

# SUPPLEMENTARY MATERIALS FOR KEQING

**Anonymous authors**

Paper under double-blind review

## A DATASETS & PREPROCESS

**MetaQA** (Zhang et al., 2018) consists of a movie ontology derived from the WikiMovies Dataset and three sets of question-answer pairs written in different levels of difficulty. It evaluates the effectiveness in a specific domain.

**WebQSP** (Yih et al., 2016) contains questions from WebQuestions that are answerable by Freebase. It tests i.i.d. generalization on simple questions.

**GrailQA** (Gu et al., 2021) is a diverse KBQA dataset built on Freebase, covering 32,585 entities and 3,720 relations across 86 domains. It is designed to test three levels of generalization of KBQA models: I.I.D., compositional, and zero-shot.

## B BASELINES

For the baselines in comparison, we have included the competitive methods that have a publication on the official leaderboard of each dataset and record their results from the paper directly with the same evaluation matrix. For ease of comparison, we have summarized the main thoughts of competitive baselines in the following:

**KB-BINDER** (Li et al., 2023a) is a training-free system, which for the first time, proposes to utilize the in-context learning capability of large language models (LLMs) to solve KBQA tasks. Particularly, it leverages the Codex (Chen et al., 2021) to generate logical forms as the draft for a specific question by imitating a few demonstrations, and then grounds on the knowledge base to bind the generated draft to an executable one with BM25 score matching.

**Pangu** (Gu et al., 2022) is developed as a generic framework for grounded language understanding that capitalizes on the discriminative ability instead of the generative ability of LLMs. Specifically, Pangu consists of a symbolic agent and a neural LLM working in a concerted fashion, where the agent explores the environment to incrementally construct valid plans, and the LLM evaluates the plausibility of the candidate plans to guide the search process.

**FlexKBQA** (Li et al., 2023b) targets at leveraging automated algorithms to sample diverse programs, such as SPARQL queries, from the knowledge base, which are subsequently converted into natural language questions via LLMs. Moreover, FlexKBQA introduces an additional execution guided self-training method to iterative leverage unlabeled user questions, which can reduce the barriers of distribution shift between synthetic data and real user questions.

## C MORE EXPERIMENTAL RESULTS

The experimental results on GrailQA dataset have been exhibited on Table. 1 and Table. 2.

Table 1: Performance comparison of different methods on the GrailQA dev set.

Method	Overall	
	EM	F1
QGG (Lan & Jiang, 2020)	-	36.7
GloVE+Transduction (Gu et al., 2021)	17.6	18.4
GloVE+Ranking (Gu et al., 2021)	39.5	45.1
BERT+Transduction (Gu et al., 2021)	33.3	36.8
BERT+Ranking (Gu et al., 2021)	50.6	58.0
RnG-KBQA (Ye et al., 2022)	68.8	74.4
DecAF (Yu et al., 2022)	68.4	78.8
TIARA (Shu et al., 2022)	73.0	78.5
Pangu (Gu et al., 2022)	<b>73.7</b>	<b>79.9</b>
KB-BINDER (Li et al., 2023a)	50.6	56.0
FlexKBQA (Li et al., 2023b)	62.8	69.4
Keqing-LLaMA (Ours)	72.5	77.8

Table 2: Performance comparison of different methods on the GrailQA dev set.

Method	IID		Compositional		Zero-shot	
	EM	F1	EM	F1	EM	F1
QGG (Lan & Jiang, 2020)	-	40.5	-	33.0	-	36.6
GloVE+Transduction (Gu et al., 2021)	50.5	51.6	16.4	18.5	3.0	3.1
GloVE+Ranking (Gu et al., 2021)	62.2	67.3	40.0	47.8	28.9	33.8
BERT+Transduction (Gu et al., 2021)	51.8	53.9	31.0	36.0	25.7	29.3
BERT+Ranking (Gu et al., 2021)	59.9	67.0	45.5	53.9	48.6	55.7
RnG-KBQA (Ye et al., 2022)	86.2	89.0	63.8	71.2	63.0	69.2
DecAF (Yu et al., 2022)	84.8	89.9	73.4	81.8	58.6	72.3
TIARA (Shu et al., 2022)	<b>88.4</b>	<b>91.2</b>	<b>66.4</b>	<b>74.8</b>	<b>73.3</b>	<b>80.7</b>
Pangu (Gu et al., 2022)	<b>82.6</b>	<b>87.1</b>	<b>74.9</b>	<b>81.2</b>	<b>69.1</b>	<b>76.1</b>
KB-BINDER (Li et al., 2023a)	51.9	57.4	50.6	56.6	49.9	55.1
FlexKBQA (Li et al., 2023b)	71.3	75.8	59.1	65.4	60.6	68.3
Keqing-LLaMA (Ours)	80.5	85.6	73.3	80.1	67.5	74.7

## REFERENCES

- Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*, 2021.
- Yu Gu, Sue Kase, Michelle Vanni, Brian Sadler, Percy Liang, Xifeng Yan, and Yu Su. Beyond iid: three levels of generalization for question answering on knowledge bases. In *Proceedings of the Web Conference 2021*, pp. 3477–3488, 2021.
- Yu Gu, Xiang Deng, and Yu Su. Don’t generate, discriminate: A proposal for grounding language models to real-world environments. *arXiv preprint arXiv:2212.09736*, 2022.
- Yunshi Lan and Jing Jiang. Query graph generation for answering multi-hop complex questions from knowledge bases. Association for Computational Linguistics, 2020.
- Tianle Li, Xueguang Ma, Alex Zhuang, Yu Gu, Yu Su, and Wenhui Chen. Few-shot in-context learning for knowledge base question answering. *arXiv preprint arXiv:2305.01750*, 2023a.
- Zhenyu Li, Sunqi Fan, Yu Gu, Xiuxing Li, Zhichao Duan, Bowen Dong, Ning Liu, and Jianyong Wang. Flexkbqa: A flexible llm-powered framework for few-shot knowledge base question answering. *arXiv preprint arXiv:2308.12060*, 2023b.
- Yiheng Shu, Zhiwei Yu, Yuhua Li, Börje F Karlsson, Tingting Ma, Yuzhong Qu, and Chin-Yew Lin. Tiara: Multi-grained retrieval for robust question answering over large knowledge bases. *arXiv preprint arXiv:2210.12925*, 2022.

- Xi Ye, Semih Yavuz, Kazuma Hashimoto, Yingbo Zhou, and Caiming Xiong. Rng-kbqa: Generation augmented iterative ranking for knowledge base question answering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022.
- Wen-tau Yih, Matthew Richardson, Christopher Meek, Ming-Wei Chang, and Jina Suh. The value of semantic parse labeling for knowledge base question answering. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pp. 201–206, 2016.
- Donghan Yu, Sheng Zhang, Patrick Ng, Henghui Zhu, Alexander Hanbo Li, Jun Wang, Yiqun Hu, William Wang, Zhiguo Wang, and Bing Xiang. Decaf: Joint decoding of answers and logical forms for question answering over knowledge bases. *arXiv preprint arXiv:2210.00063*, 2022.
- Yuyu Zhang, Hanjun Dai, Zornitsa Kozareva, Alexander Smola, and Le Song. Variational reasoning for question answering with knowledge graph. In *Proceedings of the AAAI conference on artificial intelligence*, volume 32, 2018.