QUILL: Query Intent with Large Language Models using Retrieval Augmentation and Multi-stage Distillation

Krishna Srinivasan* Google Research krishnaps@google.com

Karthik Raman* Google Research karthikraman@google.com

Anupam Samanta Google anupamsamanta@google.com

Lingrui Liao Google lingrui@google.com

Luca Bertelli Google lbertelli@google.com

Mike Bendersky Google Research bemike@google.com

Abstract

Large Language Models (LLMs) have shown impressive results on a variety of text understanding tasks. Search queries though pose a unique challenge, given their short-length and lack of nuance or context. Complicated feature engineering efforts do not always lead to downstream improvements as their performance benefits may be offset by increased complexity of knowledge distillation. Thus, in this paper we make the following contributions: (1) We demonstrate that Retrieval Augmentation of queries provides LLMs with valuable additional context enabling improved understanding. While Retrieval Augmentation typically increases latency of LMs (thus hurting distillation efficacy), (2) we provide a practical and effective way of distilling Retrieval Augmentation LLMs. Specifically, we use a novel two-stage distillation approach that allows us to carry over the gains of retrieval augmentation, without suffering the increased compute typically associated with it. (3) We demonstrate the benefits of the proposed approach (QUILL) on a billion-scale, real-world query understanding system resulting in huge gains. Via extensive experiments, including on public benchmarks, we believe this work offers a recipe for practical use of retrieval-augmented query understanding.

1 Introduction

The recent advent of billion+ parameter Large Language Models (LLMs) – such as T5 (Raffel et al., 2019), mT5 (Xue et al., 2021), GPT-3 (Brown et al., 2020) and most recently PaLM (Chowdhery et al., 2022) – has disrupted many language understanding tasks – with new benchmarks set or eclipsed routinely by these Transformer models and their variants.

Queries – especially keyword search ones – present a unique challenge though. Their short length, inherent ambiguity and lack of grammar mean query understanding tasks typically require more memorization and world knowledge than other NLP tasks (Broder et al., 2007). Consequently, despite LLMs leading performance on language and query understanding tasks – like intent classification, query parsing and relevance prediction – there is significant room for further improvement.

In this paper we leverage Retrieval-Augmentation to provide LLMs more context and grounding for search queries. We show that the titles and URLs of documents retrieved for the query, greatly help improve LLMs query understanding capabilities. While different retrieval augmentation models exist, we show that even simple concatenation of these titles / urls with the query can help improve LLM performance considerably.

However, the use of retrieval augmentation leads to a new challenge: Increased complexity of LLM inference. More specifically, the quadratic complexity of self-attention in Transformer models means that the latency of LLMs blows up given these (often 10x+) longer input sequences. This presents a significant problem as LLMs are impractical for online use and thus need to be distilled into smaller, more efficient models to be served online. However knowledge distillation (Gou et al., 2021) into these *student* models requires a lot of distillation data annotated by these LLMs – which may not be feasible for these retrieval augmented models.

Thus as a remedy we introduce a new two-stage distillation approach. In the first stage of this approach we distill the retrieval-augmented (long input) LLM (the *Professor*) into a non-retrieval augmented (short input) LLM (the *Teacher*) using a small distillation set. This second LLM *Teacher* is in turn distilled into the final *Student* using a large set.

* Corresponding Authors

Proceedings of EMNLP 2022 Industry Track, pages 502–511
December 9–11, 2020.

Via extensive experiments on a large-scale, real-world problem and data we demonstrate that the resulting QUILL system provides for an efficient and effective way of retaining the performance gains of retrieval augmented LLMs on query understanding tasks.

2 Related Work

Large language models (LLMs) such as mT5 (Xue et al., 2021) demonstrated significant performance improvements on a variety of natural language understanding (NLU) tasks. Specifically in the context of query understanding, researchers found that (a) model size significantly effects the quality of the resulting models (Nogueira et al., 2019; Han et al., 2020), and (b) using additional context in the form of query-associated documents is crucial to the model performance due to the paucity of context available in the query itself (Nogueira and Lin, 2019; Zhang et al., 2020). Retrieval augmentation of the query with the search results retrieved by it is a proven way to incorporate such context in LLM training for NLU tasks, as has been shown recently by models such as RAG (Lewis et al., 2020), REALM (Guu et al., 2020), and RETRO (Borgeaud et al., 2022).

In this paper, we leverage this insight to improve performance of query intent prediction (Broder, 2002) — a crucial query understanding task that is at the heart of modern search engines – using LLMs. Prior work by Broder et al. (2007) found importance of retrieval augmentation using statistical methods for this task. Statistical retrieval augmentation has also been found critical for other query understanding tasks including query expansion (Broder et al., 2008; Diaz and Metzler, 2006) and query tagging (Wang, 2020). We demonstrate similar benefits when using retrieval-augmented LLMs as well.

We also leverage Knowledge Distillation (Hinton et al., 2015; Mirzadeh et al., 2020; Gou et al., 2021) techniques to create a Student model that retains the LLMs gains.

3 Query Intent Understanding

While the techniques described in this paper could be applied to any query understanding task, for the sake of brevity we focus on the task of query intent classification. Query intent (QI) classification is a classical IR task studied for over two decades (Kang and Kim, 2003; Baeza-Yates et al., 2006; Jansen et al., 2008; Kathuria et al., 2010; Lewandowski et al., 2012; Figueroa, 2015; Mohasseb et al., 2019). This task is particularly important in practice, as it is at the top of the search funnel, and the entire search engine behavior may vary based on the predicted query intent. Given the centrality of this task on overall retrieval, models for this task need to be both fast (*i.e.,* low latency) and high efficacy. Thus even a single percentage point quality gain on the QI task can be considered a major accomplishment.

Data	Train	Val	Test	Unlabeled
Orcas-I	1.28M	1K	1K	10.3M
EComm	36K	4K	4K	128M

Table 1: Statistics of datasets used.

In this paper we tackle the QI task using LLMs. In particular we use two datasets in our study whose details are provided in Table 1:

- **EComm**: Our main dataset will be a **real-world** dataset. Cast as a binary classification problem, this task involves identifying queries with a specific intent – where the required intent is similar to the *transactional* intent of the Broder taxonomy (Broder, 2002) in the context of e-commerce. As common in real-world applications, the human labeled data is accompanied by a large unlabeled set – that is used for knowledge distillation.

- **Orcas-I**: The largest publicly available query intent dataset is ORCAS-I (Alexander et al., 2022). This comprises queries of the ORCAS dataset (Craswell et al., 2020) labeled with one of 5 intent classes. Note that while the test set is human-labeled, the training set labels are weak labels as detailed in ORCAS-I (Alexander et al., 2022) paper's Methodology section.

4 QUILL Methodology

The keyword nature of queries and lack of context make the QI task (like other query understanding tasks) challenging for LLMs. Thus we propose QUILL as a solution. As seen in Figure 1, QUILL consists of two stages: (a) Retrieval Augmented LLM training, (b) Multi-Stage Distillation into efficient student.

Retrieval Augmented (RA) LLM: The key insight here is that titles / urls of related documents

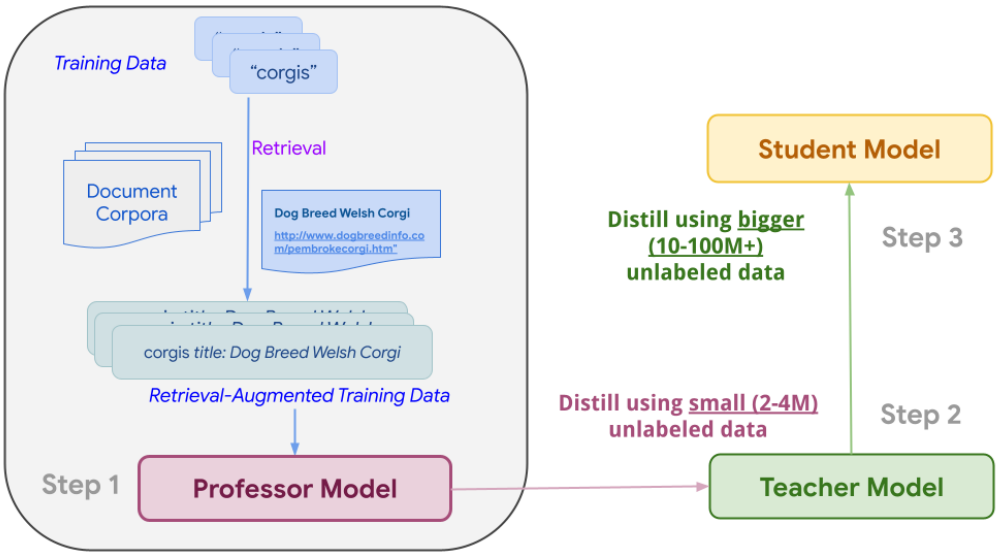

Figure 1: QUILL Architecture : Retrieval Augmentation and Multi-stage Distillation.

Feature	EComm		Orcas-I	
	Median	99%	Median	99%
Query	5	18	5	10
ExpandTerms	17	28	N/A	N/A
(Up to) 10 Titles	157	245	13	104
(Up to) 10 URLs	159	304	39	238

Table 2: mT5 sequence lengths by features.

could provide valuable context to help understand the intent of the query. For example, it may not be immediately apparent what a query like `ua 1234` may mean. However, via the retrieved documents we can understand that the query is seeking information about a United Airlines flight.

While there are multiple ways of augmenting the input via retrieved documents (example: the Fusion-in-Decoder architecture (Izacard and Grave, 2020)), we chose to study the most straightforward and popular approach of concatenating the titles / urls of the retrieved top-k documents with the original query as the input to our LLM. As shown empirically (Sec 5), this model outperforms all baselines – demonstrating the value of additional context.

Multi-stage distillation: The drawback of RA is the additional sequence length of the input. As seen from Table 2, augmenting a query with (upto) 10 titles and urls increases the sequence length by an order of magnitude. Consequently, this makes distillation far more challenging given the quadratic complexity of sequence length (due to self-attention) in transformer models. This leaves us in a dichotomy between a more effective model with a much smaller distillation set, vs. a lower performing model with a larger distillation set. Given a large distillation set is required for training an effective student, this leaves us at risk of not being able to benefit from RA, given that a very large dataset with RA will incur very long and impractical inference times.

To get the best of both worlds we propose a two-stage distillation approach. In the first stage we distill the *Professor* RA LLM into a *Teacher* LLM without RA. The Teacher model uses ExpandTerms which provide additional context to the queries. While this may not be as expressive as retrieval augmentation, this provides a good compromise of greatly reducing sequence length while giving up only a little in performance. We do so by using a small subset of the unlabeled data. As shown empirically, a LLM teacher trained in this manner performs significantly better than a non-RA LLM trained directly on the human data, while at the same time allowing us to efficient distillation.

In the second stage we use the *Teacher* LLM to annotate the entire unlabeled dataset. This is in turn used to train the final *Student* model that will be used in practice.

5 Experiments and Results

Experimental Setup: Our experiments were all conducted using the mT5 (Xue et al., 2021) checkpoints. We validate performance across three learning rates (1e-3, 1e-4, 5e-5) – selecting the best checkpoint using the validation set loss. For mod-

els trained from the provided training sets, we used a batch size of 64 in our experiments and trained for 4K steps (EComm) / 20K steps (Orcas-I).

For distilled models, we used a batch size of 128 for Teacher models and 1024 for Student models. We use different batch sizes because of the model architectures, mT5 for the Teacher vs a BERT-based model for the Student. In both cases, we trained for 1 epoch, unless mentioned otherwise. We only use the encoder of the mT5 model with an additional layer added on top to predict the classification scores. The Professor, the Teacher and the Student fine-tuning experiments are all set up as a query intent classification task. Given that the Teacher and Student models are trained on millions of examples and this itself is a time and resource intensive step, we restrict our experiments to only one epoch. We demonstrate performance gains even with one epoch via the techniques elaborated in this paper.

We studied the effect of distillation data size, for both stages of distillation. For the EComm dataset, we used an in-house retriever to find related documents. For Orcas-I, we use the provided docids (aggregated at per-query level) for retrieval augmentation. Unless specified, we use (upto) the top-10 results for retrieval augmentation[1]. Sequence length for models are based on the training set and features (set to 99%-percentile of sequence lengths).

Students and Features: Our experiments demonstrate results for a fast, efficient 4-layer transformer student architecture, with hidden dimensionality of 256. We default to using the query as the only feature in the student for simplicity. To compare against query expansion techniques, we used a sophisticated in-house memorization-based query expansion model in our Professor / Teacher experiments on EComm. This expansion model – which we refer to as **ExpandTerms** – provides a list of related terms for a given query, which are concatenated with the query (and identifiers for start / end of each feature).

Metrics: To compare performance of different models we use two metrics: **MicroF1** and **MacroF1** for Orcas-I, and **AUC-PR** and **AUC-ROC** for EComm. For EComm, we only report performance of models relative to the mT5 query-only Base-sized model[2].

[1]For Orcas-I, nearly 2/3rd of the queries only have a single provided result, while some have upwards of 2000 results, which is why the lengths for RA features on Orcas-I in Table 2 are smaller. The 10 results augmented are randomly chosen if more exist.

Model	Size	ROC	PR
query	Base	0.0%	0.0%
+ RA (titles, urls)	Base	**+4.3%**	**+4.6%**
query	XL	+2.7%	+3.1%
+ RA (titles, urls)	XL	**+6.3%**	**+6.7%**
query	XXL	+3.0%	+3.3%
+ RA (titles, urls)	XXL	**+6.4%**	**+6.9%**

Table 3: Results demonstrating the benefit of Retrieval Augmentation (RA) across all model sizes.

EComm	ROC	PR
query	0.0%	0.0%
+ Terms	+2.6%	+1.9%
+ RA (titles)	+4.8%	+4.8%
+ RA (titles) + Terms	+5.1%	+5.2%
+ RA (urls)	**+5.3%**	**+5.7%**

Table 4: Analysis of the impact of different features (using Base-sized models) for the EComm dataset. ExpandTerms abbreviated as Terms.

5.1 Effect of Retrieval Augmentation

While the use of retrieval augmentation (RA) has been known to improve query classification performance (Broder et al., 2007), the benefit of RA is unclear in the age of LLMs. Thus, we start by evaluating the first stage of QUILL *i.e., the RA model*. As seen in Table 3, RA improves performance significantly across all model sizes including the billion-parameter+ XL and XXL models. In fact the gains from RA on the Base-sized model exceed the gains obtained by increasing model size of a query-only model to XXL. Given the gains observed across all models sizes, we use Base-sized models in the rest of the paper to simplify experimentation.

[2]For a sense of scale, each 0.5% point increase in metrics on EComm is considered a significant gain.

EComm	ROC	PR
Orcas-I	**MicF1**	**MacF1**
query	69.8	69.75
+ RA (titles)	+6.3%	+5.1%
+ RA (urls)	**+8.2%**	**+6.2%**
+ RA (titles+urls)	**+9.0%**	**+7.2%**

Table 5: Analysis of the impact of different features (using Base-sized models) for the Orcas-I dataset.

EComm	ROC	PR
Baseline Teacher (Finetuned on Training Set)	+2.6%	+1.9%
QUILL Teacher (2M Prof Distilled Set)	+3.3%	+2.8%
QUILL Teacher (4M)	**+3.4%**	**+2.9%**
QUILL Teacher (8M)	+3.5%	+2.9%
QUILL Professor	**+5.3%**	**+5.7%**

Table 6: Comparison of different Teacher models trained directly or via Professor-distillation for the EComm dataset.

Orcas-I	MicF1	MacF1
Baseline Teacher (Finetuned on Training Set)	69.8	69.75
QUILL Teacher (2M Prof Distilled Set)	**+1.1%**	**+0.8%**

Table 7: Comparison of different Teacher models trained directly or via Professor-distillation for the Orcas-I dataset.

A natural question that may arise though is how do these gains from RA compare to those obtained by powerful query expansion techniques. Thus, we performed an in-depth ablation of features for the EComm dataset (on a Base-sized model for ease of experimentation) as seen in Table 4 and for the Orcas-I dataset as seen in Table 5. These results clearly demonstrate the potency of powerful query expansion models (*i.e.,* **ExpandTerms**) – as evidenced by the large ∼2% gains over query-only models. However, we find that RA adds even more value over these highly sophisticated expansion models with an a nearly 5+% increase in performance. Furthermore, we find that RA techniques can still be combined with query expansion for further gains.

The improvements on RA for Orcas-I (seen in Table 5) are even more substantial, with a nearly 9% improvement over the query-only baseline, via the use of the titles and urls of related documents. Interestingly, among RA features we find that urls tend to perform slightly better than titles on both datasets. We believe this to be because titles can have a higher variance of informativeness – with both highly verbose and very short titles commonly seen. Hence, given the simplicity and consistency of urls, we chose to use RA(urls) for subsequent experiments as the *Professor* model.

5.2 Distilling gains from RA

We next focus on the second stage of QUILL: Distilling the RA model. Typically larger amounts of distillation data lead to better performance. However, given the increased sequence length of RA models and the cost of retrieval augmentation itself, annotating large distillation sets is highly challenging. Thus to capture such practical trade-offs, we only used a small subset of the unlabeled data for the QUILL *Professor* to *Teacher* distillation. In particular, we used 4M examples for EComm (*i.e.,* 3.1% of unlabeled data) and 2M for Orcas-I (19%) for this first stage of distillation – to represent a set that is small enough set to be practical, but large enough to learn from. However, we do share results for varying this size to understand its importance.

QUILL *Teacher* models were thus trained by distilling the RA(urls) *Professor* models. Our *Teacher* models had the same capacity and architecture (*i.e., mT5-Base* as the Professor[3] – except it does not use RA (features).

As a realistic and competitive baseline, we chose a *Baseline Teacher* that resembles the QUILL *Teacher* in all aspects bar one – the data they are trained on. Specifically, the Baseline Teacher is directly trained from the gold-labeled training data, unlike the QUILL teacher. We believe this is representative of practical applications today, where LLMs are trained directly on gold-labeled sets (before being distilled into the final student models). To further challenge QUILL, we leverage the powerful **ExpandTerms** features (for the EComm dataset) in our Teacher models – both Baseline and QUILL. We believe this provides a more challenging but realistic evaluation setup, since many baseline models in use today avail of powerful features (along with the query).

As seen from the results in Table 6 and Table 7, we find the QUILL Teachers provide a significant performance improvement over the Baseline Teacher, despite having never directly seen the gold label data. On EComm, despite using an enhanced (realistic) baseline, QUILL teachers are ∼1% better on all metrics. We find a similar gap on Orcas-I despite the Teacher there being trained on only 2M examples (just 1.5x the training set size). Put differently, we now have trained our non-retrieval aug-

[3]We observed similar trends even if the Teacher had less capacity than the Professor.

Model (# Distillation)	ROC	PR
No Distillation Student	-6.3%	-7.3%
Baseline Student	-0.9%	-1.6%
QUILL Student	**+2.0%**	**+1.5%**
QUILL 1-Stage Student(4M)	+0.4%	-0.2%
QUILL 1-Stage Student(32M)	+1.1%	+0.6%

Table 8: Performance of the different student models trained from different teachers and using differing amounts of distillation data (on EComm).

mented language model to benefit from the gains of retrieval augmentation. Even though the student model does not have Retrieval Augmentation, because of the Teacher model's performance improvement, it is possible to annotate a considerably large number of training examples. We observe the Student models to close the gap (compared to the Teacher) given larger training datasets.

To test the robustness of QUILL teachers we also varied the amount of distillation data used – halving or doubling it. While there still exist distillation gaps to the professor (which can be narrowed via more distillation data) on both datasets, our proposed approach works well even when using small amounts of distillation data – which in turn allows us to save significant compute.

Query	Example URL	W/L
bengals	sports.yahoo.com/nfl/teams/cin/	✓
pah compounds	en.wikipedia.org/wiki/Polycyclic_aromatic_hydrocarbon	✓
launch tech usa	launchtechusa.com/	✓
noun university	en.wikipedia.org/wiki/Noun	✗
airbed uk	www.airbnb.co.uk/	✗

Table 9: Wins/losses examples on Orcas-I.

5.3 Final student training

So far, we have shown that QUILL can learn a better (non-RA) teacher. However, an important question remains unanswered: Can these Teacher gains be translated to the final student model? In particular, we postulate that the predictions of the QUILL Teacher may be more robust and easier to learn (for student models) than those of the Baseline. To verify this hypothesis we compared 4 fast student models (4-layer encoder-only models), with the only difference being the data they were trained on:

- *No Distillation Student*: This is the simple solution of directly training the Student using the labeled data.
- *Baseline Student*: This is the current standard involving distilling the Baseline Teacher model using the full unlabeled set.
- *QUILL Student*: This is the proposed solution involving distilling the QUILL Teacher model using the full unlabeled set.
- *QUILL 1-Stage Student*: Rather than the two stage distillation approach, this student is directly distilled from the Professor using a subset of the unlabeled data.

As seen from Table 8, all QUILL-based students significantly outperform the Baseline Student. In particular our proposed 2-stage approach leads to a $\sim$3 point gain on both metrics. This is notable in that the gap between QUILL and Baseline students is even higher than the Teachers – which we attribute to the QUILL Teacher labels being more robust.

Comparing different QUILL students, we find that there is a notable performance gain by first distilling into a non-RA teacher, before distilling into the final student. While 1-stage distillation performance improves as more data is used, even when 1/4th of all unlabeled data is retrieval-augmented and annotated by the Professor for direct distillation, it still falls short of the 2-stage approach. Together, these results show: (1) QUILL students outperform the current state-of-the-art significantly, and (2) QUILL benefits from the 2 stage distillation of Professor to Teacher to final student.

5.4 Examples of Wins/Losses from RA

While the previous sections focused on demonstrating the efficacy (and efficiency) of QUILL, we wanted to also understand **why** and **where** are some of these gains from RA stem from. To do so we used the test-set of Orcas-I and sampled illustrative examples of wins / losses (Table 9) between the baseline and the retrieval-augmented professor models. One common win pattern we found for RA models is when the query is unclear, or uses technical terms / abbreviations. In these cases, the augmented urls / titles help provide additional context for the language model to understand what the

query is about. On the flip side, we also found the biggest loss pattern to be when retrieval was inaccurate, which in turn misled the model regarding the query intent. For example, we found our retriever returned wikipedia results more often than it should, which mislead the model to believe the query had *Factual* intent.

6 Future Work

While we studied the problem of query intent classification in this paper, the approach proposed in our paper is general and could be applied to any query understanding task. Following our approach, could enable myriad query understanding tasks use retrieval augmentation in a practically realistic and efficient manner. We leave this to future work though. We should also note that our experiments reveal non-trivial distillation gaps in both stages of distillation, which we believe is another open opportunity for future improvements.

7 Conclusion

This paper provides a practical recipe for combining Retrieval Augmentation and Large Language Models. In particular, we proposed QUILL as an approach to tackle the problem of query intent classification. Our empirical study demonstrates conclusively that Retrieval Augmentation can provide significant value over existing approaches. Furthermore we show that via our two-stage distillation approach, that QUILL not only learns better performing, more robust teachers, but also leads to even bigger gains when distilled into fast, real-world capable production student models.

8 Acknowledgements

We sincerely thank Jiecao Chen, William Dennis Kunz, Austin Tarango, Lee Gardner, Yang Zhang, Constance Wang, Derya Ozkan, Nitin Nalin, Raphael Hoffmann, Iftekhar Naim, Siddhartha Brahma, Siamak Shakeri, Hongkun Yu, John Nham, Ming-Wei Chang, Marc Najork, Corinna Cortes and many others for their insightful feedback and help. We also thank the EMNLP Reviewers for their thorough review, feedback and suggestions.

Limitations

This paper focuses on efficient and effective way of improving query intent classification using Retrieval Augmentation (RA) and Multi-stage distillation. While we have made the best attempts to ensure a robust and efficient method, we would be remiss to not point out some key limitations of our work:

- **Quality of Retrieval:** A key reason for the gains seen in this paper is the use of Retrieval Augmentation. This additional context provided in the form of result titles / URLs are helpful, but are dependent on the quality of the retrieval system. While we did not get a chance to explore the dependence of performance gains on retrieval quality, we plan to explore this in future work.

- **Dependency of Retrieval:** While our approach provides for a practical and low-compute way of incorporating retrieval augmentation, it still does add some compute (to augment the datasets) and system complexity. While we considered this trade-off well worth it in our use case, this may depend on specific settings.

- **Retrieval-Augmentation techniques:** As discussed in Section 4, we used a simple concatenation based retrieval augmentation. However, there do exist more sophisticated techniques for retrieval augmentation. For example, models built on a Fusion-in-Decoder (Izacard and Grave, 2020) backbone have demonstrated great performance (Hofstätter et al., 2022b; Izacard et al., 2022) and improved efficiency (Hofstätter et al., 2022a). We believe that these more sophisticated retrieval-augmentation technique may bring further improvements in our system and leave this for future work to follow up on.

- **Datasets:** The lack of large public query sets means that we were very limited in terms of what public benchmarks we could study this problem on. While ORCAS-I is the largest such available set, they lack many alternatives that are large enough to study the effects of distillation. In the future though, we hope to use the (somewhat related) problem of question-answering where larger datasets (with large enough unlabeled data) exist for a more thorough study.

- **Distillation gaps:** Our results also clearly demonstrate large distillation gaps in both stages. While there have been innovative techniques proposed to improve distillation performance, we intentionally chose to keep things simple as those approaches are largely complementary to the problem we study in this work.

- **Limited "Large" Model Experiments:** While our work is intended for and positioned in the context of "Large" Language Models, we realize that our most common model choice (mT5-Base), may not be the most representative model in that category. This was an intentional choice on our end as we hoped doing so would make the work more relevant to use cases and applications with more limited compute. For practitioners interested in models with tens of billions of parameters, we refer them to our analysis of mT5-XXL sized models in Table 3, that demonstrates the viability of our approach on models of that scale.

Ethics Statement

In this paper, we used only publicly available Language Model and Checkpoints that have been previously published – namely mT5.

An important consideration when working with query datasets is data privacy. This is perhaps the biggest reason why there do not exist many large public query datasets. We intentionally chose ORCAS-I for this reason, as it is constructed from the ORCAS query set – which is widely regarded as a well-constructed, non-PII, sufficiently anonymized query dataset. While the EComm dataset used in this paper is proprietary, we should note that it too has been scrubbed of PII and aims to follow the same (if not higher) data privacy principles. Our data (and methodology) do not contain any information for or target any demographic or identity characteristics.

The task we focus on – query classification – is a general problem that benefits everyone. In fact, it can enable better IR systems thereby benefiting users who otherwise might not get answers. Thus, we do not anticipate any biases or misuse issues stemming from this. We believe that by using

publicly available and vetted retrieval models, the resulting retrieval augmented models should not create any new or further any existing biases.

In many ways a goal of our work is making retrieval augmentation more practical and reducing compute needs for any such applications. While we did present results with XXL sized models, we focused most of our experiments on the smaller, more efficient Base-sized models so as to benefit a wider section of our community and to reduce the computational needs of our experiments.

References

Daria Alexander, Wojciech Kusa, and Arjen P. de Vries. 2022. ORCAS-i. In *Proceedings of the 45th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM.

Ricardo Baeza-Yates, Liliana Calderón-Benavides, and Cristina González-Caro. 2006. The intention behind web queries. In *International symposium on string processing and information retrieval*, pages 98–109. Springer.

Sebastian Borgeaud, Arthur Mensch, Jordan Hoffmann, Trevor Cai, Eliza Rutherford, Katie Millican, George Bm Van Den Driessche, Jean-Baptiste Lespiau, Bogdan Damoc, Aidan Clark, et al. 2022. Improving language models by retrieving from trillions of tokens. In *International Conference on Machine Learning*, pages 2206–2240. PMLR.

Andrei Broder. 2002. A taxonomy of web search. In *ACM Sigir forum*, volume 36, pages 3–10. ACM New York, NY, USA.

Andrei Z. Broder, Peter Ciccolo, Marcus Fontoura, Evgeniy Gabrilovich, Vanja Josifovski, and Lance Riedel. 2008. Search advertising using web relevance feedback. In *Proceedings of the 17th ACM Conference on Information and Knowledge Management*, CIKM '08, page 1013–1022, New York, NY, USA. Association for Computing Machinery.

Andrei Z Broder, Marcus Fontoura, Evgeniy Gabrilovich, Amruta Joshi, Vanja Josifovski, and Tong Zhang. 2007. Robust classification of rare queries using web knowledge. In *Proceedings of the 30th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 231–238.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Aakanksha Chowdhery, Sharan Narang, Jacob Devlin, Maarten Bosma, Gaurav Mishra, Adam Roberts, Paul Barham, Hyung Won Chung, Charles Sutton, Sebastian Gehrmann, et al. 2022. Palm: Scaling language modeling with pathways. *arXiv preprint arXiv:2204.02311*.

Nick Craswell, Daniel Campos, Bhaskar Mitra, Emine Yilmaz, and Bodo Billerbeck. 2020. Orcas: 20 million clicked query-document pairs for analyzing search. In *Proceedings of the 29th ACM International Conference on Information Knowledge Management*, CIKM '20, page 2983–2989, New York, NY, USA. Association for Computing Machinery.

Fernando Diaz and Donald Metzler. 2006. Improving the estimation of relevance models using large external corpora. In *Proceedings of the 29th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 154–161.

Alejandro Figueroa. 2015. Exploring effective features for recognizing the user intent behind web queries. *Computers in Industry*, 68:162–169.

Jianping Gou, Baosheng Yu, Stephen J Maybank, and Dacheng Tao. 2021. Knowledge distillation: A survey. *International Journal of Computer Vision*, 129(6):1789–1819.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Mingwei Chang. 2020. Retrieval augmented language model pre-training. In *International Conference on Machine Learning*, pages 3929–3938. PMLR.

Shuguang Han, Xuanhui Wang, Mike Bendersky, and Marc Najork. 2020. Learning-to-rank with bert in tf-ranking. *arXiv preprint arXiv:2004.08476*.

Geoffrey Hinton, Oriol Vinyals, Jeff Dean, et al. 2015. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2(7).

Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2022a. Fid-light: Efficient and effective retrieval-augmented text generation. *arXiv preprint arXiv:2209.14290*.

Sebastian Hofstätter, Jiecao Chen, Karthik Raman, and Hamed Zamani. 2022b. Multi-task retrieval-augmented text generation with relevance sampling. *arXiv preprint arXiv:2207.03030*.

Gautier Izacard and Edouard Grave. 2020. Leveraging passage retrieval with generative models for open domain question answering. *arXiv preprint arXiv:2007.01282*.

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

Bernard J Jansen, Danielle L Booth, and Amanda Spink. 2008. Determining the informational, navigational, and transactional intent of web queries. *Information Processing & Management*, 44(3):1251–1266.

In-Ho Kang and GilChang Kim. 2003. Query type classification for web document retrieval. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 64–71.

Ashish Kathuria, Bernard J Jansen, Carolyn Hafernik, and Amanda Spink. 2010. Classifying the user intent of web queries using k-means clustering. *Internet Research*.

Dirk Lewandowski, Jessica Drechsler, and Sonja Von Mach. 2012. Deriving query intents from web search engine queries. *Journal of the American Society for Information Science and Technology*, 63(9):1773–1788.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in Neural Information Processing Systems*, 33:9459–9474.

Seyed Iman Mirzadeh, Mehrdad Farajtabar, Ang Li, Nir Levine, Akihiro Matsukawa, and Hassan Ghasemzadeh. 2020. Improved knowledge distillation via teacher assistant. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 5191–5198.

Alaa Mohasseb, Mohamed Bader-El-Den, and Mihaela Cocea. 2019. A customised grammar framework for query classification. *Expert Systems with Applications*, 135:164–180.

Rodrigo Nogueira and Jimmy Lin. 2019. From doc2query to doctttttquery. *Online preprint*, 6.

Rodrigo Nogueira, Wei Yang, Jimmy Lin, and Kyunghyun Cho. 2019. Document expansion by query prediction. *arXiv preprint arXiv:1904.08375*.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Xuanhui Wang. 2020. *Query Segmentation and Tagging*, pages 43–67. Springer International Publishing, Cham.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, Yanyan Lan, and Xueqi Cheng. 2020. Query understanding via intent description generation. In *Proceedings of the 29th ACM International Conference on Information amp; Knowledge Management*, CIKM '20, page 1823–1832, New York, NY, USA. Association for Computing Machinery.