
Knowledge Graph Prompting for Multi-Document Question Answering

Anonymous Author(s)

Affiliation

Address

email

Abstract

1 The 'pre-train, prompt, predict' paradigm of large language models (LLMs) has
2 achieved remarkable success in open-domain question answering (OD-QA). How-
3 ever, few works explore this paradigm in the scenario of multi-document question
4 answering (MD-QA), a task demanding a thorough understanding of the logical
5 associations among the contents and structures of different documents. To fill
6 this crucial gap, we propose a Knowledge Graph Prompting (KGP) method to
7 formulate the right context in prompting LLMs for MD-QA, which consists of a
8 graph construction module and a graph traversal module. For graph construction,
9 we create a knowledge graph (KG) over multiple documents with nodes symboliz-
10 ing passages or document structures (e.g., pages/tables), and edges denoting the
11 semantic/lexical similarity between passages or intra-document structural relations.
12 For graph traversal, we design an LM-guided graph traverser that navigates across
13 nodes and gathers supporting passages assisting LLMs in MD-QA. The constructed
14 graph serves as the global ruler that regulates the transitional space among passages
15 and reduces retrieval latency. Concurrently, the LM-guided traverser acts as a local
16 navigator that gathers pertinent context to progressively approach the question and
17 guarantee retrieval quality. Extensive experiments underscore the efficacy of KGP
18 for MD-QA, signifying the potential of leveraging graphs in enhancing the prompt
19 design for LLMs. Our code will be released upon publication.

20 1 Introduction

21 Due to the emergence of large language models (LLMs), the "pre-train, prompt, predict" paradigm has
22 revolutionized natural language processing (NLP) in real-world applications, such as open-domain
23 question answering (O-QA), fact-checking (FC), and arithmetic reasoning (AR) [1, 6, 2, 22, 38, 32].
24 However, no significant efforts have investigated this framework in the scenario of multi-documental
25 question answering (MD-QA), which enjoys practical usage in academic research, customer support,
26 and financial/legal inquiries that require analysis/insights derived from multiple documents [3, 37].

27 To investigate the capability of LLMs for MD-QA, we randomly sample multi-document ques-
28 tions from the development set of 2WikiMQA [14] and MuSiQue [41], and then prompt LLMs
29 in four different strategies for the answer¹. Successfully answering these questions requires
30 knowledge of multiple Wikipedia documents. As shown in Figure 1, on 2WikiMQA and
31 MuSiQue, directly prompting LLMs without providing any context, i.e., None, achieves only
32 25.07%/10.58% F1 and 18.60%/4.60% EM on 2WikiMQA and MuSiQue, which is far less than
33 59.69%/47.75% F1 and 40.20%/30.60% EM when prompting with supporting facts² provided

¹Detailed experimental setting is presented in Section 5.

²Supporting facts: passages that are assumed to contain the answer to the question.

34 as contexts, i.e., the Golden one. This demonstrates the limitation of fulfilling MD-QA using
 35 solely the knowledge encoded in LLMs. One standard solution to overcome this limitation
 36 in conventional O-QA and single document question-answering (D-QA) [27, 45] is to retrieve
 37 grounding contexts and derive faithful answers from the contexts, i.e., retrieve-and-read [21, 56].
 38

39 However, unlike O-QA and D-QA, the primary
 40 challenge of MD-QA roots in its demands for alternatively retrieving and reasoning knowledge
 41 across different documents [5, 31]. For example, successfully answering questions in Figure
 42 2(a)-(b) requires reasoning over distinct passages from two different documents (in these
 43 two cases, Wikipedia pages). Moreover, each document is essentially a compilation of multi-
 44 modality structured data (e.g., pages, sections, paragraphs, tables, and figures) and some ques-
 45 tions may specifically ask for the content in specific structures, which necessitates a comprehensive
 46 grasp of these complex document structures. For example, the question in Figure 2(c) asks about the
 47 difference between Page 1 and Table 2, which is unanswerable if leveraging heuristic methods like
 48 BM25 or deep-learning ones like DPR [22]. Building on previous challenges, the advent of LLMs
 49 introduces new complexities.
 50

55 For the challenge of alternatively retrieving and reasoning knowledge across different documents,
 56 although previous works train a multi-hop retriever [44, 52] to imitate such process by sequentially
 57 fetching the next passage based on the already-retrieved ones, none of them explore the potential of
 58 engaging LLMs into this process. More recent works design different prompting strategies such as
 59 Chain/Tree/Graph-of-thought [40, 42, 48, 49] to guide LLMs approaching answers progressively.
 60 However, prompting non-open-sourced LLMs back and forth incurs forbiddable latency as well as
 61 unaffordable consumption. In addition, how to integrate different document structures into the prompt
 62 design so that LLMs can understand them is still an open-ended question.

63 In view of the above challenges, we propose a knowledge graph prompting (KGP) method for enhanc-
 64 ing LLMs in MD-QA. Specifically, we construct a knowledge graph (KG) over the given documents
 65 with nodes symbolizing passages or document structures and edges denoting their lexical/semantic
 66 similarity between passages or intra-document structural relations. Then for the first challenge of
 67 alternative retrieving and reasoning knowledge across different documents, we address it by alterna-
 68 tively prompting LMs to generate the next evidence to approach the question, i.e., reasoning, and
 69 selecting the most promising neighbor to visit next from the constructed KG based on the generated
 70 evidence, i.e., retrieval. Moreover, we apply the instruction fine-tuning strategy to augment the
 71 reasoning capability of our own LMs and hence refrain from repeatedly prompting non-open-sourced
 72 LLMs for evidence generation. For the multi-modality challenge, we add different types of nodes
 73 to the KG characterizing different document structures and hence enabling content retrieval within
 74 those specific structures. We highlight our contributions as follows:

- 75 • **Generally-applicable KG Construction.** We propose three KG construction methods over
 76 documents, with passages or document structures as nodes and their lexical/semantical similarity
 77 or structural relations as edges. Then we empirically evaluate the quality of the constructed KGs in
 78 MD-QA by checking the level of overlap between the neighborhood and the supporting facts for
 79 each question (Figure 4). We also provide a comprehensive summary of our proposed and existing
 80 KG construction methods in Table 5 in Supplementary.
- 81 • **Engaging KG for Prompt Formulation.** We design a Knowledge Graph Prompting (KGP) method,
 82 which retrieves the question-relevant contexts by traversing the constructed KG. Meanwhile, we
 83 fine-tune LMs that guide the graph traverser to adaptively navigate the most promising neighbors
 84 for approaching the question based on the already-visited nodes (retrieved passages).
- 85 • **Case Studies Verifying MD-QA Framework.** We provide insightful analysis, including comparing
 86 the quality of the constructed KGs in MD-QA and the performance of using different LMs to guide
 87 the graph traversal. We design a user interface and conduct case studies on visualizing MD-QA in
 88 Section A.7 in Supplementary.

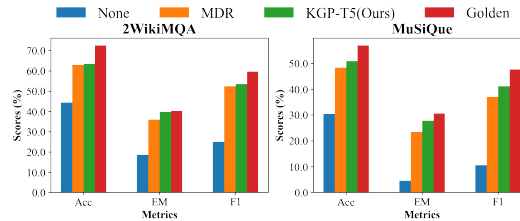


Figure 1: MD-QA performance when prompting ChatGPT with contexts retrieved in different ways.

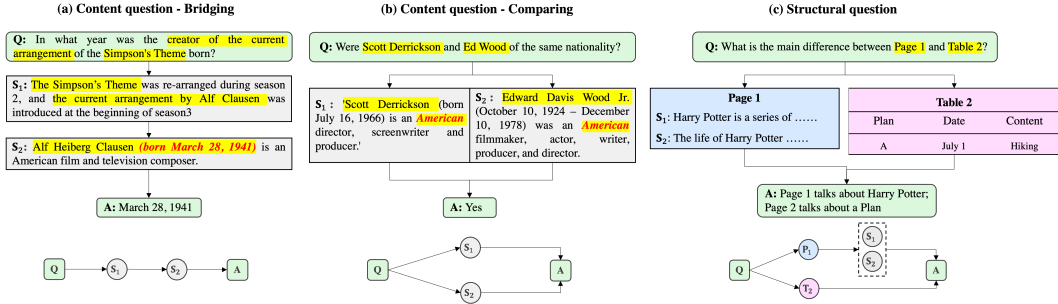


Figure 2: Questions requiring reasoning and retrieving over passages/pages/tables from multiple documents. (a) **Bridging questions** rely on sequential reasoning while (b) **Comparing questions** rely on parallel reasoning over different passages. (c) **Structural questions** rely on fetching contents in the corresponding document structures.

2 Related Work

Question answering Question Answering (QA) aims to provide answers to users' questions in natural language [30, 56], and most QA systems are composed of information retrieval (IR) and answer extraction (AE) [21, 26]. In IR, the system searches for query-relevant factual passages using heuristic methods (BM25) [36] or neural-ranking ones (DPR) [22]. In AE, the final answer is usually extracted as a textual span from related passages. Although this framework has been broadly applied in O-QA [26, 29] and D-QA [27, 45], no previous work focus on MD-QA, which demands alternatively reasoning and retrieving knowledge from multiple documents. To tackle this issue, we construct the KG to encode the logical associations among different passages across multiple documents and design an LM-guided traverser to alternatively generate the reason and visit the most matching passage node.

Pre-train, Prompt, and Predict with LLMs With the emergence of LLMs, the paradigm of 'pre-train, prompt, predict' has gained magnificent popularity in handling a wide spectrum of tasks [13, 24, 54]. This approach begins with pre-training LLMs by pretext tasks to encode world knowledge into tremendous parameters [43] followed by a prompting function to extract pertinent knowledge for downstream tasks [46]. Recent advancements explore different prompting strategies to enhance LLMs' reasoning capabilities [42, 48]. In contrast to that, our work offers a novel perspective by transforming the prompt formulation into the KG traversal.

3 Knowledge Graph Construction

Following [17], let $G = (\mathcal{V}, \mathcal{E})$ be a knowledge graph constructed from a set of documents \mathcal{D} , where the node set $\mathcal{V} = \{v_i\}_{i=1}^n$ representing document structures (e.g., passages/pages/tables, etc.) and the edge set $\mathcal{E} \subset \mathcal{V} \times \mathcal{V}$ representing the connections among different nodes (e.g., semantic/lexical similarity and belonging relations among document structures, etc.). Let $\mathcal{X} = \{\mathcal{X}_i\}_i^n$ be node features and \mathcal{X}_i corresponds to the feature of node v_i , the form of which could be the text for the passage, the markdown for the table and the page number for the page.

Despite numerous well-established KGs [15, 23], they treat nodes/edges as entities/relations, which necessitates sophisticated relational extraction techniques and thereby limits their applicability in general domains [18]. Additionally, their primary focus on the Wikipedia domain also restricts their usage for answering non-Wikipedia questions such as ones over legal or financial documents. To remedy this issue, we propose generally-applicable KG construction methods.

We first analyze two representative questions in Figure 2(a)-(b) to motivate our KG construction. Answering these two questions necessitates the deduction of logical associations among different passages. These associations are encoded either through 1) lexical similarity: common keywords shared among different passages, e.g., 'Alf Clausen' bridges passage S_1 and passage S_2 in Figure 2(a), or 2) semantic similarity: syntactic elements that convey semantic relations, e.g., 'nationality' and 'American director' in Figure 2(b). This motivates us to construct the graph by modeling passages as nodes and their lexical/semantic similarity as edges. More specifically in Figure 3, we split each document into individual passages, and for each passage S_i , we add a node v_i to the KG with its feature being the text of that passage \mathcal{X}_i . Then we add edges by checking the lexical/semantic similarity between pairs of passage nodes.

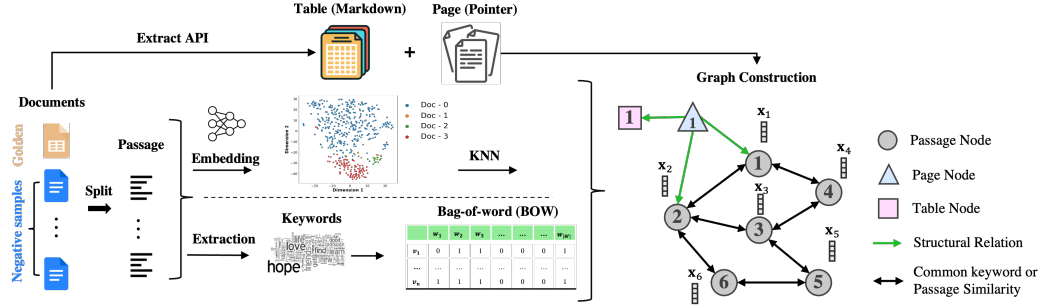


Figure 3: Knowledge Graph Construction. We split each document in the document collection into passages. For each passage, we either directly obtain their embeddings via pre-trained encoders or extract their keywords to build bag-of-words (BOW) features. Then we connect two passages based on their embedding similarity or whether they share common keywords. Additionally, we extract tables/pages via Extract-PDF API and add them as structural nodes to the KG. If pages include passages and tables, we add a directed edge to denote the belonging relations. The table nodes include the markdown formatted content of that table as Figure 8 in Supplementary has empirically shown that LLMs are able to understand tables in this format.

129 3.1 TF-IDF KG Construction

130 For adding edges according to lexical similarity, we first apply TF-IDF keyword extraction [35] over
 131 each document to filter out meaningless words such as supporting verbs and articles, which reduces
 132 the dimension of BOW features, sparsifies the constructed graph and increases the efficiency of the
 133 graph traversal. In addition, we add the document title into the extracted keyword set since some
 134 questions focus on title entities. We collect the extracted keywords from all documents to form the
 135 keyword space \mathcal{W} and then connect two passages if they share any common keyword in \mathcal{W} .

136 3.2 KNN-ST/MDR KG Construction

137 For adding edges according to semantic similarity, we can readily employ pre-existing models such as
 138 sentence transformers to generate passage embedding \mathbf{X}_i for each node v_i and subsequently compute
 139 pairwise similarity matrix to construct the K-nearest neighbor (KNN) graph. However, these off-the-
 140 shelf models, typically trained on tasks not so-related to MD-QA, may not adequately encapsulate
 141 necessary logical associations in their embedding similarity demanded by the question. To overcome
 142 this problem, we follow the training strategy of MDR [44] and train a sentence encoder by predicting
 143 the subsequent supporting facts based on previously supporting facts, thereby endowing the encoder
 144 with reasoning capability. Consequently, the embedding similarity and the corresponding constructed
 145 KNN graph fundamentally encapsulate the necessary logical associations between different passages.

146 3.3 TAGME

147 Moreover, we employ TAGME [28] to extract Wikipedia entities from each passage and construct the
 148 graph based on whether two passage nodes share common Wikipedia entities.

149 In addition to passage nodes, we further add structural nodes into the graph by extracting document
 150 structures via Extract-PDF³. In this paper, we only consider adding pages and tables but the
 151 constructed KG can include more different types of document structures. The feature of table nodes
 152 is the markdown since LLMs can understand this as demonstrated in Figure 8 in Supplementary. The
 153 feature of page nodes is the page number and we add directed edges from it to sentence/table nodes
 154 in that page. *Note that we do not aim to propose a one-size-fits-all KG construction method. Instead,*
 155 *we seek to compare the merits and limitations of various methods in Table 5, offering guidance on*
 156 *which KGs are best suited for specific scenarios.*

157 To verify the constructed KGs indeed encode the necessary information for MD-QA, we randomly
 158 sample questions from HotpotQA and construct KGs over the set of documents for each of these
 159 questions using our proposed methods. We vary the hyperparameters to control the sparsity of

³<https://developer.adobe.com/document-services/docs/overview/pdf-extract-api/>

160 the constructed graph and measure how much percentage of the supporting facts are covered by
 161 neighbors of the seeding passages initialized by TF-IDF. Details about the construction methods
 162 and their hyperparameters are included in Section A.5 in Supplementary. As shown in Figure 4,
 163 as the constructed graph becomes denser, the chance that the neighboring node passages hit the
 164 supporting facts increases (i.e., SF-EM increases) although the redundant information also increases
 165 (i.e., the precision decreases). Given the common keywords shared between one passage to all other
 166 passages are typically far less than the total number of passages across all documents, the density
 167 of the constructed graph by TF-IDF would be upper-bounded, causing lower SF-EM (evidenced by
 168 SF-EM below 0.7 in Figure 4 for TF-IDF curve). For TAGME, we empirically find it identifies a
 169 larger quantity of entities mentioned in a single passage, which leads to a denser graph and causes the
 170 starting SF-EM of TAGME to be already around 0.95. In addition, since KNN-MDR is pre-trained
 171 by predicting the next supporting facts [44] on HotpotQA, it achieves better trade-off than KNN-ST
 172 where the embeddings are directly from the sentence transformer without dataset-specific pre-training.

173 To summarize, although high SF-EM indicates
 174 that the supporting facts for most questions are
 175 fully covered by the neighbors of seeding pas-
 176 sages, low precision signifies that most of these
 177 neighboring passages are irrelevant to the ques-
 178 tion. Therefore, if we blindly perform graph
 179 traversal without any question-tailored adapta-
 180 tion, our retrieved contexts would include redun-
 181 dant passages and compromise the capability of
 182 LLMs in MD-QA (which is also verified by the
 183 low performance of KGP w/o LM in Table 3).
 184 To remedy this issue, in the next section, we
 185 introduce an LM-guided graph traverser to adap-
 186 tively visit neighboring passages that are most
 187 conducive to answering the given question.

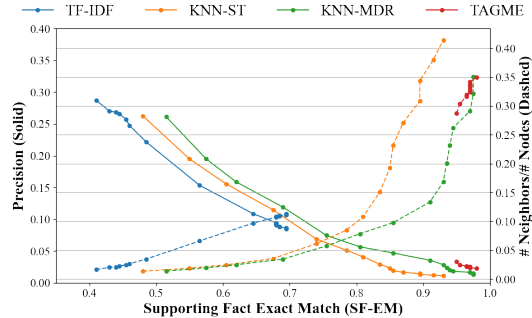


Figure 4: MD-QA performance when prompting ChatGPT with contexts retrieved in different ways.

188 4 LM-guided Graph Traverser

189 A natural solution to enable adaptive graph traversal is to rank the candidate nodes, i.e., the neighbors
 190 of the already-visited nodes in our case, thereby determining which ones to visit next. The most
 191 straightforward way is to apply heuristic-based fuzzy matching or embedding-based similarity ranking,
 192 which cannot capture the intrinsic logic relations between the already traversed paths and the nodes
 193 to visit. Instead, we fine-tune a language model (LM) to guide the graph traversal toward the next
 194 most promising passages in approaching the question based on the visited passages.

195 Given a question q asking about the document content, the LM-guided graph traverser reasons over
 196 previously visited nodes/retrieved passages $\{s_k\}_{k=0}^j$ and then generates the next passage s_{j+1} as
 197 follows:

$$s_{j+1} = \arg \max_{v \in \mathcal{N}_j} \phi(g(\mathcal{X}_v), f(\|\|_{k=0}^j \mathcal{X}_k)), \quad (1)$$

198 where $\|\|_{k=0}^j \mathcal{X}_k$ concatenates the textual information of previously retrieved passages/visited nodes.
 199 For the choice of f , one way is to employ encoder-only models like Roberta-base [2, 44, 52] and
 200 correspondingly g would be another encoder model with $\phi(\cdot)$ being the inner product measuring
 201 the embedding similarity. Another way is to employ encoder-decoder models such as T5 [4, 39]
 202 and correspondingly g would be an identity function with $\phi(\cdot)$ measuring the textual similarity. To
 203 mitigate the hallucination issue [19] and enhance the reasoning capability [42] of LMs, we further
 204 apply instruction fine-tuning to f [7] by predicting the next supporting facts based on previous
 205 supporting facts, thereby integrating commonsense knowledge encoded originally in their pre-trained
 206 parameters with the enhanced reasoning capability inherited from the instruction fine-tuning. After
 207 visiting the top-scoring nodes selected from the candidate neighbor queue by Eq (1), the candidate
 208 neighbor queue is updated by adding neighbors of these newly visited nodes. We iteratively apply
 209 this process until hitting the preset budget. Next, we illustrate the above process with an example in
 210 Figure 5 but leave the comprehensive traversal algorithm in Algorithm 1 in Supplementary.

211 In Figure 5, the content-based question asks ‘In what year was the creator of the current arrangement
 212 of Simpson’s Theme born?’. We use TF-IDF search to initialize our seeding passage Node 1, which

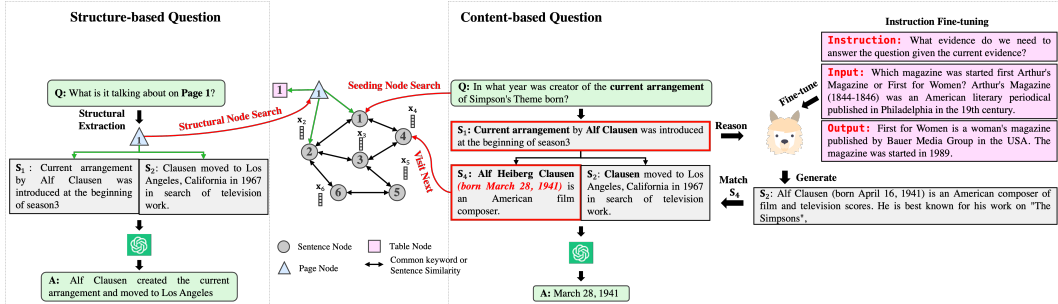


Figure 5: LM-guided graph traverser for context retrieval. For questions on document structures (left), we employ LM to extract structures and retrieve their corresponding contents (the content of pages are passages belonging to that page, and the content of tables is the markdown-formatted text). For questions on document content, we concatenate it with the currently retrieved context and prompt the LM to generate the next evidence to answer the question. By comparing the similarity between the candidate neighboring sentences and the generated passage, we determine the next passage node to traverse. Correspondingly, the candidate neighbors are updated for the next round of traversal.

213 reads: ‘Alf Heiberg Clausen (born March 28, 1941) is an American film composer’. Subsequently, we
 214 prefix the currently retrieved-context (Node 1) with the question and prompt the LM to generate the
 215 next evidence required to approach the question closer. Because we augment the reasoning capability
 216 of the LM by instruction fine-tuning, it is expected to recognize the logical associations between the
 217 question and the currently retrieved context. Consequently, it can predict the subsequent passage that
 218 *maintains logical coherence, albeit may contain factual mistakes*, i.e., ‘Alf Clausen (born April 16,
 219 1941) is an American composer of film and television scores.’ To rectify this potential factual mistake,
 220 we select nodes from the candidate neighbors that match the most with the LM generated passage, in
 221 this case, Node 4 ‘Alf Heiberg Clausen (born March 28, 1941) is an American film composer’. Since
 222 this passage sources directly from documents, it inherently ensures the validity of the information.
 223 Then we prompt LLMs along with the retrieved context Node 1 and 4 for the answer.

224 Additionally, for questions asking about document structures, we extract the document structure
 225 names and locate their corresponding structural nodes in the KG. For the table node, we retrieve its
 226 markdown formatted content while for the page node, we traverse its one-hop neighbor and obtain
 227 passages belonging to that page.

228 5 Experiment

229 In this section, we conduct experiments to verify the proposed knowledge graph prompting method
 230 (KGP) for MD-QA. In particular, we answer the following questions:

- 231 • **Q1 - Section 5.2:** How well does KGP perform MD-QA compared with existing baselines?
- 232 • **Q2 - Section 5.3-5.4:** How do the quality of the constructed KG and the LM-guided graph traverser
 233 impact the MD-QA performance?

234 Due to space limitations, we first briefly introduce our experimental setting in the following and leave
 235 comprehensive details in Supplementary A.1-A.2.

Table 1: Statistics of document collections and KGs constructed by TAGME average across questions.

Dataset	# Questions	# Passages	# Edges	Passage Avg. Length	KG Density
HotpotQA	500	715.22	70420.68	37.55	0.23
IIRC	477	1120.55	143136.17	37.24	0.20
WikiMHop	500	294.19	19235.15	37.24	0.27
MuSiQue	500	748.04	97931.28	38.56	0.29

236 **5.1 Experimental Setting**

237 **5.1.1 Dataset**

238 To explore the uncharted domain of MD-QA, we have created our own datasets to simulate real-world
239 scenarios where users maintain folders containing various documents and pose questions to which the
240 answers are only from certain parts of these documents. Specifically, we randomly sample questions
241 from the development set of four existing datasets: HotpotQA [47], IIRC [10], 2WikiMQA [14], and
242 MuSiQue [41]. For each question, we source documents from Wikipedia that encompass supporting
243 facts pertaining to the question and combine them with randomly sampled negative documents to
244 form the document collection. In addition to the content-based questions from these four existing
245 datasets, we additionally incorporate the ‘Comp’ dataset, an internal company collection of real-world
246 document-based questions. During its creation, humans were asked to read documents and pose
247 questions according to document structures. We summarize the statistics of each dataset along with
248 their KGs in Table 1 with more details in Supplementary.

249 **5.1.2 Baselines**

250 We compare KGP with retrieval baselines in three categories. The first category is the heuristic-based
251 retriever including KNN with fuzzy search, TF-IDF [35], and BM25 [36]. The second category
252 is the deep-learning-based retriever including DPR [22] and MDR [44]. The third category is the
253 prompting-based retriever including IRCOT [40]. For KGP, we explore three variants based on their
254 LM-guided graph traverser: KGP-T5, KGP-LLaMA, and KGP-MDR, using T5 (encoder-decoder),
255 LLaMA (decoder only), and MDR (encoder only) respectively as f in Eq (1).

256 **5.1.3 Evaluation Criteria**

257 Following [53], we compute F1 and EM to compare the LLM’s answer and the ground-truth one. As
258 the predicted answer may not overlap with the ground-truth one, we additionally check the correctness
259 of the answer following [9, 25, 55] by prompting the LLM. Moreover, for evaluating the quality
260 of KGs in Figure 4, we adopt SF-EM (Supporting Fact Exact Matching) and precision from [44].
261 Given the subjective nature of the questions in Comp, we devise the metric, Structure Exact Matching
262 (Struct-EM) to assess if retrieved contexts include the document structures mentioned in the question.

263 **5.2 Performance Comparison on MD-QA**

264 We compare the MD-QA performance of the proposed KGP-T5 and other baselines in Table 2. Firstly,
265 the baseline ‘None’ and ‘Golden’ achieve the worst and the best performance because one provides
266 no context and the other provides the golden context. All other baselines achieve the performance
267 in-between because the retrieved context only covers the partial of the supporting facts. Our proposed
268 methods KGP-T5 rank at the Top-1 except for the Golden baseline. The 2nd-performing baseline
269 MDR fine-tunes a RoBERTa-base encoder by predicting the next supporting fact based on the question
270 and the already retrieved contexts [44]. This next-passage prediction pretext task equips the model
271 with the reasoning capability of the knowledge across different passages and hence increases the
272 quality of the retrieved contexts. The other deep-learning-based retriever DPR achieves much worse
273 performance than MDR because it only fine-tunes the encoder by maximizing the similarity between
274 the query and its supporting facts regardless of their sequential order, demonstrating the importance of
275 understanding the logical order of different knowledge when solving MD-QA [44]. By comparing the
276 MD-QA performance across different datasets, we find that all baselines perform better on HotpotQA
277 than on IIRC. This is because questions in HotpotQA are generally simpler than in IIRC. Existing
278 works [20] have shown that some questions can be easily answered by following shortcuts while
279 questions in IIRC sometimes necessitate arithmetic skills to derive the numerical answers, e.g., ‘How
280 many years did the event last when Wingfield lost much of his fortune?’.

281 Moreover, without any particular design for document structures, no existing baselines can handle
282 structural questions in Comp, e.g. ‘What is the difference between Page 1 and Page 2’ or ‘In Table 3,
283 which station has the highest average flow rate?’. Fortunately, with the constructed KG incorporating
284 the structural nodes and our designed traversal algorithm retrieving structural contexts, our proposed
285 method achieves 67% Struct-EM.

Table 2: MD-QA performances of different baselines. The best (runner-up) are in **bold** (underlined).

Method	Metric	None	KNN	TF-IDF	BM25	DPR	MDR	IRCoT	KGP-MDR	KGP-T5	KGP-LLaMA	Golden
HotpotQA	Acc	41.80	71.57	76.64	71.95	73.43	75.30	74.36	75.72	<u>76.53</u>	75.66	82.19
	EM	19.00	40.73	<u>45.97</u>	41.46	43.61	45.55	45.29	46.09	46.51	46.22	50.20
	F1	30.50	57.97	64.64	59.73	62.11	<u>65.16</u>	64.12	65.77	66.77	66.31	71.06
IIRC	Acc	19.50	43.82	47.47	41.93	48.11	50.84	<u>49.78</u>	49.58	48.28	49.57	62.68
	EM	8.60	25.15	27.22	23.48	26.89	<u>27.52</u>	27.73	29.32	26.94	28.09	35.64
	F1	13.17	37.24	40.80	35.55	41.85	43.47	41.65	43.21	41.54	42.56	54.76
2WikiMQA	Acc	44.40	52.40	58.40	55.80	62.40	<u>63.00</u>	61.81	60.94	63.50	62.45	72.60
	EM	18.60	31.20	34.60	30.80	35.60	36.00	<u>37.75</u>	37.22	39.80	37.55	40.20
	F1	25.07	42.13	44.50	40.55	51.10	<u>52.44</u>	50.17	51.29	53.50	52.45	59.69
MuSiQue	Acc	30.40	44.70	44.40	44.47	44.27	<u>48.39</u>	45.14	51.22	50.92	50.81	57.00
	EM	4.60	18.86	21.59	21.11	20.32	<u>23.49</u>	22.46	27.76	27.90	26.72	30.60
	F1	10.58	30.04	32.50	31.15	31.64	<u>37.03</u>	34.21	41.11	41.19	40.01	47.75
Comp	Acc	0.00	-	-	-	-	-	-	-	67.00	-	100.00
Rank	w Comp	10.54	9.00	6.69	8.92	7.23	4.54	5.61	3.23	3.69	3.69	1.00
	w/o Comp	11.00	9.33	6.83	9.25	7.42	4.50	5.66	3.33	3.83	3.83	1.00

None: no passages but only the question is provided. Golden: supporting facts are provided along with the question.

286 5.3 Impact of the LM-guided Graph Traverser

287 Here we study the influence of using different LMs in guiding graph traversers over TAGME-
 288 constructed KG on MD-QA performance. Specifically, we compare the guidance by no LM (w/o
 289 LM), LLaMA, T5, and MDR in Table 3. Because TAGME w/o LM only blindly traverses in the KG
 290 without any guidance from LM, it unavoidably collects irrelevant passages and hence achieves the
 291 worst performance than others with LM guidance. This aligns with our previous observation on the
 292 generally low precision in Figure 4 and further demonstrates the necessity of using LMs to guide the
 293 graph traversal. Interestingly, we find that KGP-T5 performs better than LLaMA even though the
 294 parameters of LLaMA (7B) are more than the ones with T5 (0.7B). We hypothesize this is because
 295 models with larger amounts of parameters require more training data to avoid over-fitting.

Table 3: Statistics of document collections and KGs by TAGME average across all questions.

Dataset	Metric	HotpotQA			IIRC			2WikiMQA			MuSiQue		
		Acc	EM	F1	Acc	EM	F1	Acc	EM	F1	Acc	EM	F1
TAGME	w/o LM	73.52	43.79	63.14	46.30	27.70	41.43	58.12	35.07	45.95	44.67	21.93	32.90
	LLaMA	75.66	<u>46.22</u>	66.31	<u>49.57</u>	<u>28.09</u>	<u>42.56</u>	<u>62.45</u>	<u>37.55</u>	<u>52.45</u>	50.81	26.72	40.01
	T5	76.53	46.51	<u>66.77</u>	48.28	26.94	41.54	63.50	39.80	53.50	<u>50.92</u>	27.90	41.19
	MDR	<u>75.72</u>	46.09	65.77	49.58	29.32	43.21	60.94	37.22	51.29	51.22	<u>27.76</u>	<u>41.11</u>

296 5.4 Impact of the Constructed Graph and Branching Factor in Graph Traversal

297 Here we construct KGs with varying densities by changing the hyperparameters of TF-IDF/KNN-
 298 ST/KNN-MDR/TAGME and studying its impact on the performance and the neighbor matching time
 299 of MD-QA using KGP-T5. Since the LM-guided graph traverser selects the next node to visit from
 300 neighbors of already visited nodes, the chance that it hits the supporting facts increases as the number
 301 of neighbors increases. In contrast, the neighborhood matching efficiency decreases as the candidate
 302 pool, i.e., \mathcal{N}_j in Eq (1), becomes larger. As evidenced in Figure 6(a), we observe a similar trend, i.e.,
 303 as the KG density increases, the F1/EM increases and then stays stable while the latency for selecting
 304 the most promising neighbors to visit next also increases. KNN-MDR performs better than KNN-ST
 305 when the density of the two constructed KGs is the same. This is because the encoder in KNN-ST
 306 is pre-trained on wide-spectrum datasets while the encoder in MDR is specifically pre-trained on
 307 the HotpotQA dataset by the pretext task of predicting the next supporting facts. Therefore, the
 308 embedding similarity and the corresponding neighbor relations better reflect the logical associations
 309 among different passages, which aligns with the better constructed KG by KNN-MDR than the KG by
 310 KNN-ST in Figure 4. Compared with KNN-MDR/ST, TAGME delivers superior performance at the
 311 cost of increasing latency since the generated KG by TAGME is denser than KGs by KNN-ST/MDR.

312 Furthermore, we perform the sensitivity analysis of the branching factor (the number of nodes
 313 selected from candidate neighbors to visit next). In Figure 6(b) the performance first increases as the
 314 branching factor increases because more passage nodes selected from the candidate neighbors lead
 315 to more reasoning paths to reach the final answer. However, as we fix the context budget to ensure
 316 fair comparison (i.e., the total number of passages we are allowed to retrieve for each question is the
 317 same across all baselines), the performance declines as the branching factor increases because the
 318 number of initial seeding nodes diminishes, leading to reduced coverage of the KG.

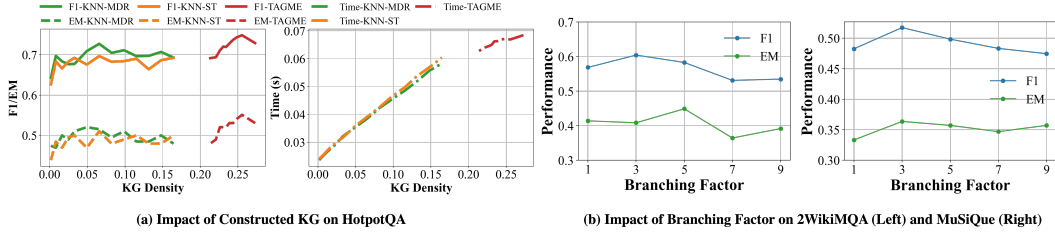


Figure 6: **(a)**: The performance/latency increases as the KG density increases. The results are averaged across 100 randomly sampled questions on HotpotQA. **(b)**: The performance first increases and then decreases as the branching factor increases. The results are averaged across 100 sampled questions on 2WikiMQA and MuSiQue.

319 5.5 Visualizing the Reasoning-and-Retrieving Process of LM-guided Graph Traverser

320 In this section, we visualize the KG-LLaMA’s reasoning-and-retrieving process in retrieving relevant
 321 context for MD-QA. Due to space limitation, for each question, we visualize the top-3 sentence nodes
 322 visited at 1-hop along with their generated evidence from LLaMA that required further to approach
 323 the answer. Based on the generated evidence, we retrieve the top-2 sentence nodes from the candidate
 324 neighbor queue. For each retrieved sentence node, we also visualize its ranking score given by
 325 TF-IDF. We can see that the first retrieved evidence suggests the academy (USMMA) where Joseph
 326 D. Stewart was appointed Superintendent. Based on that, the LLM can then rationalize correctly and
 327 suggest the next passage should include information indicating the location of USMMA, which is
 328 used further to retrieve the ground-truth passage including that information.

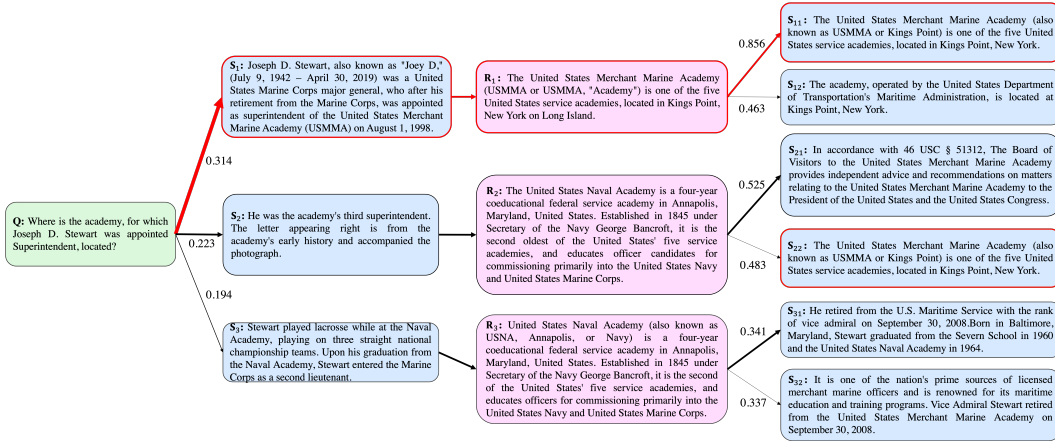


Figure 7: Visualizing the graph traversal over MD-QA.

329 6 Conclusion

330 Answering multi-document questions demands knowledge reasoning and retrieving from different
 331 documents across various modalities, presenting challenges for applying the paradigm of ‘pre-train,
 332 prompt and predict’ with LLMs. Recognizing that the logical associations among passages and
 333 structural relations within the documents can be unified into a graphical representation, we propose a
 334 Knowledge Graph Prompting method (KGP) for aiding LLMs in MD-QA. The KGP constructs KGs
 335 from documents with nodes depicting sentences or document structures and edges denoting their
 336 lexical/semantic similarity or structural relations. Since the constructed KGs may contain irrelevant
 337 neighbor information, we further design an LM-guided graph traverser that selectively visits the most
 338 promising node in approaching the question. In the future, we plan to investigate the capability of
 339 LLMs in understanding graph topology and explore the potential of fine-tuning/prompting LLMs to
 340 encode complex topological signals hidden in the graph.

References

- 341
- 342 [1] Rami Aly, Zhijiang Guo, Michael Schlichtkrull, James Thorne, Andreas Vlachos, Christos
343 Christodoulopoulos, Oana Cocarascu, and Arpit Mittal. Feverous: Fact extraction and verifica-
344 tion over unstructured and structured information. *arXiv preprint arXiv:2106.05707*, 2021.
- 345 [2] Akari Asai, Kazuma Hashimoto, Hannaneh Hajishirzi, Richard Socher, and Caiming Xiong.
346 Learning to retrieve reasoning paths over wikipedia graph for question answering. *arXiv preprint*
347 *arXiv:1911.10470*, 2019.
- 348 [3] Mark Bolino, David Long, and William Turnley. Impression management in organizations:
349 Critical questions, answers, and areas for future research. *Annual Review of Organizational*
350 *Psychology and Organizational Behavior*, 3:377–406, 2016.
- 351 [4] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal,
352 Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are
353 few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.
- 354 [5] Avi Caciularu, Matthew E Peters, Jacob Goldberger, Ido Dagan, and Arman Cohan. Peek across:
355 Improving multi-document modeling via cross-document question-answering. *arXiv preprint*
356 *arXiv:2305.15387*, 2023.
- 357 [6] Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. Reading wikipedia to answer
358 open-domain questions. *arXiv preprint arXiv:1704.00051*, 2017.
- 359 [7] Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Eric Li,
360 Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, et al. Scaling instruction-finetuned
361 language models. *arXiv preprint arXiv:2210.11416*, 2022.
- 362 [8] Junnan Dong, Qinggang Zhang, Xiao Huang, Keyu Duan, Qiaoyu Tan, and Zhimeng Jiang.
363 Hierarchy-aware multi-hop question answering over knowledge graphs. In *Proceedings of the*
364 *ACM Web Conference 2023*, pages 2519–2527, 2023.
- 365 [9] Yann Dubois, Xuechen Li, Rohan Taori, Tianyi Zhang, Ishaan Gulrajani, Jimmy Ba, Carlos
366 Guestrin, Percy Liang, and Tatsunori B Hashimoto. AlpacaFarm: A simulation framework for
367 methods that learn from human feedback. *arXiv preprint arXiv:2305.14387*, 2023.
- 368 [10] James Ferguson, Matt Gardner, Hannaneh Hajishirzi, Tushar Khot, and Pradeep Dasigi.
369 Iirc: A dataset of incomplete information reading comprehension questions. *arXiv preprint*
370 *arXiv:2011.07127*, 2020.
- 371 [11] Paolo Ferragina and Ugo Scaiella. Tagme: on-the-fly annotation of short text fragments (by
372 wikipedia entities). In *Proceedings of the 19th ACM international conference on Information*
373 *and knowledge management*, pages 1625–1628, 2010.
- 374 [12] Aristides Gionis, Piotr Indyk, Rajeev Motwani, et al. Similarity search in high dimensions via
375 hashing. In *Vldb*, volume 99, pages 518–529, 1999.
- 376 [13] Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey,
377 and Noah A Smith. Don’t stop pretraining: Adapt language models to domains and tasks. *arXiv*
378 *preprint arXiv:2004.10964*, 2020.
- 379 [14] Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. Constructing
380 a multi-hop qa dataset for comprehensive evaluation of reasoning steps. *arXiv preprint*
381 *arXiv:2011.01060*, 2020.
- 382 [15] Johannes Hoffart, Fabian M Suchanek, Klaus Berberich, and Gerhard Weikum. Yago2: A
383 spatially and temporally enhanced knowledge base from wikipedia. *Artificial intelligence*,
384 194:28–61, 2013.
- 385 [16] Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang,
386 Lu Wang, and Weizhu Chen. Lora: Low-rank adaptation of large language models. *arXiv*
387 *preprint arXiv:2106.09685*, 2021.

- 388 [17] Ziniu Hu, Yuxiao Dong, Kuansan Wang, and Yizhou Sun. Heterogeneous graph transformer. In
389 *Proceedings of the web conference 2020*, pages 2704–2710, 2020.
- 390 [18] Xiaofeng Huang, Jixin Zhang, Zisang Xu, Lu Ou, and Jianbin Tong. A knowledge graph based
391 question answering method for medical domain. *PeerJ Computer Science*, 7:e667, 2021.
- 392 [19] Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang,
393 Andrea Madotto, and Pascale Fung. Survey of hallucination in natural language generation.
394 *ACM Computing Surveys*, 55(12):1–38, 2023.
- 395 [20] Yichen Jiang and Mohit Bansal. Avoiding reasoning shortcuts: Adversarial evaluation, training,
396 and model development for multi-hop qa. *arXiv preprint arXiv:1906.07132*, 2019.
- 397 [21] Mingxuan Ju, Wenhao Yu, Tong Zhao, Chuxu Zhang, and Yanfang Ye. Grape: Knowledge graph
398 enhanced passage reader for open-domain question answering. *arXiv preprint arXiv:2210.02933*,
399 2022.
- 400 [22] Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov,
401 Danqi Chen, and Wen-tau Yih. Dense passage retrieval for open-domain question answering.
402 *arXiv preprint arXiv:2004.04906*, 2020.
- 403 [23] Jens Lehmann, Robert Isele, Max Jakob, Anja Jentzsch, Dimitris Kontokostas, Pablo N Mendes,
404 Sebastian Hellmann, Mohamed Morsey, Patrick Van Kleef, Sören Auer, et al. Dbpedia—a large-
405 scale, multilingual knowledge base extracted from wikipedia. *Semantic web*, 6(2):167–195,
406 2015.
- 407 [24] Pengfei Liu, Weizhe Yuan, Jinlan Fu, Zhengbao Jiang, Hiroaki Hayashi, and Graham Neubig.
408 Pre-train, prompt, and predict: A systematic survey of prompting methods in natural language
409 processing. *ACM Computing Surveys*, 55(9):1–35, 2023.
- 410 [25] Yang Liu, Dan Iter, Yichong Xu, Shuohang Wang, Ruochen Xu, and Chenguang Zhu. Gpteval:
411 Nlg evaluation using gpt-4 with better human alignment. *arXiv preprint arXiv:2303.16634*,
412 2023.
- 413 [26] Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and
414 Weizhu Chen. Rider: Reader-guided passage reranking for open-domain question answering.
415 *arXiv preprint arXiv:2101.00294*, 2021.
- 416 [27] Minesh Mathew, Dimosthenis Karatzas, and CV Jawahar. Docvqa: A dataset for vqa on
417 document images. In *Proceedings of the IEEE/CVF winter conference on applications of*
418 *computer vision*, pages 2200–2209, 2021.
- 419 [28] Sewon Min, Danqi Chen, Luke Zettlemoyer, and Hannaneh Hajishirzi. Knowledge guided text
420 retrieval and reading for open domain question answering. *arXiv preprint arXiv:1911.03868*,
421 2019.
- 422 [29] Kyosuke Nishida, Itsumi Saito, Atsushi Otsuka, Hisako Asano, and Junji Tomita. Retrieve-and-
423 read: Multi-task learning of information retrieval and reading comprehension. In *Proceedings*
424 *of the 27th ACM international conference on information and knowledge management*, pages
425 647–656, 2018.
- 426 [30] Hariom A Pandya and Brijesh S Bhatt. Question answering survey: Directions, challenges,
427 datasets, evaluation matrices. *arXiv preprint arXiv:2112.03572*, 2021.
- 428 [31] Jayr Pereira, Robson Fidalgo, Roberto Lotufo, and Rodrigo Nogueira. Visconde: Multi-
429 document qa with gpt-3 and neural reranking. In *European Conference on Information Retrieval*,
430 pages 534–543. Springer, 2023.
- 431 [32] Chengwei Qin, Aston Zhang, Zhuosheng Zhang, Jiaao Chen, Michihiro Yasunaga, and Diyi
432 Yang. Is chatgpt a general-purpose natural language processing task solver? *arXiv preprint*
433 *arXiv:2302.06476*, 2023.

- 434 [33] Ao Qu, Yu Wang, Yue Hu, Yanbing Wang, and Hiba Baroud. A data-integration analysis on
435 road emissions and traffic patterns. In *Driving Scientific and Engineering Discoveries Through*
436 *the Convergence of HPC, Big Data and AI: 17th Smoky Mountains Computational Sciences*
437 *and Engineering Conference, SMC 2020, Oak Ridge, TN, USA, August 26-28, 2020, Revised*
438 *Selected Papers 17*, pages 503–517. Springer, 2020.
- 439 [34] Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena,
440 Yanqi Zhou, Wei Li, and Peter J Liu. Exploring the limits of transfer learning with a unified
441 text-to-text transformer. *The Journal of Machine Learning Research*, 21(1):5485–5551, 2020.
- 442 [35] Juan Ramos et al. Using tf-idf to determine word relevance in document queries. In *Proceedings*
443 *of the first instructional conference on machine learning*, volume 242, pages 29–48. Citeseer,
444 2003.
- 445 [36] Stephen Robertson, Hugo Zaragoza, et al. The probabilistic relevance framework: Bm25 and
446 beyond. *Foundations and Trends® in Information Retrieval*, 3(4):333–389, 2009.
- 447 [37] Girolamo Tessuto. Legal problem question answer genre across jurisdictions and cultures.
448 *English for Specific Purposes*, 30(4):298–309, 2011.
- 449 [38] James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. Fever: a
450 large-scale dataset for fact extraction and verification. *arXiv preprint arXiv:1803.05355*, 2018.
- 451 [39] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timo-
452 thée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open
453 and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.
- 454 [40] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Interleaving
455 retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. *arXiv*
456 *preprint arXiv:2212.10509*, 2022.
- 457 [41] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. Musique:
458 Multihop questions via single-hop question composition. *Transactions of the Association for*
459 *Computational Linguistics*, 10:539–554, 2022.
- 460 [42] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le,
461 Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models.
462 *Advances in Neural Information Processing Systems*, 35:24824–24837, 2022.
- 463 [43] Xuansheng Wu, Kaixiong Zhou, Mingchen Sun, Xin Wang, and Ninghao Liu. A survey of graph
464 prompting methods: techniques, applications, and challenges. *arXiv preprint arXiv:2303.07275*,
465 2023.
- 466 [44] Wenhan Xiong, Xiang Lorraine Li, Srinu Iyer, Jingfei Du, Patrick Lewis, William Yang Wang,
467 Yashar Mehdad, Wen-tau Yih, Sebastian Riedel, Douwe Kiela, et al. Answering complex
468 open-domain questions with multi-hop dense retrieval. *arXiv preprint arXiv:2009.12756*, 2020.
- 469 [45] Yang Xu, Yiheng Xu, Tengchao Lv, Lei Cui, Furu Wei, Guoxin Wang, Yijuan Lu, Dinei
470 Florencio, Cha Zhang, Wanxiang Che, et al. Layoutlmv2: Multi-modal pre-training for visually-
471 rich document understanding. *arXiv preprint arXiv:2012.14740*, 2020.
- 472 [46] Jingfeng Yang, Hongye Jin, Ruixiang Tang, Xiaotian Han, Qizhang Feng, Haoming Jiang, Bing
473 Yin, and Xia Hu. Harnessing the power of llms in practice: A survey on chatgpt and beyond.
474 *arXiv preprint arXiv:2304.13712*, 2023.
- 475 [47] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W Cohen, Ruslan Salakhut-
476 dinov, and Christopher D Manning. Hotpotqa: A dataset for diverse, explainable multi-hop
477 question answering. *arXiv preprint arXiv:1809.09600*, 2018.
- 478 [48] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Thomas L Griffiths, Yuan Cao, and Karthik
479 Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *arXiv*
480 *preprint arXiv:2305.10601*, 2023.

- 481 [49] Yao Yao, Zuchao Li, and Hai Zhao. Beyond chain-of-thought, effective graph-of-thought
482 reasoning in large language models. *arXiv preprint arXiv:2305.16582*, 2023.
- 483 [50] Michihiro Yasunaga, Antoine Bosselut, Hongyu Ren, Xikun Zhang, Christopher D Manning,
484 Percy S Liang, and Jure Leskovec. Deep bidirectional language-knowledge graph pretraining.
485 *Advances in Neural Information Processing Systems*, 35:37309–37323, 2022.
- 486 [51] Michihiro Yasunaga, Hongyu Ren, Antoine Bosselut, Percy Liang, and Jure Leskovec. Qa-gnn:
487 Reasoning with language models and knowledge graphs for question answering. *arXiv preprint*
488 *arXiv:2104.06378*, 2021.
- 489 [52] Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, Nitish Shirish Keskar, and Caiming Xiong.
490 Modeling multi-hop question answering as single sequence prediction. *arXiv preprint*
491 *arXiv:2205.09226*, 2022.
- 492 [53] Wenhao Yu, Dan Iter, Shuohang Wang, Yichong Xu, Mingxuan Ju, Soumya Sanyal, Chenguang
493 Zhu, Michael Zeng, and Meng Jiang. Generate rather than retrieve: Large language models are
494 strong context generators. *arXiv preprint arXiv:2209.10063*, 2022.
- 495 [54] Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming
496 Shen, and Chao Zhang. Large language model as attributed training data generator: A tale of
497 diversity and bias. *arXiv preprint arXiv:2306.15895*, 2023.
- 498 [55] Yanzhe Zhang, Ruiyi Zhang, Jiuxiang Gu, Yufan Zhou, Nedim Lipka, Diyi Yang, and Tong Sun.
499 Llavar: Enhanced visual instruction tuning for text-rich image understanding. *arXiv preprint*
500 *arXiv:2306.17107*, 2023.
- 501 [56] Fengbin Zhu, Wenqiang Lei, Chao Wang, Jianming Zheng, Soujanya Poria, and Tat-Seng Chua.
502 Retrieving and reading: A comprehensive survey on open-domain question answering. *arXiv*
503 *preprint arXiv:2101.00774*, 2021.

504 A Supplementary

505 A.1 Dataset Collection

506 This section introduces the collection of datasets used for the experiments conducted in this paper.

507 A.1.1 Document Set Collection and Precession

508 As no previous works focus on MD-QA, we create our own datasets to simulate real-world scenarios
509 where users maintain folders containing various documents and pose questions to which the answers
510 are only from certain parts of these documents. To imitate this scenario, we randomly sample
511 questions from the development set of existing datasets: HotpotQA/IIRC/2WikiMQA/MuSiQue, and
512 then for each specific question, we fetch documents from Wikipedia that encompass supporting facts
513 pertaining to the question ⁴ and term these documents as golden documents. Then we randomly
514 sample negative documents from Wikipedia and pair them with golden documents to constitute the
515 document collection. For each document in the collected document set, we split it into multiple
516 passages with the default passage length being the sentence length. As questions from these existing
517 datasets are only focused on document contents, we additionally incorporate the ‘Comp’ dataset, an
518 internal company collection of real-world questions focusing on document structures.

519 A.1.2 Knowledge Graph Construction

520 We construct a knowledge graph for each question and its corresponding collection of documents.
521 For datasets where the questions are from Wikipedia: HotpotQA, IIRC, WikiMHop, and Musique,
522 we only have passage nodes since answering questions in these datasets does not require information
523 about document structures. For the Comp dataset, in addition to passage nodes, we apply ExtractAPI
524 to obtain the page and table information so that the constructed KG also has pages/tables as nodes.
525 For all of these datasets, we add edges following Section 3. Table 4 summarizes the average statistics
526 of the document collections across all questions with their corresponding KGs. Except for Comp, we
527 plan to release the code for collecting the documents and constructing the KGs upon publication.

Table 4: Statistics of document collections and their corresponding knowledge graph used in Table 2 and 3 average across all questions.

Dataset	#Docs	#Questions	#Passages	#Edges	Passage Avg. Length	KG Density
HotpotQA	12	500	715.22	70420.68	37.55	0.23
IIRC	12	477	1120.55	143136.17	37.24	0.20
2WikiMQA	12	500	294.19	19235.15	37.24	0.27
MuSiQue	12	500	748.04	97931.28	38.55	0.29

528 For Comp, due to privacy concerns, we omit the data statistics but only provide some question
529 examples, e.g., ‘How many more classical students in Table 2 had the mixed teaching style versus the
530 classical teaching style?’ or ‘Can you give me a simple summary about page 5?’.

531 A.1.3 Sequential Data Collection

532 Training MDR [44] requires rearranging supporting facts into the sequential order that progressively
533 approaches the answer. To fulfill this requirement, we directly follow MDR and use the pre-processed
534 HotpotQA data from the GitHub Repository⁵ to train the encoder and apply it to other datasets that
535 do not provide the sequential order of supporting facts. For instruction fine-tuning LLaMA, we
536 still use the above HotpotQA data and rearrange it into the instruction-input-output format and use
537 the instruction ‘What evidence do we need to answer the question given the current evidence’. We
538 present one example in Listing 1. For T5-large, we use the same input-output but prefix the reasoning
539 instruction to the input following the original T5 input format [34].

⁴The HotpotQA/IIRC/2WikiMQA/Musique datasets already have the supporting facts for each question.

⁵https://github.com/facebookresearch/multihop_dense_retrieval/tree/main

540 A.2 Experiment Details

541 A.2.1 Training DPR and MDR

542 For training DPR [22], we pair each question with its supporting facts as its positive passages, and
543 some randomly sampled negative passages as its negative passages. For training MDR [44], as
544 each question in HotpotQA only requires 2 supporting facts to derive the answer, we set the first
545 supporting fact as the positive pair for each question. Further, we concatenate this question and the
546 first supporting fact to form a new question and for this newly-formed question, we set the second
547 supporting fact as its positive pair. For both the original question and the concatenated one, we
548 randomly sample other passages as the negative pair. Following [44, 22], we use RoBERTa-base as
549 the default encoder and search hyperparameters for them as follows: hidden dimension 768, max
550 context length {128, 256, 350}, batch size {128, 256, 512}, epoch 50, warmup steps 300, learning
551 rate $2e - 5$, gradient clipping range 2.

552 A.2.2 Instruction Fine-tuning LLaMA⁶ and T5-Large⁷

553 We fine-tune LLaMA using instruction data in Listing 1. Due to the computational limitation, we
554 choose LLaMA-7B and use LoRA [16]. For fine-tuning T5-Large, we use the same instruction data
555 except that we remove the instruction but only prefix the reasoning instruction to the input [34]. We
556 use the default hyperparameters from their original GitHub repository to fine-tune these two LLMs.

557 A.2.3 Prompting LLMs for MD-QA - Table 2 and 3

558 Following [40], we randomly select questions from the development set for reporting the performance.
559 To ensure a fair comparison, we set the number of retrieved passages to 30 across all baselines and
560 use ChatGPT as the downstream LLM for reading the retrieved passages and generating the answer.
561 We summarize the key implementation details for each baseline as follows:

- 562 • **KNN**: We employ the sentence-transformer variant ‘multi-qa-MiniLM-L6-cos-v1’ to obtain passage
563 embeddings as it has been trained on 215M (question, answer) pairs from diverse sources. Then we
564 select the top-15 passages according to the embedding similarity and the top-15 passages according
565 to the fuzzy matching⁸.
- 566 • **MDR**: We use beam search with the inner product as the scoring function to rank passages. We
567 limit the search depth to 2 as answering questions in HotpotQA requires at most 2-hop reasoning
568 steps [44]. We set the number of passages to be 15 in the first-hop retrieval and for each of these
569 passages, we further retrieve 3 more passages in the second round, which in total generates 45
570 passage pairs. Then we rank these 45 passage pairs by the product of the scores between the
571 first-hop and the second-hop retrieval and select the top 30 ones as the final context.
- 572 • **IRCoT**: Instead of directly employing the original IRCoT code [40], we modify it based on our
573 problem setting. The first reason is that passages to be retrieved in IRCoT [40] are the pre-processed
574 Wikipedia Corpus and do not cover the whole contents of Wikipedia documents, which thereby
575 is not aligned with our MD-QA setting. The second reason is that the question-answering reader
576 employed in IRCoT requires running on A100-80G GPU, which is not affordable on our side.
577 Therefore, we modify the IRCoT by replacing the question reader with the ChatGPT and using
578 our pre-processed Wikipedia document collections as introduced in Section A.1. For the prompt
579 used in the reasoning step, we select 2 examples from ‘gold_with_2_distractors_context’ for the
580 demonstration purpose. We iteratively select top-5 passages based on the generated reason from
581 LLM along with their document titles and add them to the retrieved context until hitting the prefix
582 budget. For the prompt used in the reading step, we use exactly the same prompt as other baselines
583 as we find it empirically leads to better performance than the original one used in IRCoT [40].
- 584 • **KGP-T5/LLaMA/MDR**: We use T5-large/LLaMA-7B/MDR as the LM to guide the graph traversal
585 respectively. For content-based questions, similar to MDR, we perform a 2-hop retrieval but for
586 each hop, we only search the node to visit next from neighbor candidates. In the 1st-hop retrieval,
587 we select 10 passages and in 2nd-hop retrieval, we select 3 passages, which totally forms 30
588 reasoning paths. Note that passages in the 1st-hop retrieval are allowed to overlap with the ones in
589 the 2nd-hop retrieval. For structural-based questions, we first use ChatGPT to extract page/table

⁶<https://github.com/Lightning-AI/lit-llama>

⁷<https://shivanandroy.com/fine-tune-t5-transformer-with-pytorch/>

⁸We use Levenshtein-distance to measure the lexical distance between two passages.

590 structures and then fetch relevant contents in those structures. Future work could explore how to
 591 pre-train a structural extraction model to obtain document structures.

592 • **KGP w/o LM:** We remove the LM-guided graph traversal but select passages nodes based on their
 593 TF-IDF similarity to the given question.

594 Note that we put the prompt template for running all the above baselines in Section A.9.

595 A.3 Algorithm and Complexity for KGP

596 Here we present the algorithm for our proposed knowledge graph prompting (KGP) method for
 597 MD-QA. Given a question, we first apply LLM to classify whether the question is asking about
 598 the document structure or document content. If the question focuses on the document structure, we
 599 extract the structural keywords such as Page or Table, and retrieve the content in the corresponding
 600 structural nodes in KG. If the question focuses on the document content, we follow the step according
 601 to Algorithm 1. Specifically, we first initialize seeding passages \mathcal{V}^s and the reasoning path queue \mathcal{P}
 602 by TF-IDF search. Then for each seeding passage $v_i \in \mathcal{V}^s$, we add its neighboring passage nodes
 603 \mathcal{N}_i into the candidate neighbor queue \mathcal{C} . (lines 1-4) After that, we iteratively pop out the leftmost
 604 reasoning path/candidate neighborhood $\mathcal{P}_i/\mathcal{C}_i$ from \mathcal{P}/\mathcal{C} and employ the fine-tuned LM-guided graph
 605 traverser to rank the popped out neighbors in \mathcal{C}_i by Eq. (1) (lines 5-7). Last, we select top-k passage
 606 nodes \mathcal{V}'_i from \mathcal{C}_i to visit next based on their rank and correspondingly update the candidate neighbor
 607 queue/reasoning path queue (lines 8-13). The above process terminates when either the candidate
 608 neighbor queue becomes empty or the prefixed budget K for the retrieved passages is met.

Algorithm 1: Knowledge Graph Prompting Method for Questions on Document Contents

Input: A question q over a set of documents \mathcal{D} , the constructed knowledge Graph $G = \{\mathcal{V}, \mathcal{E}, \mathcal{X}\}$ over \mathcal{D} ,
 the fine-tuned LLM-guided graph traversal f_{GT} , the preset context budget K , the initial TF-IDF
 search function g .

- 1 Initialize seed passages $\mathcal{V}^s = g(\mathcal{V}, \mathcal{X}, q)$
- 2 Initialize the retrieved passage queue $\mathcal{P} = [\{v_i\} | v_i \in \mathcal{V}^s]$
- 3 Initialize the candidate neighbor queue $\mathcal{C} = [\mathcal{N}_i | v_i \in \mathcal{V}^s]$
- 4 Initialize the retrieved passage counter $k = \sum_{\mathcal{P}_i \in \mathcal{P}} |\mathcal{P}_i|$
- 5 **while** queue \mathcal{P} and queue \mathcal{C} are not empty **do**
- 6 $\mathcal{P}_i \leftarrow \mathcal{P}.dequeue(), \mathcal{C}_i \leftarrow \mathcal{C}.dequeue()$
- 7 $\mathcal{V}'_i = \text{Graph Traversal}(\{q\} \cup \mathcal{P}_i, \mathcal{C}_i, k)$ by Eq (1)
- 8 **for** $v \in \mathcal{V}'_i$ **do**
- 9 $\mathcal{P}.enqueue(\mathcal{P}_i \cup \{v\})$
- 10 $\mathcal{C}.enqueue(\mathcal{N}_v)$
- 11 $k \leftarrow k + 1$
- 12 **if** $k > K$ **then**
- 13 **Terminate**
- 14 **return** Retrieved Passage Queue \mathcal{P}

609 Since our algorithm can be essentially deemed as the combination of the neighborhood ranking by
 610 Eq. (1) and the breadth-first-search. The time complexity would be the multiplication between the
 611 time of bread-first-search $\mathcal{O}(|\mathcal{V}| + |\mathcal{E}|)$ and the time of neighborhood ranking $\mathcal{O}(|\mathcal{N}|\gamma) = \mathcal{O}(\hat{d}\gamma)$
 612 where γ is the time for computing the embedding similarity between a specific neighbor passage
 613 and the retrieved reasoning path and \hat{d} is the average degree of the KG. Therefore the final time
 614 complexity would be $\mathcal{O}((|\mathcal{V}| + |\mathcal{E}|)\hat{d}\gamma)$, which is in-between the linear and quadratic to the size of the
 615 graph. As users typically maintain 10-100 documents, correspondingly the number of nodes in the
 616 constructed KG would be around 1,000-10,000 (according to Table 4, a collection of 12 documents
 617 have roughly 200-1000 passage nodes), which is affordable even with the quadratic time complexity.
 618 Moreover, we can apply advanced techniques to further reduce the time complexity for neighborhood
 619 ranking, such as KD-tree [33] and LSH [12].

620 For space complexity, it takes $\mathcal{O}(|\mathcal{V}|(\alpha + \beta))$ to maintain the constructed KG on the fly where α is
 621 the average space for saving the passage embedding vector while β is the average space for saving
 622 the textual information of that passage. Although our constructed KG treats passages as nodes, which
 623 cannot scale very well when the graph is extremely large, the total number of documents a user
 624 maintains in a folder is typically around 10-100, which is still affordable.

625 **A.4 Markdown-Formatted Table**

626 Figure 8 demonstrates that by sending Tables in the markdown format, ChatGPT can successfully
 627 understand their content and perform information retrieval based on the given questions. However,
 628 we do observe that such a markdown-formatted solution is not feasible for the long table due to the
 629 input token limitation of ChatGPT, we plan to explore the solution using SQL as the prompt content
 or modeling the Table as the grid graph to solve the issue in the future.

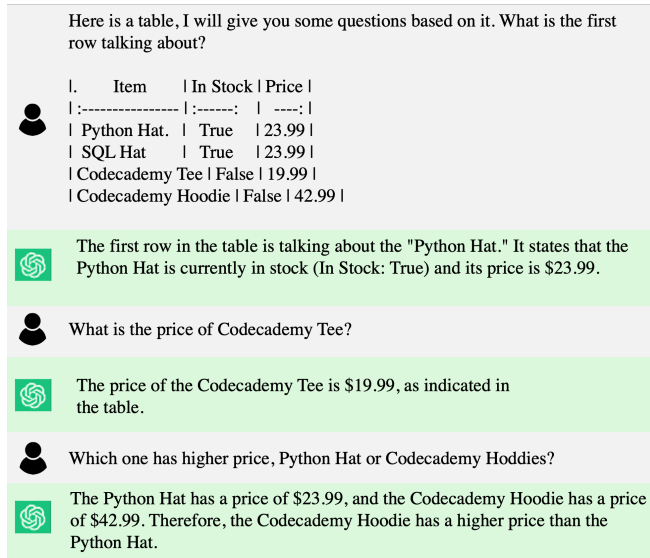


Figure 8: An example demonstrating that ChatGPT can understand table in the markdown format.

630

631 **A.5 Knowledge Graph Construction Comparison**

632 Table 5 compares different knowledge graph construction methods and their pros and cons.

Table 5: Systematically Comparison among existing and our proposed Knowledge Graphs.

KG	Node	Edge	Domain	Constructor	Scalability	Hyperparameters	Advantage	Disadvantage
TAGME	Passage	Common Wikipedia Entity	Wikipedia	/	No	Prior Threshold	Effectively Identify Wikipedia Entities	Low efficiency for Entity Identification Narrow Domain Application
TF-IDF	Passage	Common Keyword	General	/	No	# Keywords	No Domain Limitation	Common keywords irrelevant to question
KNN-ST	Passage	Semantic Similarity	General	Sentence Transformer	No	# Neighbors	No Domain Limitation	Semantic Similarity irrelevant to question
KNN-MDR	Passage	Semantic Similarity	General	MDR	No	# Neighbors	Encoding the logical association for QA	Require logically ordered supporting facts to pre-train the model
Knowledge Base	Entity	Relationship	Specific	Human	Yes	/	Powerful in encoding factual information	Relation Extraction is non-trivial Domain Specific

- 633 • **TAGME:** TAGME [11] is very effective in extracting Wikipedia Entities from a passage despite
 634 the low efficiency. In our graph construction, it usually takes more than 8 hours to extract entities
 635 of all passages for even just 12 Wikipedia documents. Even after we apply parallel processing,
 636 it still takes more than 2 hours. In addition, it can only handle entities mentioned in the existing
 637 Wikipedia system and hence cannot generalize to documents from other domains.
- 638 • **TF-IDF and KNN-ST:** Although there is no domain limitation, it is hard to guarantee the extracted
 639 keywords or the embedding semantic similarity can precisely encode the relationships that are
 640 desired for answering the given question between any two passages. We empirically find TF-IDF is
 641 more likely to extract meaningless keywords even after removing supporting verbs and articles.

- 642 • **KNN-MDR**: Since KNN-MDR pre-trains the sentence encoder by predicting the next supporting
643 passage given already-retrieved passages, the embedding similarity between two passages is more
644 likely to encode necessary logical associations required for MD-QA. However, the main bottleneck
645 here is how to obtain the logically ordered supporting facts that can progressively reach the
646 answer. Obtaining these sequential data is non-trivial and usually requires a large number of human
647 resources for well-curated annotation.
- 648 • **Existing Knowledge Base**: One common approach in the literature is to use existing knowledge
649 bases or extract subgraphs from them for specific tasks [8, 50, 51]. Because the factual information
650 is characterized as a triplet consisting of two entity nodes and their relationship, it is very powerful
651 in encoding factual information/commonsense knowledge and also avoids the scalability issue
652 (since two different passages might share the same entity). Despite its potency and ease of
653 use, constructing this type of KGs demands meticulously designed relation extractors, which is
654 still deemed a challenging task in the literature. Recent research has explored using LLMs for
655 relation extraction. However, with increasing document numbers, using non-open-sourced LLMs
656 can become prohibitively expensive. A potential solution is fine-tuning an open-sourced LLM
657 specifically for relation extraction. Detailed discussion on this is beyond the scope of this study
658 and is thus omitted.

659 To put it in a nutshell, there’s no one-size-fits-all method for KG construction. Our paper offers
660 an in-depth analysis of the proposed KG construction methods alongside other existing ones. The
661 best approach often depends on the specific use case. For broad domains containing general factual
662 information, tools like 'TAGME' or 'Knowledge Base' might be apt. However, for more niche or
663 sensitive areas, methods like TF-IDF/KNN-ST are more appropriate. In certain situations, gathering
664 domain-specific data and pre-training encoders is the most effective way to build the KG.

665 A.6 Additional Results and Discussions

666 A.6.1 Quality of KG on MuSiQue

667 Similar to the setting used for Figure 4, we change the hyperparameters to construct KGs for each
668 question in MuSiQue with varying levels of sparsity and measure how much percentage of the
669 supporting facts are covered by neighbors of the seeding passages that are initially retrieved by
670 TF-IDF. The general trend in Figure 9(a) is similar to the one in Figure 4, i.e., as the graph becomes
671 denser, the precision decreases while the SF-EM increases. However, on MuSiQue, KNN-MDR
672 achieves the worst trade-off between Precision and SF-EM compared with KNN-ST and TF-IDF.
673 This is because our KNN-MDR is pre-trained on HotpotQA and due to the distribution shift from
674 HotpotQA to MuSiQue, it is expected for the graph constructed with KNN-MDR to have less quality.
675 Note that although here KNN-ST leads to a better KG than KNN-MDR, it does not mean the KNN
676 baseline in Table 2 should perform better than MDR because the baseline name only refers to the
677 retrieval method while the name in this figure refers to the KG construction method.

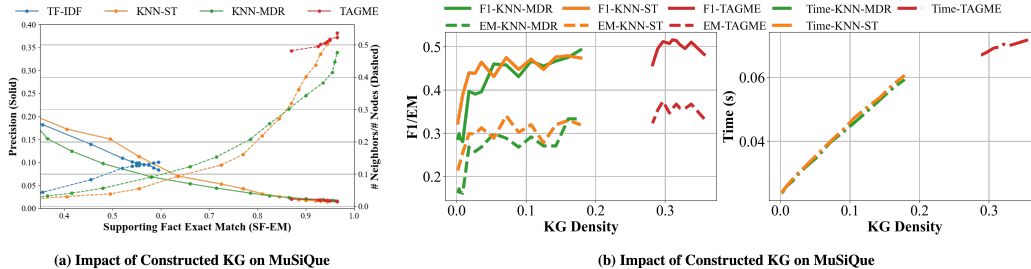


Figure 9: (a): Quality of constructed KGs with different methods on MuSiQue. **TF-IDF**: lexical similarity based on common keywords extracted by TF-IDF. **KNN-ST**: KNN graph constructed based semantic similarity of embeddings from sentence-transformer; **KNN-MDR**: KNN graph constructed based on semantic similarity of embeddings from the pre-trained MDR [44]; **TAGME**: graph constructed based on whether two passages share common Wikipedia entity mentions. (b): The performance/latency increases as the KG density increases. The results are averaged across 100 randomly sampled questions on MuSiQue.

678 **A.6.2 The impact of KG on MuSiQue**

679 Similar to the setting used for Figure 6, we compare the MD-QA performance for KGP-T5 using
680 TAGME-based KG with different levels of density. Similar to Figure 6, here we also observe that as
681 the KG becomes denser, the MD-QA performance increases while the time for the next node search
682 increases. However, on MuSiQue, in most cases, KNN-ST achieves better F1/EM than KNN-MDR,
683 which exactly aligns with the constructed KG quality observed in Figure 9, i.e., KNN-ST achieves
684 better Precision/SF-EM trade-off than KNN-MDR on MuSiQue.

685 **A.7 Case study on Structural/Content Questions**

686 In this section, we conduct six MD-QA case studies using our self-designed user interface coupled
 687 with the proposed method on the backend. Examples include two table-based QA (Figure 10-11),
 688 one page-based QA (Figure 12), one single-document content-based QA (Figure 13) and two multi-
 689 document content-based QA (Figure 14-17). In our designed interface, we can upload documents we
 690 are interested in reading and the model on the backend will split each of them into multiple passages.
 691 In addition, on the left side, we can ask questions related to the currently uploaded documents. By
 692 clicking the button 'SUBMIT', the question would be sent to the model on the backend and it retrieves
 693 relevant context and arranges them as the prompt to get the answer from ChatGPT. In the figures
 694 below, we can see our system can understand the Table/Page questions and also questions requiring
 695 knowledge across multiple documents.

Documental question

Based on Table on Page 4, how many people belong to the membership grade 'Fellow'?

Answer

There are 24 people who belong to the membership grade 'Fellow'.

Evidence

Page 4/Table 1:

Membership Grade	Number of People	Section Annual Dues
Fellow	24	\$25
Member	153	\$25
Institute Affiliate	8	\$25
Student Member	47	\$0*
Esteemed Colleague	6	\$0
TOTAL	238	

SUBMIT

Documental question

Oregon Section 2007/2008 Annual Report

DR--18286691

2. MEMBERSHIP

The following table provides a summary of the Oregon Section's membership, including grade and section dues.

Membership Grade	Number of People	Section Annual Dues
Fellow	24	\$25
Member	153	\$25
Institute Affiliate	8	\$25
Student Member	47	\$0*
Esteemed Colleague	6	\$0
TOTAL	238	

*Note: Students also receive a substantial discount on meeting registration fees.

Membership Committee:

The membership committee continued its organizational captain membership program over the past year. The program was designed to identify specific individuals within public and private organizations who might be interested in joining ITE or attending ITE functions. The captains also assist with advertising meetings, recruiting new members, and promoting ITE.

Institute of Transportation Engineers

Page 4 of 13

UPLOAD PDF

Figure 10: Table QA asking for the number of people belonging to the membership grade 'Fellow'. It is shown that ChatGPT can understand table structure in the format of markdown and successfully fetch the number of people belonging to membership 'Fellow'.

Documental question

According to Table on Page 7, where will the event occur on Date 5-18-07?

Answer

The event on Date 5-18-07 will occur at the Crowne Plaza Hotel in Lake Oswego, OR.

Evidence

Page 7/Table 1:

Date	Meeting	Attendees	Subject
4-10-07	April - Joint meeting with WTS		Kris Strickler and Danielle Cogan gave a presentation on the Columbia River Crossing project on the I-5 bridge that connects Oregon and Washington.
5-18-07	May - Joint meeting with Oregon Traffic Control Districts Conference (OTCCDC)	160	Sections newly elected officers were announced and a presentation on the Highway 25 emergency reconstruction efforts that followed a large storm on Mt. Hood early 2007.
5-18-07	May 2007 ITE - 2007 ITE District 6 Annual Meeting & Qual Conference	200	2007 ITE District 6 Annual Meeting & Qual Conference in Portland. This year's annual meeting broke many previous records with over 500 meeting participants at the meeting.
9-21-07	2007 Golf Tournament	57	Due to timing of ITE District 6 conference in Portland, this year's golf tournament was held in late summer. We had good turn out and good corporate sponsorship donations.
9-25-07	September Meeting	56	Jim Fries (ORS Associates) and Jay McCoy (City of Clatskanie) gave a presentation on the use of recent SCATS Adaptive Traffic Signal System at City of Clatskanie, Oregon.
10-23-07	October Meeting	73	Transportation committee discussed annual competition judging and his 'Moving Oregon' award nominations and efforts to build support for major investment in Oregon's transportation system.
11-15-07	2007 Student Traffic Bowl	125 including 52 students	Oregon ITE 10 th Annual Student Traffic Bowl competition featured an innovative team from the northwest. This year's 1 st place prize went to University of Portland with University of Washington and Oregon Institute of Technology (OIT) took 2 nd place.

4-10-07 | April - Joint meeting with WTS | Kris Strickler and Danielle Cogan gave a presentation on the Columbia River Crossing project on the I-5 bridge that connects Oregon and Washington. | Embassy Suites Hotel Portland, OR | 160 |

SUBMIT

Documental question

Oregon Section 2007/2008 Annual Report

DR--18286691

5. MEETINGS

The Oregon Section conducted six (6) general meetings and the summer golf tournament over the past year, as well as hosting Joint 2007 ITE District 6 annual meeting. The general meetings included luncheons with speakers, joint meetings with other professional societies, the annual traffic bowl, and a technical workshop. The table below summarizes the general meetings conducted over the past year and those scheduled for the remainder of 2008.

Date	Meeting	Subject	Location	Attendees
4-10-07	April - Joint meeting with WTS	Kris Strickler and Danielle Cogan gave a presentation on the Columbia River Crossing project on the I-5 bridge that connects Oregon and Washington.	Embassy Suites Hotel Portland, OR	60
5-18-07	May - Joint meeting with Oregon Traffic Control Districts Conference (OTCCDC)	Sections newly elected officers were announced and a presentation on the Highway 25 emergency reconstruction efforts that followed a large storm on Mt. Hood early 2007.	Crowne Plaza Hotel Lake Oswego, OR	45
5-18-07	May 2007 ITE - 2007 ITE District 6 Annual Meeting & Qual Conference	2007 ITE District 6 Annual Meeting & Qual Conference in Portland. This year's annual meeting broke many previous records with over 500 meeting participants at the meeting.	Hyatt Regency Portland, OR	200
9-21-07	2007 Golf Tournament	Due to timing of ITE District 6 conference in Portland, this year's golf tournament was held in late summer. We had good turn out and good corporate sponsorship donations.	Oregon Golf Association (OGA) Golf Course, Woodburn, OR	57
9-25-07	September Meeting	Jim Fries (ORS Associates) and Jay McCoy (City of Clatskanie) gave a presentation on the use of recent SCATS Adaptive Traffic Signal System at City of Clatskanie, Oregon.	Kelly Inn Park Portland, OR	56
10-23-07	October Meeting	Transportation committee discussed annual competition judging and his 'Moving Oregon' award nominations and efforts to build support for major investment in Oregon's transportation system.	Hotel Monaco, Portland, OR	73
11-15-07	2007 Student Traffic Bowl	Oregon ITE 10 th Annual Student Traffic Bowl competition featured an innovative team from the northwest. This year's 1 st place prize went to University of Portland with University of Washington and Oregon Institute of Technology (OIT) took 2 nd place.	McMenamins Highfield Troutdale, OR	125 including 52 students

Institute of Transportation Engineers

Page 7 of 13

UPLOAD PDF

Figure 11: Table QA asking for the place where the event on Date 5-18-07 will occur.

What is main content on Page 2?

Open Access

PARTICIPANTS AND METHODS

Participants
Sixteen participants (nine women and seven men) were included with a mean (sD) age, 23±5 years and body mass index (BMI), 26±5.5 kg/m² (↔)(table 1).
Description of chairs Standard office chair (control chair): The criterion model chair is a standard office chair (Steelcase, Grand Rapids, Michigan, USA).
FootFidget (↔)(http://footfidget.com)↔(http://footfidget.com) (FootFidget, Lake Zurich, Illinois, USA) (↔)(figure 1), is an under-desk elasticated footrest that encourages leg activity while seated. It comprises of a steel 17"×10"×10" frame support base. The elasticated central footpad consists of a 7" foam-covered cylindrical rigid tube centred on two 17" flex-ible resistance cords that run through the tube and attach to the four upright legs on the stand. The user repeatedly "bounces" their foot on the cylindrical tube that encourages resistance.
CoreChair (↔)(https://www.corechair.com)↔(https://www.corechair.com) (CoreChair, Aurora, Ontario, Canada) (↔)(figure 1), is a chair designed to promote activity while a person stays seated. It is a modified five-wheel office chair. It has a low, 9" backrest with adjustable depth and is without armrests. The seat is sculpted and covered in 2" thick foam padding. The main feature of the CoreChair is the mechanical core that allows for lateral movement while seated. Severity of seat tilt is adjustable and has a range of motion up to 14° in all directions. For the study trial, tilt severity was set to allow for the greatest range of motion. Interchangeable

Participants then continued their work-life activities and were provided with the FootFidget. Energy expenditure and heart rate were measured for 20 min. Participants then sat on the CoreChair during which their energy expenditure and heart rate were measured for 20 min. Participants continued sitting on the CoreChair, stopped their work-life activities and followed a 7 min chair-based exercise video. The participant was given a 7 min break. Subsequently, the participant followed the 7 min video for a second sample. Finally, participants walked at 2 mph for 20 min on a calibrated treadmill (Pacemaker Business, Aerotech, Westborough, New Jersey, USA).

Energy expenditure
Energy expenditure was measured using indirect calorimetry (Metamax 3B, Cosmed, Legnano, Germany). The calorimeter was calibrated using 3.0% O₂, 13.0% O₂, balance nitrogen (Praxair, Danbury, Connecticut, USA) and ambient air according to the manufacturer's specifications. In addition, a wet volume calibrated before each participant using a 3 L syringe. The calorimeter collects breath-by-breath O₂ and O₂ production and consumption, respectively, and energy expenditure is calculated using standard formulas.¹⁸

Heart rate monitoring
Participants were also fitted with a Polar Heart Rate Monitor (P7, Polar, Lake Success, New York, USA). Heart rate samples were recorded and synchronized for each breath.

RESULTS
Energy expenditure of the four seated conditions and also walking (2 mph) are shown in figure 2. While sitting in the standard office chair, as expected, resting energy expenditure sitting in a standard chair is lowest a positive correlation with body weight (r=0.53, p<0.05). The relationship was described by the equation: resting energy expenditure (kcal/hour)=0.97×weight (kg) +0.07. Energy expenditure increased significantly while using the FootFidget (↔)(p=0.02) when compared to the standard office chair. Energy expenditure increased in all participants from a mean of 26.1 to 26.4 kcal/hour (p<0.001). Heart rate did not increase significantly, however (73.0 to 76.1 bpm). Similarly, resting energy expenditure increased significantly while using the

DR-182866691
DR-1058108

Submit

Upload PDF

Figure 12: Page QA asking the main content on Page 2. The answer provides a high-level summarization of Page 2, covering the title of each section.

What is associated with chronic health conditions and impair cognitive function and obesity?

Open Access

BMJ Open Sport & Exercise Medicine

Chair-based fidgeting and energy expenditure

Gabriel A Koop,¹ Graham K Moore,² James A Levine^{1,2}

ABSTRACT
Sedentariness is associated with chronic health conditions, impaired cognitive function, and obesity. Breaking up sitting time with standing or walking while working can effectively decrease sedentariness and improve insulin sensitivity and lipids. Solutions to promote physical activity are necessary to reverse sedentariness and prevent chronic diseases.

Introduction
Sedentariness is associated with a myriad of chronic diseases, impaired cognition (↔) and obesity (↔). The mechanism by which sitting excessively causes disease is not well understood, but it is known that breaking up sitting improves insulin sensitivity and lipids. Several studies have examined the effectiveness of programmes to displace sitting with standing or walking while working. These measures can effectively decrease sitting time and improve productivity although their

Results
Energy expenditure increased significantly while using either an under-desk footpad for a 7 min fidgeting period (↔)(p=0.02) when compared to the standard office chair (control chair). 73.0 to 76.1 bpm heart rate did not increase significantly (↔)(p=0.02). However, heart rate did increase significantly in all participants (↔)(p=0.001). Energy expenditure increased significantly while using the FootFidget (↔)(p=0.02) when compared to the standard office chair. Energy expenditure increased in all participants from a mean of 26.1 to 26.4 kcal/hour (p<0.001). Heart rate did not increase significantly, however (73.0 to 76.1 bpm). Similarly, resting energy expenditure increased significantly while using the

Conclusion
Chairs and devices that promote fidgeting can increase energy expenditure by ~20–25% and not increase heart rate. Dynamic sitting may be a simple, a means of options to help people move more while at work.

Introduction
Sedentariness is associated with a myriad of chronic diseases, impaired cognitive and obesity.^{1–3} The mechanism by which sitting excessively causes disease is not well understood, but it is known that breaking up sitting improves insulin sensitivity and lipids. Several studies have examined the effectiveness of programmes to displace sitting with standing or walking while working.^{4–7} These measures can effectively decrease sitting time and improve productivity,^{8–10} although their

DR-182866691
DR-1058108

Submit

Upload PDF

Figure 13: Single Document Content QA asking Sedentariness. The 2nd retrieved sentence includes the answer and corresponds to the first sentence in the abstract of the paper.

Is the state where LeBron James grew up playing basketball rank the 34th-largest by area in US?

Answer

Yes, the state where LeBron James grew up playing basketball, Ohio, is the 34th-largest by area in the US.

Evidence

1: : 23James began playing organized basketball in the fifth grade. He later played Amateur Athletic Union (AAU) basketball for the Northeast Ohio Shooting Stars.

2: Ohio (/oʊˈhaɪoʊ/ (listen)) is a state in the Midwestern United States. Of the fifty U.S. states, it is the 34th-largest by area. With a population of nearly 11.8 million, Ohio is the seventh-most populous and tenth-most densely populated state.

3: James grew up playing basketball for St. Vincent–St. Mary High School in his hometown of Akron, Ohio. He was heavily touted by the national media as a future NBA superstar for his all-around scoring, passing, athleticism and playmaking abilities.

4: As a 6-foot-2-inch (1.88 m) tall freshman, James averaged 21 points and 6 rebounds per game for the St. Vincent–St. Mary varsity basketball team.

5: : 117 St. Vincent–St. Mary finished the year with a 23–4 record, ending their season with a loss in the Division II championship game.

6: Ohio's three largest cities are Columbus, Cleveland, and Cincinnati, all three of which anchor major metropolitan areas. Columbus is the capital of the state, located near its geographic center and is well known for Ohio State University.

WIKIPEDIA The Free Encyclopedia

LeBron James

LeBron Raymone James Sr. (/lɛɪbrɒn/ /lɛˈbroʊ/; born December 30, 1984), also known as **LEJ**, is an American professional basketball player for the Los Angeles Lakers of the National Basketball Association (NBA). Nicknamed "King James", he is widely regarded as one of the greatest players in the history of the sport and is often compared to Michael Jordan in debates over the greatest basketball player of all time.^[f] James is the all-time leading scorer in NBA history and ranks fourth in career assists. He has won four NBA championships (two with the Miami Heat, one each with the Lakers and Cleveland Cavaliers), and has competed in 10 NBA Finals.^[g] He has also won four Most Valuable Player (MVP) Awards, four Finals MVP Awards, and two Olympic gold medals, and has been named an All-Star 16 times, selected to the All-NBA Team 19 times (including 11 First Team selections^[h]) and the All-Defensive Team six times, and was a runner-up for the NBA Defensive Player of the Year Award twice in his career.^{[i][j]}

James grew up playing basketball for St. Vincent–St. Mary High School in his hometown of Akron, Ohio. He was heavily touted by the national media as a future NBA superstar for his all-around scoring, passing, athleticism and playmaking abilities.^[k] A prep-tee, he was selected by the Cleveland Cavaliers with the first overall pick of the 2003 NBA draft. Named the 2004 NBA Rookie of the Year,^[l] he soon established himself as one of the league's premier players, leading the Cavaliers to their first NBA Finals appearance in 2007 and winning the NBA MVP award in 2009 and 2010.^[l] After failing to win a championship with Cleveland, James left in 2010 as a free agent to join the Miami Heat.^[m] This was announced in a nationally televised special titled *The Decision* and is among the most controversial free agency moves in sports history.^[n]

James won his first two NBA championships while playing for the Heat in 2012 and 2013; in both of those years, he also earned the league's MVP and Finals MVP awards. After his fourth season with the Heat in 2014, James opted out of his contract and re-signed with the Cavaliers. In 2016, he led the Cavaliers to victory over the Golden State Warriors in the Finals by coming back from a 3–1 deficit, delivering the team's first championship and ending the Cleveland sports curse.^[o] In 2018, James exercised his contract option to leave the Cavaliers and signed with the Lakers, where he won the 2020 NBA championship and his fourth Finals MVP.^[p] James is the first player in NBA history to accumulate \$1 billion in

James with the Los Angeles Lakers in 2022

No. 33 – Los Angeles Lakers

Position Small forward / power forward

League NBA

Personal information

Born December 30, 1984
Akron, Ohio, U.S.

Listed height 6 ft 9 in (2.06 m)

Listed weight 250 lb (113 kg)

Center information

High school St. Vincent–St. Mary (Akron, Ohio)

NBA draft 2003: 1st round, 1st overall pick

Selected by the Cleveland Cavaliers

Playing career 2003–present

Career history

DR-182866691

DR-1058108

LeBron James

Ohio

Michael Jordan

SUBMIT

UPLOAD PDF

Figure 14: Multi-document Bridging Question asking the information about Lebron James and State Ohio. It requires to first retrieve the sentence stating the state where Lebron James grew up playing basketball.

Is the state where LeBron James grew up playing basketball rank the 34th-largest by area in US?

Answer

Yes, the state where LeBron James grew up playing basketball, Ohio, is the 34th-largest by area in the US.

Evidence

1: : 23James began playing organized basketball in the fifth grade. He later played Amateur Athletic Union (AAU) basketball for the Northeast Ohio Shooting Stars.

2: Ohio (/oʊˈhaɪoʊ/ (listen)) is a state in the Midwestern United States. Of the fifty U.S. states, it is the 34th-largest by area. With a population of nearly 11.8 million, Ohio is the seventh-most populous and tenth-most densely populated state.

3: James grew up playing basketball for St. Vincent–St. Mary High School in his hometown of Akron, Ohio. He was heavily touted by the national media as a future NBA superstar for his all-around scoring, passing, athleticism and playmaking abilities.

4: As a 6-foot-2-inch (1.88 m) tall freshman, James averaged 21 points and 6 rebounds per game for the St. Vincent–St. Mary varsity basketball team.

5: : 117 St. Vincent–St. Mary finished the year with a 23–4 record, ending their season with a loss in the Division II championship game.

6: Ohio's three largest cities are Columbus, Cleveland, and Cincinnati, all three of which anchor major metropolitan areas. Columbus is the capital of the state, located near its geographic center and is well known for Ohio State University.

WIKIPEDIA The Free Encyclopedia

Ohio

Coordinates: 40°53′N 83°0′W﻿ / ﻿40.883°N 83.0°W﻿ / 40.883; -83.0

Ohio (/oʊˈhaɪoʊ/ (listen)) is a state in the Midwestern United States and the 17th largest by area, and the 34th largest by population, in the United States. It is the 34th-largest by area. With a population of nearly 11.8 million, Ohio is the seventh-most populous and tenth-most densely populated state. Its capital and largest city is Columbus, with other large population centers including Cleveland, Cincinnati, Dayton, Akron, and Toledo. Ohio is bordered by Lake Erie to the north, Pennsylvania to the east, West Virginia to the southeast, Kentucky to the southwest, Indiana to the west, and Michigan to the northwest. Ohio is nicknamed the "Buckeye State" after its Ohio buckeye trees, and Ohioans are also known as "Buckeyes".^[f] Its state flag is the only non-rectangular flag of all the U.S. states.

Ohio takes its name from the Ohio River, which, in turn, originated from the Seneca word *ohio*, meaning "good river", "great river", or "large creek".^{[g][h]} The state arose from the lands west of the Appalachian Mountains that were ceded from colonial times through the Northwest Indian Wars of the late 18th century. It was partitioned from the resulting Northwest Territory, which was the first frontier of the new United States, becoming the 17th state admitted to the Union on March 1, 1803, and the first under the Northwest Ordinance.^{[i][j]} Ohio was the first post-colonial free state admitted to the union and became one of the earliest and most influential industrial powerhouses during the 19th century. Although it has transitioned to a more information- and service-based economy in the 21st century, it remains an industrial state, ranking seventh in GDP as of 2023,^[k] with the third-largest manufacturing sector and second-largest automobile production.^[l]

The government of Ohio is composed of the executive branch, led by the governor, the legislative branch, consisting of the bicameral Ohio General Assembly, and the judicial branch, led by the state Supreme Court. Ohio occupies 16 seats in the United States House of Representatives.^[m] The state is known for its status as both a swing state and a bellwether in national

Ohio

State

State of Ohio

Flag

Seal

Nickname(s) The Buckeye State
Bridgeway of America; The Heart of It All
Motto: "With God, all things are possible"^[n]
Anthem: "Beautiful Ohio"^[o]

Map of the United States with Ohio highlighted

Country United States

Admitted to the Union March 1, 1803^[p] (17th, declared retroactively on August 7, 1939^[q])

Capital (and largest city) Columbus^[r]

Largest metro and urban areas Greater Cleveland (combined and urban)
Cincinnati (metro)
Columbus (metro)
Toledo (metro)^[s]

Government

DR-182866691

DR-1058108

LeBron James

Ohio

Michael Jordan

SUBMIT

UPLOAD PDF

Figure 15: Multi-document Bridging Question asking the information about Lebron James and State Ohio. Then it requires to judge whether the State Ohio ranks the 34th-largest by area in the US.

Who borns later, Michael Jordan or LeBron James?

Answer

LeBron James was born later than Michael Jordan, on December 30, 1984, while Jordan was born on February 17, 1963.

Evidence

- Former NBA players, including Michael Jordan and Magic Johnson, were also critical of James, condemning him for joining with Bosh and Wade in Miami and not trying to win a championship as "the man".
- "James is also discussed within the context of being the greatest basketball player of all-time, which has resulted in frequent comparisons to Michael Jordan.[a]
- Michael Jeffrey Jordan (born February 17, 1963), also known by his initials MJ, is an American former professional basketball player and businessman.
- In a 2016 interview with Sports Illustrated, James acknowledged that his motivation was surpassing Jordan as the greatest.
- He also joined Michael Jordan as the only players to win an NBA MVP award, NBA championship, NBA Finals MVP, and Olympic gold medal in the same year.
- LeBron Raymonte James Sr. (/ləˈbrɒn/ lə-ˈBRON; born December 30, 1984), also known as LBJ, is an American professional basketball player for the Los Angeles Lakers of the National Basketball Association (NBA).

WIKIPEDIA The Free Encyclopedia

Michael Jordan

Michael Jeffrey Jordan (born February 17, 1963), also known by his initials MJ,^[a] is an American former professional basketball player and businessman. The official National Basketball Association (NBA) website states: "By acclamation, Michael Jordan is the greatest basketball player of all time."^[b] He played fifteen seasons in the NBA, winning six NBA championships with the Chicago Bulls. He was integral in popularizing the sport of basketball and the NBA around the world in the 1980s and 1990s,^[c] becoming a global cultural icon.^[d]

Jordan played college basketball for three seasons under coach Dean Smith with the North Carolina Tar Heels. As a freshman, he was a member of the Tar Heels' national championship team in 1982.^[e] Jordan joined the Bulls in 1984 as the third overall draft pick^[f] and quickly emerged as a league star, entertaining crowds with his prolific scoring while gaining a reputation as one of the game's best defensive players.^[g] His leaping ability, demonstrated by performing slam dunks from the free-throw line in Slam Dunk Contests, earned him the nicknames "Air Jordan" and "His Airness."^[h] Jordan won his first NBA title with the Bulls in 1991 and followed that achievement with titles in 1992 and 1993, securing a three-peat. Jordan abruptly retired from basketball before the 1993–94 NBA season to play Minor League Baseball but returned to the Bulls in March 1995 and led them to three more championships in 1996, 1997, and 1998, as well as a three-round 72 regular season wins in the 1995–96 NBA season.^[i] He retired for the second time in January 1999 but returned for two more NBA seasons from 2001 to 2003 as a member of the Washington Wizards.^{[j][k]} During the course of his professional career, he was also selected to play for the United States national team, winning four gold medals—the 1992 Pan American Games, 1984 Summer Olympics, 1992 Tournament of the Americas and 1992 Summer Olympics—while also being undefeated.^[l]

Jordan's individual accolades and accomplishments include six NBA Finals Most Valuable Player (MVP) awards, ten NBA scoring titles (both all-time records), five NBA MVP awards, ten All-NBA First Team designations, nine All-Defensive First Team honors, fourteen NBA All-Star Game selections, three NBA All-Star Game MVP awards, three NBA steals titles, and the 1988 NBA Defensive Player of the Year Award.^[m] He holds the NBA records for career regular season scoring average (30.1 points per game) and career playoff scoring average (32.4 points per game)^[n] in 1990. He was named the sixth century's greatest American athlete by ESPN and was second to Babe Ruth on the Associated Press' list of athletes of the century.^[o] Jordan was twice inducted into the Naamath Memorial Basketball Hall of Fame, once in 2009 for his individual career,^[p] and again in 2010 as part of the 1992

Jordan in April 2014

Personal information	
Born	February 17, 1963 New York City, New York, U.S.
Listed	6 ft 6 in (1.98 m)
Height	
Listed	216 lb (98 kg) ^[q]
Weight	
Career information	
High school	Emory A. Laney (Wilmington, North Carolina)
College	North Carolina (1981–1984)
NBA draft	1984: 1st round, 3rd overall pick
Selected by the Chicago Bulls	
Playing career	1984–1993, 1995–1998, 2001–2003
Position	Shooting guard / small forward

https://en.wikipedia.org/wiki/Michael_Jordan

DR--18286691

DR--1058108

LeBron James

Ohio

Michael Jordan

SUBMIT

UPLOAD PDF

Figure 16: Multi-document Question comparing LeBron James and Michael Jordan. It requires the birthday information of LeBron and Jordan.

Who borns later, Michael Jordan or LeBron James?

Answer

LeBron James was born later than Michael Jordan, on December 30, 1984, while Jordan was born on February 17, 1963.

Evidence

- Former NBA players, including Michael Jordan and Magic Johnson, were also critical of James, condemning him for joining with Bosh and Wade in Miami and not trying to win a championship as "the man".
- "James is also discussed within the context of being the greatest basketball player of all-time, which has resulted in frequent comparisons to Michael Jordan.[a]
- Michael Jeffrey Jordan (born February 17, 1963), also known by his initials MJ, is an American former professional basketball player and businessman.
- In a 2016 interview with Sports Illustrated, James acknowledged that his motivation was surpassing Jordan as the greatest.
- He also joined Michael Jordan as the only players to win an NBA MVP award, NBA championship, NBA Finals MVP, and Olympic gold medal in the same year.
- LeBron Raymonte James Sr. (/ləˈbrɒn/ lə-ˈBRON; born December 30, 1984), also known as LBJ, is an American professional basketball player for the Los Angeles Lakers of the National Basketball Association (NBA).

WIKIPEDIA The Free Encyclopedia

LeBron James

LeBron Raymonte James Sr. (/ləˈbrɒn/ lə-ˈBRON; born December 30, 1984), also known as LBJ, is an American professional basketball player for the Los Angeles Lakers of the National Basketball Association (NBA). Nicknamed "King James", he is widely regarded as one of the greatest players in the history of the sport and is often compared to Michael Jordan in debates over the greatest basketball player of all time.^[a] James is the all-time leading scorer in NBA history and ranks fourth in career assists. He has won four NBA championships (two with the Miami Heat, one each with the Lakers and Cleveland Cavaliers), and has competed in 10 NBA Finals.^[b] He has also won four Most Valuable Player (MVP) Awards, four Finals MVP Awards, and two Olympic gold medals, and has been named an All-Star 10 times, selected to the All-NBA Team 10 times (including 11 First Team selections)^[c] and the All-Defensive Team six times, and was a runner-up for the NBA Defensive Player of the Year Award twice in his career.^{[d][e]}

James grew up playing basketball for St. Vincent–St. Mary High School in his hometown of Akron, Ohio. He was heavily touted by the national media as a future NBA superstar for his all-around scoring, passing, athleticism and playmaking abilities.^[f] A prep-tee, he was selected by the Cleveland Cavaliers with the first overall pick of the 2003 NBA draft. Named the 2004 NBA Rookie of the Year,^[g] he soon established himself as one of the league's premier players, leading the Cavaliers to their first NBA Finals appearance in 2007 and winning the NBA MVP award in 2009 and 2010.^[h] After failing to win a championship with Cleveland, James left in 2010 as a free agent to join the Miami Heat.^[i] His was announced in a nationally televised special titled *The Decision* and is among the most controversial free agency moves in sports history.^[j]

James won his first two NBA championships while playing for the Heat in 2012 and 2013; in both of these years, he also earned the league's MVP and Finals MVP awards. After his fourth season with the Heat in 2014, James opted out of his contract and re-signed with the Cavaliers. In 2016, he led the Cavaliers to victory over the Golden State Warriors in the Finals by coming back from a 3–1 deficit, delivering the team's first championship and ending the Cleveland sports curse.^[k] In 2018, James exercised his contract option to leave the Cavaliers and signed with the Lakers, where he won the 2020 NBA championship and his fourth Finals MVP.^[l] James is the first player in NBA history to accumulate \$1 billion in

James with the Los Angeles Lakers in 2022

Personal information	
Full name	LeBron James Sr.
Position	Small forward / power forward
League	NBA
Career information	
Born	December 30, 1984 Akron, Ohio, U.S.
Listed	6 ft 9 in (2.06 m)
Height	
Listed	250 lb (113 kg)
Weight	
Career information	
High school	St. Vincent–St. Mary (Akron, Ohio)
NBA draft	2003: 1st round, 1st overall pick
Selected by the Cleveland Cavaliers	
Playing career	2003–present
Career history	

https://en.wikipedia.org/wiki/LeBron_James

DR--18286691

DR--1058108

LeBron James

Ohio

Michael Jordan

SUBMIT

UPLOAD PDF

Figure 17: Multi-document Question comparing LeBron James and Michael Jordan. It requires the birthday information of LeBron and Jordan.

696 **A.8 Visualizing the Reasoning-and-Retrieving Process of LM-guided Graph Traverser**

697 In this section, we visualize the KG-LLaMA’s reasoning-and-retrieving process in retrieving relevant
 698 context for MD-QA. Due to space limitation, for each question, we visualize the top-3 sentence
 699 nodes visited at 1-hop along with their generated evidence from LLaMA that required further to
 700 approach the answer. Based on the generated evidence, we retrieve the top-2 sentence nodes from the
 701 candidate neighbor queue. For each retrieved sentence node, we also visualize its ranking score given
 702 by TF-IDF. We can clearly see our designed LM-guided graph traversal could find the right evidence
 703 path to answer the given question.

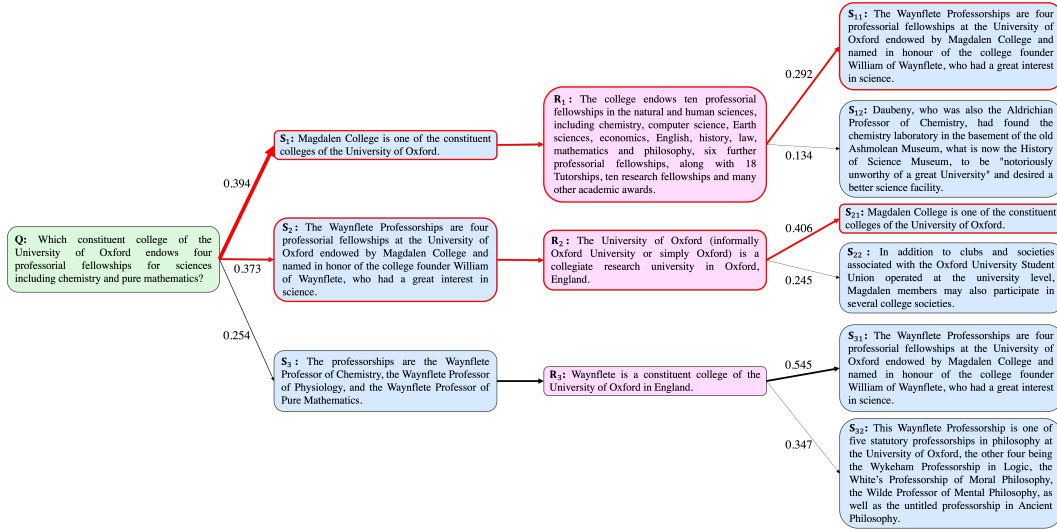


Figure 18: Visualizing the graph traversal over MD-QA-Example 1.

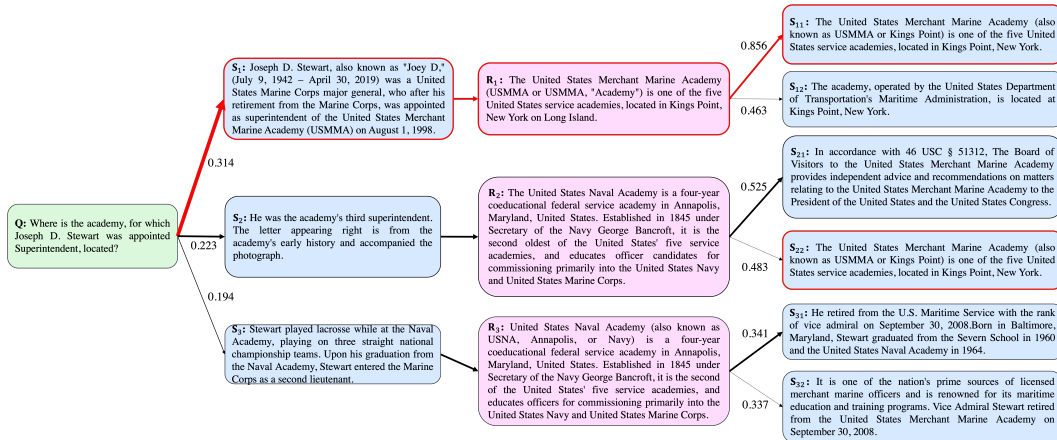


Figure 19: Visualizing the graph traversal over MD-QA-Example 2.

704 **A.9 Prompt template used throughout this work**

Listing 1: Examples of the Instruction Data for Fine-tuning LLaMA.

705 Question: Which magazine was started first Arthur’s Magazine or First for Women?
706 Answer: Arthur’s Magazine
707 Supporting Facts:
708 (1) Arthur’s Magazine (1844–1846) was an American literary periodical published in Philadelphia in the 19th
709 century.
710 (2) First for Women is a woman’s magazine published by Bauer Media Group in the USA. The magazine was
711 started in 1989.
712
713 Instruction: What evidence do we need to answer the question given the current evidence?
714 Input: Which magazine was started first Arthur’s Magazine or First for Women? Arthur’s Magazine
715 (1844–1846) was an American literary periodical published in Philadelphia in the 19th century.
716 Output: First for Women is a woman’s magazine published by Bauer Media Group in the USA. The magazine
717 was started in 1989.
718 =====
719
720 Question: In what year was the creator of the current arrangement of Simpson’s Theme born?
721 Answer: March 28, 1941
722 Supporting Facts:
723 (1) The theme was re–arranged during season 2, and the current arrangement by Alf Clausen was introduced
724 at the beginning of season 3.
725 (2) Alf Heiberg Clausen (born March 28, 1941) is an American film and television composer.
726
727 Instruction: What evidence do we need to answer the question given the current evidence?
728 Input: In what year was the creator of the current arrangement of Simpson’s Theme born? The theme was re–
729 arranged during season 2, and the current arrangement by Alf Clausen was introduced at beginning of
730 season 3.
731 Output: Alf Heiberg Clausen (born March 28, 1941) is an American film and television composer.

Listing 2: Example of the Prompt for QA without Retrieved Contexts.

732 Given the following question, create a final answer to the question.
733 =====
734 QUESTION: What is the birthday of this Anglo–Irish actress, courtesan, and mistress, who was the mother to
735 the illegitimate daughter of King William IV?
736 =====
737 ANSWER: Please answer in less than 6 words.

Listing 3: Example of the Prompt for QA with Retrieved Contexts.

738 Given the following question and contexts, create a final answer to the question.
739 =====
740 QUESTION: During which years was the model of car, featured on the cover of Earth’s "Pentastar: In the
741 Style of Demons" manufactured?
742 =====
743 CONTEXT:
744 1: Pentastar: In the Style of Demons is the third full–length studio album by the drone doom band Earth.
745 2: In 1957, he published The Interpersonal Diagnosis of Personality, which the Annual Review of Psychology
746 called the "most important book on psychotherapy of the year".
747 3: During the evanescent heyday of the cyberdelic counterculture, he served as a consultant to Billy Idol in the
748 production of the 1993 album Cyberpunk.
749 4: During the development of the Barracuda, one of the worst–kept secrets was Ford’s plan to introduce a new
750 sporty compact car based on the inexpensive Falcon chassis and running gear (which was eventually
751 released as the Mustang in mid–model year 1964); the extent of the other changes was not known.
752 5: "Peace in Mississippi" is a cover of the Jimi Hendrix song. The original vinyl release of the album has an
753 alternative take of "Peace in Mississippi".
754 6: A 1975 Barracuda had been planned before the end of the 1970–74 model cycle.
755 7: In the spring of 2021, when the third wave of the coronavirus epidemic arrived, ValAradi called their airline
756 one of the "rare rays of hope" for investors.
757 8: During this time the first U.S. Federal auto safety standards were phased in, and Chrysler’s response a
758 requirement for side–marker lights distinguishes each model year of the second–generation Barracuda:
759 As the pony–car class became established and competition increased, Plymouth began to revise the
760 Barracuda’s engine options.

- 761 9: The Barracuda sold for a base price of US\$2,512 (\$24,000 today).The 1964 model year was the first for the
762 Barracuda and also the last year for push-button control of the optional Torqueflite automatic
763 transmission.
- 764 10: In the words of symbolist poet Stel`Aphane Mallarme`A:Languages are imperfect because multiple; the
765 supreme language is missing...no one can utter words which would bear the miraculous stamp of Truth
766 Herself Incarnate...how impossible it is for language to express things...in the Poet's hands...by the
767 consistent virtue and necessity of an art which lives on fiction, it achieves its full efficacy.
- 768 11: In France, the heart of the Decadent movement was during the 1880s and 1890s, the time of fin de
769 sie`Acle, or end-of-the-century gloom.
- 770 12: Pentastar: In the Style of Demons is the third full-length studio album by the drone doom band Earth,
771 released in 1996. It has a more rock-oriented sound than their earlier drone doom work, although in a
772 very minimalist style.
- 773 13: The game was a rematch of the previous year's Russell Athletic Bowl, which Clemson won 40`A\$6.The
774 two participants for the game were two of the semifinalists which were the Clemson Tigers and
775 Oklahoma Sooners.
- 776 14: The effect of the war on Ernst was devastating; in his autobiography, he wrote of his time in the army thus:
777 "On the first of August 1914 M[ax].E[rnst]. died. He was resurrected on the eleventh of November
778 1918".
- 779 15: Plymouth's executives had wanted to name the new model Panda, an idea unpopular with its designers. In
780 the end, John Samsen's suggestion of Barracuda prevailed. Based on Chrysler's A-body, the Barracuda
781 debuted in fastback form on April 1, 1964.
- 782 16: The Scapigliati (literally meaning "unkempt" or "disheveled") were a group of writers and poets who
783 shared a sentiment of intolerance for the suffocating intellectual atmosphere between the late
784 Risorgimento (1860s) and the early years of unified Italy (1870s).
- 785 17: Recurrent themes in his literary works include the supremacy of the individual, the cult of beauty,
786 exaggerated sophistication, the glorification of machines, the fusion of man with nature, and the exalted
787 vitality coexisting with the triumph of death.
- 788 18: Disc brakes and factory-installed air conditioning became available after the start of the 1965 model year.
789 For the 1966 model year, the Barracuda received new taillamps, new front sheet metal, and a new
790 instrument panel.
- 791 19: "Perhaps the worst failing of the book is the omission of any kind of proof for the validity and reliability
792 of the diagnostic system," Eysenck wrote.
- 793 20: Based on stretched underpinnings of the rear-drive Alfa Romeo Giulia, it was rumored to be powered by
794 a turbocharged V6 and arrive within the 2019 model year.
- 795 21: Their investments are in fleet development and the construction of airports, the first of which will be
796 opened in Brasov.
- 797 22: He broke the hill record and this innovation was widely copied in the years to come.[citation needed]Mays
798 made his mark on the track in such events as the 1935 German Grand Prix (scene of a famous victory of
799 Tazio Nuvolari), sharing his ERA with Ernst von Delius.
- 800 23: There is still a question about the truth of the disclosure. In the 1968 Dagnet episode "The Big Prophet",
801 Liam Sullivan played Brother William Bentley, leader of the Temple of the Expanded Mind, a thinly
802 fictionalized Leary.
- 803 24: The Belgian Fe`Alicien Rops was instrumental in the development of this early stage of the Decadent
804 movement. A friend of Baudelaire, he was a frequent illustrator of Baudelaire's writing, at the request of
805 the author himself.
- 806 25: After taking responsibility for the controlled substance, Leary was convicted of possession under the
807 Marihuana Tax Act of 1937 on March 11, 1966, sentenced to 30 years in prison, fined \$30,000, and
808 ordered to undergo psychiatric treatment.
- 809 26: The general court delegation from Sullivan County is made up of all of the members of the New
810 Hampshire House of Representatives from the county. In total, there are 13 members from 11 different
811 districts.
- 812 27: Both teams then exchanged field goals, which brought the score to 16-10 in favor of Clemson. With 2:17
813 remaining, Oklahoma drove down the length of the field to score a touchdown, which gave the Sooners a
814 one-point lead.
- 815 28: The average household size was 2.41 and the average family size was 2.88.23.90% of the population were
816 under the age of 18, 6.40% from 18 to 24, 28.00% from 25 to 44, 25.90% from 45 to 64, and 15.80%
817 who were 65 years of age or older.
- 818 29: The band announced the release of a deluxe version of the album "How It Feels To Be Lost", which came
819 out on August 21, 2020. On June 2, 2021, the band released the single "Bloody Knuckles" from their
820 upcoming album.
- 821 30: The 82nd Orange Bowl was a College Football Playoff semifinal with the winner of the game competing
822 against the winner of the 2015 Cotton Bowl: Alabama Crimson Tide football in the 2016 College
823 Football Playoff National Championship, which took place at the University of Phoenix Stadium in
824 Glendale, Arizona.

825 =====
826 QUESTION: During which years was the model of car, featured on the cover of Earth's "Pentastar: In the
827 Style of Demons" manufactured?
828 =====
829 ANSWER: Please answer in less than 6 words.

Listing 4: Example of the Prompt for QA with Retrieved Contexts for MDR, KGP-T5, KGP-LLaMA and KGP-MDR.

830 Given the following question and contexts, create a final answer to the question.
831 =====
832 QUESTION: Anthony Avent played basketball for a High School that is located in a city approximately 8 mi
833 west of where?
834 =====
835 CONTEXT:
836 1: Newark is the second largest city in the New York metropolitan area, located approximately 8 mi west of
837 lower Manhattan. Prior to Seton Hall, Avent played at Malcolm X Shabazz High School in Newark,
838 New Jersey.
839
840 2: Newark is the second largest city in the New York metropolitan area, located approximately 8 mi west of
841 lower Manhattan. The United States District Court for the District of New Jersey is also located in the
842 city.
843
844 3: Newark is the second largest city in the New York metropolitan area, located approximately 8 mi west of
845 lower Manhattan. Near Market Street and includes a dormitory for boarding students; and Saint
846 Vincent Academy which is an all-girls Roman Catholic high school founded and sponsored by the
847 Sisters of Charity of Saint Elizabeth and operated continuously since 1869. Link Community School is a
848 non-denominational coeducational day school that serves approximately 128 students in seventh and
849 eighth grades.
850
851 4: Prior to Seton Hall, Avent played at Malcolm X Shabazz High School in Newark, New Jersey. Newark is
852 the second largest city in the New York metropolitan area, located approximately 8 mi west of lower
853 Manhattan.
854
855 5: Prior to Seton Hall, Avent played at Malcolm X Shabazz High School in Newark, New Jersey. The
856 United States District Court for the District of New Jersey is also located in the city.
857
858 6: Prior to Seton Hall, Avent played at Malcolm X Shabazz High School in Newark, New Jersey. On
859 Newark Bay, it is run by the Port Authority of New York and New Jersey and serves as the principal
860 container ship facility for goods entering and leaving the New York metropolitan area and the
861 northeastern quadrant of North America.
862
863 7: He played collegiately at Seton Hall University where he played in the 1989 NCAA championship game.
864 Prior to Seton Hall, Avent played at Malcolm X Shabazz High School in Newark, New Jersey. Prior to
865 Seton Hall, Avent played at Malcolm X Shabazz High School in Newark, New Jersey.
866
867 8: He played collegiately at Seton Hall University where he played in the 1989 NCAA championship game.
868 Prior to Seton Hall, Avent played at Malcolm X Shabazz High School in Newark, New Jersey. The
869 United States District Court for the District of New Jersey is also located in the city.
870
871 9: He played collegiately at Seton Hall University where he played in the 1989 NCAA championship game.
872 Prior to Seton Hall, Avent played at Malcolm X Shabazz High School in Newark, New Jersey. As of
873 the 2020-21 school year, the district, comprises 65 schools, had an enrollment of 40,423 students and
874 2,886.5 classroom teachers (on an FTE basis), for a student-teacher ratio of 14.0:1. Science Park
875 High School, which was the 69th-ranked public high school in New Jersey out of 322 schools statewide,
876 in New Jersey Monthly magazine's September 2010 cover story on the state's "Top Public High
877 Schools", after being ranked 50th in 2008 out of 316 schools.
878
879 10: Anthony Avent (born October 18, 1969) is an American former professional basketball player who was
880 selected by the Atlanta Hawks in the first round (15th pick overall) of the 1991 NBA draft. Newark is
881 the second largest city in the New York metropolitan area, located approximately 8 mi west of lower
882 Manhattan.
883

884 11: Anthony Avent (born October 18, 1969) is an American former professional basketball player who was
885 selected by the Atlanta Hawks in the first round (15th pick overall) of the 1991 NBA draft. The United
886 States District Court for the District of New Jersey is also located in the city.
887

888 12: Anthony Avent (born October 18, 1969) is an American former professional basketball player who was
889 selected by the Atlanta Hawks in the first round (15th pick overall) of the 1991 NBA draft. Atlanta
890 United 1, New York Red Bulls 2 The first game in Atlanta United history was played before a sellout
891 crowd of 55,297.
892

893 13: Anthony Avent (born October 18, 1969) is a retired American professional basketball player who was
894 selected by the Atlanta Hawks in the first round (15th pick overall) of the 1991 NBA Draft. The total
895 school enrollment in Newark was 77,097 in the 2013–2017 ACS, with nursery and preschool
896 enrollment of 7,432, elementary/high school (K–12) enrollment of 49,532, and total college/graduate
897 school enrollment of 20,133. The Newark Public Schools, a state-operated school district, is the largest
898 school system in New Jersey.
899

900 14: Anthony Avent (born October 18, 1969) is a retired American professional basketball player who was
901 selected by the Atlanta Hawks in the first round (15th pick overall) of the 1991 NBA Draft. As of the
902 2020–21 school year, the district, comprises 65 schools, had an enrollment of 40,423 students and
903 2,886.5 classroom teachers (on an FTE basis), for a student–teacher ratio of 14.0:1. Science Park
904 High School, which was the 69th–ranked public high school in New Jersey out of 322 schools statewide,
905 in New Jersey Monthly magazine’s September 2010 cover story on the state’s “Top Public High
906 Schools”, after being ranked 50th in 2008 out of 316 schools.
907

908 15: Anthony Avent (born October 18, 1969) is a retired American professional basketball player who was
909 selected by the Atlanta Hawks in the first round (15th pick overall) of the 1991 NBA Draft. In the
910 2013–2017 American Community Survey, 13.6% of Newark residents ages 25 and over had never
911 attended high school and 12.5% didn’t graduate from high school, while 74.1% had graduated from high
912 school, including the 14.4% who had earned a bachelor’s degree or higher.
913 =====

914 QUESTION: Anthony Avent played basketball for a High School that is located in a city approximately 8 mi
915 west of where?
916 =====

917 ANSWER: Please answer in less than 6 words.

Listing 5: Example of the Prompt for Grading QA.

918 You are an expert professor specialized in grading whether the prediction to the question is correct or not
919 according to the real answer.
920 =====

921 For example:
922 =====

923 Question: What company owns the property of Marvel Comics?
924 Answer: The Walt Disney Company
925 Prediction: The Walt Disney Company
926 Return: 1
927 =====

928 Question: Which constituent college of the University of Oxford endows four professorial fellowships for
929 sciences including chemistry and pure mathematics?
930 Answer: Magdalen College
931 Prediction: Magdalen College.
932 Return: 1
933 =====

934 Question: Which year was Marvel started?
935 Answer: 1939
936 Prediction: 1200
937 Return: 0
938 =====

939 You are grading the following question:
940 Question: Anthony Avent played basketball for a High School that is located in a city approximately 8 mi
941 west of where?
942 Answer: lower Manhattan
943 Prediction: Newark
944 If the prediction is correct according to the answer, return 1. Otherwise, return 0.
945 Return: your reply can only be one number '0' or '1'