cTBLS: Augmenting Large Language Models with Conversational Tables

Anirudh S Sundar, Larry Heck
AI Virtual Assistant (AVA) Lab
The Georgia Institute of Technology
{asundar34,larryheck}@gatech.edu

Abstract

Optimizing accuracy and performance while eliminating hallucinations of open-domain conversational large language models (LLMs) is an open research challenge. A particularly promising direction is to augment and ground LLMs with information from structured sources. This paper introduces Conversational Tables (cTBLS), a three-step architecture to retrieve and generate dialogue responses grounded on retrieved tabular information. cTBLS uses Transformer encoder embeddings for Dense Table Retrieval and obtains up to 125% relative improvement over the retriever in the previous state-of-the-art system on the HYRBIDIALOGUE dataset. cTBLS then uses a shared process between encoder and decoder models to perform a coarse+fine tabular knowledge (e.g., cell) ranking combined with a GPT-3.5 LLM response generator to yield a 2x relative improvement in ROUGE scores. Finally, human evaluators prefer cTBLs +80% of the time (coherency, fluency) and judge informativeness to be 4x better than the previous state-of-the-art.

1 Introduction

Equipping conversational AI with multimodal capabilities broadens the range of dialogues that humans have with such systems. A persisting challenge in multimodal conversational AI is the development of systems that produce conversationally coherent responses grounded in textual and non-textual modalities (Sundar and Heck, 2022).

It is well-established that large language models (LLMs) possess real-world knowledge stored within their parameters, as demonstrated by recent research (Roberts et al., 2020; Heinzerling and Inui, 2021). Nevertheless, the incorporation of conversation-specific extrinsic knowledge into these models to yield precise responses remains an active area of investigation. While humans can easily retrieve contextual information from tables by examining rows and columns, LLMs often struggle to identify relevant information amidst conversational distractions.

HYBRIDIALOGUE (Nakamura et al., 2022), a dataset of conversations grounded on structured and unstructured knowledge from tables and text, introduces the task of responding to messages by utilizing information from external knowledge and prior dialogue turns. The authors also present an approach and experimental results on HYBRIDIALOGUE that represents the current state-of-the-art (SoTA).

This paper proposes an extension to the SoTA approach of HYBRIDIALOGUE in the form of Conversational Tables (cTBLS) [1], a novel three-step encoder-decoder architecture designed to augment LLMs with tabular data in conversational settings. In the first step, cTBLS uses a dual-encoder Transformer-based (Vaswani et al., 2017) Dense Table Retriever (DTR) to retrieve the correct table from the entire corpus based on the user's query. The second step employs a fine-tuned dual-encoder Transformer to track system state and rank cells in the retrieved table according to their relevance to the conversation. Finally, cTBLS utilizes GPT-3.5 to generate a natural language response by prompting it with the ranked cells.

While previous research separated knowledge retrieval and response generation between encoder and decoder models, this paper demonstrates that LLM decoders can perform these tasks jointly when prompted with knowledge sources ranked by language model encoders. Furthermore, by pre-training the Dense Table Retriever to perform retrieval over a corpus of tables, cTBLS can be extended to new knowledge sources without re-training, by appending additional knowledge to the corpus.

Compared to the previous SoTA, experiments

[1] Our code will be available at https://github.com/avalab-gt/cTBLS

Proceedings of the 5th Workshop on NLP for Conversational AI (NLP4ConvAI 2023), pages 59–70

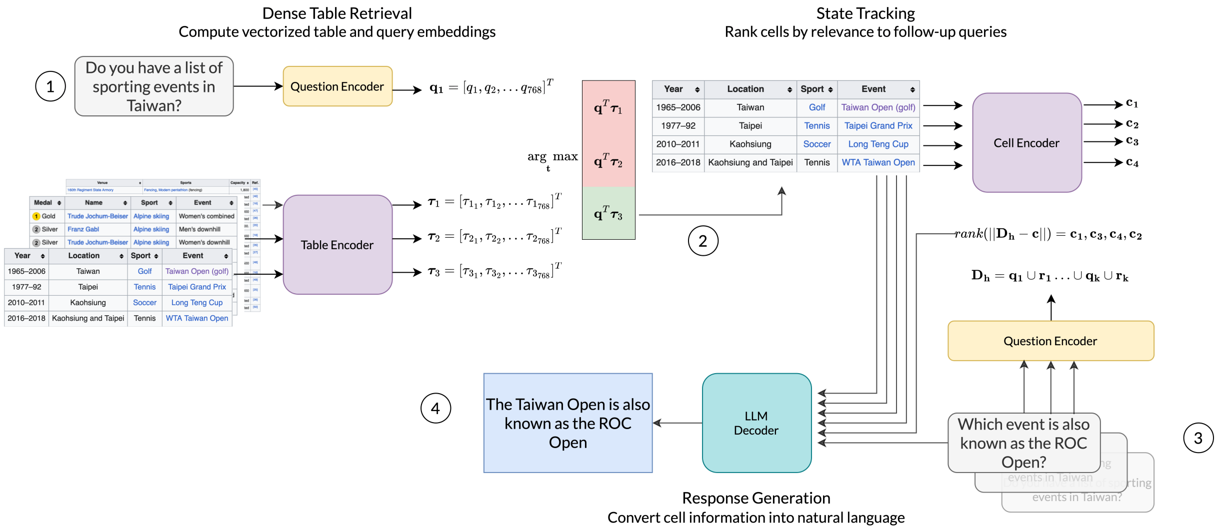


Figure 1: cTBLS for conversations on HYBRIDIALOGUE. Dense Table Retrieval identifies the table most relevant to the initial query. The retrieved table is provided to the state tracker for follow-up queries. State Tracking ranks cells in the table based on their ability to answer a follow-up query. Response Generation utilizes a LLM Decoder provided with the ranked cell information and the follow-up query to convert tabular data into a natural language response and continue the conversation. Details on individual components are provided in Section 3.

on cTBLS show up to 125% relative improvement in table retrieval and a 2x relative improvement in ROUGE scores. In addition, human evaluators prefer cTBLs +80% of the time (coherency, fluency) and judge informativeness to be 4x better than the previous SoTA.

Our contributions are as follows:

1. The introduction of Conversational Tables (cTBLS), a novel three-step encoder-decoder architecture designed to augment LLMs with tabular data in conversational settings.

2. Experimental results demonstrating that Dense Table Retrieval, which utilizes neural models fine-tuned with a summary of tabular information, outperforms sparse techniques based on keyword matching for table retrieval.

3. The presentation of evidence that augmenting state-of-the-art LLM decoders using knowledge sources ranked by encoder language models leads to better results on automatic (ROUGE-Precision) and human (Coherence, Fluency, and Informativeness) evaluation for knowledge-grounded response generation while limiting the number of API calls to these models.

This paper presents the cTBLS system and demonstrates its application to the HYBRIDIALOGUE dataset. In Section 2, we review the existing literature in the fields of Table Question Answering and Knowledge Grounded Response Generation. Section 3 describes the various components of cTBLS as presented in Figure 1. In Section 4, we evaluate the performance of cTBLS against previous methods for conversations over tables and report experimental results from automatic and human evaluations. Finally, Section 5 concludes the paper and outlines potential directions for future research.

2 Related Work

2.1 Table Question Answering

Table Question Answering is a well-researched precursor to conversations over tables. In WIKITABLEQUESTIONS, Pasupat and Liang (2015) transform HTML tables into a knowledge graph and retrieve the correct answer by converting natural language questions into graph queries. FRETS (Jauhar et al., 2016) uses a log-linear model conditioned on alignment scores between cells in tables and individual QA pairs in the training set. Cho et al. (2018) introduce NEOP, a multi-layer sequential network with attention supervision to answer queries conditioned on tables. Hannan et al. (2020) propose MANYMODALQA, which uses a modality selection network and pre-trained text-based QA, Table-based QA, and Image-based QA models to jointly answer questions over text, tables, and images. Chen et al. (2020c) present HYBRIDER, which performs multi-hop QA over tables using keyword-matching for cell linking followed by BERT (Devlin et al., 2019) for reasoning. Chen et al. (2020a) propose OTT-QA, which uses a fusion retriever to identify relevant tables and text and a cross-block reader based on a long-range Sparse Attention Transformer (Ainslie et al., 2020) to choose the correct answer. Heck and Heck (2020) perform multi-task fine-tuning of Transformer encoders by modeling slot filling as question answering over tabular and visual information in Visual Slot. Herzig et al. (2020) and Yin et al. (2020) extend BERT for Table Question Answering by pre-training a masked language model over text-table pairs in TAPAS and TaBERT, respectively. Recent work building off the Transformer architecture for Table Question Answering includes (Eisenschlos et al., 2021; Li et al., 2021; Herzig et al., 2021; Zayats et al., 2021; Zhao et al., 2022; Huang et al., 2022; Yang et al., 2022; Chen, 2022). Jin et al. (2022) provide a comprehensive survey of advancements in Table Question Answering.

2.2 Knowledge Grounded Response Generation

Early work related to grounding responses generated by language models in real-world knowledge was motivated by the need to improve prior information for open-domain dialogue (Heck et al., 2013; Hakkani-Tür et al., 2014; Hakkani-Tür et al., 2014; Huang et al., 2015; Jia et al., 2017). More recently, knowledge grounded response generation has been applied to mitigate the hallucination problem (Maynez et al., 2020; Shuster et al., 2021) in LLMs. RAG (Lewis et al., 2020) fine-tunes LLMs using Dense Passage Retrieval (Karpukhin et al., 2020) over a Wikipedia dump to ground responses for Open Domain Question Answering. KGPT (Chen et al., 2020b) and SKILL (Moiseev et al., 2022) pre-train a Transformer encoder (Vaswani et al., 2017) with English Wikidump for Natural Language Generation. Fusion-in-Decoder (Izacard and Grave, 2021) fine-tunes decoder models using evidence acquired through Dense Passage Retrieval.

Recent research also includes a dual-stage approach where LLMs generate knowledge sources based on prompts (Yu et al., 2022; Bonifacio et al., 2022; Jeronymo et al., 2023). Closest to our work, Wizard of Wikipedia (Dinan et al., 2018) jointly optimizes an encoder-decoder Transformer to produce dialogue responses conditioned on retrieved knowledge and dialogue context but does not extend their approach to the multiple modalities. REPLUG (Shi et al., 2023) ensembles output responses generated by prompting large language models with inputs from a dense retriever in a zero-shot setting. However, this requires multiple API calls to state-of-the-art LLMs. LLM-AUGMENTER (Peng et al., 2023) incorporates external knowledge in LLM responses by matching keywords in dialogue state to candidate knowledge sources obtained through web-search. A survey of knowledge fusion in LLMs is available in Colon-Hernandez et al. (2021) and Richardson and Heck (2023).

In contrast to prior research that focuses on either Table Question Answering or Knowledge Grounded Response Generation, our work, cTBLS, addresses the challenge of generating responses grounded on tabular knowledge. Moreover, while cTBLS is fine-tuned to retrieve tables and filter out incorrect references, it leverages the power of SoTA pre-trained LLMs for response generation. Furthermore, by fine-tuning open-source table and knowledge retrievers to remove inaccurate references, cTBLS reduces the number of API calls to the SoTA LLMs.

3 Method

The challenge of developing conversational systems grounded in tabular information consists of three tasks, namely table retrieval, system state tracking, and response generation. Table retrieval requires identifying the most relevant table in the dataset based on a given natural language query. System state tracking is responsible for ranking the cells in the table, enabling the system to provide responses to follow-up queries about the table. Finally, response generation involves converting the ranked cells into a natural language response.

3.1 Table Retrieval

Table retrieval is a prerequisite to answering queries when the exact table to converse over is unspecified. The objective is to identify the correct table from a vast corpus. cTBLS proposes formulating table retrieval as document retrieval by assigning a relevance score to each table based on its relevance to the natural language query. Inspired by Karpukhin et al. (2020) and Huang et al. (2013), cTBLS uses a dual-encoder-based Dense Table Retrieval (DTR) model. The DTR model pre-computes a vectorized embedding of all tables in the corpus. Given a

Figure 2: An example of table-associated text in the context of Wikipedia, where the input to the DTR text-encoder includes the page title, the introduction to the article, the section title, and the introduction paragraph.

query at inference, the retrieved table is closest to the query in the embedded space, indicated by the upper-left portion of Figure 1.

The DTR model consists of a table encoder and a question encoder, initialized from RoBERTa-base (Liu et al., 2019). The input to the table encoder comprises the table's title and, if available, textual information associated with the table. Figure 2 presents an example of table-associated text in the context of Wikipedia, where introductions from the page and section provide additional grounding. The input to the question encoder is the current query to be answered. Taking the average over the sequence of the last hidden state at the table and question encoder results in 768-dimensional embeddings of the table information and the query.

The DTR model is optimized through a contrastive prediction task, which aims to maximize the similarity between embeddings of a given query q and the table to be retrieved τ while minimizing the similarity to other incorrect tables τ_{n_i} for $i = 1, \ldots, N$. As per (Karpukhin et al., 2020), normalized embedding vectors are utilized to optimize the objective in Equation 1:

$$\arg\min_{\tau} \left(-\log \frac{e^{q \cdot \tau}}{e^{q \cdot \tau} + \sum_{i=1}^{N} e^{q \cdot \tau_{n_i}}} \right) \quad (1)$$

Given a batch B of d-dimensional query embeddings $\mathbf{Q}$ and table embeddings $\mathbf{T}$, the DTR model computes the similarity $\mathbf{Q}\mathbf{T}^T (\in \mathbb{R}^{B \times B})$ between every query and table in the batch. This similarity computation enables the sampling of negatives from other query-table pairs, resulting in B^2 training samples in each batch, consisting of B positive pairs along the diagonal and $B^2 - B$ negatives.

3.2 Coarse System State Tracking

Given a table, system state tracking involves ranking cells in the table by their relevance to conversational queries. In contrast to quesiton-answering, conversational queries require leveraging information from external modalities in conjunction with prior dialogue turns to generate coherent responses (Sundar and Heck, 2022). cTBLS addresses system state tracking through two sub-tasks - coarse and fine system state tracking. Coarse system state tracking ranks cells in the table, while fine system state tracking identifies fine-grained information in the most relevant cell to answer the query.

cTBLS uses a RoBERTa-base dual-encoder architecture for coarse system state tracking. The cell encoder embeds all cells and associated hyperlinked information, and the question encoder generates embeddings for the dialogue history ($\mathbf{D_h}$) that includes the current turn's query as well as previous queries and responses.

To rank cells based on their relevance to the follow-up query, as illustrated in the upper-right section of Figure 1, the question and cell encoders are optimized using a triplet loss configuration. This optimization aims to minimize the distance between the anchor $\mathbf{D_h}$ and the positive cell c, while pushing the negative cell $\bar{c}$ further away from $\mathbf{D_h}$ by a margin m (Equation 2).

$$\arg\min_{c_i} \left(\max\{d(\mathbf{D_h}, c) - d(\mathbf{D_h}, \bar{c}) + m, 0\} \right) \quad (2)$$

$$d(x, y) = ||x - y||_2 \quad (3)$$

For our approach, we utilize an anchor-positive-negative triplet consisting of the complete dialogue history (including queries and responses from previous turns) concatenated with the current query as the anchor, the correct cell as the positive, and other cells from the same table that are not relevant to the query as negatives. We measure the distance between the anchor and the positive and between the anchor and the negatives using the 2-norm distance function $d(\cdot)$.

3.3 Fine System State Tracking and Response Generation

In contrast to coarse system state tracking, fine system state tracking involves identifying the exact phrase that answers the query from a ranked subset. The extracted phrase is converted into a natural language response that is coherent within the context of the conversation.

cTBLS employs GPT-3.5 (Brown et al., 2020) to perform fine system state tracking and response generation jointly. GPT-3.5 is prompted to generate a natural language response to a follow-up query conditioned on cells of the table ranked by their relevance to the query as obtained from the coarse state tracker. The prompt includes the dialogue history, ranked knowledge sources, and the query to be answered. The bottom-right section of Figure 1 outlines this process.

4 Experiments

4.1 HybriDialogue

The HybriDialogue dataset (Nakamura et al., 2022) comprises 4800 natural language conversations grounded in text and tabular information from Wikipedia. Crowdsourced workers break down multi-hop questions from the OTT-QA dataset (Chen et al., 2020a) into natural questions and conversational responses related to tabular data. On average, dialogues in the dataset consist of 4-5 conversation turns, with a total of 21,070 turns available in the dataset. Examples of conversations can be found in Figures 3 and 4.

4.2 Table Retrieval

The first conversation turn of HybriDialogue requires selecting the correct table based on the input query for which we use the Dense Table Retriever outlined in Section 3.1. The Dense Table Retriever is fine-tuned for 20 epochs using Adam (Kingma and Ba, 2014) with a learning rate of 1e-6 and a linear learning schedule with five warmup steps. The loss function is a modification of the contrastive loss implementation from ConVIRT (Zhang et al., 2022), with image embeddings replaced by table embeddings. The table retriever used in the HybriDialogue paper (Nakamura et al., 2022) was the BM25Okapi Retriever (Trotman et al., 2014) from rank-bm25. According to the results presented in Table 1, cTBLS-DTR outperforms BM25 in terms of Mean Reciprocal Rank (MRR), Top-1 Accuracy, and Top-3 Accuracy on HybriDialogue.

	MRR @10	Top 1 Acc	Top 3 Acc
BM25	0.491	0.345	0.460
cTBLS-DTR	**0.846**	**0.777**	**0.901**

Table 1: BM25 vs cTBLS-DTR for retrieval on first turn of conversation, results on HybriDialogue testing dataset. cTBLS-DTR obtains up to 125% relative improvement over sparse table retrieval

4.3 Coarse State Tracking

Coarse state tracking ranks cells from a table based on their relevance to a query. As before, the dual-encoder coarse state tracker of cTBLS consists of RoBERTa-base fine-tuned using Adam with a learning rate of 1e-6 and a linear learning schedule with five warmup steps. In contrast to table retrieval, the state tracker uses triplet margin loss with a margin of 1.0 (Equation 2) instead of contrastive loss (Equation 1). The results, as demonstrated in Table 2, show that fine-tuning RoBERTa-base solely on HybriDialogue surpasses the performance of SentenceBERT (Reimers and Gurevych, 2019). Furthermore, it nearly attains the same MRR @10 as TaPas (Herzig et al., 2020), even without additional table pre-training on the SQA dataset (Iyyer et al., 2017).

	MRR@10
SentenceBERT (Reimers and Gurevych, 2019)	0.603
TaPas (Herzig et al., 2020)	**0.689**
cTBLS - RoBERTa-base	0.683

Table 2: System state tracking results on HybriDialogue. cTBLS achieves nearly the same Mean Reciprocal Rank (MRR) @ 10 as TaPaS, without additional table pre-training on SQA (Iyyer et al., 2017)

4.4 Fine State Tracking and Response Generation

cTBLS uses GPT-3.5 (text-davinci-003) with the existing dialogue context, the current query, and the retrieved references from coarse state tracking to obtain a natural language response. Since fine-tuning the best available version of the model is cost prohibitive, we opt to prompt GPT-3.5 to generate responses instead.

	Top-1	Top-3	Top-10
cTBLS - RoBERTa-base	0.559	0.778	0.925

Table 3: Top-k accuracy for cTBLS on coarse system state tracking. cTBLS ranks the correct cell as the top reference in 56% of follow-up queries on HybriDialogue. The correct cell is ranked in the Top-3 and Top-10 retrievals in approximately 78% and 93% of conversations, respectively.

Model	TR	KR	RG	ROUGE-1	ROUGE-2	ROUGE-L
-	BM25	Top-1	DialoGPT	0.207	0.042	0.181
-	BM25	Top-3	DialoGPT	0.212	0.045	0.186
-	BM25	Top-1	GPT3.5	0.428	0.207	0.369
-	BM25	Top-3	GPT3.5	0.475	0.242	0.413
-	DTR	Top-1	DialoGPT	0.222	0.051	0.195
-	DTR	Top-3	DialoGPT	0.226	0.059	0.199
-	DTR	Top-1	GPT3.5	0.494	0.255	0.424
-	DTR	Top-3	GPT3.5	0.560	0.295	0.479
HYBRIDIALOGUE	Gold	Top-1	DialoGPT	0.438	0.212	0.375
cTBLS NoK	Gold	-	GPT3.5	0.487	0.229	0.422
cTBLS Top-1	Gold	Top-1	GPT3.5	0.603	0.304	0.517
cTBLS Top-3	Gold	Top-3	GPT3.5	**0.642**	**0.322**	**0.548**

Table 4: Ablation study on automatic evaluation metrics ROUGE-1, ROUGE-2, and ROUGE-L Precision. Using Dense Table Retrieval (DTR) improves results over BM25 across Top-1 and Top-3 knowledge for DialoGPT and GPT3.5. Furthermore, using Top-3 knowledge sources results in better results than using only Top-1 knowledge sources for DialoGPT and GPT3.5 using both table retrieval methods. cTBLS No Knowledge (NoK), Top-1 Knowledge, Top-3 Knowledge, and HYBRIDIALOGUE use ground truth table retrieval. cTBLS exhibits a 2x relative improvement in ROUGE Precision over HYBRIDIALOGUE. TR: Table Retrieval, KR: Knowledge Retrieval, RG: Response Generation

The results presented in Table 3 demonstrate that the coarse state tracker successfully retrieves the correct cell in approximately 56% of conversations during inference. Furthermore, it achieves Top-3 and Top-10 retrievals in approximately 78% and 93% of conversations, respectively. Motivated by these results, the fine state tracker of cTBLS is evaluated in two different configurations by prompting GPT-3.5 augmented with the Top-1 and Top-3 knowledge references (cTBLS Top-1 and cTBLS Top-3). Due to limits on token length associated with the OpenAI API, we remove stopwords from the knowledge provided in the prompt and do not experiment with Top-10 knowledge augmentation.

Since LLMs store factual information in their weights (Roberts et al., 2020; Heinzerling and Inui, 2021), we compare to few-shot prompting (using two examples) with no knowledge sources (cTBLS-NoK). Furthermore, to enable a meaningful comparison with existing research (Nakamura et al., 2022), we measure cTBLS against the system proposed by HYBRIDIALOGUE that utilizes a fine-tuned DialoGPT-medium (Zhang et al., 2019) model augmented with Top-1 knowledge.

Table 4 presents ROUGE-1, ROUGE-2, and ROUGE-L precision (Lin, 2004) for all models assessed. The results demonstrate that superior downstream performance can be achieved through improvements in table retrieval. Specifically, when keeping the number of knowledge sources constant, we observe an improvement in ROUGE precision scores when transitioning from BM25 to DTR, and from DTR to gold table retrieval. The inclusion of additional knowledge sources leads to an improved n-gram overlap with the ground truth reference, as evidenced by the Top-3 knowledge augmented models outperforming their Top-1 counterparts utilizing the same table retriever, and cTBLS Top-1 outperforming the baseline model cTBLS NoK. Moreover, cTBLS Top-3 achieves the best performance across all automatic metrics, suggesting the benefits of splitting knowledge retrieval into coarse and fine state tracking, and utilizing additional knowledge sources. Finally, all three configurations of cTBLS demonstrate superior performance to HYBRIDIALOGUE.

4.5 Human Evaluation

To gain a deeper understanding of cTBLS, we conducted human evaluation using the metrics outlined by Nakamura et al. (2022), namely Coherence, Fluency, and Informativeness. For the evaluation of these metrics, we enlisted crowd workers from Amazon Mechanical Turk (AMT) to assess 50% of the test data. The evaluation process involved a comparison between the responses generated by HYBRIDIALOGUE and cTBLS Top-3.

	cTBLS Top-3 vs HYBRIDIALOGUE
Coherence	0.842
Fluency	0.827

Table 5: Coherence and Fluency - cTBLS Top-3 is more conversationally coherent than the best performing HYBRIDIALOGUE system 84.2% of the time and is more fluent 82.7% of the time.

	Informativeness
HYBRIDIALOGUE	0.124
cTBLS - NoK	0.306
cTBLS Top-1	0.456
cTBLS Top-3	**0.500**

Table 6: Human Evaluation Metrics - Fraction of cases where model response is semantically equivalent to ground truth response. Using more knowledge sources results in responses that are more informative, helping reduce hallucination.

In accordance with the methodology delineated in Nakamura et al. (2022), Coherence was defined as the degree to which a response continued the conversation in a logically coherent manner based on prior context. Fluency, conversely, was determined by evaluating absence of grammatical and spelling errors, and appropriate use of parts of speech.

To ensure the quality of the evaluated responses, we engaged crowd workers possessing a Masters qualification on AMT and originating from English-speaking countries (USA, Canada, Australia, New Zealand, or Great Britain). Each task required approximately 30 seconds to complete, and workers were remunerated at a rate of $0.05 per task. Moreover, to minimize bias and guarantee the dependability of the evaluations, we assigned two crowd workers to assess each response, with a response deemed more coherent or fluent only if both evaluations concurred.

The results presented in Table 5 reveal that the responses generated by cTBLS Top-3 were more coherent than those produced by HYBRIDIALOGUE in 84.2% of cases and exhibited greater fluency 82.7% of the time, suggesting that improvements in table retrieval, knowledge retrieval, and response generation lead to better downstream performance.

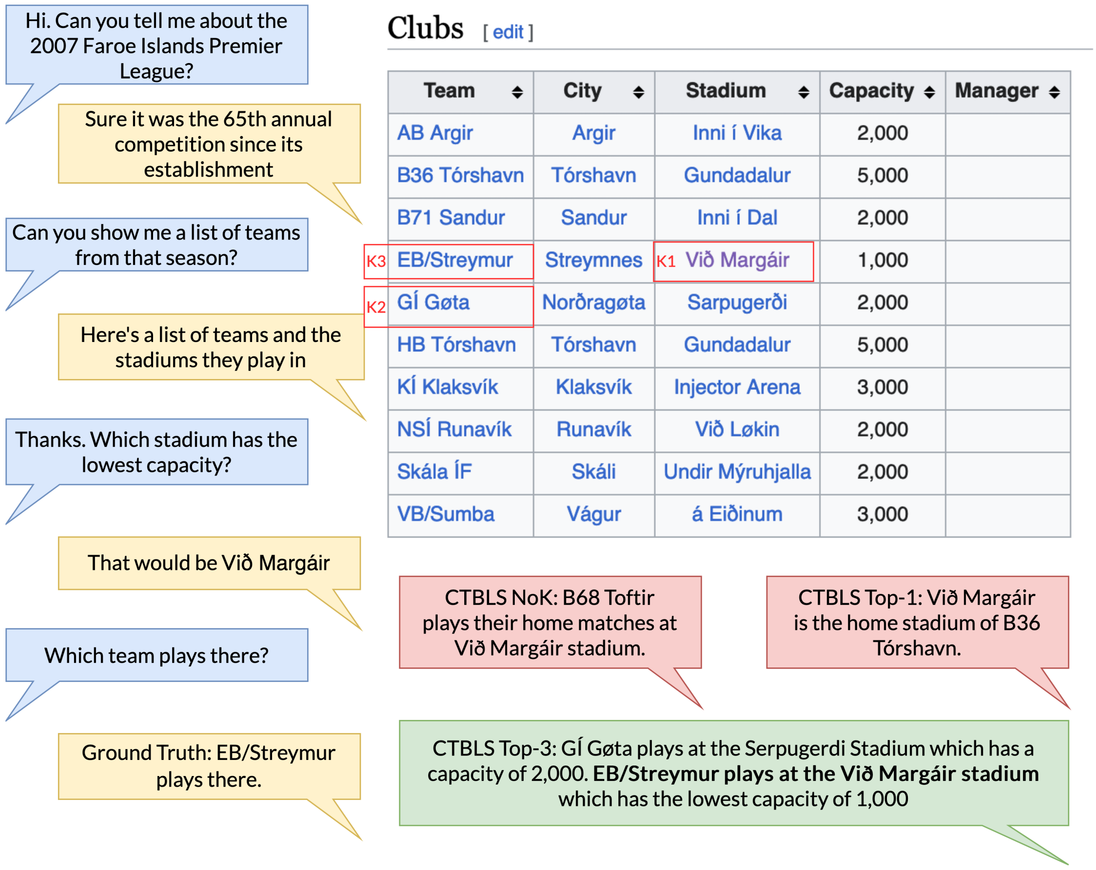
Informativeness represents the accuracy of machine-generated responses when compared to the ground-truth (Nakamura et al., 2022) and serves as a measure of hallucination in LLMs. Hallucinated responses tend to be less informative, deviating significantly from the ground-truth.

To evaluate informativeness, crowd workers determined whether generated responses were semantically equivalent to the ground truth response. Each response was assessed by two Turkers, and a response was deemed more informative only if there was inter-annotator agreement. The absence of illustrative examples in the prompting process resulted in responses generated by cTBLS Top-1 and cTBLS Top-3 being longer than the ground truth response. Consequently, the knowledge-augmented cTBLS responses were considered informative if all the information provided in the ground truth was encapsulated in the model response, even if cTBLS included supplementary information.

The data in Table 6 indicate that cTBLS Top-3 encompasses the same information as the ground truth response 50% of the time, a higher rate than cTBLS Top-1 at 45.6%, exemplifying the benefits of partitioning retrieval into coarse and fine state tracking and augmenting with additional knowledge. Based on these findings, we hypothesize that the attention mechanism in decoder models facilitates additional knowledge retrieval. cTBLS NoK generates the correct response 30.6% of the time, suggesting that HYBRIDIALOGUE comprises questions and answers predicated on general world knowledge embedded in the weights of LLMs. Responses produced by HYBRIDIALOGUE are informative in merely 12.4% of instances.

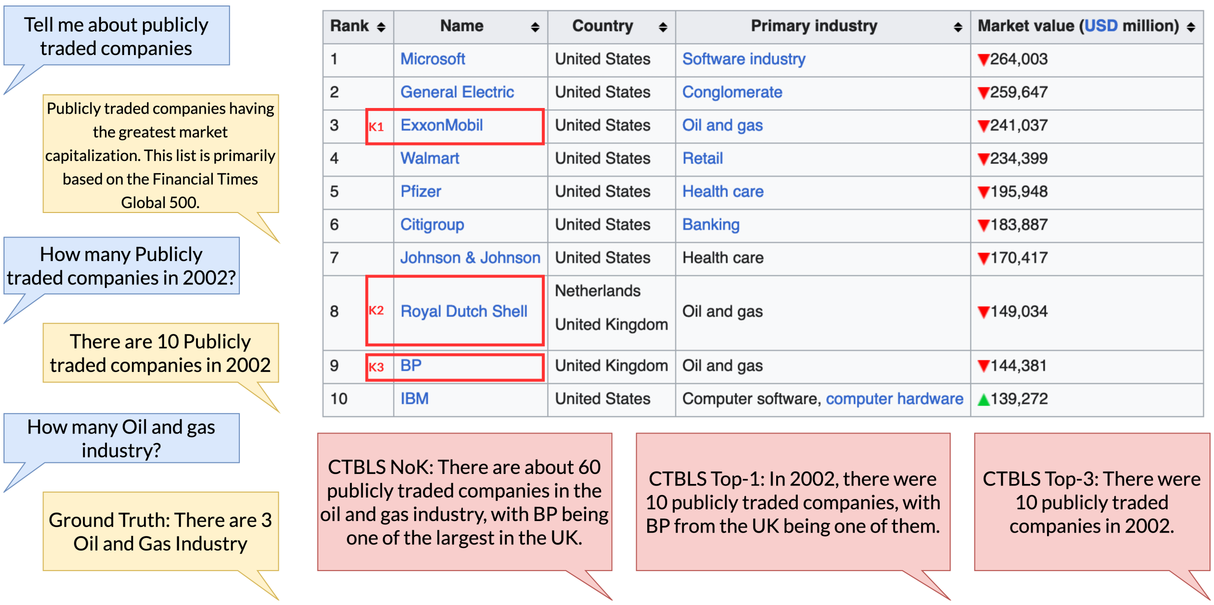
Figure 3 presents a comparison of responses generated by various configurations of cTBLS on the HYBRIDIALOGUE dataset. The entire dialogue history constitutes the context and is depicted as an exchange between the user (in blue) and the system (in yellow). The final question box represents the follow-up query to be addressed, while the last answer chat box indicates the ground truth response. Knowledge K1, K2, and K3 correspond to cells of the table retrieved during state tracking, based on which responses are produced. cTBLS NoK generates a response solely relying on the context, cTBLS Top-1 formulates a response conditioned on K1, and cTBLS Top-3 devises a response based on K1, K2, and K3.

cTBLS NoK creates a hallucinated response, answering with the random Faroese club B68 Toftir. Similarly, cTBLS Top-1 hallucinates a response, opting for B36 Tórshavn, as K1 refers to the stadium Viò Margáir rather than the correct club's



Team	City	Stadium	Capacity	Manager
AB Argir	Argir	Inni í Vika	2,000	
B36 Tórshavn	Tórshavn	Gundadalur	5,000	
B71 Sandur	Sandur	Inni í Dal	2,000	
EB/Streymur	Streymnes	Við Margáir	1,000	
GÍ Gøta	Norðragøta	Sarpugerði	2,000	
HB Tórshavn	Tórshavn	Gundadalur	5,000	
KÍ Klaksvík	Klaksvík	Injector Arena	3,000	
NSÍ Runavík	Runavík	Við Løkin	2,000	
Skála ÍF	Skáli	Undir Mýruhjalla	2,000	
VB/Sumba	Vágur	á Eiðinum	3,000	

Figure 3: Generated responses vs Ground Truth on HybriDialogue test set. Questions are in blue and responses in yellow. K1, K2, and K3 represent the Top 3 knowledge sources ranked by relevance to the query "Which team plays there?". cTBLS Top-3 is able to leverage K3 to generate the correct response while cTBLS NoK hallucinates a response and cTBLS Top-1 generates an incorrect response based on K1. Table obtained from Wikipedia available here



Rank	Name	Country	Primary industry	Market value (USD million)
1	Microsoft	United States	Software industry	▼264,003
2	General Electric	United States	Conglomerate	▼259,647
3	ExxonMobil	United States	Oil and gas	▼241,037
4	Walmart	United States	Retail	▼234,399
5	Pfizer	United States	Health care	▼195,948
6	Citigroup	United States	Banking	▼183,887
7	Johnson & Johnson	United States	Health care	▼170,417
8	Royal Dutch Shell	Netherlands United Kingdom	Oil and gas	▼149,034
9	BP	United Kingdom	Oil and gas	▼144,381
10	IBM	United States	Computer software, computer hardware	▲139,272

Figure 4: Generated responses vs Ground Truth on HybriDialogue test set. Despite selecting the rows of the table corresponding to Oil and gas industries, cTBLS NoK, Top-1, and Top-3 struggle with counting and hallucinate a response. Table obtained from Wikipedia available here

name. In contrast, cTBLS Top-3 produces the accurate response, EB/Streymur, since K3 contains the necessary information. This example demonstrates the benefits of augmenting response generation with additional pertinent knowledge, which aids in mitigating the hallucination problem (Maynez et al., 2020).

5 Conclusion

In this paper, we introduce Conversational Tables (cTBLS), a system designed to address multi-turn dialogues that are grounded in tabular data. cTBLS separates tabular dialogue into three distinct tasks, specifically table retrieval, system state tracking, and response generation. The dense table retrieval system of cTBLS yields an enhancement of up to 125% relative to keyword-matching based techniques on the HYBRIDIALOGUE dataset, with regard to Top-1 Accuracy and Mean Reciprocal Rank @ 10. Furthermore, cTBLS conducts system state tracking utilizing a two-step process shared between encoder and decoder models. This methodology results in natural language responses exhibiting a 2x relative improvement in ROUGE scores. Human evaluators favor cTBLS +80% of the time (coherency and fluency) and judge informativeness to be 4x better than the previous state-of-the-art.

6 Limitations

Although cTBLS enhances LLMs with tabular knowledge to generate grounded responses, certain limitations remain to be addressed.

Firstly, the efficacy of cTBLS is constrained by the total number of knowledge sources employed during the augmentation process. Token length restrictions in the OpenAI API limit the knowledge augmentation to the top three cells of the table. Another limitation is the incapacity of cTBLS to handle queries pertaining to the entire table. Figure 4 demonstrates one such instance in which the state tracker module accurately retrieves three rows of the table corresponding to oil and gas industries, yet the response generation module fails to utilize this information when transforming the retrieved state into a response. Generally, cTBLS encounters difficulties with counting, comparing the values of cells, and other mathematical operations, an issue we aim to address in future research.

7 Acknowledgements

We would like to thank Christopher Richardson, Benjamin Z Reichman, Atishay Jain, and Srikar Bhumireddy for their contributions. We would also like to thank the review committee for their feedback. This work was supported by NSF IIS-2112633 and by CoCoSys, one of seven centers in JUMP 2.0, a Semiconductor Research Corporation (SRC) program sponsored by DARPA.

References

Joshua Ainslie, Santiago Ontanon, Chris Alberti, Vaclav Cvicek, Zachary Fisher, Philip Pham, Anirudh Ravula, Sumit Sanghai, Qifan Wang, and Li Yang. 2020. ETC: Encoding long and structured inputs in transformers. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 268–284, Online. Association for Computational Linguistics.

Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, and Rodrigo Nogueira. 2022. Inpars: Data augmentation for information retrieval using large language models. *arXiv preprint arXiv:2202.05144*.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Wenhu Chen. 2022. Large language models are few (1)-shot table reasoners. *arXiv preprint arXiv:2210.06710*.

Wenhu Chen, Ming-Wei Chang, Eva Schlinger, William Wang, and William W Cohen. 2020a. Open question answering over tables and text. *arXiv preprint arXiv:2010.10439*.

Wenhu Chen, Yu Su, Xifeng Yan, and William Yang Wang. 2020b. Kgpt: Knowledge-grounded pre-training for data-to-text generation. *arXiv preprint arXiv:2010.02307*.

Wenhu Chen, Hanwen Zha, Zhiyu Chen, Wenhan Xiong, Hong Wang, and William Yang Wang. 2020c. HybridQA: A dataset of multi-hop question answering over tabular and textual data. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1026–1036, Online. Association for Computational Linguistics.

Minseok Cho, Reinald Kim Amplayo, Seung-won Hwang, and Jonghyuck Park. 2018. Adversarial tableqa: Attention supervision for question answering on tables. In *Asian Conference on Machine Learning*, pages 391–406. PMLR.

Pedro Colon-Hernandez, Catherine Havasi, Jason Alonso, Matthew Huggins, and Cynthia Breazeal. 2021. Combining pre-trained language models and structured knowledge. *arXiv preprint arXiv:2101.12294*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2018. Wizard of wikipedia: Knowledge-powered conversational agents. *arXiv preprint arXiv:1811.01241*.

Julian Martin Eisenschlos, Maharshi Gor, Thomas Müller, and William W Cohen. 2021. Mate: multi-view attention for table transformer efficiency. *arXiv preprint arXiv:2109.04312*.

Dilek Hakkani-Tür, Malcolm Slaney, Asli Celikyilmaz, and Larry Heck. 2014. Eye gaze for spoken language understanding in multi-modal conversational interactions. In *Proceedings of the 16th International Conference on Multimodal Interaction*, ICMI '14, page 263–266, New York, NY, USA. Association for Computing Machinery.

Dilek Hakkani-Tür, Asli Celikyilmaz, Larry Heck, Gokhan Tur, and Geoff Zweig. 2014. Probabilistic enrichment of knowledge graph entities for relation detection in conversational understanding. In *Proceedings of Interspeech*.

Darryl Hannan, Akshay Jain, and Mohit Bansal. 2020. Manymodalqa: Modality disambiguation and qa over diverse inputs. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 7879–7886.

Larry Heck, Dilek Hakkani-Tür, Madhu Chinthakunta, Gokhan Tur, Rukmini Iyer, Partha Parthasacarthy, Lisa Stifelman, Elizabeth Shriberg, and Ashley Fidler. 2013. Multimodal conversational search and browse. *First Workshop on Speech, Language and Audio in Multimedia Marseille, France*.

Larry Heck and Simon Heck. 2020. Zero-shot visual slot filling as question answering. *arXiv preprint arXiv:2011.12340*.

Benjamin Heinzerling and Kentaro Inui. 2021. Language models as knowledge bases: On entity representations, storage capacity, and paraphrased queries. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1772–1791, Online. Association for Computational Linguistics.

Jonathan Herzig, Thomas Müller, Syrine Krichene, and Julian Eisenschlos. 2021. Open domain question answering over tables via dense retrieval. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 512–519, Online. Association for Computational Linguistics.

Jonathan Herzig, Pawel Krzysztof Nowak, Thomas Müller, Francesco Piccinno, and Julian Eisenschlos. 2020. TaPas: Weakly supervised table parsing via pre-training. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4320–4333, Online. Association for Computational Linguistics.

Hongzhao Huang, Larry Heck, and Heng Ji. 2015. Leveraging deep neural networks and knowledge graphs for entity disambiguation. *arXiv preprint arXiv:1504.07678*.

Junjie Huang, Wanjun Zhong, Qian Liu, Ming Gong, Daxin Jiang, and Nan Duan. 2022. Mixed-modality representation learning and pre-training for joint table-and-text retrieval in openqa. *arXiv preprint arXiv:2210.05197*.

Po-Sen Huang, Xiaodong He, Jianfeng Gao, Li Deng, Alex Acero, and Larry Heck. 2013. Learning deep structured semantic models for web search using clickthrough data. In *Proceedings of the 22nd ACM international conference on Information & Knowledge Management*, pages 2333–2338. ACM.

Mohit Iyyer, Wen-tau Yih, and Ming-Wei Chang. 2017. Search-based neural structured learning for sequential question answering. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1821–1831, Vancouver, Canada. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Sujay Kumar Jauhar, Peter Turney, and Eduard Hovy. 2016. Tables as semi-structured knowledge for question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 474–483, Berlin, Germany. Association for Computational Linguistics.

Vitor Jeronymo, Luiz Bonifacio, Hugo Abonizio, Marzieh Fadaee, Roberto Lotufo, Jakub Zavrel, and Rodrigo Nogueira. 2023. Inpars-v2: Large language models as efficient dataset generators for information retrieval. *arXiv preprint arXiv:2301.01820*.

Robin Jia, Larry Heck, Dilek Hakkani-Tür, and Georgi Nikolov. 2017. Learning concepts through conversations in spoken dialogue systems. In *2017 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 5725–5729.

Nengzheng Jin, Joanna Siebert, Dongfang Li, and Qingcai Chen. 2022. A survey on table question answering: Recent advances. In *Knowledge Graph and Semantic Computing: Knowledge Graph Empowers the Digital Economy*, pages 174–186, Singapore. Springer Nature Singapore.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6769–6781, Online. Association for Computational Linguistics.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Advances in Neural Information Processing Systems*, volume 33, pages 9459–9474. Curran Associates, Inc.

Alexander Hanbo Li, Patrick Ng, Peng Xu, Henghui Zhu, Zhiguo Wang, and Bing Xiang. 2021. Dual reader-parser on hybrid textual and tabular evidence for open domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 4078–4088, Online. Association for Computational Linguistics.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1906–1919, Online. Association for Computational Linguistics.

Fedor Moiseev, Zhe Dong, Enrique Alfonseca, and Martin Jaggi. 2022. SKILL: Structured knowledge infusion for large language models. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1581–1588, Seattle, United States. Association for Computational Linguistics.

Kai Nakamura, Sharon Levy, Yi-Lin Tuan, Wenhu Chen, and William Yang Wang. 2022. HybriDialogue: An information-seeking dialogue dataset grounded on tabular and textual data. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 481–492, Dublin, Ireland. Association for Computational Linguistics.

Panupong Pasupat and Percy Liang. 2015. Compositional semantic parsing on semi-structured tables. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1470–1480, Beijing, China. Association for Computational Linguistics.

Baolin Peng, Michel Galley, Pengcheng He, Hao Cheng, Yujia Xie, Yu Hu, Qiuyuan Huang, Lars Liden, Zhou Yu, Weizhu Chen, and Jianfeng Gao. 2023. Check your facts and try again: Improving large language models with external knowledge and automated feedback.

Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992, Hong Kong, China. Association for Computational Linguistics.

Christopher Richardson and Larry Heck. 2023. Commonsense reasoning for conversational ai: A survey of the state of the art. *Workshop on Knowledge Augmented Methods for NLP, (KnowledgeNLP-AAAI'23)*.

Adam Roberts, Colin Raffel, and Noam Shazeer. 2020. How much knowledge can you pack into the parameters of a language model? In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5418–5426, Online. Association for Computational Linguistics.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803, Punta Cana, Dominican Republic. Association for Computational Linguistics.

Anirudh Sundar and Larry Heck. 2022. Multimodal conversational AI: A survey of datasets and approaches. In *Proceedings of the 4th Workshop on NLP for Conversational AI*, pages 131–147, Dublin, Ireland. Association for Computational Linguistics.

Andrew Trotman, Antti Puurula, and Blake Burgess. 2014. Improvements to bm25 and language models examined. In *Proceedings of the 2014 Australasian Document Computing Symposium*, ADCS '14, page 58–65, New York, NY, USA. Association for Computing Machinery.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Advances in neural information processing systems*, 30.

Jingfeng Yang, Aditya Gupta, Shyam Upadhyay, Luheng He, Rahul Goel, and Shachi Paul. 2022. Tableformer: Robust transformer modeling for table-text encoding. *arXiv preprint arXiv:2203.00274*.

Pengcheng Yin, Graham Neubig, Wen-tau Yih, and Sebastian Riedel. 2020. TaBERT: Pretraining for joint understanding of textual and tabular data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8413–8426, Online. Association for Computational Linguistics.

Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang Zhu, Michael Zeng, and Meng Jiang. 2022. Generate rather than retrieve: Large language models are strong context generators. *arXiv preprint arXiv:2209.10063*.

Vicky Zayats, Kristina Toutanova, and Mari Ostendorf. 2021. Representations for question answering from documents with tables and text. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2895–2906, Online. Association for Computational Linguistics.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, JJ (Jingjing) Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. In *arXiv:1911.00536*.

Yuhao Zhang, Hang Jiang, Yasuhide Miura, Christopher D Manning, and Curtis P Langlotz. 2022. Contrastive learning of medical visual representations from paired images and text. In *Machine Learning for Healthcare Conference*, pages 2–25. PMLR.

Yilun Zhao, Yunxiang Li, Chenying Li, and Rui Zhang. 2022. MultiHiertt: Numerical reasoning over multi hierarchical tabular and textual data. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6588–6600, Dublin, Ireland. Association for Computational Linguistics.