HiGenQA: Exploring Hint Generation Approaches for Open Domain Question Answering

Anonymous ACL submission

Abstract

Automatic Question Answering (QA) systems rely on contextual information to provide accurate answers. Commonly, contexts are prepared through either retrieval-based or generation-based methods. The former involves retrieving relevant documents from a corpus like Wikipedia, whereas the latter uses generative models such as Large Language Models (LLMs) to generate the context. In this paper, we introduce a novel context preparation approach called HIGENQA, which employs Automatic Hint Generation (HG) systems. Unlike traditional methods, HIGENQA prompts LLMs to produce hints about potential answers for the question rather than generating relevant context. We evaluate our approach across three QA datasets including TriviaQA, Natural Questions, and Web Questions, examining how the number and order of hints impact performance. Our findings show that HIGENQA surpasses both retrieval-based and generation-based approaches. We demonstrate that hints enhance the accuracy of answers more than retrieved and generated contexts.

1 Introduction

Automatic Question Answering (QA) systems (Abdel-Nabi et al., 2023) have recently garnered significant attention. They allow users posing questions and receiving direct responses. QA systems typically comprise three main components: Context-Preparator, Reranker, and Reader (Rogers et al., 2023). The Context-Preparator component aims to supply relevant context to the user question. The Reranker then prioritizes the documents based on their relevance to the question or to potential answers (Mao et al., 2021). Lastly, the Reader extracts the answer from the provided context. The Context-Preparator component is the initial step and a crucial element in QA systems. If this component fails to prepare the most relevant contexts, the entire QA system can be led astray. Therefore, the accuracy and performance of the Context-Preparator component are crucial for the overall success of QA systems. The Context-Preparator component is divided into two primary categories including Retrieval-based and Generation-based (Li et al., 2024).

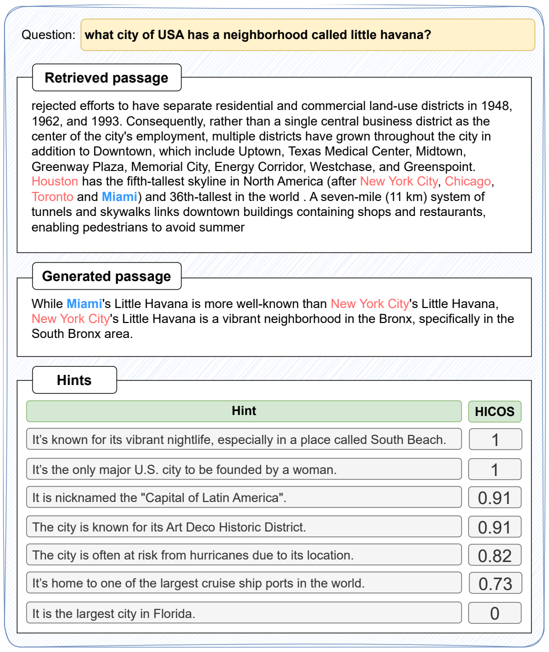

Figure 1: Example of generated hints, context produced by LLaMA-70, and a passage retrieved by MSS-DPR for a TriviaQA sample question, with convergence score (HICOS) ranging from 0 (lowest) to 1 (highest). Words in blue indicate the correct answer, while those in red represent other potential answers.

Retrieval-based methods retrieve relevant passages from document collections, such as Wikipedia, using techniques like keyword matching (Siddiqui and Tiwary, 2005) or vector space models (Gysel et al., 2018). A notable limitation of these methods is that a retrieved passage tend to be lengthy, often exceeding 100 words (Karpukhin et al., 2020). Consequently, some sentences within these passages may be irrelevant to the question (Mitra and Craswell, 2017). Figure 1 illustrates a retrieved passage where only one sentence contains the potential answers including also the correct one, while the others are irrelevant.

In contrast, generation-based methods use generative models, such as large language models (LLMs) (Workshop et al., 2022) and Seq-to-Seq techniques (Sutskever et al., 2014), to produce relevant context. A major limitation of these methods is that they typically produce only a small number of sentences as context, usually just one or two. When the number of sentences is small, there is a risk that the QA system could be mislead if the answer is incorrect, due to insufficient context to substantiate the answer. Figure 1 also displays a generated passage consisting of only two sentences, which could mislead the Reader. This is because the correct answer appears less frequently than incorrect ones, and the scant context does not provide sufficient information for the Reader component to identify the correct answer accurately.

Our research aims to overcome the shortcomings of both retrieval-based and generation-based methods. It eliminates irrelevant sentences and provides only those with useful information about the answer, addressing by this a key limitation of the retrieval-based method. Additionally, we aim to expand the number of informative sentences beyond just one or two as usually is in the case of generated context, tackling a major drawback of the generation-based approach.

We present HiGenQA[1], a novel approach that utilizes Automatic Hint Generation (HG) systems (Jangra et al., 2024) to generate hints as the context. This method generates hints per question with the aim to guide the Reader component toward the answer without directly revealing it. Figure 1 illustrates seven generated hints, each accompanied by its computed convergence score (HICOS). The convergence score is a measure that indicates how effectively a hint can narrow down or eliminate potential answers to a given question (Mozafari et al., 2024). The hints can be then subsequently reranked based on criteria such as the aforementioned convergence score or semantic relevance, setting the stage for the Reader to discern the correct answer from these prioritized hints. To assess the effectiveness of our approach, we generate hints for each question belonging to the test sets of the TriviaQA (Joshi et al., 2017), Natural Questions (NQ) (Kwiatkowski et al., 2019), and Web Questions (WebQ) (Berant et al., 2013) datasets. Our extensive experiments demonstrate that using hints leads to better performance than relying on retrieved passages or generated context. To sum up, we make the following contributions in this work:

- We propose a new approach for the Context-Preparator component in QA systems using hint generation systems.
- We generate and release hints along with their corresponding convergence scores for the questions of the test sets of the TriviaQA, NQ, and WebQ datasets.
- We conduct extensive experiments on three datasets using zero and few-shot strategies across various numbers of hints and reranking methods.

2 Related Work

2.1 Retrieval-based Methods

Retrieval-based methods can be divided into two primary categories: (1) Sparse retrieval and (2) Dense retrieval. Sparse retrieval methods rely on word-level matching to establish connections between vocabulary and documents. Notable examples are Boolean Retrieval (Salton et al., 1983), BM25 (Robertson and Zaragoza, 2009), SPLADE (Formal et al., 2021), and UniCOIL (Lin and Ma, 2021). On the other hand, dense retrieval methods capture deep semantic information from documents to understand underlying semantics and improve retrieval accuracy. Some key examples are DPR (Karpukhin et al., 2020), ANCE (Xiong et al., 2020), E5 (Wang et al., 2022), and SimLM (Wang et al., 2023).

2.2 Generation-based Methods

Generation-based systems can be broadly classified into two main categories: (1) Generative document retrieval and (2) Reliable response generation. Generative document retrieval utilizes the parametric memory of generative models to retrieve relevant documents. Unlike retrieval-based systems, this approach depends on pre-trained generative models, such as the BART (Lewis et al., 2020), to produce document identifiers directly related to the question. Some notable examples are

[1]We have included the datasets and experimental results in the supplementary data. They will also be made available on GitHub after publication.

DSI (Tay et al., 2024), DynamicRetriever (Zhou et al., 2023), SEAL (Bevilacqua et al., 2022), and NCI (Wang et al., 2024). Conversely, Reliable response generation methods provide a more dynamic form of information access by directly producing detailed, user-centric responses. Notable instances are LLaMA (Brown et al., 2020), InstructGPT (Ouyang et al., 2024), T5 (Raffel et al., 2020), PaLM (Chowdhery et al., 2024) and Copilot[2].

2.3 Hint Generation

HG systems can be categorized into two main categories: (1) Hint generation for Programming (AHGP) and (2) Hint generation for Questions (AHGQ). AHGP aims to create helpful hints for programming exercises (Rivers et al., 2016). Some notable examples are ITAP (Jin et al., 2012) and Catnip (Obermüller et al., 2021) systems. In contrast, methods for AHGQ focus on generating hints for user questions rather than programming exercises (Jangra et al., 2024). The study by Jatowt et al. (2023) explores the use of Wikipedia for generating hints without utilizing LLMs, primarily to introduce this as a new area of research. The work by Mozafari et al. (2024) advances the field by releasing the first dedicated dataset named TriviaHG, along with a novel automatic evaluation method for assessing the quality of hints.

To the best of our knowledge, no study has yet explored the use of AGHQ approaches as the Context-Preparator component for QA systems.

3 Method

In this section, we first explore the theoretical foundations underpinning our approach, followed by a detailed explanation of its implementation.

3.1 Hypothesis

Let q be a question linked to a set of candidate answers $\mathcal{A} = \{a_1, a_2, \ldots, a_n\}$, such that $q \to \mathcal{A}$, which indicates that $\mathcal{A}$ is assumed to encompass all possible answers to q. Additionally, let $\mathcal{S} = \{s_1, s_2, \ldots, s_j\}$ be the context, consisting of a series of sentences s_i provided to determine the answer to q. Each sentence s_i typically discusses or relates to certain entities or subjects, which we refer to as $\mathcal{C}'_i$. For instance, the sentence *"He was a professional."* might pertain to different possible professions such as actor, painter, athlete, etc. Consequently, the set $\mathcal{C}'_i$ could encompass, in this example, various individuals from diverse occupations. However, if the question q specifically inquires about just one particular profession, it is superfluous to consider all potential entities that the sentence might include. Therefore, we define $\mathcal{C}_i = \mathcal{C}'_i \cap \mathcal{A}$ to select only those entities that represent the intersection between the candidate answers for q and the possible entities from s_i. This process assists in eliminating irrelevant entities, retaining only valid candidate answers to q.

We define a score $\tau_{\mathcal{S}}(a)$ for a candidate answer a within the context $\mathcal{S}$ to represent how well a scores as a candidate answer in the context $\mathcal{S}$. It counts the number of supporting sentences for the candidate answer a among all sentences in the context $\mathcal{S}$:

$$\tau_{\mathcal{S}}(a) = \frac{\sum_{s \in \mathcal{S}} \chi_{\mathcal{C}_s}(a)}{|\mathcal{S}|} \tag{1}$$

where $|\mathcal{S}|$ indicates the number of sentences within $\mathcal{S}$, and $\mathcal{C}_s$ identifies the valid candidate answer set associated with sentence s. The function $\chi_{\mathcal{C}_s}(a)$ is to determine whether a candidate answer a is a member of the candidate answer set $\mathcal{C}_s$:

$$\chi_{\mathcal{C}_s}(a) = \begin{cases} 1 & \text{if } a \in \mathcal{C}_s \\ 0 & \text{if } a \notin \mathcal{C}_s \end{cases} \tag{2}$$

The candidate answer a with the highest $\tau_{\mathcal{S}}(a)$ across the context $\mathcal{S}$ is proposed as the most likely correct answer:

$$a^* = \arg\max_{a \in \mathcal{A}} \tau_{\mathcal{S}}(a) \tag{3}$$

Let's consider an example as follows. Suppose the question q is: *"What city in the USA has a neighborhood called Little Havana?"*. And suppose the context $\mathcal{S}$ consists of two sentences s_1 (red) and s_2 (blue):

```
The city is often at risk from hurricanes
due to its location. Additionally, it's
the only major U.S. city to be founded by
a woman.
```

The entities supported by s_1 are $\mathcal{C}'_1 = \{$*San Juan, Kingston, Miami, New York City, . . .*$\}$, and ones by s_2 are $\mathcal{C}'_2 = \{$*Miami*$\}$. Let us also suppose that the following candidate answers are possible for q: $\mathcal{A} = \{$*Houston, Miami, New York City*$\}$. Thus, the intersecting sets are $\mathcal{C}_1 = \mathcal{C}'_1 \cap \mathcal{A} = \{$*Miami, New York City*$\}$ and $\mathcal{C}_2 = \mathcal{C}'_2 \cap \mathcal{A} = \{$*Miami*$\}$. We calculate the score $\tau_{\mathcal{S}}$ for *Miami* using Eq. 1:

[2] https://copilot.microsoft.com/

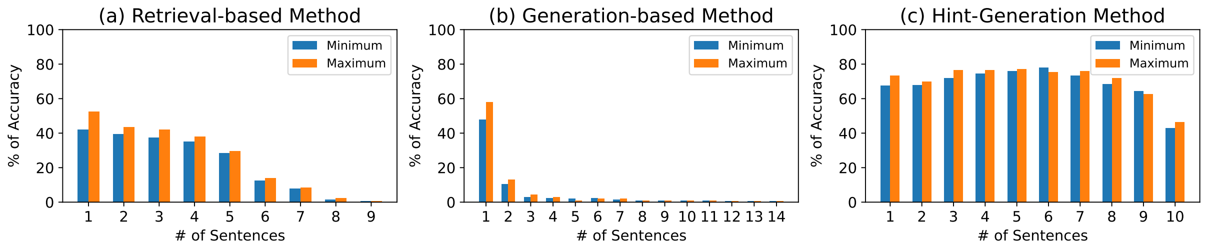

Figure 2: Accuracy results for 200 random questions from TriviaQA, NQ, and WebQ when using LLaMA-7b as the Reader and varying the numbers of context sentences. The context sentences are obtained by (a) Retrieval-based (DPR), (b) Generation-based (LLaMA-70b), and (c) Hint-Generation (HiGen-FT) methods. The blue (red) columns indicate the accuracy when the total number of potential entities across sentences is at its minimum (maximum). The number of potential entities per sentence is calculated using HICOS approach (Mozafari et al., 2024).

$$\tau_{\mathcal{S}}(\text{Miami}) = \frac{\chi_{\mathcal{C}_1}(\text{Miami}) + \chi_{\mathcal{C}_2}(\text{Miami})}{|\mathcal{S}|} = \frac{2}{2} = 1 \quad (4)$$

The scores for *Houston* and *New York City* are 0 and 0.5, respectively. Thus, according to Eq. 3, the most likely correct answer to q is **Miami** as supported by most of the sentences.

We believe that a context supporting more potential entities within its sentences can improve the performance of QA systems. As shown in Figure 2, the *Maximum* column illustrates that when the total number of potential entities across sentences is highest, the accuracy exceeds that observed with the lowest count. Figure 2b also demonstrates how a scarcity of potential entities can mislead the QA system. As discussed in Section 1, this issue is especially common in generation-based methods, which frequently produce contexts with a small number of sentences.

Moreover, Figure 2a shows that incorporating additional relevant sentences can enhance QA system performance; conversely, the inclusion of irrelevant sentences can impair it. The figure demonstrates a correlation between an increase in irrelevant sentences and a decrease in accuracy. This presents a frequent challenge for retrieval-based methods, which are prone to including irrelevant sentences in the passages they retrieve.

Nevertheless, Figure 2c demonstrates that the results of the HG method can effectively guide the QA system toward the correct answer. Table 25 in Appendix D shows some generated hints and their supported candidate answers.

3.2 Implementation

To implement our approach, we adapt the method introduced by Mozafari et al. (2024) for generating ten hints, modifying their original prompt. While they implemented an answer-aware approach, we take an answer-agnostic approach since the correct answer is unknown. The prompt we use is as follows:

```
Generate 10 concise and relevant hint
sentences for the following question. List
the hints without revealing the answers
within them.
```

We also utilize the following prompt in the Reader to extract the answer from the context:

```
According to the following context, answer
the question:
Context: Provided Context
Question: Given Question
Answer: Here is the answer
```

4 Experimental Setup

4.1 Datasets

Our evaluation is conducted using three diverse datasets: TriviaQA (Joshi et al., 2017), NQ (Natural Questions) (Kwiatkowski et al., 2019), and WebQ (Berant et al., 2013). TriviaQA dataset comprises a comprehensive collection of trivia questions, which have been curated from various trivia and quiz-league websites. NQ has been constructed from Google Search queries, providing a realistic set of questions people ask. The answers to these questions are drawn as specific spans or segments from Wikipedia articles. WebQ dataset consists of questions sourced from the Google Suggest API, which generates predictive search suggestions based on user input. The answers are tied to entities within Freebase (Bollacker et al., 2008). A more detailed description of dataset statistics, their splits (Table 8), and distributions based on the question type (Table 9) can be found in Appendix A.

Method	TriviaQA	NQ	WebQ
BM25	117.15	114.93	114.24
DPR	118.66	110.97	114.56
Contriever	117.41	107.47	113.69
MSS	118.62	113.44	117.25
MSS-DPR	118.35	109.56	115.66
LLaMA-70b	50.34	61.52	75.93
HiGen-FT	73.54	96.13	90.43
HiGen-Va	96.85	106.78	93.02

Table 1: Comparison of the average lengths of hints, generated contexts, and retrieved passages.

Method	TriviaQA[1]		NQ[2]		WebQ[3]	
	EM	F1	EM	F1	EM	F1
Zero-Shot						
BM25	23.28	27.22	3.55	5.62	10.97	18.54
Contriever	18.13	22.29	1.94	3.66	8.17	14.05
DPR	23.22	27.7	2.3	3.93	11.71	19.43
MSS	18.15	22.35	1.97	3.58	9.94	17.24
MSS-DPR	18.14	22.23	4.24	6.53	11.17	18.71
LLaMA-70b	21.45	26	3.88	6.23	12.11	20.27
HiGen-Va	22.01	26.5	9.06	12.54	13.88	21.74
HiGen-FT	**23.55**	**28.03**	**10.89**	**14.85**	**14.96**	**23.08**
Few-Shot						
BM25	25.78	30.29	4.6	7.33	11.17	18.93
Contriever	21.48	25.87	2.47	4.21	7.53	13.49
DPR	25.02	29.49	3.24	5.09	11.37	19.37
MSS	20.89	25.27	2.85	4.75	10.33	17.99
MSS-DPR	20.92	25.19	4.79	7.69	11.47	19.81
LLaMA-70b	23.64	28.86	5.1	7.9	9.4	17.86
HiGen-Va	34.19	39.74	12.85	18.06	18.9	28.97
HiGen-FT	**38.54**	**44.29**	**16.68**	**22.64**	**24.11**	**34.52**

[1] Zero-Shot→ HiGen-Va: 10 Def, HiGen-FT: 10 Def
Few-Shot→ HiGen-Va: 5 Conv, HiGen-FT: 7 Def
[2] Zero-Shot→ HiGen-Va: 10 Def, HiGen-FT: 10 Def
Few-Shot→ HiGen-Va: 5 Conv, HiGen-FT: 7 Def
[3] Zero-Shot→ HiGen-Va: 2 Conv, HiGen-FT: 10 Def
Few-Shot→ HiGen-Va: 5 Conv, HiGen-FT: 7 Conv

Table 2: The results for **T5-3b** used as the reader, utilizing zero-shot and few-shot strategies. The footnotes provide information on the optimal number of hints and the ranking method chosen to achieve the best results for each learning strategy and hint generation method.

4.2 Baseline Models

BM25 (Robertson and Zaragoza, 2009) is a probabilistic retrieval model that employs term frequency (TF) and inverse document frequency (IDF) metrics to assess the relevance of documents based on the common words in the question and the documents. **Contriever** (Izacard et al., 2022) is an unsupervised framework designed for pre-training models for retrieval tasks, utilizing contrastive learning techniques. **MSS** (Sachan et al., 2021) is a dense retrieval model trained to predict masked salient spans, such as named entities, using a reader network. **DPR** (Karpukhin et al., 2020) uses annotated question-context paragraphs and hard negative examples to train a supervised dense retriever. **MSS-DPR** (Sachan et al., 2021) enhances the performance of DPR by initially pre-training the dense retriever with MSS. This is followed by supervised fine-tuning in the style of DPR. **LLaMA-v2** (Touvron et al., 2023) is an advanced LLM tailored for scalable natural language processing tasks, providing exceptional efficiency in generating context.

We employ the preprocessed English Wikipedia dump, provided by Karpukhin et al. (2020), as a source for our evidence passages in retrieval-based methods. We also utilize the first top retrieved passage for the Reader. We use the LLaMA-70b as the generation-based baseline because it is the core for our HG system. Therefore, it is reasonable to compare the HiGenQA method directly with LLaMA-70b to ensure a fair assessment.

4.3 Hint Generation Methods

We employ two versions of HG systems to create hints for questions: The vanilla version (HiGen-Va) and the finetuned version (HiGen-FT). In the HiGen-Va, the LLaMA-70b model is simply prompted to generate hints for a specific question. For the HiGen-FT, we first finetune the LLaMA-70b model using the TriviaHG dataset (Mozafari et al., 2024), and then prompt it to generate hints. For the detailed statistics of the TriviaHG dataset, readers are referred to Table 10 in Appendix A.

Additionally, we explore three different reranking methods for reranking hints: Default (Def), RankT5 (T5), and Convergence (Conv). The Default order refers to the sequence in which the hints are originally generated by the HG system. The RankT5 method rearranges hints through pairwise and listwise ranking techniques employing the T5 model (Zhuang et al., 2023). Lastly, the Convergence method sorts the hints according to the HICOS score in descending order.

We also investigate the impact of using various quantities of hints to prepare context. In our experiments, we concatenate the first 2, 5, 7, or 10 hints in various sequences to generate a comprehensive context for the Reader component. This approach allows us to assess how the number and order of hints influence the effectiveness and performance of the QA system. To compare results, we use the

Method	ACC	EM	F1	PR	RC	CON	BERT
			Zero-Shot				
BM25	34.21	0	7.67	4.56	36.2	38.98	69.29
Contriever	20.64	0	5.57	3.28	30.71	26.47	67.13
DPR	31.19	0	7.5	4.47	35.12	37.03	69.22
MSS	20.38	0	5.43	3.19	30.4	26.11	67.06
MSS-DPR	19.73	0	5.58	3.27	30.67	26.43	67.2
LLaMA-70b	47.3	0	**9.11**	**5.44**	**42.57**	55.32	70.77
HiGen-Va[1]	**59.06**	0	8.04	4.75	41.51	54.74	70.35
HiGen-FT[2]	54.97	0	8.96	5.33	42.21	**60.93**	**71.4**
			Few-Shot				
BM25	40.5	38.15	46.7	46.2	52.8	51.06	83.32
Contriever	31.62	33.54	40.4	39.9	47.31	42.86	80.46
DPR	36.29	37.15	45.3	44.8	51.06	49.16	82.91
MSS	31.56	33.99	41.1	40.7	47.84	43.41	80.66
MSS-DPR	31.96	32.69	39.9	39.4	46.95	42.43	80.2
LLaMA-70b	52.59	41.26	48.7	48.6	52.59	51.58	83.3
HiGen-Va[3]	57.71	50.76	60.6	60.4	65.12	65.92	88.61
HiGen-FT[1]	**58.06**	**54.6**	**64.7**	**64.8**	**69.53**	**70.15**	**89.89**

[1] 7 hints, Convergence reranking.
[2] 10 hints, Default reranking.
[3] 5 hints, Convergence reranking.

Table 3: The results for **LLaMA-7b** used as the reader on **TriviaQA**, using zero-shot and few-shot strategies.

Method	ACC	EM	F1	PR	RC	CON	BERT
			Zero-Shot				
BM25	23.38	0	2.72	1.54	19	15.21	63.14
Contriever	11.52	0	1.84	1.03	15.71	10	61.14
DPR	11.36	0	1.77	0.99	15.55	9.78	61.04
MSS	11.44	0	1.67	0.94	14.75	9.36	60.94
MSS-DPR	23.21	0	2.94	1.66	21.16	17.73	63.96
LLaMA-70b	37.73	0	3.88	2.2	**31.98**	31.97	65.31
HiGen-Va[1]	**51.11**	0	3.44	1.95	25.71	26.2	64.97
HiGen-FT[1]	49.26	0	**4.38**	**2.5**	26.96	**33.19**	**66.8**
			Few-Shot				
BM25	36.65	10.33	16.6	16.1	23.28	19.14	70.32
Contriever	31.66	6.84	10.7	10.2	16.17	11.19	66.61
DPR	31.3	7.15	11.1	10.6	16.92	11.63	66.83
MSS	29.25	7.15	11.1	10.5	17.2	12.05	66.78
MSS-DPR	34.35	10.44	16.4	15.9	22.81	18.67	70.24
LLaMA-70b	50.21	10.55	16.1	15.9	21.06	18.34	68.9
HiGen-Va[2]	59.36	18.48	26.6	26.4	34.58	33.24	75.58
HiGen-FT[2]	**64.43**	**20.72**	**29.5**	**29.55**	**37.19**	**36.81**	**76.7**

[1] 10 hints, Convergence reranking.
[2] 7 hints, Convergence reranking.

Table 4: The results for **LLaMA-7b** used as the reader on **NQ**, utilizing zero-shot and few-shot strategies.

metrics mentioned in Appendix B.

4.4 Readers

We utilize two distinct language models, T5-3b (Raffel et al., 2020) and LLaMA-7b (Touvron et al., 2023), as the Reader component in our system. In addition to employing these models, we incorporate techniques such as Zero-Shot and Few-Shot[3] to enhance their capability to handle tasks with limited direct training on specific tasks. This setup allows us to explore the effectiveness of these models in adapting to new data and challenges using minimal examples.

5 Results

5.1 Context Length

We first discuss the average lengths of contexts retrieved or generated by different models within the Context-Preparator component. As noted in Section 1, our approach yields contexts that are longer than those produced by generation-based methods but shorter than those from retrieval-based methods. Table 1 provides details on the average lengths of hints, generated contexts, and retrieved passages across the TriviaQA, NQ, and WebQ datasets. The data indicates that the length of hints produced by both **HiGen-FT** and **HiGen-Va** methods are shorter than those from all retrieval-based methods. However, when compared with **LLaMA-70b** used as a generative approach, the hints are longer.

[3]The choice to limit the number of shots to only 5 in few-shot learning is motivated by the high cost associated with exploring various shot values.

5.2 Results of HiGenQA

In this section, we present and analyze the performance and results of the HiGenQA approach, comparing it against various baselines. As previously mentioned, our experimental framework encompasses a range of setups, including different datasets (Section 4.1), baseline models (Section 4.2), HG systems, orders of hints, numbers of hints (Section 4.3), and readers (Section 4.4). This comprehensive evaluation helps in assessing the robustness and effectiveness of the HiGenQA approach across multiple dimensions.

Table 2 presents the performance of the T5-3b model as the Reader component, utilizing zero-shot and few-shot learning strategies across the specified datasets, measured by Exact Match and F1 scores. The results indicate that **HiGen-FT** achieves the best performance in both learning strategies. Additionally, the outcomes from the few-shot learning strategy surpass those of the zero-shot learning strategy. For a more detailed analysis of T5-3b's performance using HiGen-Va on TriviaQA, NQ, and WebQ datasets, readers can refer to Table 11 to Table 13 in Appendix C. Tables 14 to Table 16 provide information on T5-3b's performance using

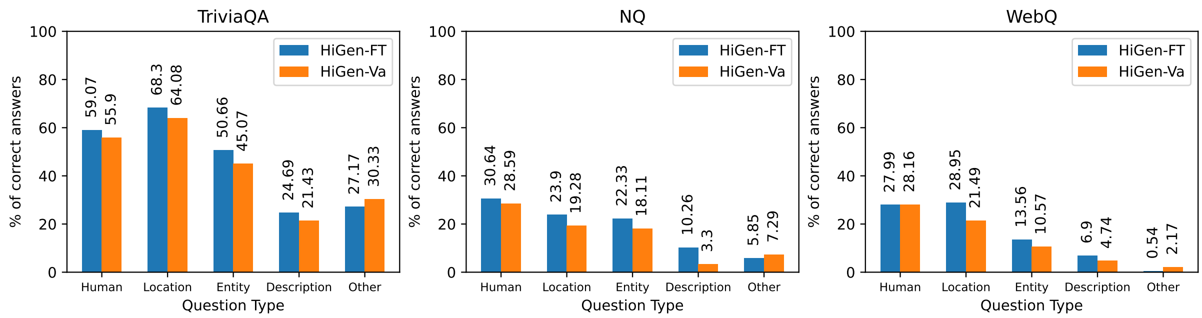

Figure 3: Exact Match values for TriviaQA, NQ, and WebQ datasets categorized by question type, based on the optimal settings for both HiGen-Va and HiGen-FT using few-shot learning on LLaMA-7b.

Method	ACC	EM	F1	PR	RC	CON	BERT
				Zero-Shot			
BM25	27.51	0	4.38	2.6	26.89	23.77	65.41
Contriever	8.22	0	2.42	1.37	21.7	14.12	62.41
DPR	26.53	0	4.8	2.79	31.7	26.57	65.63
MSS	24.9	0	4.06	2.39	27.1	21.75	64.54
MSS-DPR	30.17	0	5.08	2.98	31.42	27.36	66
LLaMA-70b	45.13	0	6.16	3.65	**44.39**	**47.39**	67.05
HiGen-Va[1]	52.95	0	5.83	3.42	38.15	40.26	67.37
HiGen-FT[1]	**54.08**	0	**7.01**	**4.14**	40.04	45.23	**68.79**
				Few-Shot			
BM25	35.33	11.42	22.7	22.8	32.42	31.55	73.04
Contriever	17.47	5.41	10.3	9.86	18.43	13.44	66.86
DPR	30.41	9.5	20.7	20.4	30.07	29.18	72.06
MSS	28.54	9.4	18.6	18.6	26.93	25.1	70.94
MSS-DPR	33.51	10.29	22.1	22.2	31.51	32.23	72.9
LLaMA-70b	48.03	8.46	16.5	16.8	21.79	22.49	68.5
HiGen-Va[2]	55.87	17.52	32.1	32.1	44.22	44.88	76.87
HiGen-FT[2]	**56.55**	**20.28**	**35.4**	**35.3**	**47.32**	**49.9**	**78.51**

[1] 10 hints, Convergence reranking.
[2] 7 hints, Convergence reranking.

Table 5: The results for **LLaMA-7b** used as the reader on **WebQ**, utilizing zero-shot and few-shot strategies.

HiGen-FT for these datasets.

Table 3, Table 4, and Table 5 show the performance of the LLaMA-7b model as the Reader component across different experimental setups. The results show that in the few-shot strategy, **HiGen-FT** consistently delivers the best outcomes across all three datasets. However, the performance under the zero-shot learning strategy varies. For the TriviaQA dataset, **LLaMA-70b** leads in F1, Precision, and Recall metrics. In the case of the NQ dataset, **LLaMA-70b** performs best regarding Recall, while for the WebQ dataset, **LLaMA-70b** excels in both Recall and Contains metrics. For other metrics across these datasets, the **HiGenQA** approach outperforms the rest. Figure 3 displays Exact Match scores for the TriviaQA, NQ, and WebQ datasets, broken down by the question type, under the optimal settings for both HiGen-Va and HiGen-FT using few-shot strategy on LLaMA-7b. The figure illustrates that **HiGen-FT** outperforms HiGen-Va across various question types. For more detailed analysis of LLaMA-7b's performance using HiGen-Va and HiGen-FT on TriviaQA, NQ, and WebQ datasets, readers can refer to tables from Table 17 to Table 22 in Appendix C.

In Appendix D, Table 24 presents a comparison of answers for a random selection of questions from the TriviaQA, NQ, and WebQ datasets, using DPR, LLaMA-70b, and HiGenQA. Additionally, Table 26 shows the answers generated from contexts retrieved by MSS-DPR, contexts generated by LLaMA-70b, and hints generated by HiGenQA using the LLaMA-7b model in a zero-shot learning strategy. Table 27, Table 28, and Table 29 illustrate the answers generated under a few-shot learning strategy by MSS-DPR, LLaMA-70b, and HiGenQA, respectively, using the LLaMA-7b model.

5.3 Ablation Study

Impact of various LLMs We investigate the impact of LLMs used as the primary component in the HG method, producing various hints for some random questions from the TriviaQA dataset. Utilizing various LLMs, we generate hints per each question. Table 6 presents the top-performing results for these LLMs as the core of the HG method across different numbers of hints and reranking methods, with LLaMA-7b serving as the Reader. The findings reveal that **Copilot** and **GPT-4** (Achiam et al., 2023) deliver the best performance for zero-shot and few-shot learning strategies, respectively, highlighting that a more knowledgeable core can produce higher-quality hints. The results when T5-3b is used as the Reader component are given in Table 23 in Appendix C.

Hint Generator	# of Params	# of Hints	Ranking	ACC	EM	F1	PR	RC	CON	BERT
					Zero-Shot					
LLaMA-Va (Touvron et al., 2023)	7b	2	Conv	68.0	0	9.37	5.57	47.48	65.0	70.52
LLaMA-Va (Touvron et al., 2023)	70b	2	Def	78.0	0	10.12	5.84	53.18	77.0	71.44
LLaMA-FT (Mozafari et al., 2024)	7b	2	T5	79.0	0	11.39	6.75	54.58	74.0	71.87
LLaMA-Va (Touvron et al., 2023)	13b	2	T5	79.0	0	9.98	5.91	48.09	73.0	71.39
WizardLM (Xu et al., 2024)	70b	5	T5	80.0	0	9.94	5.9	47.22	75.0	71.58
GPT 3.5 (Brown et al., 2020)	175b	2	Conv	81.0	0	10.6	6.13	**59.37**	81.0	71.45
LLaMA-FT (Mozafari et al., 2024)	13b	5	Conv	83.0	0	10.65	6.31	48.23	76.0	71.68
LLaMA-FT (Mozafari et al., 2024)	70b	2	Conv	83.0	0	11.28	6.77	50.72	81.0	72.41
Gemini (Team et al., 2023)	-	7	Def	88.0	0	11.83	7.05	53	88.0	72.47
GPT 4 (Achiam et al., 2023)	-	5	Def	**96.0**	0	11.5	6.8	53.97	89.0	**73.2**
Copilot	-	7	T5	92.0	0	**11.89**	**7.09**	55.32	**90.0**	72.69
					Few-Shot					
LLaMA-FT (Mozafari et al., 2024)	7b	5	Conv	76.0	67.0	72.91	71.56	76.17	78.0	92.66
LLaMA-Va (Touvron et al., 2023)	7b	7	T5	76.0	57.0	67.23	65.56	71.74	72.0	90.65
LLaMA-Va (Touvron et al., 2023)	13b	7	T5	83.0	67.0	77.04	74.87	82.17	83.0	93.33
LLaMA-FT (Mozafari et al., 2024)	13b	10	Def	84.0	67.0	74.37	72.85	78.37	82.0	92.26
LLaMA-Va (Touvron et al., 2023)	70b	7	Conv	84.0	67.0	74.29	73.18	78.87	79.0	92.09
WizardLM (Xu et al., 2024)	70b	10	T5	87.0	72.0	80.04	78.29	85.17	86.0	93.67
GPT 3.5 (Brown et al., 2020)	175b	7	Conv	88.0	72.0	79.74	78.14	83.7	84.0	93.57
Gemini (Team et al., 2023)	-	7	Def	90.0	73.0	81.24	79.73	85.5	89.0	94.58
LLaMA-FT (Mozafari et al., 2024)	70b	5	Def	91.0	69.0	80.06	78.11	85.87	87.0	94.02
Copilot	-	7	Conv	91.0	**77.0**	86.16	84.07	92	**94.0**	**95.57**
GPT 4 (Achiam et al., 2023)	-	10	Def	**93.0**	76.0	**87.29**	**85.03**	**92.17**	92.0	95.46

Table 6: The results of LLaMA-7b across different LLMs as the core of the HiGenQA system, generating hints for 100 questions. Def, Conv, and T5 indicate Default, Convergence, and RankT5 methods, respectively.

Method	TriviaQA			NQ		
	EM	RC	CON	EM	RC	CON
			Without using rerankers			
BM25	38.15	52.8	51.06	10.33	23.28	19.14
Contriever	33.54	47.31	42.86	6.84	16.17	11.19
DPR	37.15	51.06	49.16	7.15	16.92	11.63
MSS	33.99	47.84	43.41	7.15	17.2	12.05
			With using rerankers			
MSS+UPR	53.1	67.3	60.6	25.4	40.7	31
DPR+UPR	53.9	68.7	62	**25.6**	**42**	33.1
			Our method			
HiGen-Va	50.76	65.12	65.92	18.48	34.58	33.24
HiGen-FT	**54.62**	**69.53**	**70.15**	20.72	37.19	**36.81**

Table 7: Comparison of reults between baselines with rerankers, baselines without rerankers, and HiGenQA.

Impact of Rerankers Finally, we evaluate the impact of rerankers on retrieval-based methods and the HIGENQA approach to determine how HIGENQA performs relative to other methods when rerankers are used. Table 7 displays the results for retrievers without rerankers, with the UPR-reranker (Sachan et al., 2022), and HIGENQA for both the TriviaQA and NQ datasets. The results show that HIGENQA surpasses others on TriviaQA dataset. Yet, while HIGENQA achieves the best results with the Contains metric for the NQ dataset, UPR-reranker performs better in other metrics.

6 Conclusion

In this paper, we introduced a novel approach to the Context-Preparator in QA systems that generates hints instead of relying on retrieved passages or generated contexts. To thoroughly test its effectiveness, we designed a variety of experimental setups, aiming to cover a broad spectrum of possible scenarios. Our findings reveal that this new approach consistently surpasses traditional baseline methods, including both retrieval-based and generation-based approaches, on the TriviaQA, NQ, and WebQ datasets across multiple evaluation metrics. Moreover, we demonstrated that different configurations, such as employing various LLMs as the core of the HG method and adjusting ranking methods and the number of hints, significantly boost the performance of our approach. Our future work will integrate retrieval-based and generation-based methods to further enhance hint quality. The hybrid approach would seek to better utilize the extensive knowledge stored in LLMs, producing more accurate hints for complex QA tasks.

Limitations

Our study has the following limitations:

- The proposed HiGenQA approach capitalizes on the capabilities of LLMs to significantly influence the quality of the hints it generates. By drawing on the extensive stored knowledge within these models, HiGenQA provides useful hints for a variety of questions. However, this strategy also carries inherent limitations, primarily because the hints are based on the data available up to the point when the LLMs were last trained. Consequently, the hints might be out-of-date, as they cannot reflect the latest documents or current information that emerges after the training period. This issue could be particularly concerning in fields where knowledge and data are frequently updated, such as technology, medicine, and science, potentially reducing the relevance and accuracy of the hints over time.

- The computational cost and time required to calculate HICOS scores using LLMs pose significant challenges. The results demonstrate that arranging hints in descending order of their HICOS scores yields the best performance. However, the process of computing these scores for hints is both time-intensive and computationally expensive. This adds a layer of complexity and resource demand, potentially constraining scenarios that require quick or cost-effective solutions. Moreover, the need for substantial computational resources may limit the deployment of such systems in environments with restricted hardware capabilities or where minimizing operational costs is crucial.

- The LLMs used in the reader component were deliberately not fine-tuned on the TriviaQA, NQ, and WebQ datasets. This approach was chosen to purely assess the effectiveness of the Hint Generation (HG) method as a Context-Preparator tool, ensuring that the results would be free from any potential biases that could arise if the reader component had prior familiarity with these specific datasets. This setup allows us to more accurately evaluate how well the HG method can enhance the reader's performance based purely on its ability to prepare context, rather than on any pre-existing knowledge of the dataset content.

Ethical Considerations

Our study employs the GPT models, governed by the OpenAI License and Apache-2.0 license, and the LLaMA model, distributed under Meta's LLaMA 2 Community License Agreement. We adhere to these licenses for all applications. Moreover, the datasets we use are sourced from repositories authorized for academic purposes. The artifacts developed during our research are released under the MIT license to promote easy modification and use by the research community. We have ensured that our data handling, model training, and dissemination of results comply with ethical standards and legal requirements related to each utilized artifact.

References

Heba Abdel-Nabi, Arafat Awajan, and Mostafa Z. Ali. 2023. Deep learning-based question answering: a survey. *Knowledge and Information Systems*, 65(4):1399–1485.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. 2023. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on Freebase from question-answer pairs. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*, pages 1533–1544, Seattle, Washington, USA. Association for Computational Linguistics.

Michele Bevilacqua, Giuseppe Ottaviano, Patrick Lewis, Scott Yih, Sebastian Riedel, and Fabio Petroni. 2022. Autoregressive search engines: Generating substrings as document identifiers. In *Advances in Neural Information Processing Systems*, volume 35, pages 31668–31683. Curran Associates, Inc.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD International Conference on Management of Data*, SIGMOD '08, page 1247–1250, New York, NY, USA. Association for Computing Machinery.

Tom B. Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, and Others. 2020. Language models are few-shot learners. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*, NIPS '20, Red Hook, NY, USA. Curran Associates Inc.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, Parker Schuh, and Others. 2024. Palm: scaling language modeling with pathways. *J. Mach. Learn. Res.*, 24(1).

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

Thibault Formal, Benjamin Piwowarski, and Stéphane Clinchant. 2021. Splade: Sparse lexical and expansion model for first stage ranking. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '21, page 2288–2292, New York, NY, USA. Association for Computing Machinery.

Christophe Van Gysel, Maarten de Rijke, and Evangelos Kanoulas. 2018. Neural vector spaces for unsupervised information retrieval. *ACM Trans. Inf. Syst.*, 36(4).

Gautier Izacard, Mathilde Caron, Lucas Hosseini, Sebastian Riedel, Piotr Bojanowski, Armand Joulin, and Edouard Grave. 2022. Unsupervised dense information retrieval with contrastive learning. *Transactions on Machine Learning Research*.

Anubhav Jangra, Jamshid Mozafari, Adam Jatowt, and Smaranda Muresan. 2024. Navigating the Landscape of Hint Generation Research: From the Past to the Future. *arXiv e-prints*, arXiv:2404.04728.

Adam Jatowt, Calvin Gehrer, and Michael Färber. 2023. Automatic hint generation. In *Proceedings of the 2023 ACM SIGIR International Conference on Theory of Information Retrieval*, ICTIR '23, page 117–123, New York, NY, USA. Association for Computing Machinery.

Wei Jin, Tiffany Barnes, John Stamper, Michael John Eagle, Matthew W. Johnson, and Lorrie Lehmann. 2012. Program representation for automatic hint generation for a data-driven novice programming tutor. In *Proceedings of the 11th International Conference on Intelligent Tutoring Systems*, ITS'12, page 304–309, Berlin, Heidelberg. Springer-Verlag.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Ehsan Kamalloo, Nouha Dziri, Charles Clarke, and Davood Rafiei. 2023. Evaluating open-domain question answering in the era of large language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5591–5606, Toronto, Canada. Association for Computational Linguistics.

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

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Jacob Devlin, Kenton Lee, Kristina Toutanova, Llion Jones, Matthew Kelcey, Ming-Wei Chang, Andrew M. Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: A benchmark for question answering research. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Xiaoxi Li, Jiajie Jin, Yujia Zhou, Yuyao Zhang, Peitian Zhang, Yutao Zhu, and Zhicheng Dou. 2024. From Matching to Generation: A Survey on Generative Information Retrieval. *arXiv e-prints*, arXiv:2404.14851.

Jimmy Lin and Xueguang Ma. 2021. A Few Brief Notes on DeepImpact, COIL, and a Conceptual Framework for Information Retrieval Techniques. *arXiv e-prints*, arXiv:2106.14807.

Yuning Mao, Pengcheng He, Xiaodong Liu, Yelong Shen, Jianfeng Gao, Jiawei Han, and Weizhu Chen. 2021. Reader-guided passage reranking for open-domain question answering. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 344–350, Online. Association for Computational Linguistics.

Bhaskar Mitra and Nick Craswell. 2017. Neural Models for Information Retrieval. *arXiv e-prints*, arXiv:1705.01509.

Jamshid Mozafari, Anubhav Jangra, and Adam Jatowt. 2024. TriviaHG: A Dataset for Automatic Hint Generation from Factoid Questions. *arXiv e-prints*, arXiv:2403.18426.

Florian Obermüller, Ute Heuer, and Gordon Fraser. 2021. Guiding next-step hint generation using automated tests. In *Proceedings of the 26th ACM Conference on Innovation and Technology in Computer Science Education V. 1*, ITiCSE '21, page 220–226, New York, NY, USA. Association for Computing Machinery.

Long Ouyang, Jeff Wu, Xu Jiang, Diogo Almeida, Carroll L. Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul Christiano, Jan Leike, and Ryan Lowe. 2024. Training language models to follow instructions with human feedback. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. 2011. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12:2825–2830.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *J. Mach. Learn. Res.*, 21(1).

Kelly Rivers, Erik Harpstead, and Ken Koedinger. 2016. Learning curve analysis for programming: Which concepts do students struggle with? In *Proceedings of the 2016 ACM Conference on International Computing Education Research*, ICER '16, page 143–151, New York, NY, USA. Association for Computing Machinery.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Anna Rogers, Matt Gardner, and Isabelle Augenstein. 2023. Qa dataset explosion: A taxonomy of nlp resources for question answering and reading comprehension. *ACM Comput. Surv.*, 55(10).

Devendra Sachan, Mike Lewis, Mandar Joshi, Armen Aghajanyan, Wen-tau Yih, Joelle Pineau, and Luke Zettlemoyer. 2022. Improving passage retrieval with zero-shot question generation. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3781–3797, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Devendra Sachan, Mostofa Patwary, Mohammad Shoeybi, Neel Kant, Wei Ping, William L. Hamilton, and Bryan Catanzaro. 2021. End-to-end training of neural retrievers for open-domain question answering. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6648–6662, Online. Association for Computational Linguistics.

Gerard Salton, Edward A. Fox, and Harry Wu. 1983. Extended boolean information retrieval. *Commun. ACM*, 26(11):1022–1036.

Tanveer J. Siddiqui and Uma Shanker Tiwary. 2005. Integrating relation and keyword matching in information retrieval. In *Proceedings of the 9th International Conference on Knowledge-Based Intelligent Information and Engineering Systems - Volume Part IV*, KES'05, page 64–73, Berlin, Heidelberg. Springer-Verlag.

Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'14, page 3104–3112, Cambridge, MA, USA. MIT Press.

Yi Tay, Vinh Q. Tran, Mostafa Dehghani, Jianmo Ni, Dara Bahri, Harsh Mehta, Zhen Qin, Kai Hui, Zhe Zhao, Jai Gupta, Tal Schuster, William W. Cohen, and Donald Metzler. 2024. Transformer memory as a differentiable search index. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

Gemini Team, Rohan Anil, Sebastian Borgeaud, Yonghui Wu, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, et al. 2023. Gemini: A Family of Highly Capable Multimodal Models. *arXiv e-prints*, arXiv:2312.11805.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. LLaMA: Open and Efficient Foundation Language Models. *arXiv e-prints*, arXiv:2302.13971.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2022. Text Embeddings by Weakly-Supervised Contrastive Pre-training. *arXiv e-prints*, arXiv:2212.03533.

Liang Wang, Nan Yang, Xiaolong Huang, Binxing Jiao, Linjun Yang, Daxin Jiang, Rangan Majumder, and Furu Wei. 2023. SimLM: Pre-training with representation bottleneck for dense passage retrieval. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2244–2258, Toronto, Canada. Association for Computational Linguistics.

Yujing Wang, Yingyan Hou, Haonan Wang, Ziming Miao, Shibin Wu, Hao Sun, Qi Chen, Yuqing Xia, Chengmin Chi, Guoshuai Zhao, Zheng Liu, Xing Xie, Hao Allen Sun, Weiwei Deng, Qi Zhang, and Mao Yang. 2024. A neural corpus indexer for document retrieval. In *Proceedings of the 36th International Conference on Neural Information Processing Systems*, NIPS '22, Red Hook, NY, USA. Curran Associates Inc.

BigScience Workshop, Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, et al. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model. *arXiv e-prints*, arXiv:2211.05100.

Lee Xiong, Chenyan Xiong, Ye Li, Kwok-Fung Tang, Jialin Liu, Paul Bennett, Junaid Ahmed, and Arnold Overwijk. 2020. Approximate Nearest Neighbor Negative Contrastive Learning for Dense Text Retrieval. *arXiv e-prints*, arXiv:2007.00808.

Can Xu, Qingfeng Sun, Kai Zheng, Xiubo Geng, Pu Zhao, Jiazhan Feng, Chongyang Tao, Qingwei Lin, and Daxin Jiang. 2024. WizardLM: Empowering large pre-trained language models to follow complex instructions. In *The Twelfth International Conference on Learning Representations*.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q. Weinberger, and Yoav Artzi. 2020. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yu-Jia Zhou, Jing Yao, Zhi-Cheng Dou, Ledell Wu, and Ji-Rong Wen. 2023. Dynamicretriever: A pre-trained model-based ir system without an explicit index. *Machine Intelligence Research*, 20(2):276–288.

Honglei Zhuang, Zhen Qin, Rolf Jagerman, Kai Hui, Ji Ma, Jing Lu, Jianmo Ni, Xuanhui Wang, and Michael Bendersky. 2023. Rankt5: Fine-tuning t5 for text ranking with ranking losses. In *Proceedings of the 46th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '23, page 2308–2313, New York, NY, USA. Association for Computing Machinery.

A Dataset Details

In this section, we present several tables that detail the statistics of the datasets utilized in our study. The tables include comprehensive data such as sample sizes, feature counts, and other relevant metrics, providing an overview of the datasets' composition and scope.

B Metrics

In this section, we provide a detailed explanation of the metrics employed in our study to evaluate the effectiveness of our methods. We utilize the scikit-learn library (Pedregosa et al., 2011) to compute the metrics.

Dataset	Train	Dev	Test
TriviaQA	78,785	8,837	11,313
NQ	79,168	8,757	3,610
WebQ	3,417	361	2,032

Table 8: Statistics of TriviaQA, NQ and WebQ datasets.

Question Type	TriviaQA	NQ	WebQ
Human	36%	40%	30%
Location	21%	14%	28%
Entity	32%	11%	21%
Description	6%	8%	11%
Other	5%	27%	10%

Table 9: Distribution of TriviaQA, NQ, and WebQ datasets based on the question type.

	Training	Validation	Test
Number of questions	14,645	1,000	1,000
Number of hints	140,973	9,638	9,619
Avg. question length (words)	14.18	14.08	13.95
Avg. hint length (words)	14.98	15.07	15.14
Avg. #hints / question	9.62	9.63	9.61
Avg. #entities / question	1.35	1.40	1.35
Avg. #entities / hint	0.96	1.00	0.98
Avg. #sources / question	6.27	6.17	6.71

Table 10: Statistics of the TriviaHG dataset (Mozafari et al., 2024)

- **Accuracy (ACC):** This metric leverages LLMs to determine the correctness of the answers (Kamalloo et al., 2023).

- **Exact Match (EM):** This metric evaluates whether the retrieved passage perfectly includes the answer text without modifications.

- **Precision (PR):** This metric quantifies the proportion of words in the retrieved passage that are relevant to the answer.

- **Recall (RC):** This metric measures the extent to which words from the answer are present in the retrieved passage.

- **F1-measure (F1):** This metric is the harmonic mean of precision and recall.

- **Contains (CON):** This metric checks if the retrieved passage encompasses all vital elements of the correct answer or essential information.

- **BERTScore (BERT):** This metric (Zhang et al., 2020) calculates the semantic similarity between words in the retrieved passage and the answer, utilizing the contextual embeddings from BERT (Devlin et al., 2019).

C Additional Experimental Results

In this section, we provide a detailed presentation of the results from our experiments across various scenarios. We will explore how different conditions and variables influenced the outcomes, highlighting both expected trends and surprising findings.

# of Hints	Ranking	EM	F1	PR	RC	CON	BERT
			Zero-Shot				
2	default	20.76	25.14	26.26	25.05	26.37	75.92
2	convergence	21.1	25.43	26.57	25.32	26.62	76.03
2	t5	20.22	24.6	25.73	24.48	25.77	75.78
5	default	21.44	25.82	27.04	25.68	27.25	76.13
5	convergence	21.44	25.72	26.97	25.54	27.15	76.03
5	t5	20.98	25.4	26.64	25.23	26.81	76.05
7	default	21.57	26.01	27.21	25.89	27.48	76.22
7	convergence	21.52	25.86	27.05	25.71	27.33	76.14
7	t5	21.64	26	27.21	25.83	27.37	76.26
10	default	22.01	26.5	27.77	26.32	27.9	76.48
10	convergence	21.59	26.05	27.32	25.87	27.54	76.33
10	t5	21.82	26.25	27.49	26.07	27.76	76.35
			Few-Shot				
2	default	31.78	37.36	38.54	37.49	39.06	80.85
2	convergence	32.9	38.29	39.43	38.42	39.86	81.11
2	t5	30.12	35.5	36.51	35.68	36.73	80.23
5	default	33.44	38.92	40.09	39.07	40.77	81.36
5	convergence	34.19	39.74	40.92	39.89	41.58	81.54
5	t5	32.29	37.86	39	38.01	39.33	81
7	default	33.25	38.78	39.97	38.91	40.78	81.25
7	convergence	33.9	39.41	40.53	39.59	41.32	81.37
7	t5	32.89	38.36	39.47	38.56	40.1	81.05
10	default	33.78	39.12	40.31	39.24	41.42	81.23
10	convergence	33.7	39.17	40.34	39.31	41.34	81.24
10	t5	33.21	38.64	39.76	38.81	40.6	81.11

Table 11: The results of **T5-3b** used as the reader, employing zero-shot and few-shot learning strategies on the **TriviaQA** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-Va** method.

D Case Studies

In this section, we delve into several case studies that illustrate the prompts we have chosen, along with examples from our experiments and their respective outcomes. The case studies are designed to demonstrate the practical application of our theoretical framework and to showcase the effectiveness of our chosen methodologies in real-world scenarios.

# of Hints	Ranking	EM	F1	PR	RC	CON	BERT
			Zero-Shot				
2	default	7.92	11.03	13.18	10.4	13.32	69.81
2	convergence	8.17	11.4	13.46	10.84	13.55	69.91
2	t5	8.06	11.22	13.14	10.66	13.21	69.9
5	default	8.84	12.04	14.24	11.39	14.4	70.29
5	convergence	8.73	11.98	14.11	11.36	14.18	70.21
5	t5	8.12	11.46	13.52	10.87	13.6	70.11
7	default	9.03	12.48	14.63	11.81	14.76	70.57
7	convergence	8.81	12.21	14.33	11.58	14.38	70.4
7	t5	8.53	11.94	13.99	11.32	14.02	70.37
10	default	9.06	12.54	14.74	11.89	14.93	70.68
10	convergence	8.67	12.19	14.39	11.53	14.52	70.39
10	t5	8.59	12.01	14.15	11.37	14.4	70.42
			Few-Shot				
2	default	11.63	16.47	19.01	15.73	19.67	72.79
2	convergence	12.19	17.01	19.44	16.28	19.86	73.02
2	t5	11.08	15.75	18.09	15.04	18.61	72.57
5	default	12.33	17.42	20.1	16.61	20.22	73.42
5	convergence	12.85	18.06	20.74	17.23	20.89	73.56
5	t5	12.22	16.94	19.35	16.25	19.53	73.2
7	default	12.27	17.3	19.86	16.53	19.92	73.26
7	convergence	12.85	17.92	20.49	17.14	20.5	73.48
7	t5	12.35	17	19.32	16.31	19.31	73.27
10	default	12.47	17.57	20.18	16.78	20.17	73.3
10	convergence	12.47	17.38	19.89	16.65	19.97	73.23
10	t5	12.49	17.3	19.77	16.57	19.7	73.29

Table 12: The results of **T5-3b** used as the reader, employing zero-shot and few-shot learning strategies on the **NQ** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-Va** method.

# of Hints	Ranking	EM	F1	PR	RC	CON	BERT
			Zero-Shot				
2	default	13.24	21.09	24.99	19.65	25.15	73.1
2	convergence	13.88	21.74	25.7	20.28	25.89	73.19
2	t5	13.29	21.04	25.02	19.57	24.95	73.03
5	default	13.44	21.07	25.09	19.6	25.25	73.11
5	convergence	13.09	20.75	24.82	19.23	25.05	72.96
5	t5	12.8	20.7	24.8	19.17	24.61	73.02
7	default	13.78	21.3	25.11	19.92	25.49	73.2
7	convergence	13.44	20.93	24.81	19.5	25.2	73
7	t5	13.09	20.67	24.58	19.2	24.56	73.04
10	default	13.39	21.32	25.28	19.88	25.39	73.26
10	convergence	13.04	20.74	24.72	19.26	24.75	73.04
10	t5	13.24	21.02	25.09	19.54	25.25	73.25
			Few-Shot				
2	default	17.32	27.12	31.33	25.61	30.36	75.87
2	convergence	18.45	28.21	32.43	26.72	31.15	76.22
2	t5	16.14	26.03	30.26	24.45	29.08	75.46
5	default	17.52	27.76	32.51	26.07	31.5	76.27
5	convergence	18.9	28.97	33.63	27.25	32.43	76.52
5	t5	17.77	27.54	32.21	25.8	31.15	76.18
7	default	18.31	28.24	32.8	26.62	32.14	76.42
7	convergence	18.31	28.58	33.18	26.89	31.94	76.44
7	t5	17.96	27.8	32.5	26.1	31.55	76.37
10	default	18.06	28.21	32.9	26.51	32.14	76.42
10	convergence	18.26	28.61	33.26	26.87	32.33	76.42
10	t5	17.86	27.9	32.65	26.15	31.64	76.34

Table 13: The results of **T5-3b** used as the reader, employing zero-shot and few-shot learning strategies on the **WebQ** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-Va** method.

# of Hints	Ranking	EM	F1	PR	RC	CON	BERT
			Zero-Shot				
2	default	21.71	26.09	27.21	25.97	27.18	76.42
2	convergence	21.95	26.27	27.44	26.15	27.53	76.39
2	t5	21.4	25.79	26.94	25.67	26.83	76.3
5	default	23.27	27.63	28.82	27.53	28.98	76.8
5	convergence	23.13	27.46	28.68	27.33	28.89	76.79
5	t5	22.71	27.14	28.35	27.02	28.57	76.65
7	default	23.45	27.89	29.09	27.8	29.24	76.94
7	convergence	23.15	27.54	28.74	27.44	28.81	76.8
7	t5	22.95	27.29	28.47	27.18	28.76	76.7
10	default	23.55	28.03	29.29	27.9	29.52	76.99
10	convergence	23.38	27.85	29.1	27.73	29.2	76.92
10	t5	23.27	27.76	28.98	27.65	29.18	76.88
			Few-Shot				
2	default	35.28	41.12	42.38	41.27	43.08	82.06
2	convergence	36.14	41.99	43.27	42.13	44.14	82.19
2	t5	33.93	39.55	40.81	39.66	41.45	81.63
5	default	38.29	43.98	45.3	44.07	46.22	82.94
5	convergence	38.01	43.75	45.07	43.87	45.94	82.8
5	t5	36.7	42.54	43.9	42.63	44.78	82.5
7	default	38.54	44.29	45.62	44.39	46.5	82.94
7	convergence	38.05	43.81	45.12	43.93	45.96	82.82
7	t5	37.62	43.43	44.76	43.54	45.66	82.67
10	default	38.23	43.96	45.29	44.06	46.3	82.77
10	convergence	37.79	43.72	45.08	43.84	45.93	82.72
10	t5	37.85	43.64	45	43.75	45.87	82.74

Table 14: The results of **T5-3b** used as the reader, employing zero-shot and few-shot learning strategies on the **TriviaQA** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-FT** method.

# of Hints	Ranking	EM	F1	PR	RC	CON	BERT
			Zero-Shot				
2	default	8.86	12.4	14.4	11.82	14.18	70.8
2	convergence	9.28	12.69	14.74	12.09	14.65	70.96
2	t5	8.81	12.35	14.46	11.72	14.35	70.93
5	default	10.19	13.97	16.18	13.33	15.87	71.68
5	convergence	10.44	14.32	16.59	13.67	16.4	71.72
5	t5	10.19	13.88	16.21	13.18	16.04	71.64
7	default	10.64	14.43	16.74	13.75	16.45	71.94
7	convergence	10.47	14.26	16.6	13.57	16.34	71.85
7	t5	10.61	14.54	16.93	13.85	16.81	71.84
10	default	10.89	14.85	17.28	14.16	16.95	72.03
10	convergence	10.08	14.03	16.42	13.32	16.07	71.8
10	t5	10.22	14.29	16.8	13.56	16.62	71.86
			Few-Shot				
2	default	14.79	20.33	22.9	19.58	23.27	74.7
2	convergence	16.01	21.49	24.07	20.7	24.24	75.08
2	t5	13.66	18.95	21.47	18.23	21.55	74.29
5	default	16.54	22.38	25.14	21.6	25.32	75.51
5	convergence	16.65	22.36	25.08	21.58	25.35	75.48
5	t5	15.46	21.11	23.74	20.41	23.74	75.22
7	default	16.68	22.64	25.56	21.74	25.51	75.63
7	convergence	16.32	22.12	24.91	21.31	25.04	75.3
7	t5	15.6	21.33	24.12	20.53	23.77	75.33
10	default	16.2	21.98	24.77	21.13	24.82	75.26
10	convergence	16.01	21.7	24.6	20.82	24.76	75.21
10	t5	16.12	21.89	24.66	21.1	24.52	75.37

Table 15: The results of **T5-3b** used as the reader, employing zero-shot and few-shot learning strategies on the **NQ** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-FT** method.

# of Hints	Ranking	EM	F1	PR	RC	CON	BERT
			Zero-Shot				
2	default	13.19	21.16	25.3	19.6	25.94	72.9
2	convergence	13.58	21.82	26.09	20.2	26.38	73.07
2	t5	13.24	21.63	25.81	20.11	26.18	73.31
5	default	13.88	22.07	26.38	20.4	27.17	73.56
5	convergence	14.12	22.23	26.45	20.65	26.82	73.53
5	t5	13.78	22.07	26.36	20.44	26.57	73.42
7	default	14.42	22.78	27.19	21.09	27.41	73.77
7	convergence	13.98	21.75	25.92	20.21	26.43	73.44
7	t5	14.17	22.18	26.35	20.55	26.87	73.65
10	default	14.96	23.08	27.26	21.45	27.46	73.92
10	convergence	14.27	22.37	26.66	20.72	27.07	73.77
10	t5	14.17	22.04	26.1	20.46	26.33	73.59
			Few-Shot				
2	default	21.51	32.39	36.86	30.75	35.19	78.02
2	convergence	22	32.96	37.5	31.3	35.78	78.11
2	t5	20.37	30.89	35.4	29.35	34.06	77.64
5	default	23.43	34.41	39.14	32.64	37.65	78.74
5	convergence	23.52	34.44	39.08	32.75	37.75	78.78
5	t5	22.88	33.55	38.09	31.9	37.11	78.54
7	default	23.47	34.33	38.93	32.65	37.75	78.64
7	convergence	24.11	34.52	39.15	32.9	38.19	78.8
7	t5	23.67	34.35	38.92	32.68	37.84	78.68
10	default	23.97	34.46	39	32.87	37.84	78.66
10	convergence	23.62	34.49	39.07	32.84	38.04	78.7
10	t5	23.82	34.21	38.79	32.58	37.5	78.57

Table 16: The results of **T5-3b** used as the reader, employing zero-shot and few-shot learning strategies on the **WebQ** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-FT** method.

# of Hints	Ranking	ACC	EM	F1	PR	RC	CON	BERT
				Zero-Shot				
2	default	56.12	0	7.99	4.69	43.36	53.48	70.25
2	convergence	55.98	0	8	4.7	43.49	54.11	70.26
2	t5	54.16	0	7.93	4.67	41.85	51.55	70.08
5	default	58.25	0	7.91	4.66	41.87	54.48	70.28
5	convergence	58.25	0	7.93	4.67	42.34	54.76	70.26
5	t5	57.56	0	7.85	4.64	41.33	53	70.12
7	default	58.63	0	7.98	4.72	40.99	54.17	70.33
7	convergence	59.06	0	8.04	4.75	41.51	54.74	70.35
7	t5	59.12	0	7.98	4.71	41.13	53.29	70.24
10	default	59.52	0	8.14	4.82	40.28	53.95	70.42
10	convergence	59.46	0	8.1	4.8	40.86	54.22	70.37
10	t5	59.69	0	8.03	4.75	41.04	53.66	70.29
				Few-Shot				
2	default	55.62	49.49	58.69	58.53	63.34	64.6	88
2	convergence	55.32	50.05	59.29	59.19	63.84	65	88.07
2	t5	55.85	48.52	58.06	57.79	62.75	63.59	87.66
5	default	57.62	50.42	60.15	59.94	64.5	65.54	88.55
5	convergence	57.71	50.76	60.6	60.4	65.12	65.92	88.61
5	t5	57.96	49.35	59.36	58.96	63.89	64.64	88.25
7	default	58.24	50.1	60.11	59.79	64.51	65.41	88.53
7	convergence	58.5	50.3	60.52	60.23	64.96	65.85	88.59
7	t5	58.27	49.88	59.88	59.54	64.34	65.33	88.41
10	default	58.32	49.48	59.44	59.16	63.87	64.9	88.31
10	convergence	58.4	49.97	59.85	59.49	64.31	65.12	88.39
10	t5	58.49	49.44	59.5	59.11	64.2	64.86	88.33

Table 17: The results of **LLaMA-7b** used as the reader, employing zero-shot and few-shot learning strategies on the **TriviaQA** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-Va** method.

# of Hints	Ranking	ACC	EM	F1	PR	RC	CON	BERT
				Zero-Shot				
2	default	48.25	0	3.16	1.77	25.37	23.52	64.47
2	convergence	50.55	0	3.29	1.84	25.97	24.88	64.56
2	t5	47.56	0	3.18	1.78	24.15	22.08	64.34
5	default	48.67	0	3.35	1.89	25.22	25.18	64.86
5	convergence	50.36	0	3.48	1.97	25.74	26.12	65
5	t5	48.75	0	3.3	1.87	25.11	24.43	64.68
7	default	48.95	0	3.39	1.92	25.06	25.54	65.02
7	convergence	50.58	0	3.4	1.92	25.5	26.04	65.02
7	t5	49.97	0	3.36	1.9	24.89	25.54	64.88
10	default	50	0	3.49	1.98	25.14	26.04	65.05
10	convergence	51.11	0	3.44	1.95	25.71	26.2	64.97
10	t5	51.86	0	3.36	1.9	25.5	25.82	64.91
				Few-Shot				
2	default	54.68	14.04	21.34	21.03	29.8	28.75	72.98
2	convergence	55.48	14.76	21.83	21.5	30.18	28.89	72.78
2	t5	53.8	14.24	21.37	21.09	29.61	28.06	72.77
5	default	57.81	17.59	25.36	25.08	33.14	31.94	75.06
5	convergence	58.45	18.31	26.41	26.27	34.33	32.74	75.21
5	t5	57.42	17.42	25.3	25.07	33.04	31.36	74.92
7	default	58.75	17.92	25.99	25.66	34.01	32.74	75.46
7	convergence	59.36	18.48	26.61	26.36	34.58	33.24	75.58
7	t5	58.45	18.12	26.28	26	34.02	31.94	75.41
10	default	58.14	18.06	26.14	25.8	34.29	33.19	75.55
10	convergence	58.75	18.34	26.52	26.18	34.71	33.68	75.61
10	t5	59.31	18.2	26.15	25.9	34.09	32.3	75.44

Table 18: The results of **LLaMA-7b** used as the reader, employing zero-shot and few-shot learning strategies on the **NQ** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-Va** method.

# of Hints	Ranking	ACC	EM	F1	PR	RC	CON	BERT
				Zero-Shot				
2	default	51.62	0	5.51	3.23	36.43	35.73	66.98
2	convergence	53.49	0	5.4	3.17	35.32	35.93	66.9
2	t5	52.36	0	5.41	3.18	35.35	33.96	66.75
5	default	50.94	0	5.75	3.38	36.99	39.07	67.5
5	convergence	51.43	0	5.73	3.36	37.33	39.27	67.43
5	t5	50.15	0	5.65	3.32	35.91	36.91	67.28
7	default	52.12	0	5.82	3.42	37.14	38.98	67.5
7	convergence	52.02	0	5.77	3.39	37.16	38.98	67.42
7	t5	51.48	0	5.82	3.42	36.63	37.8	67.29
10	default	52.21	0	5.86	3.45	38.04	39.76	67.36
10	convergence	52.95	0	5.83	3.42	38.15	40.26	67.37
10	t5	51.87	0	5.76	3.39	36.52	38.24	67.28
				Few-Shot				
2	default	49.56	13.93	26.1	26.17	37.77	38.53	74.39
2	convergence	53.1	13.68	25.96	25.79	36.95	36.96	74.22
2	t5	48.87	14.52	27.18	27.08	38	38.44	74.74
5	default	55.27	16.29	30.02	30.13	42.1	43.55	76.17
5	convergence	56	17.22	30.82	30.79	43.3	44.93	76.38
5	t5	54.23	16.68	30.52	30.4	42.05	43.16	76.3
7	default	55.56	16.54	31	31.03	43.09	44.34	76.54
7	convergence	55.87	17.52	32.1	32.13	44.22	44.88	76.87
7	t5	55.12	16.73	30.76	30.88	42.54	43.9	76.39
10	default	55.76	16.68	30.71	30.71	42.77	44.64	76.45
10	convergence	55.95	16.49	30.84	30.87	43.11	44.69	76.5
10	t5	55.51	16.39	30.15	30.4	41.91	44	76.31

Table 19: The results of **LLaMA-7b** used as the reader, employing zero-shot and few-shot learning strategies on the **WebQ** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-Va** method.

# of Hints	Ranking	ACC	EM	F1	PR	RC	CON	BERT
				Zero-Shot				
2	default	49.89	0	8.85	5.23	44.22	59.71	71.18
2	convergence	49.34	0	8.69	5.14	44.1	59.19	71.02
2	t5	50.69	0	8.68	5.13	43.23	56.85	70.93
5	default	53.58	0	8.86	5.26	43.17	61.02	71.33
5	convergence	53.91	0	8.74	5.18	43.48	60.74	71.21
5	t5	54	0	8.69	5.15	42.95	59.49	71.14
7	default	54.54	0	8.92	5.31	42.5	61.12	71.38
7	convergence	54.76	0	8.85	5.24	42.98	61.16	71.33
7	t5	54.36	0	8.8	5.22	42.5	60.13	71.27
10	default	54.97	0	8.96	5.33	42.21	60.93	71.4
10	convergence	55.26	0	8.86	5.25	42.3	60.72	71.36
10	t5	55.04	0	8.87	5.27	42.31	60.69	71.35
				Few-Shot				
2	default	54.9	53.93	63.29	63.57	66.87	68.54	89.49
2	convergence	55.91	53.24	62.7	62.94	66.64	68.31	89.18
2	t5	55.7	52.79	62.22	62.33	66.06	67.29	89.14
5	default	57.22	54.31	64.4	64.56	68.17	69.64	89.85
5	convergence	57.35	54.57	64.48	64.61	68.15	69.55	89.8
5	t5	57.57	54.07	63.91	64	67.55	68.88	89.71
7	default	57.66	54.3	64.39	64.49	68.22	69.73	89.87
7	convergence	58.06	54.62	64.66	64.75	69.53	70.15	89.89
7	t5	57.47	54.06	64	64.02	67.86	69.52	89.78
10	default	57.55	54.09	64.14	64.16	68.08	69.65	89.76
10	convergence	57.52	54.61	64.58	64.59	68.46	69.83	89.88
10	t5	58.13	54.06	64.17	64.18	68.11	69.72	89.8

Table 20: The results of **LLaMA-7b** used as the reader, employing zero-shot and few-shot learning strategies on the **TriviaQA** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-FT** method.

# of Hints	Ranking	ACC	EM	F1	PR	RC	CON	BERT
				Zero-Shot				
2	default	39.62	0	3.87	2.19	26.58	29.86	66.12
2	convergence	38.62	0	3.82	2.16	26.27	30.36	65.95
2	t5	42.14	0	3.86	2.18	25.74	27.48	65.82
5	default	47.32	0	4.32	2.45	27.45	32.94	66.7
5	convergence	46.65	0	4.27	2.43	27.77	32.91	66.55
5	t5	47.76	0	4.11	2.34	26.34	30.94	66.36
7	default	48.76	0	4.3	2.45	26.73	32.94	66.74
7	convergence	48.12	0	4.29	2.44	26.81	32.8	66.66
7	t5	49.15	0	4.2	2.39	26.68	32.3	66.57
10	default	49.51	0	4.34	2.47	26.74	33.02	66.79
10	convergence	49.26	0	4.38	2.5	26.96	33.19	66.8
10	t5	49.76	0	4.28	2.44	26.52	33.02	66.78
				Few-Shot				
2	default	58.95	18.28	26	26.24	32.76	30.69	74.73
2	convergence	62.13	17.42	25	25.16	32.47	30.64	74.24
2	t5	60.47	16.76	24.59	24.74	31.64	28.98	74.18
5	default	64.68	19.92	28.59	28.95	35.41	35.54	76.28
5	convergence	64.96	20.25	28.83	28.95	35.77	36.01	76.25
5	t5	64.07	18.53	27.17	27.45	33.77	32.63	75.63
7	default	64.43	20.22	28.99	29.04	36.01	36.09	76.61
7	convergence	64.43	20.72	29.47	29.55	37.19	36.81	76.7
7	t5	64.82	19.7	28.6	28.82	35.62	34.85	76.17
10	default	63.96	20	28.98	29.05	36.77	36.2	76.48
10	convergence	64.38	20.55	29.3	29.53	36.17	36.32	76.58
10	t5	64.76	20	29.11	29.36	36.4	35.65	76.47

Table 21: The results of **LLaMA-7b** used as the reader, employing zero-shot and few-shot learning strategies on the **NQ** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-FT** method.

# of Hints	Ranking	ACC	EM	F1	PR	RC	CON	BERT
				Zero-Shot				
2	default	47.33	0	6.23	3.65	38.76	41.49	67.96
2	convergence	46.25	0	6.19	3.64	39.29	41.54	67.81
2	t5	49.75	0	6.2	3.62	37.08	38.48	67.91
5	default	52.7	0	6.7	3.93	39.31	43.7	68.58
5	convergence	52.45	0	6.66	3.92	39.18	43.21	68.48
5	t5	52.6	0	6.68	3.94	38.38	42.18	68.48
7	default	53.58	0	6.9	4.06	39.84	44.34	68.66
7	convergence	53.44	0	6.83	4.03	39.42	44.14	68.56
7	t5	53.39	0	6.87	4.05	38.43	43.01	68.59
10	default	53.93	0	6.98	4.11	40.02	44.39	68.65
10	convergence	54.08	0	7.01	4.14	40.04	45.23	68.79
10	t5	53.98	0	6.95	4.1	39.62	44.09	68.65
				Few-Shot				
2	default	48.97	16.68	29.06	29.21	40.11	41.58	76
2	convergence	54.13	16.83	28.81	28.89	40.12	41.14	75.65
2	t5	51.03	16.58	29.13	29.17	39.69	40.75	75.84
5	default	55.56	18.95	33.73	33.64	45.62	47.93	77.89
5	convergence	56.15	19.69	34.52	34.36	46.04	48.38	78.12
5	t5	55.51	19.88	34.7	34.57	46.11	47.54	78.1
7	default	56.05	18.9	34.42	34.2	46.44	48.57	78.34
7	convergence	56.55	20.28	35.35	35.31	47.32	49.9	78.51
7	t5	55.76	20.23	35.22	34.93	46.75	49.02	78.5
10	default	56.25	18.85	34.68	34.51	47.15	49.75	78.36
10	convergence	55.91	18.9	33.68	33.5	46.21	48.77	78.02
10	t5	56.94	18.85	34.4	34.13	46.98	49.66	78.09

Table 22: The results of **LLaMA-7b** used as the reader, employing zero-shot and few-shot learning strategies on the **WebQ** dataset, are analyzed based on different ranking methods and a range of hint quantities. These hints were generated using the **HiGen-FT** method.

Hint Generator	# of Parameters	# of Hints	Ranking	EM	F1
		Zero-Shot			
LLaMA-Va	7b	10	default	33.0	36.82
WizardLM 70b	70b	7	t5	34.0	37.02
LLaMA-FT	7b	2	default	35.0	39.43
LLaMA-Va	13b	7	default	35.0	40.39
LLaMA-Va	70b	10	default	35.0	37.63
LLaMA-FT	13b	2	t5	38.0	40.73
LLaMA-FT	70b	7	default	38.0	41.23
Copilot	-	10	default	39.0	43.62
GPT 3.5	175b	5	default	39.0	43.2
Gemini	-	10	t5	42.0	**46.19**
GPT 4	-	10	default	**43.0**	44.5
		Few-Shot			
LLaMA-Va	7b	2	default	37.0	40.37
LLaMA-FT	7b	10	t5	42.0	44.72
LLaMA-Va	13b	5	default	45.0	48.44
LLaMA-Va	70b	2	default	45.0	47.17
LLaMA-FT	13b	7	default	46.0	49.22
LLaMA-FT	70b	7	t5	47.0	48.22
WizardLM	70b	5	default	48.0	52.4
GPT 3.5	175b	10	t5	49.0	51.47
Copilot	-	5	t5	52.0	56.03
Gemini	-	10	default	56.0	57.67
GPT 4	-	10	default	**59.0**	**60.5**

Table 23: The performance of T5-3b across different LLMs as the central component of the HiGenQA system, generating hints for 100 questions.

Question	Retriever	LLaMA-70b	HiGenQA	True Answer
		TriviaQA		
How many dot positions are usually used in each letter of the Braille system?	**6**	**six**	**six**	**6, six**
Who was the leader of the gang whose members included Benny the Ball ,Brain and Choo Choo?	the bowery boys	**top cat**	**top cat**	**top cat**
Which Glasgow group signed to Creation Records and recorded their debut single "All Fall Down", in 1985?	**primal scream**	the pastels	the jesus and mary chain	**primal scream**
Who is the only man to win a best actor Oscar playing brothers?	jack nicholson	daniel day	henry fonda	**lee marvin**
		NQ		
who played taylor on the bold and beautiful?	**hunter tylo**	**hunter tylo**	**hunter tylo**	**hunter tylo**
who wrote the song going to kansas city?	bo diddley	**jerry leiber**	**jerry leiber**	**jerry leiber**
what part of the brain is in the middle cranial fossa?	**the pituitary gland**	temporal lobe region	the hippocampus	**the pituitary gland**
who did the broncos beat in the super bowl?	the packers	green bay	the falcons	**carolina panthers**
		WebQ		
where are boeing headquarters?	**chicago**	**chicago**	**seattle**	**seattle, chicago**
what university did obama graduated from?	harvard law school	harvard law school	**columbia university**	**columbia university**
what country did buddha come from?	**india**	india	nepal	**india**
who played amy squirrel in bad teacher?	cameron diaz	**lucy punch**	cameron diaz	**lucy punch**

Table 24: Comparison of answers for randomly selected questions from the TriviaQA, NQ, and WebQ datasets.

Question: what city of USA has a neighborhood called little havana?

Answer: Miami

Candidate Answers:
1. Havana
2. Washington D.C.
3. San Francisco
4. Chicago
5. New York City
6. Los Angeles
7. Houston
8. Dallas
9. Atlanta
10. Philadelphia
11. Miami

Vanilla

Hints	**Candidates**
This city is located in the southeastern United States.	0 0 0 0 0 0 0 0 1 1 1
The city is known for its vibrant Cuban culture and cuisine.	1 0 0 0 0 0 0 0 0 0 0
The city has a large Hispanic population.	0 1 0 1 1 1 1 1 1 1 1
The neighborhood is home to many Cuban restaurants and shops.	0 0 0 0 0 0 0 0 0 0 0
The city is located near the Atlantic Ocean.	0 1 0 0 1 0 0 0 1 1 1
The city has a tropical climate.	1 0 0 0 0 0 0 0 0 0 1
The neighborhood is known for its colorful street art and murals.	0 1 1 0 0 1 0 0 0 1 1
The city has a rich history of Cuban immigration.	0 0 0 0 1 0 0 0 0 0 1
The neighborhood is a popular destination for foodies and cultural enthusiasts.	0 0 1 0 1 0 0 0 0 0 1

Finetuned

Hints	**Candidates**
It is the largest city in Florida.	0 0 0 0 0 0 0 0 0 0 0
It's known for its vibrant nightlife, especially in a place called South Beach.	0 0 0 0 0 0 0 0 0 0 1
It's the only major U.S. city to be founded by a woman.	0 0 0 0 0 0 0 0 0 0 1
It's home to one of the largest cruise ship ports in the world.	0 0 0 0 1 1 1 0 0 0 1
It is nicknamed the "Capital of Latin America".	1 0 0 0 0 0 0 0 0 0 1
The city is known for its Art Deco Historic District.	1 0 0 0 0 0 0 0 0 0 1
The city is often at risk from hurricanes due to its location.	1 0 0 0 1 0 1 0 0 0 1
It is located in the southeastern part of the state.	0 0 0 0 0 0 0 0 1 0 0
It is the 44th-most populous city in the United States.	0 0 0 1 1 0 1 0 1 1 0
It is the 16th-most populous metropolitan area in the United States.	0 0 0 1 1 1 1 1 1 1 0

Table 25: Hints generated by the HiGen-Va and HiGen-FT methods and the candidate answers they encompass. The 'Candidates' column displays which candidate answers are included in each hint, indicated by the index of candidate answers; for example, the first bit confirms the inclusion of 'Havana' and so on.

Question: what city of USA has a neighborhood called little havana?
Answer: Miami

MSS-DPR:
`According to the following context, answer the question:`
`Context:` rejected efforts to have separate residential and commercial land-use districts in 1948, 1962, and 1993. Consequently, rather than a single central business district as the center of the city's employment, multiple districts have grown throughout the city in addition to Downtown, which include Uptown, Texas Medical Center, Midtown, Greenway Plaza, Memorial City, Energy Corridor, Westchase, and Greenspoint. Houston has the fifth-tallest skyline in North America (after New York City, Chicago, Toronto and Miami) and 36th-tallest in the world . A seven-mile (11 km) system of tunnels and skywalks links downtown buildings containing shops and restaurants, enabling pedestrians to avoid summer
`Question:` what city of USA has a neighborhood called little havana?
`Answer:` *There is no mention of a neighborhood called "Little Havana" in the provided context.*

LLaMA-70b:
`According to the following context, answer the question:`
`Context:` While Miami's Little Havana is more well-known than New York City's Little Havana, New York City's Little Havana is a vibrant neighborhood in the Bronx, specifically in the South Bronx area.
`Question:` what city of USA has a neighborhood called little havana?
`Answer:` *According to the context, the city with a neighborhood called Little Havana is New York City (specifically, the Bronx)*

HiGenQA:
`According to the following context, answer the question:`
`Context:` It's known for its vibrant nightlife, especially in a place called South Beach. It's the only major U.S. city to be founded by a woman. It's home to one of the largest cruise ship ports in the world. The city is known for its Art Deco Historic District, which showcases colorful buildings and neon signs. The city is often at risk from hurricanes due to its location.
`Question:` what city of USA has a neighborhood called little havana?
`Answer:` *Based on the context, the city with a neighborhood called Little Havana is Miami.*

Table 26: Case study of the retrieved passage from MSS-DPR, generated context by LLaMA-70b, and hints generated by HiGenQA on LLaMA 7b in Zero-Shot. Words in blue indicate the correct answer, while those in red represent other potential answers.

Question: what city of USA has a neighborhood called little havana?
Answer: Miami

MSS-DPR:
`According to the following context, answer the question:`
`Context:` Red Sandy Spika dress of Reba McEntire American recording artist Reba McEntire wore a sheer red dress to the 1993 Country Music Association Awards ceremony on September 29, 1993. The sheer fabric was covered with sequins, and cut with a low neckline. The garment was designed by stylist Sandy Spika, and McEntire wore it during a duet performance of "Does He Love You" with Linda Davis. McEntire later said, "I got more press off that dress than if I'd won entertainer of the year." According to McEntire, when her little sister, Susie, saw her on stage she leaned over and.
`Question:` who sings does he love me with reba?
`Answer:` Linda Davis

`According to the following context, answer the question:`
`Context:` the introduction of The National Endowment for the Oceans, Coasts, and Great Lakes Act. This proposal is meant to preserve the ecosystems that coastal communities and economies depend on. Ocean Champions Ocean Champions, a 501(c)(4) environmental organization in the United States with a connected political action committee (Ocean Champions PAC), is the first national organization of its kind focused solely on oceans and ocean wildlife. Their goal is to create a political environment where protecting and restoring the oceans is a national government priority. They do this by helping to elect pro-ocean Congressional candidates and working to defeat the others.
`Question:` where do the great lakes meet the ocean?
`Answer:` the Saint Lawrence River

`According to the following context, answer the question:`
`Context:` would be joining the cast as Melissa Shield and Katsuhisa Namase would play David Shield, both original characters. On June 11, 2018, "Weekly Shōdnen Jump" announced that Rikiya Koyama had been cast as the film's villain, Wolfram. Masaki Suda performs the film's theme song , which was written and composed by Hiromu Akita of amazarashi. Funimation and Toho premiered the film at Anime Expo in Los Angeles on July 5, 2018, and it was later released in Japan on August 3 of that year. The first one million audience members to see the movie will receive a special book containing.
`Question:` when does the new my hero academia movie come out?
`Answer:` July 5 , 2018

`According to the following context, answer the question:`
`Context:` Sphenic number In number theory, a sphenic number (from , 'wedge') is a positive integer that is the product of three distinct prime numbers. A sphenic number is a product "pqr" where "p", "q", and "r" are three distinct prime numbers. This definition is more stringent than simply requiring the integer to have exactly three prime factors. For instance, $60 = 2 \times 3 \times 5$ has exactly 3 prime factors, but is not sphenic. The smallest sphenic number is $30 = 2 \times 3 \times 5$, the product of the smallest three primes. The first few sphenic numbers are.
`Question:` what is the smallest prime number that is greater than 30?
`Answer:` 31

`According to the following context, answer the question:`
`Context:` She then cleans up the room and leaves. The next day, soon after Letty and Jerry have arrived at the home of his parents, a detective from New York arrives looking for Letty and demanding that she come with him. Jerry, Mrs. Lynton and Letty's maid accompany her to see District Attorney John J. Haney, who produces the letters and accuses Letty of murder. After she admits that she went to see Emile, Jerry lies by saying that he and Letty spent the night together at his apartment after she left Emile's, and that he knew all about the letters.
`Question:` who plays letty in bring it on all or nothing?
`Answer:` Francia Raisa

`According to the following context, answer the question:`
`Context:` rejected efforts to have separate residential and commercial land-use districts in 1948, 1962, and 1993. Consequently, rather than a single central business district as the center of the city's employment, multiple districts have grown throughout the city in addition to Downtown, which include Uptown, Texas Medical Center, Midtown, Greenway Plaza, Memorial City, Energy Corridor, Westchase, and Greenspoint. Houston has the fifth-tallest skyline in North America (after New York City, Chicago, Toronto and Miami) and 36th-tallest in the world . A seven-mile (11 km) system of tunnels and skywalks links downtown buildings containing shops and restaurants, enabling pedestrians to avoid summer
`Question:` what city of USA has a neighborhood called little havana?
`Answer:` *Houston*

Table 27: Case study of the retrieved passage from MSS-DPR retriever on LLaMA-7b in Few-Shot. Words in blue indicate the correct answer, while those in red represent other potential answers.

Question: what city of USA has a neighborhood called little havana?
Answer: Miami

LLaMA-70b:
`According to the following context, answer the question:`
`Context:` The Nobel Peace Prize for 2009 was awarded to President Barack Obama "for his extraordinary efforts to strengthen international diplomacy and cooperation between nations."
`Question:` Who won the Nobel Peace Prize in 2009?
`Answer:` Barack Obama

`According to the following context, answer the question:`
`Context:` Edouard Daladier became Prime Minister of France in 1933, not in 1933. He actually served as Prime Minister of France from 1933 to 1934, and then again from 1938 to 1940.
`Question:` Edouard Daladier became Prime Minister of which country in 1933?
`Answer:` France

`According to the following context, answer the question:`
`Context:` Mahonia japonica, also known as Japanese Mahonia or Oregon Grape, is a deciduous shrub that produces yellow flowers in the late winter to early spring. In most regions, you can expect to see it in bloom from February to April, depending on your location and climate. In warmer climates, such as USDA Hardiness Zones 8-10, the flowers may appear as early as January. In cooler climates, such as USDA Hardiness Zones 5-7, the flowers may bloom later, from March to May. After the flowers have faded, Mahonia japonica produces blue-black berries that attract birds and other wildlife. The berries are an important food source for many animals, making this shrub a valuable addition to your garden or landscape. Keep in mind that the exact blooming period may vary depending on your specific location, weather conditions, and the age and health of your shrub.
`Question:` If you have a 'Mahonia Japonica', in which season will it be in flower?
`Answer:` Winter

`According to the following context, answer the question:`
`Context:` The sport played under the "Harvard Rules" is Rugby. The "Harvard Rules" refer to the rules of Rugby that were used in the United States in the late 19th and early 20th centuries, particularly at Harvard University.
`Question:` Which sport is played under the 'Harvard Rules'?
`Answer:` Rugby

`According to the following context, answer the question:`
`Context:` After some research, I found that the person made an honorary citizen of Haslach, Austria in 1938, and had the honor withdrawn in 2004, was Adolf Hitler.
`Question:` Who was made an honorary citizen of Haslach, Austria, in 1938, an honour withdrawn in 2004?
`Answer:` Adolf Hitler

`According to the following context, answer the question:`
`Context:` While Miami's Little Havana is more well-known than New York City's Little Havana, New York City's Little Havana is a vibrant neighborhood in the Bronx, specifically in the South Bronx area.
`Question:` what city of USA has a neighborhood called little havana?
`Answer:` *New York City*

Table 28: Case study of the context generated using LLaMA-70b on LLaMA-7b in Few-Shot. Words in blue indicate the correct answer, while those in red represent other potential answers.

Question: what city of USA has a neighborhood called little havana?
Answer: Miami

HiGenQA:
`According to the following context, answer the question:`
`Context:` He was the 44th President of the United States. He served as President from 2009 to 2017. He was the first African-American President of the United States. He was a member of the Democratic Party. He was born on August 4, 1961 in Honolulu, Hawaii. Before becoming President, he represented Illinois in the U.S. Senate from 2005 to 2008. He was awarded the Nobel Peace Prize in 2009 for his extraordinary efforts to strengthen international diplomacy and cooperation between peoples. He is married to Michelle Robinson and they have two children, Malia and Sasha. He graduated from Columbia University with a BA and from Harvard University with a JD. After graduating from Columbia University, he worked as a community organizer in Chicago.
`Question:` Who won the Nobel Peace Prize in 2009?
`Answer:` Barack Obama

`According to the following context, answer the question:`
`Context:` The capital city of this country is Paris. This country is located in northwestern Europe. This country has a long history and has played a significant role in international affairs. The official language of this country is French. The currency used in this country is the Euro. This country has several major mountain ranges, including the Alps and the Pyrenees. This country is known for its cuisine, fashion, art, and architecture. This country is a member of the United Nations Security Council and is an official nuclear-weapon state. This country has a diverse landscape, with beautiful beaches, towering mountains, and fertile plains. This country has a rich cultural heritage and is home to many world-renowned museums and landmarks.
`Question:` Edouard Daladier became Prime Minister of which country in 1933?
`Answer:` France

`According to the following context, answer the question:`
`Context:` Its the coldest season of the year. Its the season when snow falls in many regions. Its the season when many people celebrate Christmas and New Year's Eve. Its the season when days are shorter and nights are longer. Its the season when many animals hibernate. Its the season when people often wear warm clothes like coats, hats, and gloves. Its the season when many trees lose their leaves. Its the season when many people go skiing or snowboarding. Its the season that comes after autumn and before spring. Its the season when hot cocoa and warm fires are popular.
`Question:` If you have a 'Mahonia Japonica', in which season will it be in flower?
`Answer:` Winter

`According to the following context, answer the question:`
`Context:` It is a team sport that originated in the United States. It is played with an oval-shaped ball. The objective of the game is to score points by advancing the ball into the opposing team's end zone. Points can be scored by carrying the ball across the opponent's goal line, throwing it to a teammate in the end zone, or kicking it through the opponent's goalposts. The game is divided into four quarters, each lasting 15 minutes. The team with possession of the ball, known as the offense, attempts to advance down the field by running or passing the ball.The opposing team, known as the defense, tries to stop the offense and take control of the ball for themselves.The offense must advance at least 10 yards in four downs or plays; if they fail, they turn over the ball to the opposing team. If the offense succeeds in advancing 10 yards or more, they are given a new set of four downs to continue their drive towards the end zone. The game is played on a rectangular field that measures 120 yards long and 53.3 yards wide.
`Question:` Which sport is played under the 'Harvard Rules'?
`Answer:` AMERICAN FOOTBALL

`According to the following context, answer the question:`
`Context:` He was born on April 20, 1889 in Braunau am Inn, Austria. He was the leader of the Nazi Party. He became the chancellor of Germany in 1933. He took the title of Führer und Reichskanzler in 1934. He initiated World War II in Europe by invading Poland on September 1, 1939. He was closely involved in military operations throughout the war. He was central to the perpetration of the Holocaust. He committed suicide on April 30, 1945.His father's name was Alois and he was born in 1837. His mother's name was Klara and she died after much suffering in 1907.
`Question:` Who was made an honorary citizen of Haslach, Austria, in 1938, an honour withdrawn in 2004?
`Answer:` Adolf Hitler

`According to the following context, answer the question:`
`Context:` It's known for its vibrant nightlife, especially in a place called South Beach. It's the only major U.S. city to be founded by a woman. It's home to one of the largest cruise ship ports in the world. The city is known for its Art Deco Historic District, which showcases colorful buildings and neon signs. The city is often at risk from hurricanes due to its location.
`Question:` what city of USA has a neighborhood called little havana?
`Answer:` *Miami*

Table 29: Case study of the hints generated using HiGenQA on LLaMA-7b in Few-Shot. Words in blue indicate the correct answer.