larger than models like Llama [Touvron et al., 2023] and its successors (typically 7 / 13B parameters). This means that those 7B models might help in retrieving more efficiently than models like DPRs. Yet, they are harder to train and use for inference. It might make sense to have multiple specialized LLMs / DPRs, rather than a single big one to encode all documents.

   Additional thoughts

   - A mixture of experts model, where each model is a Llama-like model fine-tuned on a particular corpus (e.g. labour law, chemistry).

[13] https://irlab.science.uva.nl/

- What is the efficiency of different model architectures in terms of how much information they contain per billion parameters? Broadly speaking, we can list:
  - DPR index for whole MS MARCO [Nguyen et al., 2016]: 25M x 768 = 20B parameters.
  - Recent autoregressive LLMs: 7 / 13 B or more parameters [Touvron et al., 2023].
  - The original DSI T5 [Tay et al., 2022] that can only handle 300k docs: 9B parameters.

- ColBert [Khattab and Zaharia, 2020] and T5 are the norm in retrieval, but can we use recent > 10B parameters autoregressive LLMs (e.g. Llama)?
- Can Llama be the architecture for any of the retrieval stages: document encoding like DSI, for the dual encoder in DPR (e.g. instead of sentence-T5 [Ni et al., 2022]), for reranking [Nogueira et al., 2020], for learned sparse retrieval (LSR) (e.g. instead of sentence-T5)?

2. **We want to memorise more with models like DSI.** That might mean that we want more queries per document for a whole corpus.

   Additional thoughts

   - Can we release more complete resource papers that DSI can make use of?
   - Do more queries per document help for memorization, for document understanding or both?
   - Query generation will not help avoiding catastrophic forgetting and overfitting.
   - Should we use document embeddings as model weights?

3. **Atomic IDs are very arbitrary.** While it might work on non-changing corpus, it might not work in dynamic settings, with inference on new documents. Hence why it is important for the output to be more semantic and more generalized. Once we add more documents in those models, we see catastrophic forgetting hence the need for adapters [Houlsby et al., 2019].

   Additional thoughts

   - Investigate the use of smaller LLMs as adapters [Houlsby et al., 2019].
   - Ensure that documents used to update the LLM come from different enough distributions.
   - We could perform hierarchical clustering on the metadata via the topics, domain name and ontology. However, this requires curated metadata that is distinct enough from the ID.
   - Could we use several IDs per document, for instance one per generated question? This might help in encompassing several aspects of the document.

4. **Can we get the first word right?** It is very similar to classification problems. Of course, beam search can help but it's no panacea.

Additional thoughts

- Can we use uncertainty quantification methods used in classification in order to perform "safe" first-word generation? For instance, conformal prediction [Shafer and Vovk, 2008] might help in assessing how good a first word is.
- Does the inherent greediness of beam search result in sub-optimal generations for the first word [Kruszewski et al., 2023]?

5. **LLMs are not competitive on MS Marco.** We don't see the size increase performance impact that we expected. Should we use dual encoders to improve performance?

   Additional thoughts

   - Dual encoders that are trained in a dual Q-learning fashion [Hasselt, 2010].
   - What is the best architecture combination of encoders? LLM as encoder, encoder with teacher forcing? Or something else?
   - Can we easily distil larger models without a loss of generalization, if the size increase is not striking?
   - Can we use FAISS [Johnson et al., 2019]-related methods that can perform filtering at inference time (e.g. 'who is Michael Jordan?'; search only in academic literature)?

6. **People in industry are not making use of those DSI models.** A reason is that there are clear limitations, but also because there is no tool to easily use it.

   Additional thoughts

   - Can we ensure generalizability to avoid the cost of re-training?
   - Adding those models to user-friendly libraries like SimpleTransformers[14] might speed-up adoption, as professionals need to be able to experiment and prototype quickly.
   - If DSI is good for certain tasks, the usage in industry will trivially follow.

4.1.2 Model Behavior

Panelists. Chua Tat-Seng, Omar Khattab, Nazneen Fatema Rajani and Fabio Petroni

Discussion summary. The session discussed the current trends in Gen-IR, focusing on generative document retrieval (GDR), e.g., differentiable search index (DSI) models, and grounded answer generation (GAG). The panelists see a potential for GDR and GAG that has yet be unleashed, along multiple dimensions: mixing natural language and document ids in answers, perform multiple steps of reasoning prior to delivering an answer, representing documents as snippets, and combining LLMs and retrieval systems' strengths. However, they agreed that some ingredients are currently missing in order to enable accurate and trustworthy GDR systems: stronger reasoning and decoding capabilities, better fact-checking and attribution, as well as better metrics, tasks and tools that would enable faster iterations. The panelists also discussed how to make such models

[14] https://github.com/ThilinaRajapakse/simpletransformers

robust to adversarial attacks, whether reinforcement learning from human feedback (RLHF) is necessary and sufficient, how to quantify model confidence, and the link with cross-encoders.

Hot takes.

1. **DSI might help in, tip-of-the-tongue tasks, for example**; by making larger reasoning leaps from the query to the relevant documents (as opposed to just comparing embeddings in DPR).

   Additional thoughts:

   - Reasoning may be enabled via subquery reformulation with an LLM.

2. **Answer generation and document retrieval should be combined in a single model.**

   Additional thoughts:

   - Training data for such models could be provided by NQ-like datasets which contain both documents and answers.

3. **Combining classic IR methods with LLMs** seems like a promising way forward (e.g., SEAL [Bevilacqua et al., 2022] uses an FM index).

   Additional thoughts:

   - Query expansion could be another such example.

4. **Generating atomic identifiers for whole documents might be sub-optimal**, we should consider retrieving smaller units of texts, like snippets.

   Additional thoughts:

   - Models like SEAL [Bevilacqua et al., 2022] attempt to do that.
   - Techniques can be borrowed from passage retrieval and entity linking, e.g., GENRE [Cao et al., 2020].
   - Multiview identifiers (n-grams, urls, titles) can be used, but identifier length should be carefully studied (longer IDs cover more content of the document but are harder to learn)

5. **Gen-IR models lack the ability to actively clarify questions.**

   Additional thoughts:

   - Can we look more closely at retrieval-enhanced multi-turn conversation?
   - Could clarifying questions be used as training data for DSI?
   - Also, when should the model ask for clarifying questions?
   - When should it decide to not just show the most probable answer?

6. **LLMs should judge the trustworthiness of the content they are consuming.**

7. **We don't know yet how to force Gen-IR models actively fact-check their answers** (e.g., using knowledge graphs) and provide how confident they are.

   Additional thoughts:

   - Knowledge graphs are very restrictive in the amount of relationships they encode, but retrieval augmented models may be a solution to this.
   - Models can also be trained to be calibrated on NQ datasets.

8. **On model confidence:** evaluating the consistency of responses and asking counterfactual questions might be more attainable than getting model confidence scores.

   Additional thoughts:

   - To perform consistency checks, we could ask questions in different ways, for instance by modifying some parts of the question and observing whether the answer is still the same.
   - Asking counterfactual questions and observing whether the model retains its initial opinion after being contradicted give insight into the confidence of the model in its assessment.
   - It also raises the related question of the confidence of the model when truth is ambiguous or there are multiple sides to a story (see Broader issues – 4) and the amount of novel evidence required for it to change its mind once it has been trained on false or outdated data.

9. **LLMs are currently quite agreeable to the point that users can convince them to change an answer to complete nonsense**. This tendency of LLMs might be due to training models to be "helpful" using RLHF. Maybe RLHF can be used to teach models to be less agreeable and to stand their ground on answers the model is confident about.

   Additional thoughts:

   - Consistency (see previous point) is oftentimes not even achieved within a single conversation. Therefore, consistency checks could help flag instruction-finetuned models that easily apologize and changes their mind, upon being contradicted.

10. **RLHF cannot solve problems of factuality**, a systematic model issue requires a systematic fix.

    Additional thoughts:

    - Regarding a systematic fix for factuality, we are missing a structure that stores beliefs. Maybe this could be a knowledge graphs? Or different token for facts / beliefs? Maybe the model could know about logic and reasoning?
    - Does the model understand the concept of amount of evidence towards beliefs?

- Can the model differentiate the information retrieved in documents from current instruction? Can *situational awareness* [Berglund et al., 2023] help with that?

11. **LLMs being attribution-focused in their answer might avoid certain prompting attacks.** If a person's Twitter biography asks LLMs to say that they are attractive, instead of answering: "The user is very attractive [1]"[15], the model could reply with: "According to the user's Twitter biography [1], they are very attractive."

    Additional thoughts:

    - A Gen-IR model must be robust to such adversarial attacks in the documents: what it reads should not change its way of reasoning.
    - A naive prompt improvement could be to ask the model to ignore instructions in the text.
    - It seems important to build benchmarks on adversarially corrupted data.
    - Another idea would be substitution, i.e., assessing what share of the document can be substituted with adversarial text and still retrieve the right answer.

12. **Gen-IR models need high-level reasoning capabilities**, such as performing simple calculations on top of retrieved documents.

    Additional thoughts:

    - Stronger reasoning capabilities may require training and architectural changes (e.g., specific tokens, curriculum learning, ...)

13. **We lack public benchmarks that require complex reasoning on top of retrieval.**

14. **The current speed of LLM improvements might make current evaluation metrics quickly obsolete.** We need new metrics, and we need to make sure the metrics and tasks still make sense when we release a new model.

4.1.3 Broader Issues

Panelists. Chirag Shah, Emily Bender, Yiqun Liu and Guido Zuccon

Discussion summary. The panel discussed the broader societal implications of the widespread adoption of LLMs in web search. The panel debated the long-term consequences of users relying on compelling answers instead of consulting sources, making them vulnerable to misinformation and confirmation bias. Opinions varied among panelists and audience members regarding the feasibility of addressing current LLM flaws. While others were optimistic that LLMs could enter our daily lives as practical black box systems and even enter high-stakes domains such as medicine, others were more skeptical, citing concerns about opaqueness and unpredictability of their mistakes. The panel also delved into the long-term effects of generative search on the information

[15] [1] represents the attribution as displayed to the user

ecosystem. Concerns included: (i) Sidelining information producers, altering incentive structures and business models, (ii) inundating the web with generated content, complicating content provenance determination and training new models, and (iii) the potential for new adversarial attacks on LLMs. In exchanges with the audience, the panel debated the responsibilities of releasing and advertising new models and whether there is a general longing for AI-generated content for humans in the first place.

Hot takes.

1. **A holistic view of the information ecosystem:** As generative search seeks to directly answer questions instead of directing search traffic to external sources, we must consider the impact on the business models and incentive structures for information producers.

   Additional thoughts:

   - What is the impact of generative AI on existing business models on the web?
   - Should model responses include paid links as advertisement?
   - How do incentives change for unpaid providers of information, e.g., editors on Wikipedia, travel blogs, or people sharing recipes?
   - Can generated responses encourage users to contribute to the information ecosystem?
   - What is the impact of generative AI on copyright?

2. **A useful black box?** Can LLMs serve as useful black box systems in our daily lives, similar to how we use navigational apps?

   Additional thoughts:

   - Isn't there a fundamental difference between an incorrect GPS route because of an untracked roadblock and the unpredictable ways current LLMs fail?
   - LLMs might arrive at the right answer for the wrong reasons.
   - Is there a difference between considering trustworthiness from the ground up when designing models instead of nudging model outputs using RLHF?
   - Should the black box nature of current LLMs prohibit certain high-stakes applications?
   - Restricting the use of LLMs in search to summarizing retrieved documents might make evaluation and verification more tractable.

3. **Attribution is not a silver bullet:** How many users consult linked source material, especially when the generated response feels compelling enough?

   Additional thoughts:

   - Confirmation bias in web search is already a societal problem.
   - Checking if a source contains the generated information is not trivial.
   - Attribution shifts the responsibility for fact-checking from search engines to users.
   - The attributed link might be hallucinated. And aren't there more effective ways to include source material?

4. **All sides of the story?** Search engines are confronted with many questions without objective answers, e.g., because a topic might be subjective or facts are historically contended. In the debate about truthful information, how do we handle ambiguous cases?

   Additional thoughts:

   - Presenting all sides of a story might lead to users picking the answer they want to hear. How do you deal with conspiracy theories or heavily contested societal issues?
   - Who decides which opinions are covered and which are not?
   - Even if a system could provide true answers (like a calculator), we might be asking the wrong question.
   - This problem exists already in today's web search and we might leverage existing solutions, including policies or compliance teams.

5. **Mechanisms for trust:** We often cannot personally verify information. Society has developed mechanisms for trusting experts, e.g., in the form of university degrees. What authority do LLMs have to summarize and present information, particularly in specialized domains?

   Additional thoughts:

   - Grounding LLMs might be enough for language models to be useful.
   - How would we measure trust?

6. **The curse of recursion:** The widespread adoption of generative AI might flood the web with generated content.

   Additional thoughts:

   - How is the training of LLMs impacted when a significant portion on the web is generated content?
   - How can we differentiate human versus AI generated content? Should watermarking content be mandated for commercial LLMs?
   - As we can never rely on watermarking to be used, e.g. for open models, are there other ways to determine content provenance?

7. **The audience for AI art:** Humans currently prefer to watch/read/listen to human-generated content in many creative domains, such as music or poetry. Are people even interested in generative content in the long term?

   Additional thoughts:

   - While computers have dominated humans in chess for over twenty years, more people watch chess grandmasters compete than chess bots.
   - Does this trend change for children growing up with generative AI? Will they still seek out human-generated content?

8. **New adversarial attacks:** As people have always tried to game web search, we need more research on new attack vectors enabled by generative search.

9. **The role of user education:** While thoroughly evaluating models before release is crucial, what is the role of educating the general public on AI safety?

   Additional thoughts:

   - Society has had to learn the limits of other technological innovations in the past, are LLMs different?
   - Are the limits of current LLMs appropriately reflected in their advertisement?

4.2 Lessons Learned

The following lessons were learned as a result of organizing this workshop. We share these in hopes they will be beneficial to others who organize similar workshops in the future.

- Attendees were engaged during all three panel sessions, with questions coming from both the moderators and from the audience. Preparing a diverse bank of potential questions beforehand helped the moderators to get the discussion started and to followup on topics the attendees or panelists found interesting.
- The joint poster session held with two related workshops went well. We like that this gave people the opportunity to engage with posters from all three workshops, rather than asking them to choose to attend a single poster session. We would recommend this poster session format whenever possible.
- The poster session was hybrid: presenters who could not make it to Taipei were assigned to a Jitsi call (processed described in Section 3.3). This setup was not very successful: only 1-2 people were caught talking with the online presenter.
- While the workshop had a hybrid format, there were significantly more attendees in person. Most of the participation in the panels came from in-person attendees. This was not a problem given the high in-person attendance, but the panels would have been much less engaging without people present to ask questions and join the discussion.
- The day ended with an informal, open-ended roundtable discussion among the remaining attendees. Attendees were less engaged than during the panels, perhaps due to the time of day and relatively low attendance at this point. To improve the usefulness of a roundtable, we would recommend giving discussion groups a concrete assignment on what to discuss and holding the roundtable earlier in the day.
- Code and reproducibility efforts were encouraged in the call for paper, but were not often shared in submitted papers. Maybe more stringent guidelines or checklists (e.g. at NeurIPS[16]) are necessary.
- The reviewing platform EasyChair is not double-blind by default. Actually, it requires a few minutes of play-testing to figure out that three different settings need to be changed. This might nudge workshop and conference organizers into practicing single-blind reviewing.

[16] https://neurips.cc/Conferences/2022/PaperInformation/PaperChecklist

5 Conclusion

The Gen-IR 2023 workshop was crafted to serve as a platform for fostering collaboration between academia and industry researchers with diverse backgrounds, all of whom share a keen interest in the concept, development, and application of generative information retrieval. This commitment to inclusiveness is evident in our workshop program, which features three panels comprising 12 researchers, a poster session showcasing 15 accepted papers, and a roundtable discussion.

The immense potential of (large) language models in information retrieval, with subsequent applications in downstream services, is widely acknowledged. However, the exact definition of a generative information retrieval system remains fluid; e.g., should it be closed-book or open-book? model-centered or user-centered? The resolution to this question is likely contingent on the specific application context. We look forward to sharing efforts that help push the agenda on Gen-IR further.

Acknowledgments

We would like to thank ACM and SIGIR for hosting this workshop, extend our appreciation to the SIGIR2023 workshop chairs, Liana Ermakova and Maura R. Grossman, as well as our exceptional panelists, program committee members, paper authors, and participants. The IRLab[17] NeuralIR reading group attendees provided most of the content for the additional thoughts to the panel hot takes.

Contribution Statement

Ruqing Zhang, Donald Metzler and Gabriel Bénédict organized the Gen-IR workshop. Andrew Yates joined as an organizer a few weeks before the workshop. Philipp Hagger, Romain Defayet and Sami Jullien transcribed the panels in person at the workshop. Gabriel Bénédict singled out hot takes from the transcription to be discussed in the IRLab NeuralIR reading group. Gabriel Bénédict handed the notes back to Philipp Hagger, Romain Defayet and Sami Jullien; so they could write down a section per panel session in this paper. Ruqing Zhang wrote the paper summaries and the conclusion. Gabriel Bénédict wrote the abstract, sections 1 and 2. Gabriel Bénédict, Donald Metzler and Andrew Yates proof read and corrected the draft.

[17] https://irlab.science.uva.nl/

References

Amin Abolghasemi, Suzan Verberne, Arian Askari, and Leif Azzopardi. Retrievability bias estimation using synthetically generated queries. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Arian Askari, Mohammad Aliannejadi, Evangelos Kanoulas, and Suzan Verberne. Generating synthetic documents for cross-encoder re-rankers: A comparative study of chatgpt and human experts. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Lukas Berglund, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo, and Owain Evans. Taken out of context: On measuring situational awareness in llms, 2023.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. Autoregressive search engines: Generating substrings as document identifiers. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 31668–31683. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/cd88d62a2063fdaf7ce6f9068fb15dcd-Paper-Conference.pdf.

Nicola De Cao, Gautier Izacard, Sebastian Riedel, and Fabio Petroni. Autoregressive entity retrieval. *CoRR*, abs/2010.00904, 2020. URL https://arxiv.org/abs/2010.00904.

Shang Gao, Divyanshu Murli, Javed Qadrud-Din, and Martin Gajek. Gpt-4 synthetic data improves generalizability for contract clause retrieval. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Hado Hasselt. Double q-learning. *Advances in neural information processing systems*, 23, 2010.

Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. Parameter-efficient transfer learning for NLP. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 2790–2799. PMLR, 09–15 Jun 2019.

Rolf Jagerman, Honglei Zhuang, Zhen Qin, Xuanhui Wang, and Michael Bendersky. Query expansion by prompting large language models. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Jeff Johnson, Matthijs Douze, and Hervé Jégou. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data*, 7(3):535–547, 2019.

Omar Khattab and Matei Zaharia. Colbert: Efficient and effective passage search via contextualized late interaction over bert. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '20, page 39–48, New York, NY, USA, 2020. Association for Computing Machinery. ISBN 9781450380164.

Germán Kruszewski, Jos Rozen, and Marc Dymetman. disco: a toolkit for distributional control of generative models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 144–160, Toronto, Canada, July 2023. Association for Computational Linguistics.

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

Yu-An Liu, Ruqing Zhang, Jiafeng Guo, Wei Chen, and Xueqi Cheng. On the robustness of generative retrieval models: An out-of-distribution perspective. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Pei-Chi Lo, Yi-Hang Tsai, Ee-Peng Lim, and San-Yih Hwang. On exploring the reasoning capability of large language models with knowledge graphs. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Thong Nguyen and Andrew Yates. Generative retrieval as dense retrieval. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. Ms marco: A human generated machine reading comprehension dataset. In *CoCo@ NIPs*, 2016.

Jianmo Ni, Gustavo Hernandez Abrego, Noah Constant, Ji Ma, Keith Hall, Daniel Cer, and Yinfei Yang. Sentence-t5: Scalable sentence encoders from pre-trained text-to-text models. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1864–1874. Association for Computational Linguistics, May 2022.

Rodrigo Nogueira, Zhiying Jiang, Ronak Pradeep, and Jimmy Lin. Document ranking with a pretrained sequence-to-sequence model. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 708–718, Online, November 2020. Association for Computational Linguistics.

Aleksandr V Petrov and Craig Macdonald. Generative sequential recommendation with gptrec. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Ronak Pradeep, Kai Hui, Jai Gupta, Adam D Lelkes, Honglei Zhuang, Jimmy Lin, Donald Metzler, and Vinh Q Tran. How does generative retrieval scale to millions of passages? *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Stephen Robertson, S. Walker, M. M. Hancock-Beaulieu, M. Gatford, and A. Payne. Okapi at trec-4. In *The Fourth Text REtrieval Conference (TREC-4)*, pages 73–96. Gaithersburg, MD: NIST, January 1996. URL https://www.microsoft.com/en-us/research/publication/okapi-at-trec-4/.

Glenn Shafer and Vladimir Vovk. A tutorial on conformal prediction. *Journal of Machine Learning Research*, 9(3), 2008.

Sajad Sotudeh and Nazli Goharian. Qontsum: On contrasting salient content for query-focused summarization. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. Transformer memory as a differentiable search index. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=Vu-B0clPfq.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. Llama: Open and efficient foundation language models, 2023.

Xiao Wang, Sean MacAvaney, Craig Macdonald, and Iadh Ounis. Generative query reformulation for effective adhoc search. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Fan Yang, Zheng Chen, Ziyan Jiang, Eunah Cho, Xiaojiang Huang, and Yanbin Lu. Palr: Personalization aware llms for recommendation. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Weijia Zhang, Svitlana Vakulenko, Thilina Rajapakse, Yumo Xu, and Evangelos Kanoulas. Tackling query-focused summarization as a knowledge-intensive task: A pilot study. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.

Shengyao Zhuang, Houxing Ren, Linjun Shou, Jian Pei, Ming Gong, Guido Zuccon, and Daxin Jiang. Bridging the gap between indexing and retrieval for differentiable search index with query generation. *Gen-IR@SIGIR 2023: The First Workshop on Generative Information Retrieval*, 2023.