Modeling Exemplification in Long-form Question Answering via Retrieval

Anonymous ACL submission

Abstract

Exemplification is a process by which writers explain or clarify a concept by providing an example. While common in all forms of writing, exemplification is particularly useful in the task of long-form question answering (LFQA), where a complicated answer can be made more understandable through simple examples. In this paper, we provide the first computational study of exemplification in QA, performing a fine-grained annotation of different types of examples (e.g., hypotheticals, anecdotes) in three corpora. We show that not only do state-of-the-art LFQA models struggle to generate relevant examples, but also that standard evaluation metrics such as ROUGE are insufficient to judge exemplification quality. We propose to treat exemplification as a *retrieval* problem in which a partially-written answer is used to query a large set of human-written examples extracted from a corpus. Our approach allows a reliable ranking-type automatic metrics that correlates well with human evaluation. Human evaluation shows that examples retrieved from our retriever are more relevant than examples generated from state-of-the-art LFQA model.

1 Introduction

When an author introduces a complicated concept, they commonly follow it up with a concrete example to help clarify their intended meaning. This process, known as *exemplification*, occurs in diverse forms, including hypothetical examples, personal anecdotes, and analogies (Clouse, 2013). Exemplification is particularly common within the NLP task of *long-form question answering* (LFQA), where an author wants to communicate a concept unknown to the question asker. Consider the following QA pair from the `r/explainlikeimfive` subreddit (Fan et al., 2019):

> **Q:** How does the ground not cave in while being under the heavy weight of cities?
> **A:** It's all about what they're building on, and occasionally they get it wrong... For example, San Francisco's Millennium Tower, built on mud and sand, has already sunk 18 inches into the ground.

Here, the answerer uses a specific example to emphasize the importance of building on a solid foundation. In general, explaining via exemplification is a fundamental technique in these kinds of pedagogical scenarios (Hyland, 2007), and it warrants separate study due to its importance in LFQA and the challenges in evaluating it. However, existing work on building LFQA models (Fan et al., 2019) does not give special treatment to exemplification or any other discourse phenomena, choosing instead to evaluate model outputs against reference answers using metrics such as ROUGE (Lin, 2004) that are not meaningful for this task (Krishna et al., 2021). In the above QA pair, any other structurally-unstable building (e.g., the Leaning Tower of Pisa) could serve as a valid example, but an LFQA model would be unfairly penalized for generating one of these acceptable alternatives.

In this paper, we first conduct a detailed study of exemplification across three different domains: Wikipedia, web articles, and community answers to questions from the ELI5 LFQA dataset. We extract sentences and clauses associated with exemplification by matching on explicit markers such as *"for example"* and *"(e.g., ...)"* and annotate 300 examples from this dataset. Our analysis reveals significant variation in occurrence frequencies of different forms of exemplification (e.g., hypothetical vs. specific examples) across the three domains.

Next, we focus on improving the modeling and evaluation of the subtask of exemplification within LFQA. We propose to treat it as a *retrieval* problem rather than a generation problem: given a question and a prefix of a reference answer (in the above QA pair, all of the text before *"For example"*), a model must retrieve the ground-truth example that follows the prefix (the sentence about the Millenium Tower) from the set of all exemplifying sentences

and clauses in the dataset. We can use retrieval metrics such as recall@k to evaluate a model's ability to select the ground-truth example, which are far more informative than metrics such as ROUGE.

We demonstrate that pretraining our retriever on a large-scale dataset of exemplifying units extracted from the Books3 corpus (Gao et al., 2020) and then fine-tuning it on ELI5 examples results in substantial improvements on these ranking metrics. Finally, we crowdsource human evaluation comparing our retriever's outputs with those generated by the state-of-the-art ELI5 model of Krishna et al. (2021) and find that workers prefer the retriever's outputs far more frequently than those of the generation model. We hope that our work spurs more research into the evaluation and modeling of exemplification and other complex discourse phenomena present in LFQA.

2 Exploring Exemplification

In this section, we first describe our data extraction process, which we use to mine instances of exemplification from three datasets (and domains): ELI5 (Fan et al., 2019), Natural Questions (Kwiatkowski et al., 2019) and Books3 (Gao et al., 2020). This process involves matching on a set of "exemplification markers" and collecting both the text of the matching example (a sentence or clause) as well as the surrounding context on both sides. We then conduct a fine-grained human annotation study on the extracted data, breaking exemplification down into different types and exploring how they are used across the different domains.

2.1 Extracting a Dataset of Exemplification

Exemplification markers: Hyland (2007) annotated a diverse collection of articles from multiple disciplines with a variety of rhetorical practices[1] and found that more than 75% of "examples" are signalled parenthetically or lexically with the use of the three most frequent "exemplification markers": *"such as"*, *"for example"*, *"e.g."*. Empirically, we find that the *"such as"* marker is noisy at signalling exemplification and often leads to ambiguous cases where it is hard to automatically detect the example boundary. Hence, we take the other two most frequent exemplification markers, namely *"for example"* and *"e.g."*, and extract the parentheses-enclosed clauses and sentences that contain these exemplification markers as examples.

	ELI5	NQ	Books3
# training examples	65,157	1,209	2,848,171
# validation examples	1,185	52	712,043
avg. # context words	123.7	74.0	155.7
avg. # example words	23.3	33.1	27.1
avg. # right words tokens	135.3	54.8	107.5

Table 1: Statistics of extracted example-in-context data.

Examples from Diverse Domains Using these two exemplification markers, we extract a dataset of examples in context from two popular LFQA datasets that come from different domains, ELI5 (Fan et al., 2019, Reddit answers) and Natural Questions (Kwiatkowski et al., 2019, Wikipedia passages).[2] To study the exemplification phenomenon from a more diverse perspective, we also extract examples along with their surrounding context from the Books3 Corpus (Gao et al., 2020), a large-scale 100GB collection of books spanning a variety of topics and genres. Table 1 contains detailed statistics for the extracted example-in-context datasets.

2.2 Fine-grained Annotation Study

With the extracted dataset of examples, we conduct an in-depth analysis to understand different uses of exemplification in various domains. We (the authors) annotate a total of 300 examples extracted using exemplification markers from Natural Questions, ELI5 and Books3 as below. Fifty examples are annotated by two annotators (for purposes of computing agreement, reported in Section 2.3) and the rest are annotated by one annotator.

Given an extracted example and its left and right context, we first filter out around 7% of the extracted examples, either because they are extraction artifacts or because the marker is used for functions other than exemplification (e.g., referring to a figure or table). After this basic check, we annotate both structural information about the example (e.g., discourse units such as the anchor of the example) and semantic information about how it is used in the context. Table 2 contains statistics of the annotated subset.

[1] The annotation included texts from physics, biology, mechanical & electric engineering, philosophy, sociology, applied linguistics, and marketing.

[2] We consider questions with *only* a long answer span (i.e. paragraph answer) since they cannot be addressed solely by entity names or a boolean, and are suitable for studying LFQA.

Dataset	Valid	Extracted	% Valid
ELI5	87	93	94%
NQ	89	95	94%
Books3	85	94	90%
Total	261	282	93%

Table 2: Statistics of annotated examples.

Dataset	# samples	Anchor	Example
ELI5	87	1.1/16.6	1.3/29.2
NQ	89	1.0/15.4	1.4/25.1
Books3	85	1.1/17.0	1.2/25.1
Type			
Real	209	1.1/14.5	1.3/24.6
Hypothetical	52	1.2/23.4	1.2/33.7
Personal	13	1.1/18.1	2/49.6
Not-personal	248	1.1/16.3	1.2/25.2

Table 3: Length of the discourse units per dataset and type, presented as average # sentences / # words.

2.2.1 Discourse units

Exemplification is usually expressed through three discourse units (Meyer, 1992; Triki, 2021): the anchor (also known as "exemplified unit"), the exemplification marker, and the example text itself ("exemplifying unit"). We annotate the anchor (marked as **bold**) and example (marked as *italics*). Concretely, in example (1) below, the anchor is "euryhaline species", the exemplifying marker is "e.g.", and the exemplifying unit is "Flounder". As in the study of Triki (2021), we find that these units mainly come in two forms: (1) nominal groups that refer to entities, or (2) clauses that represent statements.

> (1) However , some fish show a tremendous ability to effectively osmoregulate across a broad range of salinities ; fish with this ability are known as **euryhaline species** , e.g., *Flounder*.

> (2) **Players earn you points, depending on their performance.** For example, *your QB might throw for 200 yards, earning 1 point per 10 yards, for 20 points.*

During our initial investigation, we also noticed examples (3) which are signalled *implicitly* (i.e. without exemplification markers). Identifying such examples automatically is beyond beyond the scope of our study and warrants future work.

> (3) **The biggest driver of disparity in tech jobs is cost of living.** *If it costs 2000 a month to live in Boston, and 200 a month to live in India, then salaries will reflect that.*

Table 3 shows that the length of the discourse units is roughly the same across the three datasets, which supports our experimental decision in Section 3 of using sentences as the base unit for example retrieval. We also find that all of the anchors we annotated occur *before* the examples, suggesting that using preceding context context to retrieve examples gives the model sufficient information.

2.2.2 Real vs. Hypothetical Examples

One notable categorization found during our investigation and also identified by Triki (2021) is whether the examples are real, specific scenarios, or hypothetically-constructed scenarios. We detail the definition of the two types here:

Real examples: These examples are either real entities (4) or specific scenarios (5) that are constructed as fact clauses.

> (4) CEOs lead a range of organizations , including public and private corporations, non-profit organizations and even **some government organizations** (e.g., *Crown corporations*).

> (5) **For a given pressure , different liquids boil at different temperatures**. For example, *water boils at 100° C (212° F) at sea level, but at 93.4° C (200.1° F) at 2,000 metres (6,600 ft) altitude.*

Hypothetical examples: In contrast, hypothetical examples are scenarios constructed by the author. According to Triki (2021), hypothetical examples often come with the use of conditional clauses *if* or signalled via *assume*. These examples are generally more complicated and are specifically constructed for the purpose of exemplification.

> (6) **The reasoning is that if you share your life with someone then anything you do during that time is made possible by their support.** For example, *if your wife is a stay at home wife and your a business man making lots of money, the reasoning is that you would not have the same amount of time to dedicate to work if you had to look after your own house and/or children.*

We observe a different distribution of the two types of examples in the three datasets (Figure 1, top). ELI5 contains more hypothetical examples (32%) than the other two datasets (16% for NQ, 12% for Books3), showing that hypothetical examples are commonly used to explain complicated concepts. We also note that hypothetical examples are generally longer than real ones (33.7 v.s. 24.6 words, as seen in Table 3), aligning with our observation that these examples are more complicated.

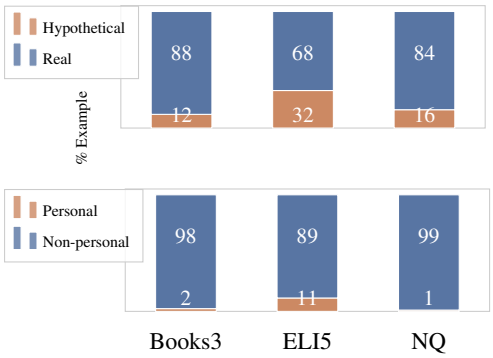


Figure 1: Distribution of different types of examples across three datasets.

2.2.3 Personal Information

Previous work found that exemplification is "rarely personalized" in academic texts (Ädel, 2006), referring to the uncommon use of first person pronouns or reference to the author. In contrast, we found a consistent presence of personal information in the examples we examined, in the form of either personal anecdotes (7) or an example situated in the author's own circumstance (e.g., *in my city...*). We thus also annotate whether the exemplifying units contain personal information or not.

> (7) ... **But they also give advice a doctor might have forgotten.** For example *about 6 months ago I went to the urgent care center for ear pain and was prescribed an ear drop antibiotic.* ...

We observe differences in the presence of personal information across the three datasets (Figure 1, bottom) – ELI5 answers, which are written by users from online community forum, contain a substantial portion of examples with personal information (12%), while such information is rarely present in the other two datasets. There is also a notable length difference between examples with and without personal information (49.6 v.s. 25.2 words), showing that more detailed description is provided for personal anecdotes. The observation that ELI5 contains many personal examples raises concerns that language models trained on such datasets will generate personal examples that cannot be verified or meaningfully interpreted.

2.3 Annotation agreement

We report agreement for the three annotation tasks performed on the 50 two-way annotated samples. For discourse unit annotation, we measure unigram overlap between the two annotations, and find high agreements of 0.81 for the anchor and 0.92 for the example. For annotation of real vs. hypothetical examples and the presence of personal information, we find a modest to high agreement with a Cohen's kappa of 0.48 for both.[3]

3 Retrieving Examples in LFQA

Our annotation and analysis of the extracted exemplification dataset reveals the diversity and complexity of this discourse phenomenon. In this section, we shift our focus towards building retrieval-based models to produce examples based on a given LFQA context. First, we define our example-retrieval task formally and introduce our evaluation metrics. Then, we describe our contrastive-learning based retriever model(EGRET) and baseline retrieval models, and report their performances.

3.1 Task Definition

Given a context (part of the answer to a given LFQA question) with a masked out exemplifying unit, model is asked to retrieve that masked unit from a retrieval corpus. We consider two settings for context: (1) concatenation of left and right contexts surrounding the exemplifying unit and (2) left context preceding the exemplifying unit only (as in Figure 2). Both left and right contexts are truncated to 256 tokens surrounding the exemplifying unit. We use all 66K exemplifying units extracted from the 272K QA instances in the training and development portion of ELI5 dataset (Petroni et al., 2021) as the retrieval corpus. To illustrate the input format, consider the below answer to the question "Why didn't anyone discover dinosaur fossils before the 1800s?":

> Prior to the 1800's, when people dug up fossils (and more frequently, subfossil bones from ice age animals, which are more common and easier to find) they tended to interpret them in light of their existing myths and legends. **[MASK]** Because fossils are almost always found as a jumble of bones rather than a neat skeleton and because they are incomplete... nobody looked at dinosaur skeletons and realized what the animals that made them actually looked like.

where the **[MASK]** token corresponds to

> For example, when a wooly rhino skull was dug up near Klagenfurt, it was thought to be the skull of a dragon.

The first quote block showing the masked answer will be used as a query to retrieve the exemplifying unit in the second quote block.

[3] We present examples with split annotation in Appendix A.

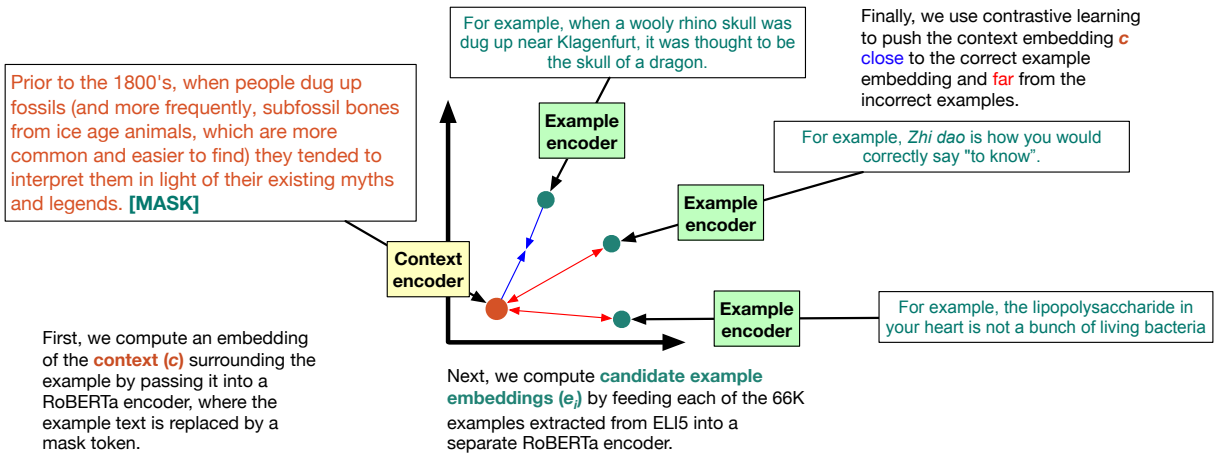


Figure 2: Our EGRET model uses dual RoBERTa encoders to embed (1) the context surrounding an exemplifying unit and (2) all 66K exemplifying units extracted from ELI5. A contrastive objective is then used to move the context embedding close to the embedding of the ground-truth exemplifying unit that occurs within that context, and far from all other negative exemplifying units.

This retrieval task is challenging due not only to the size of the candidate set but also because of the topical similarity between many ELI5 questions, which was previously noted by (Krishna et al., 2021). Retrieving based on lexical overlap, as in BM-25 and other string-matching based approaches, cannot identify exemplifying units that are relevant but share little overlap with the context.

3.2 Evaluation Data / Metric

We use 1,185 context-example pairs identified from 1,507 QA instances in the ELI5 development set (Petroni et al., 2021) as evaluation data. We evaluate retrieval model performance by measuring how reliable they are at retrieving the ground-truth examples from the candidate example set. Concretely, given the candidate set of examples, each model should output a ranked list of all examples according to their fit for the context. We evaluate these rankings using the recall@k with $k = 1, 3, 5, 10, 50, 100$ of the ground-truth example from the set of all 66K examples in ELI5. By using retriever-based evaluation metrics, we can directly measure a model's ability to understand and exemplify a context, in contrast to string overlap metrics like ROUGE (Lin, 2004) which are uninformative for this task.

3.3 Models

We first introduce our model (EGRET) and describe baseline retrieval methods. We use all baseline models in a zero-shot manner (without additional fine-tuning on exemplification dataset).

EGRET: an example retriever for LFQA We train an example retriever model (EGRET) on our extracted exemplification dataset from training portion of ELI5 (Fan et al., 2019), which contains $65K$ extracted context-example pairs. Our retriever consists of dual Transformer encoders (Vaswani et al., 2017), one to encode the context query and the other to encode candidate examples (Figure 2). Both encoders are initialized as pretrained RoBERTa-base model (Liu et al., 2019). To obtain a query embedding $\boldsymbol{c}_i$, we feed the query encoder the surrounding context of a masked-out example, as in Figure 2. Similarly, we use the other encoder to compute embeddings of the ground-truth example ($\boldsymbol{e}_i^+$) as well as negative examples sampled from other contexts, forming a set E of example embeddings. We fine-tune both encoders in EGRET with a contrastive learning objective (van den Oord et al., 2018; Chen et al., 2020):

$$\mathcal{L}(\theta) = -\sum_{(\boldsymbol{c}_i, \boldsymbol{e}_i) \in E} \log \frac{\exp \boldsymbol{c}_i \cdot \boldsymbol{e}_i^+}{\sum_{\boldsymbol{e}_j \in E} \exp \boldsymbol{c}_i \cdot \boldsymbol{e}_j} \quad (1)$$

This objective places the context vector $\boldsymbol{c}_i$ *close* to that of the ground-truth example vector $\boldsymbol{e}_i^+$ of an example, and *far* from other examples $\boldsymbol{e}_j$ in the batch E ("in-batch" negative samples). We train both the left-context-only and the left-and-right-context models on a single RTX-8000 GPU for 10 epochs, using the Adam optimizer (Kingma and Ba, 2015) with learning rate initialized at $1e-5$ for 10 epochs with early stopping. Both models converge in 4 epochs of training over the ELI5 dataset.

Model	Context	Recall@k (↑)					
		1	3	5	10	50	100
baselines, including pretrained dense retrievers							
random		0.0	0.0	0.01	0.02	0.08	0.15
BM25 (Robertson et al., 1995)	L	4.6	9.5	12.1	16.2	25.6	30.4
DPR (Karpukhin et al., 2020)	L	2.7	5.2	7.1	9.7	20.3	27.5
ColBERT (Khattab and Zaharia, 2020)	L	6.0	11.8	14.3	18.2	31.2	36.3
SBERT (Reimers and Gurevych, 2021)	L	5.7	11.6	15.0	20.4	34.3	42.2
BM25 (Robertson et al., 1995)	L+R	8.7	16.1	20.0	24.8	38.0	42.7
DPR (Karpukhin et al., 2020)	L+R	4.4	8.6	11.1	15.7	27.2	33.5
ColBERT (Khattab and Zaharia, 2020)	L+R	8.9	**16.4**	18.9	23.3	36.3	42.2
SBERT (Reimers and Gurevych, 2021)	L+R	**9.2**	16.3	**21.2**	**27.1**	**44.2**	**51.3**
our models, trained on exemplification data							
EGRET (ELI5)	L	13.0	22.8	29.3	36.5	55.2	64.0
EGRET (Books3 only)	L	19.3	30.4	36.8	44.1	63.1	69.0
EGRET (Books3 + ELI5)	L	21.1	33.5	39.2	46.8	66.7	73.0
EGRET (ELI5)	L+R	23.5	35.6	41.9	51.0	71.0	77.6
EGRET (Books3 only)	L+R	33.4	47.8	53.8	61.5	76.5	82.5
EGRET (Books3 + ELI5)	L+R	**36.9**	**50.6**	**58.2**	**64.8**	**80.2**	**85.8**

Table 4: Our EGRET model outperforms pretrained (or non-parametric) baselines on the example retrieval task, indicating that exemplification cannot be solved by term matching or coarse query-context similarity alone. Pretraining EGRET on out-of-distribution examples from Books3 results in large improvements in recall@k. Finally, including context to the right of the exemplifying unit significantly boosts performance.

Pretraining EGRET on a huge set of examples: While the EGRET model described above is trained on the ELI5 dataset, exemplification is pervasive in many kinds of texts, as shown by our annotation in Section 2. Thus, we also experiment with a *transfer* learning scenario by pretraining EGRET on a dataset of 3.5 million examples extracted from Books3 (Gao et al., 2020), and then fine-tuning the resulting model on the ELI5 examples. We perform Books3 pretraining for both left-context-only and left-and-right-context models on a single RTX-8000 GPU for 5 epochs using Adam with learning rate initialized at $1e-5$. Both models converge after one epoch of fine-tuning over the ELI5 dataset.

Baselines: We compare EGRET to a term matching method as well as three publicly-available pretrained dense retrievers.

BM25 (Robertson et al., 1995): BM25 retrieves text via a scoring function reliant on lexical overlap. We use the implementation from the rank_bm25 library,[4] with the default BM25Okapi as the similarity algorithm.

DPR (Karpukhin et al., 2020): DPR is a retriever model trained on Natural Questions (Kwiatkowski et al., 2019) that computes dense representations of queries and evidence paragraphs for retrieval.

ColBERT (Khattab and Zaharia, 2020): ColBERT similarly uses pretrained language models to embed text, but contextualizes query and candidate documents using late interaction and was trained on MS MARCO (Bajaj et al., 2018).

SBERT (Reimers and Gurevych, 2021): SBERT is a sentence-BERT-based encoder model with down-project layers that outperforms DPR on the MS MARCO dataset.

4 Results & Analysis

We report results from baseline retrievers and our trained models in Table 4. We also conduct a human evaluation to compare retrieved examples from EGRET (**L**) to examples *generated* by the state-of-the-art LFQA Routing Transformer model of Krishna et al. (2021).

4.1 Automatic Evaluation

While all models outperform random baselines, our trained EGRET models (Fan et al., 2019) consistently outperform both the lexical baseline (BM25) and other neural baselines. Among the baseline models, SBERT model, trained on MSMARCO dataset, consistently outperforms other models and DPR model lags behind the lexical baseline.

[4] https://github.com/dorianbrown/rank_bm25

Including context after the exemplifying unit improves recall: As a sanity check, we observe that including more context (**L+R**) significantly boosts recall for both EGRET and all baselines compared to including just context before the exemplifying unit (**L**). In addition to providing more constraints over the exemplifying unit, we also observe improvements on multi-sentence exemplifying units due to term matching; as this **L+R** setting is not particularly realistic, we analyze only the **L** configuration moving forward.

Pretraining improves example retrieval: Pretraining on out-of-distribution Books3 examples substantially boosts EGRET's performance, with the best left-context-only model achieving a recall@1 of 21.1% compared to 13.0% without pretraining. In fact, an EGRET model pretrained on Books3 *without* fine-tuning on in-domain ELI5 data (19.3% recall@1) outperforms the ELI5-only EGRET. While Figure 1 shows that the distribution over exemplification types differs based on the dataset/domain, our results suggest that many aspects of exemplification apply generally to wide forms of writing, and that the pretrained Books3 EGRET could be useful for many other applications.

4.2 Human evaluation of retrieved examples vs. generated examples

How do the examples retrieved by EGRET compare to examples generated by a state-of-the-art LFQA model? In theory, generative models for LFQA should be able to produce examples tailored to any input context; in practice, however, they struggle to generate relevant and informative examples. Retriever model will always produce human-written examples, but it may not always be possible to retrieve an example for an arbitrary held-out context. To explore this trade-off, we conduct a human evaluation by providing Mechanical Turk workers with an ELI5 question and context, and asking them to both *rank* and *rate* (on a 5 point Likert scale) three candidate exemplifying units: (1) the ground-truth; (2) the top-ranked retrieval from EGRET, restricted to only cases where this retrieval is *not* the ground-truth; and (3) a generated output from the state-of-the-art c-REALM-RT model of Krishna et al. (2021).[5]

Task setup: In the ranking task, we ask workers to produce a ranking of the three choices (e.g., 1>2>3). We allow equality (e.g., 1=2>3) since multiple candidates can be equally valid for a given context. In the rating task, we ask workers to evaluate how well each example fits with the given context on a scale of 1 to 5. For both tasks, we collect three annotations per item for 100 total items, and we pay $0.35 per item for an estimated hourly wage of $21 per hour.[6] While a completely fair comparison of EGRET to c-REALM-RT is infeasible due to differences in training objective and architecture, we choose to focus only on sentence-level exemplifying units that begin with *"For example"*. We provide the question, left context, and *"For example"* marker to c-REALM-RT and decode using nucleus sampling with $p = 0.9$ until a sentence boundary (e.g., period) is generated. For the retrieved output, we use EGRET(Books3 + ELI5, **L**) since the RT model has access to only the left context.

EGRET retrievals are preferred over generated exemplifying units: In both tasks, crowdworkers exhibit a clear preference for exemplifying units retrieved by EGRET compared to those generated by c-REALM-RT. While both tasks are fairly subjective, as shown by the low interannotator agreement measured by Krippendorf's alpha, the general trend is apparent. These results indicate that as of now, exemplification in LFQA is better handled by retrieval models than generative models, and that research into hybrid generation/retrieval LFQA models is a promising direction. We also note that our human study is extremely coarse, and future experiments could perform more fine-grained ratings of properties such as grammaticality and relevance instead of overall fit with context.

5 Related Work

Linguistic studies of exemplification: Early work (Kurohashi and Nagao, 1994) studies automatic detection of discourse relations including exemplification. Several works have studied exemplification in the domains of academic writing and teaching (Hyland, 2007; Oliveira and Brown, 2016; Triki, 2021). Li and Nenkova (2016) examine the closely-related *instantiation* discourse

[5] This model is pretrained on the PG-19 dataset (Rae et al., 2020) and fine-tuned on ELI5, conditioned on retrieved Wikipedia documents.

[6] We restrict workers to those in English speaking countries who have completed at least 1000 HITs with an acceptance rate of 97%.

Left context	Ground-truth	EGRET-retrieved vs others	Analysis
Evolution is not a force towards the optimum, it's a force towards the minimum necessary.	For example, if grass was poisonous, it would be better for its survival, as less animals would come eat it.	**[EGRET-retrieved]**: For example, we move incredible slowly when compared to the maximum speed allowed in the universe. (**R**:0.111 / **H**:5.0) **[Highest ROUGE-L]:** For example, if a certain percentage of snakes are venomous, then the more snakes in a given area the more venomous. (**R**:0.22)	ROUGE is not a viable evaluation for example quality. Our EGRET's retrieved example was rated a 5/5 by all three crowdworkers but achieves lower ROUGE than an irrelevant example.
... You're brain is asleep and not paying any attention to your body so it ignores all of these stimuli unless they become too hard to ignore.	For example if the touching turns to slapping, the talking turns to yelling, or the light in the eyes turns to really bright light in the eyes.	**[EGRET-retrieved]**: For example, if you're in a room with a clock ticking you don't notice the ticking after a while. (**H**:4.0) **[c-REALM-RT Generated]**: For example, its not just that your brain is dead. (**H**:2.0)	The **EGRET-retrieved** example effectively illustrates the phenomenon in the context and receives a higher average rating from crowdworkers than the generated example and even the ground-truth (3.3).
... Multiple births mean less time per offspring. Each individual offspring therefore has a lower chance of survival, ... Seems like the larger mammals tend to have **single births.**	For example, polar bears and elephants usually have single births.	**[EGRET-retrieved]**: For example, in mammals, a typical litter will be one offspring per pair of nipples as this is as many individuals a female can reasonably sustain.	EGRET retrieves an example based on a key entity from the context (mammals) but fails to address the concept to be exemplified (**"single births"**)

Table 5: Instances where EGRET retrieves exemplifying units that are rated highly by **H**umans but have low **R**OUGE score with the ground-truth example (top); where model retrievals are rated as more meaningful than those generated by the c-REALM-RT model (middle); and where the model fails to retrieve a relevant example by relying too much on lexical overlap (bottom).

	$\textbf{Rating}_{\text{STD}}$ (↑)	Krip. α
c-REALM-RT	$2.80_{0.775}$	0.058
EGRET (Books3 + ELI5, **L**)	$3.55_{0.636}$	0.125
Ground-truth	$3.70_{0.597}$	0.128

Table 6: Crowdworkers rate EGRET retrievals higher (on a scale of 1 to 5 of how well the exemplifying unit fits with the context) than the SOTA generative LFQA model.

	$\textbf{Ranking}_{\text{STD}}$ (↓)	Krip. α
c-REALM-RT	$2.26_{0.271}$	0.168
EGRET (Books3 + ELI5, **L**)	$1.88_{0.252}$	0.154
Ground-truth	$1.71_{0.284}$	0.200

Table 7: On our ranking task (1=best, 3=worst), crowdworkers prefer EGRET retrievals over the generative LFQA model.

relation, where one text span explains in further detail the events described in another text span.

Long-form question answering: Our work studies exemplification mainly within the task of long-form question answering (LFQA), which involves generating paragraph-length answers to open-ended questions. Previous work has approached this problem using retrieval-augmented generation (Fan et al., 2019; Lewis et al., 2020), while Nakano et al. (2021) set up an interactive language model that learns LFQA through human interaction. Krishna et al. (2021) demonstrate that lexical overlap metrics such as ROUGE are not meaningful for this task.

Neural retrieval models: Our EGRET retriever builds on recently-developed neural models that retrieve evidence documents for open-retrieval question answering (Karpukhin et al., 2020; Guu et al., 2020) and fact-checking (Samarinas et al., 2021). These models demonstrate superior performance compared to non-neural methods like BM25 (Robertson et al., 1995); that said recent sparse/dense hybrid retrievers (Luan et al., 2021) could be interesting to explore on our exemplification task in the future.

6 Conclusion

In this work, we present the first computational study of exemplification in long-form question answering. We perform a detailed annotation over the use of exemplification across various domains and observe different distributions over complex exemplification types and units. While existing LFQA systems are conditional language models that do not give special treatment to exemplification, we propose to *retrieve* examples based on their context instead of generating them. We develop EGRET, a simple dual encoder trained with a contrastive learning objective, that outperforms a diverse set of baselines on this task of example retrieval, which we can meaningfully evaluate using simple ranking metrics instead of unsuitable metrics like ROUGE. We hope that our work spurs researchers to consider separately modeling and evaluating the fine-grained linguistic and discourse phenomena found in LFQA data.

7 Ethical Considerations

We make use of pretrained language models to both generate and retrieve text in this work. Representations from pretrained language models are known to cause ethical concerns, such as perpetuating racial or gender bias (Field et al., 2021; Gala et al., 2020). We advise using caution and adopting a post-processing strategy to filter potentially offensive text produced by pretrained language models before releasing text content to users. Additionally, we note that most existing LFQA datasets (including the ELI5 dataset used in this work) and benchmarks are collected from English text sources. We hope future works can explore the use of exemplification in other languages.

References

Annelie Ädel. 2006. Metadiscourse in l1 and l2 english.

Payal Bajaj, Daniel Campos, Nick Craswell, Li Deng, Jianfeng Gao, Xiaodong Liu, Rangan Majumder, Andrew McNamara, Bhaskar Mitra, Tri Nguyen, Mir Rosenberg, Xia Song, Alina Stoica, Saurabh Tiwary, and Tong Wang. 2018. Ms marco: A human generated machine reading comprehension dataset.

Ting Chen, Simon Kornblith, Mohammad Norouzi, and Geoffrey Hinton. 2020. A simple framework for contrastive learning of visual representations. In *Proceedings of the International Conference of Machine Learning*.

Barbara Fine Clouse. 2013. *The student writer*. McGraw-Hill.

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.

Anjalie Field, Su Lin Blodgett, Zeerak Waseem, and Yulia Tsvetkov. 2021. A survey of race, racism, and anti-racism in NLP. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1905–1925, Online. Association for Computational Linguistics.

Dhruvil Gala, Mohammad Omar Khursheed, Hannah Lerner, Brendan O'Connor, and Mohit Iyyer. 2020. Analyzing gender bias within narrative tropes. In *Workshop on NLP and CSS at EMNLP*.

Leo Gao, Stella Biderman, Sid Black, Laurence Golding, Travis Hoppe, Charles Foster, Jason Phang, Horace He, Anish Thite, Noa Nabeshima, Shawn Presser, and Connor Leahy. 2020. The Pile: An 800gb dataset of diverse text for language modeling. *arXiv preprint arXiv:2101.00027*.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: Retrieval-augmented language model pre-training. *ArXiv*, abs/2002.08909.

Ken Hyland. 2007. Applying a gloss: exemplifying and reformulating in academic discourse. *Applied Linguistics*, 28:266–285.

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

Omar Khattab and Matei Zaharia. 2020. Colbert: Efficient and effective passage search via contextualized late interaction over bert.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*.

Kalpesh Krishna, Aurko Roy, and Mohit Iyyer. 2021. Hurdles to progress in long-form question answering. In *North American Association for Computational Linguistics*.

Sadao Kurohashi and Makoto Nagao. 1994. Automatic detection of discourse structure by checking surface information in sentences. In *COLING 1994 Volume 2: The 15th International Conference on Computational Linguistics*.

Tom Kwiatkowski, Jennimaria Palomaki, Olivia Redfield, Michael Collins, Ankur Parikh, Chris Alberti, Danielle Epstein, Illia Polosukhin, Matthew Kelcey, Jacob Devlin, Kenton Lee, Kristina N. Toutanova, Llion Jones, Ming-Wei Chang, Andrew Dai, Jakob Uszkoreit, Quoc Le, and Slav Petrov. 2019. Natural questions: a benchmark for question answering research. *Transactions of the Association of Computational Linguistics*.

Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Kuttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. *ArXiv*, abs/2005.11401.

Junyi Jessy Li and Ani Nenkova. 2016. The instantiation discourse relation: A corpus analysis of its properties and improved detection. In *NAACL*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

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

Yi Luan, Jacob Eisenstein, Kristina Toutanova, and Michael Collins. 2021. Sparse, Dense, and Attentional Representations for Text Retrieval. *Transactions of the Association for Computational Linguistics*, 9:329–345.

Charles F. Meyer. 1992. Apposition in contemporary english.

Reiichiro Nakano, Jacob Hilton, Suchir Balaji, Jeff Wu, Long Ouyang, Christina Kim, Christopher Hesse, Shantanu Jain, Vineet Kosaraju, William Saunders, Xu Jiang, Karl Cobbe, Tyna Eloundou, Gretchen Krueger, Kevin Button, Matthew Knight, Benjamin Chess, and John Schulman. 2021. Webgpt: Browser-assisted question-answering with human feedback. *CoRR*, abs/2112.09332.

Alandeom W. Oliveira and Adam Oliver Brown. 2016. Exemplification in science instruction: Teaching and learning through examples. *Journal of Research in Science Teaching*, 53:737–767.

Fabio Petroni, Aleksandra Piktus, Angela Fan, Patrick Lewis, Majid Yazdani, Nicola De Cao, James Thorne, Yacine Jernite, Vladimir Karpukhin, Jean Maillard, Vassilis Plachouras, Tim Rocktäschel, and Sebastian Riedel. 2021. KILT: a benchmark for knowledge intensive language tasks. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2523–2544, Online. Association for Computational Linguistics.

Jack W. Rae, Anna Potapenko, Siddhant M. Jayakumar, Chloe Hillier, and Timothy P. Lillicrap. 2020. Compressive transformers for long-range sequence modelling. In *International Conference on Learning Representations*.

Nils Reimers and Iryna Gurevych. 2021. The curse of dense low-dimensional information retrieval for large index sizes. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*, pages 605–611, Online. Association for Computational Linguistics.

Stephen E Robertson, Steve Walker, Susan Jones, Micheline M Hancock-Beaulieu, Mike Gatford, et al. 1995. Okapi at trec-3. *Nist Special Publication Sp*, 109:109.

Chris Samarinas, Wynne Hsu, and Mong Li Lee. 2021. Improving evidence retrieval for automated explainable fact-checking. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies: Demonstrations*, pages 84–91, Online. Association for Computational Linguistics.

Nesrine Triki. 2021. Exemplification in research articles: Structural, semantic and metadiscursive properties across disciplines. *Journal of English for Academic Purposes*, 54:101039.

Aäron van den Oord, Yazhe Li, and Oriol Vinyals. 2018. Representation learning with contrastive predictive coding. *ArXiv*, abs/1807.03748.

Ashish Vaswani, Noam M. Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *ArXiv*, abs/1706.03762.

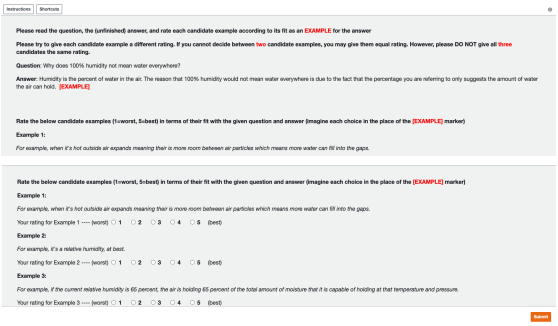
A More Annotated Examples

We present *nuanced* examples, on which annotators disagree with. When the example refers to abstract actions, annotations on whether it is real or hypothetical are split.

> (1) Carpets are used for a variety of purposes , including insulating a person 's feet from a cold tile or concrete floor , **making a room more comfortable** as a place to sit on the floor (*e.g. , when playing with children or as a prayer rug*) , reducing sound from walking (particularly in apartment buildings) and adding decoration or colour to a room .

> (2)The power of the purse can be used **positively** (e.g. *awarding extra funding to programs that reach certain benchmarks*) or **negatively** (e.g. *removing funding for a department or program , effectively eliminating it*) .

B Evaluation Interface on Mechanic Turk



Given a question, a partial answer (context) and three candidate examples, a worker is asked to rate all three candidate examples, according to their fit with the question and the given context.

Dataset	Type	Personal	Left context	Extracted Example
NQ	Real		Group Areas Act was the title of three acts of the Parliament of South Africa enacted under the apartheid government of South Africa. The acts assigned racial groups to different residential and business sections in urban areas in a system of urban apartheid.**An effect of the law was to exclude non-Whites from living in the most developed areas , which were restricted to Whites**	(e.g. , *Sea Point , Lansdowne , Cape Town , Claremont , Cape Town*).
NQ	Hypothetical		Although the safest way to recognize a chord 's root is , after having reduced the chord to close spacing , to rearrange it as a stack of thirds , there are shortcuts to this : [...] **With chord types, such as chords with added sixths or chords over pedal points, more than one possible chordal analysis may be possible.**	For example, *in a tonal piece of music , the notes C, E, G, A, sounded as a chord , could be analyzed as a C major sixth chord in root position (a major triad – C, E, G – with an added sixth – A – above the root) or as a first inversion A minor seventh chord (the A minor seventh chord contains the notes A, C, E and G, but in this example, the C note, the third of the A minor chord, is in the bass).*
ELI5	Real	✓	My uncle owns a pretty large recycling business. They export the majority of their newly created raw materials to the places that produce with the materials (China).[...] **Raw material are often remade into base products several times over before it gets to a manufacturing plant.**	For example: *My uncle's business is primarily plastics. They get cast offs, seconds, etc plastic from all kinds of US manufacturers. They then sort, filter and break down the plastic to the most basic starting point (often really small non died beads) and ship it to China. [...]*
ELI5	Hypothetical	✓	OP I guess you are coming from movies/ace attorney but avoiding that. Let's say you have the most cut and dry murder case [...] **There are a limited number of prosecutors, judges, and defense attorneys**	(for example *I currently intern at a medical malpractice firm and if we were forced to do criminal defense I would actually be the most qualified one there to do so- at a firm where the youngest attorney still has 15 years of experience*)
Books3	Real		**People in a second group were given a verbal description, with which they were to construct an image of walking along the two segments**	For example, *people were told to imagine they would "Go forward 3 m, turn clockwise 90°, then go forward 3 m."*
Books3	Hypothetical		When we cook together, I have to stay alert because she is always throwing a lemon at me—**sometimes double down on acid and mix lemon juice with a little bit of vinegar to get the sunny sweet-sour note of the citrus along with earthy, apple, or wine notes of a vinegar for greater complexity**	For example, *if you toss roasted beets (a notoriously earthy and sweet vegetable that some might say tastes like soil) with just lemon juice, olive oil, and salt, it would no doubt be good, but if you supplement the sunny lemon juice with a tiny splash of sherry vinegar for its woodsy earthiness, you get a roasted beet dish that is far more complex and delicious than if you had used only one or the other.*

Table A1: Different types of annotated examples in the three datasets. The anchor (mark in **bold**) and example (mark in *italics*) are highlighted.

Left context	Ground truth	Retrieved/Generated	Error Analysis
Dog's do not pass a mirror test so it's highly unlikely they have a sense of self.[...]**They do recognize there names, it's a bit of an illusion though.**	For example if you have two dogs, and you give one of them a treat when you say Fido, and the other when you say Clifford, they learn that the respective words only apply to them.	[*EGRET*] For example people can identify their own dog	*EGRET* retrieved relevant but semantically incorrect example (people identify their dog, instead of how dog identify themselves).
An economist would say healthcare has a positive externality.[...]**There are some things you can buy that make everyone better off.**	For example: going to the doctor every time you are sick will make you less likely to make other people sick.	[*EGRET*] For example, a butterfly house, a free cinema, games consoles etc.	*EGRET* retrieved examples related to the immediate preceding context ("some things you can buy...") but failed to retrieved examples based on earlier context (about healthcare).
4 billion - Economic and Military aid for Pakistan, Egypt, and Jordan. The goal is to have a few people in the mideast who call us allies. Essentially, we buy their cooperation. **That cooperation is sometimes useful.**	For example, when we killed Osama Bin Laden, we sent troops into Pakistan. Normally, countries don't tolerate troops from other countries. The Pakistanis did complain a little, but they didn't do anything about it.	[*c-REALM-RT*] For example, For everyone here talking about how a lot of aid works: If we put money towards helping foreign countries rebuild, we are imposing restrictions on domestic activity. [...]	*c-REALM-RT* generated an on-topic hypothetical example, which contradicts with the context.
You don't usually work on the same files because everything is split up between the departments. **I haven't used USD yet but I have encountered the following workflow in different studios (using Maya).**	For example: a character that has been rigged by one (or more, but not at the same time) rigger goes to the animators. Every animator works with the same character rig BUT each animator works on his/her own shot.[...]	[*c-REALM-RT*] For example, As a starting point: I'm a post graduate and work in the final sector of the project, not the project itself. Most of the work is done with other studios around the world who are made up of multiple departments.[...]	*c-REALM-RT* generated a personal example that is irrelevant to the context.

Table A2: Error analysis of retrieved and generated examples.

Left context	Ground-truth	Retrieved	Top ROUGE
evolution is not a force towards the optimum, it's a force towards the minimum necessary.	for example, if grass was poisonous, it would be better for its survival, as less animals would come eat it.	For example, we move incredible slowly when compared to the maximum speed allowed in the universe. (**R**:0.111/**H**:5.0)	For example, if a certain percentage of snakes are venomous, then the more snakes in a given area the more venomous. (**R**:0.22)

Table A3: Retrieved examples with high **H**uman rating and highest **R**OUGE-L w.r.t. ground truth.