RELiC: Retrieving Evidence for Literary Claims

Anonymous ACL submission

Abstract

Humanities scholars commonly provide evidence for claims that they make about a work of literature (e.g., a novel) in the form of quotations from the work. We collect a large-scale dataset (RELiC) of 90K literary quotations and surrounding critical analysis and use it to formulate the novel task of *literary evidence retrieval*, in which models are given an excerpt from a literary analysis surrounding a masked quotation and asked to retrieve the quoted passage from the set of all passages in the work. Solving this retrieval task requires a deep understanding of complex literary and linguistic phenomena, which proves challenging to methods that overwhelmingly rely on lexical and semantic similarity matching. We implement a RoBERTa-based dense passage retriever for this task that outperforms existing pretrained information retrieval baselines; however, experiments and analysis by human domain experts indicate that there is substantial room for improvement.

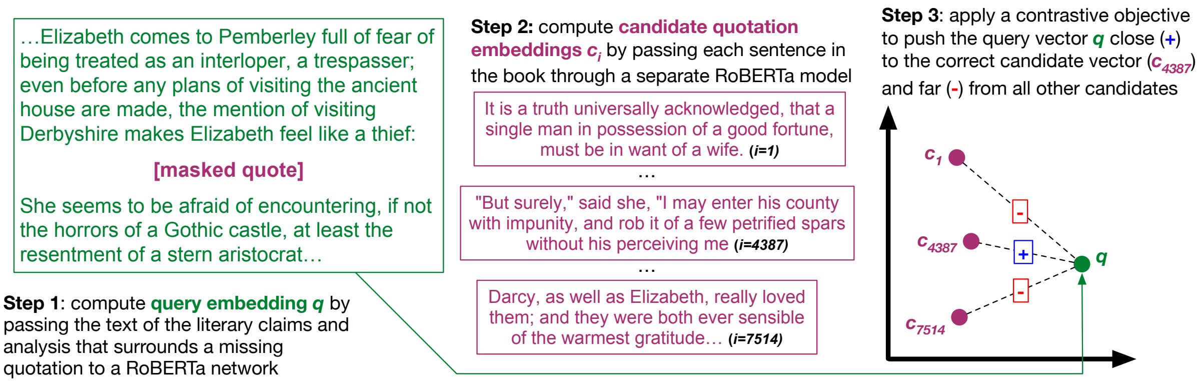
1 Introduction

When analyzing a literary work (e.g., a novel or short story), scholars make claims about the text and provide supporting evidence in the form of quotations from the work (Thompson, 2002; Finnegan, 2011; Graff et al., 2014). For example, McNichol (1990) claims that Jacob, the titular character in Virginia Woolf's *Jacob's Room*, has "all the arrogance of youth in his belief that he knows what life is and where meaning and true values are to be found", and then directly quotes Jacob's outburst against elderly people as evidence: "Had they never read Homer, Shakespeare, the Elizabethans?".

Human readers decipher literary arguments like these by making complex connective inferences between claims and quotes (e.g., connecting the arrogance of youth to the belief that meaning and true values are found in the works of Homer and Shakespeare). This process requires a deep understanding of both literary phenomena (e.g., metaphor and symbolism) and linguistic phenomena (coreference, paraphrasing, and stylistics). In this paper, we computationally study the relationship between literary claims and quotations by collecting a large-scale dataset (RELiC) of 90K scholarly excerpts of literary analysis, each of which contains a quotation from one of 79 widely-read English texts.

The complexity of the claims and quotations in RELiC makes it a challenging testbed for modern neural retrievers: given just the text of the claim and analysis that surrounds a masked quotation, can a model retrieve the quoted passage from the set of all possible passages in the literary work? This *literary evidence retrieval* task (see Figure 1) differs considerably from retrieval problems commonly studied in NLP, such as those used for fact checking (Thorne et al., 2018), open-domain QA (Chen et al., 2017; Chen and Yih, 2020), and text generation (Krishna et al., 2021), in the relative lack of lexical or even semantic similarity between claims and queries. Instead of latching onto surface-level cues, our task requires models to understand complex devices in literary writing and apply general theories of interpretation. RELiC is also challenging because of the large number of retrieval candidates: for *War and Peace*, the longest literary work in the dataset, models must choose from one of $\sim$ 32K candidate passages.

How well do state-of-the-art retrievers perform on RELiC? Inspired by recent research on dense passage retrieval (Guu et al., 2020; Karpukhin et al., 2020), we build a neural model (dense-RELiC) by embedding both scholarly claims and candidate literary quotations with pretrained RoBERTa networks (Liu et al., 2019), which are then fine-tuned using a contrastive objective that encourages the representation for the ground-truth quotation to lie nearby to that of the claim. Both sparse retrieval methods such as BM25 as well as pretrained dense retrievers such as DPR and REALM perform



…Elizabeth comes to Pemberley full of fear of being treated as an interloper, a trespasser; even before any plans of visiting the ancient house are made, the mention of visiting Derbyshire makes Elizabeth feel like a thief:

[masked quote]

She seems to be afraid of encountering, if not the horrors of a Gothic castle, at least the resentment of a stern aristocrat…

Step 1: compute **query embedding *q*** by passing the text of the literary claims and analysis that surrounds a missing quotation to a RoBERTa network

Step 2: compute **candidate quotation embeddings *c_i*** by passing each sentence in the book through a separate RoBERTa model

It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife. *(i=1)*

…

"But surely," said she, "I may enter his county with impunity, and rob it of a few petrified spars without his perceiving me *(i=4387)*

…

Darcy, as well as Elizabeth, really loved them; and they were both ever sensible of the warmest gratitude… *(i=7514)*

Step 3: apply a contrastive objective to push the query vector ***q*** close (+) to the correct candidate vector (c_{4387}) and far (-) from all other candidates

Figure 1: An example of our *literary evidence retrieval* task and the model we built to solve it. The model must retrieve a missing quotation from *Pride and Prejudice* given the literary claims and analysis that surround the quotation. The retrieval candidate set for this example consists of all 7,514 sentences from *Pride and Prejudice*. Our dense-RELiC model is trained with a contrastive loss to push a learned representation of the surrounding context close to a representation of the ground-truth missing quotation (here, the 4,387th sentence from the novel).

poorly on RELiC, which underscores the difference between our dataset and existing information retrieval benchmarks (Thakur et al., 2021) on which these baselines are much more competitive. Our dense-RELiC model fares better than these baselines but still lags far behind human performance, and an analysis of its errors suggests that it struggles to understand complex literary phenomena.

Finally, we qualitatively explore whether our dense-RELiC model can be used to support evidence-gathering efforts by researchers in the humanities. Inspired by prompt-based querying (Jiang et al., 2020), we issue our own out-of-distribution queries to the model by formulating simple descriptions of events or devices of interest (e.g., *symbols of Gatsby's lavish lifestyle*) and discover that it often returns relevant quotations. To facilitate future research in this direction, we publicly release our dataset and models.[1]

2 Collecting a Dataset for Literary Evidence Retrieval

We collect a dataset for the task of **R**etrieving **E**vidence for **Li**terary **C**laims, or RELiC, the first large-scale retrieval dataset that focuses on the challenging literary domain. Each example in RELiC consists of two parts: (1) the context surrounding the quoted material, which consists of literary claims and analysis, and (2) a quotation from a widely-read English work of literature. This section describes our data collection and preprocessing steps, as well as a fine-grained analysis of 200 examples from RELiC to shed light on the types of quotations it contains.

2.1 Collecting and Preprocessing RELiC

Selecting works of literature: We collect 79 primary source works written or translated into English[2] from Project Gutenberg and Project Gutenberg Australia.[3] These public domain sources were selected because of their popularity and status as members of the Western literary canon, which also yield more scholarship (Porter, 2018). All primary sources were published in America or Europe between 1811 and 1949. 77 of the 79 are fictional novels or novellas, one is a collection of short stories (*The Garden Party and Other Stories* by Katherine Mansfield), and one is a collection of essays (*The Souls of Black Folk* by W. E. B. Du Bois).

Collecting quotations from literary analysis: We queried all documents in the HathiTrust Digital Library,[4] a collaboration between academic and research libraries, for exact matches of all sentences of ten or more tokens from each of the 79 works. The overwhelming majority of HathiTrust documents are scholarly in nature, so most of these matches yielded critical analysis of the 79 primary source works. We received permission from the HathiTrust to publicly release short windows of

[1] Our RELiC dataset is attached to this submission.

[2] Of the 79 primary sources in RELiC, 72 were originally written in English, 3 were written in French, and 4 were written Russian. RELiC contains the corresponding English translations of these 7 primary source works. The complete list of primary source works is available in Appendix Tables A7 and A8.

[3] https://www.gutenberg.org/

[4] https://www.hathitrust.org/

# training examples	71,395
# validation examples	9,036
# test examples	9,034
average query length (words)	154.1
average candidate length (words)	45.5
average # of candidates per query	10,648.8
# primary sources	79
# unique sec. sources	9,127

Table 1: RELiC statistics. Primary sources are texts from Project Gutenberg or Project Gutenberg Australia and secondary sources are scholarly works from HathiTrust.

text surrounding each matching quotation.

Filtering and preprocessing: The scholarly articles we collected from our HathiTrust queries were filtered to exclude duplicates and non-English sources. We then preprocessed the resulting text to remove pervasive artifacts such as in-line citations, headers, footers, page numbers, and word breaks using a pattern-matching approach (details in Appendix A). Finally, we applied sentence tokenization using spaCy's dependency parser-based sentence segmenter[5] to standardize the size of the windows in our dataset. Each window in RELiC contains the identified quotation and four sentences of claims and analysis[6] on each side of the quotation (see Table 2 for examples). Finally, to avoid asking the model to retrieve a quote it has already seen during training, we create training, validation, and test splits such that primary sources in each fold are mutually exclusive. Statistics of our dataset sources are provided in Appendix A.3.

2.2 Comparison to other retrieval datasets

Table 1 contains detailed statistics of RELiC. To the best of our knowledge, RELiC is the first retrieval dataset in the literary domain, and the only one that requires understanding complex phenomena like symbolism and metaphor. We provide a detailed comparison of ReLiC to other retrieval datasets in the recently-proposed BEIR retrieval benchmark (Thakur et al., 2021) in Appendix Table A10. RELiC has a much longer query length (154.1 tokens on average) compared to all BEIR datasets except ArguAna (Wachsmuth et al., 2018). Furthermore, our results in Section 3.3 show that while these longer queries confuse pretrained retriever models (which heavily rely on token overlap), a model trained on RELiC is able to leverage the longer queries for better retrieval.

2.3 Analyzing different types of quotation

What are the different ways in which literary scholars use direct quotation in RELiC? We perform a manual analysis of 200 held-out examples to gain a better understanding of quotation usage, categorizing each quote into the following three types:

Claim-supporting evidence: In 153 of the 200 annotated examples, literary scholars used direct quotation to provide evidence for a more general claim about the primary source work. In the first row of Table 2, Hartstein (1985) claims that "this whale... brings into focus such fundamental questions as the knowability of space:" and then quotes the following metaphorical description from *Moby Dick* as evidence: "And as for this whale spout, you might almost stand in it, and yet be undecided as to what it is precisely." When quoted material is used as **claim-supporting evidence**, the context before and after usually refers directly to the quoted material;[7] for example, the paradoxes of reality and uncertainties of this world are exemplified by the vague nature of the whale spout.

Paraphrase-supporting evidence: In 25 of the examples, we observe that scholars used the primary source work to support their own paraphrasing of the plot in order to contextualize later analysis. In the second row of Table 2, Blackstone (1972) uses the quoted material to enhance a summary of a specific scene in which Jacob's mind is wandering during a chapel service. Jacob's daydreaming is later used in an analysis of Cambridge as a location in Virginia Woolf's works, but no literary argument is made in the immediate context. When quoted material is being employed as **paraphrase-supporting evidence**, the surrounding context does not refer directly to the quotation.

[5] https://spacy.io/, the default segmenter in spaCy is modified to use ellipses, colons, and semicolons as custom sentence boundaries, based on the observation that literary scholars often only quote part of what would typically be defined as a sentence.

[6] The HathiTrust permitted us to release windows consisting of up to eight sentences of scholarly analysis. While more context is of course desirable, we note that (1) conventional model sizes are limited in input sequence length, and (2) context further away from the quoted material has diminishing value, as it is likely to be less relevant to the quoted span.

[7] In 8 of the 153 **claim-supporting evidence** examples, scholars introduce quoted material by explicitly referring to a specific "sentence," "motif," "scene," or similar delineation.

Quote type	Preceding context, **primary source quotation**, subsequent context
Claim-supporting evidence (*153*)	If this whale inspires the most lyrical passages in the novel, it also brings into focus such fundamental questions as the knowability of space: **And as for this whale spout, you might almost stand in it, and yet be undecided as to what it is precisely.** But Ishmael stands before the paradoxes of reality with historical and scientific intellect, wisdom, and comic elasticity that accommodates–however tenuously–the uncertainties of this world (Hartstein, 1985).
Paraphrase-supporting evidence (*25*)	But then, suddenly, Jacob's thought switches back to the lantern under the tree, with the old toad and the beetles and the moths crossing from side to side in the light, senselessly.**Now there was a scraping and murmuring. He caught Timmy Durrant's eye; looked very sternly at him; and then, very solemnly, winked.** From a boat on the Cam there is another sort of beauty to be seen. There are buttercups gilding the meadows, and cows munching, and the legs of children deep in the grass. Jacob looks at all these things and becomes absorbed (Blackstone, 1972).

Table 2: Examples of the two major types of evidence identified in our manual analysis of 200 RELiC examples.

Miscellaneous: 13 of the 200 samples were not literary analysis, though many of these were still related to literature (for example, analysis of the the film adaptation of *The Age of Innocence*). An additional ten of the samples suffered from severe OCR artifacts which are sometimes compounded by our preprocessing methods (see Appendix A.2).

3 Literary Evidence Retrieval

Having established that the examples in RELiC contain complex interplay between literary quotes and scholarly analysis, we now shift to measuring how well neural models can understand these interactions. In this section, we first formalize our evidence retrieval task, which provides the scholarly context *without* the quote as input to a model, along with a set of candidate passages that come from the same book, and asks the model to retrieve the ground-truth missing quotation from the candidates. Then, we describe standard information retrieval baselines as well as a RoBERTa-based ranking model that we implement to solve our task.

3.1 Task formulation

Formally, we represent a single window in RELiC from book b as $(..., l_{-2}, l_{-1}, q_n, r_1, r_2, ...)$ where q_n is the quoted n-sentence long passage, and l_i and r_j correspond to individual sentences before and after the quotation in the scholarly article, respectively. The window size on each side is bounded by hyperparameters l_{max} and r_{max}, each of which can be up to 4 sentences. Given a set of candidates $C_{b,n}$ that consists of all n-sentence long passages in book b, we ask a model to identify the missing quotation q_n from the candidate set $C_{b,n}$, given the surrounding $l_{-l_{max}:-1}$ and $r_{1:r_{max}}$ sentences (see Figure 1). This is a particularly challenging retrieval task because the candidates are part of the same overall narrative and thus mention the same overall set of entities (e.g., characters, locations) and other plot elements, which is a disadvantage for methods based on string overlap.

Evaluation: Models built for our task must produce a ranked list of candidates $C_{b,n}$ for each example. We evaluate these rankings using both recall@k for $k = 1, 3, 5, 10, 50, 100$ and *mean rank* of q in the ranked list. Both types of metrics focus on the position of the correct quote q in the ranked list, and neither gives special treatment to candidates that overlap with the ground-truth quote q. As such, recall@1 alone is overly strict when the quote length $l > 1$, which is why we show recall at multiple values of k. An additional motivation is that there may be multiple different candidates that fit a single context equally well. We also report accuracy on a proxy task with only three candidates, which allows us to compare with human performance as described in Section 4.

3.2 Models

Baselines: BM25 (Robertson et al., 1995) is a standard bag-of-words information retrieval method. We form queries by concatenating left and right context and use the implementation from rank_bm25 library[10] to build a BM25 model for each unique candidate set $C_{b,n}$, tuning the $k1$ and b parameters as recommended by Kamphuis et al. (2020).[11] Meanwhile, our dense retrieval baselines

[8] ColBERT does not provide a ranking for candidates outside the top 1000, so we cannot report mean rank.

[9] We do not report BM25's accuracy on the proxy task because its top-ranked quotes were used as candidates in the proxy task in addition to the ground-truth quotation.

[10] https://github.com/dorianbrown/rank_bm25, a library that implements multiple BM25-based algorithms

[11] The best configuration after tuning on the validation set is $k1 = 0.5$, $b = 0.9$.

Model	**L/R**	**Recall@*k* (↑)**						**Avg rank (↓)**	**Proxy task acc (↑)**
		1	3	5	10	50	100		
				(non-parametric / pretrained zero-shot)					
random		0.0	0.1	0.1	0.2	1.0	2.0	2520.5	33.3
BM25	1/1	1.4	3.5	4.6	6.4	12.7	17.2	1602.2	–[9]
BM25	4/4	1.2	2.7	4.1	6.7	14.6	19.6	1435.6	–
SIM (Wieting et al., 2019)	1/1	1.3	3.1	4.1	6.0	13.4	18.9	1397.6	23.0
SIM (Wieting et al., 2019)	4/4	0.9	2.1	3.1	4.7	12.1	17.5	1405.1	11.0
DPR (Karpukhin et al., 2020)	1/1	1.4	3.2	4.5	6.7	15.6	22.3	1253.1	25.5
DPR (Karpukhin et al., 2020)	4/4	1.1	2.2	3.2	5.2	13.5	20.4	1254.7	22.5
c-REALM (Krishna et al., 2021)	1/1	1.6	3.5	4.8	7.1	15.4	21.3	1383.4	23.0
c-REALM (Krishna et al., 2021)	4/4	1.0	2.2	3.3	5.1	12.7	18.6	1382.5	17.5
ColBERT (Khattab and Zaharia, 2020)	**1/1**	**2.9**	**5.8**	**7.8**	**11.0**	**21.2**	**27.7**	N/A[8]	**38.8**
ColBERT (Khattab and Zaharia, 2020)	4/4	1.8	3.9	5.4	8.2	18.4	25.1	N/A	18.9
				(trained on RELiC training set)					
dense-RELiC	0/1	2.9	6.2	8.5	11.7	22.5	29.2	1189.1	42.5
	0/4	5.4	10.1	13.2	18.1	32.5	40.5	849.4	46.5
	1/0	4.8	9.8	13.0	18.0	33.4	41.4	793.1	**67.5**
	4/0	7.2	14.9	19.4	26.2	45.0	54.0	529.8	65.5
	1/1	7.0	14.1	18.2	24.3	41.2	50.0	616.4	67.0
	4/4	**9.6**	**19.8**	**25.5**	**33.5**	**52.9**	**61.9**	**370.9**	65.0
Human domain experts	4/4								**93.5**

Table 3: Overall comparison of different systems and context sizes (L/R indicates the number of sentences on the left and right side of the missing quote) on the test set of RELiC using recall@k metrics, normalized to a maximum score of 100. Our trained dense-RELiC retriever significantly outperforms BM25 and all pretrained dense retrieval models. The average number of candidates per example is 5,041. We report the accuracy of different systems on a proxy task that we administered to human domain experts, which shows that there is huge room for improvement.[9]

are pretrained neural encoders which map queries and candidates to vectors. We compute vector similarity scores (e.g., cosine similarity) between every query/candidate pair, which are used to rank candidates for every query and perform retrieval.

We consider four pretrained dense retriever baselines in our work, which we deploy in a zero-shot manner (i.e., not fine-tuned on RELiC). **DPR** (Dense Passage Retrieval) is a dense retrieval model from Karpukhin et al. (2020) trained to retrieve relevant context paragraphs in open-domain question answering. We use the DPR context encoder[12] pretrained on Natural Questions (Kwiatkowski et al., 2019) with dot product as a similarity function. **SIM** is a semantic similarity model from Wieting et al. (2019) that is effective on semantic textual similarity benchmarks (Agirre et al., 2016). SIM is trained on ParaNMT (Wieting and Gimpel, 2018), a dataset containing 16.8M paraphrases; we follow the original implementation,[13] and use cosine similarity as the similarity function. **c-REALM** (contrastive Retrieval Augmented Language Model) is a dense retrieval model from Krishna et al. (2021) trained to retrieving relevant contexts in open-domain long-form question answering, shown to be a better retriever than REALM (Guu et al., 2020) on the ELI5 KILT benchmark (Fan et al., 2019; Petroni et al., 2021). Finally, **ColBERT** is a ranking model from Khattab and Zaharia (2020) that estimates relevance between a query and a document using contextualized late interaction, trained on the MS MARCO Ranking dataset (Bajaj et al., 2018).

Training retrievers on RELiC (dense-RELiC): Both BM25 and the pretrained dense retriever baselines perform similarly poorly on RELiC (Table 3). These methods are unable to capture more complex interactions within RELiC that do not exhibit extensive string overlap between quotation and context. As such, we also implement a strong neural retrieval model that is actually *trained* on RELiC, using a similar setup to DPR and REALM. We first form a context string c by concatenating a window of sentences on either side of the quotation q (replaced by a MASK token),

$$c = (l_{-l_{max}}, ..., l_{-1}, \text{[MASK]}, r_1, ..., r_{r_{max}})$$

[12] https://huggingface.co/facebook/dpr-ctx_encoder-single-nq-base

[13] https://github.com/jwieting/beyond-bleu

Surrounding context	**Correct candidate**	**Incorrect candidate**	**Analysis**
She is caught up for a moment or two in a fantasy of possession: **[masked quote]** The thought that she would not have been allowed to invite the Gardiners is a lucky recollection it save[s] her from something like regret. (Paris, 1978)	[*dense-RELiC*]: "And of this place," thought she, "I might have been mistress! With these rooms I might now have been familiarly acquainted!"	[*BM25*]: I should not have been allowed to invite them." This was a lucky recollection-it saved her from something very like regret.	dense-RELiC correctly retrieves the quotation that shows the "fantasy of possession," while BM25 retrieves a quote that is paraphrased in the surrounding context.
It is delicious from the opening sentence: **[masked quote]** Mr. Bingley, with his four or five thousand a year, had settled at Netherfield Park. (Masefield, 1967)	[*Human*]: It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife.	[*dense-RELiC*]: "My dear Mr. Bennet," said his lady to him one day, "have you heard that Netherfield Park is let at last?"	Human readers can immediately identify the first sentence of *Pride and Prejudice*.
Sometimes we hear Mrs Bennet's idea of marriage as a market in a single word: **[masked quote]** Her stupidity about other people shows in all her dealings with her family... (McEwan, 1986)	[*Human*]: "I do not blame Jane," she continued, "for Jane would have got Mr. Bingley if she could."	[*dense-RELiC*]: You must and shall be married by a special licence.	Human readers understood the uncommon usage of "got" to convey a transaction.

Table 4: Examples that show failure cases of BM25 (top row) and our dense-RELiC retriever (bottom two rows) from our proxy task on *Pride and Prejudice*. BM25 is easily misled by string overlap, while dense-RELiC lacks world knowledge (e.g., knowing the famous first sentence) and complex linguistic understanding (e.g., the relationship between marriage as a market and got) that humans can easily rely on to disambiguate the correct quotation.

We train two encoder neural networks to project the literary context and quote to fixed 768-d vectors. Specifically, we project c and q using **separate** encoder networks initialized with a pretrained RoBERTa-base model (Liu et al., 2019). We use the <s> token of RoBERTa to obtain 768-d vectors for the context and quote, which we denote as $\mathbf{c}_i$ and $\mathbf{q}_i$. To train this model, we use a contrastive objective (Chen et al., 2020) that pushes the context vector $\mathbf{c}_i$ close to its quote vector $\mathbf{q}_i$, but away from all other quote vectors $\mathbf{q}_j$ in the same minibatch ("in-batch negative sampling"):

$$\text{loss} = -\sum_{(c_i,q_i)\in B} \log \frac{\exp \mathbf{c}_i \cdot \mathbf{q}_i}{\sum_{q_j\in B} \exp \mathbf{c}_i \cdot \mathbf{q}_j}$$

where B is a minibatch. Note that the size of the minibatch $|B|$ is an important hyperparameter since it determines the number of negative examples.[14] Finally, all elements of the minibatch are context/quote pairs sampled from the same book. During inference, we rank all quotation candidate vectors by their dot product with the context vector.

3.3 Results

We report results from the baselines and our dense-RELiC model in Table 3 with varying context sizes where L/R refers to L preceding context sentences and R subsequent context sentences. While all models substantially outperform random candidate selection, all pretrained neural dense retrievers perform similarly to BM25, with ColBERT being the best neural retriever (2.9 recall@1). This result indicates that matching based on string overlap or semantic similarity is not enough to solve RELiC, and even powerful neural retrievers struggle at this benchmark. Training on RELiC is crucial: our best-performing dense-RELiC model performs 8x better than BM25 (9.6 vs 1.2 recall@1).

Context size and location matters for model performance: Table 3 shows that dense-RELiC effectively utilizes longer context — keeping only one sentence on each side of the quote (1/1) is not as effective as a longer context (4/4) of four sentences on each side (7.0 vs 9.6 recall@1). However,

[14] We set $|B| = 100$, and train all models for 10 epochs on a single RTX8000 GPU with an initial learning rate of 1e-5 using the Adam optimizer (Kingma and Ba, 2015), early stopping on validation loss. Models typically took 2 hours to complete 10 epochs. Our implementation uses the HuggingFace `transformers` library (Wolf et al., 2020). The total number of model parameters is 249M.

the longer contexts hurt performance for pretrained dense retrievers in the zero-shot setting (1.6 vs 1.0 recall@1 for c-REALM), perhaps because context further away from the quotation is less likely to be helpful. Finally, we observe that dense-RELiC performance is strictly better (5.4 vs 7.2 recall@1) when the model is given only preceding context (4/0 or 1/0) compared to when the model is given only subsequent context (0/4 or 0/1).

Dense vs. sparse retrievers: As expected, BM25 retrieves the correct quote when there is significant string overlap between the quotation and context, as in the following example from *The Great Gatsby*, in which the terms *sky*, *bloom*, *Mrs. McKee*, *voice*, *call*, and *back* appear in both places:

> Yet his analogy also implicitly unites the two women. Myrtle's expansion and revolution in the smoky air are also outgrowths of her surreal attributes, stemming from her residency in the Valley of Ashes. **The late afternoon sky bloomed in the window for a moment like the blue honey of the Mediterranean-then the shrill voice of Mrs. McKee called me back into the room.** The objective talk of Monte Carlo and Marseille has made Nick daydream. In Chapter I Daisy and the rooms had bloomed for him, with him, and now the sky blooms. The fact that Mrs. McKee's voice "calls him back" clearly reveals the subjective daydreamy nature of this statement.

However, this behavior is undesirable for most examples in RELiC, since string overlap is generally not predictive of the relationship between quotations and claims. The top row of Table 4 contains one such example, where dense-RELiC correctly chooses the missing quote while BM25 is misled by string overlap.

4 Human performance and analysis

How well do humans actually perform on RELiC? To compare the performance of our dense retriever to that of humans, we hired six domain experts with degrees in English literature from the Upwork[15] freelancing platform. Because providing thousands of candidates to a human evaluator is infeasible, we instead measure human performance on a simplified proxy task: we provide our evaluators with four sentences on either side of a missing quotation from *Pride and Prejudice*[16] and ask them to select one of only three candidates to fill in the blank. We obtain human judgments both to measure a *human upper bound* on this proxy task as well as to evaluate whether humans struggle with examples that fool our model.

Human upper bound: First, to measure a human upper bound on this proxy task, we chose 200 test set examples from *Pride and Prejudice* and formed a candidate pool for each by including BM25's top two ranked answers along with the ground-truth quotation. As the task is trivial to solve with random candidates, we decided to use a model to select harder negatives, and we chose BM25 to see if humans would be distracted by high string overlap in the negatives. Each of the 200 examples was separately annotated by three experts, and they were paid $100 for annotating 100 examples. The last column of Table 3 compares all of our baselines along with dense-RELiC against human domain experts on this proxy task. Humans substantially outperform all models on the task, with at least two of the three domain experts selecting the correct quote 93.5% of the time;[17] meanwhile, the highest score for dense-RELiC is 67.5%, which indicates huge room for improvement. Interestingly, all of the zero-shot dense retrievers except ColBERT 1/1 underperform random selection on this task; we theorize that this is because all of these retrievers are misled by the high string overlap of the negative BM25-selected examples.

Human error analysis of dense-RELiC: To evaluate the shortcomings of our dense-RELiC retriever, we also administered a version of the proxy task where the candidate pool included the ground-truth quotation along with dense-RELiC's two top-ranked candidates, where for all examples the model ranked the ground-truth outside of the top 1000 candidates. Three domain experts attempted 100 of these examples and achieved an accuracy of 94%, demonstrating that humans can easily disambiguate cases on which our model fails. The bottom two rows of Table 4 contain instances in which all human annotators agreed on the correct candidate but dense-RELiC failed to rank it in the top 1000. In one, all human annotators immediately recognized the opening line of *Pride and Prejudice*, one of the most famous in English literature. In the other, the claim mentions that the interpretation hinges on a single word's ("got") connotation of "a

[15] https://upwork.com

[16] We decided to keep our proxy task restricted to the most well-known book in our test set because of the ease with which we could find highly-qualified workers who self-reported that they had read (and often even re-read) *Pride and Prejudice*.

[17] As a measure of agreement between experts, we report a Krippendorf's alpha value of 0.238.

From *Frankenstein*, given **Victor does not consider the consequences of his actions:** our model's top-ranked candidates are:

1. It is even possible that the train of my ideas would never have received the fatal impulse that led to my ruin.
2. The threat I had heard weighed on my thoughts, but I did not reflect that a voluntary act of mine could avert it.
3. Now my desires were complied with, and it would, indeed, have been folly to repent.

From *The Great Gatsby*, given **A symbol of Gatsby's lifestyle:** our model's top-ranked candidates are:

1. His movements-he was on foot all the time-were afterward traced to Port Roosevelt and then to Gad's Hill where he bought a sandwich that he didn't eat and a cup of coffee.
2. Every Friday five crates of oranges and lemons arrived from a fruiterer in New York-every Monday these same oranges and lemons left his back door in a pyramid of pulpless halves.
3. On week-ends his Rolls-Royce became an omnibus, bearing parties to and from the city, between nine in the morning and long past midnight, while his station wagon scampered like a brisk yellow bug to meet all trains.

Table 5: Given a novel and a short out-of-distribution prompt, this table shows the top 3 quotations from the novel that dense-RELiC returns as evidence. The relevance of many of the returned quotations, even without string overlap between the prompt and candidates, indicates the model is learning some non-trivial relationships that could have potential impact for building tools that support humanities research.

market," which humans understood.

Issuing out-of-distribution queries to the retriever: Does our dense-RELiC model have potential to support humanities scholars in their evidence-gathering process? Inspired by prompt-based learning, we manually craft simple yet out-of-distribution prompts and queried our dense-RELiC retriever trained with 1 sentence of left context and no right context. A qualitative inspection of the top-ranked quotations in response to these prompts (Table 5) reveals that the retriever is able to obtain evidence for distinct character traits, such as the ignorance of the titular character in *Frankenstein* or Gatsby's wealthy lifestyle in *The Great Gatsby*. Additionally, the retriever's top-ranked quotes have little to no string overlap with the prompts. An exciting future direction is to integrate our retriever into a user study with humanities scholars to see if it can be useful for their research needs.

5 Related Work

Datasets for literary analysis: Our work relates to previous efforts to apply NLP to literary datasets such as LitBank (Bamman et al., 2019; Sims et al., 2019), an annotated dataset of 100 works of fiction with annotations of entities, events, coreferences, and quotations. Papay and Padó (2020) introduced RiQuA, an annotated dataset of quotations in English literary text for studying dialogue structure, while Chaturvedi et al. (2016) label character relationships in novels. Our work also relates to quotability identification (MacLaughlin and Smith, 2021), which focuses on ranking passages in a literary work by how often they are quoted in a larger collection. Unlike RELiC, however, these datasets do not contain literary analysis about the works.

Retrieving cited material: Citation retrieval closely relates to RELiC and has a long history of research, mostly on scientific papers: O'Connor (1982) formulated the task of document retrieval using "citing statements", which Liu et al. (2014) revisit to create a reference retrieval tool that recommends references given context. Bertin et al. (2016) examine the rhetorical structure of citation contexts. Perhaps closest to RELiC is the work of Grav (2019), which concentrates on the quotation of secondary sources in other secondary sources, unlike our focus on quotation from primary sources. Finally, as described in more detail in Section 2.2 and Appendix A10, RELiC differs significantly from existing NLP and IR retrieval datasets in domain, linguistic complexity, and query length.

6 Conclusion

In this work, we introduce the task of *literary evidence retrieval* and an accompanying dataset, RELiC. We find that direct quotation of primary sources in literary analysis is most commonly used as evidence for literary claims or arguments. We build a dense retriever model for the task of retrieving quotations from claims and show that while it significantly outperforms baselines, there still remains large room for improvement, as evidenced by human performance on a proxy task. An important direction for future work on RELiC is to build better models of *primary sources* that integrate narrative and discourse structure into the candidate representations instead of computing them out-of-context as in our current retriever. Furthermore, integrating RELiC models into actual tools that humanities scholars find useful for evidence retrieval is an area of high potential impact.

7 Ethical Considerations

We acknowledge that the group of authors from whom we selected primary sources lacks diversity because we selected from among digitized, public domain sources in the Western literary canon, which is heavily biased towards white, male writers. We made this choice because there are relatively few primary sources in the public domain that are written by minority authors and also have substantial amounts of literary analysis written about them. We hope that our data collection approach will be followed by those with access to copyrighted texts in an effort to collect a more diverse dataset. The experiments involving humans were IRB-approved.

References

Eneko Agirre, Carmen Banea, Daniel Cer, Mona Diab, Aitor Gonzalez-Agirre, Rada Mihalcea, German Rigau, and Janyce Wiebe. 2016. SemEval-2016 task 1: Semantic textual similarity, monolingual and cross-lingual evaluation. In *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*.

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.

David Bamman, Sejal Popat, and Sheng Shen. 2019. An annotated dataset of literary entities. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 2138–2144.

Marc Bertin, Iana Atanassova, Cassidy R Sugimoto, and Vincent Lariviere. 2016. The linguistic patterns and rhetorical structure of citation context: an approach using n-grams. *Scientometrics*, 109(3):1417–1434.

Bernard Blackstone. 1972. *Virginia Woolf: A Commentary*. London.

Alexander Bondarenko, Maik Fröbe, Meriem Beloucif, Lukas Gienapp, Yamen Ajjour, Alexander Panchenko, Chris Biemann, Benno Stein, Henning Wachsmuth, Martin Potthast, and Matthias Hagen. 2020. Overview of Touché 2020: Argument Retrieval. In *Working Notes Papers of the CLEF 2020 Evaluation Labs*, volume 2696 of *CEUR Workshop Proceedings*.

Vera Boteva, Demian Gholipour, Artem Sokolov, and Stefan Riezler. 2016. A full-text learning to rank dataset for medical information retrieval. In *Proceedings of the 38th European Conference on Information Retrieval (ECIR 2016)*, pages 716–722.

Snigdha Chaturvedi, Shashank Srivastava, Hal Daume III, and Chris Dyer. 2016. Modeling evolving relationships between characters in literary novels. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Danqi Chen, Adam Fisch, Jason Weston, and Antoine Bordes. 2017. Reading Wikipedia to answer open-domain questions. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1870–1879, Vancouver, Canada. Association for Computational Linguistics.

Danqi Chen and Wen-tau Yih. 2020. Open-domain question answering. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics: Tutorial Abstracts*, pages 34–37, Online. Association for Computational Linguistics.

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*.

Arman Cohan, Sergey Feldman, Iz Beltagy, Doug Downey, and Daniel Weld. 2020. SPECTER: Document-level representation learning using citation-informed transformers. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2270–2282, Online. Association for Computational Linguistics.

Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. Climate-fever: A dataset for verification of real-world climate claims.

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.

Ruth Finnegan. 2011. *Why do we quote?: the culture and history of quotation*. Open Book Publishers.

Gerald Graff, Cathy Birkenstein, and Cyndee Maxwell. 2014. *They say, I say: The moves that matter in academic writing*. Gildan Audio.

Peter F. Grav. 2019. Harnessing Sources in the Humanities: A Corpus-based Investigation of Citation Practices in English Literary Studies. *Discourse and Writing/Rédactologie*, 29:24–50.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. REALM: Retrieval-augmented language model pre-training. In *Proceedings of the International Conference of Machine Learning*.

Arnold M. Hartstein. 1985. Myth and History in Moby Dick. *American Transcendental Quarterly*, 57:31–43.

Faegheh Hasibi, Fedor Nikolaev, Chenyan Xiong, Krisztian Balog, Svein Erik Bratsberg, Alexander Kotov, and Jamie Callan. 2017. Dbpedia-entity v2: A test collection for entity search. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '17, pages 1265–1268. ACM.

Evelyn Thomas Helmick. 1968. Myth in the Works of Willa Cather. *Midcontinent American Studies Journal*, 9(2):63–69.

Mark M. Hennelly, Jr. 1983. The Eyes Have It. *Jane Austen: New Perspectives*, 3.

Doris Hoogeveen, Karin M Verspoor, and Timothy Baldwin. 2015. Cqadupstack: A benchmark data set for community question-answering research. In *Proceedings of the 20th Australasian document computing symposium*, pages 1–8.

Zhengbao Jiang, Frank F. Xu, Jun Araki, and Graham Neubig. 2020. How Can We Know What Language Models Know? *Transactions of the Association for Computational Linguistics*, 8:423–438.

Chris Kamphuis, Arjen P de Vries, Leonid Boytsov, and Jimmy Lin. 2020. Which bm25 do you mean? a large-scale reproducibility study of scoring variants. In *European Conference on Information Retrieval*, pages 28–34. Springer.

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*.

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*.

Colin Legum. 1972. *Congo Disaster*. Peguin Books Ltd.

Shengbo Liu, Chaomei Chen, Kun Ding, Bo Wang, Kan Xu, and Yuan Lin. 2014. Literature retrieval based on citation context. *Scientometrics*, 101(2):1293–1307.

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*.

Ansel MacLaughlin and David A Smith. 2021. Content-based models of quotation. In *Proceedings of the European Chapter of the Association for Computational Linguistics*, pages 2296–2314.

Deborah L. Madsen. 2000. *Feminist Theory and Literary Practice*. London.

Hena Maes-Jelinek. 1970. *Criticism of Society in the English Novel Between the Wars*. Paris.

Macedo Maia, Siegfried Handschuh, André Freitas, Brian Davis, Ross McDermott, Manel Zarrouk, and Alexandra Balahur. 2018. Www'18 open challenge: Financial opinion mining and question answering. In *Companion Proceedings of the The Web Conference 2018*, WWW '18, page 1941–1942, Republic and Canton of Geneva, CHE. International World Wide Web Conferences Steering Committee.

Muriel Agnes Bussell Masefield. 1967. *Women Novelists from Fanny Burney to George Eliot*. Books for Libraries Press, New York.

Neil McEwan. 1986. *Style in English prose.* York handbooks. Longman, Harlow, Essex.

Stella McNichol. 1990. *Virginia Woolf and the Poetry of Fiction*. Routledge.

Tri Nguyen, Mir Rosenberg, Xia Song, Jianfeng Gao, Saurabh Tiwary, Rangan Majumder, and Li Deng. 2016. Ms marco: A human generated machine reading comprehension dataset. *choice*, 2640:660.

John O'Connor. 1982. Citing statements: Computer recognition and use to improve retrieval. *Information Processing & Management*, 18(3):125–131.

Sean Papay and Sebastian Padó. 2020. RiQuA: A corpus of rich quotation annotation for English literary text. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 835–841, Marseille, France. European Language Resources Association.

Bernard J. Paris. 1978. *Character and Conflict in Jane Austen's Novels: A Psychological Approach*. Wayne State University Press, Detroit.

Kenneth Parker. 1985. The Revelation of Caliban: 'The Black Presence' in the Classroom. In David Dabydeen, editor, *The Black Presence in English Literature*. Manchester University Press.

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.

J.D. Porter. 2018. Literary Lab Pamphlet 17: Popularity/Prestige. Pamphlet.

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.

Matthew Sims, Jong Ho Park, and David Bamman. 2019. Literary event detection. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3623–3634.

Ian Soboroff, Shudong Huang, and Donna Harman. 2018. Trec 2018 news track overview. In *TREC*.

Axel Suarez, Dyaa Albakour, David Corney, Miguel Martinez, and Jose Esquivel. 2018. A data collection for evaluating the retrieval of related tweets to news articles. In *40th European Conference on Information Retrieval Research (ECIR 2018), Grenoble, France, March, 2018.*, pages 780–786.

Nandan Thakur, Nils Reimers, Andreas Rücklé, Abhishek Srivastava, and Iryna Gurevych. 2021. Beir: A heterogenous benchmark for zero-shot evaluation of information retrieval models. *arXiv preprint arXiv:2104.08663*.

Jennifer Wolfe Thompson. 2002. The death of the scholarly monograph in the humanities? citation patterns in literary scholarship. *Libri*, 52.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

George Tsatsaronis, Georgios Balikas, Prodromos Malakasiotis, Ioannis Partalas, Matthias Zschunke, Michael R Alvers, Dirk Weissenborn, Anastasia Krithara, Sergios Petridis, Dimitris Polychronopoulos, et al. 2015. An overview of the bioasq large-scale biomedical semantic indexing and question answering competition. *BMC bioinformatics*, 16(1):138.

Ellen Voorhees. 2005. Overview of the trec 2004 robust retrieval track.

Ellen Voorhees, Tasmeer Alam, Steven Bedrick, Dina Demner-Fushman, William R Hersh, Kyle Lo, Kirk Roberts, Ian Soboroff, and Lucy Lu Wang. 2021. Trec-covid: constructing a pandemic information retrieval test collection. In *ACM SIGIR Forum*, volume 54, pages 1–12. ACM New York, NY, USA.

Henning Wachsmuth, Shahbaz Syed, and Benno Stein. 2018. Retrieval of the best counterargument without prior topic knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 241–251. Association for Computational Linguistics.

David Wadden, Shanchuan Lin, Kyle Lo, Lucy Lu Wang, Madeleine van Zuylen, Arman Cohan, and Hannaneh Hajishirzi. 2020. Fact or fiction: Verifying scientific claims. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7534–7550, Online. Association for Computational Linguistics.

John Wieting, Taylor Berg-Kirkpatrick, Kevin Gimpel, and Graham Neubig. 2019. Beyond BLEU:training neural machine translation with semantic similarity. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4344–4355, Florence, Italy. Association for Computational Linguistics.

John Wieting and Kevin Gimpel. 2018. ParaNMT-50M: Pushing the limits of paraphrastic sentence embeddings with millions of machine translations. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 451–462, Melbourne, Australia. Association for Computational Linguistics.

Brian Wilkie. 1992. Jane Austen: Amore and Amoralism. *Journal of English and German Philology*, 91(1):529–555.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-art natural language processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 38–45, Online. Association for Computational Linguistics.

James Woodress. 1975. Willa Cather: The World and the Parish. *Architectural Association Quarterly*, 7:51–59.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Appendices for "ReLiC: Retrieving Evidence from Literature in Context"

A Dataset Collection & Statistics

Filtering secondary sources: The HathiTrust is not exclusively a repository of literary analysis, and we observe that many matching quotes come from different editions of a primary source, writing manuals, and even advertisements. Because we are seeking only scholarly work that directly analyzes the quoted sentences, we performed a combination of manual and automatic filtering to remove such extraneous matches. For each primary source, we first aggregate all secondary sources matches by the their unique HathiTrust-assigned identifier. From manual inspection of the secondary source titles, most sources that quote a particular literary work only once or twice are not likely to be literary scholarship, while sources with hundreds of matches are almost always a different edition of the primary source itself. For each primary source, we create upper and lower thresholds for number of matches, discarding sources that fall outside of these bounds. Additionally, we discard secondary sources whose titles contain the words "dictionary", "anthology", "encyclopedia," and others that indicate that a secondary source is not literary scholarship.

Preprocessing: After the above filtering, we identified and removed all non-English secondary sources using langid,[18] a Python tool for language identification. Next, because the secondary source texts in the HathiTrust are digitized via OCR, various artifacts appear throughout the pages we download. Some of these, such as citations that include the page number of primary source quotes, allow models trained on our task to "cheat" to identify the proper quote (see Table A1), necessitating their removal. Using a pattern-matching approach, we eliminate the most pervasive: in-line citations, headers, footers, and word breaks. Finally, we apply sentence tokenization in order to standardize the length of preceding and subsequent context windows for the final dataset. Specifically, we feed the preprocessed text through spaCy's[19] dependency parser-based sentence segmenter on the cleaned text. The default segmenter in spaCy is modified to use ellipses, colons, and semicolons as custom sentence boundaries, based on the observation that literary scholars often only quote part of what would typically be defined as a sentence (Table A2).

Raw text from HathiTrust:
The prejudice in these same eyes, however, keeps them "less clear-sighted" **(p. 149)** to Bingley's feelings for Jane and totally closed to the real **worth- lessness** of Wickham and worth of Darcy. When Jane's letter reporting **196 Mark M. Hennelly, Jr.** Lydia's disappearance with Wickham confirms Darcy's earlier indictment of him, though, Elizabeth's "eyes were opened to his real character" **(p. 277)**.

Table A1: An analysis of Jane Austen's *Pride and Prejudice* from Hennelly (1983) that contains artifacts (bold) such as citations and page numbers that we remove during preprocessing.

Quoted span in context of literary analysis: Edna tries to discuss this issue of possession versus self-possession with Madame Ratignolle but to no avail; '**the two women did not appear to understand each other or to be talking the same language.**' Madame Ratignolle cannot comprehend that there might be something more that a mother could sacrifice for her children beyond her life...
Quote in original context from The Awakening: Edna had once told Madame Ratignolle that she would never sacrifice herself for her children, or for any one. Then had followed a rather heated argument; **the two women did not appear to understand each other or to be talking the same language.** Edna tried to appease her friend, to explain.

Table A2: An analysis of Kate Chopin's *The Awakening* from Madsen (2000) that quotes part of a sentence (following a semi-colon) from the primary source. We detect such partial matches during preprocessing.

Identifying quoted sentences: As previously mentioned, HathiTrust does not provide the exact indices corresponding to the primary source quote. As such, we identify which secondary source sentences (from the output of the sentence tokenizer) include quotes from primary source works using RapidFuzz, [20] a fuzzy string match library, with the QRatio metric and a score threshold of 80.0. Fuzzy match is essential for detecting quotes with OCR mistakes or with author modifications; in Appendix Table A3, for instance, the author adds clarification [the natives] and omits "he would say" when citing two sentences from Joseph Conrad's *Heart of Darkness*. Once a fuzzy match is identified in a secondary source document, we replace it with its corresponding primary source sentence.

[18] https://github.com/saffsd/langid.py

[19] https://spacy.io/

[20] https://github.com/maxbachmann/RapidFuzz

Secondary source material: Kurtz’s credo, like his royal employer’s, was a simple one. 1. “You show them [the natives] you have in you something that is really profitable, and then there will be no limits to the recognition of your ability. 2. Of course you must take care of the motives—right motives—always.” Kurtz dies screaming: "The Horror! The Horror!" Leopold, so far as one knows, died more peacefully (Legum, 1972).
Window in RELiC with standardized quote: Kurtz’s credo, like his royal employer’s, was a simple one. **‘You show them you have in you something that is really profitable, and then there will be no limits to the recognition of your ability,’ he would say. ‘Of course you must take care of the motives—right motives—always.’** Kurtz dies screaming: "The Horror! The Horror!" Leopold, so far as one knows, died more peacefully.

Table A3: This example demonstrates the necessity of fuzzy match and block quote identification. Consecutive sentences are quoted and one is slightly modified from its original form in the primary source.

Identifying block quotes: While we query HathiTrust at a sentence level, many of the returned results are actually *block quotes* in which multiple contiguous sentences from the primary source are quoted. Correct identification of these block quotes is integral to the quality of our dataset and formulated task: if the preceding or subsequent context contains part of the quoted span, our evidence retrieval task becomes trivial because part of the answer exists in the input. In our approach, if the fuzzy match yields consecutive matches in secondary source documents for sentences that also appear consecutively in the primary source, we concatenate them together and consider them a single block quote.

Handling ellipses: One prevalent technique for direct quotation in literary analysis is the use of ellipses to condense primary source material. As our fuzzy match method still falls short in detecting block quotes that contain ellipses, we implement an additional method for insuring that block quotes are properly delineated. Once the fuzzy match approach fails to identify any more consecutively quoted sentences in a secondary source, we continue to search for matches adjacent to the block quote using the Longest Common Substring (LCS) metric. If a block-quote-adjacent sentence in the secondary source shares an LCS of 15 or more characters with the block-quote-adjacent sentence in the primary source, this is considered a match and concatenated with the block quote (see Appendix A.1 for an example).

A.1 LCS example

For example, in Parker (1985), Kenneth Parker cites a passage from Joseph Conrad’s *Heart of Darkness*: “The narrator, Marlow, informs us, approvingly:...**I met a white man, in such an unexpected elegance of get-up that in the first moment I took him for a sort of vision.** I saw a high starched collar, white cuffs, a light alpaca jacket, snowy trousers, a clean necktie, and varnished boots.” Fuzzy match alone is insufficient for detecting the first sentence in this block quote that contains an ellipse in place of primary source text. With our LCS approach, we are able to replace the first sentence of block quote above with “**When near the buildings I met a white man, in such an unexpected elegance of get-up that in the first moment I took him for a sort of vision.**”

A.2 Noise when standardizing quotes:

In a small number of cases, our quote standardization process removes important context. For example, the analysis of Maes-Jelinek (1970) quotes a sentence from D.H. Lawrence’s *The Rainbow* as “As to Will, **his intimate life was so violently active, that it set another man free in him.**”. After standardization, the example in our dataset becomes “**His intimate life was so violently active, that it set another man free in him.**”, dropping the critical “As to Will” necessary for the integration of the quote in the surrounding analysis.

Model-predicted quotes are sometimes as valid as the gold quote: Human raters also identify cases in which multiple quotes appear to be appropriate evidence for a literary claim, which illustrate the model’s potential in helping humanities scholars find evidence. In Table A4, both model and experts failed to identify the correct quote that both depicts Elizabeth’s “discomfiture” and has a “Greek ring to it:” “Till this moment I never knew myself.” However, the experts all selected the model’s second ranked choice which mentions Elizabeth’s “anger” at “herself.” This quote also shows Elizabeth’s displeasure while referring to the Greek idea of self.

Window of secondary source analysis:
For example, Elizabeth's anger with herself, after reading Darcy's letter, is couched largely in the vocabulary of rectifiable intellectual error"blind, partial, prejudiced, absurd, and the like-rather than in the relentless, coercive vocabulary of moral contrition. Her discomfiture, though profound, has a Greek ring to it: **Till this moment I never knew myself.** Heuristically, the distinction between moral and other spheres of value throws light also on other Austen novels that we can only glance at here (Wilkie, 1992).
Best model's top ranked candidate:
that loss of virtue in a female is irretrievable;
Best model's second ranked candidate
but when she considered how unjustly she had condemned and upbraided him, her anger was turned against herself;

Table A4: The model ranked the correct quote outside of the top ten percent of 5,278 candidates, but all 3 domain experts selected the model's second ranked candidate over the ground-truth quote.

A.3 More dataset statistics

Each primary source has relevant windows from an average of 157 unique secondary sources, and an average of 12.58% of the sentences in each primary source are quoted in secondary sources. On average, each primary source has 673 corresponding windows in our dataset, and each secondary source produced an average of 6 windows. The top three secondary sources for our dataset (Appendix Table A5) are books focusing on specific authors (Dickens, Woolf). Figure 2 shows the distribution of quote lengths in RELiC, suggesting that successful models will have to learn to understand both sentence and block quotes in context.

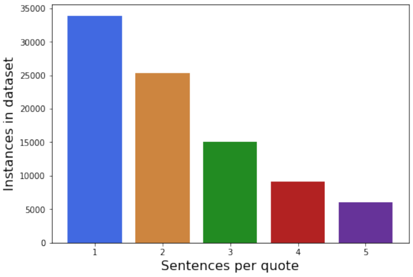


Figure 2: Distribution of RELiC quote lengths.

B Best model detailed results

Candidate length does not significantly affect model performance: We observe in Table A9 that the length of the ground-truth quote and the candidates does not significantly impact model performance — for a fixed k, model performance is within 10% for any candidate length. Model performance is slightly worse for longer candidates of length 4 or 5, and for the shortest single sentence contexts (possibly due to under-specification).

Title of secondary source	**#**
Dickens and Thackeray: Punishment and Forgiveness	138
Virginia Woolf: Strategist of Language	118
The Houses that James Built, and Other Literary Studies	108
Twentieth-century Literary Criticism, v. 32 1989	105

Table A5: The top secondary sources in RELiC.

Claim-supporting evidence	The relationship between Alexandra and the earth is an intensely personal one: **For the first time, perhaps, since that land emerged from the waters of geologic ages, a human face was set toward it with love and yearning...** The religious connotations of the more lyrical descriptions of the land prepare us for the emergence of Alexandra as its goddess (Helmick, 1968).
Paraphrase-supporting evidence	O Pioneers! is the story of a Swedish immigrant, Alexandra Bergson, who some to Nebraska with her parents when she is young. Her father dies, and she has to take over the farm and look after her younger brothers. Her courage, vision, and energy bring life and civilization to the wilderness. As Alexandra faces the future after her father's death, Willa Cather writes: **For the first time, perhaps, since that land emerged from the waters of geologic ages, a human face was set toward it with love and yearning.** The history of every country begins in the heart of a man or a woman. Alexandra succeeds in taming the wild land, and after a heaping measure of material success and personal tragedy, she faces the future calmly. At the end of the novel Alexandra's childhood lover comes back to her, but the land remains the ultimate heroine (Woodress, 1975).

Table A6: The most quoted primary source sentence in RELiC, from Willa Cather's *O Pioneers!*, is quoted 49 times.

		Training Set		
Year	**Title**	**Author (Translator)**	**Type**	**Language**
1811	Sense and Sensibility	Jane Austen	novel	English
1814	Mansfield Park	Jane Austen	novel	English
1818	Frankenstein	Mary Shelley	novel	English
1837	The Pickwick Papers	Charles Dickens	novel	English
1839	Nicholas Nickleby	Charles Dickens	novel	English
1839	Oliver Twist	Charles Dickens	novel	English
1843	A Christmas Carol	Charles Dickens	novella	English
1844	Martin Chuzzlewit	Charles Dickens	novel	English
1847	Jane Eyre	Charlotte Brontë	novel	English
1847	Wuthering Heights	Emily Brontë	novel	English
1850	David Copperfield	Charles Dickens	novel	English
1850	The Scarlet Letter	Nathaniel Hawthorn	novel	English
1851	Moby Dick	Herman Melville	novel	English
1852	Uncle Tom's Cabin	Harriet Beecher Stowe	novel	English
1853	Bleak House	Charles Dickens	novel	English
1856	Madame Bovary	Gustave Flaubert (Eleanor Marx-Avelin)	novel	French
1857	Little Dorrit	Charles Dickens	novel	English
1859	Adam Bede	George Eliot	novel	English
1861	Great Expectations	Charles Dickens	novel	English
1865	Alice's Adventures in Wonderland	Lewis Carroll	novel	English
1866	Crime and Punishment	Fyodor Dostoevsky (Garnett)	novel	Russian
1867	War and Peace	Leo Tolstoy (Constance Garnett)	novel	Russian
1871	Middlemarch	George Eliot	novel	English
1878	Daisy Miller	Henry James	novella	English
1880	Brothers Karamazov	Fyodor Dostoevsky (Garnett)	novel	Russian
1884	Adventures of Huckleberry Finn	Mark Twain	novel	English
1890	The Picture of Dorian Gray	Oscar Wilde	novel	English
1893	Maggie: A Girl of the Streets	Stephen Crane	novella	English
1895	The Red Badge of Courage	Stephen Crane	novel	English
1892	Iola Leroy	Frances Harper	novel	English
1897	What Maisie Knew	Henry James	novel	English
1898	The Turn of the Screw	Henry James	novella	English
1899	The Awakening	Kate Chopin	novel	English
1900	Sister Carrie	Theodore Dreiser	novel	English
1902	The Sport of the Gods	Paul Laurence Dunbar	novel	English
1903	The Ambassadors	Henry James	novel	English
1903	The Call of the Wild	Jack London	novel	English
1903	The Souls of Black Folk	W. E. B. Du Bois	collection (nonfiction)	English
1905	House of Mirth	Edith Wharton	novel	English
1913	O Pioneers!	Willa Cather	novel	English
1916	A Portrait of the Artist as a Young Man	James Joyce	novel	English
1915	The Rainbow	D. H. Lawrence	novel	English
1918	My Antonia	Willa Cather	novel	English
1920	The Age of Innocence	Edith Wharton	novel	English
1920	This Side of Paradise	F. Scott Fitzgerald	novel	English
1922	Jacob's Room	Virginia Woolf	novel	English
1922	Swann's Way	Marcel Proust (C. K. Scott Moncrieff)	novel	French
1925	An American Tragedy	Theodore Dreiser	novel	English
1925	Mrs Dalloway	Virginia Woolf	novel	English
1927	To the Lighthouse	Virginia Woolf	novel	English
1928	Lady Chatterly's Lover	D. H. Lawrence	novel	English
1932	Brave New World	Aldous Huxley	novel	English
1936	Gone with the Wind	Margaret Mitchell	novel	English
1931	The Waves	Virginia Woolf	novel	English
1945	Animal Farm	George Orwell	novel	English
1949	1984	George Orwell	novel	English

Table A7: Primary sources from which training set windows were derived.

Validation Set

Year	Title	Author (Translator)	Type	Language
1815	Emma	Jane Austen	novel	English
1817	Northanger Abbey	Jane Austen	novel	English
1830	The Red and the Black =	Stendhal (Horace B. Samuel)	novel	French
1841	Barnaby Rudge	Charles Dickens	novel	English
1847	Agnes Grey	Anne Brontë	novel	English
1848	The Tenant of Wildfell Hall	Anne Brontë	novel	English
1854	Hard Times	Charles Dickens	novel	English
1859	A Tale of Two Cities	Charles Dickens	novel	English
1869	Little Women	Louisa May Alcott	novel	English
1877	Anna Karenina	Leo Tolstoy (Garnett)	novel	Russian
1883	Treasure Island	Robert Louis Stevenson	novel	English
1898	The War of the Worlds	H. G. Wells	novel	English
1911	Ethan Frome	Edith Wharton	novel	English
1915	The Song of the Lark	Willa Cather	novel	English
1920	Main Street	Sinclair Lewis	novel	English
1922	Babbitt	Sinclair Lewis	novel	English
1922	The Garden Party and Other Stories	Katherine Mansfield	collection (fiction)	English
1925	Arrowsmith	Sinclair Lewis	novel	English

Test Set

Year	Title	Author (Translator)	Type	Language
1813	Pride and Prejudice	Jane Austen	novel	English
1817	Persuasion	Jane Austen	novel	English
1899	Heart of Darkness	Joseph Conrad	novella	English
1925	The Great Gatsby	F. Scott Fitzgerald	novel	English
1934	Tender Is the Night	F. Scott Fitzgerald	novel	English

Table A8: Primary sources from which validation and test set windows were derived.

# of sents	# instances	recall@k						mean rank	avg. # candidates
in quote		1	3	5	10	50	100		
1	3682	9.4	17.4	22.1	29.0	47.5	56.5	456.6	5057.5
2	2478	10.9	22.6	28.4	36.6	56.5	65.9	315.7	5145.1
3	1408	10.2	22.3	28.8	38.6	56.3	64.0	310.6	5011.7
4	894	7.7	19.8	26.7	35.0	58.6	66.8	301.6	4944.0
5	572	7.0	17.5	25.0	34.4	54.9	65.9	314.9	4712.2

Table A9: A breakdown of performance by quote length in sentences of the performance of our best model, the dense retriever with 4 context sentences on each side. All numbers are on the test set of RELiC.

Split (→)					**Train**	**Dev**	**Test**			**Avg. Word Lengths**	
Task (↓)	**Domain** (↓)	**Dataset** (↓)	**Title**	**Relevancy**	**#Pairs**	**#Query**	**#Query**	**#Corpus**	**Avg. D / Q**	**Query**	**Document**
Passage-Retrieval	Misc.	MS MARCO (Nguyen et al., 2016)	✗	Binary	532,761	—-	6,980	8,841,823	1.1	5.96	55.98
Bio-Medical Information Retrieval (IR)	Bio-Medical	TREC-COVID (Voorhees et al., 2021)	✓	3-level	—-	—-	50	171,332	493.5	10.60	160.77
	Bio-Medical	NFCorpus (Boteva et al., 2016)	✓	3-level	110,575	324	323	3,633	38.2	3.30	232.26
	Bio-Medical	BioASQ (Tsatsaronis et al., 2015)	✓	Binary	32,916	—-	500	14,914,602	4.7	8.05	202.61
Question Answering (QA)	Wikipedia	NQ (Kwiatkowski et al., 2019)	✓	Binary	132,803	—-	3,452	2,681,468	1.2	9.16	78.88
	Wikipedia	HotpotQA (Yang et al., 2018)	✓	Binary	170,000	5,447	7,405	5,233,329	2.0	17.61	46.30
	Finance	FiQA-2018 (Maia et al., 2018)	✗	Binary	14,166	500	648	57,638	2.6	10.77	132.32
Tweet-Retrieval	Twitter	Signal-1M (RT) (Suarez et al., 2018)	✗	3-level	—-	—-	97	2,866,316	19.6	9.30	13.93
News Retrieval	News	TREC-NEWS (Soboroff et al., 2018)	✓	5-level	—-	—-	57	594,977	19.6	11.14	634.79
	News	Robust04 (Voorhees, 2005)	✗	3-level	—-	—-	249	528,155	69.9	15.27	466.40
Argument Retrieval	Misc.	ArguAna (Wachsmuth et al., 2018)	✓	Binary	—-	—-	1,406	8,674	1.0	**192.98**	166.80
	Misc.	Touché-2020 (Bondarenko et al., 2020)	✓	3-level	—-	—-	49	382,545	19.0	6.55	292.37
Duplicate-Question Retrieval	StackEx.	CQADupStack (Hoogeveen et al., 2015)	✓	Binary	—-	—-	13,145	457,199	1.4	8.59	129.09
	Quora	Quora	✗	Binary	—-	5,000	10,000	522,931	1.6	9.53	11.44
Entity-Retrieval	Wikipedia	DBPedia (Hasibi et al., 2017)	✓	3-level	—-	67	400	4,635,922	38.2	5.39	49.68
Citation-Prediction	Scientific	SCIDOCS (Cohan et al., 2020)	✓	Binary	—-	—-	1,000	25,657	4.9	9.38	176.19
Fact Checking	Wikipedia	FEVER (Thorne et al., 2018)	✓	Binary	140,085	6,666	6,666	5,416,568	1.2	8.13	84.76
	Wikipedia	Climate-FEVER (Diggelmann et al., 2020)	✓	Binary	—-	—-	1,535	5,416,593	3.0	20.13	84.76
	Scientific	SciFact (Wadden et al., 2020)	✓	Binary	920	—-	300	5,183	1.1	12.37	213.63
Literary evidence retrieval	**Literature**	ReLiC (this work)	✗	Binary	71395	9036	9034	5041	1.0	**154.1**	45.5

Table A10: A comparison between datasets in the BEIR benchmark and our ReLiC dataset. Ours is the first retrieval dataset in the literary domain, formulating a new task of literary evidence retrieval.