南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Generate-then-Ground in Retrieval-Augmented Generation for Multi-hop Question Answering

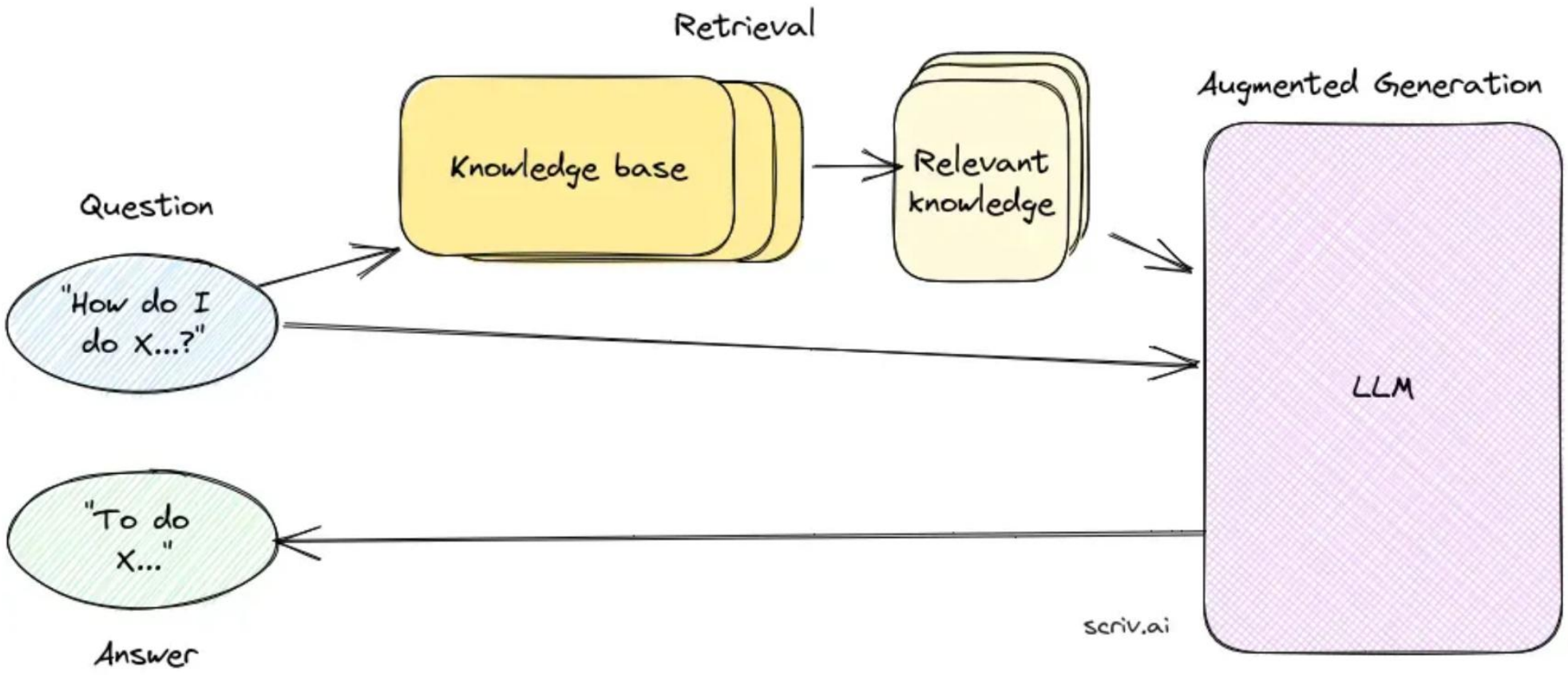
Zhengliang Shi[1] **Shuo Zhang**[2] **Weiwei Sun**[1] **Shen Gao**[3]
Pengjie Ren[1] **Zhumin Chen**[1] **Zhaochun Ren**[4*]
[1]Shandong University, Qingdao, China [2]Bloomberg, London, United Kingdom
[3]University of Electronic Science and Technology of China, Chengdu, China
[4]Leiden University, Leiden, The Netherlands
shizhl@mail.sdu.edu.cn szhang611@bloomberg.net
sunnweiwei@gmail.com z.ren@liacs.leidenuniv.nl

ACL24

Background

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

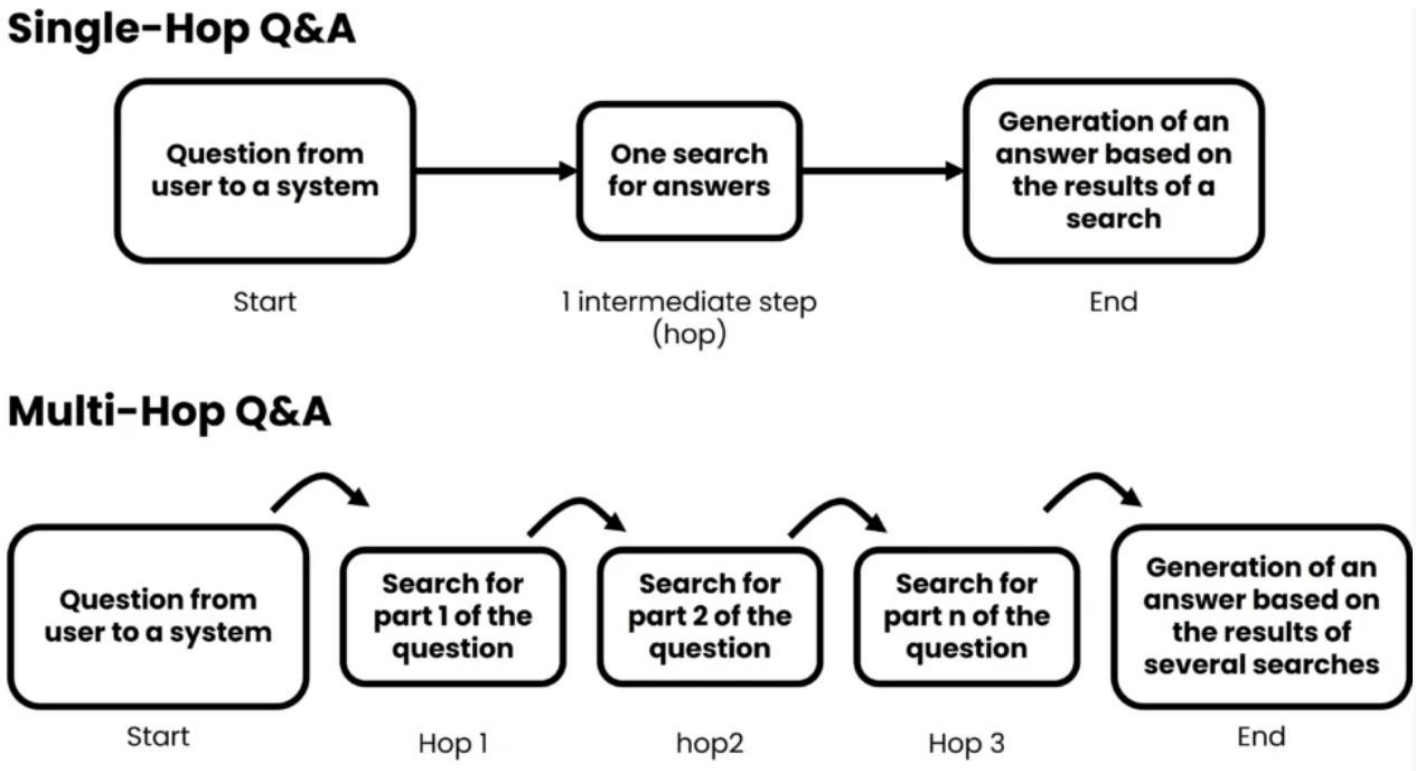
Retrieval-Augmented Generation



Background

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Multi-Hop Question



示例：

问题：“爱因斯坦的出生地所在国家的首都是哪里？”

解析：
1.先回答：“爱因斯坦出生在哪里？”

答案是 **德国。**

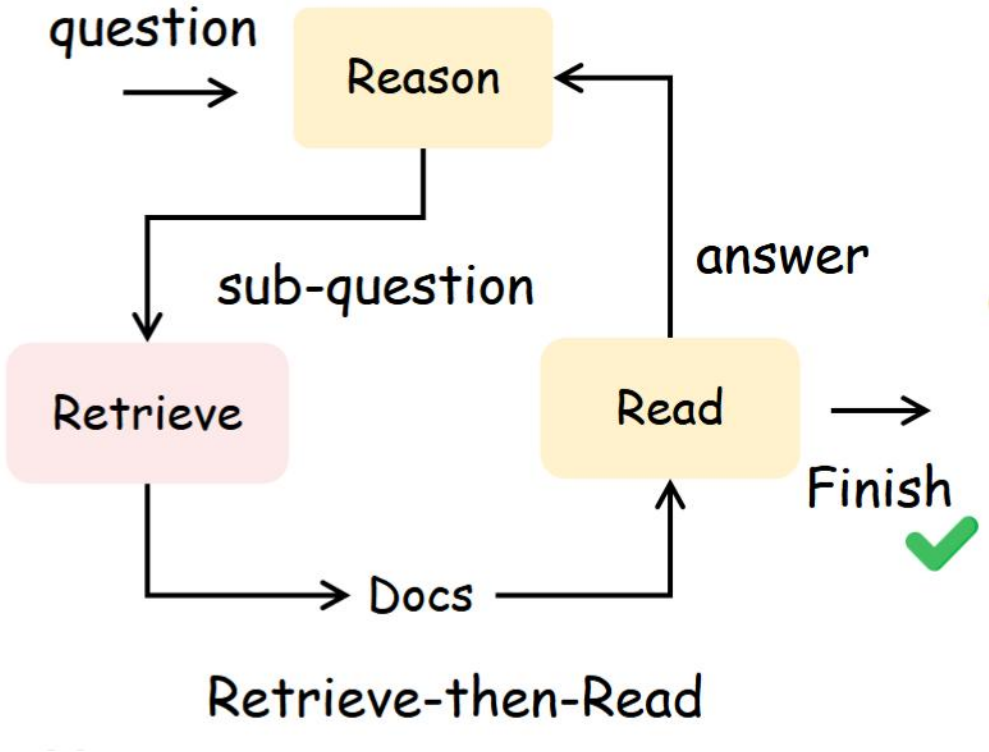
2.再回答：“德国的首都是哪里？”

答案是 **柏林。**

3.最终得出答案：**柏林。**

Background

Retrieve-then-Read



Retrieve-then-Read 的缺陷

检索器主导信息获取：
LLM 直接基于检索器返回的文档进行推理。
如果检索器返回**错误或不相关信息**，LLM 可能会受到误导。

LLM 直接面对大量无关信息：
文档可能**冗长**，包含大量噪声信息。
LLM 需要自己从中找到相关部分，容易受到错误信息干扰。

Generate-then-ground

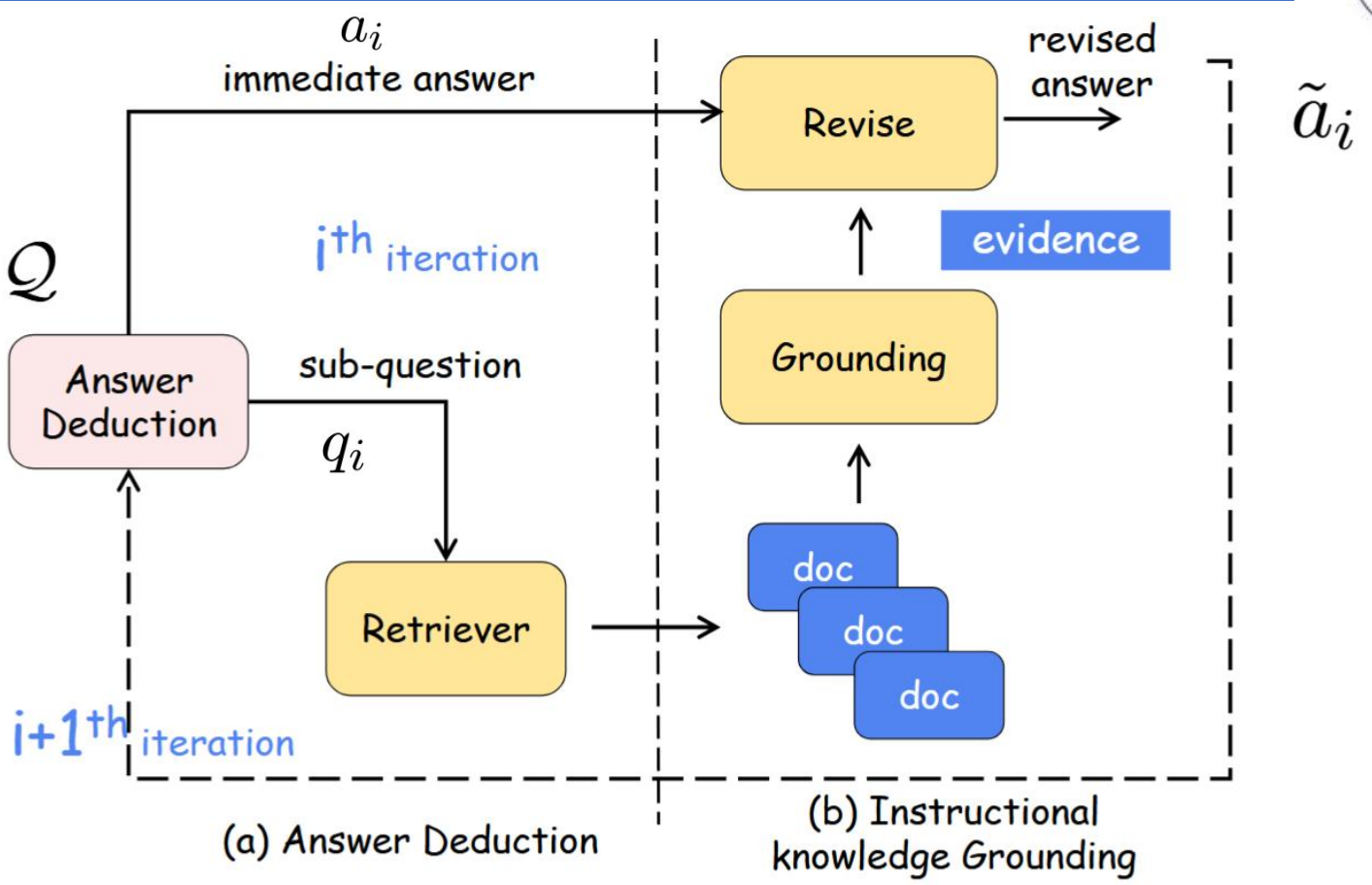


Figure 2: The architecture of the proposed generate-then-ground framework.

Answer Deduction

南京航空航天大學 Nanjing University of Aeronautics and Astronautics

IA

Please decompose a multi-hop question into sub-questions and answer the sub-questions step by step.

Starting below, you should interleave Deduce and Answer until deriving at the final answer.
- Deduce: deduce the current context and then formulate a sub-question
- Answer: answer the deduced question

Here are some examples: {examples}

{question} q_i

{answer} a_i

(a) Answer Generation

当前推理上下文 H，其中包含之前生成的子问题及其修正后的答案：

$$H = \{(q_j, \tilde{a}_j) | j < i\}$$

$$q_i, a_i = M_\theta(IA, Q, H_i)$$

Instructional Knowledge Grounding

南京航空航天大學 Nanjing University of Aeronautics and Astronautics

IG

Here is an answer to the question. Please cite evidence from the documents list to revise the answer. You should encapsulate the evidence using "<ref></ref>", and the revised answer using "<revise> </revise>".

If no evidence can be found, just give "<ref> Empty </ref>".

<ref> Evidence </ref>

<revise> {Revised answer} </revise>

(b) Instructional Knowledge Grounding

1. **检索外部文档** D_i:

$$D_i = \text{Retrieval}(q_i)$$

2. **基于证据修正答案**:

$$\tilde{a}_i = M_\theta(IG, Q, q_i, a_i)$$

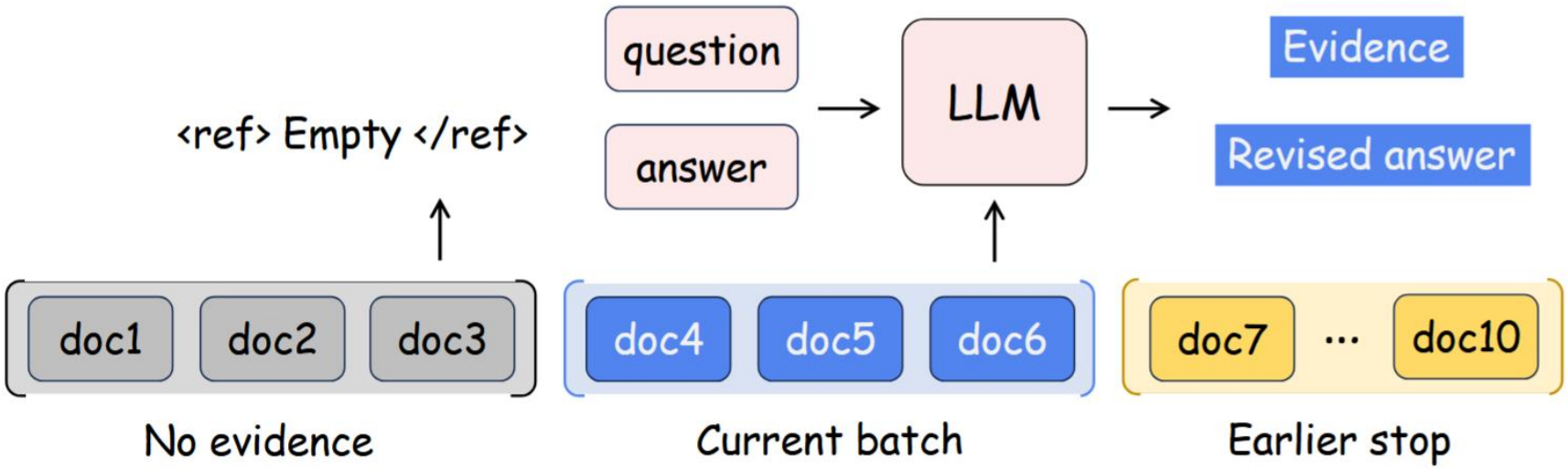
3. **更新上下文**:

$$H_{i+1} = H_i \cup \{(q_i, \tilde{a}_i)\}$$

4. **特殊情况**（如果没有找到相关证据）：

$$\tilde{a}_i = a_i$$

Batch Knowledge Ground



Experimental Setup

数据集（Datasets)
实验选用了四个多跳问答（MHQA）基准数据集：
1.HotpotQA (HQA)（Yang et al., 2018）
2.MuSiQue (MQA)（Trivedi et al., 2022）
3.2WikiMultiHopQA (WQA)（Ho et al., 2020）
4.StrategyQA (SQA)（Geva et al., 2021，来自 BIG-bench）

这些数据集涵盖了 不同类型的多跳推理问题，确保实验的多样性。

基线方法（Baselines)
对比方法分类
•不依赖外部检索（Generation w/o Retrieval)
•依赖外部检索（Generation w/ Retrieval)

Experimental Setup

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

评估指标（Evaluation Metrics)
评估方式采用：
1.Accuracy (Acc)：检查生成的答案是否包含真实答案。

2.F1 分数：衡量 **生成答案和真实答案之间的重叠度**

3.语义准确率 (Acc†)：
1. 使用 GPT-3.5-turbo-instruct2 进行语义匹配评估，确保答案语义上正确，即使表述不同。

4.人工评估：
1. 由 **三名受过教育的评审员** 评估 **随机抽样的 120 个案例**，采用 **三分制评分**（正确/部分正确/错误）。
2. 目的是降低自动评估的偏差（Shi et al., 2023）。

Experimental Results

Methods	HotpotQA			MuSiQue			2Wikimultihopqa			StrategyQA
	F1	Acc	Acc†	F1	Acc	Acc†	F1	Acc	Acc†	Acc
Generate w/o Retrieval										
CoT (Wei et al., 2022)	35.28	30.79	37.07	23.35	13.21	17.85	35.41	32.46	34.52	67.83
CoT-SC (Wang et al., 2022b)	42.25	38.68	39.07	15.61	10.02	12.42	40.37	36.57	38.59	70.84
GenRead (Yu et al., 2022)	35.21	36.81	37.54	9.77	9.29	10.32	23.13	20.62	28.31	67.13
GenRead *w/ decomposition*	42.28	43.32	45.31	20.13	17.58	20.62	41.19	41.63	43.24	68.13
Generate w/ Retrieval										
VE (Zhao et al., 2023)	29.64	22.64	24.64	6.5	11.14	15.57	13.76	31.57	32.64	63.07
ReAct (Yao et al., 2022)	40.70	33.10	37.12	15.34	17.32	19.32	35.50	30.10	33.41	68.37
GRG *w/ decomposition*	50.21	45.18	50.80	24.87	17.91	22.33	40.42	40.48	43.05	75.21
RetGen (Shao et al., 2023)	28.30	41.04	44.10	21.04	17.69	20.19	36.00	42.17	45.21	73.42
SearChain (Xu et al., 2024)	-	46.76	48.12	-	17.07	20.45	-	42.14	46.27	76.95
DSPy (Khattab et al., 2023)	47.80	42.43	50.07	20.11	13.40	17.40	44.77	43.43	45.43	71.78
GenGround (Ours)	**52.26**	**47.27**	**55.73**	**27.36**	**20.24**	**24.77**	**50.21**	**45.61**	**48.58**	**77.12**

Results with Different Retrievers

Method	HQA	MQA	WQA	Average Δ ↓
Retriever → BM25				
Ours	42.21	18.32	40.32	-
DSPy	40.86	15.32	30.85	5.27↓
GRG *w/ dq*	41.31	15.62	38.84	2.36↓
RetGen	39.12	8.41	35.83	6.50↓
SearChain	39.57	14.93	37.41	3.65↓
Retriever → Google Search				
Ours	48.95	21.54	46.87	-
DSPy	46.86	20.71	39.92	3.29↓
GRG *w/ dq*	42.57	18.41	43.21	4.39↓
RetGen	42.82	14.27	44.31	5.32↓
SearChain	44.35	19.76	44.39	2.95↓

Table 5: Accuracy (Acc) on three datasets using BM25

Ablation Study

南京航空航天大學
Nanjing University of Aeronautics and Astronautics

Ablation	HQA			SQA
	F1	Acc	Acc†	Acc
w/o deduction	42.65 $\downarrow_9$	41.08$\downarrow_6$	43.14 $\downarrow_{12}$	66.51 $\downarrow_{10}$
w/o grounding	45.14$\downarrow_7$	41.35$\downarrow_4$	43.23$\downarrow_5$	72.34$\downarrow_5$
w/o batch	47.27$\downarrow_5$	45.03$\downarrow_2$	51.19$\downarrow_4$	71.72$\downarrow_5$

Table 4: Evaluation results of our ablation study on two MHQA benchmarks.

Thanks