On the Role of Long-tail Knowledge in Retrieval Augmented Large Language Models

Dongyang Li[1,2] [*], **Junbing Yan**[1,2][*], **Taolin Zhang**[2][*], **Chengyu Wang**[2][†], **Xiaofeng He**[1,3][†], **Longtao Huang**[2], **Hui Xue**[2], **Jun Huang**[2]
[1] School of Computer Science and Technology, East China Normal University
[2] Alibaba Group, [3] NPPA Key Laboratory of Publishing Integration Development, ECNUP
dongyangli0612@gmail.com, {yanjunbing.yjb, zhangtaolin.ztl, chengyu.wcy, kaiyang.hlt, hui.xueh, huangjun.hj}@alibaba-inc.com, hexf@cs.ecnu.edu.cn

Abstract

Retrieval augmented generation (RAG) exhibits outstanding performance in promoting the knowledge capabilities of large language models (LLMs) with retrieved documents related to user queries. However, RAG only focuses on improving the response quality of LLMs via enhancing queries indiscriminately with retrieved information, paying little attention to what type of knowledge LLMs really need to answer original queries more accurately. In this paper, we suggest that long-tail knowledge is crucial for RAG as LLMs have already remembered common world knowledge during large-scale pre-training. Based on our observation, we propose a simple but effective long-tail knowledge detection method for LLMs. Specifically, the novel Generative Expected Calibration Error (GECE) metric is derived to measure the "long-tailness" of knowledge based on both statistics and semantics. Hence, we retrieve relevant documents and infuse them into the model for patching knowledge loopholes only when the input query relates to long-tail knowledge. Experiments show that, compared to existing RAG pipelines, our method achieves over 4x speedup in average inference time and consistent performance improvement in downstream tasks.

1 Introduction

Large language models (LLMs), equipped with retrieval augmented generation (RAG), perform well in various tasks (Izacard et al., 2023; Cheng et al., 2023; Shao et al., 2023). RAG retrieves supplement knowledge by retrievers and enhances prompts for LLMs by retrieved documents, in order to generate more accurate contents (Borgeaud et al., 2022; Cheng et al., 2023; Shao et al., 2023).

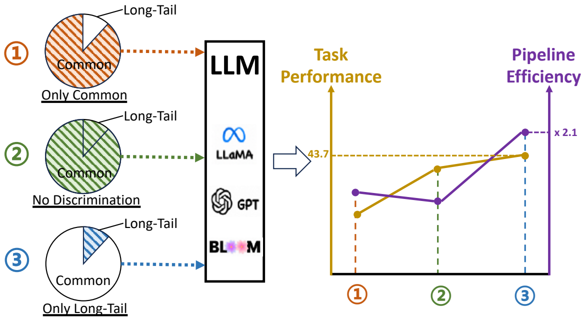

Figure 1: Comparison between different RAG strategies over the NQ dataset (Kwiatkowski et al., 2019).

However, previous RAG works concentrate on improving the task performance, without fine-grained process of knowledge (Wang et al., 2023a; Trivedi et al., 2023). In this case, redundant computation is performed on well-learned common knowledge, which does not require further enhancement. Therefore, more consideration should be given to long-tail knowledge that LLMs really need, which rarely occurs during pre-training (Kandpal et al., 2023). [1]

In the literature, RAG can be roughly divided into two categories: (1) *Once Retrieval*. Wang et al. (2023a); Cheng et al. (2023); Shi et al. (2023) retrieve external knowledge just once by different retrievers and enhance the model with recalled related content for more effective generation. They treat all queries equally and do not model the familiarity of different queries to LLMs. (2) *Iterative Retrieval*. Shao et al. (2023); Feng et al. (2023); Asai et al. (2023) construct multi-step retrieval-then-augmentation process to generate accurate results by synergistic feedback of LLMs. Yet, as shown in Figure 1, augmenting LLMs with common knowledge that the models do not need results

[*] D. Li, J. Yan and T. Zhang contributed equally to this work.

[†] Co-corresponding authors.

[1] Note that Long-tail knowledge is in low individual sample frequencies but high aggregated quantities, which implies a certain amount of significance (Jansen, 2007).

Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), pages 120–126
August 11-16, 2024 ©2024 Association for Computational Linguistics

in low efficiency and redundant computation. To our knowledge, there is a lack of research on the use of long-tail knowledge for RAG.

Building upon our observation, we explore the role of long-tail knowledge in RAG. We suggest that long-tail knowledge is crucial for RAG and propose an improved RAG pipeline. Specifically, to measure the "long-tailness" of knowledge in terms of LLMs, we largely extend Expected Calibration Error (ECE) for classification tasks (Aimar et al., 2023; Zhong et al., 2021; Xu et al., 2021), and propose Generative Expected Calibration Error (GECE). It leverages METEOR (Banerjee and Lavie, 2005) and the output probability of LLMs to characterize "long-tailness", which considers both continuous gradient-based semantics and discrete frequency-based statistics. Based on GECE, our pipeline retrieves relevant documents and performs RAG only when user queries relate to long-tail knowledge. Our approach outperforms current RAG pipelines, providing a 4x speedup in inference and improved performance in retrieval tasks.

2 Related Work

2.1 Retrieval Augmentation

The augmentation stage of RAG can be divided into three stages: pre-training, fine-tuning, and inference. Atlas (Izacard et al., 2023) is a retrieval-augmented pre-trained LLM and works well in few-shot settings. Borgeaud et al. (2022); Wang et al. (2023a) retrieve neighbor-related, chunk-grained knowledge from memory and inject the knowledge during the pre-training stage. Cheng et al. (2023); Lin et al. (2023); Shi et al. (2023) fine-tune both the retriever and the generator synergistically and boost each other mutually. Shao et al. (2023); Feng et al. (2023); Trivedi et al. (2023) insert knowledge at the inference stage by iterative guiding with frozen retrievers and LLMs. These methods introduce knowledge without detecting knowledge "long-tailness" and redundancy.

2.2 Long-Tail Processing

Zhao et al. (2023); Yao et al. (2024); Zheng et al. (2023) design repeat-sampling, under-sampling, and other strategies to access the unbalanced problem. They concentrate on classification tasks and consider less about the recent popular tendency of text generation tasks. Liang et al. (2023); Zhou et al. (2023); Wang et al. (2024) leverage compositional operation to synthesize head and tail instances together by attention, graph-connection, and other fusion mechanisms. Wang et al. (2023c); Li et al. (2023); Xu et al. (2023) import extra features to tail classes for patching the demand of more information. To our knowledge, existing works touch less on distinguishing whether the instance is long-tail or not because of the existence of labeled training datasets.

3 Preliminaries

Traditional works rely on text frequencies to define whether the instance is long-tail or not; thus, low-frequency texts tend to be classified into long-tail classes. For LLMs, computing text frequencies of previously unknown user queries is by no means an easy task. As in (Aimar et al., 2023; Zhong et al., 2021; Xu et al., 2021), *Expected Calibration Error* (ECE) provides a new perspective to measure "long-tailness". ECE measures how well a model's estimated probabilities match true (observed) probabilities (Guo et al., 2017). In the calculation of ECE, the confidence of each instance is allocated to a specific interval and obtained by the model predicted probability. The accuracy is determined by the comparison of the predicted label and the ground truth. The absolute margin between confidence and accuracy of each instance represents the calibration degree. The expected calibration degree of the whole dataset indicates the reliance of the model. Formally, ECE can be formulated as:

$$ECE = \sum_{i=1}^{B} \frac{n_{b_i}}{N} |acc(b_i) - conf(b_i)| \quad (1)$$

where i denotes i-th bin, N is the total instance count of the dataset, $acc(b_i)$ and $conf(b_i)$ represent the accuracy and confidence of the bin b_i, and n_{b_i} is the instance number of the bin b_i. B is the count of bins in the interval of $[0, 1]$. In our work, we extend ECE for NLP, particularly for the LLM text generation scenario.

4 Methodology

4.1 Metric-based Long-tailness Detection

As long-tail knowledge is crucial for RAG, we propose the GECE metric to detect the instance "long-tailness". Here, we transform the traditional ECE formula with METEOR (Banerjee and Lavie, 2005) and average prediction probability:

- Accuracy in ECE is to measure the agreement between prediction and ground truth. In the generative scenario, we utilize METEOR (Banerjee and Lavie, 2005) to measure coherence and relevance between predicted candidates and ground truth.

- Confidence in ECE is the predicted probability produced by the model itself. Similarly, we employ the average token probability output by LLMs.

Moreover, to enhance our metric with long-tail detection abilities, we further integrate the following two factors, which assist us to further separate common and long-tail instances apart:

- Average word frequency, as word frequency is a basic indication of long-tail texts.
- Dot product between the mean gradient of the total dataset and the gradient of a specific instance is leveraged to evaluate the discrepancy (Chen et al., 2022). This is because the gradient of a long-tail instance has a large disparity with the mean gradient of the total dataset, and vice versa.

From the above analysis, we construct GECE as:

$$GECE = \frac{|M(pred, ref) - \frac{1}{n}\sum_{i=1}^{n} p(t_i)|}{\alpha \cdot [E(\bigtriangledown_{ins}) \cdot \bigtriangledown_{ins}]} \quad (2)$$

where $pred$ and ref represent the generated text and the referenced ground truth, respectively. $M(pred, ref)$ is the METEOR score (Banerjee and Lavie, 2005). The average token probability is formulated as $\frac{1}{n}\sum_{i=1}^{n} p(t_i)$ where $p(t_i)$ denotes the i-th token's probability produced by LLM, and n is the token sequence length. For the denominator part, α is the average word frequency. We can see that a long-tail instance has a smaller α value and hence its reciprocal will be larger. In addition, $\bigtriangledown_{ins}$ is the gradient w.r.t. the current instance, and $E(\bigtriangledown_{ins})$ is the mean gradient of the total dataset. To obtain the gradient, we run a forward and a backward pass only through fine-tuning the LLM using the dataset. We can see that a long-tail instance has a smaller gradient $\bigtriangledown_{ins}$, compared to the mean score of the dataset, and thus obtains a smaller dot product $E(\bigtriangledown_{ins}) \cdot \bigtriangledown_{ins}$.

Larger GECE value implies larger degree of long-tailness. For example, if we apply GECE to the query of NQ “Who was named African footballer of the year 2014”, the value is 34.6. In contrast, for a long-tail, more professional NQ query “Who has played Raoul in The Phantom of the Opera”, the GECE value is 112.7.

4.2 Improved RAG Pipeline

As an extension to vanilla RAG pipelines, we only retrieve documents related to long-tail queries from the data source, disregarding common instances. The retrieval process is implemented by a dense passage retriever to retrieve related WikiPedia[2] documents. For long-tail instances, we input the query concatenated with the recalled related documents to LLMs for answer attainment. For common instances, we only input the query itself to LLMs.

5 Experiments

In this section, we briefly describe the experimental results and leave detailed experimental settings in Appendix A, and supplementary experimental results in Appendix B.

5.1 Datasets

NQ (Kwiatkowski et al., 2019) is a large-scale question answering dataset and constructed by human-labeled answers from Wikipedia web pages. We utilize the short answer type of NQ in this paper. **TriviaQA** (Joshi et al., 2017) is a relatively complex dataset containing syntactic and lexical differences between questions and answers. **MMLU** (Hendrycks et al., 2021) is a typical model evaluation benchmark that includes various-domain samples and it ranges in multiple degrees of difficulty from primary to advanced professional level.

5.2 Baselines

Llama2-7B (Wang et al., 2023d) is a pre-trained LLM with large-scale parameters and performs well on most benchmarks. **IRCoT** (Trivedi et al., 2023) introduces an interleaves retrieval approach, exploiting Chain-of-Thought (CoT) to assist the retrieval and leveraging the retrieval results to support CoT. **SKR** (Wang et al., 2023b) utilizes LLMs to distinguish whether the query can be resolved or not, and only retrieve the knowledge out of the model's self-knowledge. **SELF-RAG** (Asai et al., 2023) introduces special reflection tokens to help the model to determine the retrieval requirement and retrieved content quality. **FILCO** (Wang et al., 2023d) refines the retrieved context by a filter that is trained by string inclusion, lexical overlap relationship and conditional cross-mutual information. **ITER-RETGEN** (Shao et al., 2023) proposes a mutual promotion manner via the retrieval-augmented generation and generation-augmented retrieval.

[2] https://www.wikipedia.org/

Model	Type	Rouge-1				Bleu-4				Speed-up
		10	15	20	Avg.	10	15	20	Avg.	
Llama2-7B	w/o GECE	41.2	42.2	42.9	42.1$_{(\pm 0.2)}$	7.19	7.31	7.40	7.30$_{(\pm 0.22)}$	1.0 ×
	w GECE	41.9	43.1	43.7	42.9$_{(\pm 0.2)}$	7.27	7.40	7.48	7.38$_{(\pm 0.15)}$	2.1 ×
IRCoT	w/o GECE	45.5	45.8	46.3	45.9$_{(\pm 0.3)}$	7.52	7.73	7.70	7.65$_{(\pm 0.31)}$	1.0 ×
	w GECE	45.7	46.4	46.5	46.2$_{(\pm 0.3)}$	7.56	7.75	7.74	7.68$_{(\pm 0.26)}$	6.7 ×
SKR	w/o GECE	46.3	47.0	47.2	46.8$_{(\pm 0.2)}$	7.57	7.65	7.79	7.67$_{(\pm 0.11)}$	1.0 ×
	w GECE	46.9	47.1	47.6	47.2$_{(\pm 0.1)}$	7.66	7.78	7.85	7.76$_{(\pm 0.09)}$	5.5 ×
SELF-RAG	w/o GECE	42.1	43.3	43.7	43.0$_{(\pm 0.3)}$	7.12	7.35	7.44	7.30$_{(\pm 0.28)}$	1.0 ×
	w GECE	44.8	45.0	45.3	45.0$_{(\pm 0.2)}$	7.48	7.63	7.62	7.58$_{(\pm 0.22)}$	3.3 ×
FILCO	w/o GECE	43.6	44.2	44.7	44.2$_{(\pm 0.3)}$	7.46	7.48	7.52	7.49$_{(\pm 0.17)}$	1.0 ×
	w GECE	43.7	44.5	44.8	44.3$_{(\pm 0.2)}$	7.49	7.51	7.53	7.51$_{(\pm 0.15)}$	2.4 ×
ITER-RETGEN	w/o GECE	45.5	46.4	47.1	46.3$_{(\pm 0.2)}$	7.63	7.75	7.78	7.72$_{(\pm 0.31)}$	1.0 ×
	w GECE	46.5	47.0	47.3	46.9 $_{(\pm 0.1)}$	7.76	7.81	7.82	7.80$_{(\pm 0.26)}$	7.0 ×

Table 1: Experimental results on NQ. T-tests show the improvements are statistically significant with $p < 0.05$.

Model	Type	Rouge-1				Bleu-4				Speed-up
		10	15	20	Avg.	10	15	20	Avg.	
Llama2-7B	w/o GECE	22.5	24.6	24.9	24.0$_{(\pm 0.3)}$	6.68	6.92	7.17	6.92$_{(\pm 0.18)}$	1.0 ×
	w GECE	23.3	25.2	25.8	24.8$_{(\pm 0.3)}$	6.74	6.99	7.25	6.99$_{(\pm 0.32)}$	2.2 ×
IRCoT	w/o GECE	25.4	26.0	26.5	26.0$_{(\pm 0.2)}$	7.11	7.24	7.28	7.21$_{(\pm 0.24)}$	1.0 ×
	w GECE	25.9	26.7	26.7	26.4$_{(\pm 0.1)}$	7.18	7.26	7.31	7.25$_{(\pm 0.17)}$	6.2 ×
SKR	w/o GECE	26.6	27.2	27.5	27.1$_{(\pm 0.2)}$	7.51	7.57	7.62	7.57$_{(\pm 0.09)}$	1.0 ×
	w GECE	27.1	27.3	27.6	27.3$_{(\pm 0.2)}$	7.54	7.60	7.63	7.59$_{(\pm 0.15)}$	6.0 ×
SELF-RAG	w/o GECE	26.3	26.2	26.7	26.4$_{(\pm 0.2)}$	7.46	7.47	7.51	7.48$_{(\pm 0.19)}$	1.0 ×
	w GECE	26.4	26.5	27.0	26.6$_{(\pm 0.1)}$	7.55	7.55	7.56	7.55$_{(\pm 0.26)}$	3.5 ×
FILCO	w/o GECE	25.8	25.9	26.5	26.1$_{(\pm 0.3)}$	7.43	7.49	7.50	7.47$_{(\pm 0.16)}$	1.0 ×
	w GECE	26.3	26.6	26.8	26.6$_{(\pm 0.1)}$	7.48	7.52	7.54	7.51$_{(\pm 0.23)}$	2.3 ×
ITER-RETGEN	w/o GECE	26.8	26.7	27.2	26.9$_{(\pm 0.1)}$	7.36	7.41	7.57	7.45$_{(\pm 0.12)}$	1.0 ×
	w GECE	27.1	27.3	27.4	27.3$_{(\pm 0.2)}$	7.49	7.55	7.59	7.54$_{(\pm 0.13)}$	7.3 ×

Table 2: Experimental results on TriviaQA. T-tests show the improvements are statistically significant with $p < 0.05$.

5.3 General Results

We validate our method on the three datasets and the performance is listed in Table 1, Table 2, and Table 4. Due to space limitation, we move the result of MMLU to Appendix B.1. From the results, we can observe that: (1) All baseline models have better process speed when the data is filtered with GECE. Especially, the iterative methods are accelerated significantly (i.e., ITER-RETGEN and IRCoT). This improvement owes to the filter operation of GECE and the fine discrimination of the need or not for extra augmentation. (2) With GECE, the task performance is also promoted by introducing less noise of the common instances. (3) As the number of augmentation documents increases, i.e., from 10 to 20, the performance is boosted because of the substantial knowledge supplementation.

	NQ Rouge-1	TriviaQA Rouge-1	MMLU Accuracy
Ours	43.7	25.8	86.4
Item Replacement	42.3	24.2	84.8
w/o Statistics only	43.5	25.7	86.0
w/o Semantics only	41.6	24.9	85.5

Table 3: Results of ablation study.

5.4 Ablation Study

In Table 3, (1) Item Replacement means that we utilize chrF (Popovic, 2015) and TER (Snover et al., 2006) to replace METEOR, two other metrics for text generation with the same value scale as METEOR. The replaced mean results of these two alternative metrics decline, indicating that METEOR is more accurate. (2) For removing Statistics and Semantics, we delete the two items outside the absolute margin of GECE. The dropped scores demonstrate the importance of the two indicators.

6 Conclusion

In summary, our research highlights the significance of long-tail knowledge to enhance the efficacy of RAG for LLMs. We introduced the Generative Expected Calibration Error (GECE) to identify long-tail knowledge, which accelerates the inference process by more than fourfold in average and improves performance on downstream tasks without compromising the quality of responses. This demonstrates the benefits of selectively augmenting LLMs with targeted information, paving the way for more efficient and accurate RAG systems.

Acknowledgements

We would like to thank anonymous reviewers for their valuable comments. This work was supported in part by National Key R&D Program of China (No. 2022ZD0120302) and Alibaba Group through Alibaba Research Intern Program.

Limitations

While our method shows considerable promise for improving the efficiency and accuracy of RAG-augmented language models, it is important to acknowledge several limitations. The long-tail knowledge detection method we propose is based on the GECE metric, which may not capture all dimensions of “long-tailness”. Given that long-tail knowledge can be multi-faceted and context-specific, there may be instances where our method fails to detect, leading to suboptimal retrieval results. In addition, the applicability of GECE to more models and settings has not been thoroughly investigated. Further research is required to validate its effectiveness and adaptability across diverse LLMs and knowledge retrieval scenarios.

Ethical Considerations

Our research on RAG for LLMs aims to enhance the precision and efficiency of knowledge retrieval, hence we believe that there are no direct negative social impacts associated with our contributions. Yet, it is important to acknowledge that any generative AI technology, including our application based on LLMs, must be deployed with careful consideration of its broader implications.

References

Emanuel Sanchez Aimar, Arvi Jonnarth, Michael Felsberg, and Marco Kuhlmann. 2023. Balanced product of calibrated experts for long-tailed recognition. In *CVPR*, pages 19967–19977.

Akari Asai, Zeqiu Wu, Yizhong Wang, Avirup Sil, and Hannaneh Hajishirzi. 2023. Self-rag: Learning to retrieve, generate, and critique through self-reflection. *CoRR*, abs/2310.11511.

Satanjeev Banerjee and Alon Lavie. 2005. METEOR: an automatic metric for MT evaluation with improved correlation with human judgments. In *ACL*, pages 65–72.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George van den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, Diego de Las Casas, Aurelia Guy, Jacob Menick, Roman Ring, Tom Hennigan, Saffron Huang, Loren Maggiore, Chris Jones, Albin Cassirer, Andy Brock, Michela Paganini, Geoffrey Irving, Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. 2022. Improving language models by retrieving from trillions of tokens. In *ICML*, pages 2206–2240.

Zhao Chen, Vincent Casser, Henrik Kretzschmar, and Dragomir Anguelov. 2022. Gradtail: Learning long-tailed data using gradient-based sample weighting. *CoRR*, abs/2201.05938.

Daixuan Cheng, Shaohan Huang, Junyu Bi, Yuefeng Zhan, Jianfeng Liu, Yujing Wang, Hao Sun, Furu Wei, Weiwei Deng, and Qi Zhang. 2023. UPRISE: universal prompt retrieval for improving zero-shot evaluation. In *EMNLP*, pages 12318–12337.

Zhangyin Feng, Xiaocheng Feng, Dezhi Zhao, Maojin Yang, and Bing Qin. 2023. Retrieval-generation synergy augmented large language models. *CoRR*, abs/2310.05149.

Chuan Guo, Geoff Pleiss, Yu Sun, and Kilian Q. Weinberger. 2017. On calibration of modern neural networks. In *ICML*, pages 1321–1330.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. 2021. Measuring massive multitask language understanding. In *ICLR*.

Gautier Izacard, Patrick S. H. Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. 2023. Atlas: Few-shot learning with retrieval augmented language models. *J. Mach. Learn. Res.*, pages 251:1–251:43.

Bernard J. Jansen. 2007. Chris anderson, the long tail: Why the future of business is selling less or more, hyperion, new york (2006) ISBN 1-4013-0237-8 $24.95. *Inf. Process. Manag.*, (4):1147–1148.

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. 2017. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension. In *ACL*, pages 1601–1611.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *ICML*, pages 15696–15707.

Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick S. H. Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *EMNLP*, pages 6769–6781.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur P. 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. *Trans. Assoc. Comput. Linguistics*, pages 452–466.

Jing Li, Qiu-Feng Wang, Kaizhu Huang, Xi Yang, Rui Zhang, and John Yannis Goulermas. 2023. Towards better long-tailed oracle character recognition with adversarial data augmentation. *Pattern Recognit.*, page 109534.

Tianming Liang, Yang Liu, Xiaoyan Liu, Hao Zhang, Gaurav Sharma, and Maozu Guo. 2023. Distantly-supervised long-tailed relation extraction using constraint graphs. *IEEE Trans. Knowl. Data Eng.*, (7):6852–6865.

Xi Victoria Lin, Xilun Chen, Mingda Chen, Weijia Shi, Maria Lomeli, Rich James, Pedro Rodriguez, Jacob Kahn, Gergely Szilvasy, Mike Lewis, Luke Zettlemoyer, and Scott Yih. 2023. RA-DIT: retrieval-augmented dual instruction tuning. *CoRR*, abs/2310.01352.

Maja Popovic. 2015. chrf: character n-gram f-score for automatic MT evaluation. In *EMNLP*, pages 392–395.

Zhihong Shao, Yeyun Gong, Yelong Shen, Minlie Huang, Nan Duan, and Weizhu Chen. 2023. Enhancing retrieval-augmented large language models with iterative retrieval-generation synergy. In *Findings of EMNLP*, pages 9248–9274.

Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. 2023. REPLUG: retrieval-augmented black-box language models. *CoRR*, abs/2301.12652.

Matthew G. Snover, Bonnie J. Dorr, Richard M. Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *AMTA*, pages 223–231.

Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. 2023. Interleaving retrieval with chain-of-thought reasoning for knowledge-intensive multi-step questions. In *ACL*, pages 10014–10037.

Boxin Wang, Wei Ping, Peng Xu, Lawrence McAfee, Zihan Liu, Mohammad Shoeybi, Yi Dong, Oleksii Kuchaiev, Bo Li, Chaowei Xiao, Anima Anandkumar, and Bryan Catanzaro. 2023a. Shall we pretrain autoregressive language models with retrieval? A comprehensive study. In *EMNLP*, pages 7763–7786.

Haoran Wang, Yajie Wang, Baosheng Yu, Yibing Zhan, Chunfeng Yuan, and Wankou Yang. 2024. Attentional composition networks for long-tailed human action recognition. *ACM Trans. Multim. Comput. Commun. Appl.*, (1):8:1–8:18.

Yile Wang, Peng Li, Maosong Sun, and Yang Liu. 2023b. Self-knowledge guided retrieval augmentation for large language models. In *Findings of EMNLP*, pages 10303–10315.

Yuning Wang, Pu Zhang, Lei Bai, and Jianru Xue. 2023c. FEND: A future enhanced distribution-aware contrastive learning framework for long-tail trajectory prediction. In *CVPR*, pages 1400–1409.

Zhiruo Wang, Jun Araki, Zhengbao Jiang, Md. Rizwan Parvez, and Graham Neubig. 2023d. Learning to filter context for retrieval-augmented generation. *CoRR*, abs/2311.08377.

Pengyu Xu, Lin Xiao, Bing Liu, Sijin Lu, Liping Jing, and Jian Yu. 2023. Label-specific feature augmentation for long-tailed multi-label text classification. In *AAAI*, pages 10602–10610.

Zhengzhuo Xu, Zenghao Chai, and Chun Yuan. 2021. Towards calibrated model for long-tailed visual recognition from prior perspective. In *NeurIPS*, pages 7139–7152.

Yitong Yao, Jing Zhang, Peng Zhang, and Yueheng Sun. 2024. A dual-branch learning model with gradient-balanced loss for long-tailed multi-label text classification. *ACM Trans. Inf. Syst.*, (2):34:1–34:24.

Yaochi Zhao, Sen Chen, Qiong Chen, and Zhuhua Hu. 2023. Combining loss reweighting and sample resampling for long-tailed instance segmentation. In *ICASSP*, pages 1–5.

Shanshan Zheng, Yachao Zhang, Hongyi Huang, and Yanyun Qu. 2023. Sample-aware knowledge distillation for long-tailed learning. In *ICASSP*, pages 1–5.

Zhisheng Zhong, Jiequan Cui, Shu Liu, and Jiaya Jia. 2021. Improving calibration for long-tailed recognition. In *CVPR*, pages 16489–16498.

Xuesong Zhou, Junhai Zhai, and Yang Cao. 2023. Feature fusion network for long-tailed visual recognition. *Pattern Recognit.*, page 109827.

Model	Type	Accuracy 10	 15	 20	 Avg.	Speed-up
Llama2-7B	w/o GECE	84.9	85.4	85.5	$85.3_{(\pm 0.3)}$	1.0 ×
	w GECE	85.3	86.1	86.4	$85.9_{(\pm 0.3)}$	2.4 ×
IRCoT	w/o GECE	87.3	87.8	88.2	$87.8_{(\pm 0.5)}$	1.0 ×
	w GECE	87.4	88.1	88.6	$88.0_{(\pm 0.4)}$	6.5 ×
SKR	w/o GECE	87.8	89.2	89.6	$88.9_{(\pm 0.1)}$	1.0 ×
	w GECE	89.2	89.6	89.7	$89.5_{(\pm 0.2)}$	6.3 ×
SELF-RAG	w/o GECE	86.3	87.1	87.5	$87.0_{(\pm 0.4)}$	1.0 ×
	w GECE	87.4	87.9	88.0	$87.8_{(\pm 0.3)}$	3.1 ×
FILCO	w/o GECE	86.5	86.6	87.1	$86.7_{(\pm 0.2)}$	1.0 ×
	w GECE	86.0	86.9	87.2	$86.7_{(\pm 0.3)}$	2.2 ×
ITER-RETGEN	w/o GECE	88.7	89.5	89.4	$89.2_{(\pm 0.1)}$	1.0 ×
	w GECE	89.2	89.6	89.8	$89.5_{(\pm 0.2)}$	7.1 ×

Table 4: Experimental results on MMLU. T-tests show the improvements are statistically significant with $p < 0.05$.

A Experimental Settings

For a fair comparison, we set baselines to the same backbone and retriever, i.e., Llama2-7B (Wang et al., 2023d) and DPR (Karpukhin et al., 2020), respectively. The utilization of GECE on SKR replaces the known/unknown judgment with GECE with other baseline operations set as usual. Our experiment results are averaged over multiple runs. The number of retrieved documents by DPR is set to $\{10, 15, 20\}$. The gradient of Equation 2 is obtained from the average gradient of Feed-Forward Networks (FFN) in 29-32 layers. We categorize the instances with the top 20% of large GECE values as long-tail instances and the rest as common instances. The max related document token length is limited to 512. The temperature hyper-parameter of Llama2 is assigned as 0.6, top-p is set to 0.9. Our ablation study is based on the baseline of Llama2-7B and the setting of 20 retrieved documents.

B Supplementary Experimental Results

B.1 Additional Results on the MMLU Dataset

The results over the MMLU dataset are shown in Table 4. The conclusion is also consistent with the results over other datasets, showing the efficacy of the proposed method.

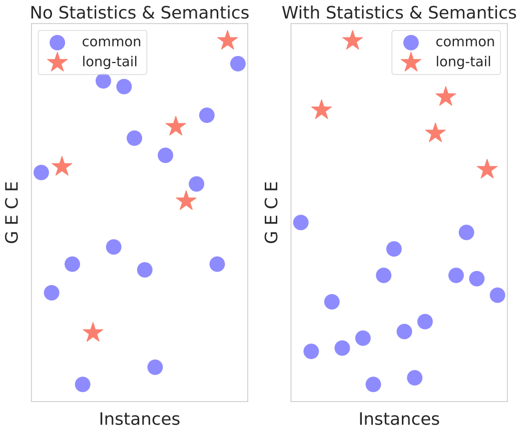

Figure 2: Comparison between absence and presence of statistics and semantics information in GECE.

B.2 Detailed Analysis of Statistics & Semantics Information

To probe the influence of statistics and semantics information, we sample 15 common instances and 5 long-tail instances from NQ and plot the GECE value of the sampled instance in Figure 2. Removing the statistics and semantics information leads to mixed and scattered instance distribution. With the help of the statistics and semantics information, we can separate common and long-tail instances apart distinctly.