Multi-hop Question Answering

Other titles in Foundations and Trends® in Information Retrieval

Multi-hop Question Answering
Vaibhav Mavi, Anubhav Jangra and Adam Jatowt
ISBN: 978-1-63828-374-4

Conversational Information Seeking
Hamed Zamani, Johanne R. Trippas, Jeff Dalton and Filip Radlinski
ISBN: 978-1-63828-200-6

Perspectives of Neurodiverse Participants in Interactive Information Retrieval
Laurianne Sitbon, Gerd Berget and Margot Brereton
ISBN: 978-1-63828-202-0

Efficient and Effective Tree-based and Neural Learning to Rank
Sebastian Bruch, Claudio Lucchese and Franco Maria Nardini
ISBN: 978-1-63828-198-6

Multi-hop Question Answering

Vaibhav Mavi
New York University
vaibhavg152@gmail.com

Anubhav Jangra
Indian Institute of Technology Patna
anubhav0603@gmail.com

Adam Jatowt
University of Innsbruck
jatowt@acm.org

the essence of knowledge
Boston — Delft

Foundations and Trends® in Information Retrieval

Published, sold and distributed by:
now Publishers Inc.
PO Box 1024
Hanover, MA 02339
United States
Tel. +1-781-985-4510
www.nowpublishers.com
sales@nowpublishers.com

Outside North America:
now Publishers Inc.
PO Box 179
2600 AD Delft
The Netherlands
Tel. +31-6-51115274

The preferred citation for this publication is

V. Mavi *et al.*. *Multi-hop Question Answering*. Foundations and Trends® in Information Retrieval, vol. 17, no. 5, pp. 457–586, 2024.

ISBN: 978-1-63828-375-1

Editorial Scope

Foundations and Trends® in Information Retrieval publishes survey and tutorial articles in the following topics:

- Applications of IR
- Architectures for IR
- Collaborative filtering and recommender systems
- Cross-lingual and multilingual IR
- Distributed IR and federated search
- Evaluation issues and test collections for IR
- Formal models and language models for IR
- IR on mobile platforms
- Indexing and retrieval of structured documents
- Information categorization and clustering
- Information extraction
- Information filtering and routing
- Metasearch, rank aggregation and data fusion
- Natural language processing for IR
- Performance issues for IR systems, including algorithms, data structures, optimization techniques, and scalability
- Question answering
- Summarization of single documents, multiple documents, and corpora
- Text mining
- Topic detection and tracking
- Usability, interactivity, and visualization issues in IR
- User modelling and user studies for IR
- Web search

Information for Librarians

Foundations and Trends® in Information Retrieval, 2024, Volume 17, 5 issues. ISSN paper version 1554-0669. ISSN online version 1554-0677. Also available as a combined paper and online subscription.

Contents

Multi-hop Question Answering

Vaibhav Mavi[1], Anubhav Jangra[2] and Adam Jatowt[3]

[1] *New York University, USA; vaibhavg152@gmail.com*
[2] *Indian Institute of Technology Patna, India; anubhav0603@gmail.com*
[3] *University of Innsbruck, Austria; jatowt@acm.org*

ABSTRACT

The task of Question Answering (QA) has attracted significant research interest for a long time. Its relevance to language understanding and knowledge retrieval tasks, along with the simple setting, makes the task of QA crucial for strong AI systems. Recent success on simple QA tasks has shifted the focus to more complex settings. Among these, Multi-Hop QA (MHQA) is one of the most researched tasks over recent years. In broad terms, MHQA is the task of answering natural language questions that involve extracting and combining multiple pieces of information and doing multiple steps of reasoning. An example of a multi-hop question would be *"The Argentine PGA Championship record holder has won how many tournaments worldwide?"*. Answering the question would need two pieces of information: *"Who is the record holder for Argentine PGA Championship tournaments?"* and *"How many tournaments did [Answer of Sub Q1] win?"*. The ability to answer multi-hop questions and perform multi step reasoning can significantly improve the utility of NLP systems. Consequently, the field has seen a surge of high quality datasets, models and evaluation strategies. The notion of 'multiple hops' is somewhat abstract

---

Vaibhav Mavi, Anubhav Jangra and Adam Jatowt (2024), "Multi-hop Question Answering", Foundations and Trends® in Information Retrieval: Vol. 17, No. 5, pp 457–586. DOI: 10.1561/1500000102.

which results in a large variety of tasks that require multi-hop reasoning. This leads to different datasets and models that differ significantly from each other and make the field challenging to generalize and survey. We aim to provide a general and formal definition of the MHQA task, and organize and summarize existing MHQA frameworks. We also outline some best practices for building MHQA datasets. This monograph provides a systematic and thorough introduction as well as the structuring of the existing attempts to this highly interesting, yet quite challenging task.

1

Introduction

1.1 Question Answering

An eventual goal of artificial intelligence (AI) is to impart the ability to reason over natural language to machines. In order to achieve this, several natural language understanding and generation tasks have been proposed that require an agent to do some reasoning to get to the goal. One such example is the task of Question Answering (QA) where given a question and some relevant context, the goal is to predict the correct answer. The question answering task provides a quantifiable way to evaluate a system's capability of language understanding and reasoning (Qiu *et al.*, 2019; Rajpurkar *et al.*, 2016a; Hermann *et al.*, 2015). It is a critical problem in the fields of natural language processing (NLP) and information retrieval (IR), and a long-standing AI milestone.

Abundance of readily-available, high-quality information on the internet facilitates the need of automated QA systems that help probe this rich content based on individual needs. Due to recent advancements in Deep Learning techniques (Lan *et al.*, 2019), the machines have become able to successfully beat human performance on datasets like SQUAD 2.0 (Rajpurkar *et al.*, 2016b). However, we have only scratched the surface of what these modern systems are capable of achieving.

Depending on the user requirements, the complexity of QA tasks may vary. Some questions can be answered in brief (*e.g.*, "*Which color do you get when you mix red and yellow paints?*") - such questions are called *objective questions* or *factoid questions.* On the other hand, there exist *subjective questions* that demand detailed explanations to meet user requirements (*e.g.*, "*Why does mixing red, green and blue paints give black color paint, but projecting red, green, and blue light on a white surface return white light?*"). A question can also be considered complex if it requires a very niche domain expertise to answer the question (*e.g.*, "*What symptoms help diagnose chickenpox?*").

1.2 What is Multi-hop Question Answering (MHQA)?

For questions mentioned above, there might exist a single document or a single passage (formally referred to as a '*context*') that can provide a justifiable answer. However, there exist certain questions that cannot be answered using a single *context* (*e.g.*, "*What is the national bird of the nation that has a negative carbon footprint?*"). The task of answering such questions is called multi-hop question answering (MHQA). The goal of MHQA is to predict the correct answer to a question that requires multiple reasoning 'hops' across given contexts (text, table, knowledge graph etc). We look at a more detailed definition of the task in Section 2.

The success in simple QA systems (also referred to as *single hop QA*) does not necessarily entail success of MHQA systems. Min *et al.* (2018) and Qiu *et al.* (2019) observe that most questions in existing single-hop QA datasets are answerable without much reasoning, by retrieving a small set of sentences. Moreover, multi-step reasoning is required by the models to answer complex questions (refer to Table 1.1). Humans can easily perform these multi-step reasoning in their everyday tasks, yet this is still a difficult task for machines. An agent can be said to perform multi-step reasoning if it reaches one or more intermediate conclusions before deriving the final answer and each of the intermediate conclusions serves as a necessary premise for some other conclusion. This sequence of intermediate conclusions, including the final answer, is called a *reasoning chain* and each step from one conclusion to the next can be referred to as a *hop*.

Table 1.1: Examples of various types of multi-hop questions.

Type of question	Question	Answer
Bridge Entity-based (temporal entity)	Who was the president of United States in the year in which Mike Tyson declared his retirement?	George W. Bush
Bridge Entity-based (geographical entity)	What is the national bird of the nation that has a negative carbon footprint?	The Raven
Bridge Entity-based (named entity)	What is the birth place of the tennis player who has won the most grand slams?	Belgrade, Serbia
Intersection	Who is the only person to win an olympic medal and a Nobel prize?	Philip John Noel-Baker
Comparison	Which country has won more soccer world cups - Argentina or Brazil?	Brazil
Commonsense Reasoning	If A prefers fruits over meat, when given an option of apple and chicken sandwich, what will A prefer?	Apple

It is important to note that the inability of AI systems to perform multiple steps of reasoning can be severely limiting, significantly reducing their usability. One such instance can be as shown in Figure 1.1. Say a user is interested in knowing more about 'the daughter of A' and the only relevant information available in this context is 'B's father is C and her mother is A'. In this case, the AI system has to first infer that B is female and her mother is A. The system will then have to use common sense reasoning to conclude that B is the entity of interest and then retrieve the required information (refer to Figure 1.1 for visual aid). Something like this seems trivial to humans but it may fatally confuse many existing AI systems. Therefore, we argue that multi-step reasoning is a crucial challenge and solving it can be a giant leap towards the goals of AI.

1.3 Applications of MHQA

As discussed above, MHQA serves as an appropriate benchmark task for evaluating an agent's ability to perform multi-step reasoning. Along with this scientific significance, the task of MHQA has various practical applications. Queries given to current web search systems can often require multi-hop reasoning to reach the relevant documents. User satisfaction when using such systems can be greatly improved by utilizing multi-hop reasoning models. Furthermore, conversations between humans and agents can be smoother and more informative if

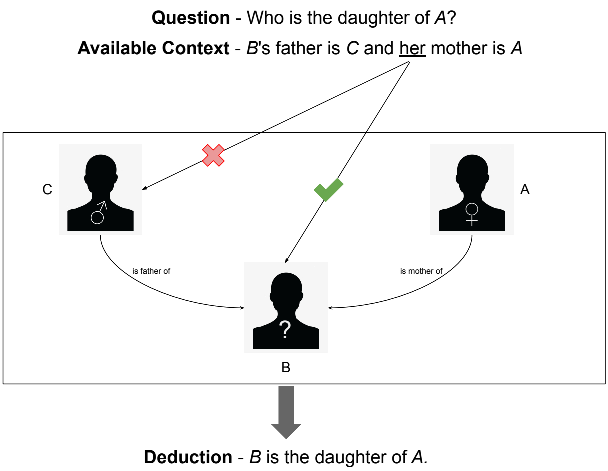

Figure 1.1: An example of multi-hop reasoning

the latter can handle complex questions. Answering a multi-hop question requires systems to aggregate information over multiple contexts. Therefore, techniques that are successful for MHQA can inspire progress in tasks such as sentence fusion (Weiss *et al.*, 2021; Geva *et al.*, 2019b) and abstractive summarization (Nayeem *et al.*, 2018; Lebanoff *et al.*, 2019), event occurrence time prediction (Wang *et al.*, 2021c), as well as multi-document summarization (Ma *et al.*, 2020; Goldstein *et al.*, 2000; Haghighi and Vanderwende, 2009; Barzilay *et al.*, 1999) or timeline summarization (Yan *et al.*, 2011; Ghalandari and Ifrim, 2020; Steen and Markert, 2019; Yu *et al.*, 2021) that require information aggregation over multiple documents. Additionally, most applications of QA such as information extraction (IE) and entailment, can be immensely benefited by multi-hop reasoning abilities (Boros *et al.*, 2021).

Kumar *et al.* (2019) argue that MHQA is a challenging task to an extent that they quantify the difficulty of a question as the number of inference steps (or hops) required to answer the question. This illustrates the direct utility of MHQA for the task of Difficulty controllable Question Generation (DQG) (Gao *et al.*, 2018) that has various applications including curriculum-learning based methods for QA systems (Kurdi *et al.*, 2019) and designing school exams of certain difficulty levels (Sachan and Xing, 2016).

Another problem closely related to MHQA consists of generating clarifying questions for conversational QA (chatbots) (Sun *et al.*, 2021; Zaib *et al.*, 2021). In this setting, the original question/query can be ambiguous and hence more information is needed to disambiguate it. The model is supposed to generate a clarifying question in natural language, asking the user for the missing information. This can be considered as another task involving multi-step reasoning and can be greatly helped by improvements in MHQA.

1.4 Overview

Recently, a variety of datasets and techniques have been proposed for MHQA, including ones designed for MHQA over Knowledge Bases and Knowledge Graphs as well as those designed for QA over tables and text. A substantial number of recent works have focused on the task of MHQA and contributed to significant advancements. High quality datasets (Yang *et al.*, 2018; Welbl *et al.*, 2018; Kočiskỳ *et al.*, 2018; Mihaylov *et al.*, 2018; Khashabi *et al.*, 2018; Chen *et al.*, 2020b; Khot *et al.*, 2020) have encouraged better models to be proposed which in turn have achieved impressive accuracy on these benchmarks. There has been a significant research in the recent years to solve the task. A variety of methods model the task as performing inference over static or dynamic graphs to find the reasoning paths (Ding *et al.*, 2019; Fang *et al.*, 2020; Zhang *et al.*, 2021; Cao *et al.*, 2019; Thayaparan *et al.*, 2019; De Cao *et al.*, 2019; Zhang *et al.*, 2020; Qiu *et al.*, 2019; Huang and Yang, 2021; Shao *et al.*, 2020; Cao and Liu, 2021). A number of works have also attempted to decompose the multi-hop questions into single hop questions or generate follow-up questions based on the retrieved information (Min *et al.*, 2019b; Cao and Liu, 2021; Sun *et al.*, 2021; Zhang *et al.*, 2021; Malon and Bai, 2020). The recent success of large language models (LLMs) has significantly influenced MHQA as well, with multiple attempts of using LLMs' strong natural understanding and emergent abilities for answering complex multi-hop questions (Zhao *et al.*, 2023b; Patel *et al.*, 2022; Balepur *et al.*, 2023; Wang *et al.*, 2023; Rahgouy *et al.*, 2023; Xu *et al.*, 2021). We discuss all these methods in a detailed and organized manner in Sections 4, 5 and 6.

Due to the surge in the attention received by the task over the last decade, we believe that the community would benefit from an extensive survey encompassing recent advancements in MHQA. In this work, we closely cover ∼75 works from top venues including but not limited to EMNLP, ACL, NAACL, TACL, AAAI, EACL, SIGIR, ICLR, COLING, CoRR etc. published from 2016 to 2024. The research community has already several surveys in the field of question-answering, such as for single-hop QA (Allam and Haggag, 2012; Bouziane *et al.*, 2015; Mishra and Jain, 2016; Höffner *et al.*, 2017; Soares and Parreiras, 2020; Dimitrakis *et al.*, 2020), open-domain QA (Roy and Anand, 2021; Etezadi and Shamsfard, 2023; Zhu *et al.*, 2021), medical QA (Lin *et al.*, 2021; Jin *et al.*, 2022), visual QA (Srivastava *et al.*, 2020; Wu *et al.*, 2017), etc. The surveys that are most relevant to MHQA are the ones focused on QA over knowledge bases (Fu *et al.*, 2020; Lan *et al.*, 2021; Diefenbach *et al.*, 2018; Roy and Anand, 2021) and visual QA (Srivastava *et al.*, 2020; Lin *et al.*, 2021; Wu *et al.*, 2017). However, these can be considered as sub-domains of the more general formulation of the MHQA field that this monograph aims to survey. Since the existing works go a long way in summarizing their intended domains, we choose to exclude Visual MHQA and MHQA over Knowledge Bases and Knowledge Graphs from the scope of this work.

We observe that despite the impressive accuracy of recent models on MHQA benchmarks, significant concerns have been raised regarding whether the models are actually able to perform multi-step reasoning in order to answer the multi-hop questions. Several works (Jansen, 2018; Wang *et al.*, 2019; Chen and Durrett, 2019; Min *et al.*, 2019a; Trivedi *et al.*, 2020; Jhamtani and Clark, 2020; Inoue *et al.*, 2020; Tang *et al.*, 2021; Tu *et al.*, 2020) conduct experiments and demonstrate that a significant portion of the accuracy can be ascribed to pattern matching and single step reasoning (also termed as *shortcut reasoning*). This points to new challenges and future directions for research in MHQA. Above all, it is fair to say that despite the inspiring progress made so far, the task of MHQA is still a long way from being solved.

A promising direction for solving some of these challenges is the task of explainable MHQA, a particular setting of MHQA that requires the model to output the correct reasoning chain (or equivalently, some kind of representation of the reasoning chain) along with the correct answer. This increases the model's accountability and interpretability to the end user since the model now has to also explain how it reached the answer. Interpretability of the AI systems is crucial for their wide adoption for most high-stake applications such as finance, law and healthcare (Samek *et al.*, 2017; Alvarez-Melis and Jaakkola, 2017; Arras *et al.*, 2016; Biran and Cotton, 2017; Gilpin *et al.*, 2018). Consequently, more recent works (Feng *et al.*, 2020; Chen *et al.*, 2019; Yang *et al.*, 2018; Inoue *et al.*, 2020; Jhamtani and Clark, 2020) have focused on this setting. Yang *et al.* (2018) have also argued that training the model to output reasoning chain can further help in training to predict the correct answer as it serves as a useful auxiliary task. Tu *et al.* (2020) also find that using the reasoning chain as a supervision signal during training improves the performance on adversarial examples as well.

The remainder of this monograph is structured as follows: Section 2 aims to formalize the task of MHQA in a way that encompasses most existing variants. Section 3 describes existing MHQA datasets, their creation techniques, critiques and challenges.[1] Section 4 discusses traditional pre-LLM models in-depth in a structured way that leads to a taxonomy for existing methods in Section 6. Section 5 is dedicated to recent LLM based methods for MHQA, challenges of incorporating LLMs and their proposed solutions. Section 7 discusses the standard evaluation metrics along with evaluation methods specifically designed for evaluating multi-step reasoning/retrieval. Section 8 touches upon the multi-hop question generation problem. Section 9 then summarizes the insights of the monograph and critiques of the existing methods and datasets, to propose promising directions for future research in MHQA.

[1]We discuss the datasets before methods as doing so provides on overview of the existing variants of the tasks which would be helpful to understand the intuition behind the proposed architectures.

Appendices

A

Background

A.1 BM25

BM25 is a ranking function used to retrieve documents given a search query. BM25 stands stands for *Best Match 25*.[1] It uses a bag-of-words mechanism to score proximity between the search query and the documents. Given a query $Q = q_1, q_2, ..., q_n$, where q_i denotes a keyword in the query Q, the BM25 score of the document D is defined as follows -

$$BM25(D,Q) = \sum_{i=1}^{n} IDF(q_i).\frac{freq(q_i,D).(k_1+1)}{freq(q_i,D) + k_1.(1-b+b.\frac{|D|}{avg.doc.len.})} \tag{A.1}$$

where $freq(q_i, D)$ is the number of times q_i occurs in D, $|.|$ denotes the number of words in D, $avg.doc.len.$ denotes the average number of words in the document, k_1 and b are free parameters,[2] and $IDF(q_i)$ denotes the inverse document frequency weight of query term usually computed as follows -

[1]BM25 is also known as Okapi BM25, which was used first by the Okapi information retrieval system implemented by London's City University (https://en.wikipedia.org/wiki/Okapi_BM25).

[2]Typically $k_1 \in [1.2, 2.0]$ and $b = 0.75$.

$$IDF(q_i) = ln(\frac{N - n(q_i) + 0.5}{n(q_i) + 0.5}) + 1 \tag{A.2}$$

where N is the total number of documents in the collection, and $n(q_i)$ is the number of documents containing q_i.

Even though the technique was devised in 1970s-80s, BM25 and its variations are still widely adopted for document retrieval, especially when the document corpus is very large and using *dense retrievers*[3] has a big computational overhead.

A.2 Recurrent Neural Networks

Recurrent Neural Networks (RNNs) are a class of artificial neural networks that have loop connections that allow information propagation across time through the same neurons. Prior to transformer networks (Vaswani *et al.*, 2017b), RNNs were the most popular framework class to process sequential information, and are still widely adopted in real-world systems. Most practical RNN-based architectures have additional stored states that allow the vanilla RNN architecture to overcome its shortcoming of short-term memory loss. Gated recurrent units (GRU) cells (Cho *et al.*, 2014) and long short term memory (LSTM) cells (Hochreiter and Schmidhuber, 1997b) are two of the most popular stateful RNN cells that use gated mechanism to handle long term memory. See *et al.* (2017b) proposed a pointer generator network to overcome the over-repetition of RNN generated output using coverage loss. We point the readers to the comprehensive survey of recurrent neural networks by Lipton *et al.* (2015) for extensive explanation on the topic.

A.3 Transformers for Language Modeling

Even the advanced RNN models like LSTMs and GRUs have a tough time dealing with long sequences. Luong *et al.* (2015) introduced the attention mechanism which allows the model to focus on certain parts

[3] *Dense retriever* is a general umbrella term used to refer to the neural network based retrieval systems.

of the input when predicting a particular output token. Doing so significantly helps with tasks like machine translation where certain words of the input sequence are directly related to a word in the output sequence. Many forms of attention have since been used effectively for various tasks.

Vaswani *et al.* (2017a) extended the idea of attention by removing the recurrent component of the model altogether and proposed the transformer model where both the encoder and the decoder consist of several self-attention and feed forward layers. The transformer model also introduced the multi-head attention. These components allows for very large models which can have a lot more parameters without comprising on the performance. Transformers are also proved to be very versatile, having great success in a large number of natural language applications.

While the original transformers model was trained using the next-token prediction task implying the unidirectionality of the encoder model, BERT (Kenton and Toutanova, 2019) was a bidirectional encoder based transformer which was trained using the masked language modeling task. BERT has proved to be a versatile model and the word representations learned using BERT have been used as embeddings for almost all natural language tasks.

Success of transformer models including BERT led to their use as large pre-training models and several models like AlBERT (Lan *et al.*, 2019), RoBERTa (Liu *et al.*, 2019) and GPT were proposed. AlBERT uses parameter reduction techniques which allow for smaller and faster training of the BERT models while achieving a similar level of accuracy as BERT. RoBERTa is a much more robustly optimized version of BERT, trained with optimized design and hyperparameters choices, which could significantly outperform the originally trained BERT model.

Pre-training of large language models (LLMs) has become increasingly popular leading to larger and larger models trained on huge corpora of natural language. The different versions of the model follow the same principle, with GPT-1 having 117 million parameters and GPT-4 having about a 100 trillion parameters. GPTs are trained on huge corpora using the next token prediction task. An extensively detailed explanation of different architectures and training techniques for transformer based models is neither feasible nor in the scope for this work. Therefore, we point the readers to the comprehensive survey of transformers by Lin *et al.* (2022) for further details on the topic.

A.4 Graph Neural Networks

Graphs are a very simple and versatile method of representing data and its inherent structure. Neural Networks could be adapted to incorporate this structure leading to Graph Neural Networks (GNNs). GNNs can be adopted for various different types of data and tasks, leading to several improvements increasing their capabilities. The integral part of all these models is the message passing algorithm briefly explained below.

Given a graph $G = (V, E)$ having $n = |V|$ nodes, the representation of each node is updated following the given steps:

- **Initialization:** The representation of every node v is initialized as $h_v^0 = X_v$, where X_v is the feature vector.
- **Update:** For each layer i, the representations of each node v is updated as:

$$h_v^i = \sigma_{u \epsilon N(v)}(W_i \Sigma \frac{h_u^{i-1}}{N(v)} + U_i h_v^{i-1}) \tag{A.3}$$

  where σ is the activation function, W_i and U_i are the weight matrices corresponding to the layer i and $N(v)$ is the set of neighbouring nodes of the node v.
- **Prediction:** The representations after layer K are passed to a linear network for the eventual prediction task.

At every layer, the representation of node v is updated with an activation applied to the weighted average of representations of the nodes directly connected to v. Therefore, after k layers, the node v is supposed to receive the 'message' from all nodes having a path to v of length $\leq k$. The weighted average also ensures that the nodes that are closer to v in the graph end up affecting its representation more.

A layer of a Graph Convolutional Network (GCN) (Kipf and Welling, 2016b) consists of a GNN layer followed by a Linear layer. Relation

GCN (R-GCN) (Schlichtkrull *et al.*, 2017) allow for different kinds of edges by having different weight matrices for nodes connected to v via different kind of edges. Graph Attention Networks (GAN) (Veličković *et al.*, 2018) incorporate self attention into GNNs by using the attention weights while performing the message passing algorithm. Several other modifications of GNNs are proposed for different tasks.

We point the readers to the comprehensive survey of graph neural networks by Wu *et al.* (2020) for further reading on the topic.

A.5 Large Language models

Language models refer to a class of self-supervised NLP models that are trained on large unlabeled datasets to learn to predict the likelihood of a word or sequence of words occurring based on the context provided by the preceding words. This ability to estimate the probability of a word given its context forms the foundation of language modeling. These models undergo training on various tasks, such as next-word prediction (Brown *et al.*, 2020), masked language modeling (the task of predicting randomly missing tokens), and next-sentence prediction (Kenton and Toutanova, 2019), without the need for labeled data. Due to their reliance on extensive training data, language models develop a strong grasp of underlying language patterns and concepts. Generally, language models are not designed for specific tasks and can be fine-tuned with minimal data for various downstream applications. Extensive research has shown that utilizing large language models (LLMs) pre-trained on vast amounts of data yields impressive results in language understanding and generation tasks (Tan *et al.*, 2023; Wang *et al.*, 2021d; Hendy *et al.*, 2023; Blair-Stanek *et al.*, 2023). The advent of transformer models has made it possible to train such highly advanced language models, resulting in popular models like BERT, T5, and GPT-3 (Kenton and Toutanova, 2019; Raffel *et al.*, 2019; Brown *et al.*, 2020).

A.5.1 Generative Pre-trained Transformer (GPT)

GPT, a series of generative pre-trained large language models (Brown *et al.*, 2020), is characterized by its decoder-only transformer architecture.

Unlike other transformer models that have both encoder and decoder blocks, GPT models consist solely of decoder blocks, eliminating the encoder-decoder cross-attention layer from each block. The different versions of GPT, namely GPT, GPT-2, GPT-3, and GPT-4, vary in terms of model size and training data. For example, GPT-3 has 175 billion model parameters and is trained on a massive corpus of 499 billion tokens, while GPT-2 has 1.5 billion parameters and is trained on a dataset of 10 billion tokens.

A.5.2 Prompting GPT-3

GPT-3 has achieved remarkable success in various downstream natural language tasks, including question answering (Tan *et al.*, 2023), Machine Translation (Hendy *et al.*, 2023) and Entailment prediction (Wang *et al.*, 2021d), with minimal supervision required. During a typical run of the model, an incomplete piece of text is provided as a ‘prompt’, and the model iteratively generates the most likely tokens to complete the text. This prompting technique has demonstrated impressive performance in the zero-shot setting, where the model is not provided with any in-context examples and is expected to predict the correct output for the given question in the prompt (Figure A.1).

On the other hand, few-shot prompting (Fei-Fei *et al.*, 2006) involves including a small number of sample input-output pairs within the prompt as references for the model (Figure A.1). The inclusion of a few reference examples provides valuable guidance to the model, allowing it to generate more accurate and relevant responses.

In their work, Wei *et al.* (2023) introduced the concept of chain-of-thought (CoT) prompting, which goes a step beyond simply providing input-sample output pairs. CoT prompting includes a coherent sequence of reasoning steps that gradually build up to the correct answer. By presenting the model with a step-by-step thought process, CoT prompting offers explicit examples of how to arrive at the correct answer based on the given input facts. This method is particularly valuable for tackling complex tasks that demand multiple layers of reasoning including the task that this study focuses on. Figure A.1 shows examples of zero-shot, few-shot, and CoT prompts for an arithmetic question. Here, the prompt consists of 2 in-context examples is 2.

Question: Adam walks 40m north, 25m west and then 35m south-west. How far is he from his starting point?
Answer:

Question: Emily walks 10m east, 5m south, 15m west and then 17m north. How far is she from her starting point?
Answer: 13m

Question: John walks 50m north, 20m east and then 30m north-west. How far is he from his starting point?
Answer: 54.45m

Question: Adam walks 40m north, 25m west and then 35m south-west. How far is he from his starting point?
Answer:

Question: Emily walks 10m east, 5m south, 15m west and then 17m north. How far is she from her starting point?
Answer:
Let's say Emily starts walking from (0, 0). After walking 10m east, her position is (10, 0). After walking 5m south, her position is (10, -5). After walking 15m west, her position is (-5, -5). After walking 17m north, her position is (-5, 12). The distance from the starting point is: $\sqrt{(-5)^2 + (12)^2} = 13$m.
Therefore, the answer is **13m**.

Question: John walks 50m north, 20m east and then 30m north-west. How far is he from his starting point?
Answer:
Let's say John starts walking from (0, 0). After walking 50m north, his position is (0, 50). After walking 20m east, his position is (25, 50). After walking 30m north-west, his position is $(25+15\sqrt{2}, 50-15\sqrt{2})$. The distance from the starting point is: $\sqrt{(25+15\sqrt{2})^2 + (50-15\sqrt{2})^2} = 54.45$m.
Therefore, the answer is **54.45m**.

Question: Adam walks 40m north, 25m west and then 35m south-west. How far is he from his starting point?
Answer:

Figure A.1: Example of zero-shot (top), few-shot (middle), and CoT (bottom) prompting for the same question.

For further background and details, we refer the readers to the comprehensive survey on LLMs by Zhao *et al.* (2023b)

References

Ainslie, J., S. Ontanon, C. Alberti, V. Cvicek, Z. Fisher, P. Pham, A. Ravula, S. Sanghai, Q. Wang, and L. Yang. (2020). "ETC: Encoding Long and Structured Inputs in Transformers". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics. 268–284. DOI: 10.18653/v1/2020.emnlp-main.19.

Allam, A. M. N. and M. H. Haggag. (2012). "The question answering systems: A survey". *International Journal of Research and Reviews in Information Sciences (IJRRIS)*. 2(3).

Alvarez-Melis, D. and T. Jaakkola. (2017). "A causal framework for explaining the predictions of black-box sequence-to-sequence models". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics. 412–421. DOI: 10.18653/v1/D17-1042.

Arras, L., F. Horn, G. Montavon, K.-R. Müller, and W. Samek. (2016). "What is Relevant in a Text Document?": An Interpretable Machine Learning Approach. CoRR abs/1612.07843 (2016)". *arXiv preprint arXiv:1612.07843*.

Balepur, N., J. Huang, S. Moorjani, H. Sundaram, and K. C.-C. Chang. (2023). "Mastering the ABCDs of Complex Questions: Answer-Based Claim Decomposition for Fine-grained Self-Evaluation". arXiv: 2305.14750 [cs.CL].

Banarescu, L., C. Bonial, S. Cai, M. Georgescu, K. Griffitt, U. Hermjakob, K. Knight, P. Koehn, M. Palmer, and N. Schneider. (2013). "Abstract Meaning Representation for Sembanking". In: *Proceedings of the 7th Linguistic Annotation Workshop and Interoperability with Discourse*. Ed. by A. Pareja-Lora, M. Liakata, and S. Dipper. Sofia, Bulgaria: Association for Computational Linguistics. 178–186. URL: https://aclanthology.org/W13-2322.

Banerjee, S. and A. Lavie. (2005). "METEOR: An Automatic Metric for MT Evaluation with Improved Correlation with Human Judgments". In: *Proceedings of the ACL Workshop on Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*. Ann Arbor, Michigan: Association for Computational Linguistics. 65–72. URL: https://aclanthology.org/W05-0909.

Barzilay, R., K. McKeown, and M. Elhadad. (1999). "Information fusion in the context of multi-document summarization". In: *Proceedings of the 37th annual meeting of the Association for Computational Linguistics*. 550–557.

Bauer, L., Y. Wang, and M. Bansal. (2018). "Commonsense for Generative Multi-Hop Question Answering Tasks". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 4220–4230.

Becker, R., F. Corò, G. D'Angelo, and H. Gilbert. (2020). "Balancing Spreads of Influence in a Social Network". *Proceedings of the AAAI Conference on Artificial Intelligence*. 34(Apr.): 3–10. DOI: 10.1609/aaai.v34i01.5327.

Bhagavatula, C. S., T. Noraset, and D. Downey. (2013). "Methods for exploring and mining tables on wikipedia". In: *Proceedings of the ACM SIGKDD workshop on interactive data exploration and analytics*. 18–26.

Biran, O. and C. Cotton. (2017). "Explanation and justification in machine learning: A survey". In: *IJCAI-17 workshop on explainable AI (XAI)*. Vol. 8. No. 1. 8–13.

Blair-Stanek, A., N. Holzenberger, and B. V. Durme. (2023). "Can GPT-3 Perform Statutory Reasoning?" arXiv: 2302.06100 [cs.CL].

Boros, E., J. G. Moreno, and A. Doucet. (2021). "Event Detection as Question Answering with Entity Information". *CoRR*. abs/2104.06969. URL: https://arxiv.org/abs/2104.06969.

Bouziane, A., D. Bouchiha, N. Doumi, and M. Malki. (2015). "Question answering systems: survey and trends". *Procedia Computer Science*. 73: 366–375.

Bowman, S. R., G. Angeli, C. Potts, and C. D. Manning. (2015). "A large annotated corpus for learning natural language inference". In: *Conference on Empirical Methods in Natural Language Processing, EMNLP 2015*. Association for Computational Linguistics (ACL). 632–642.

Brown, T., B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei. (2020). "Language Models are Few-Shot Learners". In: *Advances in Neural Information Processing Systems*. Ed. by H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, and H. Lin. Vol. 33. Curran Associates, Inc. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.

Cao, X. and Y. Liu. (2021). "Coarse-grained decomposition and fine-grained interaction for multi-hop question answering". *Journal of Intelligent Information Systems*: 1–21.

Cao, Y., M. Fang, and D. Tao. (2019). "BAG: Bi-directional Attention Entity Graph Convolutional Network for Multi-hop Reasoning Question Answering". 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)*. 357–362.

Chen, D., A. Fisch, J. Weston, and A. Bordes. (2017a). "Reading Wikipedia to Answer Open-Domain Questions". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics. 1870–1879. DOI: 10.18653/v1/P17-1171.

Chen, J. and G. Durrett. (2019). "Understanding Dataset Design Choices for Multi-hop Reasoning". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics. 4026–4032. DOI: 10.18653/v1/N19-1405.

Chen, J., S.-t. Lin, and G. Durrett. (2019). "Multi-hop question answering via reasoning chains". *arXiv preprint arXiv:1910.02610*.

Chen, Q., X. Zhu, Z.-H. Ling, S. Wei, H. Jiang, and D. Inkpen. (2017b). "Enhanced LSTM for Natural Language Inference". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics. 1657–1668. DOI: 10.18653/v1/P17-1152.

Chen, W., J. Chen, Y. Su, Z. Chen, and W. Y. Wang. (2020a). "Logical Natural Language Generation from Open-Domain Tables". *CoRR*. abs/2004.10404. URL: https://arxiv.org/abs/2004.10404.

Chen, W., H. Zha, Z. Chen, W. Xiong, H. Wang, and W. Y. Wang. (2020b). "HybridQA: A Dataset of Multi-Hop Question Answering over Tabular and Textual Data". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 1026–1036.

Chen, Z., W. Chen, C. Smiley, S. Shah, I. Borova, D. Langdon, R. Moussa, M. Beane, T.-H. Huang, B. Routledge, and W. Y. Wang. (2021). "FinQA: A Dataset of Numerical Reasoning over Financial Data". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics. 3697–3711. DOI: 10.18653/v1/2021.emnlp-main.300.

Cho, K., B. van Merrienboer, C. Gulcehre, F. Bougares, H. Schwenk, and Y. Bengio. (2014). "Learning phrase representations using RNN encoder-decoder for statistical machine translation". In: *Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*.

Choi, E., H. He, M. Iyyer, M. Yatskar, W. Yih, Y. Choi, P. Liang, and L. Zettlemoyer. (2018). "QuAC : Question Answering in Context". *CoRR*. abs/1808.07036. URL: http://arxiv.org/abs/1808.07036.

Chung, H. W., L. Hou, S. Longpre, B. Zoph, Y. Tay, W. Fedus, E. Li, X. Wang, M. Dehghani, S. Brahma, A. Webson, S. S. Gu, Z. Dai, M. Suzgun, X. Chen, A. Chowdhery, S. Narang, G. Mishra, A. Yu, V. Zhao, Y. Huang, A. Dai, H. Yu, S. Petrov, E. H. Chi, J. Dean, J. Devlin, A. Roberts, D. Zhou, Q. V. Le, and J. Wei. (2022). "Scaling Instruction-Finetuned Language Models". DOI: 10.48550/ARXIV.2210.11416.

Church, K. W. and P. Hanks. (1990). "Word Association Norms, Mutual Information, and Lexicography". *Computational Linguistics*. 16(1): 22–29. URL: https://aclanthology.org/J90-1003.

Clark, E., A. Celikyilmaz, and N. A. Smith. (2019). "Sentence mover's similarity: Automatic evaluation for multi-sentence texts". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. 2748–2760.

Clark, P., I. Cowhey, O. Etzioni, T. Khot, A. Sabharwal, C. Schoenick, and O. Tafjord. (2018). "Think you have solved question answering? try arc, the ai2 reasoning challenge". *arXiv preprint arXiv:1803.05457*.

Cobbe, K., V. Kosaraju, M. Bavarian, M. Chen, H. Jun, L. Kaiser, M. Plappert, J. Tworek, J. Hilton, R. Nakano, C. Hesse, and J. Schulman. (2021). "Training Verifiers to Solve Math Word Problems". arXiv: 2110.14168 [cs.LG].

Dai, H., B. Dai, Y. Zhang, S. Li, and L. Song. (2017). "Recurrent Hidden Semi-Markov Model". In: *ICLR*.

Das, R., A. Godbole, D. Kavarthapu, Z. Gong, A. Singhal, M. Yu, X. Guo, T. Gao, H. Zamani, M. Zaheer, *et al.* (2019). "Multi-step entity-centric information retrieval for multi-hop question answering". In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. 113–118.

Dasigi, P., K. Lo, I. Beltagy, A. Cohan, N. A. Smith, and M. Gardner. (2021). "A Dataset of Information-Seeking Questions and Answers Anchored in Research Papers". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics. 4599–4610. DOI: 10.18653/v1/2021.naacl-main.365.

De Cao, N., W. Aziz, and I. Titov. (2019). "Question Answering by Reasoning Across Documents with Graph Convolutional Networks". 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)*. 2306–2317.

Demszky, D., K. Guu, and P. Liang. (2018). "Transforming Question Answering Datasets Into Natural Language Inference Datasets". *CoRR*. abs/1809.02922. URL: http://arxiv.org/abs/1809.02922.

Deng, Z., Y. Zhu, Y. Chen, M. Witbrock, and P. Riddle. (2022). "Interpretable AMR-Based Question Decomposition for Multi-hop Question Answering". In: *Proceedings of the Thirty-First International Joint Conference on Artificial Intelligence, IJCAI-22*. Ed. by L. D. Raedt. International Joint Conferences on Artificial Intelligence Organization. 4093–4099. DOI: 10.24963/ijcai.2022/568.

Diefenbach, D., V. Lopez, K. Singh, and P. Maret. (2018). "Core techniques of question answering systems over knowledge bases: a survey". *Knowledge and Information systems*. 55(3): 529–569.

Dimitrakis, E., K. Sgontzos, and Y. Tzitzikas. (2020). "A survey on question answering systems over linked data and documents". *Journal of intelligent information systems*. 55(2): 233–259.

Ding, M., C. Zhou, Q. Chen, H. Yang, and J. Tang. (2019). "Cognitive Graph for Multi-Hop Reading Comprehension at Scale". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics. 2694–2703. DOI: 10.18653/v1/P19-1259.

Du, X., J. Shao, and C. Cardie. (2017). "Learning to Ask: Neural Question Generation for Reading Comprehension". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Vancouver, Canada: Association for Computational Linguistics. 1342–1352. DOI: 10.18653/v1/P17-1123.

Dua, D., C. dos Santos, P. Ng, B. Athiwaratkun, B. Xiang, M. Gardner, and S. Singh. (2021). "Generative Context Pair Selection for Multi-hop Question Answering". In: *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. 7009–7015.

Dua, D., S. Singh, and M. Gardner. (2020). "Benefits of Intermediate Annotations in Reading Comprehension". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics. 5627–5634. DOI: 10.18653/v1/2020.acl-main.497.

Dua, D., Y. Wang, P. Dasigi, G. Stanovsky, S. Singh, and M. Gardner. (2019). "DROP: A Reading Comprehension Benchmark Requiring Discrete Reasoning Over Paragraphs". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics. 2368–2378. DOI: 10.18653/v1/N19-1246.

Etezadi, R. and M. Shamsfard. (2023). "The state of the art in open domain complex question answering: a survey". *Applied Intelligence*. 53(4): 4124–4144.

Fan, A., C. Gardent, C. Braud, and A. Bordes. (2019). "Using Local Knowledge Graph Construction to Scale Seq2Seq Models to Multi-Document Inputs". 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)*. Hong Kong, China: Association for Computational Linguistics. 4186–4196. DOI: 10.18653/v1/D19-1428.

Fang, Y., S. Sun, Z. Gan, R. Pillai, S. Wang, and J. Liu. (2020). "Hierarchical Graph Network for Multi-hop Question Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 8823–8838.

Fei-Fei, L., R. Fergus, and P. Perona. (2006). "One-Shot Learning of Object Categories". *IEEE transactions on pattern analysis and machine intelligence*. 28(May): 594–611. DOI: 10.1109/TPAMI.2006.79.

Feldman, Y. and R. El-Yaniv. (2019). "Multi-Hop Paragraph Retrieval for Open-Domain Question Answering". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics. 2296–2309. DOI: 10.18653/v1/P19-1222.

Feng, S., W. Shi, Y. Bai, V. Balachandran, T. He, and Y. Tsvetkov. (2024a). "Knowledge Card: Filling LLMs' Knowledge Gaps with Plug-in Specialized Language Models". arXiv: 2305.09955 [cs.CL].

Feng, S., W. Shi, Y. Wang, W. Ding, V. Balachandran, and Y. Tsvetkov. (2024b). "Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration". arXiv: 2402.00367 [cs.CL].

Feng, Y., M. Yu, W. Xiong, X. Guo, J. Huang, S. Chang, M. Campbell, M. Greenspan, and X. Zhu. (2020). "Learning to recover reasoning chains for multi-hop question answering via cooperative games". *arXiv preprint arXiv:2004.02393*.

Fu, B., Y. Qiu, C. Tang, Y. Li, H. Yu, and J. Sun. (2020). "A survey on complex question answering over knowledge base: Recent advances and challenges". *arXiv preprint arXiv:2007.13069*.

Gao, L., A. Madaan, S. Zhou, U. Alon, P. Liu, Y. Yang, J. Callan, and G. Neubig. (2023). "PAL: Program-aided Language Models". arXiv: 2211.10435 [cs.CL].

Gao, Y., J. Wang, L. Bing, I. King, and M. R. Lyu. (2018). "Difficulty Controllable Question Generation for Reading Comprehension". *CoRR*. abs/1807.03586. URL: http://arxiv.org/abs/1807.03586.

Gatt, A. and E. Krahmer. (2018). "Survey of the state of the art in natural language generation: Core tasks, applications and evaluation". *Journal of Artificial Intelligence Research*. 61: 65–170.

Geva, M., Y. Goldberg, and J. Berant. (2019a). "Are We Modeling the Task or the Annotator? An Investigation of Annotator Bias in Natural Language Understanding Datasets". 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)*. Hong Kong, China: Association for Computational Linguistics. 1161–1166. DOI: 10.18653/v1/D19-1107.

Geva, M., E. Malmi, I. Szpektor, and J. Berant. (2019b). "DiscoFuse: A Large-Scale Dataset for Discourse-Based Sentence Fusion". 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)*. 3443–3455.

Ghalandari, D. G. and G. Ifrim. (2020). "Examining the state-of-the-art in news timeline summarization". *arXiv preprint arXiv:2005.10107*.

Gilpin, L. H., D. Bau, B. Z. Yuan, A. Bajwa, M. Specter, and L. Kagal. (2018). "Explaining explanations: An overview of interpretability of machine learning". In: *2018 IEEE 5th International Conference on data science and advanced analytics (DSAA)*. IEEE. 80–89.

Goldstein, J., V. O. Mittal, J. G. Carbonell, and M. Kantrowitz. (2000). "Multi-document summarization by sentence extraction". In: *NAACL-ANLP 2000 workshop: automatic summarization*.

Graves, A., G. Wayne, and I. Danihelka. (2014). "Neural turing machines". *arXiv preprint arXiv:1410.5401*.

Gu, J., Z. Lu, H. Li, and V. O. Li. (2016). "Incorporating Copying Mechanism in Sequence-to-Sequence Learning". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics. 1631–1640. DOI: 10.18653/v1/P16-1154.

Gulcehre, C., S. Ahn, R. Nallapati, B. Zhou, and Y. Bengio. (2016). "Pointing the Unknown Words". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics. 140–149. DOI: 10.18653/v1/P16-1014.

Guo, Y. (2023). "ArthModel: Enhance Arithmetic Skills to Large Language Model". arXiv: 2311.18609 [cs.CL].

Gupta, D., H. Chauhan, R. T. Akella, A. Ekbal, and P. Bhattacharyya. (2020). "Reinforced Multi-task Approach for Multi-hop Question Generation". In: *Proceedings of the 28th International Conference on Computational Linguistics*. 2760–2775.

Gururangan, S., S. Swayamdipta, O. Levy, R. Schwartz, S. Bowman, and N. A. Smith. (2018). "Annotation Artifacts in Natural Language Inference Data". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*. New Orleans, Louisiana: Association for Computational Linguistics. 107–112. DOI: 10.18653/v1/N18-2017.

Haghighi, A. and L. Vanderwende. (2009). "Exploring content models for multi-document summarization". In: *Proceedings of human language technologies: The 2009 annual conference of the North American Chapter of the Association for Computational Linguistics*. 362–370.

Haji, S., K. Suekane, H. Sano, and T. Takagi. (2023). "Exploratory Inference Chain: Exploratorily Chaining Multi-hop Inferences with Large Language Models for Question-Answering". In: *2023 IEEE 17th International Conference on Semantic Computing (ICSC)*. 175–182. DOI: 10.1109/ICSC56153.2023.00036.

Han, S., H. Schoelkopf, Y. Zhao, Z. Qi, M. Riddell, L. Benson, L. Sun, E. Zubova, Y. Qiao, M. Burtell, D. Peng, J. Fan, Y. Liu, B. Wong, M. Sailor, A. Ni, L. Nan, J. Kasai, T. Yu, R. Zhang, S. Joty, A. R. Fabbri, W. Kryscinski, X. V. Lin, C. Xiong, and D. Radev. (2022). "FOLIO: Natural Language Reasoning with First-Order Logic". arXiv: 2209.00840 [cs.CL].

He, P., X. Liu, J. Gao, and W. Chen. (2021). "DeBERTa: Decoding-enhanced BERT with Disentangled Attention". arXiv: 2006.03654 [cs.CL].

Hendrycks, D., C. Burns, S. Kadavath, A. Arora, S. Basart, E. Tang, D. Song, and J. Steinhardt. (2021). "Measuring mathematical problem solving with the math dataset". *arXiv preprint arXiv:2103.03874*.

Hendy, A., M. Abdelrehim, A. Sharaf, V. Raunak, M. Gabr, H. Matsushita, Y. J. Kim, M. Afify, and H. H. Awadalla. (2023). "How Good Are GPT Models at Machine Translation? A Comprehensive Evaluation". arXiv: 2302.09210 [cs.CL].

Hermann, K. M., T. Kocisky, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom. (2015). "Teaching machines to read and comprehend". *Advances in neural information processing systems*. 28.

Hochreiter, S. and J. Schmidhuber. (1997a). "Long short-term memory". *Neural computation*. 9(8): 1735–1780.

Hochreiter, S. and J. Schmidhuber. (1997b). "Long short-term memory". *Neural computation*. 9(8): 1735–1780.

Höffner, K., S. Walter, E. Marx, R. Usbeck, J. Lehmann, and A.-C. Ngonga Ngomo. (2017). "Survey on challenges of question answering in the semantic web". *Semantic Web*. 8(6): 895–920.

Huang, Y. and M. Yang. (2021). "Breadth First Reasoning Graph for Multi-hop Question Answering". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Online: Association for Computational Linguistics. 5810–5821. DOI: 10.18653/v1/2021.naacl-main.464.

Imani, S., L. Du, and H. Shrivastava. (2023). "MathPrompter: Mathematical Reasoning using Large Language Models". arXiv: 2303.05398 [cs.CL].

Inoue, N., P. Stenetorp, and K. Inui. (2020). "R4C: A Benchmark for Evaluating RC Systems to Get the Right Answer for the Right Reason". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics. 6740–6750. DOI: 10.18653/v1/2020.acl-main.602.

Jansen, P. (2018). "Multi-hop Inference for Sentence-level TextGraphs: How Challenging is Meaningfully Combining Information for Science Question Answering?" In: *Proceedings of the Twelfth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-12)*. 12–17.

Jansen, P., E. Wainwright, S. Marmorstein, and C. Morrison. (2018). "WorldTree: A Corpus of Explanation Graphs for Elementary Science Questions supporting Multi-hop Inference". In: *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC 2018)*. Miyazaki, Japan: European Language Resources Association (ELRA). URL: https://aclanthology.org/L18-1433.

Järvelin, K. and J. Kekäläinen. (2002). "Cumulated Gain-Based Evaluation of IR Techniques". *ACM Trans. Inf. Syst.* 20(4): 422–446. DOI: 10.1145/582415.582418.

Jatowt, A., C. A. Yeung, and K. Tanaka. (2013). "Estimating document focus time". In: *22nd ACM International Conference on Information and Knowledge Management, CIKM'13, San Francisco, CA, USA, October 27 - November 1, 2013*. Ed. by Q. He, A. Iyengar, W. Nejdl, J. Pei, and R. Rastogi. ACM. 2273–2278. DOI: 10.1145/2505515.2505655.

Jhamtani, H. and P. Clark. (2020). "Learning to Explain: Datasets and Models for Identifying Valid Reasoning Chains in Multihop Question-Answering". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics. 137–150. DOI: 10.18653/v1/2020.emnlp-main.10.

Jia, R. and P. Liang. (2017). "Adversarial Examples for Evaluating Reading Comprehension Systems". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2021–2031.

Jin, Q., Z. Yuan, G. Xiong, Q. Yu, H. Ying, C. Tan, M. Chen, S. Huang, X. Liu, and S. Yu. (2022). "Biomedical Question Answering: A Survey of Approaches and Challenges". *ACM Computing Surveys (CSUR)*. 55(2): 1–36.

Joshi, N., H. Zhang, K. Kalyanaraman, Z. Hu, K. Chellapilla, H. He, and L. E. Li. (2023a). "Improving Multi-Hop Reasoning in LLMs by Learning from Rich Human Feedback". In: *Neuro-Symbolic Learning and Reasoning in the era of Large Language Models*. URL: https://openreview.net/forum?id=wxfqhp9bNR.

Joshi, N., H. Zhang, K. Kalyanaraman, Z. Hu, K. Chellapilla, H. He, and L. E. Li. (2023b). "Improving Multi-Hop Reasoning in LLMs by Learning from Rich Human Feedback". In: *Neuro-Symbolic Learning and Reasoning in the era of Large Language Models*. URL: https://openreview.net/forum?id=wxfqhp9bNR.

Kadavath, S., T. Conerly, A. Askell, T. Henighan, D. Drain, E. Perez, N. Schiefer, Z. Hatfield-Dodds, N. DasSarma, E. Tran-Johnson, S. Johnston, S. El-Showk, A. Jones, N. Elhage, T. Hume, A. Chen, Y. Bai, S. Bowman, S. Fort, D. Ganguli, D. Hernandez, J. Jacobson, J. Kernion, S. Kravec, L. Lovitt, K. Ndousse, C. Olsson, S. Ringer, D. Amodei, T. Brown, J. Clark, N. Joseph, B. Mann, S. McCandlish, C. Olah, and J. Kaplan. (2022). "Language Models (Mostly) Know What They Know". arXiv: 2207.05221 [cs.CL].

Kadlec, R., O. Bajgar, and J. Kleindienst. (2017). "Knowledge Base Completion: Baselines Strike Back". In: *Proceedings of the 2nd Workshop on Representation Learning for NLP*. Vancouver, Canada: Association for Computational Linguistics. 69–74. DOI: 10.18653/v1/W17-2609.

Kenton, J. D. M.-W. C. and L. K. Toutanova. (2019). "BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding". In: *Proceedings of NAACL-HLT*. 4171–4186.

Khashabi, D., S. Chaturvedi, M. Roth, S. Upadhyay, and D. Roth. (2018). "Looking beyond the surface: A challenge set for reading comprehension over multiple sentences". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. 252–262.

Khot, T., P. Clark, M. Guerquin, P. Jansen, and A. Sabharwal. (2020). "Qasc: A dataset for question answering via sentence composition". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 05. 8082–8090.

Khot, T., A. Sabharwal, and P. Clark. (2019). "What's Missing: A Knowledge Gap Guided Approach for Multi-hop Question Answering". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. 2814–2828.

Khot, T., H. Trivedi, M. Finlayson, Y. Fu, K. Richardson, P. Clark, and A. Sabharwal. (2023). "Decomposed Prompting: A Modular Approach for Solving Complex Tasks". arXiv: 2210.02406 [cs.CL].

Kim, J.-H., J. Jun, and B.-T. Zhang. (2018). "Bilinear attention networks". *Advances in Neural Information Processing Systems*. 31.

Kipf, T. N. and M. Welling. (2016a). "Semi-Supervised Classification with Graph Convolutional Networks". *CoRR*. abs/1609.02907. arXiv: 1609.02907. URL: http://arxiv.org/abs/1609.02907.

Kipf, T. N. and M. Welling. (2016b). "Semi-Supervised Classification with Graph Convolutional Networks". *CoRR*. abs/1609.02907. arXiv: 1609.02907. URL: http://arxiv.org/abs/1609.02907.

Kočiskỳ, T., J. Schwarz, P. Blunsom, C. Dyer, K. M. Hermann, G. Melis, and E. Grefenstette. (2018). "The narrativeqa reading comprehension challenge". *Transactions of the Association for Computational Linguistics*. 6: 317–328.

Kovaleva, O., A. Romanov, A. Rogers, and A. Rumshisky. (2019). "Revealing the Dark Secrets of BERT". 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)*. Hong Kong, China: Association for Computational Linguistics. 4365–4374. DOI: 10.18653/v1/D19-1445.

Kumar, V., Y. Hua, G. Ramakrishnan, G. Qi, L. Gao, and Y.-F. Li. (2019). "Difficulty-controllable multi-hop question generation from knowledge graphs". In: *International Semantic Web Conference*. Springer. 382–398.

Kurdi, G., J. Leo, B. Parsia, U. Sattler, and S. Al-Emari. (2019). "A Systematic Review of Automatic Question Generation for Educational Purposes". *International Journal of Artificial Intelligence in Education*. 30(Nov.). DOI: 10.1007/s40593-019-00186-y.

Kusner, M., Y. Sun, N. Kolkin, and K. Weinberger. (2015). "From word embeddings to document distances". In: *International conference on machine learning*. PMLR. 957–966.

Lan, Y., G. He, J. JIANG, J. JIANG, W. X. ZHAO, and J.-R. WEN. (2021). "A survey on complex knowledge base question answering: Methods, challenges and solutions". In: IJCAI.

Lan, Z., M. Chen, S. Goodman, K. Gimpel, P. Sharma, and R. Soricut. (2019). "ALBERT: A Lite BERT for Self-supervised Learning of Language Representations". In: *International Conference on Learning Representations*.

Lebanoff, L., J. Muchovej, F. Dernoncourt, D. S. Kim, S. Kim, W. Chang, and F. Liu. (2019). "Analyzing Sentence Fusion in Abstractive Summarization". In: *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. 104–110.

Lei, F., X. Li, Y. Wei, S. He, Y. Huang, J. Zhao, and K. Liu. (2023). "S^{3}HQA: A Three-Stage Approach for Multi-hop Text-Table Hybrid Question Answering". arXiv: 2305.11725 [cs.CL].

Lewis, M., Y. Liu, N. Goyal, M. Ghazvininejad, A. Mohamed, O. Levy, V. Stoyanov, and L. Zettlemoyer. (2019). "BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension". *CoRR*. abs/1910.13461. URL: http://arxiv.org/abs/1910.13461.

Li, J., M. Ren, Y. Gao, and Y. Yang. (2023). "Ask to Understand: Question Generation for Multi-hop Question Answering". In: *Chinese Computational Linguistics*. Ed. by M. Sun, B. Qin, X. Qiu, J. Jing, X. Han, G. Rao, and Y. Chen. Singapore: Springer Nature Singapore. 19–36.

Li, J., W. Monroe, T. Shi, S. Jean, A. Ritter, and D. Jurafsky. (2017). "Adversarial Learning for Neural Dialogue Generation". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics. 2157–2169. DOI: 10.18653/v1/D17-1230.

Li, R. and X. Du. (2023). "Leveraging Structured Information for Explainable Multi-hop Question Answering and Reasoning". arXiv: 2311.03734 [cs.CL].

Liang, P., R. Bommasani, T. Lee, D. Tsipras, D. Soylu, M. Yasunaga, Y. Zhang, D. Narayanan, Y. Wu, A. Kumar, B. Newman, B. Yuan, B. Yan, C. Zhang, C. Cosgrove, C. D. Manning, C. Ré, D. Acosta-Navas, D. A. Hudson, E. Zelikman, E. Durmus, F. Ladhak, F. Rong, H. Ren, H. Yao, J. Wang, K. Santhanam, L. Orr, L. Zheng, M. Yuksekgonul, M. Suzgun, N. Kim, N. Guha, N. Chatterji, O. Khattab, P. Henderson, Q. Huang, R. Chi, S. M. Xie, S. Santurkar, S. Ganguli, T. Hashimoto, T. Icard, T. Zhang, V. Chaudhary, W. Wang, X. Li, Y. Mai, Y. Zhang, and Y. Koreeda. (2023). "Holistic Evaluation of Language Models". arXiv: 2211.09110 [cs.CL].

Lin, C.-Y. (2004). "ROUGE: A Package for Automatic Evaluation of Summaries". In: *Text Summarization Branches Out*. Barcelona, Spain: Association for Computational Linguistics. 74–81. URL: https://aclanthology.org/W04-1013.

Lin, T., Y. Wang, X. Liu, and X. Qiu. (2022). "A survey of transformers". *AI Open*. 3: 111–132. DOI: https://doi.org/10.1016/j.aiopen.2022.10.001.

Lin, Z., D. Zhang, Q. Tac, D. Shi, G. Haffari, Q. Wu, M. He, and Z. Ge. (2021). "Medical Visual Question Answering: A Survey". *arXiv preprint arXiv:2111.10056*.

Lindberg, D., F. Popowich, J. Nesbit, and P. Winne. (2013). "Generating Natural Language Questions to Support Learning On-Line". In: *Proceedings of the 14th European Workshop on Natural Language Generation*. Sofia, Bulgaria: Association for Computational Linguistics. 105–114. URL: https://aclanthology.org/W13-2114.

Lipton, Z. C., J. Berkowitz, and C. Elkan. (2015). "A critical review of recurrent neural networks for sequence learning". *arXiv preprint arXiv:1506.00019*.

Liu, L. and M. T. Özsu. (2009). "Mean average precision". *Encyclopedia of Database Systems2009*. 1703.

Liu, Y., M. Ott, N. Goyal, J. Du, M. Joshi, D. Chen, O. Levy, M. Lewis, L. Zettlemoyer, and V. Stoyanov. (2019). "Roberta: A robustly optimized bert pretraining approach". *arXiv preprint arXiv:1907.11692*.

Luong, T., H. Pham, and C. D. Manning. (2015). "Effective Approaches to Attention-based Neural Machine Translation". In: *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Lisbon, Portugal: Association for Computational Linguistics. 1412–1421. DOI: 10.18653/v1/D15-1166.

Ma, C., W. E. Zhang, M. Guo, H. Wang, and Q. Z. Sheng. (2020). "Multi-document summarization via deep learning techniques: A survey". *arXiv preprint arXiv:2011.04843*.

Madaan, A., N. Tandon, P. Gupta, S. Hallinan, L. Gao, S. Wiegreffe, U. Alon, N. Dziri, S. Prabhumoye, Y. Yang, S. Gupta, B. P. Majumder, K. Hermann, S. Welleck, A. Yazdanbakhsh, and P. Clark. (2023). "Self-Refine: Iterative Refinement with Self-Feedback". arXiv: 2303.17651 [cs.CL].

Malon, C. and B. Bai. (2020). "Generating followup questions for interpretable multi-hop question answering". *arXiv preprint arXiv:2002.12344*.

Maltoni, D. and M. Ferrara. (2024). "Arithmetic with Language Models: from Memorization to Computation". arXiv: 2308.01154 [cs.AI].

Mavi, V., A. Saparov, and C. Zhao. (2023). "Retrieval-Augmented Chain-of-Thought in Semi-structured Domains". In: *Proceedings of the Natural Legal Language Processing Workshop 2023*. Ed. by D. Preoțiuc-Pietro, C. Goanta, I. Chalkidis, L. Barrett, G. (Spanakis, and N. Aletras. Singapore: Association for Computational Linguistics. 178–191. DOI: 10.18653/v1/2023.nllp-1.18.

Melo, F. (2013). "Area under the ROC Curve". In: *Encyclopedia of Systems Biology*. Ed. by W. Dubitzky, O. Wolkenhauer, K.-H. Cho, and H. Yokota. New York, NY: Springer New York. 38–39. DOI: 10.1007/978-1-4419-9863-7_209.

Mihaylov, T., P. Clark, T. Khot, and A. Sabharwal. (2018). "Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2381–2391.

Miller, G. A. (1995). "WordNet: A Lexical Database for English". *Commun. ACM*. 38(11): 39–41. DOI: 10.1145/219717.219748.

Min, S., J. Michael, H. Hajishirzi, and L. Zettlemoyer. (2020). "AmbigQA: Answering ambiguous open-domain questions". *arXiv preprint arXiv:2004.10645*.

Min, S., E. Wallace, S. Singh, M. Gardner, H. Hajishirzi, and L. Zettlemoyer. (2019a). "Compositional Questions Do Not Necessitate Multi-hop Reasoning". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics. 4249–4257. DOI: 10.18653/v1/P19-1416.

Min, S., V. Zhong, R. Socher, and C. Xiong. (2018). "Efficient and Robust Question Answering from Minimal Context over Documents". In: *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Melbourne, Australia: Association for Computational Linguistics. 1725–1735. DOI: 10.18653/v1/P18-1160.

Min, S., V. Zhong, L. Zettlemoyer, and H. Hajishirzi. (2019b). "Multi-hop Reading Comprehension through Question Decomposition and Rescoring". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics. 6097–6109. DOI: 10.18653/v1/P19-1613.

Mishra, A. and S. K. Jain. (2016). "A survey on question answering systems with classification". *Journal of King Saud University-Computer and Information Sciences*. 28(3): 345–361.

Nair, I., S. Somasundaram, A. Saxena, and K. Goswami. (2023). "Drilling Down into the Discourse Structure with LLMs for Long Document Question Answering". arXiv: 2311.13565 [cs.CL].

Nallapati, R., B. Zhou, C. dos Santos, C. Gulcehre, and B. Xiang. (2016). "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond". In: *Proceedings of the 20th SIGNLL Conference on Computational Natural Language Learning*. Ed. by S. Riezler and Y. Goldberg. Berlin, Germany: Association for Computational Linguistics. 280–290. DOI: 10.18653/v1/K16-1028.

Nayeem, M. T., T. A. Fuad, and Y. Chali. (2018). "Abstractive unsupervised multi-document summarization using paraphrastic sentence fusion". In: *Proceedings of the 27th International Conference on Computational Linguistics*. 1191–1204.

Nema, P. and M. M. Khapra. (2018). “Towards a Better Metric for Evaluating Question Generation Systems”. In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics. 3950–3959. DOI: 10.18653/v1/D18-1429.

Novikova, J., O. Dušek, A. Cercas Curry, and V. Rieser. (2017). “Why We Need New Evaluation Metrics for NLG”. In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics. 2241–2252. DOI: 10.18653/v1/D17-1238.

Ouyang, L., J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe. (2022). “Training language models to follow instructions with human feedback”. arXiv: 2203.02155 [cs.CL].

Pan, L., W. Chen, W. Xiong, M.-Y. Kan, and W. Y. Wang. (2021). “Unsupervised Multi-hop Question Answering by Question Generation”. In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 5866–5880.

Papineni, K., S. Roukos, T. Ward, and W.-J. Zhu. (2002). “Bleu: a method for automatic evaluation of machine translation”. In: *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*. 311–318.

Patel, A., S. Bhattamishra, and N. Goyal. (2021). “Are NLP Models really able to Solve Simple Math Word Problems?” In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 2080–2094.

Patel, P., S. Mishra, M. Parmar, and C. Baral. (2022). “Is a Question Decomposition Unit All We Need?” arXiv: 2205.12538 [cs.CL].

Pennington, J., R. Socher, and C. D. Manning. (2014). “Glove: Global vectors for word representation”. In: *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*. 1532–1543.

Peters, M. E., M. Neumann, M. Iyyer, M. Gardner, C. Clark, K. Lee, and L. Zettlemoyer. (2018). "Deep Contextualized Word Representations". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics. 2227–2237. DOI: 10.18653/v1/N18-1202.

Qi, P., X. Lin, L. Mehr, Z. Wang, and C. D. Manning. (2019). "Answering Complex Open-domain Questions Through Iterative Query Generation". 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)*. Hong Kong, China: Association for Computational Linguistics. 2590–2602. DOI: 10.18653/v1/D19-1261.

Qiu, L., Y. Xiao, Y. Qu, H. Zhou, L. Li, W. Zhang, and Y. Yu. (2019). "Dynamically Fused Graph Network for Multi-hop Reasoning". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics. 6140–6150. DOI: 10.18653/v1/P19-1617.

Radford, A., J. Wu, R. Child, D. Luan, D. Amodei, I. Sutskever, *et al.* (2019). "Language models are unsupervised multitask learners". *OpenAI blog*. 1(8): 9.

Raffel, C., N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, Y. Zhou, W. Li, and P. J. Liu. (2019). "Exploring the Limits of Transfer Learning with a Unified Text-to-Text Transformer". *CoRR*. abs/1910.10683. URL: http://arxiv.org/abs/1910.10683.

Rahgouy, M., H. B. Giglou, D. Feng, T. Rahgooy, G. Dozier, and C. D. Seals. (2023). "Navigating the Fermi Multiverse: Assessing LLMs for Complex Multi-hop Queries". In:

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. (2016a). "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Austin, Texas: Association for Computational Linguistics. 2383–2392. DOI: 10.18653/v1/D16-1264.

Rajpurkar, P., J. Zhang, K. Lopyrev, and P. Liang. (2016b). "SQuAD: 100,000+ Questions for Machine Comprehension of Text". In: *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. 2383–2392.

Rennie, S. J., E. Marcheret, Y. Mroueh, J. Ross, and V. Goel. (2017). "Self-Critical Sequence Training for Image Captioning". In: *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Roy, R. S. and A. Anand. (2021). *Question Answering for the Curated Web: Tasks and Methods in QA over Knowledge Bases and Text Collections. Synthesis Lectures on Information Concepts, Retrieval, and Services*. Morgan & Claypool Publishers. DOI: 10.2200/S0113ED1V01Y202109ICR076.

Rumelhart, D. E., G. E. Hinton, and R. J. Williams. (1985). "Learning internal representations by error propagation". *Tech. rep.* California Univ San Diego La Jolla Inst for Cognitive Science.

Sachan, D. S., L. Wu, M. Sachan, and W. Hamilton. (2020). "Stronger Transformers for Neural Multi-Hop Question Generation". *arXiv preprint arXiv:2010.11374*.

Sachan, M. and E. Xing. (2016). "Easy Questions First? A Case Study on Curriculum Learning for Question Answering". In: *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. Berlin, Germany: Association for Computational Linguistics. 453–463. DOI: 10.18653/v1/P16-1043.

Saeidi, M., M. Bartolo, P. Lewis, S. Singh, T. Rocktäschel, M. Sheldon, G. Bouchard, and S. Riedel. (2018). "Interpretation of Natural Language Rules in Conversational Machine Reading". In: *EMNLP*.

Samek, W., T. Wiegand, and K.-R. Müller. (2017). "Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models". *arXiv preprint arXiv:1708.08296*.

Saparov, A. and H. He. (2023). "Language Models Are Greedy Reasoners: A Systematic Formal Analysis of Chain-of-Thought". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=qFVVBzXxR2V.

Schlichtkrull, M., T. N. Kipf, P. Bloem, R. Van Den Berg, I. Titov, and M. Welling. (2017). "Modeling Relational Data with Graph Convolutional Networks (2017)". *Preprint*.

Schubotz, M., P. Scharpf, K. Dudhat, Y. Nagar, F. Hamborg, and B. Gipp. (2018). "Introducing mathqa: a math-aware question answering system". *Information Discovery and Delivery*.

Scialom, T., B. Piwowarski, and J. Staiano. (2019). "Self-Attention Architectures for Answer-Agnostic Neural Question Generation". In: *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*. Florence, Italy: Association for Computational Linguistics. 6027–6032. DOI: 10.18653/v1/P19-1604.

See, A., P. J. Liu, and C. D. Manning. (2017a). "Get To The Point: Summarization with Pointer-Generator Networks". In: *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 1073–1083.

See, A., P. J. Liu, and C. D. Manning. (2017b). "Get to the point: Summarization with pointer-generator networks". *arXiv preprint arXiv:1704.04368*.

Seo, M. J., A. Kembhavi, A. Farhadi, and H. Hajishirzi. (2016). "Bidirectional Attention Flow for Machine Comprehension". *CoRR*. abs/1611.01603. URL: http://arxiv.org/abs/1611.01603.

Shao, N., Y. Cui, T. Liu, S. Wang, and G. Hu. (2020). "Is Graph Structure Necessary for Multi-hop Question Answering?" In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. 7187–7192.

Shao, N., Y. Cui, T. Liu, S. Wang, and G. Hu. (2021). "Memory augmented sequential paragraph retrieval for multi-hop question answering". *arXiv preprint arXiv:2102.03741*.

Shi, Q., H. Cui, H. Wang, Q. Zhu, W. Che, and T. Liu. (2024). "Exploring Hybrid Question Answering via Program-based Prompting". arXiv: 2402.10812 [cs.CL].

Sidiropoulos, G., N. Voskarides, S. Vakulenko, and E. Kanoulas. (2021a). "Combining Lexical and Dense Retrieval for Computationally Efficient Multi-hop Question Answering". In: *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*. 58–63.

Sidiropoulos, G., N. Voskarides, S. Vakulenko, and E. Kanoulas. (2021b). "Combining Lexical and Dense Retrieval for Computationally Efficient Multi-hop Question Answering". In: *Proceedings of the Second Workshop on Simple and Efficient Natural Language Processing*. Virtual: Association for Computational Linguistics. 58–63. DOI: 10.18653/v1/2021.sustainlp-1.7.

Slobodkin, A., O. Goldman, A. Caciularu, I. Dagan, and S. Ravfogel. (2023). "The Curious Case of Hallucinatory (Un)answerability: Finding Truths in the Hidden States of Over-Confident Large Language Models". In: *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*. Ed. by H. Bouamor, J. Pino, and K. Bali. Singapore: Association for Computational Linguistics. 3607–3625. DOI: 10.18653/v1/2023.emnlp-main.220.

Soares, M. A. C. and F. S. Parreiras. (2020). "A literature review on question answering techniques, paradigms and systems". *Journal of King Saud University-Computer and Information Sciences*. 32(6): 635–646.

Song, L., Z. Wang, M. Yu, Y. Zhang, R. Florian, and D. Gildea. (2018). "Exploring graph-structured passage representation for multi-hop reading comprehension with graph neural networks". *arXiv preprint arXiv:1809.02040*.

Speer, R., J. Chin, and C. Havasi. (2016). "ConceptNet 5.5: An Open Multilingual Graph of General Knowledge. Singh 2002 (2016)". *arXiv preprint arxiv:1612.03975*.

Srivastava, Y., V. Murali, S. R. Dubey, and S. Mukherjee. (2020). "Visual question answering using deep learning: A survey and performance analysis". In: *International Conference on Computer Vision and Image Processing*. Springer. 75–86.

Steen, J. and K. Markert. (2019). "Abstractive timeline summarization". In: *Proceedings of the 2nd Workshop on New Frontiers in Summarization*. 21–31.

Su, D., Y. Xu, W. Dai, Z. Ji, T. Yu, and P. Fung. (2020). "Multi-hop Question Generation with Graph Convolutional Network". In: *Findings of the Association for Computational Linguistics: EMNLP 2020*. 4636–4647.

Sun, H., W. W. Cohen, and R. Salakhutdinov. (2021). "Iterative Hierarchical Attention for Answering Complex Questions over Long Documents". *arXiv preprint arXiv:2106.00200*.

Talmor, A. and J. Berant. (2018). "The Web as a Knowledge-Base for Answering Complex Questions". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics. 641–651. DOI: 10.18653/v1/N18-1059.

Tan, Y., D. Min, Y. Li, W. Li, N. Hu, Y. Chen, and G. Qi. (2023). "Evaluation of ChatGPT as a Question Answering System for Answering Complex Questions". arXiv: 2303.07992 [cs.CL].

Tang, D., N. Duan, T. Qin, and M. Zhou. (2017). "Question Answering and Question Generation as Dual Tasks". *CoRR*. abs/1706.02027. URL: http://arxiv.org/abs/1706.02027.

Tang, Y., H. T. Ng, and A. Tung. (2021). "Do Multi-Hop Question Answering Systems Know How to Answer the Single-Hop Sub-Questions?" In: *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*. Online: Association for Computational Linguistics. 3244–3249. DOI: 10.18653/v1/2021.eacl-main.283.

Thayaparan, M., M. Valentino, V. Schlegel, and A. Freitas. (2019). "Identifying Supporting Facts for Multi-hop Question Answering with Document Graph Networks". In: *Proceedings of the Thirteenth Workshop on Graph-Based Methods for Natural Language Processing (TextGraphs-13)*. 42–51.

Tian, K., E. Mitchell, A. Zhou, A. Sharma, R. Rafailov, H. Yao, C. Finn, and C. D. Manning. (2023). "Just Ask for Calibration: Strategies for Eliciting Calibrated Confidence Scores from Language Models Fine-Tuned with Human Feedback". arXiv: 2305.14975 [cs.CL].

Trischler, A., T. Wang, X. Yuan, J. Harris, A. Sordoni, P. Bachman, and K. Suleman. (2016). "NewsQA: A Machine Comprehension Dataset". *CoRR*. abs/1611.09830. URL: http://arxiv.org/abs/1611.09830.

Trivedi, H., N. Balasubramanian, T. Khot, and A. Sabharwal. (2020). "Is Multihop QA in DiRe Condition? Measuring and Reducing Disconnected Reasoning". In: *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Online: Association for Computational Linguistics. 8846–8863. DOI: 10.18653/v1/2020.emnlp-main.712.

Trivedi, H., N. Balasubramanian, T. Khot, and A. Sabharwal. (2023). "Interleaving Retrieval with Chain-of-Thought Reasoning for Knowledge-Intensive Multi-Step Questions". arXiv: 2212.10509 [cs.CL].

Trivedi, H., H. Kwon, T. Khot, A. Sabharwal, and N. Balasubramanian. (2019). "Repurposing Entailment for Multi-Hop Question Answering Tasks". 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)*. 2948–2958.

Tu, M., K. Huang, G. Wang, J. Huang, X. He, and B. Zhou. (2020). "Select, answer and explain: Interpretable multi-hop reading comprehension over multiple documents". In: *Proceedings of the AAAI Conference on Artificial Intelligence*. Vol. 34. No. 05. 9073–9080.

Van den Oord, A., Y. Li, and O. Vinyals. (2018). "Representation learning with contrastive predictive coding". *arXiv e-prints*: arXiv–1807.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. (2017a). "Attention is all you need". *Advances in neural information processing systems*. 30.

Vaswani, A., N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin. (2017b). "Attention is all you need". *Advances in neural information processing systems*. 30.

Vedantam, R., C. L. Zitnick, and D. Parikh. (2014). "Cider: consensus-based image description evaluation. CoRR". *arXiv preprint arXiv: 1411.5726*.

Veličković, P., G. Cucurull, A. Casanova, A. Romero, P. Liò, and Y. Bengio. (2018). "Graph Attention Networks". In: *International Conference on Learning Representations*.

Wang, B. (2021). "Mesh-Transformer-JAX: Model-Parallel Implementation of Transformer Language Model with JAX". https://github.com/kingoflolz/mesh-transformer-jax.

Wang, H., M. Yu, X. Guo, R. Das, W. Xiong, and T. Gao. (2019). "Do Multi-hop Readers Dream of Reasoning Chains?" In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. 91–97.

Wang, J., A. Jatowt, M. Färber, and M. Yoshikawa. (2020). "Answering Event-Related Questions over Long-Term News Article Archives". In: *Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part I*. Vol. 12035. *Lecture Notes in Computer Science*. Springer. 774–789.

Wang, J., A. Jatowt, M. Färber, and M. Yoshikawa. (2021a). "Improving question answering for event-focused questions in temporal collections of news articles". *Inf. Retr. J.* 24(1): 29–54.

Wang, J., A. Jatowt, and M. Yoshikawa. (2021b). "ArchivalQA: A Large-scale Benchmark Dataset for Open Domain Question Answering over Archival News Collections". *CoRR*. abs/2109.03438. URL: https://arxiv.org/abs/2109.03438.

Wang, J., A. Jatowt, and M. Yoshikawa. (2021c). "Event Occurrence Date Estimation based on Multivariate Time Series Analysis over Temporal Document Collections". In: *SIGIR '21: The 44th International ACM SIGIR Conference on Research and Development in Information Retrieval, Virtual Event, Canada, July 11-15, 2021*. ACM. 398–407.

Wang, S. and J. Jiang. (2016). "Machine comprehension using match-lstm and answer pointer". *arXiv preprint arXiv:1608.07905*.

Wang, S., H. Fang, M. Khabsa, H. Mao, and H. Ma. (2021d). "Entailment as Few-Shot Learner". arXiv: 2104.14690 [cs.CL].

Wang, X., J. Wei, D. Schuurmans, Q. Le, E. Chi, S. Narang, A. Chowdhery, and D. Zhou. (2023). "Self-Consistency Improves Chain of Thought Reasoning in Language Models". arXiv: 2203.11171 [cs.CL].

Wei, J., X. Wang, D. Schuurmans, M. Bosma, B. Ichter, F. Xia, E. Chi, Q. Le, and D. Zhou. (2023). "Chain-of-Thought Prompting Elicits Reasoning in Large Language Models". arXiv: 2201.11903 [cs.CL].

Weiss, D. B., P. Roit, O. Ernst, and I. Dagan. (2021). "Extending Multi-Text Sentence Fusion Resources via Pyramid Annotations". *arXiv preprint arXiv:2110.04517*.

Welbl, J., P. Stenetorp, and S. Riedel. (2018). "Constructing datasets for multi-hop reading comprehension across documents". *Transactions of the Association for Computational Linguistics*. 6: 287–302.

Welleck, S., I. Kulikov, S. Roller, E. Dinan, K. Cho, and J. Weston. (2019). "Neural Text Generation With Unlikelihood Training". In: *International Conference on Learning Representations*.

Williams, A., N. Nangia, and S. Bowman. (2018). "A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference". In: *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. New Orleans, Louisiana: Association for Computational Linguistics. 1112–1122. DOI: 10.18653/v1/N18-1101.

Williams, R. J. (1992). "Simple Statistical Gradient-Following Algorithms for Connectionist Reinforcement Learning". *Mach. Learn.* 8(3–4): 229–256. DOI: 10.1007/BF00992696.

Wishart, D. S., C. Knox, A. C. Guo, D. Cheng, S. Shrivastava, D. Tzur, B. Gautam, and M. Hassanali. (2008). "DrugBank: a knowledgebase for drugs, drug actions and drug targets". *Nucleic acids research*. 36(suppl_1): D901–D906.

Wu, J., L. Yang, Y. Ji, W. Huang, B. F. Karlsson, and M. Okumura. (2024). "GenDec: A robust generative Question-decomposition method for Multi-hop reasoning". arXiv: 2402.11166 [cs.CL].

Wu, Q., D. Teney, P. Wang, C. Shen, A. Dick, and A. van den Hengel. (2017). "Visual question answering: A survey of methods and datasets". *Computer Vision and Image Understanding*. 163: 21–40.

Wu, Y., M. Schuster, Z. Chen, Q. V. Le, M. Norouzi, W. Macherey, M. Krikun, Y. Cao, Q. Gao, K. Macherey, J. Klingner, A. Shah, M. Johnson, X. Liu, L. Kaiser, S. Gouws, Y. Kato, T. Kudo, H. Kazawa, K. Stevens, G. Kurian, N. Patil, W. Wang, C. Young, J. Smith, D. Riesa, A. Rudnick, O. Vinyals, G. Corrado, M. Hughes, and J. Dean. (2016). "Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation". *CoRR*. abs/1609.08144. URL: http://arxiv.org/abs/1609.08144.

Wu, Z., S. Pan, F. Chen, G. Long, C. Zhang, and S. Y. Philip. (2020). "A comprehensive survey on graph neural networks". *IEEE transactions on neural networks and learning systems*. 32(1): 4–24.

Xiong, W., X. L. Li, S. Iyer, J. Du, P. S. H. Lewis, W. Y. Wang, Y. Mehdad, W. Yih, S. Riedel, D. Kiela, and B. Oguz. (2020). "Answering Complex Open-Domain Questions with Multi-Hop Dense Retrieval". *CoRR*. abs/2009.12756. URL: https://arxiv.org/abs/2009.12756.

Xiong, W., M. Yu, X. Guo, H. Wang, S. Chang, M. Campbell, and W. Y. Wang. (2019). "Simple yet Effective Bridge Reasoning for Open-Domain Multi-Hop Question Answering". In: *Proceedings of the 2nd Workshop on Machine Reading for Question Answering*. 48–52.

Xu, W., Y. Deng, H. Zhang, D. Cai, and W. Lam. (2021). "Exploiting Reasoning Chains for Multi-hop Science Question Answering". arXiv: 2109.02905 [cs.CL].

Yadav, V., S. Bethard, and M. Surdeanu. (2019a). "Alignment over Heterogeneous Embeddings for Question Answering". 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)*. Minneapolis, Minnesota: Association for Computational Linguistics. 2681–2691. DOI: 10.18653/v1/N19-1274.

Yadav, V., S. Bethard, and M. Surdeanu. (2019b). "Quick and (not so) Dirty: Unsupervised Selection of Justification Sentences for Multi-hop Question Answering". In: *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*. Hong Kong, China: Association for Computational Linguistics. 2578–2589. DOI: 10.18653/v1/D19-1260.

Yadav, V., S. Bethard, and M. Surdeanu. (2020). "Unsupervised Alignment-based Iterative Evidence Retrieval for Multi-hop Question Answering". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. 4514–4525.

Yadav, V., S. Bethard, and M. Surdeanu. (2021). "If You Want to Go Far Go Together: Unsupervised Joint Candidate Evidence Retrieval for Multi-hop Question Answering". In: *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. 4571–4581.

Yan, R., X. Wan, J. Otterbacher, L. Kong, X. Li, and Y. Zhang. (2011). "Evolutionary timeline summarization: a balanced optimization framework via iterative substitution". In: *Proceedings of the 34th international ACM SIGIR conference on Research and development in Information Retrieval*. 745–754.

Yang, Z., P. Qi, S. Zhang, Y. Bengio, W. Cohen, R. Salakhutdinov, and C. D. Manning. (2018). "HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. 2369–2380.

Yao, K., L. Zhang, T. Luo, L. Tao, and Y. Wu. (2018). "Teaching Machines to Ask Questions". In: *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, IJCAI-18*. International Joint Conferences on Artificial Intelligence Organization. 4546–4552. DOI: 10.24963/ijcai.2018/632.

Yavuz, S., K. Hashimoto, Y. Zhou, N. S. Keskar, and C. Xiong. (2022). "Modeling Multi-hop Question Answering as Single Sequence Prediction". arXiv: 2205.09226 [cs.CL].

Ye, D., Y. Lin, Z. Liu, Z. Liu, and M. Sun. (2019). "Multi-paragraph reasoning with knowledge-enhanced graph neural network". *arXiv preprint arXiv:1911.02170*.

Yu, J., W. Liu, S. Qiu, Q. Su, K. Wang, X. Quan, and J. Yin. (2020). "Low-Resource Generation of Multi-hop Reasoning Questions". In: *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Online: Association for Computational Linguistics. 6729–6739. DOI: 10.18653/v1/2020.acl-main.601.

Yu, Y., A. Jatowt, A. Doucet, K. Sugiyama, and M. Yoshikawa. (2021). "Multi-TimeLine Summarization (MTLS): Improving Timeline Summarization by Generating Multiple Summaries". 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)*. Online: Association for Computational Linguistics. 377–387.

Zaib, M., W. E. Zhang, Q. Z. Sheng, A. Mahmood, and Y. Zhang. (2021). "Conversational question answering: A survey". *arXiv preprint arXiv: 2106.00874*.

Zelikman, E., Y. Wu, J. Mu, and N. D. Goodman. (2022). "STaR: Bootstrapping Reasoning With Reasoning". arXiv: 2203.14465 [cs.LG].

Zhang, M., F. Li, Y. Wang, Z. Zhang, Y. Zhou, and X. Li. (2020). "Coarse and Fine Granularity Graph Reasoning for Interpretable Multi-Hop Question Answering". *IEEE Access*. 8: 56755–56765. DOI: 10.1109/ACCESS.2020.2981134.

Zhang, X. and M. Lapata. (2017). "Sentence Simplification with Deep Reinforcement Learning". In: *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Copenhagen, Denmark: Association for Computational Linguistics. 584–594. DOI: 10.18653/v1/D17-1062.

Zhang, Y., P. Nie, A. Ramamurthy, and L. Song. (2021). "Answering Any-hop Open-domain Questions with Iterative Document Reranking". In: *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 481–490.

Zhao, R., X. Li, S. Joty, C. Qin, and L. Bing. (2023a). "Verify-and-Edit: A Knowledge-Enhanced Chain-of-Thought Framework". arXiv: 2305.03268 [cs.CL].

Zhao, W. X., K. Zhou, J. Li, T. Tang, X. Wang, Y. Hou, Y. Min, B. Zhang, J. Zhang, Z. Dong, Y. Du, C. Yang, Y. Chen, Z. Chen, J. Jiang, R. Ren, Y. Li, X. Tang, Z. Liu, P. Liu, J.-Y. Nie, and J.-R. Wen. (2023b). "A Survey of Large Language Models". arXiv: 2303.18223 [cs.CL].

Zhao, Y., X. Ni, Y. Ding, and Q. Ke. (2018a). "Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics. 3901–3910. DOI: 10.18653/v1/D18-1424.

Zhao, Y., X. Ni, Y. Ding, and Q. Ke. (2018b). "Paragraph-level Neural Question Generation with Maxout Pointer and Gated Self-attention Networks". In: *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Brussels, Belgium: Association for Computational Linguistics. 3901–3910. DOI: 10.18653/v1/D18-1424.

Zhou, B., K. Richardson, X. Yu, and D. Roth. (2022). "Learning to Decompose: Hypothetical Question Decomposition Based on Comparable Texts". arXiv: 2210.16865 [cs.CL].

Zhou, D., N. Schärli, L. Hou, J. Wei, N. Scales, X. Wang, D. Schuurmans, C. Cui, O. Bousquet, Q. V. Le, and E. H. Chi. (2023). "Least-to-Most Prompting Enables Complex Reasoning in Large Language Models". In: *The Eleventh International Conference on Learning Representations*. URL: https://openreview.net/forum?id=WZH7099tgfM.

Zhu, F., W. Lei, C. Wang, J. Zheng, S. Poria, and T.-S. Chua. (2021). "Retrieving and reading: A comprehensive survey on open-domain question answering". *arXiv preprint arXiv:2101.00774*.