Towards Effective Long-Form QA with Evidence Augmentation

Mengxia Yu[1]*, Sara Rosenthal[2], Mihaela Bornea[2], Avirup Sil[2]
[1]University of Notre Dame, [2]IBM Research AI
myu2@nd.edu, {sjrosenthal, mabornea, avi}@us.ibm.com

Abstract

In this study, we focus on the challenge of improving Long-form Question Answering (LFQA) by extracting and effectively utilizing knowledge from a large set of retrieved passages. We first demonstrate the importance of accurate evidence retrieval for LFQA, showing that optimal extracted knowledge from passages significantly benefits the generation. We also show that the choice of generative models impacts the system's ability to leverage the evidence and produce answers that are grounded in the retrieved passages. We propose a Mixture of Experts (MoE) model as an alternative to the Fusion in Decoder (FiD) used in state-of-the-art LFQA systems and we compare these two models in our experiments.

1 Introduction

Long-form question answering (LFQA) is a generative QA task that produces informative and comprehensive answers, often requiring models to leverage external knowledge sources. Retrieving supportive passages from large text corpora, e.g., Wikipedia, is a prevalent approach to provide external knowledge for the generation model. However, the retrieved passages often suffer from noise and excessive length that poses challenges for the model.

LFQA has largely been explored using ELI5 (Fan et al., 2019), a community QA dataset, where the answers are provided by domain experts. Finding supporting evidence for these questions is often challenging because the relevant information is usually fragmented across multiple documents. While several models (Fan et al., 2019; Su et al., 2022; Krishna et al., 2021) have been proposed for LFQA, Krishna et al. (2021) reveals that state-of-the-art models generate answers that are not grounded in the retrieved documents. Our study examines the information contained in the retrieved documents and uncovers that the top 3 passages contain 13.7% of the correct answer words, while the top 20 passages contain 38.1%. An example is shown in Table 1. This suggests that retrieving more passages can yield more useful information. However, processing more retrieved passages brings challenges for the generation model, making it difficult to discern evidence from noise.

Question	What are dentists actually doing when they scrape at your teeth with those metal picks?
Sample Answer	Most of the time they're looking at how much plaque is on or around your teeth. In other instances they're determining texture of the top layer of enamel. Both of these factor into the health of your teeth and help with diagnosing any problems you may have.
Doc 1	Dental extraction
Doc 2	Denticle (tooth feature)
Doc 3	**Calculus (dental)**
Doc 4	**Teeth cleaning**
Doc 5	Dentures
Doc 6	Dental surgery
Doc 7	**Tooth polishing**
Doc 8	Dental braces
...	...

Table 1: Example of an ELI5 question, one of its answers and the titles of the top retrieved documents. Relevant documents are in **bold**. Useful documents also appear at lower ranks. Reducing the impact of irrelevant information is important for all LFQA systems.

In this work we ask the following research questions: 1) Is it possible to obtain the appropriate knowledge from a large set of passages to improve the generation model? 2) How can the appropriate knowledge be effectively utilized in a model? Specifically, we explore the Fusion-in-Decoder (FiD) model (Izacard and Grave, 2021) and introduce a Mixture-of-Experts (MoE) model for the LFQA setting.

To verify whether it is possible to use the ideal relevant information successfully, we design an optimal setting, which we call the oracle evidence. We compare the performance of the FiD model

[*] This work was conducted while the first author was doing internship at IBM Research AI.

Proceedings of the Third Workshop on Natural Language Generation, Evaluation, and Metrics (GEM), pages 155–164
December 6, 2023 ©2023 Association for Computational Linguistics

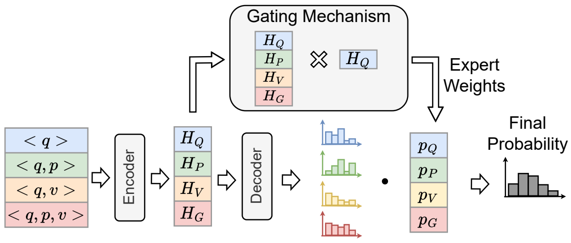

Figure 1: The Mixture of Experts (MoE) Framework.

with oracle evidence against with no evidence. We show that using the optimal evidence in FiD outperforms the baseline on two datasets, ELI5 and ASQA. These results highlight the critical role of obtaining appropriate knowledge from a large set of passages to enhance the effectiveness of question-answering systems. Our findings are relevant to prior SOTA work (Su et al., 2022; Liu et al.; Narayan et al., 2022) as applying better evidence can improve those approaches.

We also explore model architectures that can leverage the optimal evidence the most effectively. As an alternative to FiD, we propose an MoE model for the LFQA task. Our intuition is that MoE allows an advanced learning scheme capable of identifying the importance of different input sources, balancing between the retrieved passages and the extracted evidence. Our findings show that MoE captures the oracle evidence more effectively and yields better results than FiD on the ASQA dataset.

2 Method

Given a question q, we retrieve a collection of passages $P = p_i \dots p_m$ from external corpora. We consider the optimal situation where only the most relevant information for the correct answers is extracted from the large set of retrieved passages P. We refer to this as the oracle. We use the oracle evidence in our models during training and inference.

2.1 Oracle Evidence Extraction

The retrieved passages often contain relevant information at different granularity. Thus, we employ three types of oracle evidence: (1) word-based, consisting of a set of words; (2) triple-based, consisting of a set of triples that represent structured information in the form of <$Subject, Relation, Object$>; and (3) sentence-based, consisting of a set of sentences. During training we create an oracle for each gold answer, and during inference we build the oracle against the gold answer that had the largest overlap of passage words. Examples on both datasets are provided in Table 6 and 7 in the appendix.

The word-based oracle (WO) consists of a set of overlapping words between the retrieved passages and the gold answer[1]. It provides the necessary words for composing the answer, however it lacks the semantic information regarding how the words are related.

The triple-based oracle (TO) uses triples instead of words to compute the overlap between the retrieved passages and gold answers. We leverage OpenIE (Angeli et al., 2015) to extract triples on each of the retrieved passages. OpenIE produces a large number of triples for a passage, with redundant information. We apply a filtering process: keep the triples that contain oracle words in either the subject, relation, or object; if two triples contain the same oracle words, keep the one that appears first in the order of ranked passages. Finally, the triples are sorted based on the number of oracle words they contain. Triples are converted to statements: the triple <*two minute drill, refers with, little time remaining*> becomes *two minute drill refers with little time remaining*. The TO oracle has more semantic information, including relations between entities, but it's quality might be limited to the OpenIE accuracy.

The sentence-based oracle (SO) is at the sentence level. Similar to TO, we filter the sentences in the retrieved passages based on their overlap with the oracle words. This oracle has the most semantic information but significantly more noise.

2.2 Generation Model Architecture

We explore two model enhancements of the BART (Lewis et al., 2020) model architecture: Fusion-in-Decoder (FiD) (Izacard and Grave, 2021) and a novel Mixture of Experts approach inspired by prior work (Yu et al., 2022; Dai et al., 2022).

We implement the FiD model to enhance the model's capability of encoding multiple passages and evidence. FiD relies on the cross-attention mechanism to leverage the retrieved passages and the evidence.

2.3 Mixture-of-Experts

In this section we introduce a novel Mixture-of-Experts (MoE) model for LFQA aiming to handle

[1] In all oracles we exclude stop words and ignore case for matching.

and integrate diverse types of input data via the expert gating mechanism.

The model, as shown in Figure 1, consists of individual experts, designed to handle specific type of input, and a gating mechanism that selects the appropriate expert or a combination of experts.

We design four experts to focus on different parts of the input representations: 1) A question expert <Q> that only takes the question q as input. 2) An evidence expert <Q, V> which represents the question q and evidence, v. 3) A passage expert <Q, P> which represents q and the first m passages (e.g. 3 passages), and 4) a global expert <Q, V, P> which takes all input into consideration.

With our experts, q, P, and v can support the final prediction through joint interactions or separately. Each expert will encode its input separately:

$$\begin{aligned} H_Q &= Encoder(q) \\ H_P &= Encoder(q + p_1 + .. + p_N) \\ H_V &= Encoder(q + v) \\ H_G &= Encoder(q + v + p_1 + ... + p_N) \end{aligned}$$

We keep the encoder blocks shared by all the experts to capture the general features that are shared (e.g., the low-level text features).

The gating module computes the affinity scores for each expert using the cosine similarity between the hidden representations. We assign the affinity scores for the evidence, passage and global experts using their representations: $a_k = cos(H_Q, H_k)$, where $H_k \in \{H_P, H_V, H_G\}; k \in \{P, V, G\}$. The score for the question expert is the average score of the other three experts: $a_Q = \frac{1}{K}\sum_{i=1}^{K} a_k$, where $k \in \{P, V, G\}$. We assign expert weights by applying the $softmax$ function over the affinity scores: $s_i = \frac{exp(a_i)}{\sum_{k \in Q,P,V,G} exp(a_k)}$. For the final probabilities for next token sampling, the generation probabilities of all the experts are integrated: $Pr(w) = \sum_{k \in \{Q,P,V,G\}} s_k \cdot Pr_k(w)$.

3 Experiments

3.1 Datasets

ELI5 ELI5 (Fan et al., 2019) is a long form question answering dataset from the Reddit discussion forum **E**xplain **L**ike **I**'m **5**[2] where people ask for simple explanations to questions and get responses from other users. The responses tend to be long and free form. We used the KILT-ELI5 (Petroni et al., 2020) version of the task.

ASQA ASQA (Stelmakh et al., 2022) is an LFQA dataset of **A**nswer **S**ummaries for **Q**uestions which are **A**mbiguous. It was built using AmbigQA. The long-form answers are created by annotators using passages from Wikipedia that each contain different yet relevant information.

3.2 Experimental Setup

For retrieval we use DPR (Karpukhin et al., 2020) trained on the Natural Questions (Kwiatkowski et al., 2019) dataset and we index KILT Wikipedia (Petroni et al., 2020). For generation, we adopt BART-large as our base model and implement FiD and MoE models based on it. We used the question, the passages and the evidence as model input. We take the top 3 retrieved passages following prior work (Su et al., 2022). We also experimented with a larger number of passages, see Appendix Table 5, and we did not notice significant gains. For evidence, we consider WO, TO, and SO, as described in §2.1. With FiD, the evidence is given as input to the model as an additional "passage". All results are reported using Rouge-L (see Appendix C)

4 Experimental Results

Passages	Evidence	Model	ASQA	ELI5
		BART	35.0	29.9
✓		FiD	44.3	30.0
	WO	FiD	50.0	36.6
✓	WO	FiD	52.7	**36.7**
✓	WO	MoE	**55.7**	36.6
	TO	FiD	47.7	**33.9**
✓	TO	FiD	49.7	33.8
✓	TO	MoE	**50.4**	33.6
	SO	FiD	46.6	**30.7**
✓	SO	FiD	46.9	**30.7**
✓	SO	MoE	**47.9**	30.3

Table 2: Results (ROUGE-L) on ASQA and ELI5 with various input settings and different models.

4.1 Main Results

The main results are shown in Table 2. Our results show that FiD with evidence consistently outperforms BART and FiD without evidence on both datasets. The evidence brings complimentary knowledge to the passages. We do find that for

[2] https://www.reddit.com/r/explainlikeimfive/

Question Who plays patrick in 10 things i hate about you?
Gold In the 1999 American romantic comedy film 10 things i hate about you, film character Patrick is portrayed by Australian actor Heath Ledger. In the 2009 tv series by the same name, Patrick is played by American actor Ethan Peck.
FiD + WO (R-L: 48.6) **10 Things** I Hate About You is a **2009** American **romantic comedy film** directed by **Heath Ledger**. The **character Patrick** was **played** by **actor Heath Ledgers** in the **1999 film** and the **2009 series**.
FiD + TO (R-L: 44.4) **10 Things** I Hate About You is a **romantic comedy** film that was released in **1999**. **Heath Ledger plays Patrick** Verona, the lead **character** in the **film**. The **film** was later adapted into a television series that aired on February 14, **2009**. Larry Miller plays **Patrick** in the television **series**.
FiD + SO (R-L: 28.8) **10 Things** I Hate About You is a **1999** American **romantic comedy film** directed by Michael Bay and written by David Levithan. **Heath Ledger plays Patrick** Verona, the "bad boy" who is hired to date Kat and falls for her in the process. Ten years later, the **film** was adapted into a television **series** of the same title, which ran for twenty episodes and featured Larry Miller reprising his role as Walter Stratford from the **film**.

Table 3: Generated answers by FiD for all oracles. Gold answer words are in **bold**.

ELI5 the evidence alone is sufficient and including the first 3 passages does not provide an improvement. The performance gains achieved in the oracle setting underline the importance of developing effective methods for evidence extraction.

Results show that the MoE model consistently outperforms FiD on the ASQA dataset, given the same source inputs. These performance gains highlight the potential of specialized sub-models and ensemble learning approaches. However on the ELI5 dataset, the MoE shows no significant change to FiD. This might be attributed to the fact that the ELI5 passages do not provide additional gains (Krishna et al., 2021), indicating that the evidence expert alone is always sufficient.

4.2 Oracle Analysis

In our experiments we observe improvements for all oracles over the baselines with WO > TO > SO. This is because the level of noise increases for the oracles from left to right. However, the semantic relationship between words decreases from left to right. These findings remain consistent for both models and datasets.

Table 3 provides a generated answer for a question in ASQA. The generations of all three oracles provide reasonable answers that cover similar oracle words but only correctly answer the movie actor. All three answers hallucinate with WO hallucinating the most. On the other hand, the length of the answer increases as the oracle contains more semantic information. For instance, SO contains more information, that while in some cases is not always relevant to answering the question. The Rouge-L score does not capture the correctness in all the answers (Krishna et al., 2021).

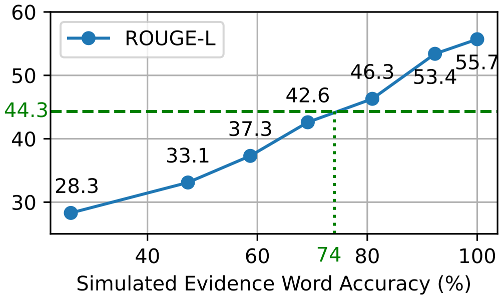

Figure 2: Performance (ROUGE-L) on ASQA as a function of the simulated evidence samples.

4.3 Impact of Evidence Accuracy

We conduct an experiment to simulate the impact of the evidence accuracy on generation. This experiment suggests what the LFQA performance would be in the non-oracle setting. We sample multiple sets of words as word-based evidence, each comprising of a combination of gold evidence words (answer words) and noise words (non-answer words in the passages). The quality of the samples compared to WO is measured using the F1 score. We use the MoE WO model to do inference on the evidence samples. Fig 2 shows that the performance increases with the quality of the evidence. The results indicate that when the simulated evidence prediction achieves an F1 of 74, the generation improves over the non evidence setting. Evidence prediction in a non-oracle setting is subject to future work.

5 Conclusion

Our study investigates the impact of including optimal evidence from external knowledge in LFQA. By employing three forms of evidence in oracle scenarios, we demonstrate that optimal evidence extracted from retrieved passages significantly improves the performance of LFQA systems. In addition, we propose an MoE model for incorporating passages and extracted evidence. Experimental results showed that our MoE design uses the evidence more efficiently on the ASQA dataset. We believe that these findings are encouraging for further exploration in the intelligent use of external evidence for LFQA to improve generation in non-oracle settings. These findings can be applied to improve all state-of-the-art approaches including LLMs.

Limitations

Our work shows that relevant information can be found from external knowledge making generated answers more grounded on retrieved passages. The main limitation of our work is that we have explored an oracle setting which promotes the usefulness of evidence extraction. In future work we will be exploring evidence extraction methods in a non-oracle setting, which will make our approach effective in many real-word applications and is compatible with existing SOTA methods. We are using the public ELI5 and ASQA datasets and we initialized our models from the BART pre-trained model. Any limitations associated with these resources are likely carried over in our work.

References

Gabor Angeli, Melvin Jose Johnson Premkumar, and Christopher D. Manning. 2015. Leveraging linguistic structure for open domain information extraction. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 344–354, Beijing, China. Association for Computational Linguistics.

Damai Dai, Wenbin Jiang, Jiyuan Zhang, Weihua Peng, Yajuan Lyu, Zhifang Sui, Baobao Chang, and Yong Zhu. 2022. Mixture of experts for biomedical question answering. *arXiv preprint arXiv:2204.07469*.

Angela Fan, Yacine Jernite, Ethan Perez, David Grangier, Jason Weston, and Michael Auli. 2019. ELI5: Long form question answering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3558–3567, Florence, Italy. Association for Computational Linguistics.

Gautier Izacard and Edouard Grave. 2021. Leveraging passage retrieval with generative models for open domain question answering. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 874–880, Online. Association for Computational Linguistics.

Vladimir Karpukhin, Barlas Oğuz, Sewon Min, Patrick Lewis, Ledell Yu Wu, Sergey Edunov, Danqi Chen, and Wen tau Yih. 2020. Dense passage retrieval for open-domain question answering. In *Conference on Empirical Methods in Natural Language Processing*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4940–4957, Online. Association for Computational Linguistics.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur 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. *Transactions of the Association for Computational Linguistics*, 7:452–466.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Veselin Stoyanov, and Luke Zettlemoyer. 2020. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7871–7880, Online. Association for Computational Linguistics.

Ruibo Liu, Guoqing Zheng, Shashank Gupta, Radhika Gaonkar, Chongyang Gao, Soroush Vosoughi, Milad Shokouhi, and Ahmed Hassan Awadallah. Knowledge infused decoding. In *International Conference on Learning Representations*.

Shashi Narayan, Joshua Maynez, Reinald Kim Amplayo, Kuzman Ganchev, Annie Louis, Fantine Huot, Dipanjan Das, and Mirella Lapata. 2022. Conditional generation with a question-answering blueprint. *arXiv preprint arXiv:2207.00397*.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick S. H. Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2020. KILT: a benchmark for knowledge intensive language tasks. *CoRR*, abs/2009.02252.

Ivan Stelmakh, Yi Luan, Bhuwan Dhingra, and Ming-Wei Chang. 2022. ASQA: Factoid questions meet long-form answers. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 8273–8288, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Dan Su, Xiaoguang Li, Jindi Zhang, Lifeng Shang, Xin Jiang, Qun Liu, and Pascale Fung. 2022. Read before generate! faithful long form question answering with machine reading. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 744–756, Dublin, Ireland. Association for Computational Linguistics.

Wenhao Yu, Chenguang Zhu, Lianhui Qin, Zhihan Zhang, Tong Zhao, and Meng Jiang. 2022. Diversifying content generation for commonsense reasoning with mixture of knowledge graph experts. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 1896–1906, Dublin, Ireland. Association for Computational Linguistics.

A Hyperparameters

The FiD and MoE models are initialized with the pre-trained `facebook/bart-large` model, which contains 400M parameters. We finetuned the models using the Adam optimizer. We conduct hyperparameter tuning with the following range: learning rates $\{3e-5, 6e-5, 1e-4\}$, batch sizes $\{16, 32, 64\}$, beam size $\{1, 4\}$. For BART, we set the max input length as 1024. For FiD and MoE, we set the input length (for each passage) as 256. We set the output length as 256 for ELI5 and 128 for ASQA. We found an optimal learning rate of 6e-5 for ELI5 and 3e-5 for ASQA.

B KILT vs Google Rouge

We report all results using the Google implementation[3] which is available through the HuggingFace library, since it's more accurate for reflecting the quality of the generation.

The following differences highlight the inconsistencies between the Python Rouge package used by KILT in contrast to the Google implementation. These cause significant variation in results.

- The Google version is case insensitive while the rouge package is not (e.g. KILT will give zero credit for the correct answer "Graphical User Interface" where capitalizing the words is appropriate)

- The rouge package does not do much tokenization cleanup, so you do not get credit for a word when there is a comma at the end of it (e.g. "user,")

- The rouge package calculates the size of m and n (the denominators for R_{LCS} and F_{LCS} using the length of the *set* of words and only gives credit per word once. On the other hand, Google calculates the size of m and n using the *list* of all words and gives credit for words based on how many times they appear in the gold answer. (e.g. if the word "user" appears 3 times you can get credit for having it 3 times in the generated answer.)

Model	R-L (KILT)	R-L (HF)
BART (Su et al., 2022)	22.69	-
FiD (Su et al., 2022)	25.70	-
RBG (Su et al., 2022))	26.46	-
BART (reproduced)	26.18	29.8
FiD (reproduced)	**26.48**	30.0

Table 4: Baselines results on ELI5 validation set reported in previous paper and results reproduced by us. Previous papers reported the KILT ROUGE-L.

C Results of Prior Work

In this paper we do not focus on comparing to prior art as our work is complementary to SOTA approaches. For reference we include a comparison with the related work and we show both the KILT rouge (the python package) as well as the rouge score with the Google implementation. These results are in Table 4.

Table 5 shows the generation performance when increasing the number of passages in the FiD input. Based on these results we decided to use 3 input passages in our experimental study, which is also consistent with prior work. As more passages are added, there are only negligible improvements in generation.

# Passages	ASQA	ELI5
0	35.0	29.9
3	44.1	30.0
5	45.0	30.1
10	45.2	30.2

Table 5: FiD results with different number of retrieved passages. Scaling up to 5 or 10 only brings a minor performance gain. With the oracle evidence the performance improves to 50.0 for ASQA and 36.6 for ELI5.

D Examples

In Tables 6-9 we show an example of the question and gold answers as well as input features and generated output for ELI5 and ASQA respectively. These show the difference of how the input features and generated output look depending on the oracle.

[3] `https://github.com/google-research/google-research/tree/master/rouge`

Question
In Trading Places (1983, Akroyd/Murphy) how does the scheme at the end of the movie work? Why would buying a lot of OJ at a high price ruin the Duke Brothers?
Best Gold Answer
If I remember correctly, they knew that the price of orange juice was going to fall. Normally this wouldn't matter, because you are supposed to buy and hold stocks, but they were buying what's called 'futures'. In a nutshell, they were buying contracts that afford them the legal right to purchase units of OJ at a specific price. Since they knew the price of OJ would fall (remember the dude with the locked briefcase?) they were buying option contracts to purchase OJ at a higher price. Anyone with half a brain would sell them these and of course that's what happened. For in depth knowledge, look up how futures trading works.
WO
"future", "1983", "Duke", "know", matter", "orange", "purchase", "Murphy", "place", work", "high", "movie", "contract", "scheme", "call", "juice", "since", "lot", "would", "go", "sell", "stock", "buy", "higher", "trading", "option", price", "hold", "end"
TO (count >= 2)
1. Dukes commit holdings to frozen orange juice futures contracts 2. prices go down just as they had expected 3. price is why higher for example 4. Canadians would buy their cars 5. buy hold antithesis of is concept 6. if geologist knows is high likelihood 7. bubble involves rising prices for example stock 8. O'Hagan used information by buying call options resulting
SO (count >= 2)
1. On the commodities trading floor, the Dukes commit all their holdings to buying frozen concentrated orange - juice futures contracts; other traders follow their lead, inflating the price. 2. A bubble involves ever - rising prices in an open market (for example stock, housing, cryptocurrency, or tulip bulbs) where prices rise because buyers bid more, and buyers bid more because prices are rising.
Top 3 passages
Trading Places Trading Places is a 1983 American comedy film directed by John Landis and starring Dan Aykroyd and Eddie Murphy. It tells the story of an upper-class commodities broker and a homeless street hustler whose lives cross paths when they are unknowingly made part of an elaborate bet...
During the firm's Christmas party, Winthorpe is caught planting drugs in Valentine's desk in an attempt to frame him, and he brandishes a gun to escape. Later, the Dukes discuss their experiment and settle their wager for one dollar, before plotting to return Valentine to the streets. Valentine overhears the conversation, and seeks out Winthorpe, who attempts suicide by overdosing on pills...
Winthorpe is publicly framed as a thief, drug dealer and philanderer by Clarence Beeks, a man on the Dukes' payroll. Winthorpe is fired from Duke & Duke, his bank accounts are frozen, he is denied entry to his Duke-owned home, and he quickly finds himself vilified by Penelope and his former friends...

Table 6: An example from the ELI5 dev set showing the best gold answer (based on overlap with passages) and the resulting features for the three oracles based on overlap with the best answerq. We also provide the top 3 passages (shortened) for comparison - in this case they are not relevant to the question.

Question What kind of car in to catch a thief?
Best Gold Answer
The car driven by Grace Kelly in T̈o Catch a Thiefẅas a metallic blue 1953 Sunbeam Alpine Mk I. The Series I used a engine and was styled by the Loewy Studios for the Rootes Group.
WO
"drive", "Series", "Sunbeam", "use", "blue", "Alpine", "car", "engine", "style"
TO (count >= 2)
1. Sunbeam Alpine was chosen car In novel 2. sapphire blue Alpine Catch Thief 3. car was shipped to USA for use 4. cars supercharged 1.6 litre engine coupled
SO (count >= 2)
1. The Alpine name was resurrected in 1976 by Chrysler (by then the owner of Rootes) , on a totally unrelated vehicle : the UK - market version of the Simca 1307 , a French - built family hatchback . 2. The car was initially badged as the Chrysler Alpine , and then finally as the Talbot Alpine following Chrysler Europe 's takeover by Peugeot in 1978 . 3. According to JLR Special Vehicle Operations chief John Edwards , the cars are c̈onstructed around a spaceframe built to World Rally Championship spec änd powered by a turbocharged and supercharged 1.6 - litre engine coupled with two electric motors . 4. However , a sapphire blue Alpine featured prominently in the 1955 Alfred Hitchcock film T̈o Catch a Thief s̈tarring Cary Grant and Grace Kelly .
Top 3 passages
1. The Alpine name was resurrected in 1976 by Chrysler (by then the owner of Rootes), on a totally unrelated vehicle: the UK-market version of the Simca 1307, a French-built family hatchback. The car was initially badged as the Chrysler Alpine, and then finally as the Talbot Alpine following Chrysler Europe's takeover by Peugeot in 1978. The name survived until 1984, although the design survived (with different names) until 1986. Section: Notable film and television appearances. BULLET: - A metallic blue 1953 Sunbeam Alpine Mk I is driven by Grace Kelly in T̈o Catch a Thief(̈1955) with Cary Grant.
2. List of James Bond vehicles Throughout the J̈ames Bonds̈eries of films and novels, Q Branch has given Bond a variety of vehicles with which to battle his enemies. Among the most noteworthy gadgets, Bond has been equipped with various vehicles that have numerous modifications to include elaborate weapons and anti-pursuit systems, alternative transportation modes, and various other functions. One car in particular that has been linked to Mr. Bond's collection is the Aston Martin DB5. This is a list of the most noteworthy vehicles seen in James Bond, used by either Bond himself, his allies or his enemies.
3. The first car shown in the film with the title of Pursuit Special is a 1972 HQ Holden Monaro V8 coupe stolen by Nightrider (played by Vince Gil), an escaped cop killer, who dies in an accident that destroys the vehicle. The more famous Pursuit Special is a heavily modified Ford Falcon XB GT, built on a vehicle originally assembled stock at the Ford plant in November 1973. Max Rockatansky (Mel Gibson) is offered the black Pursuit Special, as an incentive to stay on the force as their top pursuit man after he reveals his desire to resign.

Table 7: An example from the ASQA dev set.

Question
In Trading Places (1983, Akroyd/Murphy) how does the scheme at the end of the movie work? Why would buying a lot of OJ at a high price ruin the Duke Brothers?
FiD + p Baseline HF: 0.30 KiLT: 0.26
The OJ was a commodity. It was a good investment. The Duke brothers were rich and had a lot of money. They were also rich and knew that the price of OJ would go up. So they bought a lot. Then they sold it to the government for a lot more than they paid for it. The government then decided that they were going to sell it for a higher price than they bought it for. This was a big deal because they were losing money. So they bought the OJ at a high price. But then the government decided that the government was going to stop selling OJ. And the government said that they would sell it at a lower price. So the government sold it for less than they were paid for. The price of the Oj went up. So the government had to sell OJ for less then they paid. That was a huge deal. Because they were making a lot less money than they had paid for OJ, they had to pay a lot higher than they thought they would. So now they had a huge loss. Which is why they were so angry.
FiD + p + WO HF: 0.39 KiLT: 0.34
The orange juice is a stock called a futures contract. You purchase a contract to buy a certain amount of orange juice at a certain price. If the price of orange orange juice goes up, you can sell the contract for more than you paid for it. Since the price is higher than the contract price, you don't have to pay for it, no matter what. The Duke Brothers knew that the price was going to go up, so they purchased a lot of the contract. They then held onto the contract, and sold it at a higher price than they paid for the contract (the price of the orange juice).
MoE + p + WO HF: 0.36 KiLT: 0.36
OJ is a stock in orange juice. The Duke Brothers own a lot of orange juice futures contracts. These contracts are called čall options.. You purchase a contract to buy a certain amount of orange Juice at a certain price. If the price of orange juices goes up, you can sell the contract for more than you bought it for. Since the Duke Brothers hold a lot more orange juice than they purchased, they can sell it for more then they bought it at. The Duke Brothers know that the price will go up, so they know that they can't sell the contracts for more. So they sell the option contracts for less than they bought them for. So the Duke brothers sell the call options for less then they purchased them for, and the price goes up. They then buy the contract at a higher price, and sell it at a lower price. It doesn't matter if the price is higher or lower, they are still losing money.
MoE + p + TO HF: 0.34 KiLT: 0.31
The Duke brothers bought a stock called an orange juice futures contract. They knew that orange juice was going to be a very high demand commodity, so they bought a bunch of orange juice at a high price. The stock was going up, so the Duke brothers made a bet that the stock would go up. If the stock went up, they would sell the orange juice. Since they knew the stock was likely to go up, the Duke Brothers bought a lot of orange juices at a higher price than they expected. When the stock fell, they sold the orange juices to make up for the loss. The Duke Brothers knew that the price of orange Juice was going down, so when they bought the orange Juice futures contract, they knew that they would have to sell the stock to make back the difference. So they held onto the stock and bought the stock at a lower price than expected.
MoE + p + SO HF: 0.34 KiLT: 0.28
The Duke brothers are a bunch of rich guys who are very good at trading. They have a lot of money and they want to make money. So they hire a guy named Eddie Murphy to buy a bunch more stocks and futures. Eddie Murphy is a very smart guy and he knows that if he buys a lot, he can make a lot more money than he would if he sold it. He knows that the price of OJ is going to go up, so he buys it at a high price. The price goes up, and Eddie Murphy sells it to the Duke brothers. The Duke Brothers then use that money to buy more stocks, futures, and other things. This is called a šhorting.. Basically, they buy a lot and sell them at a low price, hoping that the market will go up. If the market goes up they can buy more stock and futures and sell it at higher prices. When the market is up, they can make money, but if the market stays up, then they lose money. So Eddie Murphy buys a bunch and sells it. The price goes down, and the Duke Brothers buy more shares and sell those shares and buy more

Table 8: An example from the ELI5 dev set showing the generated answers for the baseline and a sample of our oracle models.

Question
What kind of car in to catch a thief?
Best Gold Answer
The make of the car in the 1953 film To Catch a Thief was Rootes Group, and its model was the 1953 Sunbeam Alpine Mk I. The car driven by Grace Kelly, who stars in the film, was a metallic blue 195 3 Sunbeam Alpine Mk I. The Sunbeam Alpine is a two-seater sports drophead coupé that was produced by the Rootes Group, a British automobile manufacturer, from 1953 to 1955 and from 1959 to 1968.
The car used to catch a thief is a large SUV, which is a boxy, boxy SUV with a rear-wheel drive. The car is usually a mid-size SUV, but can be a full-size or full-sized SUV.
FiD + p Baseline HF: 0.26 KiLT: 0.14
Throughout the James Bond series of films and novels, Q Branch has given Bond a variety of vehicles with which to battle his enemies. Among the most noteworthy gadgets, Bond has been equipped with various vehicles that have numerous modifications to include elaborate weapons and anti-pursuit systems, alternative transportation modes, and various other functions. The first car shown in the film with the title of Pursuit Special is a 1972 Holden Monaro V8 coupe stolen by Nightrider, played by Vince Gil, an escaped cop killer, who dies in an accident that destroys the vehicle. The more famous Pursuit
FiD + p + WO HF:0.22 KiLT: 0.18
Throughout the J̈ames Bonds̈eries of films and novels, Q Branch has given Bond a variety of vehicles with which to battle his enemies. The first car shown in the film with the title of Pursuit Special is a 1972 Holden Monaro V8 coupe stolen by Nightrider, played by Vince Gil. The more famous Pursuit special is a heavily modified Ford Falcon XB GT, built on a vehicle originally assembled stock at the Ford plant in November 1973.
MoE + p + WO HF: 0.45 KiLT: 0.34
In To Catch a Thief, Grace Kelly drives a 1953 Sunbeam Alpine Mk I, a metallic blue Sunbeam Series III car with a V8 engine."
MoE + p + TO HF: 0.43 KiLT: 0.40
To Catch a Thief is a 1955 American sports car crime novel by American author John Updike. The car in the novel is a 1953 Sunbeam Alpine. The novel was written in the style of a sports car, but the car was actually a convertible.
MoE + p + SO HF: 0.36 KiLT: 0.29
In the 1955 Alfred Hitchcock film To Catch a Thief, a Sunbeam Alpine was the chosen car of assassinated Secret Service agent John Strangways. However, a sapphire blue Alpine featured prominently in the 1955 film To catch a Thief starring Cary Grant and Grace Kelly. The Alpine name was resurrected in 1976 by Chrysler, by then the owner of Rootes, on a totally unrelated vehicle: the UK-market version of the Simca 1307, a French-built family hatchback. The car was initially badged as the Chrysler Alpine, and then finally as the Talbot Alpine following Chrysler Europe's takeover by

Table 9: An example from the ASQA dev set showing the best gold answer (based on overlap with passages) and the resulting features for the three oracles based on overlap.