Contextualizing and Expanding Conversational Queries without Supervision

ANTONIOS MINAS KRASAKIS, ANDREW YATES, and EVANGELOS KANOULAS,
University of Amsterdam, The Netherlands

Most conversational passage retrieval systems try to resolve conversational dependencies by using an intermediate query resolution step. To do so, they synthesize conversational data or assume the availability of large-scale question rewriting datasets. To relax those conditions, we propose a zero-shot unified resolution–retrieval approach, that (i) contextualizes and (ii) expands query embeddings using the conversation history and without fine-tuning on conversational data. Contextualization biases the last user question embeddings towards the conversation. Query expansion is used in two ways: (i) abstractive expansion generates embeddings based on the current question and previous history, whereas (ii) extractive expansion tries to identify history term embeddings based on attention weights from the retriever. Our experiments demonstrate the effectiveness of both contextualization and unified expansion in improving conversational retrieval. Contextualization does so mostly by resolving anaphoras to the conversation and bringing their embeddings closer to the important resolution terms that were omitted. By adding embeddings to the query, expansion targets phenomena of ellipsis more explicitly, with our analysis verifying its effectiveness on identifying and adding important resolutions to the query. By combining contextualization and expansion, we find that our zero-shot unified resolution–retrieval methods are competitive and can even outperform supervised methods.

CCS Concepts: • **Information systems** → **Information retrieval**; **Users and interactive retrieval**; **Retrieval models and ranking**; *Information retrieval query processing;*

Additional Key Words and Phrases: Information retrieval, conversational search, dense retrieval, query expansion

ACM Reference format:
Antonios Minas Krasakis, Andrew Yates, and Evangelos Kanoulas. 2023. Contextualizing and Expanding Conversational Queries without Supervision. *ACM Trans. Inf. Syst.* 42, 3, Article 77 (December 2023), 30 pages.
https://doi.org/10.1145/3632622

This article is an extended version of Krasakis et al. [41].

This research was supported by the NWO Innovational Research Incentives Scheme Vidi (016.Vidi.189.039), the NWO Smart Culture - Big Data / Digital Humanities (314-99-301), the H2020-EU.3.4. - SOCIETAL CHALLENGES - Smart, Green And Integrated Transport (814961), and the Hybrid Intelligence Center, a 10-year program funded by the Dutch Ministry of Education, Culture and Science through the Netherlands Organisation for Scientific Research. All content represents the opinion of the authors, which is not necessarily shared or endorsed by their respective employers and/or sponsors.

Authors' address: A. M. Krasakis, A. Yates, and E. Kanoulas, University of Amsterdam, The Netherlands; e-mails: a.m.krasakis@uva.nl, a.c.yates@uva.nl, e.kanoulas@uva.nl.

1046-8188/2023/12-ART77 $15.00
https://doi.org/10.1145/3632622

1 INTRODUCTION

The introduction of commercial voice assistants along with advances in natural language understanding have enabled users to interact with retrieval systems in richer and more natural ways through conversations. While those interactions can ultimately lead to increased user satisfaction, they are inherently complex as they require an understanding of the entire dialogue semantics by the retrieval system [49, 59]. For instance consider the sequence of queries of Table 1. Initially, the user requests information regarding the Bronze Age collapse, but his follow-up queries omit vital information on the basis of it being mentioned before. That could happen either explicitly by **anaphora** (*id 34_2*: "evidence for *it*") or implicitly by **ellipsis** (*id 34_3*: "possible causes [of bronze age collapse]").

Large-scale datasets have greatly improved the ability of models to understand language across a variety of tasks [11, 53, 73], but such datasets are lacking when it comes to conversational retrieval. Constructing a large and diverse conversational retrieval dataset can be quite challenging. One issue is that each query can depend on the previous conversation. Therefore as conversations evolve queries in later turns occur in the long-tail [26, 75]. Hence, multi-turn queries are hard to aggregate and anonymise, making it unlikely that publicly available resources can be built from real user search sessions or conversational interactions. Therefore, datasets need to be built using human experts in controlled environments or simulations. However, the former leads to small-scale datasets [15–17, 61] and requires explicit conversation development instructions which bias the nature of the constructed dataset and hurt the generalizability of models to new types of conversations [2]. The latter cannot guarantee fidelity to real scenarios and is a field of study on its own [28, 30, 33, 38, 63].

On the other hand there is a plethora of data resources for ad-hoc retrieval, e.g., Craswell et al. [12]. Therefore, most conversational retrieval approaches so far introduce a query rewriting step, which essentially decomposes the conversational search problem into a query resolution problem and an ad-hoc retrieval problem. Query resolution attempts to place the user's question in the context of the conversation. One set of methods does so by generating [23, 44, 67, 74], updating [56] or expanding [42, 70] the words of the last user question/query, while others try to disambiguate the query on the latent space [14, 45, 75]. In all cases, supervision is required and performed against CANARD [21], a dataset of $40K$ curated rewrites of conversational questions, QReCC [5], a dataset that complements CANARD with additional queries and answers, or synthetically created conversations for training models [14, 49, 74] that cannot guarantee fidelity towards real scenarios. Unsupervised approaches expand the user's question by extracting general informative terms from the conversation history [10, 46].

Recent advances in instruction-tuned generative LLMs have given rise to models with notable generalization capabilities [29, 39, 54, 65]. Although little work exists so far in using such LLMs for Conversational Retrieval, preliminary evidence shows promising results for query rewriting in few-shot settings [48]. Other work tries to enhance conversational retrieval by using generative LLMs to generate potential answers and fuse these answers in the query to improve passage ranking [52]. Nonetheless, performing the entire conversational retrieval task with a generative model remains underexplored, while even performing retrieval comes with many challenges. Those challenges are related to hallucinations and issues around answer verifiability, which are crucial topics in Information Retrieval [47]. Some works try to mitigate those issue, by generating an answer and trying to attribute it to a reference document from the corpus post-hoc [9], but further research is required in this area.

Instead, in this article we focus on Dense Retrievers and inspect their abilities to handle both query resolution and retrieval in a zero-shot setting. We pose the following key research question:

Table 1. Example of Conversational Queries

id	Query
34_1	Tell me about the Bronze Age collapse.
34_2	What is the evidence for it?
34_3	What are some of the possible causes?
34_4	Who were the Sea Peoples?
34_5	What was their role in it?
34_6	What other factors led to a breakdown of trade?
34_7	What about environmental factors?
34_8	What empires survived?
34_9	What came after it?

to what extent can Dense Retrieval models transfer knowledge from ad-hoc retrieval to conversational retrieval, where data scarcity is and will likely remain an imminent problem?

To answer this question we adapt ColBERT [37], the state-of-the-art BERT-based token-level dense retriever pre-trained on ad-hoc search data. We propose ***Ze**ro-shot **Co**nversational **Co**ntextualization* (*ZeCo*^{2}), a variant of ColBERT which contextualizes all embeddings within the conversation but performs matching using only the contextualized terms of the user's last question (Figure 1). Since conversational queries can omit crucial information and contextualization operates implicitly by influencing embeddings, we hypothsize that more explicit methods can improve retrieval. Therefore we also explore conversational query expansion, which explicitly adds missing terms to the query to resolve conversational dependencies. We introduce two Query Expansion techniques that are unified in the Dense Retriever's Query Encoder, which differ from other methods that use external models or algorithms to perform expansion [46, 67, 70, 74]. The former method, **abstractive query expansion**, operates entirely in the latent space and predicts term embeddings relevant to the user's information need using *MASK* tokens. The latter, **extractive query expansion**, detects important terms from the conversation history based on attention weights from the query encoder and adds them to the query. We expect that attention weights can be a meaningful signal for identifying resolution terms, especially when phenomena of coreference appear. For example, when the anaphora "What is the evidence of *it*" occurs (referring to the Bronze Age collapse), we expect high attention scores between *it* and the *Bronze Age collapse*.

Observing that contextualization and expansion work in different ways, we argue they can complement each other and combine them. For example, consider query id 34_2 ("What is the evidence for it"). Contextualization will ideally influence the embedding of the anaphora term *it* to go closer to the resolution terms "Bronze Age Collapse", whereas expansion will explicitly add these term embeddings to the query. Therefore, in the last part of this work, we compare the effectiveness of various expansion methods when used in combination with contextualization, but also in isolation. We benchmark the unified abstractive and extractive expansion methods introduced here, as well as a number of baseline extractive expansion methods. From the latter we consider only extractive expansion methods that add terms from the conversation history, and specifically HQE [46], an unsupervised keyword-extraction method and QuReTeC [70], a state of the art BERT-based query resolution method.

To sum up, this article introduces zero-shot conversational retrieval–expansion methods unified into the Dense Retriever's Query Encoder. Our approaches are zero-shot in the conversational domain, that is, they do not use any conversational search data, neither rewritten queries nor relevance judgements, to retrieve relevant passages. In contrast to the aforementioned unsupervised keyword extraction works, that use multi-stage retrieval pipelines and hence depend on

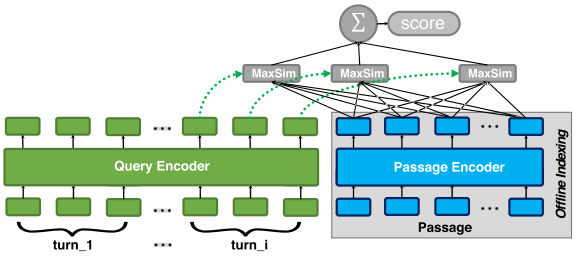


Fig. 1. Zero-shot Conversational Dense Retriever (figure adapted from [37] with permission)

the retrieval corpus to solve the query resolution problem, we use a single Dense Retriever to perform retrieval and conversational query resolution (contextualization and query expansion).

Specifically, we aim to answer the following research questions:

RQ1 Can **zero-shot contextualisation** of conversational queries improve dense passage retrieval?
- **RQ1.1** How does zero-shot contextualization affect embeddings of the last turn's query?
- **RQ1.2** How robust is zero-shot contextualization across turns?

RQ2 Can we contextualise **abstractive query expansion** with the conversation to improve dense passage retrieval?
- **RQ2.1** How does zero-shot contextualiszation affect embeddings of abstractive expansion tokens?
- **RQ2.2** How robust is abstractive query expansion across turns?

RQ3 Can we perform **extractive query expansion** from the conversation history using the Dense Retriever's encoder to improve dense passage retrieval?

RQ4 How do zero-shot unified resolution–retrieval methods compare to baseline query expansion methods and can contextualisation of independently created expansions improve performance?

The remainder of this article is structured as follows: In Section 2, we outline the Related Work, while Sections 3 and 4 contain our Methodology and Experimental Setup respectively. In Section 5, we investigate **RQ1** Conversational Contextualization, along with the corresponding analysis that looks at its effect on token embeddings and robustness across turns (Sections 5.1 and 5.2). In Section 6 we investigate **RQ2** that complements Contextualization with Abstractive Query Expansion, followed by a similar analysis on Expansion embeddings (Section 6.1) and robustness (Section 6.2). We proceed by investigating **RQ3**, Extractive Query Expansion in Section 7. Finally, in Section 8, we examine **RQ4** by examining the performance of various types of expansion mechanisms in isolation or combined with contextualization. We conclude in Section 9 with key takeaways and directions for future work.

2 RELATED WORK

Conversational agents have become increasingly ubiquitous with the introduction of commercial voice-assistants (e.g., Alexa, Google Assistant, etc.). This is fueled by advances in Machine Learning, voice recognition as well as Language Understanding. Many works have tried to understand the area by defining tasks, challenges and opportunities around Conversational Information Seeking systems [27, 76]. Radlinski and Craswell [59] provided a theoretical framework for conversational search, analyzing in depth desirable characteristics and capabilities a conversational search agent should possess. Azzopardi et al. [6] discussed **conversational information-seeking (CIS)**

systems in the scope of different actions and intents for both its users and systems. A large number of researchers came together to discuss CIS systems, define them and set a future roadmap, coming up with many diverse research areas and agendas [4].

While different systems, tasks and conceptual frameworks have been proposed, the task of Open-domain Conversational Question Answering [5, 57] was one of the earliest, most concrete and explicitly related to information retrieval. It involves a user trying to satisfy an information need while searching through a vast collection of documents to find supporting evidence while answering the question in hand. Such systems usually comprise of a retriever, that looks for a document containing the answer, and a reader which extracts the answer given the retrieved document. In this article, we focus on the task of **Conversational Passage Retrieval (ConvPR)**, which involves evaluating conversational systems based on the relevance of their retrieved results.

2.1 Query Rewriting

Query rewriting has been a crucial step in most approaches for Conversational Passage Retrieval and its goal is to make the last user question/query self-contained and therefore independent of the previous conversation. Various methods approach ConvPR as a query rewriting problem, followed by an ad-hoc retrieval pipeline.

A first set of rewriting methods can be characterised as "extractive", in the sense that they expand the last user question with terms from the conversation history. They do so either using a BERT-based token classification approach [42, 70] or using keyword extraction techniques [10, 46]. Among the first class of methods, Voskarides et al. [70] proposed a distant supervision technique [51] to automatically identify relevant terms from the conversation history using the relevant passages. Kumar and Callan [42] use a similar approach while also combining it with Multi-View Reranking, which constructs multiple alternative queries and fuse their scores at the reranking phase. When it comes to keyword extraction methods, Lin et al. [46] devise a Historical Query Expansion (HQE) pipeline, that first uses Query Performance Prediction to determine whether a conversational query is ambiguous and if so, they expand it using keywords from the conversation history. Both the QPP and the keyword-extraction algorithms here are based on BM25 scores and hence unsupervised. Borisov et al. [10] also use keyword extraction to improve conversational search, in a slightly different setting. They investigate document ranking with mixed-initiative conversations containing clarifying questions [3] and collect two keyword extraction datasets from news titles and conversations.

On the other hand, some methods use encoder-decoder architectures [64] to generate the disambiguated query entirely, instead of appending terms to the last question. Many methods fine-tune GPT-2 [58] or T5 [60] models as rewriters using CANARD [21], a query rewriting dataset consisting of 40K manually curated rewrites [23, 44, 46, 66]. In contrast, Yu et.al. [74] identify that conversational queries mostly exhibit anaphoras and ellipsis and create their own synthetic dataset for training an abstractive question rewriter. To construct the dataset, they corrupt session queries by either creating co-references or omitting terms that appear in previous turns/queries in the session. More recently, Qian and Dou [56] propose a hybrid approach, which modifies the query using insertions and replacements of words from the original query to perform the resolution. They propose a unifying framework between query rewriting and context modelling, using supervision signals from the query rewrite to amplify their influence on the learned query representations in the latent space. In a similar spirit, Mo et al. [52] try to use the predictive power of LLMs in answer generation to boost retrieval. Specifically, they learn to generate a potential answer from the query and use it to expand the query during the retrieval.

In a comparative study, Vakulenko et al. [67] benchmarks rewriters using a common retrieve-then-rerank setup, concluding that different rewriters are optimal for the retrieval and the

reranking phase. Del Tredici et al. [18] performs a similar study for open-domain QA, investigating separately passage retrieval and QA performance. They conclude that extractive rewriters methods (e.g., QuReTeC [70]) work best when used with sparse retrievers, while generative rewriters work best for the reader.

2.2 Dense Retrieval Methods for Conversational Search

A number of works have also investigated using a Dense Retriever directly with conversational queries (i.e., conversation history and last utterance), which offers advantages in terms of latency as well as a simpler pipeline at inference time. To train such a conversational dense retriever, Yu et al. [75] uses a teacher-student framework, where the student learns how to produce a good embedding/representation of the conversational query. To do so, it tries to mimic the embedding produced by the teacher network, given the human reformulated query. Essentially, their method can also be thought of as rewriting directly in the embedding space, which is also used for document/passage retrieval using approximate nearest-neighbor search architectures [72]. Lin et al. [45] also focus on efficiency and propose an end-to-end Dense Retrieval pipeline. Their method still relies on the query rewrite dataset CANARD, but in contrast to the previous mentioned work, they use the rewrites to create pseudo-relevance labels. Hence, this is a weak-supervision setting, based on the assumption that a top-ranked document for the rewritten query is relevant. More recently Dai et al. [14] create a large synthetic dataset of conversational questions and answer passages, using pre-trained **Large Language Models (LLM)** and prompting. They do so by turning Wikipedia articles to information seeking conversations, essentially creating synthetic user questions from section titles. This results in a synthetic, yet large dataset, which enables them to train a bi-encoder retriever and a cross-encoder reranker. Mao et al. [49] also generate synthetic conversations, but start from web search sessions, to ensure that the relevance target is not the same throughout the conversation.

2.3 Learned Sparse Retrieval for Conversational Search

Besides Conversational Dense Retrieval models, there have been contemporaneous works that adapt SPLADE [24], a **Learned Sparse Retrieval (LSR)** model for Conversational Search. Specifically, both Le Hai et al. [43] and Mao et al. [50] leverage a teacher-student setup as the aforementioned Yu et al. [75], where the teacher has access to human rewritten queries, while the student can only observe the conversation, trying to reproduce a faithful representation from the teacher. Le Hai et al. [43] tries to retrofitting the sparse term representations (bag-of-words output of SPLADE) closer to the ones of the teacher model, using an asymetric MSE loss that encourages term expansion from past answers. LeCoRE [50] similarly uses an Adaptive Sparsity Regularization prior to achieve this, but additionally to trying to predict the teacher's tokens, it also adds a Teacher-Proxy Distillation loss bringing the [CLS] tokens closer.

Our work differs from the aforementioned LSR works since it is entirely zero-shot (not requiring human rewritten queries), and focuses on a Dense retriever. It also differs from the rest of the literature in investigating contextualization and expansion of conversational queries, with a single model and without using any conversational data for training. That is in contrast to methods that use question rewriting datasets [70, 75], or create synthetic conversation datasets [14, 45, 74]. The most closely related method to ours is HQE [46], a keyword-based expansion method that relies on BM25 scores to rewrite queries through conversational expansion. In contrast, our method (a) contextualizes conversational embeddings without supervision (Section 5) and (b) also expands queries with identified term embeddings (Sections 6 and 7), solely based on "knowledge"encoded in the weights of a pre-trained ranking transformer.

3 METHODOLOGY

In this section, we describe our zero-shot dense retriever for conversational search. Our approach consists of two main components: an encoder that produces token embeddings of a document or query and a matching component that compares query and document token embeddings to produce a relevance score.

3.1 Task & Notation

Let q_t be the user utterance/query to the system at the tth turn, and r_t the corresponding canonical passage response[1] provided by the dataset annotators. We formulate our passage retrieval task as follows: Given a user utterance/query q_t and the previous context of the conversation at turn t: $ctx_t = (q_0, r_0, \ldots, q_{t-1}, r_{t-1})$, we produce a ranking of top-k passages $R_{q_t} = (p_t^1, p_t^2, \ldots, p_t^k)$ from a collection C that is most likely to satisfy the users' information need.

3.2 Token-Level Dense Retrieval

In this section, we briefly describe ColBERT [37], a dense retrieval model that serves as our query and document encoder f_{Enc}. In contrast to other dense retrievers that construct global query and document representations (e.g., DPR [36] or ANCE [72]), ColBERT generates embeddings of all input tokens. This allows us to perform matching at the token-level. To generate token embeddings, ColBERT passes each token through multiple attention layers in a transformer encoder architecture, which contextualizes each token with respect to its surroundings [19, 68]. We use E_q to denote the embeddings produced for a query q and E_d to denote the embeddings produced for a document d. To compute a query-document score, ColBERT performs a soft-match between the embeddings of a query token w_q and a document token w_d by taking their inner product. Specifically, each query token is matched with the most similar document token and the summation is computed:

$$Score(q, d) := \sum_{w_q \in q} \max_{w_d \in d} E_{w_q} \cdot E_{w_d}^T \tag{1}$$

where E_{w_q} and E_{w_d} are embeddings of query tokens w_q and document tokens w_d, respectively.

3.3 Conversational Contextualization with Token-Level Dense Retrievers

In this section, we describe our **Zero-shot Conversational Contextualization ($ZeCo^2$)** method.

When dealing with conversations, it is crucial for each turn to be contextualized with respect to the conversation. This is the case because conversational queries have continuity and contain ellipsis or anaphoras to previous turns [66, 70, 74]. Therefore, users omit important information on individual turns, such as the overall topic of the conversation (e.g., "the Bronze Age collapse" in Table 1), which harms ranking performance significantly. The key idea here is to bring back those important terms into the query by contextualising the last user question with the conversation history. We expect that, by doing so, we can implicitly bias the embeddings of the last question towards previously mentioned words, entities or concepts that are related to the topic of the conversation, yet might not be explicitly mentioned in the last utterance. For instance question "What are some of the possible causes" (id 34_2 of Table 1) does not include the topic "Bronze Age collapse", which would cause ranking to fail. We achieve this by extending the idea of token contextualization from tokens to multi-turn conversations. Instead of contextualising query tokens only with respect to surrounding tokens from this utterance/turn, we contextualize them using the entire conversation history. In the example discussed above, we expect the token embeddings of the last question to be—to some extent—influenced by the "Bronze Age collapse".

[1]Canonical passage responses answer previous questions, and the user can refer upon them in his next question.

In practice, we use ColBERT [37] as our query and document encoder f_{Enc}. We follow ColBERT's standard approach for encoding documents, i.e., we prepend a special token $[D]$ and generate the document token embeddings E_d: $E_d = f_{Enc}([D] \circ d \circ [SEP])$. However, to encode a query at turn t, we concatenate the conversational context ctx_t with the last query utterance q_t before generating contextualized query token embeddings $E^*_{q^t}$.

$$E^*_{q^t} := f_{Enc}([Q] \circ ctx_t \circ [SEP] \circ q_t) \tag{2}$$

While $E^*_{q^t}$ constitutes token embeddings of the entire conversation (i.e., the input to f_{Enc}), our goal is to perform ranking based on only tokens from the last utterance q_t. To do so, we (1) replace E_{w_q} with $E^*_{w_q}$ in the token-level matching function (Equation (1)) and (2) compute the score as $Score(q_t, d)$, so that only query tokens from the last turn contribute to it.

$$Score(q_t, d) := \sum_{w_q \in q_t} \max_{w_d \in d} E^*_{w_q} \cdot E^T_{w_d} \tag{3}$$

The approach of contextualising q_t with respect to the conversation history ctx_t avoids the need for resolution supervision from conversational tasks. That is, our method does not need manually rewritten questions of conversational queries, or even relevance judgements of conversational queries which are scarce. Instead, it relies on the pre-training of three different tasks: (a) Masked Language Modelling, (b) next sentence prediction tasks (pre-training of BERT [19]), and (c) ad-hoc ranking task (pre-training of ColBERT [37]).

3.4 Abstractive Query Expansion with [MASK] Tokens

At times, contextualization might not be enough to fully provide the required context from the conversation history (e.g., when ellipsis phenomena occur). To mitigate this, we could explicitly add matching embeddings to the scoring function (Equation (3)), and one way of doing so is by using ColBERTs' query expansion technique. ColBERT appends [MASK] tokens to the end of the query and uses their corresponding embeddings for matching document tokens, which has been shown to improve ranking [37]. Through its training objective, the query encoder is trained to generate embeddings from [MASK] tokens that are likely to match the corresponding relevant document tokens. Therefore, the input to the query encoder (Equation (2)) becomes:

$$E^*_{q^t} := f_{Enc}([Q] \circ ctx_t \circ [SEP] \circ q_t \circ [SEP] \circ [MASK] \circ ... \circ [MASK]) \tag{4}$$

where we treat the number of $[MASK]$ tokens as a hyperparameter and study it in Section 6.

In the context of conversational search, we explore whether and how this expansion method can utilise both the conversation history and the most recent user query to generate relevant embeddings that enhance ranking.

3.5 Extractive Query Expansion with Attention Weights

In addition to generating query expansion embeddings based on the conversation history and query (Section 3.4), we can also identify relevant previous terms from the conversation, and add their embeddings to the matching function (Equation (3)). This is similar to prior work that tries to identify salient terms from the conversation to include in the query [42, 46, 70], but has two major differences. First, it is unsupervised and integrated in the Dense Retriever's query encoder. Second, instead of adding terms to a query, we add their already contextualized embeddings.

To identify expansion terms from the conversation history, we use attention weights from the query encoder as an indication of relevance of history terms to the current utterance. Attention is a mechanism that allows models to focus on specific parts of the input when processing another part of the input, enabling them to capture dependencies and relationships between words in a

dynamic and flexible manner [8, 68]. Previous works have used attention as a source of explanation [1, 35, 71], or even as a proxy for document relevance[34]. In our case, we use attention to capture relationships of tokens within a conversation. Our hypothesis is that attention weights can help us identify resolution words from the conversation, especially when phenomena of coreference or anaphora occur. For instance, we expect that an anaphora term (e.g., "it") will have a high attention weight to the corresponding resolution (e.g., "bronze age collapse").

To capture this, we measure cross-attention weights in the query encoder between *last utterance* and *conversation history* tokens. We measure attention scores from a source token S of the last utterance towards a target token T from the history. For a token S of the last user utterance at position i, and a token T of the conversation history at position j, we compute a simple sum-of-squares attention score from token S to T, aggregated across all attention heads h:

$$\text{attn-score}(S \rightarrow T) := \sum_{h=0}^{H} (a_{ij}^{h})^2 \tag{5}$$

We measure attention weights on the 11th (second-to-last) layer, based on related work and preliminary experiments [19, 32]. We calculate this score based on only one source token (S) from the last user utterance (see details in Section 7).

Finally, we add the token embeddings of the terms we identified to expand with in query and Equation (3) becomes:

$$Score(q_t, d) := \sum_{w_q \in (q_t \cup q_{a^+})} \max_{w_d \in d} E_{w_q}^{*} \cdot E_{w_d}^{T} \tag{6}$$

where q_{a^+} are the top-N history tokens that maximize the attention score (Equation (5)). We treat the number of added expansion tokens as a hyperparameter and study it in Section 7.

3.6 Combining Contextualisation with Query Expansion

Contextualisation and methods performing query expansion from the conversation history can be combined, aiming for improved performance. Since (extractive) query expansion terms come from the conversation history, they have already been processed in the query encoder. Therefore, we only need to include their embeddings when matching, which is accomplished by changing the scoring function (Equation (3)):

$$Score(q_t, d) := \sum_{w_q \in (q_t \cup q_{r^+})} \max_{w_d \in d} E_{w_q}^{*} \cdot E_{w_d}^{T} \tag{7}$$

where q_{r^+} consists of identified history terms that are not in the last query ($q_{r^+} = q_r \backslash q_t$). If tokens in q_{r^+} appear more than once in the history, we use the embedding of the last token occurrence.

4 EXPERIMENTAL SETUP

In this section, we outline our experimental setup.

4.1 Datasets and Evaluation

We test our approach on the TREC CAsT '19, '20 and '21 [15–17] datasets. Each dataset consists of about 25 conversations, with an average of 10 turns per conversation. CAsT '20 and '21 include canonical passage responses to previous questions, that the user can refer to or give feedback. The corpus consists of the MSMarco Passages and Documents, Wikipedia, and Washington Post news articles [20, 53, 55]. TREC CAsT '19 and '20 includes relevance judgements at passage level, whereas CAsT '21 at the document level. For CAsT '21, we split the documents into passages and score each document based on its highest scored passage (*MaxP* [13]).

To quantify retrieval performance we use two metrics: NDCG@3 and Recall at two different depths (R@100 & R@1K). The former quantifies effectiveness at the top ranks, which is important for a user, while the latter expresses the ability of a first-stage ranker to retrieve relevant passages that can be later re-ranked with a more effective second-stage ranker.

4.2 Methods & Baselines

ColBERT. ColBERT was trained to contextualize tokens through (a) the self-supervision of BERT's masked language model and next-sentence prediction [19] and (b) the training to optimize ad-hoc ranking [37]. In our experiments, we use the weights of a ColBERT retriever pre-trained on the MSMarco passage ranking dataset [53]. We use ColBERT v1, while our method remains applicable to v2 [62], where the main novelty is optimizations to reduce the index size. To avoid any spill-over effects, we perform matching only on the query tokens; we deactivate matching on the *CLS*, query indicator ([*Q*]), and expansion tokens used in the original work [37].

Baselines. To assess the effect of our contextualization method $ZeCo^2$ (Section 5), we compare with appropriate conversational dense retrieval baselines. The dense retrieval baselines use either ColBERT [37] or ANCE [72] as their basis.

ColBERT-based baselines:

- *last-turn* uses only the last user question and ignores the conversation history; embeddings are therefore not contextualised (i.e., $ctx_t = \emptyset$ in Equation (2)).
- *all-history* uses the entire conversation as a query; in this case, embeddings are contextualized in conversation, and the matching function includes terms across the entire history (i.e., $Score(ctx_t \circ q_t, d)$).
- *human* uses the human rewritten queries directly, and is therefore an oracle indicating the upper-bound retrieval performance of *ColBERT*.

ANCE-based baselines (taken from *ConvDR* [75]):

- *zero-shot* is an unsupervised model, using the entire conversation as a query (similar to *ColBERT all-history*)
- *few-shot* is a model trained with Knowledge-Distillation on a query rewrite dataset.
- *human* uses the human rewritten queries directly, and is therefore an oracle indicating the upper-bound retrieval performance of *ANCE*.

To evaluate the effectiveness of our query expansion methods (Sections 6 and 7), we provide additional zero-shot and supervised query expansion baselines in an overview table in Section 8. For consistency, we choose extractive methods that pick terms from the conversation history to add to the query. Specifically, we consider the following baselines:

- *Historical Query Expansion (HQE)* [46]: A zero-shot expansion method based on keyword extraction. For fair comparison, we use the rewrites from HQE but do the retrieval step using *ColBERT* (in contrast to the original article, that used the BM25 retrieval).
- *QuReTeC* [70]: A state-of-the-art BERT-based classifier that picks expansion terms, trained on the conversational question rewriting dataset CANARD [21].

5 ZERO-SHOT CONVERSATIONAL CONTEXTUALIZATION

To answer **RQ1**, which asks whether the last user utterance (question) can be effectively contextualised with respect to the conversation history, we compare the performance of the non-contextualised utterance (*last-turn*) with our contextualised approach ($ZeCo^2$) in Table 2. It is clear that contextualisation helps in all cases, especially in terms of *Recall* with relative improvements of

Table 2. Effectiveness of Zero-shot Embedding Contextualization on TREC-CAsT Datasets

base-retriever	variant	zero-shot	CAsT'19		CAsT'20		CAsT'21	
			NDCG@3	R@100	NDCG@3	R@100	NDCG@3	R@100
ColBERT	last-turn [a]	✔	0.214	0.157	0.155	0.124	0.140	0.154
	all-history [b]	✔	0.190	0.165	0.150	0.166	**0.237**	0.265
	ZeCo^2 *(ours)*	✔	0.238 [b]	**0.216** [a,b,c]	**0.176** [b]	**0.200** [a,b,c]	0.234 [a]	**0.267** [a]
	human		0.430	0.363	0.443	0.408	0.431	0.403
ConvDR [75]	zero-shot [c]	✔	**0.247**	0.183	0.150	0.150	–	–
	few-shot		0.466	0.362	0.340	0.345	0.361	0.376
	human		0.461	0.389	0.422	0.454	0.548	0.451

Bold font indicates the best zero-shot performing model. Superscripts indicate statistically significant improvements (paired t-test, $p - value < 0.05$) of $ZeCo^2$ over zero-shot models: *last-turn* [a], *all-history* [b] and *ConvDR zero-shot* [c].

37% - 73%. We further observe that our approach significantly outperforms the *all-history* baseline, which uses the entire conversation as the query, in the first two datasets and yields comparable performance on CAsT'21. We hypothesise that the baseline's improved performance on CAsT'21 is due to its document-level annotations, with one document satisfying multiple turns of the conversation. We also observe that *all-history* performs worse than *last-turn* regarding *NDCG*@3 but better regarding *R*@100. Furthermore, $ZeCo^2$ outperforms the zero-shot ConvDR in most cases, especially with respect to *Recall*. Last, while the supervised versions of ConvDR clearly outperfom $ZeCo^2$ in *NDCG*@3, $ZeCo^2$ remains competitive in terms of *Recall*.

5.1 Effect of Contextualisation on Last Turn Embeddings

Next we consider the effect of contextualisation on the user's query, by looking into how this changes the last turn's query embeddings so that we answer **RQ1.1**.

What are the most influenced terms? As a first step, we measure how much term embeddings change when contextualised with conversation history. To do so, we define token embedding change as the cosine distance between a term before and after contextualisation: $\Delta\vec{tok} = 1 - cos(\vec{tok}_{ZeCo^2}, \vec{tok}_{last-turn})$.

We report frequent terms ranked by average embedding change (avg$\Delta\vec{tok}$) on Table 3(a). We observe that terms indicating anaphora ("they", "it", etc.), punctuation symbols and special tokens are the ones most influenced. This is expected, since users often use anaphoras referring to previous conversation rounds. Regarding punctuation and special tokens, one plausible explanation is that a global representation of a turn is aggregated in those tokens after contextualization.

Similarly, we report embedding changes per part-of-speech tags on Table 3(b). We notice that words that tend to have a more steady meaning across different contexts such as verbs (*VERB*), nouns (*NOUN*) and especially proper nouns (*PROPN*) change much less as a result of our contextualization. The same stands for Adjectives (*ADJ*), probably due to the fact that their dependent nouns are expected to be in the same turn, eg. *"What are the origins of popular music? "*. In contrast, punctuation (*PUNCT*), pronouns (*PRON*) and auxiliary (*AUX*) terms change the most after the contextualization. Regarding pronouns, that mostly consist of anaphora terms, this is expected since they refer to terms from the previous conversation. Regarding punctuation and auxiliary terms, eg. *"Why did Ben Franklin want it to be the national symbol?"*, we hypothesize that those words are quite frequent in English language, but are not very discriminative towards relevant documents— especially in the context of bi-encoders – and therefore are more prone to be affected by their surrounding words [25]. One subsequent hypothesis that requires further investigation is that such words could encapsulate more general topic or conversation embeddings in ColBERT.

Table 3. Average Embedding Change after Contextualization (all CAsT Datasets)

token	frequency	avg($\Delta\vec{tok}$)
they	60	0.501
it	196	0.480
[*SEP*]	934	0.464
?	873	0.458
that	52	0.440
.	192	0.424
...	...	...
macro-avg	–	0.185
micro-avg	–	0.323

(a) per (most frequent) token

POS tag	frequency	avg($\Delta\vec{tok}$)
PROPN	**245**	0.138
ADJ	502	0.187
NOUN	**1187**	0.199
VERB	**1000**	0.243
ADV	195	0.263
ADP	561	0.304
AUX	648	0.374
PRON	**1739**	0.391
PUNCT	**1219**	0.44

(b) per POS tag

token type	frequency	avg($\Delta\vec{tok}$)
stopwords	3913	0.356
non-stopwords	4032	0.269

(c) stopwords vs. non-stopwords

Lastly, Table 3(c) shows that conversational contextualization affects stopword embeddings more than non-stopwords. This is in line with previous research, that found that in ad-hoc search settings ColBERT embeddings of low-frequency terms vary less across different contexts [25].

How Do Terms Change when Contextualised? To illustrate how contextualisation changes the term embeddings and how this in turn helps ranking, we perform both qualitative and quantitative analyses on how the embedding of one of the most influenced anaphora terms ("it") changes and how resolution terms are brought into play. Resolution terms are highly discriminative, low-frequency terms, which are often omitted in subsequent conversational turns due to the linguistic phenomena of ellipsis and anaphoras. In contrast, anaphora words that are present in the last user utterance are completely uninformative in the absence of the conversational context (e.g., omitting "bronze age collapse"on query id 32_3 of Table 1). The effect of ignoring such discriminative, low-frequency terms has been shown to have detrimental effect in ranking in search, as well as particularly in Neural IR models [31, 69].

Qualitative Analysis. In Table 4 we focus on a highly influenced anaphora term ("it") and match it to the most similar token embeddings from the conversation history. We observe that in certain cases, zero-shot contextualization resolves anaphoras successfully, bringing anaphora embeddings very close to the referred term ("sociology", "popular music", "throat cancer"). The first row shows one noteworthy example where the matching term is always *cancer*, but contextualization allows it to resolve to the correct embedding of *throat cancer* instead of *lung cancer*. Further, this leads to an increase in the token embedding similarity (from 0.48 to 1) – in other words, the token embedding of "it"‘ becomes identical to "(throat) cancer"‘. Lastly, we see cases where embeddings come closer to punctuation symbols, indicating that those might preserve some sort of global query representation, or a multi-token concept (e.g., "the neverending story film .").

Quantitative Analysis. Additionally, we quantitatively study to what extent $ZeCo^2$ brings token embeddings of anaphoras closer to resolutions and how this affects retrieval performance. To do so,

Table 4. Examples of Best Term Matches of Anaphora Terms in Conversation History (Before & after Query Contextualization)

Utterance	Human resolution	$\Delta\vec{tok}$	Δ Recall	closest match (non-contextualized)		closest match (contextualized)	
				term	similarity	term	similarity
what is the first sign of **it** ?	throat cancer	0.52	+0.31	tell me about lung **cancer**.	0.48	what causes throat **cancer** ?	1
What is the role of positivism in **it**?	sociology	0.44	+0.67	what is taught in **sociology** ?	0.55	what is taught in **sociology** ?	1
what technological developments enabled **it** ?	popular music	0.54	+0.42	... the origins of popular **music** ?	0.46	... the origins of popular **music** ?	1
what is the evidence for **it** ?	bronze age collapse	0.36	0.00	tell me about the bronze **age** collapse.	0.64	tell me about the bronze age **collapse**.	1
why did ben franklin want **it** to be the national symbol?	turkeys	0.22	+0.29	where are turkeys from **?**	0.55	where are turkeys from **?**	0.85
what is **it** about?	neverending story film	0.39	+0.12	the neverending story film **.**	0.58	the neverending story film **.**	0.88

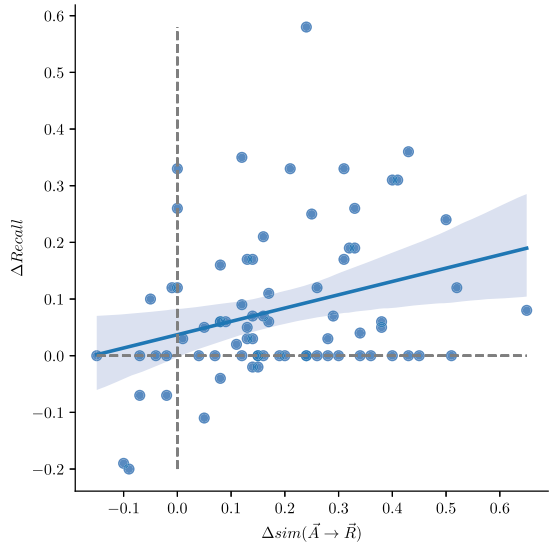


Fig. 2. Correlation between $\Delta Recall$ and similarity change of anaphoras towards resolutions (CAsT '19).

we automatically identify anaphoras and resolutions by comparing terms between human rewrites and raw user utterances. We define the effect of contextualization in bringing anaphora embeddings ($\vec{A}$) closer to resolution embeddings ($\vec{R}$) as:

$$\Delta sim(\vec{A} \rightarrow \vec{R}) = sim(\vec{A}_{ZeCo^2}, \vec{R}) - sim(\vec{A}_{last-turn}, \vec{R})$$

where anaphoras are contextualised within the last turn ($\vec{A}_{last-turn}$) or the entireconversation ($\vec{A}_{ZeCo^2}$). We encode resolutions ($\vec{R}$) independently of the conversation to ensure they retain their original representations. On queries with multi-token anaphoras or resolutions, we pick the highest match.

We present the scatterplot of this change of similarity towards resolutions and $\Delta Recall$ on Figure 2. We observe a correlation (Pearson's $R = 0.31, p-value = 0.005$) between those two terms, meaning that as anaphora embeddings go closer to resolutions, Recall improves. Additionally, we

Table 5. Embedding Similarities between Anaphoras and Resolution or Random Terms from the Conversations (CAsT '19)

	$\vec{R}$ (resolution)	$\vec{Rnd}$ (random term)
$\vec{A}_{last-turn}$	0.178	0.255
$\vec{A}_{ZeCo^2}$	0.372	0.204

observe that for most queries $\Delta Recall$ and Δsim are positive. This means that contextualization almost always improves ranking and brings anaphoras closer to resolutions.

Since bringing anaphoras closer to resolutions could simply be a by-product of being encoded together, we perform a secondary analysis, trying to quantify whether resolution terms are more important than other terms found in the conversation for contextualization. To do so, we measure similarities between anaphoras ($\vec{A}$) and (a) resolution terms ($\vec{R}$) vs (b) random terms from the same conversation ($\vec{Rnd}$) on Table 5. For consistency, we also encode random terms independently.

When anaphoras are contextualized within only the *last-turn*, they are more similar to random terms than to resolutions on average. However, our method ($ZeCo^2$) brings anaphoras closer to their resolutions, while pushing them away from other (random) words from the same conversation. This confirms that resolutions have a high impact on contextualizing anaphoras, in contrast to other random conversation words. The mechanism behind this effect requires further investigation. It could be that simply the lower frequency of resolution terms has an effect here, but it is also possible that pre-trained transformers have certain co-reference resolution capabilities (e.g., by relating "it"to a noun).

5.2 Contextualization Robustness across Turns

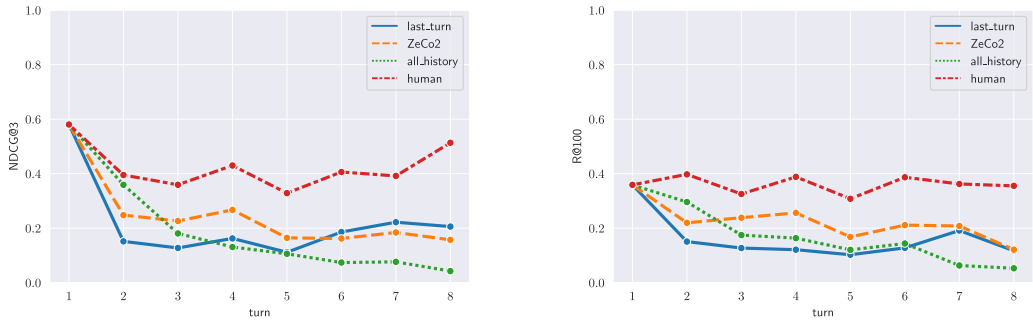
As conversations go deeper queries become more complex: later queries have more dependencies and ambiguity, and the length of the conversation history grows longer [27, 50]. This can have negative effects upon automated conversational search methods, making them fragile towards later turns. In this subsection, we investigate how robust our contextualisation method is, at each turn.

In Figure 3, we measure retrieval performance of the various contextualised and non-contextualised baselines we introduced per conversation turn. We see that performance of *human* queries is high and relatively stable across turns, with performance averaging between 0.4 and 0.6 $NDCG$ points in most cases. In contrast, *last_turn* dramatically degrades from the second turn onward, with $NDCG$ being below 0.2 points in most cases. This indicates the importance of taking previous turns into account, either via rewriting or contextualization.

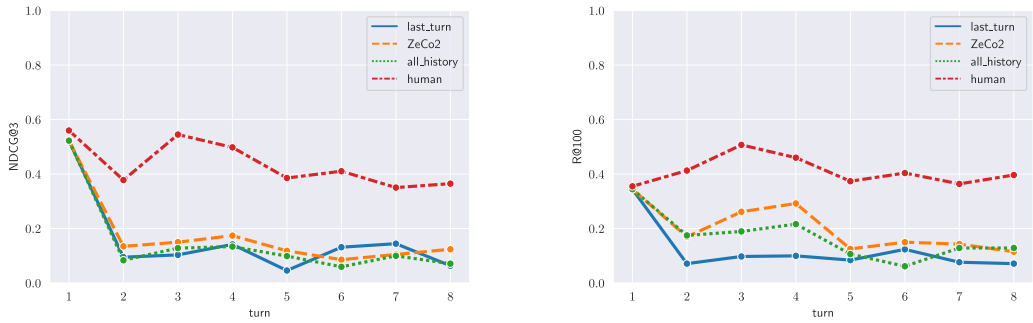
On the other hand, $ZeCo^2$ performs noticeably better compared to the non-contextualized variant (*last_turn*) at all conversation depths, especially in terms of *Recall*. When compared to the *all_history* baseline, it also shows more robustness and better performance with the exception of some early turns in specific datasets (turn 2 and 3 on CAsT'19 and CAsT'21). However, $ZeCo^2$ performance seems to follow a downward trend too as the conversation advances. This difficulty could relate to longer sequences during the query encoding phase, or might as well be a consequence of more complex semantics that appear after multiple conversation turns and (sub-) topic shifts. Nonetheless, $ZeCo^2$ remains relatively robust and competitive, despite some decline in performance as conversations evolve.

Contextualisation Conclusions

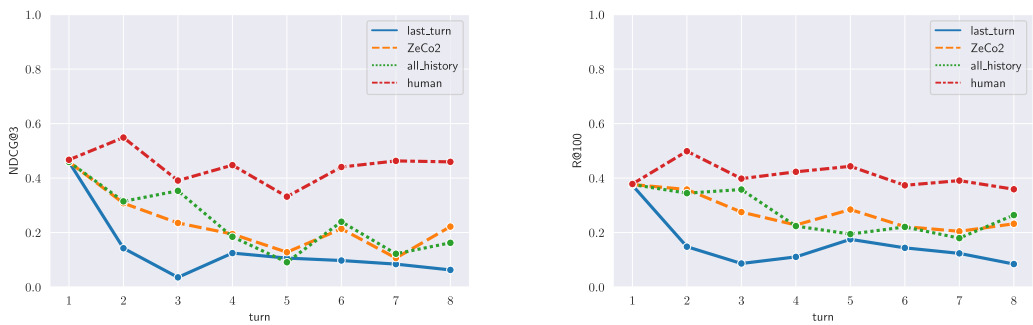
In this part we provide an overview of our conclusions from this section. We observed that contextualisation significantly improved ranking performance, especially in terms of Recall. We examined the mechanism through which this happens, concluding that when contextualising queries



(a) CAsT '19



(b) CAsT '20



(c) CAsT '21

Fig. 3. Performance comparison of different methods at various conversation depths.

wrt the conversation, query embeddings tend to go closer to the missing resolution terms. We show that this effect is stronger between anaphora terms and their corresponding resolutions, driving the increase of ranking performance. As conversations evolve and become longer the effectiveness of contextualisation dampens, but in general, improvements still occur in the last turns of the conversation.

6 ABSTRACTIVE EXPANSION BASED ON MASKED LANGUAGE MODELLING

Following the conversational contextualization, we explore whether query expansion from the conversation history can also be beneficial towards retrieval. Query expansion methods add terms from previous turns to the last user query to make the last user query self-contained. These

Table 6. Abstractive Expansion Results

	CAsT'19		CAsT'20		CAsT'21	
Query variant	NDCG@3	R@100	NDCG@3	R@100	NDCG@3	R@100
last_turn	0.214	0.157	0.155	0.124	0.140	0.154
last_turn + abstr. expansion	0.249	0.224	0.165	0.194	0.188	0.208
$ZeCo^2$	0.238	0.216	**0.176**	0.200	0.234	0.267
$ZeCo^2$ + abstr. expansion	**0.279**	**0.310**	0.168	**0.237**	**0.273**	**0.332**
human (oracle)	0.430	0.363	0.443	0.408	0.431	0.403

methods been one of the main approaches used to tackle conversational search [67, 70]. This line of work mostly tries to identify terms from the history and add them to the query. Instead, we introduce an abstractive expansion method that generates new token embeddings based on the query and conversation history and uses them to match documents. In this section, we aim to answer **RQ2**, that is, whether abstractive query expansion can be contextualized with the conversation to improve dense passage retrieval.

Concretely, our abstractive expansion method takes advantage of the ColBERT [37] embedding-based query expansion method, which adds [*MASK*] tokens to queries, allowing it to generate additional embeddings that will match document tokens and change the ranking. This expansion mechanism is learned through (a) the Masked-Language-Modeling objective of BERT's pre-training, where the encoder tries to predict the following query term, and (b) the ad-hoc ranking fine-tuning task, where the retriever tries to bring closer this embedding to tokens from relevant documents [19, 37]. At inference time, the expansion mechanism tries to generate matching embeddings being conditioned on the query. In the context of conversational search and contextualization, we apply this abstractive expansion in a similar, straightforward way: We add expansion embeddings in the end of the query, and allow them to be contextualized by both the conversation history and last user utterance (Equation (4)), as explained in Section 3.4.

Note that this approach differs from previous generative query expansion/rewriting works [44, 74] that use seq2seq models to rewrite the query text in various aspects. Most importantly, our method generates embeddings that can be used directly for matching document tokens and does not require an additional query rewriting model or phase. Additionally, this method does not require supervision from query rewriting datasets (e.g., CANARD [21]).

Abstractive Expansion Results. To study **RQ2**, we add 25 abstractive expansion ([*MASK*]) tokens on top of the previously introduced *last_turn* and contextualised ($ZeCo^2$) variants. We omit the previously used "all_history"baseline, since it already uses the entire history for matching without demonstrating strong performance. We study the number of expansion tokens as a hyperparameter in the follow-up part 6.

We present results on Table 6 and observe that adding abstractive expansion tokens is very beneficial for retrieval, both for the last_query and contextualized variant. The variant that includes both conversational contextualization along with abstractive expansion from the conversation outperforms the rest, improving *Recall* substantially and closing the gap from the human oracle variant. Therefore, we answer **RQ2** positively, concluding that abstractive expansion tokens can be contextualised with the conversation history successfully to improve ranking.

Number of expansion tokens. In this part, we investigate what is the optimal number of expansion ([*MASK*]) tokens to add in the query.

This is important, since query and expansion tokens are linearly combined in the scoring function (Equation (3)). That means that for a query consisting of 10 tokens and 10 expansion tokens,

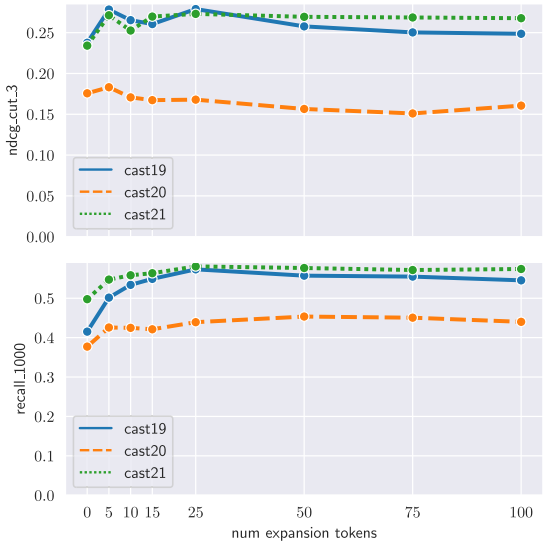


Fig. 4. Ranking performance with respect to the number of (MLM) expansion tokens.

the expansion tokens would contribute 50% of the score. Hence, adding more expansion tokens would bias the score towards query expansion matches and down-weigh the importance of the tokens originating from the last user question, possibly leading to performance degradation.

We measure performance per number of expansion ([*MASK*]) tokens added in Figure 4 and observe that adding MLM expansion tokens improves ranking metrics, with one exception (*NDCG*@3 in CAsT'20). Adding five expansion tokens already provides a reasonable performance boost, while adding 25 tokens performs reasonably well across all settings. One possible explanation for that is that during pre-training, the average number of [*MASK*] tokens added per query was closer to this number (the maximum query length of 32 minus the average query length).

NDCG@3 seems to be much more volatile than *R*@1*K*, since it generally focuses on the top ranking positions and has a lower cutoff. However, Recall seems to have an upward trend as we add more expansion tokens. This seems to be in line with previous work, that showed that query rewrites that achieve a higher *Rouge* − *R* has better Recall effectiveness. This is reasonable, since it suggests that rewrites that add more history terms than others (and hence focus on *Rouge* − *Recall* rather than *Rouge*−*Precision*) tend to achieve better Recall on ranking metrics [67]. Finally, a somewhat surprising finding is that even in the extreme case of adding 100 expansion tokens, expansion helps. Note that in that case, most of the scoring is contributed by expansion tokens rather than query or conversation history terms. It could be possible that the expansion embeddings are replicating some of the embeddings of the last turn as well, or that they are able to detect reasonably good expansion matches from the entire conversation context.

6.1 Effect of Contextualization on Abstractive Expansion Embeddings

To understand the effect of contextualization on the expansion embeddings, we perform an analysis similar to Section 5.1. However, instead of focusing on embeddings of anaphora turns, here we focus on how contextualization changes the expansion (*MASK*) embeddings.

Where Do [*MASK*] *Tokens Resolve to?* As a starting point, we perform a qualitative analysis of conversational queries, where we match expansion embeddings with the most similar embedding from the conversation. Note that we follow the same technique as the previous Table 4, but only

Table 7. Examples of Best Term Matches of ([*MASK*]) Expansion Embeddings to the Conversation History

Utterance	Human resolution	Δ*Recall* (+*ZeCo*^{2})	Δ*Recall* (+*ZeCo*^{2} +*MASK*)	closest match to [*MASK*]	
				term	simil
how did it originally work? [SEP] **[MASK]**	netflix	+0.11	+0.39	how was **Netflix** started?	0.95
why did it create tension with the US? [SEP] **[MASK]**	galileo system	+0.31	+0.52	what is the **GALILEO** system ..	0.76
what are examples of important ones? [SEP] **[MASK]**	real-time databases	+0.07	+0.19	what is a real –time **database** ?	0.88
how does the relationship influence biodiversity? [SEP] **[MASK]**	predator and pray	+0.16	+0.12	what do **predator** plants eat?	0.85
what is different compared to previous legislation? [SEP] **[MASK]**	between GDPR and EU Data Protection Directive 95/46/EC	+0.56	+0.25	**The** gdpr is expected to	0.7
how much does an owner typically make? [SEP] **[MASK]**	a Burger King franchise	0.00	+0.22	purchasing **a** burger king franchise.	0.87
that's later than i expected. who were the winners? [SEP] **[MASK]**	snowboarding winners winter olympics 1998	+0.28	+0.14	snowboarding became a winter olympic sport **in** 1998.	0.82
tell me more about angel rounds. [SEP] **[MASK]**	investment	+0.45	+0.12	**seed** money options include friends	0.91
why did the a380 stop being produced? [SEP] **[MASK]**	Airbus	+0.57	−0.57	responding to lagging a300 - 600f and a310f**sales**, airbus began marketing...	.73
Tell me about the Firebase DB. [SEP] **[MASK]**	–	−0.07	−0.14	what is a real-time **database** ?	0.66
Where was Parmesan cheese created? [SEP] **[MASK]**	–	+0.30	−0.30	what is **mortadella** and [...]	0.75
tell me more about traceability tools. [SEP] **[MASK]**	for GMO foods in the EU	+0.66	−0.30	whether GM products should be **labeled**. Labeling of GMO [..]	0.82
tell me about the origins of the JusticeLeague. [SEP] **[MASK]**	(in the DC universe)	+0.07	−0.23	who are the **Avengers** ?	0.75
are there tourism activities related to trucks or trains? [SEP] **[MASK]**	(in downtown Chattanooga)	N/A	N/A	What is **Chattanooga** famous for?	0.84

The reported $\Delta Recall$ improvements are additive, ie. $\Delta Recall(+ZeCo^2 + MASK)$ measures improvement over $\Delta Recall(+ZeCo^2)$.

report the closest matches after contextualisation for clarity. We choose the top improving and harming cases, as well as some representative examples demonstrating ellipsis or anaphoras.

For each query, we report Recall improvements when contextualization is added on top of the raw queries (+*ZeCo*^{2}) and when one *MASK* token is added on top of the contextualized query (+*ZeCo*^{2}+*MASK*). Improvements are additive, i.e., +*ZeCo*^{2}+*MASK* indicates improvement over +*ZeCo*. We observe a number of cases where the mask token comes close to resolution token embeddings, resulting in improved performance which often surpasses the improvements of contextualisation. Another interesting observation is that this token might often be a non-descriptive article (e.g., *a, the*) of the missing entity or resolution token, or simply a term that is topically related to the conversation topic (e.g., seed money and investment rounds).

On the other hand, we observe a few cases where adding generative expansion tokens harms performance. This is often a result of expanding with tokens different than the missing resolution terms (e.g., *sales* instead of *Airbus*) or in cases when expansion is not needed, yet an incorrect entity term is added (e.g., *mortadella*).

More interestingly though, in some of those examples, we can also observe how [*MASK*] tokens can help in queries that contain the linguistic phenomenon of **ellipsis**; i.e., that certain terms are implied in the context of the previous conversation. We see one such example in the last query of Table 7. The user ask the system about tourism activities, but the location of interest is only implied. In this case, the [*MASK*] token embedding goes closer to the location term "Chattanooga", that would probably boost retrieval (relevance judgements are missing on this query). Other similar examples are the "(*predator and prey*) relationship", where the [*MASK*] embedding goes closer to predator and boosts *Recall* an extra +0.12 points and '(*a Burger King Franchise* owner)" which results in +0.22 Recall points. In all these cases, there was no explicit correference in the last

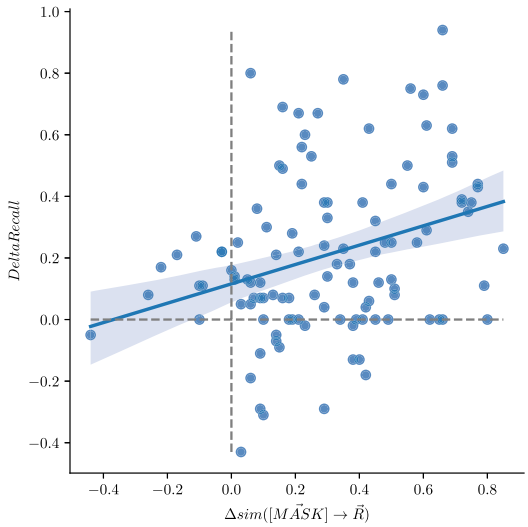


Fig. 5. Correlation between $\Delta Recall$ and similarity change of [MASK] tokens towards resolutions (CAsT '19).

Table 8. Embedding Similarities between Anaphora ($[\vec{A}]$) or $[\vec{MASK}]$ Tokens and Resolution ($\vec{R}$) or Random ($\vec{Rnd}$) Terms from the Conversations (CAsT '19)

	$\vec{R}$ (resolution)	$\vec{Rnd}$ (random term)
$\vec{A}_{last-turn}$	0.178	0.255
$\vec{A}_{ZeCo^2}$	0.372	0.204
$[\vec{MASK}]_{last-turn}$	0.187	0.210
$[\vec{MASK}]_{ZeCo^2}$	0.474	0.230

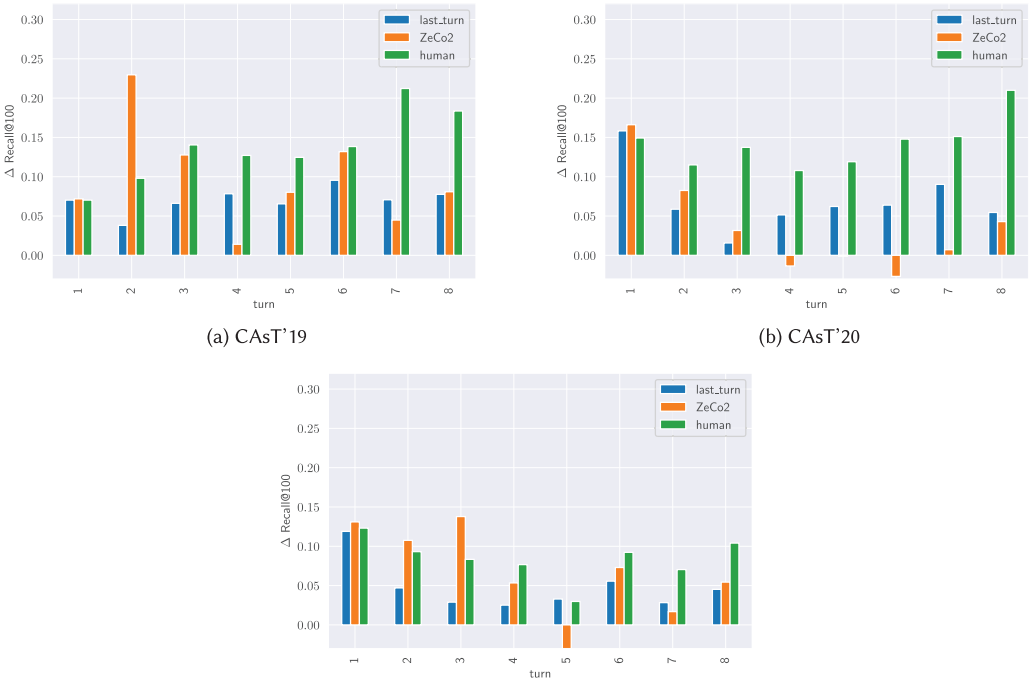
The first two rows are copied from Table 5 for easier reference.

utterance and hence contextualisation alone could not fully capture the relevant resolution terms, without the $[MASK]$ token.

Contextualization Effect on MLM Expansion Embeddings. Further, we seek to investigate how conversational contextualisation affects the $[MASK]$ embeddings. This is important, since this MLM-expansion method has proven to improve performance in ad-hoc search and might be irrelevant with regards to conversational expansion. In a similar spirit of the quantitative analyses performed in Section 5, we define the effect of contextualisation on bringing $[MASK]$ embeddings towards resolutions as $\Delta sim([\vec{MASK}] \rightarrow \vec{R})$.

We observe the scatter plot of this Δsim against $\Delta Recall$ in Figure 5 and detect a significant correlation between those terms (Pearson's $R = 0.31$, $p - value = 0.007$). Note that this correlation is in fact stronger than the one detected between anaphora embeddings and performance improvements. We observe again that, in most cases, adding a contextualised expansion embedding improves $Recall$ ($\Delta Recall > 0$), while contextualisation brings those embeddings closer to resolutions ($\Delta sim > 0$).

In Table 8, we follow the same analysis of Table 5 and measure similarities between contextualized, non-contextualized $[MASK]$ tokens and random ($\vec{Rnd}$) vs. resolution terms ($\vec{R}$) from the conversation. We confirm that resolution terms have a greater effect on the expansion embeddings.



(a) CAsT'19

(b) CAsT'20

(c) CAsT'21

Fig. 6. Per turn improvements when Abstractive expansion tokens are added upon different query variants (*last_turn*, $ZeCo^2$ and *human*).

Comparing the numbers that correspond to [*MASK*] tokens (last two rows) to the ones corresponding to Anaphoras (first two rows), we can conclude that that contextualization affect more and bring closer to resolutions the [*MASK*] tokens (0.474), rather than the anaphoras (0.372; Table 5).

In practice, this strengthens our intuition that abstractive expansion can be a very effective technique for identifying and implicitly adding salient conversation terms for conversational search. In Section 7 we build on this finding to help us identify relevant expansion terms in an extractive way.

6.2 Robustness of Abstractive Expansion across Turns

To better understand whether the abstractive expansion from conversations can be applied robustly to all conversation depths, we compare the per-turn improvements in performance when expansion is added, on the $ZeCo^2$, *last_turn* and *human* variants (Figure 6). Improvements over the *last_turn* variant (in blue) and human variant (in green) do not involve conversations or the conversational contextualization and can therefore be thought of the added benefit of using the ColBERT expansion. However, the improvements of using [*MASK*] expansion tokens also involve the capacity of the encoder to take advantage of the conversation as an expansion source.

First, we notice that adding the abstractive expansion is almost always beneficial, across depths, variants and datasets ($\Delta Recall > 0$). In general, we see that expanding upon the *last turn* brings small but stable improvements, which is expected. Expanding upon the *human* variant brings robust improvements, of larger magnitudes. This demonstrates that expansion is helpful even upon

oracle queries and is in agreement with the original results of ColBERT [37]. The larger magnitude of the improvements is probably due to the fact that oracle queries provide a clearer signal for expansion compared to the noisy signal of the whole conversation.

Looking more closely into the improvements upon $ZeCo^2$ variant, we see they can also be very strong although not entirely robust across turns. Specifically, those improvements seem to follow a U-shaped distribution: they start strong in early turns, drop in middle turns and finally recover in the few last turns. Stagnation in middle turns is expected, since after 3-4 conversation rounds the context accumulates but it is unclear whether the recovery of the improvements towards the end of the conversation has a clear explanation or is related to some dataset artefact.

In principle, expansion upon $ZeCo^2$ brings more improvements compared to expanding upon *last_turn* queries. This indicates that the abstractive expansion mechanism can effectively utilize the conversation history in most datasets and depths. However, the added benefit is generally lower than what we observe upon the *human* variant, where the expansion cannot be confused with irrelevant terms since it can only be affected by the oracle queries. In any case, we see that the expansion technique is beneficial at different conversation depths. However, our analysis suggests that it might be less effective when the context is lengthier or in the last few turns where queries can have more complex dependencies. This suggests that a better way to handle complex dependencies and lengthier context is needed and supervision might be beneficial.

Abstractive Expansion Conclusions

In this part, we summarise our findings on the abstractive expansion technique we introduced. We conclude that the generated abstractive expansion token embeddings improve ranking, and when those are contextualized jointly with respect to the conversation and current turn we achieve better performance. This means that the abstractive expansion can effectively use the conversation history to bring important missing terms and concepts into account. In fact, we find that the ability of [MASK] tokens seem more effective in doing so compared to anaphora terms (e.g., *it*). Lastly, we conclude that the amount of expansion tokens added does not play an important role, while the benefits of contextualising expansion tokens fade as conversations grow longer.

7 EXTRACTIVE EXPANSION BASED ON ATTENTION WEIGHTS

Expanding queries with terms from the conversation history has been one of the main approaches used for Conversational Passage Retrieval. The goal is to make the query of the last turn self-contained and independent of the conversation [46, 70]. Selecting the right expansion terms is very important as we saw in Table 2, where we observed big gaps across the performance of the *last_turn*, *all_history*, and *human* variants. However, most query expansion models [43, 50, 70, 75] so far rely on CANARD [21] for supervision, or synthetic data. In this section we explore to what extent a dense retriever can effectively select the expansion terms without supervision. To do so, we rely on attention weights for identifying missing words from the conversation history. Attention scores have been used multiple times in the past as a means of explaining the predictions of transformers. In fact, the value of attention weights for explainability is an on-going debate in the research community [7, 35, 71], while various methods have been introduced to measure the flow of information in the transformer through attention weights [1, 22].

For the case of conversational passage retrieval, we saw in earlier sections that the dense retriever can effectively contextualize queries since it introduces salient terms from the conversation history. One effect of that was bringing anaphora term embeddings closer to their resolutions. This indicates that the attention mechanism might be a useful resource towards distinguishing history terms that complement the last turn. To identify terms that are referred or connected with the current context to previous, we measure attention weights from the current (last_turn) to the history.

Our preliminary experiments showed little success when aggregating scores from all last_turn tokens, so we focus on measuring attentions from 3 different types of tokens from the conversation history:

- *Oracle Anaphoras:* Anaphora terms, identified as as the ones missing from the human resolutions (this is an oracle setting because it requires knowing the human resolution)
- *Vocab. Anaphoras:* Anaphora terms from a vocabulary (e.g., "they", "it")
- *[MASK] Token:* The Masked-Language-Modelling token, which was also used in Section 6 for abstractive expansion

We perform experiments using a fixed number of terms (1, 5, 10 and 15) in Table 9 for all CAsT datasets. Following previous work [67, 70] we report ranking performance and intrinsic evaluation metrics, which measures performance on successfully retrieving the word-level resolutions from the conversation. Regarding the latter we report $Rouge-1$, which measures unigram overlap, using the human rewritten queries as reference text and generated queries as candidate text.

Comparison of Expansion Source Token in Attention-based Methods. In terms of Rouge-Precision and F1, we observe that expanding from a vocabulary of anaphora terms (e.g., "he", "it", etc.) performs best. Intuitively, it is natural that the encoder would be more successful in picking up resolution terms (especially co-references) from these predefined anaphora terms. However, while this is consistent across different settings, it rarely outperforms expanding from other sources (oracle anaphoras or MASK tokens) in terms of ranking metrics. This could be due to the fact that fewer queries are expanded comparably to other methods, since the predefined anaphora vocabulary is quite limited.

Expanding from oracle anaphoras leads to marginally better $NDCG@3$ in CAsT'20. However, focusing on $[MASK]$ token attentions allows us to achieve higher Rouge-Recall, leading to always higher (ranking) *Recall* and competitive $NDCG@3$ performance. Results indicate that expanding from $[MASK]$ tokens is capable of adding the most relevant terms from the history, preserving a good trade-off between precision and recall based metrics. Additionally, as we discussed on Section 6, this is the only method that would be able to detect the phenomenon of ellipsis, which is very common in this setting.

Optimal Number of Expansion Source Tokens in Attention-Based Methods. As expected, we observe that, as we add more expansion terms to the query, Rouge-P deteriorates and Rouge-R improves. A good tradeoff is achieved across metrics when adding five terms, which is also intuitively a reasonable number. In any case, our extractive expansion method hardly ever harms ranking performance and usually provides improvements of several $NDCG$ and *Recall* points. Hence, we can conclude that expanding based on attention scores is a simple yet effective method for zero-shot conversational query expansion, based only on the query encoder of a Dense Retriever model.

For the remainder of this article, we add five extracted expansion terms based on $MASK$s, since this method allows us to abolish dependences on specific anaphora terms and ranking metrics confirm its effectiveness. We compare the performance of the extractive expansion method we introduced to other extractive baselines in Section 8.

Extractive Expansion Conclusions

In this section, we investigated to what extent attention scores from the query encoder can be used as a means to identify resolution terms from the history, in a zero-shot way. We found that attention can be successfully used to identify such terms from the history, and adding their embeddings to the query improves ranking performance. We also concluded that measuring the attention weights from the abstractive expansion tokens ($[MASK]$) is more effective and flexible compared to relying

Table 9. Extractive Expansion Results

expansion source	# expansion terms	# expanded queries	Rouge-P	Rouge-R	Rouge-F1	NDCG@3	R@100	R@1K
CAsT'19								
No expansion	0	0/479				0.238	0.216	0.415
oracle anaphoras	1	213/479	**0.89**	**0.79**	**0.83**	0.244	0.220	0.445
vocab. anaphoras		161/479	**0.90**	0.78	0.82	0.244	0.218	0.423
[MASK] token		429/479	0.83	0.79	0.80	**0.247**	**0.231**	**0.439**
oracle anaphoras	5	213/479	0.81	0.83	**0.80**	0.248	0.218	0.444
vocab. anaphoras		161/479	**0.83**	0.81	**0.80**	0.247	0.220	0.433
[MASK] token		429/479	0.63	**0.85**	0.71	**0.279**	**0.256**	**0.495**
oracle anaphoras	10	213/479	0.75	0.84	0.75	0.259	0.217	0.450
vocab. anaphoras		161/479	**0.79**	0.82	**0.77**	0.258	0.223	0.438
[MASK] token		429/479	0.50	**0.88**	0.62	**0.266**	**0.241**	**0.485**
oracle anaphoras	15	213/479	0.72	0.85	0.73	0.250	0.215	0.439
vocab. anaphoras		161/479	**0.76**	0.83	**0.75**	**0.253**	0.220	0.430
[MASK] token		429/479	0.44	**0.89**	0.56	0.245	**0.225**	**0.458**
CAsT'20								
No expansion	0	0/216				0.176	0.200	0.377
oracle anaphoras	1	121/216	0.82	**0.67**	0.72	**0.179**	0.208	0.388
vocab. anaphoras		57/216	**0.84**	0.66	**0.73**	**0.179**	0.208	0.384
[MASK] token		190/216	0.78	**0.67**	0.71	0.176	**0.210**	**0.388**
oracle anaphoras	5	121/216	0.73	0.69	0.69	**0.178**	0.216	0.394
vocab. anaphoras		57/216	**0.79**	0.68	**0.71**	0.177	0.213	0.393
[MASK] token		190/216	0.62	**0.71**	0.65	0.170	**0.230**	**0.420**
oracle anaphoras	10	121/216	0.67	0.72	0.66	**0.177**	0.213	0.397
vocab. anaphoras		57/216	**0.77**	0.69	**0.70**	0.175	0.210	0.399
[MASK] token		190/216	0.52	**0.74**	0.59	0.175	**0.225**	**0.430**
oracle anaphoras	15	121/216	0.63	0.74	0.63	**0.173**	0.214	0.399
vocab. anaphoras		57/216	**0.75**	0.70	**0.69**	**0.173**	0.210	0.399
[MASK] token		190/216	0.45	**0.77**	0.55	0.170	**0.225**	**0.425**
CAsT'21								
No expansion	0	0/239				0.234	0.267	0.498
oracle anaphoras	1	108/239	0.85	**0.68**	**0.74**	0.238	**0.272**	0.505
vocab. anaphoras		61/239	**0.86**	**0.68**	**0.74**	0.239	**0.272**	0.502
[MASK] token		184/239	0.82	**0.68**	0.73	**0.247**	0.271	**0.513**
oracle anaphoras	5	108/239	0.78	**0.70**	0.71	0.248	0.285	0.511
vocab. anaphoras		61/239	**0.83**	0.69	**0.73**	0.247	0.279	0.507
[MASK] token		184/239	0.68	**0.70**	0.67	**0.262**	**0.300**	**0.535**
oracle anaphoras	10	108/239	0.73	0.71	0.69	0.244	0.282	0.512
vocab. anaphoras		61/239	**0.79**	0.69	**0.71**	0.242	0.275	0.510
[MASK] token		184/239	0.58	**0.72**	0.62	**0.264**	**0.293**	**0.529**
oracle anaphoras	15	108/239	0.71	0.72	0.67	0.247	0.282	0.512
vocab. anaphoras		61/239	**0.78**	0.70	**0.69**	0.247	0.275	0.513
[MASK] token		184/239	0.54	**0.73**	0.59	**0.264**	**0.296**	**0.538**

Best overall results are underlined, while best across each group are **boldfaced**.

on anaphora terms, while adding 5 expansion terms from the history seems to achieve a good trade-off between intrinsic and extrinsic evaluation metrics.

8 COMBINING CONTEXTUALIZATION WITH UNIFIED AND INDEPENDENT EXPANSIONS

The methods introduced in this paper can be roughly categorized into two families: (i) **contextualization** methods that influence last turn embeddings using the conversation history, and (ii) **expansion** methods that explicitly try to resolve the conversational query dependencies by appending term embeddings. While these methods work differently, they can be combined to rank documents with conversational queries. We explore this possibility in this section. Apart from the unified expansion–retrieval methods we introduced, we also include baseline extractive expansion methods that do not depend on the Dense Retriever. Specifically, we consider the following expansion methods:

— *Abstractive-U*: Abstractive Expansion, Unified in the retriever and zero-shot (from Section 6).
— *Extractive-U*: Extractive Expansion, Unified in the retriever and zero-shot (from Section 7)
— *Extractive-CORP*: Corpus-based extractive expansion HQE [46], a zero-shot keyword-extraction method that relies on BM25 scores to (a) model the need for query expansion and (b) select expansion keywords.
— *Extractive-TRAIN*: QuReTeC [70], a trained query expansion model.

We only consider extractive expansion methods and not generative methods, since the latter can add words that do not exist in the conversation history. We choose HQE as a competing zero-shot method and QuReTeC as a trained state-of-the-art baseline. We run all rewritten queries across the same Dense Retriever (ColBERT), ensuring for fair comparison. That also allows us to assess expansion methods independently and regardless of the contextualization, which allows us to contrast unified expansion–retrieval models with (i) methods that use the retrieval corpus and its statistics through BM25 scores (HQE) and (ii) trained expansion methods (QuReTeC).

To combine query rewrites with contextualization, we follow the same process described in Section 3.6: We encode the entire conversation including the current turn, but match using only the embeddings of the current turn, including the historical expansion terms as indicated from the methods above. We perform extrinsic evaluation of methods (passage ranking performance) in Table 10. For extractive methods, we also provide expansion statistics, as well as an intrinsic evaluation (*Rouge*) in Table 11.

Unified Expansion-Retrieval Methods. We observe that among the methods that work solely based on the query encoder, the abstractive method is the most effective. Combining our abstractive and extractive query-encoder based methods expansion does not yield additive performance improvements, however, it ensures more robust performance in CAsT'20, where methods individually are not so strong.

Benchmarking Expansion Methods without Contextualization. Looking at the results of the other two expansion methods, some interesting observations arise. The trained expansion method (QuReTeC), which we expect to be superior, works much better in terms of *NDCG* in CAsT'19 and '20. However, the unsupervised keyword-based method HQE demonstrates stronger Recall performance in those two datasets and outperforms the trained expansion model. This is likely has to do with HQE expanding queries much more aggressively compared to other methods (see Table 11). This phenomenon has also been observed in mixed-initiative conversations [40].

Looking into CAsT'21, however, we see that both of these models fail to outperform our unsupervised variants. QuReTeC is on par with our method in terms of *NDCG*, where its performance

Table 10. Combining Contextualization with Various Expansion Methods

Query					CAsT'19			CAsT'20			CAsT'21		
Contextualization	Abstractive-U	Extractive-U	Extractive-CORP [46]	Extractive-TRAIN [70]	NDCG@3	R@100	R@1K	NDCG@3	R@100	R@1K	NDCG@3	R@100	R@1K
					0.214	0.157	0.365	0.155	0.124	0.224	0.140	0.154	0.277
✔					0.2377	0.216	0.415	**0.176**	0.200	0.378	0.234	0.267	0.498
✔	✔				**0.279**	**0.31**	**0.573**	0.168	**0.237**	**0.440**	**0.273**	**0.332**	**0.581**
✔		✔			**0.279**	0.256	0.495	0.170	0.230	0.420	0.2617	0.300	0.535
✔	✔	✔			**0.275**	**0.307**	**0.573**	**0.178**	**0.24**	**0.438**	**0.275**	**0.330**	**0.584**
			✔		**0.304**	**0.334**	**0.590**	**0.200**	**0.288**	**0.537**	0.205	0.290	0.526
✔			✔		0.267	0.267	0.532	0.185	0.265	0.487	**0.254**	**0.316**	**0.563**
				✔	**0.388**	**0.319**	0.512	**0.234**	**0.290**	0.472	0.235	0.280	0.468
✔				✔	0.283	0.294	**0.553**	0.183	**0.290**	**0.536**	**0.271**	**0.324**	**0.552**
Human					0.430	0.363	0.547	0.443	0.408	0.620	0.431	0.403	0.636

Best results across each group shown in **bold** and best overall results underlined.

Table 11. Intrinsic Evaluation and Statistics for Various Expansion Methods

	# expansion terms	# expanded queries	Rouge-P	Rouge-R	Rouge-F1
CAsT'19					
Extractive-QE	5	429/479	0.63	**0.85**	0.71
Extractive-CORPUS [46]	7.1 (avg)	429/479	0.63	**0.97**	0.74
Extractive-TRAIN [70]	1.9 (avg)	356/479	**0.95**	0.96	**0.95**
CAsT'20					
Extractive-QE	5	190/216	0.62	**0.71**	0.65
Extractive-CORPUS [46]	17.5 (avg)	190/216	0.40	**0.80**	0.50
Extractive-TRAIN [70]	2.6 (avg)	146/216	**0.81**	0.76	**0.78**
CAsT'21					
Extractive-QE	5	184/239	0.68	**0.70**	0.67
Extractive-CORPUS [46]	75.5 (avg)	213/239	0.22	**0.80**	0.30
Extractive-TRAIN [70]	2.4 (avg)	147/239	**0.83**	0.74	**0.76**

is stronger, but starts lacking further in terms of *Recall*. This is likely due to the gradual changes in conversations introduced across years in CAsT, the most important being the long canonical document responses and user provided feedback. This could result in a gap between the conversations seen from the QuReTeC model at training time (from CANARD dataset) versus inference, highlighting the importance in the existence of unsupervised methods too.

When it comes to HQE, the corpus-dependent unsupervised method, we see that its aggressive expansion tendency backfires in CAsT'21, where canonical responses consist of long documents. Specifically, HQE adds as many as 75 expansion terms per query on average (Table 11), which

makes it less competitive—especially in top ranking positions. Nonetheless, its overall performance indicates the value of using ranking signals (e.g., BM25 scores) for query expansion.

Contextualizing Independent Query Expansion Methods. Following up, we now explore whether the non-query-encoder based expansion methods can benefit from contextualization. We observe that contextualization does not seem to further help HQE or QuReTeC in most cases. This comes with the exception of CAsT'21, where both the corpus-based and trained expansion lack in comparison to contextualization and their combination bridges the gap from it. However, the only additive improvement we observe is in the case of CAsT'20 (QuReTeC + Contextualization), for Recall and especially at higher cutoffs.

Those results raise the question on whether incorporating expansion terms and contextualization could be improved.

9 CONCLUSIONS

In this paper, we explore the possibility of performing conversational search in a zero-shot setting, solely using the "knowledge"encoded in the weights of a Dense Retriever. We do so by (a) contextualizing embeddings with respect to the conversation history, and (b) expanding conversational queries using abstractive and extractive methods that are unified inside the Dense Retriever's query encoder.

We answer **RQ1** positively, that is, zero-shot contextualisation of conversational queries improves dense passage retrieval. We find that $ZeCo^2$ introduces a desirable bias towards salient terms from the conversation, mostly in the basis of bringing anaphora terms closer to their resolution terms. Contextualisation improves performance in most conversational depths, but demonstrates stronger improvements in earlier turns. We also introduce expansion to improve performance further, specifically targeting ellipsis phenomena (omitting content from previous rounds without explicit reference to it). In **RQ2**, we explore abstractive expansion, which concatenates [*MASK*] tokens to the end of the conversation, allowing them to be contextualised through it and use their embeddings for matching. We find that contextualizing the abstractive expansion also improves dense passage retrieval by introducing a similar, yet even stronger bias towards salient conversation terms.

Next, we investigate **RQ3** that uses a simple heuristic based on attention weights to extract important previous terms (and their embeddings) from the Dense Retriever. We conclude that this method also improves dense passage retrieval and measuring the attention weights from the abstractive expansion ([MASK]) tokens is most effective on identifying resolutions and improving retrieval.

Finally, in **RQ4** we compare our unified resolution–retrieval methods with other zero-shot or trained extractive expansion baselines, concluding that performance of our zero-shot unified methods is competitive and outperforms the baselines in CAsT'21, where the conversation includes different characteristics compared to previous years. In general, we find that while trained expansion methods to be more precise and effective in the top of the ranking (i.e., *NDCG*), unsupervised methods can outperform them in *Recall*. Additionally, we find that our unified zero-shot resolution–retrieval methods outperform the rest in CAsT'21, where new discourse characteristics are added in the queries. Lastly, we find that combining contextualization with expansion is effective when expansion is unified on the Dense Retriever and is more challenging with external expansion systems.

As a future direction, we aim to extend this work to a few-shot setting. Another valuable direction is to improve the extractive expansion, by choosing a more sophisticated attention scoring method and relaxing the condition of manual selection of number of terms to be added. Other important directions are related to the finding that supervised methods do not always outperform

their zero-shot counterparts. This is related to (i) a trade-off between Precision and Recall, as well as (ii) changes in the characteristics of conversations, such as richer interactions (user feedback) and handling longer context (introducing canonical document responses). The former indicates that different rewriting or expansion methods might be needed for first-stage rankers to reach a higher recall. The latter highlights that further research is needed in handling lengthier interactions as well as richer ones. This calls for a set of more generalizable methods that can be more invariant to changes in the forms and characteristics of conversations. This is needed since current conversational datasets [2, 5, 15–17] have fixed characteristics defined by dataset creators and thus can fall short in real-world scenarios, considering the full amount of interactions a conversation with a real user can cover.

REFERENCES

[1] Samira Abnar and Willem Zuidema. 2020. Quantifying attention flow in transformers. *arXiv preprint arXiv:2005.00928* (2020).

[2] Vaibhav Adlakha, Shehzaad Dhuliawala, Kaheer Suleman, Harm de Vries, and Siva Reddy. 2021. TopiOCQA: Open-domain conversational question answeringwith topic switching. *arXiv preprint arXiv:2110.00768* (2021).

[3] Mohammad Aliannejadi, Hamed Zamani, Fabio Crestani, and W. Bruce Croft. 2019. Asking clarifying questions in open-domain information-seeking conversations. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 475–484.

[4] Avishek Anand, Lawrence Cavedon, Matthias Hagen, Hideo Joho, Mark Sanderson, and Benno Stein. 2020. Conversational search–a report from Dagstuhl seminar 19461. *arXiv preprint arXiv:2005.08658* (2020).

[5] Raviteja Anantha, Svitlana Vakulenko, Zhucheng Tu, Shayne Longpre, Stephen Pulman, and Srinivas Chappidi. 2020. Open-domain question answering goes conversational via question rewriting. *arXiv preprint arXiv:2010.04898* (2020).

[6] Leif Azzopardi, Mateusz Dubiel, Martin Halvey, and Jeffery Dalton. 2018. Conceptualizing agent-human interactions during the conversational search process. In *Proceedings of the 2nd International Workshop on Conversational Approaches to Information Retrieval*.

[7] Joris Baan, Maartje ter Hoeve, Marlies van der Wees, Anne Schuth, and Maarten de Rijke. 2019. Do transformer attention heads provide transparency in abstractive summarization? *arXiv preprint arXiv:1907.00570* (2019).

[8] Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473* (2014).

[9] Bernd Bohnet, Vinh Q. Tran, Pat Verga, Roee Aharoni, Daniel Andor, Livio Baldini Soares, Jacob Eisenstein, Kuzman Ganchev, Jonathan Herzig, Kai Hui, et al. 2022. Attributed question answering: Evaluation and modeling for attributed large language models. *arXiv preprint arXiv:2212.08037* (2022).

[10] Oleg Borisov, Mohammad Aliannejadi, and Fabio Crestani. 2021. Keyword extraction for improved document retrieval in conversational search. *arXiv preprint arXiv:2109.05979* (2021).

[11] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question answering in context. *arXiv preprint arXiv:1808.07036* (2018).

[12] Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. ORCAS: 20 million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*. 2983–2989.

[13] Zhuyun Dai and Jamie Callan. 2019. Deeper text understanding for IR with contextual neural language modeling. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 985–988.

[14] Zhuyun Dai, Arun Tejasvi Chaganty, Vincent Y. Zhao, Aida Amini, Qazi Mamunur Rashid, Mike Green, and Kelvin Guu. 2022. Dialog inpainting: Turning documents into dialogs. In *Proceedings of the International Conference on Machine Learning*. PMLR, 4558–4586.

[15] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. CAsT 2020: The conversational assistance track overview. *Proceedings of TREC* (2020).

[16] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2020. TREC CAsT 2019: The conversational assistance track overview. *arXiv preprint arXiv:2003.13624* (2020).

[17] Jeffrey Dalton, Chenyan Xiong, and Jamie Callan. 2022. CAsT 2021: The conversational assistance track overview. *Proceedings of TREC* (2022).

[18] Marco Del Tredici, Gianni Barlacchi, Xiaoyu Shen, Weiwei Cheng, and Adrià de Gispert. 2021. Question rewriting for open-domain conversational QA: Best practices and limitations. In *Proceedings of the 30th ACM International Conference on Information & Knowledge Management*. 2974–2978.

[19] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. BERT: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805* (2018).
[20] Laura Dietz, Manisha Verma, Filip Radlinski, and Nick Craswell. 2017. TREC complex answer retrieval overview.. In *Proceedings of TREC*.
[21] Ahmed Elgohary, Denis Peskov, and Jordan Boyd-Graber. 2019. Can you unpack that? Learning to rewrite questions-in-context. *Can You Unpack That? Learning to Rewrite Questions-in-Context* (2019).
[22] Javier Ferrando, Gerard I. Gállego, and Marta R. Costa-Jussà. 2022. Measuring the mixing of contextual information in the transformer. *arXiv preprint arXiv:2203.04212* (2022).
[23] Rafael Ferreira, Mariana Leite, David Semedo, and Joao Magalhaes. 2021. Open-domain conversational search assistant with transformers. In *Proceedings of the European Conference on Information Retrieval*. Springer, 130–145.
[24] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. SPLADE: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2288–2292.
[25] Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. A white box analysis of ColBERT. In *Proceedings of the European Conference on Information Retrieval*. Springer, 257–263.
[26] Jianfeng Gao, Chenyan Xiong, and Paul Bennett. 2020. Recent advances in conversational information retrieval. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 2421–2424.
[27] Jianfeng Gao, Chenyan Xiong, Paul Bennett, and Nick Craswell. 2022. Neural approaches to conversational information retrieval. *arXiv preprint arXiv:2201.05176* (2022).
[28] Yifan Gao, Piji Li, Irwin King, and Michael R Lyu. 2019. Interconnected question generation with coreference alignment and conversation flow modeling. *arXiv preprint arXiv:1906.06893* (2019).
[29] Xinyang Geng, Arnav Gudibande, Hao Liu, Eric Wallace, Pieter Abbeel, Sergey Levine, and Dawn Song. 2023. Koala: A dialogue model for academic research. *Blog Post, April* 1 (2023).
[30] Jing Gu, Mostafa Mirshekari, Zhou Yu, and Aaron Sisto. 2021. ChainCQG: Flow-aware conversational question generation. *arXiv preprint arXiv:2102.02864* (2021).
[31] Sebastian Hofstätter, Navid Rekabsaz, Carsten Eickhoff, and Allan Hanbury. 2019. On the effect of low-frequency terms on neural-IR models. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1137–1140.
[32] Neil Houlsby, Andrei Giurgiu, Stanislaw Jastrzebski, Bruna Morrone, Quentin De Laroussilhe, Andrea Gesmundo, Mona Attariyan, and Sylvain Gelly. 2019. Parameter-efficient transfer learning for NLP. In *Proceedings of the International Conference on Machine Learning*. PMLR, 2790–2799.
[33] Seonjeong Hwang, Yunsu Kim, and Gary Geunbae Lee. 2022. Multi-type conversational question-answer generation with closed-ended and unanswerable questions. *arXiv preprint arXiv:2210.12979* (2022).
[34] Gautier Izacard and Edouard Grave. 2020. Distilling knowledge from reader to retriever for question answering. *arXiv preprint arXiv:2012.04584* (2020).
[35] Sarthak Jain and Byron C. Wallace. 2019. Attention is not explanation. *arXiv preprint arXiv:1902.10186* (2019).
[36] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. *arXiv preprint arXiv:2004.04906* (2020).
[37] Omar Khattab and Matei Zaharia. 2020. ColBERTt: 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*. 39–48.
[38] Gangwoo Kim, Sungdong Kim, Kang Min Yoo, and Jaewoo Kang. 2022. Generating information-seeking conversations from unlabeled documents. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2362–2378.
[39] Andreas Köpf, Yannic Kilcher, Dimitri von Rütte, Sotiris Anagnostidis, Zhi-Rui Tam, Keith Stevens, Abdullah Barhoum, Nguyen Minh Duc, Oliver Stanley, Richárd Nagyfi, et al. 2023. OpenAssistant conversations–democratizing large language model alignment. *arXiv preprint arXiv:2304.07327* (2023).
[40] Antonios Minas Krasakis, Mohammad Aliannejadi, Nikos Voskarides, and Evangelos Kanoulas. 2020. Analysing the effect of clarifying questions on document ranking in conversational search. In *Proceedings of the 2020 ACM SIGIR on International Conference on Theory of Information Retrieval*. 129–132.
[41] Antonios Minas Krasakis, Andrew Yates, and Evangelos Kanoulas. 2022. Zero-shot query contextualization for conversational search. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1880–1884.
[42] Vaibhav Kumar and Jamie Callan. 2020. Making information seeking easier: An improved pipeline for conversational search. In *Findings of the Association for Computational Linguistics: EMNLP 2020*. 3971–3980.

[43] Nam Le Hai, Thomas Gerald, Thibault Formal, Jian-Yun Nie, Benjamin Piwowarski, and Laure Soulier. 2023. CoS-PLADE: Contextualizing SPLADE for conversational information retrieval. *arXiv e-prints* (2023), arXiv–2301.
[44] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. [n. d.]. TREC 2020 notebook: CAsT track. ([n. d.]).
[45] Sheng-Chieh Lin, Jheng-Hong Yang, and Jimmy Lin. 2021. Contextualized query embeddings for conversational search. *arXiv preprint arXiv:2104.08707* (2021).
[46] Sheng-Chieh Lin, Jheng-Hong Yang, Rodrigo Nogueira, Ming-Feng Tsai, Chuan-Ju Wang, and Jimmy Lin. 2021. Multi-stage conversational passage retrieval: An approach to fusing term importance estimation and neural query rewriting. *ACM Transactions on Information Systems (TOIS)* 39, 4 (2021), 1–29.
[47] Nelson F Liu, Tianyi Zhang, and Percy Liang. 2023. Evaluating verifiability in generative search engines. *arXiv preprint arXiv:2304.09848* (2023).
[48] Kelong Mao, Zhicheng Dou, Haonan Chen, Fengran Mo, and Hongjin Qian. 2023. Large language models know your contextual search intent: A prompting framework for conversational search. *arXiv preprint arXiv:2303.06573* (2023).
[49] Kelong Mao, Zhicheng Dou, Hongjin Qian, Fengran Mo, Xiaohua Cheng, and Zhao Cao. 2022. ConvTrans: Transforming web search sessions for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 2935–2946.
[50] Kelong Mao, Hongjin Qian, Fengran Mo, Zhicheng Dou, Bang Liu, Xiaohua Cheng, and Zhao Cao. 2023. Learning denoised and interpretable session representation for conversational search. In *Proceedings of the ACM Web Conference 2023*. 3193–3202.
[51] Mike Mintz, Steven Bills, Rion Snow, and Dan Jurafsky. 2009. Distant supervision for relation extraction without labeled data. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*. 1003–1011.
[52] Fengran Mo, Kelong Mao, Yutao Zhu, Yihong Wu, Kaiyu Huang, and Jian-Yun Nie. 2023. ConvGQR: Generative query reformulation for conversational search. *arXiv preprint arXiv:2305.15645* (2023).
[53] 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. In *CoCo@ NIPS*.
[54] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35 (2022), 27730–27744.
[55] Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, et al. 2020. KILT: A benchmark for knowledge intensive language tasks. *arXiv preprint arXiv:2009.02252* (2020).
[56] Hongjin Qian and Zhicheng Dou. 2022. Explicit query rewriting for conversational dense retrieval. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*. 4725–4737.
[57] Chen Qu, Liu Yang, Cen Chen, Minghui Qiu, W Bruce Croft, and Mohit Iyyer. 2020. Open-retrieval conversational question answering. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 539–548.
[58] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. 2019. Language models are unsupervised multitask learners. *OpenAI Blog* 1, 8 (2019), 9.
[59] Filip Radlinski and Nick Craswell. 2017. A theoretical framework for conversational search. In *Proceedings of the 2017 Conference on Conference Human Information Interaction and Retrieval*. 117–126.
[60] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, Peter J Liu, et al. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.* 21, 140 (2020), 1–67.
[61] Pengjie Ren, Zhumin Chen, Zhaochun Ren, Evangelos Kanoulas, Christof Monz, and Maarten De Rijke. 2021. Conversations with search engines: SERP-based conversational response generation. *ACM Trans. Inf. Syst.* 39, 4, Article 47 (Aug 2021), 29 pages. https://doi.org/10.1145/3432726
[62] Keshav Santhanam, Omar Khattab, Jon Saad-Falcon, Christopher Potts, and Matei Zaharia. 2021. Colbertv2: Effective and efficient retrieval via lightweight late interaction. *arXiv preprint arXiv:2112.01488* (2021).
[63] Ivan Sekulić, Mohammad Aliannejadi, and Fabio Crestani. 2022. Evaluating mixed-initiative conversational search systems via user simulation. In *Proceedings of the 15th ACM International Conference on Web Search and Data Mining*. 888–896.
[64] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. *Advances in Neural Information Processing Systems* 27 (2014).
[65] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971* (2023).

[66] Svitlana Vakulenko, Shayne Longpre, Zhucheng Tu, and Raviteja Anantha. 2021. Question rewriting for conversational question answering. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 355–363.

[67] Svitlana Vakulenko, Nikos Voskarides, Zhucheng Tu, and Shayne Longpre. 2021. A comparison of question rewriting methods for conversational passage retrieval. *arXiv preprint arXiv:2101.07382* (2021).

[68] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in Neural Information Processing Systems*. 5998–6008.

[69] Nikos Voskarides. 2021. Supporting search engines with knowledge and context. *arXiv preprint arXiv:2102.06762* (2021).

[70] Nikos Voskarides, Dan Li, Pengjie Ren, Evangelos Kanoulas, and Maarten de Rijke. 2020. Query resolution for conversational search with limited supervision. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 921–930.

[71] Sarah Wiegreffe and Yuval Pinter. 2019. Attention is not not explanation. *arXiv preprint arXiv:1908.04626* (2019).

[72] Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate nearest neighbor negative contrastive learning for dense text retrieval. *arXiv preprint arXiv:2007.00808* (2020).

[73] Andrew Yates, Rodrigo Nogueira, and Jimmy Lin. 2021. Pretrained transformers for text ranking: BERT and beyond. In *Proceedings of the 14th ACM International Conference on Web Search and Data Mining*. 1154–1156.

[74] Shi Yu, Jiahua Liu, Jingqin Yang, Chenyan Xiong, Paul Bennett, Jianfeng Gao, and Zhiyuan Liu. 2020. Few-shot generative conversational query rewriting. In *Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval*. 1933–1936.

[75] Shi Yu, Zhenghao Liu, Chenyan Xiong, Tao Feng, and Zhiyuan Liu. 2021. Few-shot conversational dense retrieval. *arXiv preprint arXiv:2105.04166* (2021).

[76] Hamed Zamani, Johanne R. Trippas, Jeff Dalton, and Filip Radlinski. 2022. Conversational information seeking. *arXiv preprint arXiv:2201.08808* (2022).

Received 14 June 2023; revised 17 September 2023; accepted 6 November 2023