KCS: Diversify Multi-hop Question Generation with Knowledge Composition Sampling

Yangfan Wang[1], **Jie Liu**[1,3], **Chen Tang**[2], **Lian Yan**[1], **Jingchi Jiang**[1,3]*
[1]Harbin Institute of Technology, [2]MemTensor (Shanghai) Technology Co., Ltd.
[3]National Key Laboratory of Smart Farm Technologies and Systems
{yf.wang,23b903008}@stu.hit.edu.cn, travistang@foxmail.com
{jieliu,jiangjingchi}@hit.edu.cn

Abstract

Multi-hop question answering faces substantial challenges due to data sparsity, which increases the likelihood of language models learning spurious patterns. To address this issue, prior research has focused on diversifying question generation through content planning and varied expression. However, these approaches often emphasize generating simple questions and neglect the integration of essential knowledge, such as relevant sentences within documents. This paper introduces the **Knowledge Composition Sampling (KCS)**, an innovative framework designed to expand the diversity of generated multi-hop questions by sampling varied knowledge compositions within a given context. KCS models the knowledge composition selection as a sentence-level conditional prediction task and utilizes a probabilistic contrastive loss to predict the next most relevant piece of knowledge. During inference, we employ a stochastic decoding strategy to effectively balance accuracy and diversity. Compared to competitive baselines, our KCS improves the overall accuracy of knowledge composition selection by 3.9%, and its application for data augmentation yields improvements on HotpotQA and 2WikiMultihopQA datasets. Our code is available at: https://github.com/yangfanww/kcs.

1 Introduction

Multi-hop Question Answering (MHQA) presents unique challenges in natural language processing (Panda et al., 2024), requiring the selection and integration of multiple knowledge pieces to accurately answer complex questions. Despite advancements facilitated by high-quality MHQA datasets (Yang et al., 2018; Ho et al., 2020), data sparsity remains a significant issue, increasing the risk for language models learning spurious patterns and compromising robustness and generalization. Question Generation (QG) has been proposed to augment QA datasets (Guo et al., 2022). Recent efforts to diversify QG focus on enhancing question diversity through content planning and expression, aiming to capture the one-to-many nature of QG tasks (Gou et al., 2023; Narayan et al., 2022; Deschamps et al., 2021; Cao and Wang, 2021; Holtzman et al., 2020; Shen et al., 2019; Cho et al., 2019). These methods often concentrate on the generation of simple questions and fail to introduce external knowledge, such as relevant sentences from documents, to generate more complex and diverse questions. In reality, without incorporating external knowledge and relying solely on simple textual methods, the diversity of questions doesn't truly improve. Identifying important sentences that encapsulate contextually relevant knowledge is essential for facilitating the model's understanding of the underlying logic in the data, thus generating questions deemed semantically diverse and meaningful by humans (Du and

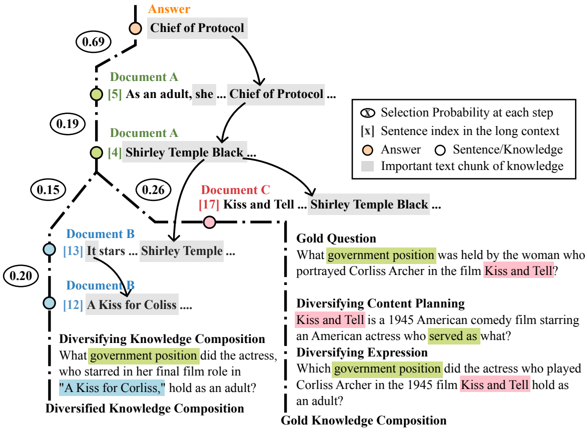

Answer
Chief of Protocol
0.69
Document A
[5] As an adult, she ... Chief of Protocol ...
0.19
Document A
[4] Shirley Temple Black ...
0.15
0.26
Document C
[17] Kiss and Tell ... Shirley Temple Black ...
Document B
[13] It stars ... Shirley Temple ...
0.20
Document B
[12] A Kiss for Coliss ....
Ⓧ Selection Probability at each step
[x] Sentence index in the long context
● Answer ○ Sentence/Knowledge
▨ Important text chunk of knowledge
Diversifying Knowledge Composition
What government position did the actress, who starred in her final film role in "A Kiss for Corliss," hold as an adult?
Diversified Knowledge Composition
Gold Question
What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?
Diversifying Content Planning
Kiss and Tell is a 1945 American comedy film starring an American actress who served as what?
Diversifying Expression
Which government position did the actress who played Corliss Archer in the 1945 film Kiss and Tell hold as an adult?
Gold Knowledge Composition

Figure 1: This figure illustrates the diversity of contextual knowledge through a tree structure. Starting from an answer (root), each subsequent knowledge piece is selected from the long context based on a probability (edge). Each knowledge composition is depicted as a branch. Unlike prior methods that consistently select the same knowledge composition as the gold standard (right), KCS tends to select varied branches to enhance the diversity of generated multi-hop questions (left).

[*] Corresponding author.

Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing, pages 23173–23185
November 4-9, 2025 ©2025 Association for Computational Linguistics

Cardie, 2017). Given the inherent complexity of MHQA (a task that integrates multiple pieces of knowledge), this step is crucial for diversifying multi-hop QG. We assert that data sparsity stems from the *underutilization of contextual knowledge*. As illustrated in Figure 1, selecting a different set of slightly varied knowledge pieces (termed a knowledge composition) can lead a QG model to generate a significantly distinct multi-hop question at the knowledge level, even when the answer remains unchanged. This observation suggests that contextual knowledge is not fully utilized in existing MHQA datasets. The detialed example is shown in Appendix B.

To address these challenges, we propose the **Knowledge Composition Sampling (KCS)**, a novel framework based on knowledge diversity that facilitates the utilization of contextual knowledge by sampling varied knowledge compositions within a given context, thus enhancing multi-hop question diversity. Our framework contains three main components: (1) knowledge composition selection to accurately select knowledge compositions from context; (2) diversifying knowledge composition to efficiently sample accurate and diverse knowledge compositions; and (3) multi-hop question generation using a vanilla model to generate multi-hop questions from the given answer and sampled knowledge compositions.

However, as shown in Figure 5, arbitrary knowledge compositions pose the risk of *degeneration*[1]. To mitigate this risk, we frame knowledge composition selection as a sentence-level conditional prediction problem and utilize a probabilistic contrastive loss to learn the potential knowledge coherence. During training, the selection model maximizes mutual information between the latent prediction representation of the current timestep and the latent representation of the next timestep, while minimizing mutual information with other latent representations within the context. To balance the accuracy and diversity, we employ a stochastic decoding strategy that truncates the unreliable tail of the probability distribution and samples the next knowledge piece from a dynamic nucleus. As shown in Figure 1, the selection of each subsequent sentence is conditional on the answer, context, and previously selected sentences. By modeling the conditional probabilities of each timestep and employing stochastic sampling, KCS effectively obtain diversified and accurate knowledge compositions, thereby producing diverse and high-quality multi-hop questions. In summary, unlike traditional methods that rely on structured graphs or textual methods on fixed knowledge compositions, our KCS leverages unstructured text to discern the potential knowledge coherence, enhancing flexibility and scalability. Our contributions are as follows:

- We propose the KCS framework designed to expand the diversity of generated multi-hop questions by sampling varied knowledge compositions within a given context;
- We introduce a novel sentence-level conditional prediction task and a probabilistic contrastive loss to discern potential knowledge coherence, and verify the effectiveness of a stochastic decoding strategy;
- Experiments on HotpotQA and 2WikiMultihopQA demonstrate that KCS improves the overall accuracy of knowledge composition selection by 3.9%, and its use for data augmentation achieves consistent improvements of downstream performance.

2 Related Works

Important Sentence Selection The initial step in QG task involves identifying sentences within the context that are question-worthy, i.e., sentences that humans consider valuable for generating questions. Previous research Du and Cardie (2017) predominantly formalizes this task as sentence classification, often focusing on simple question generation. In our study, we introduce sentence-level sequence prediction to enhance the selection of knowledge pieces for multi-hop question generation (MHQG).

Multi-hop Question Generation Previous research (Kumar et al., 2019; Pan et al., 2020; Fei et al., 2022; Hwang et al., 2024) frequently relies on pre-constructed knowledge graphs, entity graphs, or document graphs to enable controllable MHQG. This dependency often results in errors and increased costs associated with entity extraction and graph construction. Alternative approaches leverage in-context learning with large language models (Lin et al., 2024) or employ supervised fine-tuning of pretrained language models

[1] Holtzman et al. (2020) define “degeneration” as the production of automatically generated text that is generic and repetitive. Narayan et al. (2022) define it as text that is unfaithful or inconsistent with the input. In our context, “degeneration” refers to irrelevant knowledge leading to simpler or inconsistent multi-hop questions.

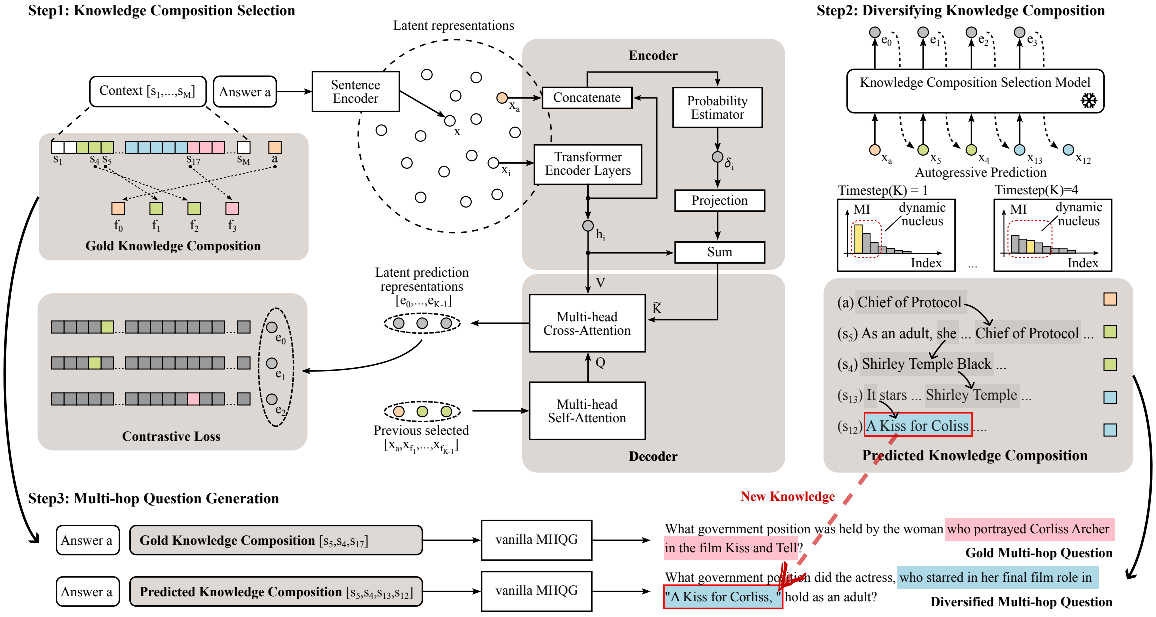

Step1: Knowledge Composition Selection
Context [s1,...,sM]
Answer a
Sentence Encoder
Latent representations
x
xa
xi
s1
s4 s5
s17
sM
a
f0
f1
f2
f3
Gold Knowledge Composition
Encoder
Concatenate
Probability Estimator
δi
Transformer Encoder Layers
Projection
hi
Sum
V
K̃
Multi-head Cross-Attention
Q
Multi-head Self-Attention
Decoder
Latent prediction representations [e0,...,eK-1]
Previous selected [xa,xf1,...,xfK-1]
e0
e1
e2
Contrastive Loss
Step2: Diversifying Knowledge Composition
e0
e1
e2
e3
Knowledge Composition Selection Model
xa
x5
x4
x13
x12
Autogressive Prediction
Timestep(K) = 1
Timestep(K)=4
MI
dynamic nucleus
Index
...
(a) Chief of Protocol
(s5) As an adult, she ... Chief of Protocol ...
(s4) Shirley Temple Black ...
(s13) It stars ... Shirley Temple ...
(s12) A Kiss for Coliss ....
Predicted Knowledge Composition
New Knowledge
Step3: Multi-hop Question Generation
Answer a
Gold Knowledge Composition [s5,s4,s17]
vanilla MHQG
What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?
Gold Multi-hop Question
Answer a
Predicted Knowledge Composition [s5,s4,s13,s12]
vanilla MHQG
What government position did the actress, who starred in her final film role in "A Kiss for Corliss, " hold as an adult?
Diversified Multi-hop Question

Figure 2: Illustration of our KCS framework.

(Murakhovs'ka et al., 2022; Xia et al., 2023) to achieve consistent MHQG. Our framework samples varied knowledge compositions from the given context for diversifying MHQG, allowing for seamless integration with existing MHQG methods. In our study, we fine-tune a vanilla model proposed by (Murakhovs'ka et al., 2022) for MHQG.

Diversifying Question Generation Recent works in diversifying QG have aimed to enhance the diversity of generated questions across two dimensions: content planning and expression. For content planning, methods such as (Cho et al., 2019) sample diverse token-level content, while others (Narayan et al., 2022; Holtzman et al., 2020) employ advanced decoding strategies at the token level. To achieve expression diversity, studies such as (Deschamps et al., 2021; Gou et al., 2023) utilize external knowledge to generate multiple expressions of the same question. However, these methods primarily focus on simple questions and assume that important sentences (knowledge compositions) in context are pre-identified and fixed. In contrast, our work addresses the entire QG process and expands the diversity of generated multi-hop questions through sampling diverse knowledge-level compositions.

3 Method

3.1 Overview

We formally define the problem of diversifying multi-hop QG as follows. Given a long context D, which consists of a set of documents, and an answer $a = [a_1, \dots, a_{T_a}]$, the objective is to generate a set of diverse multi-hop questions $Q = \{q_1, \dots, q_{N_q}\}$, conditioned on both D and a. Here, T_a denotes the number of tokens in a, and N_q represents the number of generated questions. Each document within the context is composed of multiple sentences, such that the context $D = [s_1, \dots, s_M]$ contains M sentences in total. We distinguish between the long context D, which encapsulates extensive knowledge, and a knowledge composition $c = [f_1, \dots, f_K]$, a subset of D containing K question-worthy sentences. Previous approaches term c as "context" and typically model the conditional probability $p(q|c, a)$, with the assumption that c is pre-identified and fixed. In contrast, we address the entire pipeline of diversifying multi-hop QG without such assumptions, modeling the problem as follows:

$$p(q|D, a) = \mathbb{E}_c[p(q|c, a) \times p(c|D, a)] \quad (1)$$

The overall architecture of our KCS framework is illustrated in Figure 2. The knowledge composition selection component learns a Transformer-based sentence-level sequence prediction model to accurately select a knowledge composition c for the given answer a and long context D. The diversifying knowledge composition component employs a stochastic decoding strategy to sample diverse knowledge compositions $C = \{c_1, \dots, c_{N_q}\}$ from D. The multi-hop question generation component utilizes a vanilla MHQG model to generate a multi-hop question q conditioned on each sampled knowledge composition $c \in C$ and the given answer a. The algorithm of diversifying multi-hop QG is shown in Appendix E.

3.2 Knowledge Composition Selection

We employ a hierarchical neural network architecture for sentence-level sequence prediction to select accurate knowledge compositions. As depicted in Step1 of Figure 2, the hierarchical neural network architecture includes a sentence encoder $\mathcal{M}_{\text{enc}}$, a sentence-level sequence model $\mathcal{M}_{\text{seq}}$, and two objectives $\mathcal{L}_{\text{cls}}$ and $\mathcal{L}_{\text{seq}}$, which are aligned with human intuitions. Formally, the classification objective $\mathcal{L}_{\text{cls}}$ involves assigning the correct label $z \in [0, 1]$ to each sentence $s \in D$, resulting in $Z = [z_1, \ldots, z_M]$. The sequence prediction objective $\mathcal{L}_{\text{seq}}$ involves predicting the correct knowledge composition c auto-regressively, with $f_0 = a$ for convenience:

$$\begin{aligned} p(c|D, a) &= p([f_1, \ldots, f_K]|D, a) \\ &= \prod_{k=1}^{K} p(f_k|D, Z, f_{0:k-1}) \end{aligned} \tag{2}$$

First, the BERT-based $\mathcal{M}_{\text{enc}}$ (Devlin et al., 2019) maps each sentence $s \in D \cup \{a\}$ to a latent representation $x = \mathcal{M}_{\text{enc}}(s)$. For the long context D, we obtain $X = [x_1, \ldots, x_M] \in \mathbb{R}^{M \times d}$, where d is the hidden state dimension. Next, the sentence-level Transformer-based model $\mathcal{M}_{\text{seq}}$ (Vaswani et al., 2017) extends the Transformer's encoder to produce knowledge classification probabilities Z, which are then infused from its encoder into its decoder.

$$\begin{aligned} H &= \text{EncoderLayers}(X) \\ z_i &= \text{Softmax}(\text{Linear}(h_i; x_a)) \\ e_{k-1} &= \text{DecoderLayers}(x_{f_{\leq k-1}}, H, Z) \end{aligned} \tag{3}$$

Specifically, the encoder of $\mathcal{M}_{\text{seq}}$ encodes the latent representations X to hidden states $H = [h_1, \ldots, h_M] \in \mathbb{R}^{M \times d}$. Each h_i is then concatenated with the latent answer representation x_a, and a linear network with softmax serves as the probability estimator to obtain the knowledge classification probability $p(z_i|x_a, h_i)$. We input hidden states H to the multi-head cross-attention layer of the decoder as the value state but modify the key state as $\tilde{K} = H + \delta W^{\delta}$ where $\delta_i = [1 - z_i, z_i]$, using a linear projection with parameter $W^{\delta} \in \mathbb{R}^{2 \times d}$. The decoder consumes the latent representations of previously selected sentences $f_{\leq k-1} \subset D$ to produce a latent prediction representation e_{k-1} and predict the conditional selection probability $p(s|e_{k-1})$.

Inspired by the noise-contrastive estimation in (Oord et al., 2018), we discard low-level information and noise and introduce a probabilistic contrastive loss for next step prediction to learn the potential knowledge coherence. Formally, given the long context $D = [s_1, \ldots, s_M]$, we treat the next f_k as the sole positive sample from $p(s|e_{k-1})$ and others as negative samples from the distribution $p(s)$ at the $(k-1)^{\text{th}}$ timestep. We maximize mutual information MI between the latent prediction representation e_{k-1} of the current timestep and the latent representation x_{f_k} of the next timestep, while minimizing mutual information with other latent representations in the context, optimizing the probabilistic contrastive loss $\mathcal{L}_{\text{seq}}$:

$$\mathcal{L}_{\text{seq}} = -\mathbb{E}_c \left[\log \frac{\text{MI}(x_{f_k}, e_{k-1})}{\sum\limits_{s \in D, s \neq f_k} \text{MI}(x_s, e_{k-1})} \right] \tag{4}$$

where MI is a mutual information function. The final loss is $\mathcal{L} = \mathcal{L}_{\text{cls}} + \lambda \mathcal{L}_{\text{seq}}$, where $\mathcal{L}_{\text{cls}}$ is the classification loss, and λ is a hyper-parameter.

3.3 Diversifying Knowledge Composition

While maximization-based decoding strategies are effective in selecting accurate knowledge compositions, they often lack diversity. As shown in Figure 3, the conditional probability distribution $p(s|e_{k-1})$ of our knowledge composition selection model exhibits an unreliable tail that requires truncation during generation. We introduce a stochastic decoding strategy to efficiently sample accurate and diverse knowledge compositions. Figure 3 aligns with the example in Figure 1 and demonstrates the effectiveness of this stochastic decoding strategy.

Inspired by (Holtzman et al., 2020), we employ a nucleus sampling to the hierarchical neural network, using the shape of the probability distribution to determine the set of sentences to be sampled at each timestep. We truncate the unreliable tail of the conditional probability distribution at each timestep, and then sample the next sentence from a dynamic nucleus of sentences that contains the majority of the probability mass. Formally, given the conditional probability distribution $p(s|e_{k-1})$, we define its top-p nucleus $D^{(p)} \subset D$ as the smallest set such that $\sum_{s \in D^{(p)}} p(s|e_{k-1}) \geq p$. The original distribution is then rescaled to form a new distribution p_{new} from which the next sentence is sampled.

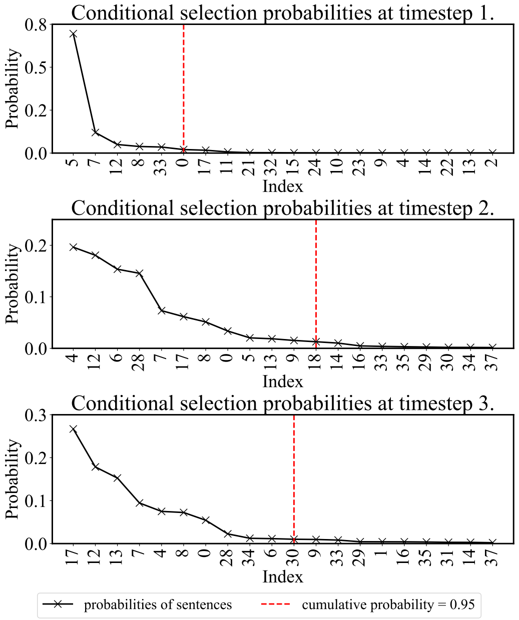


Figure 3: Conditional probabilities for knowledge selection at each timestep, with sentence indexes in descending order of probabilities. The red dashed line indicates the cutoff of $p = 0.95$. We sequentially select the sentence with the highest probability as the next knowledge for greedy sampling, i.e., [5,4,17] form the knowledge composition.

3.4 Multi-hop Question Generation

As illustrated in Figure 2, our KCS framework separates the diversification process from the generation process. The one-to-many process is accomplished at the knowledge level through the previous two components, resulting in accurate and diverse knowledge compositions. Given each identified knowledge composition c and the answer a, the multi-hop question generation component aims to generate a consistent multi-hop question q. For the same answer and varied knowledge compositions $C = [c_1, \ldots, c_{N_q}]$, we generate a set of diverse multi-hop questions $Q = \{q_1, \ldots, q_{N_q}\}$.

Specifically, we fine-tune a vanilla MixQG-base model (Murakhovs'ka et al., 2022) on the training data, and employ the standard CrossEntropy loss to generate consistent multi-hop questions. It is noteworthy that although we use the MixQG-base model, many advanced multi-hop question generation methods can be applied to improve the consistency of the generated multi-hop questions.

4 Experiments

4.1 Experimental Settings

Datasets We evaluate our method on two popular benchmark MHQA datasets: HotpotQA (Yang et al., 2018) and 2WikiMultihopQA (Ho et al., 2020). For knowledge composition selection, due to the inaccessibility of the test set for HotpotQA and the absence of supporting facts in the test set of 2WikiMultihopQA, we designate the original development set as the new test set and randomly extract 500 samples from the training set to serve as the new development set. Samples with answers of "yes" or "no" are excluded, as they do not contribute to the selection of subsequent knowledge. Detailed statistics of these datasets are provided in Appendix A. For diversifying question generation, since the evaluation datasets do not contain all possible valid questions for an answer, we indirectly evaluate the quality of generated questions through performance on downstream tasks, as a positive exploration. We randomly extract 5000 samples from the training set to serve as the original training set for data augmentation and further filter 200 samples from the new test set where each large language model (LLM) achieves the lowest Recall score to construct a test set to exclude the effect of the LLMs' base capabilities.

Metrics For knowledge composition selection, we employ Precision (P), Recall (R), and F1-Score (F1) across different lengths of knowledge composition ($K = 2, 3$) as automatic metrics to assess the accuracy of the selected knowledge compositions. For diversifying question generation, we use Exact Match (EM), Precision (P), Recall (R), and F1-Score (F1) to evaluate the correctness of the predicted answers in the downstream MHQA task. Additionally, we compute BERTScore (BSc) and Human Evaluation Score (HSc) for semantic evaluation. Following (Gou et al., 2023), we also employ Pairwise-BLEU to measure the diversity by averaging sentence-level metrics of pairs within generated N_q questions, and LLM-based metrics to measure the diversity and consistency.

4.2 Baselines

Knowledge Composition Selection we compare our model against five categories of baselines to identify question-worthy sentences that encapsulate contextually relevant knowledge:

(1) RETRIEVAL: We retrieve K relevant sentences

Method	Base Model	HotpotQA						2WikiMultihopQA					
		P@2	*R@2*	*F1@2*	*P@3*	*R@3*	*F1@3*	*P@2*	*R@2*	*F1@2*	*P@3*	*R@3*	*F1@3*
Retrieval	*bge-small*	48.83	41.49	44.33	38.19	48.82	42.22	40.04	35.50	37.02	30.69	41.01	34.53
	bge-large-v1.5	53.56	45.79	48.66	41.27	52.78	45.63	44.02	39.12	40.75	33.47	44.73	37.66
	+step by step	53.78	46.03	48.90	42.52	54.29	46.98	50.91	43.58	46.03	40.67	52.24	44.80
	bge-m3	49.29	41.97	44.68	37.98	48.35	41.90	41.06	36.53	38.04	30.55	41.02	34.47
	text-ada-002	47.49	40.19	42.88	37.05	46.91	40.76	42.29	37.41	39.03	31.65	42.14	35.54
	BM25	54.44	46.70	49.57	40.79	52.34	45.18	48.10	43.42	44.98	35.10	47.38	39.71
	+step by step	54.08	46.60	49.37	40.75	52.44	45.20	53.90	47.96	49.94	41.29	54.44	46.12
CLS	*BERT-base*	62.70	54.05	57.26	48.62	62.42	53.87	<u>79.90</u>	<u>69.68</u>	<u>73.09</u>	<u>65.06</u>	<u>82.55</u>	<u>71.20</u>
	BERT-large	58.00	49.52	52.64	44.83	59.69	49.30	73.25	64.25	67.25	61.50	79.00	67.74
	RoBERTa-large	23.00	19.52	20.83	20.33	25.93	22.47	44.75	36.88	39.50	37.83	46.25	40.60
Sentence Graph	*CommonEntity*	46.39	40.03	42.40	33.11	42.74	36.79	51.20	45.24	47.23	37.09	48.68	41.33
	Similarity	40.95	35.40	37.46	27.39	35.50	30.50	41.62	37.87	39.12	27.88	38.07	31.74
LLM	*Qwen2.5-14B*	18.10	15.18	16.26	14.77	18.49	16.16	19.24	16.09	17.14	16.68	20.84	18.11
	Qwen2.5-7B	19.26	16.12	17.28	15.90	19.92	17.40	25.05	21.58	22.74	19.61	25.20	21.61
	Llama3.1-8B	7.21	5.95	6.41	6.81	8.45	7.42	10.96	8.84	9.55	10.61	12.85	11.33
	Llama3.2-3B	9.62	8.08	8.65	8.11	10.18	8.88	13.73	10.91	11.85	11.71	14.00	12.42
	GPT-4	52.10	43.73	46.83	45.10	56.76	49.47	49.74	43.59	45.64	40.10	52.56	44.65
	DeepSeek-V3	49.93	41.66	44.71	43.08	53.95	47.14	45.77	40.15	42.02	37.96	50.09	42.43
BASE	*Random*	6.10	5.02	5.41	6.13	7.60	6.67	8.60	6.95	7.50	9.01	10.96	9.64
	MAX	*100.00*	*87.01*	*91.85*	*78.00*	*97.52*	*85.33*	*100.00*	*89.68*	*93.12*	*73.55*	*94.85*	*81.19*
KCS	*BERT-base*	**64.18**	**55.46**	**58.70**	**50.12**	**64.33**	**55.52**	**84.79**	**75.18**	**78.39**	**66.21**	**84.66**	**72.75**
	BERT-large	<u>63.29</u>	<u>54.71</u>	<u>57.90</u>	<u>49.09</u>	<u>62.97</u>	<u>54.37</u>	75.34	66.65	69.55	56.54	74.44	63.11
	RoBERTa-large	45.42	39.32	41.59	37.04	47.70	41.11	74.21	65.74	68.56	55.58	73.75	62.30

Table 1: Main results of knowledge composition selection on HotpotQA and 2WikiMultihopQA. *MAX* and *Random* represent the upper and lower bounds of knowledge composition selection performance, respectively. The **Bold** and <u>underline</u> mark the best and second-best results, excluding *MAX*.

in context as a knowledge composition either all at once or step by step, using only the answer or the answer concatenated previously retrieved sentences as input.

(2) CLS: We concatenate the answer with each sentence in context to perform a binary classification, then select the top-K sentences that exhibit the positive label as a knowledge composition.

(3) Sentence Graph: We construct a sentence graph based on common entities or sentence similarity, then randomly select one sentence containing the answer as the start node and perform a random walk on this graph to select $K-1$ sentences. These K sentences construct a knowledge composition.

(4) LLM: We employ a zero-shot approach to prompt the large language models to generate K question-worthy sentences as a knowledge composition for each sample, with the context and answer as input.

(5) BASE: We randomly select K sentences as a knowledge composition (*Random*) or assume that all sentences selected are question-worthy (*MAX*). *MAX* and *Random* represent the upper and lower bounds of knowledge composition selection performance, respectively.

Diversifying Question Generation We evaluate the effect of diversifying question generation for data augmentation in the downstream MHQA task. Two popular LLMs, Llama3.1 (8B) and Qwen2.5 (7B), are employed as baseline models. Beyond supervised finetuning the baseline models on the original training data (*ORI*), we compare three typical approaches for diversifying question generation:

(1) *RAST* (Gou et al., 2023): This approach utilizes external knowledge to generate multiple expressions of the same question.

(2) *Composition* (Narayan et al., 2022): This approach employs nucleus sampling to extract diverse entity chains from the context and beam search to guide question generation.

(3) *GPT-4*: We employ a zero-shot approach to prompt GPT-4 (Achiam et al., 2023) to generate N_q diverse multi-hop questions with the context and answer as input for each sample in original training data.

4.3 Implementation Details

We utilize BERT-base[2] and MixQG-base[3] as our foundational models for knowledge composition selection and multi-hop question generation, re-

[2] https://huggingface.co/google-bert/bert-base-uncased

[3] https://huggingface.co/Salesforce/mixqg-base

LLM	Diversifying Method	HotpotQA						2WikiMultihopQA					
		EM	*P*	*R*	*F1*	*BSc*	*HSc*	*EM*	*P*	*R*	*F1*	*BSc*	*HSc*
Llama3.1	*ORI*	54.00	66.07	65.29	64.46	58.74	69.05	49.00	58.32	58.54	57.64	51.36	62.65
	+*RAST*	55.50	67.80	68.33	66.49	59.93	71.40	55.00	63.50	63.75	62.88	57.83	67.65
	+*Composition*	55.00	67.17	67.48	66.09	61.82	72.65	60.50	67.45	67.75	66.92	63.91	72.45
	+*GPT4*	49.50	60.81	61.45	59.75	52.43	65.55	55.00	62.62	63.95	62.51	58.28	67.60
	+*KCS*	**61.00**	**73.59**	**72.58**	**72.04**	**66.51**	**77.75**	**66.50**	**74.25**	**75.20**	**73.91**	**69.53**	**78.10**
Qwen2.5	*ORI*	39.50	54.90	51.46	51.36	47.16	60.40	37.50	44.45	44.02	43.77	37.74	52.60
	+*RAST*	34.50	49.57	46.66	45.96	39.86	54.85	40.00	45.62	45.57	45.23	39.93	53.80
	+*Composition*	40.50	55.70	52.29	52.02	46.43	59.70	44.50	49.52	49.72	49.27	43.20	56.90
	+*GPT-4*	32.50	45.93	42.51	42.33	35.14	50.75	41.50	46.27	47.41	46.27	39.74	53.90
	+*KCS*	**45.00**	**60.49**	**58.54**	**57.40**	**51.56**	**67.55**	**51.50**	**56.19**	**57.16**	**56.35**	**51.44**	**64.00**

Table 2: Comparison of performance for Llama3.1 and Qwen2.5 on HotpotQA and 2WikiMultihopQA. Original represents training LLMs using the original training data. The **Bold** and underline mark the best and second-best results.

spectively. To train the knowledge composition selection model, we preprocess the supporting facts in the training data to to obtain the (gold) knowledge composition. Specifically, for each training sample, we prioritize the document containing the answer, and within the same document, we adhere to the contextual sentence order. The sentence containing the answer is used to split the sentences of the document into two sequences that maintain the contextual order, with the sequence containing the answer positioned earlier. The model is trained with λ of 1 and MI is a cosine function. For knowledge composition selection, we employ a greedy sampling strategy for evaluation. For diversifying knowledge composition, KCS generates $N_q = 5$ knowledge compositions for each sample, using a nucleus sampling strategy with top-$p = 0.95$, and each composition contains $K = 3$ pieces of knowledge. These obtained knowledge compositions are then used to generate N_q diverse multi-hop questions. Our implementation is in PyTorch[4], using AdamW for optimization with a learning rate of 3×10^{-5} and a linear warmup ratio of 0.1.

For the downstream MHQA task, we focus on the distractor setting, where supporting documents include distractor documents, challenging the model to handle noise in the input. After obtaining the augmented training data by diversifying methods with $N_q = 5$, we fine-tune LLMs using LoRA (Hu et al., 2022), with a learning rate of 5×10^{-5}, a cosine warmup ratio of 0.1, a LoRA rank of 8, and a LoRA alpha of 32. We conduct experiments with Composition Sampling (*Composition*) (Narayan et al., 2022) by ourselves, fine-tuning Pegasus[5], and employing the nucleus sampling and beam search to obtain diverse entity compositions and generate the most-likely multi-hop questions, respectively.

4.4 Main Results

Knowledge Composition Selection Table 1 presents the results of knowledge composition selection on HotpotQA and 2WikiMultihopQA datasets. Each block includes a category of baselines. Our KCS method consistently achieves the highest scores for P, R, and F1 metrics across all datasets (HotpotQA and 2WikiMultihopQA) and knowledge composition lengths ($K = 2, 3$). Among the baselines, the classification (CLS) category yields the best results due to the high correlation between knowledge compositions and answers in datasets. Compared to CLS using BERT-base, our KCS improves the overall accuracy by 3.9%. Specifically, on HotpotQA, KCS achieves 58.70 (F1@2) and 55.52 (F1@3) with the knowledge composition length $K = 2, 3$, respectively. These scores represent 63.90% and 65.06% of the upper bound (*MAX* of BASE), and outperform the most competitive CLS baseline by about 1.5% and 3%. On 2WikiMultihopQA, KCS demonstrates even stronger performance, with 84.18% and 89.60% of the upper bound and about 7.2% and 2.1% outperform to CLS. Surprisingly, the performance of KCS is further improved when the length of knowledge composition is increased from 2 to 3 during inference. This phenomenon indicates that KCS has higher accuracy in predicting longer knowledge compositions and has effectively learned the potential knowledge coherence. Visual comparison of the accuracy is shown in Figure 4. Our ablation study further investigates the significant impact of knowledge composition length and order on model

[4] https://pytorch.org/

[5] https://huggingface.co/google/pegasus-large

performance during training. Despite lack of annotated sequential relationships between knowledge and answers in training data, results indicate that KCS efficiently achieves cost-effectively and high-performance knowledge composition selection.

Diversifying Question Generation The MHQA performance of LLAMA3.1 and QWEN2.5 fine-tuned on data augmented by recent diversifying methods is shown in Table 2. The results indicate that using KCS for data augmentation achieves consistent improvements on both HotpotQA and 2WikiMultihopQA datasets, which illustrate that sampled knowledge compositions are meaningful to a certain extent. And we provide a detailed case study in Figure 6. To further illustrate that the sampled knowledge compositions are logically coherent, we conduct a LLM evaluation and an actual human analysis in Appendix D. Collectively, these results substantiate that the pieces of knowledge in sampled knowledge compositions can be logically combined to generate valid multi-hop questions.

To better understand the impact of question diversity on downstream MHQA performance, we assess the consistency and diversity of generated questions in Table 3. The results indicate that *KCS* achieves high diversity, with consistency limited by the vanilla MHQG method. As discussed in Section 3.4, advanced multi-hop question generation methods can address this issue. *RAST* and *Composition* exhibit higher consistency due to repeated high-consistency questions, which negatively impacts diversity. Although *GPT-4* achieves high diversity and consistency, it incurs prohibitive costs. We attribute the helpful improvement on MHQA performance to the effectiveness of KCS in balancing high question diversity with minimal noise.

Metrics	**Diversifying Methods**			
	Composition	*RAST*	*GPT-4*	*KCS*
Pairwise-BLEU (↓)	89.5	71.4	**15.0**	68.1
LLM-Diversity (↑)	29.0	38.6	**79.8**	46.2
LLM-Consistency (↑)	74.8	69.6	**91.8**	68.0

Table 3: Diversity and consistency of different diversifying methods on HotpotQA. The **Bold** and underline mark the best and second-best results of each row.

4.5 Ablation Study

We investigate the impact of various factors on the performance of the KCS framework.

Decoder-Only vs. Encoder-Decoder Architectures We compare decoder-only and encoder-decoder architectures for knowledge composition selection. We remove the encoder of $\mathcal{M}_{\text{seq}}$, and retrain the $\mathcal{M}_{\text{enc}}$ and the decoder of $\mathcal{M}_{\text{seq}}$ to form the decoder-only architecture (*Decoder*). The *Decoder* autoregressively selects the next sentence from the long context, using the answer and previously selected sentences as input. As shown in Table 4, the decoder-only model (*Decoder*, the top block) achieves 47.74 (F1@1), 55.94 (F1@2) and 52.89 (F1@3), indicating the critical role of the decoder in KCS performance. Incorporating the encoder and a classification objective (*Encoder-Decoder*, the bottom block) further improves performance by 1.26, 2.76 and 2.63 points, respectively. The *-decoder* setup in the bottom block, which uses encoder classification probabilities for knowledge composition selection, demonstrates a significant drop by 16.12, 15.08 and 11.62, also highlighting the importance of the decoder.

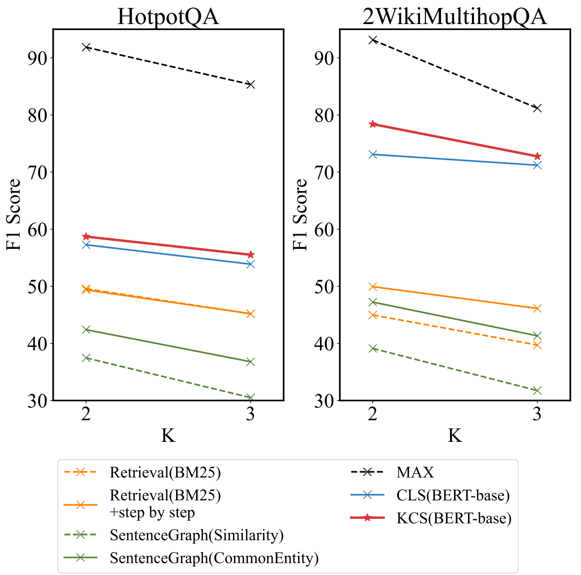


Figure 4: Accuracy comparison of competitive baselines for knowledge composition selection on HotpotQA and 2WikiMultihopQA.

Knowledge Composition Order and Length in Training We assess the *Decoder* performance across various sentence arrangements within knowledge compositions: (1) *original*: Original sequence of sentences in the dataset. (2) *shuffle*: Randomized order of sentences at each training epoch. (3) *sorted*: Sentences arranged according to their contextual order. (4) *cluster*: Consecutive sentences grouped together, prioritizing clusters that contain answers. (5) *cropping*: A maximum of 2 sentences per composition. (6) *document*: Documents containing answers are ranked highly and then reordered based on their contextual order in the document to create a (gold) knowledge composition. As shown in Table 4, the *document* arrangement proves most efficient, with different orders causing a maximum performance drop by 17.2 (F1@1), 5.56 (F1@2) and 5.29 (F1@3). A cropping length of 2 results in a performance drop by 4.49 (F1@1), 4.46 (F1@2) and 3.43 (F1@3). These results indicate that both order and length of knowledge composition during training significantly influence KCS performance.

Pre-Encoder vs. Post-Encoder Concatenation We compare accuracy using answer representation concatenated with context representations before (*pre*) and after (*post*) encoder layers in the *Encoder-Decoder* architecture. Results in Table 4 indicate that *post* outperforms *pre* by 0.2 (F1@1), 0.96 (F1@2) and 0.97 (F1@3) points, respectively.

Hyper-parameters of Architecture and Training Based on the best setup (*Encoder-Decoder* with *post*), we further explore the effects of expanding transformer layers and attention heads on model performance (*4l8h*). Our experiments show that expanding transformer layers (2 → 4) and attention heads (4 → 8) does not yield improvements, resulting in drops by 1.77 (F1@1), 2.47 (F1@2) and 2.27 (F1@3) points. Additionally, adjusting the λ parameter from 1 to 0.5 (*0.5*λ) results in performance drops of 0.76 (F1@1), 0.99 (F1@2) and 1.05 (F1@3) points, highlighting the importance of the probabilistic contrastive loss.

w/o Integration of Knowledge Classification Probabilities The *cls&gen* setup does not integrate the knowledge classification probabilities into the decoder, i.e. hidden states H are used as both the key and the value states. Results in Table 4 show that *cls&gen* results in performance drops of 0.08 (F1@1), 0.77 (F1@2) and 0.82 (F1@3) points, indicating that the integration of knowledge classification probabilities is more effective.

4.6 Case Study

To further analyze the performance of KCS on diversifying question generation, we present a case study in Appendix C. As illustrated in Figure 6, our KCS method diversifies subsequent knowledge based on the given answer, context, and previously selected knowledge. As the length of knowledge compositions increases, the generated multi-hop questions become more specific and complex. Compared to other baselines for diversifying QG, questions generated by KCS exhibit greater diversity at the knowledge level. Although the question generated based on the knowledge composition `[10, 43]` is inconsistent with the answer "Loveless", we attribute this to the absence of robust and advanced multi-hop question generation methods rather than knowledge selection issues, since we can easily utilize "My Bloody Valentine" as bridging content in sentences `10` and `43` to formulate a multi-hop question.

Model	*F1@1*	*F1@2*	*F1@3*
Decoder			
+*original*	30.54	50.38	47.60
+*shuffle*	42.74	51.48	50.05
+*sorted*	42.72	51.66	49.90
+*cluster*	47.21	54.63	51.68
+*cropping*	42.79	51.48	49.46
+*document*	47.74	55.94	52.89
Encoder-Decoder			
+*pre*	48.80	57.74	54.55
+*after*	**49.00**	**58.70**	**55.52**
+*4l8h*	47.23	56.23	53.25
+*cls&gen*	<u>48.92</u>	<u>57.93</u>	<u>54.70</u>
+*0.5*λ	48.24	57.71	54.47
−*decoder*	32.88	43.62	43.90

Table 4: Results of ablation study for the KCS. The gray row indicate the best for each block.

5 Conclusion

This paper introduces KCS, a novel framework designed to expand the diversity of generated multi-hop questions by sampling varied knowledge compositions within a given long context. Unlike prior methods that rely on structured graphs or fixed knowledge compositions, KCS leverages unstructured text to discern the potential knowledge coherence, making it more flexible and scalable. To mitigate the risk of degeneration, we propose sentence-level conditional prediction and a probabilistic contrastive loss to learn the potential knowledge coherence. To balance the accuracy and diversity, we employ a stochastic decoding strategy that truncates the unreliable tail of the probability distribution and samples the next knowledge piece from a dynamic nucleus. Comprehensive experiments show that KCS improves the overall accuracy of knowledge composition selection and its application for data augmentation enhances downstream performance.

Limitations

Our work currently exists several limitations and future directions: (1) KCS still has room for optimization. Explore different ways to calculate mutual information (such as cosine or inner product), various model architectures[6], advanced multi-hop question generation methods and the availability of high-quality annotated data may help improve KCS; (2) When use for data augmentation, investigating the performance of KCS on domain-specific data can help mitigate data sparsity challenges in domains with low resources; (3) KCS demonstrates the potential of Transformer model in high-level representation prediction and the advantages of probabilistic contrastive loss, which may inspire other similar works; (4) Not any diverse data is beneficial to downstream task improvement, and figuring out what kind of data is valuable for downstream tasks is also a direction worth studying.

Acknowledgments

We gratefully acknowledge the support of the National Natural Science Foundation of China (Grant No. 62350710797), the Key Research and Development Program of Heilongjiang Province, China [2024ZX01A07], and the Science and Technology Innovation Award of Heilongjiang Province, China [JD2023GJ01].

References

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*.

Shuyang Cao and Lu Wang. 2021. Controllable open-ended question generation with a new question type ontology. In *ACL 2021*, pages 6424–6439.

Jaemin Cho, Minjoon Seo, and Hannaneh Hajishirzi. 2019. Mixture content selection for diverse sequence generation. In *EMNLP 2019*, pages 3121–3131.

Arthur Deschamps, Sujatha Das Gollapalli, and See-Kiong Ng. 2021. On generating fact-infused question variations. In *RANLP 2021*, pages 335–345.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL 2019*, pages 4171–4186.

Xinya Du and Claire Cardie. 2017. Identifying where to focus in reading comprehension for neural question generation. In *EMNLP 2017*, pages 2067–2073.

Zichu Fei, Qi Zhang, Tao Gui, Di Liang, Sirui Wang, Wei Wu, and Xuanjing Huang. 2022. CQG: A simple and effective controlled generation framework for multi-hop question generation. In *ACL 2022*, pages 6896–6906.

Qi Gou, Zehua Xia, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li, and Nguyen Cam-Tu. 2023. Diversify question generation with retrieval-augmented style transfer. In *EMNLP 2023*, pages 1677–1690.

Shasha Guo, Jing Zhang, Yanling Wang, Qianyi Zhang, Cuiping Li, and Hong Chen. 2022. DSM: Question generation over knowledge base via modeling diverse subgraphs with meta-learner. In *EMNLP 2022*, pages 4194–4207.

Xanh Ho, Anh-Khoa Duong Nguyen, Saku Sugawara, and Akiko Aizawa. 2020. Constructing a multi-hop QA dataset for comprehensive evaluation of reasoning steps. In *ICCL 2020*, pages 6609–6625.

Ari Holtzman, Jan Buys, Li Du, Maxwell Forbes, and Yejin Choi. 2020. The curious case of neural text degeneration. In *ICLR 2020*.

Edward J Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. 2022. Lora: Low-rank adaptation of large language models. *ICLR 2022*, 1(2):3.

Seonjeong Hwang, Yunsu Kim, and Gary Geunbae Lee. 2024. Explainable multi-hop question generation: An end-to-end approach without intermediate question labeling. In *LREC-COLING 2024*, pages 6855–6866.

Vishwajeet Kumar, Yuncheng Hua, Ganesh Ramakrishnan, Guilin Qi, Lianli Gao, and Yuan-Fang Li. 2019. Difficulty-controllable multi-hop question generation from knowledge graphs. In *ISWC*, pages 382–398.

Zefeng Lin, Weidong Chen, Yan Song, and Yongdong Zhang. 2024. Prompting few-shot multi-hop question generation via comprehending type-aware semantics. In *Findings of ACL: NAACL 2024*, pages 3730–3740.

Lidiya Murakhovs'ka, Chien-Sheng Wu, Philippe Laban, Tong Niu, Wenhao Liu, and Caiming Xiong. 2022. MixQG: Neural question generation with mixed answer types. In *Findings of ACL: NAACL 2022*, pages 1486–1497.

Shashi Narayan, Gonçalo Simões, Yao Zhao, Joshua Maynez, Dipanjan Das, Michael Collins, and Mirella Lapata. 2022. A well-composed text is half done! composition sampling for diverse conditional generation. In *ACL 2022*, pages 1319–1339.

Aaron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *arXiv preprint arXiv:1807.03748*.

[6] For example, to use a MoE LoRA architecture (Tang et al., 2025).

Liangming Pan, Yuxi Xie, Yansong Feng, Tat-Seng Chua, and Min-Yen Kan. 2020. Semantic graphs for generating deep questions. In *ACL 2020*, pages 1463–1475.

Pranoy Panda, Ankush Agarwal, Chaitanya Devaguptapu, Manohar Kaul, and Prathosh Ap. 2024. HOLMES: Hyper-relational knowledge graphs for multi-hop question answering using LLMs. In *ACL 2024*, pages 13263–13282.

Tianxiao Shen, Myle Ott, Michael Auli, and Marc'Aurelio Ranzato. 2019. Mixture models for diverse machine translation: Tricks of the trade. In *ICML 2019*, pages 5719–5728.

Chen Tang, Bo Lv, Zifan Zheng, Bohao Yang, Kun Zhao, Ning Liao, Xiaoxing Wang, Feiyu Xiong, Zhiyu Li, Nayu Liu, et al. 2025. Graphmoe: Amplifying cognitive depth of mixture-of-experts network via introducing self-rethinking mechanism. *arXiv preprint arXiv:2501.07890*.

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

Zehua Xia, Qi Gou, Bowen Yu, Haiyang Yu, Fei Huang, Yongbin Li, and Nguyen Cam-Tu. 2023. Improving question generation with multi-level content planning. In *Findings of ACL: EMNLP 2023*, pages 800–814.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *EMNLP 2018*, pages 2369–2380.

A Dataset statistics

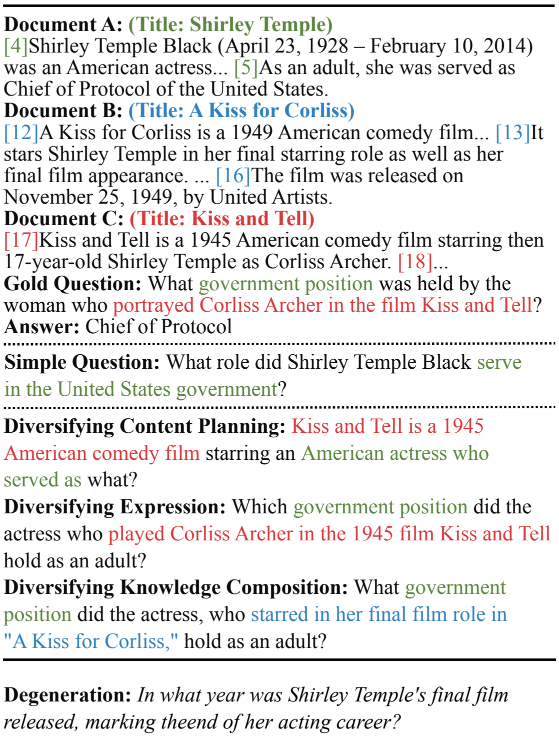
The detailed statistics of both HotpotQA and 2WikiMultihopQA datasets are provided in Table 5.

Dataset	HotpotQA	2WikiMultihopQA
	Min/Mean/Max	*Min/Mean/Max*
Answer Len.	3/5.2/171	3/5.1/28
Question Len.	6/24.7/143	8/17.6/58
Sentence Len.	2/30.8/590	3/25.0/375
Context Len.	50/1273.9/3764	125/825.6/5967
KC Len.	9/83.6/424	16/72.0/314
Context Num.	2/41.2/147	10/33.0/209
KC Num.	2/2.4/9	2/2.2/5
	Train/Dev/Test	*Train/Dev/Test*
Sample Num.	84487/479/6947	109589/306/11281

Table 5: Dataset statistics of HotpotQA and 2WikiMultihopQA, where Len. and Num. denote the number of tokens and the number of sentences, respectively. KC denotes the knowledge composition.

B Detailed Example

The detailed example is shown in Figure 5, which is algined with the example in Figure 1.



Document A: **(Title: Shirley Temple)**
[4]Shirley Temple Black (April 23, 1928 – February 10, 2014) was an American actress... [5]As an adult, she was served as Chief of Protocol of the United States.
Document B: **(Title: A Kiss for Corliss)**
[12]A Kiss for Corliss is a 1949 American comedy film... [13]It stars Shirley Temple in her final starring role as well as her final film appearance. ... [16]The film was released on November 25, 1949, by United Artists.
Document C: **(Title: Kiss and Tell)**
[17]Kiss and Tell is a 1945 American comedy film starring then 17-year-old Shirley Temple as Corliss Archer. [18]...
Gold Question: What government position was held by the woman who portrayed Corliss Archer in the film Kiss and Tell?
Answer: Chief of Protocol

Simple Question: What role did Shirley Temple Black serve in the United States government?

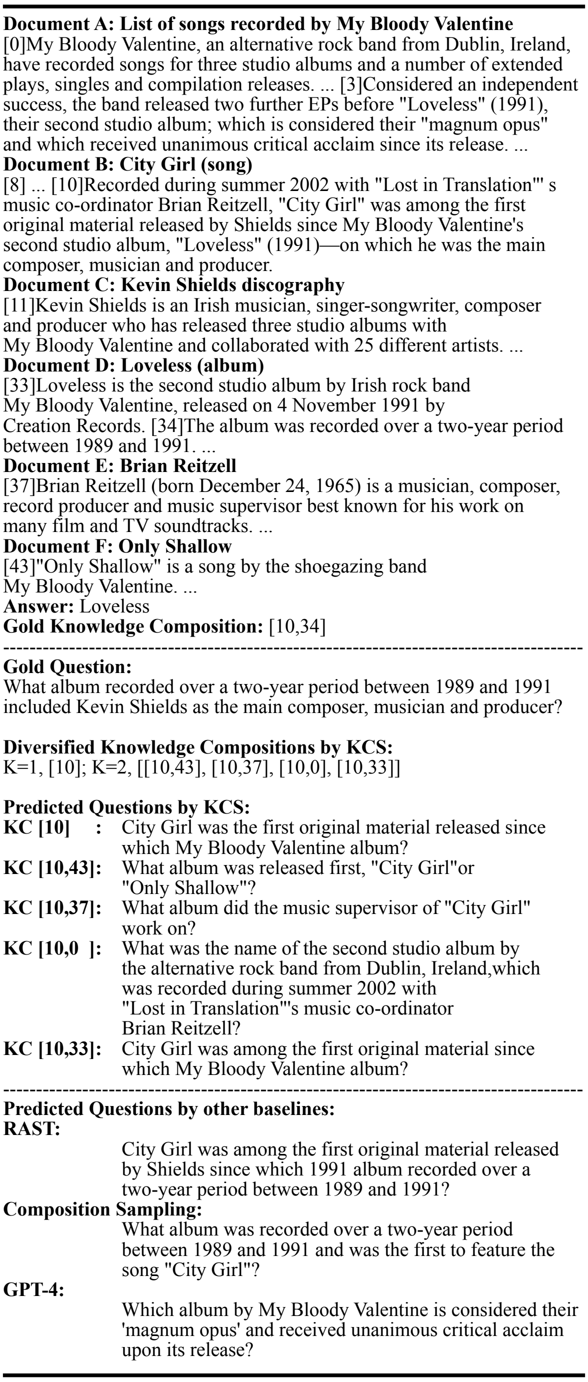
Diversifying Content Planning: Kiss and Tell is a 1945 American comedy film starring an American actress who served as what?
Diversifying Expression: Which government position did the actress who played Corliss Archer in the 1945 film Kiss and Tell hold as an adult?
Diversifying Knowledge Composition: What government position did the actress, who starred in her final film role in "A Kiss for Corliss," hold as an adult?

Degeneration: *In what year was Shirley Temple's final film released, marking theend of her acting career?*

Figure 5: A detailed example. Colored numbers ●●● indicating sentences from different documents.

C Case Study

To further analyze the impact of diverse knowledge compositions obtained through our Knowledge Composition Sampling (KCS) method on diversifying multi-hop question generation, we present a case study in Figure 6.



Document A: List of songs recorded by My Bloody Valentine
[0]My Bloody Valentine, an alternative rock band from Dublin, Ireland, have recorded songs for three studio albums and a number of extended plays, singles and compilation releases. ... [3]Considered an independent success, the band released two further EPs before "Loveless" (1991), their second studio album; which is considered their "magnum opus" and which received unanimous critical acclaim since its release. ...
Document B: City Girl (song)
[8] ... [10]Recorded during summer 2002 with "Lost in Translation"' s music co-ordinator Brian Reitzell, "City Girl" was among the first original material released by Shields since My Bloody Valentine's second studio album, "Loveless" (1991)—on which he was the main composer, musician and producer.
Document C: Kevin Shields discography
[11]Kevin Shields is an Irish musician, singer-songwriter, composer and producer who has released three studio albums with My Bloody Valentine and collaborated with 25 different artists. ...
Document D: Loveless (album)
[33]Loveless is the second studio album by Irish rock band My Bloody Valentine, released on 4 November 1991 by Creation Records. [34]The album was recorded over a two-year period between 1989 and 1991. ...
Document E: Brian Reitzell
[37]Brian Reitzell (born December 24, 1965) is a musician, composer, record producer and music supervisor best known for his work on many film and TV soundtracks. ...
Document F: Only Shallow
[43]"Only Shallow" is a song by the shoegazing band My Bloody Valentine. ...
Answer: Loveless
Gold Knowledge Composition: [10,34]

Gold Question:
What album recorded over a two-year period between 1989 and 1991 included Kevin Shields as the main composer, musician and producer?

Diversified Knowledge Compositions by KCS:
K=1, [10]; K=2, [[10,43], [10,37], [10,0], [10,33]]

Predicted Questions by KCS:
KC [10] : City Girl was the first original material released since which My Bloody Valentine album?
KC [10,43]: What album was released first, "City Girl"or "Only Shallow"?
KC [10,37]: What album did the music supervisor of "City Girl" work on?
KC [10,0]: What was the name of the second studio album by the alternative rock band from Dublin, Ireland,which was recorded during summer 2002 with "Lost in Translation"'s music co-ordinator Brian Reitzell?
KC [10,33]: City Girl was among the first original material since which My Bloody Valentine album?

Predicted Questions by other baselines:
RAST:
City Girl was among the first original material released by Shields since which 1991 album recorded over a two-year period between 1989 and 1991?
Composition Sampling:
What album was recorded over a two-year period between 1989 and 1991 and was the first to feature the song "City Girl"?
GPT-4:
Which album by My Bloody Valentine is considered their 'magnum opus' and received unanimous critical acclaim upon its release?

Figure 6: Case study.

D Logical Coherence Study

To demonstrate the logical coherence of the sampled knowledge compositions, we conducted evaluations using both a large language model (LLM) and human analysis.

Initially, we employ GPT-4 to assess whether the sampled compositions can be logically combined to form valid multi-hop questions. Specifically, we sample 200 examples from the original training data of the downstream MHQA task. We compare the LLM evaluation scores of KCS, which involves sampling five compositions per example, against the LLM evaluation scores of the ground truth, which includes one annotated meaningful knowledge composition per example. As illustrated in Table 6, the LLM evaluation scores for KCS closely approximate those of the ground truth, while significantly enhancing scale and diversity. Quantitatively, the LLM evaluation scores for KCS achieve 91.55% and 93.32% of the ground truth scores on HotpotQA and 2WikiMultihopQA, respectively. This finding confirms that the knowledge compositions selected by KCS can be logically combined to form valid multi-hop questions.

Data(Example num)	**HotpotQA**	**2WikiMultihopQA**
Ground Truth(200)	80.95	75.00
KCS(1000)	74.11	69.99

Table 6: LLM evaluation scores for original ground truth and KCS-diversified knowledge compositions on HotpotQA and 2WikiMultihopQA.

Since the LLM evaluation has a 20% to 25% deviation from the human-annotated ground truth, we conduct a manual review and in-depth analysis of these samples. Some cases are contentious, which does not invalidate the efficacy of LLM-Eval, as human evaluators may also produce divergent judgments. Other cases contain errors in human annotations, predominantly characterized by redundant knowledge components. Overall, these samples can be categorized into three distinct groups: (1) Illogical Knowledge Composition: Some combinations of knowledge components are redundant and lack coherent logical relationships; (2) Superficial Similarity with Underlying Discrepancies: Some knowledge components appear similar in form but represent fundamentally distinct facts, leading to ambiguity in their relationship (even for human evaluators); (3) Pronoun-Induced Ambiguity: Some knowledge components contain excessive pronominal references, resulting in semantic ambiguity that complicates relationship identification (even for human evaluators). Examples are provided in Figure 7.

The results of logical coherence study demonstrate that the selected knowledge components in KCS can be logically combined to generate valid multi-hop questions. The improved performance by KCS on downstream MHQA task also provides additional validation for the logical coherence in sampled knowledge compositions.

Example 1: Illogical Knowledge Composition
Answer: "Nausikaa Lake"
Knowledge1: "Nausikaa Lake is a lake in northeastern Ontario, Canada."
Knowledge2: "The Odyssey is one of two major ancient Greek epic poems attributed to Homer."
Question: "What lake in Canada was named after a character in a famous poem written by Homer?"

Example 2: Superficial Similarity with Underlying Discrepancies
Answer: "Ryan Rider"
Knowledge1: "Simon Sandberg (born 1994) is a Swedish footballer who plays as a defender for Hammarby IF."
Knowledge2: "Ryan Rider (born 1988) is a Canadian professional wrestling commentator and radio broadcaster."
Question: "Who was born first, Simon Sandberg or Ryan Rider?"

Example 3: Pronoun-Induced Ambiguity
Answer: "Maria of Cleves"
Knowledge1: "She was a daughter of the French king Louis XII of France and Anne of Brittany."
Knowledge2: "The son of Charles, Duke of Orléans, and Maria of Cleves, he succeeded his cousin Charles VIII, who died without a closer heir in 1498."
Question: "Who is the paternal grandmother of Claude Of France?"

Figure 7: Examples from the logical coherence study.

E Algorithm of Diversifying Multi-hop QG

The diversifying multi-hop QG algorithm is detailed in Algorithm 1, which is discussed in Section 3.1.

Algorithm 1: Diversify Multi-hop QG

```
Input: Context $D$, answer $a$, and $K, N_q, p$,
Output: Diversified multi-hop questions $Q$
$Q \leftarrow \emptyset$;
for $i \leftarrow 1$ to $N_q$ do
    $c_i \leftarrow \emptyset$;
    for $k \leftarrow 1$ to $K$ do
        $e_k \leftarrow$ generate latent prediction
          representation with input $(D, a, c_i)$;
        $p_{\text{new}} \leftarrow$ rescaled $p(s|e_k)$ to a new
          distribution with $p$;
        $s \leftarrow$ randomly sample a sentence from $D$
          based on $p_{\text{new}}$;
        $c_i \leftarrow c_i \cup \{s\}$;
    end
    $q \leftarrow$ generate question from $(c_i, a)$;
    $Q \leftarrow Q \cup \{q_i\}$;
end
return $Q$;
```