Thrust: Adaptively Propels Large Language Models with External Knowledge (Appendix)

Xinran Zhao[1,2] **Hongming Zhang**[1] **Xiaoman Pan**[1] **Wenlin Yao**[1]
Dong Yu[1] **Jianshu Chen**[1]
[1]Tencent AI Lab, Bellevue, [2]Language Technologies Institute, Carnegie Mellon University

A Appendix

A.1 Limitations

Cold Start. In the ideal case, a module distinguishes if a query requires external can do it in a zero-shot manner. However, as we show in Section 4.1 of our paper, we practically find that the distribution of ***Thrust*** scores of various tasks can be very different due to the essence of the task collection and the type of external knowledge needed. Considering such an effect, we need 200 examples to estimate the clusters needed to set up the computation, which is still lightweight in real-world scenarios. In the future, we will explore the usage of meta-learning to allow a few-shot start or even a cold start of IAPEK.

Black-box LLM: At the current stage, our model can work with first-layer or last-layer representations (as discussed in Figure 6 of our submitted paper), which are provided by some black-box models. To adapt to completely black-box LLMs, there are two ways we can think at this stage: (1) similar to Black-Box Tuning [4], we use the prompt embeddings adjusted by the derivative-free optimizer optimized over the black-box model outputs as our representation; (2) we use original or distilled smaller models from the same family to acquire representation (e.g., original GPT-2 or GPT-2 fine-tuned by a set of query and answers from GPT-4). We experimented with using T5-base representation to conduct IAPEK for T5-large models. It showed slightly worse but not completely ruined performance.

Extension to other Retrieval-augmented Models. In this paper, we propose a new module for the pipeline of retrieval-augmented models. We first comprehensively examined if and how external knowledge is useful with language models. Next, we examine the performance of the module **IAPEK** with the lightweight ***Thrust*** score we define as a potential implementation. We compare ***Thrust*** with BM25 with the default setting of retrieval augmented language models [2] and show its effectiveness. Since queries, not answers nor retrieved knowledge are required to set up ***Thrust***, it can be applied to various other frameworks of retrieval augmented models [1]. However, it is beyond the scope of the project at the current stage, and the contribution of our module and the frameworks are orthogonal. We will extend ***Thrust*** to other retrieval-augmented models in future work.

A.2 Implementation details

We conduct our experiments on a machine with 8 Nvidia P40 (24G) GPUs with CUDA 11 installed.

We use the Scikit-learn package [1] to measure the clusters with K-means and compute the distance between the query and cluster representations. The involved hyperparameters (including the number of clusters per class) are selected by Grid search on a smaller set of experiments. We initialize all parameters randomly or as the default of the Hugginface transformers package [2]. On average, each

---

[1] `https://scikit-learn.org/`

[2] `https://huggingface.co/`

37th Conference on Neural Information Processing Systems (NeurIPS 2023).

Table 1: Statistics of the selected datasets. Sample # denotes the number of examples used to calculate the clusters for ***Thrust*** scores. ARC-E and ARC-C denote the easy and hard ARC datasets. Q/K/A Len denotes the average number of words for the Queries/Knowledge/Answers, respectively.

Dataset	**Source**	**Sample #**	**Test #**	**Q Len**	**K Len**	**A Len**
AGNews	gold	200	7,600	8.1	35.9	1.0
e-SNLI	human	200	9,824	24.9	14.3	1.0
StrategyQA	human	200	229	10.8	33.5	1.0
CIKQA	KG	200	604	18.2	28.0	1.0
BoolQ	retriever	200	3,270	9.8	113.8	1.0
ARC-E	retriever	200	570	23.1	238.2	4.2
ARC-C	retriever	200	299	26.2	240.5	5.5
HotpotQA	gold	200	7,405	19.0	56.3	2.5
NQ	retriever	200	6,468	10.1	588.9	2.3
Web Questions	retriever	200	278	7.8	117.3	4.3
Curated TREC	retriever	200	116	8.4	116.5	7.7
TriviaQA	retriever	200	6,760	15.0	117.6	27.5

run of extracting the results for all the tasks under with/without knowledge cases takes around 20 hours. We run all experiments 3 times and report the averaged performance in the main content. For hyperparameters of the inference models, for the QA task, we set the maximum knowledge length as 480 tokens to ensure that query sentences stay in the input.

The generated answer for QA tasks for all the models is typically within 30 tokens. For classification tasks, for binary classification tasks (CIKQA, StrategyQA, BoolQ, and e-SNLI), we follow previous work to use "Yes or No?" as the suffix to the original query to guide the generative models. For AGNews, we use "political news, sports news, business news, and technology news" as the label words. We found that the default label word "word news" will largely degrade the performance of generative models on AGNews. We add "the news is about?" and provide the candidate categories as the suffix for AGNews. More details of our implementations can be found in the code attached.

A.3 Dataset Details

The detailed statistics of the involved datasets are shown in Table 1. We sample 200 data points from each dataset to conduct the clustering step of ***Thrust***. Difference datasets have different average query lengths and knowledge lengths due to the essence of the task creation and knowledge collection. Answer length 1 denotes tasks with yes and no answers. Otherwise, the answers with more than one token are either choices (for ARC-E and ARC-C) or free-form text sentences (for open-domain QA tasks). Examples of the dataset can be found in the attached data.

A.4 Experiment with Flan-T5

Figure 1 presents the performance of ***Thrust*** on CIKQA with different models. From the figure, we can observe that ***Thrust***performs better with instruction fine-tuned Flan-T5 compared to the original T5 and UnifieedQA. With Flan-T5 ***Thrust*** achieves better performance with 40% examples rejecting external knowledge usage compared to external knowledge used either on no or all examples. Such observations show the potential of using ***Thrust*** on current instruction-finetuned models.

A.5 Ablation on the design choices of *Thrust*

Following [5], we use a few-shot multitask binary NLI dataset to test the influence of each factor of ***Thrust*** (i.e., FS-NLI), through measuring how well the metric and its variants can measure the with the hardness of a diverse set of datasets. From Table 2, we can observe that all the design choices are crucial to the success of using ***Thrust*** to detect how hard a query is for a given task and model.

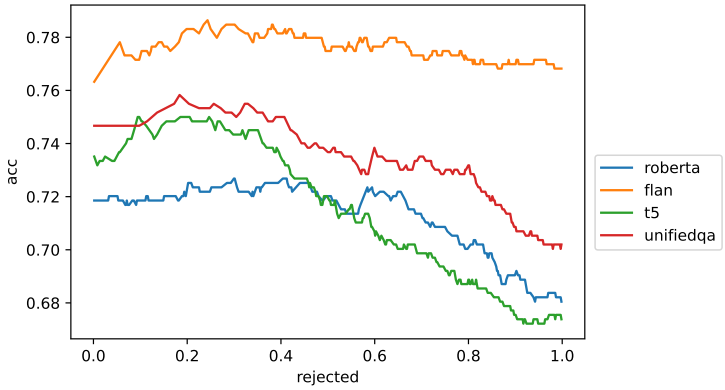

Figure 1: Performance of different models on CIKQA with different thresholds of ***Thrust***. The X-axis denotes the portion of test examples that are selected to not use external knowledge. All model names denote the large versions of the model parameters.

Table 2: Compare ***Thrust*** to its various variants following the setting of the original work [5], where higher correlation denotes that the metric can better capture the hardness of tasks with respect to a given model (RoBERTa-large [3]). *without direction* denotes the variant to use scalar instead of vectors for ***Thrust***. The best-performing entry is marked in bold.

Metric	Correlation
Thrust	**0.45**
without cluster size	0.23
without direction	0.19
without distance	0.06
cosine distance	0.08
one cluster per label class	0.32
ten clusters per label class	0.12
cluster size to inertia	0.03

A.6 Full distribution of *Thrust* across tasks

Figure 2 demonstrates the distribution of ***Thrust*** scores for each of the involved tasks. Besides the findings in the main content that ***Thrust*** can help identify the knowledge necessity for various tasks through viewing the distribution, we can also observe that ***Thrust*** can lead to a diverse distribution of scores that may contain multiple peaks.

A.7 Performance of using external knowledge (in table)

Table 3 presents the performance in Figure 3 of the original submission in a table format. Similarly, we can observe that it is not trivial to use external knowledge, especially in the zero-shot settings, it is possible that models get worse performance with external knowledge, for example, for ARC-C, both T5-base and T5-large show worse performance with the extra knowledge injected. Also, we can observe that external knowledge is crucial for open-domain QA tasks. The gain can be huge, for example, for UnifiedQA-3b, the performance is improved from 18.6 to 80.0, in terms of QA-F1 on TriviaQA, with the external knowledge.

References

[1] 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,

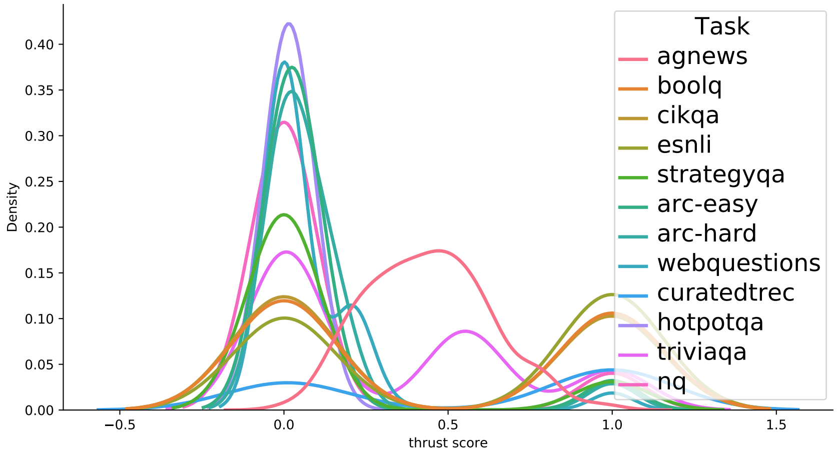

Figure 2: Distribution of ***Thrust*** scores for all involved tasks with UnifiedQA-3b to create the instance representation. The distribution is normalized by Kernel Density Estimation. Low scores denote the cases where internal knowledge is not enough, and vice versa.

Table 3: Performance of various models on the MC classification tasks (accuracy) and open-domain QA tasks (QA-F1). Performances without/with knowledge external knowledge are presented before/after the vertical bar, respectively. UnifiedQA-X denotes T5 models with corresponding sizes fine-tuned on the UnifiedQA dataset.

Model	parameters	AGNews	e-SNLI	CIKQA	StrategyQA	BoolQ	ARC-E	ARC-C
Zero-shot								
T5-base	220M	30.2 | 44.4	65.2 | 65.1	51.5 | 51.8	54.1 | 50.2	48.3 | 38.9	27.8 | 28.8	31.4 | 29.4
T5-large	770M	25.8 | 25.2	65.7 | 65.7	50.0 | 50.0	53.3 | 53.3	37.8 | 38.6	25.1 | 27.7	27.7 | 24.7
T5-3b	3B	27.9 | 39.1	57.6 | 61.5	52.6 | 50.5	44.5 | 48.9	56.6 | 45.3	25.8 | 26.0	26.4 | 28.4
GPT-J	6B	25.1 | 26.9	40.8 | 37.0	49.8 | 50.7	47.2 | 55.9	60.2 | 47.2	25.4 | 29.5	28.4 | 27.1
OPT-30b	30B	25.0 | 25.0	65.7 | 65.7	50.0 | 50.0	53.3 | 53.3	37.8 | 37.8	27.4 | 27.7	25.8 | 26.4
Transfer-learning								
UnifiedQA-base	220M	46.6 | 35.7	38.5 | 70.2	56.0 | 59.6	48.5 | 57.2	60.4 | 80.8	50.2 | 61.6	44.8 | 45.2
UnifiedQA-large	770M	71.0 | 67.9	42.8 | 74.2	59.6 | 62.1	48.5 | 66.4	59.8 | 84.5	64.0 | 66.0	55.2 | 49.5
UnifiedQA-3b	3B	75.7 | 84.5	62.2 | 89.6	61.3 | 66.9	57.6 | 83.4	61.5 | 87.8	73.7 | 76.5	64.5 | 64.2

Model	parameters	Web Questions	Curated TREC	HotpotQA	NQ	TriviaQA
Zero-shot						
T5-base	220M	6.7 | 8.7	2.5 | 3.8	6.0 | 9.1	1.9 | 6.0	8.9 | 13.1
T5-large	770M	5.7 | 7.4	1.9 | 3.0	5.1 | 6.6	1.6 | 2.7	8.3 | 9.0
T5-3b	3B	4.9 | 4.0	2.0 | 1.0	4.9 | 6.8	1.7 | 6.6	8.3 | 5.6
GPT-J	6B	4.3 | 6.3	7.4 | 2.7	5.9 | 5.1	10.9 | 6.9	1.7 | 2.1
OPT-30b	30B	18.3 | 6.3	16.0 | 2.4	11.4 | 2.4	5.3 | 2.1	16.3 | 6.9
Transfer-learning						
UnifiedQA-base	220M	10.6 | 44.2	3.6 | 36.9	13.1 | 40.3	3.0 | 34.6	11.7 | 65.6
UnifiedQA-large	770M	12.9 | 46.5	8.2 | 36.6	14.2 | 42.7	3.8 | 36.3	13.7 | 74.6
UnifiedQA-3b	3B	11.5 | 48.1	9.9 | 41.8	17.0 | 47.0	4.4 | 37.6	18.6 | 80.0

Oriol Vinyals, Simon Osindero, Karen Simonyan, Jack W. Rae, Erich Elsen, and Laurent Sifre. Improving language models by retrieving from trillions of tokens, 2022.

[2] Gautier Izacard, Patrick Lewis, Maria Lomeli, Lucas Hosseini, Fabio Petroni, Timo Schick, Jane Dwivedi-Yu, Armand Joulin, Sebastian Riedel, and Edouard Grave. Few-shot learning with retrieval augmented language models. *arXiv preprint arXiv:2208.03299*, 2022.

[3] Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. Roberta: A robustly optimized BERT pretraining approach. *CoRR*, abs/1907.11692, 2019.

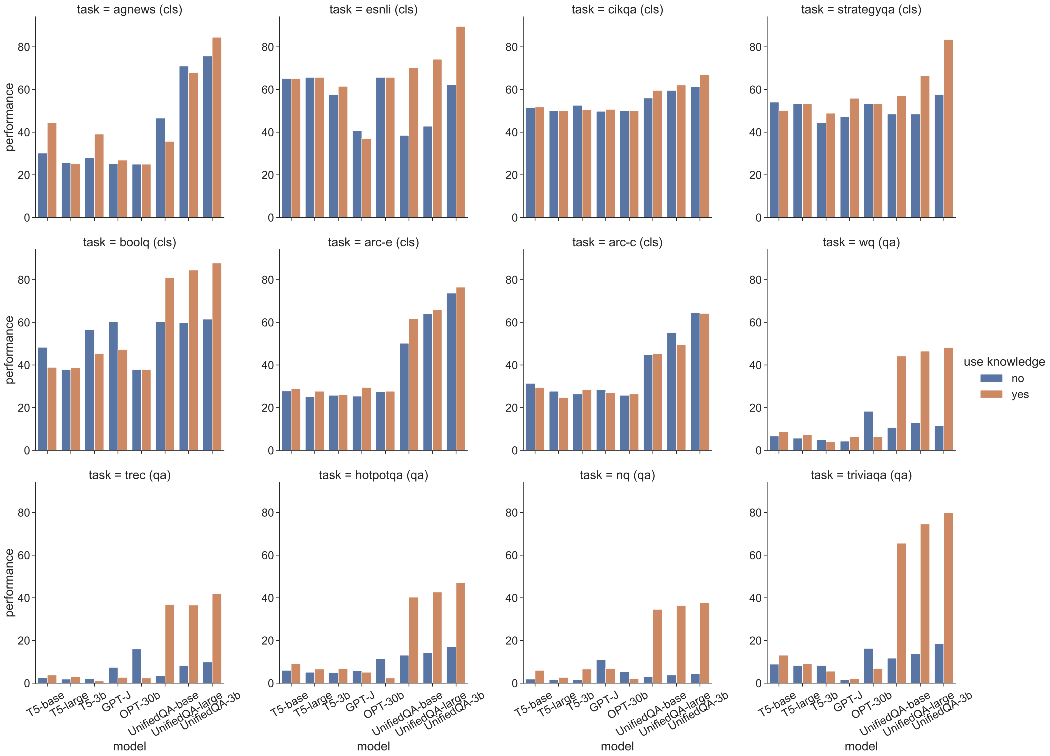

Figure 3: Performance of various models on MC classification tasks (accuracy) and open-domain QA tasks (QA-F1), denoted by (cls) and (qa), respectively. The x-axis represents the model names, which are shared across sub-figures. Use knowledge: yes or no denotes using full knowledge or not for all queries. UnifiedQA denotes T5 models with different sizes fine-tuned on the UnifiedQA dataset.

[4] Tianxiang Sun, Yunfan Shao, Hong Qian, Xuanjing Huang, and Xipeng Qiu. Black-box tuning for language-model-as-a-service. In *Proceedings of ICML*, 2022.

[5] Xinran Zhao, Shikhar Murty, and Christopher D. Manning. On measuring the intrinsic few-shot hardness of datasets. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2022.