It Is a Truth Individually Acknowledged: Cross-references On Demand

Piper Vasicek
Brigham Young University
Provo, UT
piper.vasicek@byu.edu

Courtni Byun
Brigham Young University
Provo, UT
courtni.byun@gmail.com

Kevin Seppi
Brigham Young University
Provo, UT
kseppi@cs.byu.edu

Abstract

Cross-references link source passages of text to other passages that elucidate the source passage in some way and can deepen human understanding. Despite their usefulness, however, good cross-references are hard to find, and extensive sets of cross-references only exist for the few most highly studied books such as the Bible, for which scholars have been collecting cross-references for hundreds of years.

Therefore, we propose a new task: generate cross-references for user-selected text on demand. We define a metric, *coverage*, to evaluate task performance. We adapt several models to generate cross-references, including an Anchor Words topic model, SBERT Sentence-Transformers, and ChatGPT, and evaluate their *coverage* in both English and German on existing cross-reference datasets. While ChatGPT outperforms other models on these datasets, this is likely due to data contamination. We hand-evaluate performance on the well-known works of Jane Austen and a less-known science fiction series *Sons of the Starfarers* by Joe Vasicek, finding that ChatGPT does not perform as well on these works; sentence embeddings perform best. We experiment with newer LLMs and large context windows, and suggest that future work should focus on deploying cross-references on-demand with readers to determine their effectiveness in the wild.

1 Introduction

A cross-reference is a connection between a source passage of text and another passage with bearing on the source passage. A cross-reference may contextualize, define, reinforce, restate, or even rebut the source passage, but a good cross-reference always elucidates the source passage in some way.

Good cross-references are hard to find because it takes familiarity with the text as a whole and a focused search through the text to find connections. While scholar-created sets of cross-references are rare, the concept of self-reference within a text is ubiquitous. Consider the following two conversations from the film, "The Emperor's New Groove" (Dindal, 2000):

> Yzma: Fired? What do you mean, fired?
>
> Kuzco: Um, how else can I say it? You're being let go. Your department is being downsized. You're part of an outplacement. We're going in a different direction. We're not picking up your option. Take your pick. I got more.
>
> <later>
>
> Yzma: Just think of it as you're being let go, that your life's going in a different direction, that your body's part of a permanent outplacement.
>
> Kronk: Hey, that's kind of like what he said to you when you got fired.

This type of cross-reference, known as a "callback" in screenwriting, reuses or paraphrases a previous line. This type of cross-reference is easy to identify, since its express purpose is to *be* identified by the audience and to recontextualize the earlier line, often for laughs.

Most cross-references are subtler and more difficult to find, especially when not intended as cross-references by the writer(s).[1] Subtle cross-references can be found in one of two ways:

1. While reading, the cross-referencer happens to remember another connected passage–like Kronk recognizing the callback in "The Emperor's New Groove"

2. The cross-referencer performs a focused search through the text specifically looking for connections to a source passage. When applied to each passage in a text, this translates to a complexity of $O(n^2)$.

[1] See Appendix A for examples of subtler cross-references.

Proceedings of the 4th International Conference on Natural Language Processing for Digital Humanities, pages 63–74
November 16, 2024.

For print books, all desired cross-references must be discovered prior to printing since they cannot be added later. Electronic texts make it possible to discover cross-references dynamically instead. We therefore propose a fundamentally new task: provide cross-references on demand directly to readers for any passage they select.

Our contributions are as follows:

- Define the task: cross-references on demand.
- Define a metric to evaluate the task.
- Adapt several models to accomplish the task.
- Evaluate the model performance on three works in two languages, English and German.

We discuss prior work with cross-references in Section 2, define the task, cross-references on demand, in Section 3, discuss our methodology in Section 4, and present results in Section 5.

2 Prior Work

Before the advent of computers, all cross-reference sets were necessarily compiled manually. The more than 500,000 biblical cross-references originally published as the Treasury of Scripture Knowledge around 1830, were collected by "many authors . . . over centuries" (Morton, 2010).

In 1973, the Church of Jesus Christ of Latter-day Saints formed a committee to create a new edition of the Bible for use by their congregations. This committee and "hundreds of workers"–mostly volunteers–took six years to create a cross-reference set for the new edition, despite evaluating existing Bible cross-references and using concordance software (Anderson, 1979).

The labor-intensive nature of these manual cross-referencing projects highlights the reasons why scholarly cross-reference sets are so rare.

Lund et al. (2019) investigated reducing the cost to create a cross-reference set by using topic modeling to suggest cross-references and crowdsourcing to evaluate them. However, creating a *set* of cross-references will be labor-intensive no matter how much technology improves. Even the Qur'an, although revered, was not extensively cross-referenced until 2022, the culmination of a project that took a decade to complete despite access to modern technology (Sirry, 2022).

There are over 130,000,000 books in publication (Taycher, 2010). It would be impossible to create sets of cross-references for all of them, but it might be possible to cross-reference them on demand instead.

Source Passage: **"It is a truth universally acknowledged, that a single man in possession of a good fortune, must be in want of a wife."** – Pride and Prejudice, ch.1

ChatGPT: **"Happy families are all alike; every unhappy family is unhappy in its own way."** – Anna Karenina, ch.1

Sentence Embeddings: **"But there certainly are not so many men of large fortune in the world as there are pretty women to deserve them."** – Mansfield Park, ch.1

Topic Modeling: **"It must make you better satisfied that your other four are single."** – Pride and Prejudice, ch.53

Figure 1: The best examples from the top five suggestions generated by our best-performing models on the first sentence of Pride and Prejudice. ChatGPT suggests first lines from various literary works.

3 Defining Cross-references On Demand

What would cross-references on-demand look like? A reader selects a desired source passage, and a model returns suggested cross-references; see Figure 1 for an example of what this might look like. The difficulty of this task lies in finding *good* cross-references and evaluating cross-reference quality.

A good cross-reference enhances the reader's understanding of the source passage. However readers are not monolithic, so we cannot expect every suggested cross-reference to be a good cross-reference for every reader. Even scholar-produced sets–which should only contain good cross-references–contain cross-references that some readers find unhelpful. If experts cannot produce universally good cross-references, we do not expect models to do so. Instead, we consider a model successful if a reader finds a satisfactory number of good cross-references in a relatively small number of suggestions.

To encapsulate this concept, we define the following metric which we refer to as *coverage*:

$$C = \frac{\sum_{d=1}^{n} f(d)}{n} \tag{1}$$

$$f(d) = \begin{cases} 1, & \text{if } |\{x_{d1}, x_{d2}, \ldots, x_{di}\} \cap G| \geq t \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

where C is the coverage, d is a source passage; n is the total number of source passages for which we evaluate suggested cross-references; x is a suggested cross-reference from the model; i is the number of cross-references suggested, G is the set of good cross-references, and t is the number of cross-references required to satisfy the reader. It is similar to precision@k, but instead of calculating the ratio of good suggestions to total suggestions, we determine whether or not critical mass of good suggestions has been achieved.

Coverage can be calculated for the entire text, or, conveniently, for a sample of the text with the result extrapolated to the rest of the text. Using *coverage* we can now compare the performance of various models to determine which produce the most satisfactory results.

4 Methodology

We consider which models to apply to cross-reference generation in Section 4.1; we discuss the Datasets to which we will apply them in Section 4.2; and we discuss using *coverage* to evaluate model performance in Section 4.3.

4.1 Model Selection

While any number of models could be adapted to cross-references on demand, we choose three to represent them.

Since Lund et al. (2019) applies topic modeling to static cross-reference set creation with good effect, it is logical to adapt their models to our task–see Section 4.1.1. These models rank cross-reference suggestions using topical similarity.

Sentence embeddings are a more mainstream way of comparing semantic similarity; we therefore adapt them as well–see Section 4.1.3.

Finally, it is unclear whether semantic similarity is the most effective way to find good cross-references. The context of the passage–including context external to the work–may prove essential to finding good cross-references. Large language models (LLMs) such as ChatGPT have proven astonishingly good at performing many difficult language tasks (OpenAI, 2022), and ChatGPT has recently been incorporated into a Bible study tool with an option to suggest cross-references (Norton, 2024). Therefore, we experiment with ChatGPT–see Section 4.1.5.

4.1.1 Topic-based Models

We adapt variations of two different topic-modeling-based models from Lund et al. (2019) to use as cross-reference generators and create a third, unique model using randomized topic-words.

We adapt the most successful model from Lund et al. (2019) to serve as a baseline. This model is based on Tandem Anchors (Lund et al., 2017), an extension of the Anchor Words algorithm for topic modeling (Arora et al., 2013) and uses 3000 topics. In order to generate cross-references, the model is given the entire text split up into passages (verses for the Bible and sentences for the other two works). The topics are chosen by randomly selecting a number of passages from the text equal to the desired number of topics. For each of these passages, we take the harmonic mean of the vector representation of all the words in the passage and add a small epsilon of $1e^{-10}$ in each dimension to avoid zero weights. The topics and text are then processed using the Anchor Words algorithm to produce topical weights for each passage. This topic-weight vector for each passage is then compared to the topic-weight vector for the source passage and the most topically similar passages are suggested as cross-references. We adapt this model further by sweeping the space of topics to determine an optimum number–which has not been done previously. We refer to models based on Tandem Anchors with *tand_n*, where "n" represents the number of topics.

The Anchor Words algorithm chooses words as topics based on a variation of the Gram-Schmidt process (Arora et al., 2013). Lund et al. (2019) also employed this model, which performed on par with their Tandem Anchors model. We therefore also adapt this model to our task. Topics are chosen by representing words in a high-dimensional space and attempting to pick n words to use as topics that maximally span that space, very similar to a convex hull. We refer to variations on this model with *gram_n*, where "n" represents the number of topics chosen for the model, and we sweep the topic space since this has not been done previously.

Using 3,000 or more topics for gram_n may not be sensical. For perspective, our datasets have a vocabulary between 9,000 and 20,000 words depending on preprocessing, meaning 3,000 topics is 15-30% of the vocabulary. This was not the expected use of topic modeling or Anchor Words, and it is unclear how well the modified Gram-Schmidt

process for selecting topic words will work at such high saturation. We therefore employ a model that selects n words from the vocabulary at random, and adds a small epsilon of $1e^{-10}$ in each dimension. We refer to this model as *rand_n*, where "n" represents the number of topics chosen for the model.

4.1.2 Preprocessing

For topic-based models–which use a bag-of-words approach–we remove stopwords and employ a stemmer: the Porter stemmer (Porter, 1980) for all English datasets, and the Snowball German stemmer for *Bible-GER* (Porter, 2001).

4.1.3 Sentence Embedding

We employ SBERT sentence-transformer models for sentence embedding (Reimers and Gurevych, 2019, 2020). For English data, we use all-mpnet-base-v2 which currently has the best average performance among available SBERT models. For German data we use paraphrase-multilingual-mpnet-base-v2, which currently performs best among multilingual SBERT models. Similar to the topic-based algorithms, we use the embedding vectors to rank the similarity of each passage to the source passage and suggest the most similar passages.

4.1.4 Distance Metrics

Lund et al. (2019) explored several distance metrics and ultimately evaluated their models using cosine similarity. However, a close viewing of their metric comparison results suggests that cityblock distance performs on par with cosine similarity for the most similar passages evaluated. We therefore include both cosine similarity and cityblock/Manhattan distance to determine semantic similarity.

4.1.5 ChatGPT

Finally, we employ the GPT-3.5-turbo model using default hyperparameters (OpenAI, 2022) to generate cross-references with the following prompt:

> I am reading [TEXT_NAME] and want to find some good cross-references for [REFERENCE_AND/OR_QUOTE]. Can you suggest some good cross-references?

replacing [TEXT_NAME] with the name of the book (e.g. the Bible) and replacing [REFERENCE_AND/OR_QUOTE] with a reference to the passage (only for Bible data), the text of the passage, or both (e.g. "Genesis 1:1 - In the beginning, God created the heavens and the earth.")[2]. We evaluate the cross-references generated, ignoring other generated text. We evaluate the first generation produced for each passage.

4.2 Datasets

We evaluate our models on their ability to suggest good cross-references for three different texts. First, the Bible for which there are expansive, freely available cross-reference sets. Second, for the works of Jane Austen which are widely known and analyzed, but for which there are no existing cross-reference sets. Third, for a science fiction series, The *Sons of the Starfarers*, by Joe Vasicek, which is less well-known and unlikely to be included in ChatGPT's training data.

4.2.1 The Bible

The Bible allows us to evaluate results at scale using existing cross-reference sets. "The Treasury of Scripture Knowledge, Enhanced" (TSKE) (Morton, 2010) is a cross-reference resource based on the original TSK. This set of cross-references is especially useful since it contains an impressive 670,000 cross-references and contains cross-references for 96% of verses. While not the most expansive cross-reference set–there is at least one Biblical cross-reference set that boasts over 900,000 cross-references (Smith, 2016), the TSKE is freely available for download.

`OpenBible.info` has cross-references seeded from the TSK and other open source cross-reference sets and allows users to upvote helpful cross-references and downvote unhelpful ones. We use the OpenBible.info cross references and attached up/downvotes. This yields multiple cross-reference sets, allowing us to simulate readers with different views of what is a good cross-reference. We use the set of cross-references from OpenBible.info that have at least as many upvotes as downvotes–and call this *Open*. We also use the set of cross-references that have a minimum of 5 net upvotes and call this *Open5*.

Beyond the existence of expansive cross-reference sets, the Bible is a useful text because of its many translations. At the time of this writing, there were 3,035 bible translations available in 2,014 languages on `Bible.com`. For English

[2]When prompting ChatGPT with the passage text without the reference, we change the last sentence to "Can you suggest some good *quotes* as cross-references?" Without adjusting it, ChatGPT tends to ask for a reference or describe a scene.

	tand_2000		tand_3000		tand_4000		tand_5000	
	cos	city	cos	city	cos	city	cos	city
TSKE	0.307	0.393	0.328	0.416	0.357	**0.445**	0.353	0.442
Open	0.231	0.299	0.248	0.320	0.272	**0.345**	0.270	0.340
Open5	0.141	0.190	0.146	0.202	0.161	**0.216**	0.158	0.211

Table 1: The results for the tandem anchors model using cosine and cityblock distance metrics. Tandem anchors with 4,000 topics performs best for each dataset using the cityblock distance metric.

	gram_2000		gram_3000		gram_4000		gram_5000		gram_10241	
	cos	city	cos	city	cos	city	cos	city	cos	city
TSKE	0.290	0.327	0.334	0.379	0.381	**0.447**	0.380	0.443	0.379	0.445
Open	0.217	0.248	0.252	0.293	0.289	**0.348**	0.289	0.345	0.288	0.346
Open5	0.108	0.127	0.135	0.160	0.163	0.204	0.161	**0.206**	0.162	0.203

Table 2: Results for the gram_n model using cosine and cityblock metrics. 4,000 topics performs best, except for the Open5 cross-reference set with the cityblock metric for which 5,000 topics performs best.

Bible experiments, we use the text of the English Standard Version (ESV) of the Bible, since that is the version of the Bible on OpenBible.info. In 2022, the Luther Bible was the bestselling German translation of the Bible according to the German Bible Society (Bigl, 2023). We evaluate the cross-reference sets using this translation of the Bible which we refer to as *Bible-GER*.

4.2.2 The Works of Jane Austen

We know good cross-references exist for the Bible, but it is unclear to what extent this is true for other texts. Also, data contamination, specifically the ubuiquity of Bible-cross-references, almost certainly gives ChatGPT an edge when it comes to Bible data. Therefore, Jane Austen's novels are a great dataset to explore ChatGPT's ability to perform on a text with which it is very familiar but for which there are not existing cross-reference sets.

We ask our three best-performing models to suggest 5 cross-references each for 100 randomly selected sentences from the works of Jane Austen. Two of our authors then hand-evaluate each suggestion.

4.2.3 Sons of the Starfarers

We include *Sons of the Starfarers* by Joe Vasicek because it is data that ChatGPT has never seen. We experiment with both one-shot prompting, and fine-tuning chatGPT on this text, to see what ChatGPT can do with an entirely unknown text.

4.3 *Coverage*

We report coverage at $n = 1$ and $i = 5$. In other words, a passage adds to the coverage if at least one good cross-reference appears in the first 5 suggestions. We assume that a typical reader, interested in the source passage, will be willing to read 5 suggested cross-references before giving up on finding a good one. Future research could be done with actual readers to validate these assumptions. *Coverage* results on Bible data for a range of values of n and i are given in Appendix B.

5 Results

We report results on the English Bible, followed by Bible-GER, followed by the works of Jane Austen.

5.1 Topic sweeps

We sweep the topic space with a low of 50 topics, increasing until we no longer see improvements in coverage. Performance gains steadily from 50 topics with best results around 4,000 topics. We report results for best-performing number of topics.

For the tand_n model, 4,000 topics always performs best. See Table 1 for results for the tand_n model surrounding 4,000 topics.

	rand_4000		rand_8000		rand_9000		rand_10241		rand_14730	
	cos	city	cos	city	cos	city	cos	city	cos	city
TSKE	0.363	0.421	0.443	0.506	0.461	0.529	0.465	**0.547**	0.469	0.546
Open	0.278	0.326	0.346	0.400	0.360	0.419	0.364	0.434	0.368	**0.435**
Open5	0.154	0.193	0.208	0.256	0.229	0.270	0.219	0.280	0.225	**0.286**

Table 3: Bible *coverage* results for the rand_n model using both cosine and cityblock metrics. This model performs best when given the entire vocabulary as topics, sometimes with stemmed vocabulary (rand_10241), sometimes with unstemmed vocabulary (rand_14730).

	embeddings		ChatGPT		
	cos	city	ref	quote	both
TSKE	0.564	0.562	0.50	0.60	**0.66**
Open	0.440	0.437	0.44	**0.55**	0.54
Open5	0.294	0.288	0.47	**0.50**	0.45

Table 4: Bible *coverage* results for the sentence embedding and ChatGPT models.

The gram_n model also tops out at 4,000 topics, except on *Open5*, where 5,000 topics using the cityblock distance has a very slight advantage, gaining 0.2% *coverage*. Interestingly, despite the fact that the tand_n model outperforms the gram_n model when creating a static set of cross-references in Lund et al. (2019), here the gram_n model outperforms the tand_n model in terms of coverage for nearly all numbers of topics. See Table 2 for the results for the topic sweeps surrounding 4,000 topics for the gram_n model.

Perhaps the most interesting of the topic-modeling-based results is for the rand_n model which continues to increase its performance beyond 4,000 topics, and in fact increases in *coverage* until we max out the vocabulary at rand_10241. After maxing out the stemmed vocabulary, we run the model without stemming. Maxing out the unstemmed vocabulary at gram_14730, we see very small gains in most instances. See Table 3 for the results for the rand_n topic sweeps.

Surprised by this outcome, we run gram_10241–the whole vocabulary–to ensure that gram_4,000 is not a local minimum. However, gram_10241 fails to improve on gram_4000, performing slightly worse in all cases. This is surprising since rand_10241 and gram_10241 use exactly the same set of topic words at this point. The only difference between the two models is the epsilon of $1e^{-10}$ added in each dimension to the rand_10241 topics. We do not know why this epsilon produces such a large increase in coverage (7-10% depending on the cross-referencing set). The epsilon represents uncertainty about the exact position a particular topic word should occupy in the topic space. Perhaps this allows the model the leeway it needs to tie topic words to words they might otherwise ignore if they contained zero weights in some dimensions.

The best topic-based model is rand_14730.

5.2 Sentence Embeddings

The sentence embedding model outperforms rand_14730 by 0.5-1.4% using cosine similarity. See Table 4 for sentence embedding results.

5.3 Distance Metrics

Cityblock distance outperforms cosine similarity for all topic-modeling models by a wide margin–up to 9%. This is not too surprising since the cityblock metric (L1 norm) has been shown to outperform other norms in high-dimensional spaces (Aggarwal et al., 2001). Cosine similarity outperforms cityblock distance by a small margin for all sentence embedding models. This is likely due to the way each model represents data. In topic models each vector component represents a discrete topic. Sentence embedding vector components represent data more abstractly, with no single concept attached to a particular component.

5.4 ChatGPT

ChatGPT outperforms all other models we employ on the Bible dataset, achieving a performance 3.8-20.6% better than the next best performance. See Table 4 for the ChatGPT results. ChatGPT particularly outperforms other models on the Open5 set

	tand_3000 (cos)	rand_20984 (city)	embeddings (cos)	ChatGPT (quote+ref)
TSKE	0.270	0.490	0.432	**0.588**
Open	0.200	0.383	0.342	**0.490**
Open5	0.109	0.245	0.241	**0.419**

Table 5: Bible-GER *coverage* results for the baseline model (tand_3000), rand_20984 (the entire unstemmed German vocabulary), multilingual sentence embeddings, and ChatGPT given both the reference and passage text.

of cross-references. This is most likely because the Open5 cross-references tend to be very popular cross-references, and since ChatGPT almost certainly contains Bible cross-references in its training data, very popular cross-references are highly likely to reappear in ChatGPT-generated text.

5.5 German

For the Bible-GER dataset, ChatGPT also performs best, followed by rand_20984 (using the entire unstemmed German vocabulary), followed by the sentence embedding model, see Table 5. Each model performs worse on the German data than on the English data. Some of this may be because the cross-referencing set we are using to evaluate the data was compiled using English Bibles. Likely, it is also because of the language-specific nature of many of the models and tools we are employing, including the stemmer we use for preprocessing, the sentence embedding model, and ChatGPT.

We also saw one particular recurring error in the ChatGPT generations for the Bible-GER dataset. ChatGPT often merged the number in front of a Bible reference into the structure of the list, (e.g. "5. 5.1 Thessalonicher 2:3-4", instead of "5. 1 Thessalonicher 2:3-4"). We do not see this behavior in English ChatGPT generations, but for the Bible-GER dataset it occurs in 37% of generations with an average of 2.3 errors in those generations.

5.6 Jane Austen

Our authors rated cross-references for Jane Austen wildly differently. However, sentence embedding widely outperformed rand_20948 (the full Jane Austen vocabulary), which widely outperformed by ChatGPT. See Appendix D for the numerical results for the Jane Austen data. Because ChatGPT performed so much worse for our Jane Austen data, we perform an error analysis of ChatGPT's responses below.

5.6.1 ChatGPT Error Analysis

For 28 source passages, ChatGPT did not produce 5 total suggestions.

In 19 suggestions (4% of the time), ChatGPT gave recommendations for how to find cross-references for a particular sentence instead of suggesting cross-references (e.g 'Social Class and Morality: "Mansfield Park" explores themes of social class and morality. You can look for quotes that delve into the moral values and social hierarchies of the characters.').

In 67 suggestions (14% of the time), ChatGPT recommended an entire scene instead of the explicitly requested quote (e.g. "Mr. Collins's proposal to Elizabeth is a comical but cringe-worthy moment that relates to the theme of marriage and the importance of character in choosing a spouse.")

Of the 377 quotes that ChatGPT suggested, 149 (40%) were not accurate quotes; 42 contained pieces of recognizable quotes and 107 appeared to be entirely fabricated (e.g. "Truth is always truth, either in the shape of a woman or a rhinoceros;").

Of the remaining 227 verifiably accurate quotes, 47 (21%) were from other sources besides Jane Austen's work including movies based on her novels and other literary works, and 142 (63%) had at least one duplicate among cross-reference suggestions for other sentences (e.g. "It is a truth universally acknowledged . . . " was suggested 12 times). In other words, it seems that the more well-known a quote is, the more likely it is to be suggested as a cross-reference by ChatGPT, indicating that ChatGPT may not be cross-referencing related passages so much as suggesting popular quotes regardless of the context. This so-called "Matthew Effect" (Merton, 1968) of ChatGPT has been reported in citation generation and in environmental science (Salleh, 2023; Petiska, 2023). ChatGPT's performance on this task suggests that unless the cross-referencing task has already been performed manually for a text, ChatGPT struggles

to find cross-references.

5.7 Sons of the Starfarers

As expected, when applying a one-shot prompt to Sons of the Starfarers, ChatGPT cannot suggest any real quotes from the novel. It either suggests quotes from other literature–often classic literature, which seems odd for the space genre. Or it suggests unattributed quotes that seem to be entirely fabricated.

After finetuning ChatGPT on a chapter completion task (see Appendix C for details), ChatGPT was still unable to generate any real quotes from the novel. The finetuned model also stopped suggesting quotes from other literary works and only suggested fabricated quotes. This may indicate that ChatGPT's performance on Jane Austen's works can be attributed to the prevalence of criticism and well-known quotes in ChatGPT's training data.

6 Discussion

The big winner on Bible data is ChatGPT. However, ChatGPT has an unfair advantage over the other models in this use case. Specifically, ChatGPT has undoubtedly seen cross-references and discussion of cross-references in its training data. We should obviously use this unfair advantage to our benefit when possible (for a Bible study tool, ChatGPT may be the best option). However, we also need to consider how ChatGPT will perform on texts that do not already have extensive cross-references available (the majority of texts).

For the works of Jane Austen and Sons of the Starfarers, ChatGPT performs significantly worse than the semantic similarity methods. There are almost certainly some ways in which these problems could be mitigated, including using newer models or embedding the entire corpus of text in the prompt.

We do some prelimary experiments using large contexts with ChatGPT and OpenAI's new gpt-4o model on quotes from *Sons of the Starfarers*. When including large portions of the text in the prompt both ChatGPT and gpt-4o were able to suggest quotes that were recognizably from the novel, with gpt-4o seemingly able to capture more nuanced connections. These methods may quickly become expensive (approximately $0.25 per source passage when employing the widest possible context length) which could make cross-references on demand less widely available. The desired cross-referencing corpus for a particular work may also be too large for even the largest context windows currently available, Google currently boasts a Gemini model with a 1 million token context (Pichai, 2024).[3] However, as LLM models improve and costs come down, this may indeed be a viable solution. It may also be possible to implement some form of retrieval augmented generation (RAG) to achieve better results without needing to use exceedingly large context windows (Lewis et al., 2020). Indeed, we are hopeful that this task will be adopted broadly, and new models and methods will be adapted and developed to improve on our results. Meanwhile, for non-Biblical texts, we recommend using sentence embeddings for cross-references on demand.

7 Related Works

Our work is most similar to work on intertextuality, source attribution, and literary evidence retrieval. Forstall and Scheirer provide an in-depth description of the use of computational tools, including topic-modeling, to discover literary intertextuality. Source attribution has been long-studied, but recent work by Muther and Smith (2023) is similar to our work in that it uses language models to rank candidate text. The most similar task to ours is likely literary evidence retrieval as explored by Thai et al. (2022) who created a novel dataset for literary evidence retrieval to test the ability of models to match literary analysis with the quotation described by that analysis. Source attribution, intertextuality, and literary evidence retrieval all seek to find the source for a statement, working from the assumption that a ground-truth source exists, generally in another body of work. For cross-references, there is no assumed source. In stark constrast, the "ground-truth" is how useful the cross-reference is to the reader.

8 Conclusion

We proposed cross-references on demand, defined *coverage*–a metric to evaluate performance on this task, and showed the efficacy of three different models on producing cross-references for three texts in English and German.

ChatGPT outperforms other models on the Bible. However, it performs significantly worse on texts that do not have existing cross-references, includ-

[3] Some authors, such as Brandon Sanderson, routinely write hundreds of thousands of words per year (Sanderson, 2018).

ing those represented in its training data. Topic-modeling and sentence embedding models perform comparably on the Bible, but hand-evaluation of these models suggests sentence embedding performs significantly better. We suggest applying sentence-embedding models when implementing interactive cross-references for texts for which no cross-reference sets exist.

Further research could focus on validating model performance directly with readers as well as adapting other models to cross-reference on demand.

9 Limitations

This work is still exploratory, and as such has several limitations. First, and foremost is our heavy reliance on Bible data. We are largely restricted in the ability to evaluate texts at scale by the limited existence of other large-scale cross-referencing resources.

Second, while we did use ChatGPT to produce cross-references, many other LLM models are available including more sophisticated models. Future work could explore these and ways to improve large language model performance for cross-referencing. It may be that fine-tuning a model on a large dataset of cross-references from a variety of sources could yield better results. However, the lack of available cross-referencing resources outside the Bible could make this a difficult endeavor.

Thirdly, while *coverage* is a useful quantitative representation of overall reader satisfaction, it does not take into account more qualitative aspects such as the relevance or explanatory power or cross-references. Future work with users should both validate the *coverage* metric and explore qualitative attributes of the cross-references suggested when determining the success of an on-demand cross-reference system.

Finally, we limit ourselves to cross-references from within a single body of work. Future work could assess whether these approaches are as effective when texts from multiple sources are included in the corpus.

References

Charu C Aggarwal, Alexander Hinneburg, and Daniel A Keim. 2001. On the surprising behavior of distance metrics in high dimensional space. In *Database Theory—ICDT 2001: 8th International Conference London, UK, January 4–6, 2001 Proceedings 8*, pages 420–434. Springer.

Lavina Fielding Anderson. 1979. Church publishes first lds edition of the bible. *Ensign*, 10.

Sanjeev Arora, Rong Ge, Yoni Halpern, David Mimno, Ankur Moitra, David Sontag, Yichen Wu, and Michael Zhu. 2013. A practical algorithm for topic modeling with provable guarantees. In *Proceedings of the 30th International Conference on International Conference on Machine Learning-Volume 28*, pages II–280.

Sven Bigl. Bestseller luther-bibel: Nachfrage im jubiläumsjahr 2022 um ein drittel gestiegen [online]. 2023.

Mark (director) Dindal. 2000. The emperor's new groove. Film.

Christopher W Forstall and Walter J Scheirer. Quantitative intertextuality.

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

Jeffrey Lund, Piper Armstrong, Wilson Fearn, Stephen Cowley, Emily Hales, and Kevin Seppi. 2019. Cross-referencing using fine-grained topic modeling. In *Proceedings of NAACL-HLT*, pages 3978–3987.

Jeffrey Lund, Connor Cook, Kevin Seppi, and Jordan Boyd-Graber. 2017. Tandem anchoring: A multiword anchor approach for interactive topic modeling. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 896–905.

Robert K Merton. 1968. The matthew effect in science: The reward and communication systems of science are considered. *Science*, 159(3810):56–63.

Timothy Morton. 2010. *Treasury of Scripture Knowledge, Enhanced*. BibleAnalyzer.com.

Ryan Muther and David Smith. 2023. Citations as queries: Source attribution using language models as rerankers. *Preprint*, arXiv:2306.17322.

Oak Norton. 2024. Introducing daniel, your scripture notes ai research assistant " scripture.

OpenAI. Introducing chatgpt [online]. 2022.

Eduard Petiska. 2023. Chatgpt cites the most-cited articles and journals, relying solely on google scholar's citation counts. as a result, ai may amplify the matthew effect in environmental science. *arXiv preprint arXiv:2304.06794*.

Sundar Pichai. 2024. Our next-generation model: Gemini 1.5.

Martin F Porter. 1980. An algorithm for suffix stripping. *Program*, 14(3):130–137.

$n\backslash i$	1	2	3	4	5	6	7	8	9	10
1	0.337	0.428	0.481	0.517	0.546	0.570	0.589	0.604	0.619	0.633
2	-	0.119	0.181	0.224	0.256	0.283	0.306	0.328	0.345	0.361
3	-	-	0.045	0.078	0.104	0.127	0.148	0.165	0.182	0.198
4	-	-	-	0.018	0.034	0.050	0.064	0.076	0.089	0.100
5	-	-	-	-	0.007	0.016	0.025	0.033	0.041	0.049
6	-	-	-	-	-	0.004	0.008	0.013	0.017	0.021
7	-	-	-	-	-	-	0.002	0.004	0.007	0.009
8	-	-	-	-	-	-	-	0.001	0.002	0.004
9	-	-	-	-	-	-	-	-	0.000	0.001
10	-	-	-	-	-	-	-	-	-	0.000

Table 6: This table shows the *coverage* as evaluated for various values of n and i on Bible data using rand_14730 and the cityblock distance metric, the random words topic model using the entire unstemmed vocabulary as topics; this was the most successful topic-based model.

Martin F Porter. 2001. Snowball: A language for stemming algorithms.

Nils Reimers and Iryna Gurevych. 2019. Sentence-bert: Sentence embeddings using siamese bert-networks. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3982–3992.

Nils Reimers and Iryna Gurevych. 2020. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525.

Kim Renfro. 2020. 17 details you might have missed in the'harry potter' books. *Business Insider*.

J.K. Rowling. 1997. *Harry Potter and the Philosopher's Stone*. Bloomsbury Publishing, London, UK.

J.K. Rowling. 2003. *Harry Potter and the Order of the Phoenix*. Bloomsbury Publishing, London, UK.

Hamidah M Salleh. 2023. Errors of commission and omission in artificial intelligence: contextual biases and voids of chatgpt as a research assistant.

Brandon Sanderson. 2018. What is your daily word-count/time goal?

Mun'im Sirry. 2022. *The Qur'an with Cross-References*. De Gruyter.

Jerome H. Smith. 2016. *The Ultimate Cross-Reference Treasury, for e-Sword*. Thomas Nelson.

Leonid Taycher. Books of the world, stand up and be counted! all 129,864,880 of you. [online]. 2010.

Katherine Thai, Yapei Chang, Kalpesh Krishna, and Mohit Iyyer. 2022. Relic: Retrieving evidence for literary claims. *Preprint*, arXiv:2203.10053.

A Examples of Subtle Cross-references

Below we give two examples of cross-references more subtle than the Emperor's New Groove callback in the Introduction.

A.1 Harry Potter Example

In J.K. Rowling's first book *Harry Potter and the Philosopher's Stone* (Rowling, 1997), at one point Harry thinks:

> Could Snape possibly know they'd found out about the Philosopher's Stone? Harry didn't see how he could–yet he sometimes had the horrible feeling that Snape could read minds.

In the fifth Harry Potter book, *Harry Potter and the Order of the Phoenix* (Rowling, 2003), we find the following quote from Snape:

> Those who have mastered Legilimency are able, under certain conditions, to delve into the minds of their victims and to interpret their findings correctly.

In other words, in an early book Harry wonders if Snape can read minds, and in a later book he finds out that Snape can read minds, at least after a fashion. However, unlike the callback from the

$n\backslash i$	1	2	3	4	5	6	7	8	9	10
1	0.341	0.436	0.494	0.534	0.564	0.589	0.611	0.628	0.642	0.655
2	-	0.114	0.176	0.221	0.258	0.289	0.313	0.336	0.354	0.372
3	-	-	0.039	0.073	0.099	0.122	0.143	0.162	0.179	0.195
4	-	-	-	0.013	0.027	0.042	0.056	0.069	0.081	0.093
5	-	-	-	-	0.005	0.012	0.019	0.027	0.033	0.041
6	-	-	-	-	-	0.002	0.005	0.009	0.012	0.016
7	-	-	-	-	-	-	0.001	0.002	0.004	0.006
8	-	-	-	-	-	-	-	0.000	0.001	0.002
9	-	-	-	-	-	-	-	-	0.000	0.000
10	-	-	-	-	-	-	-	-	-	0.000

Table 7: This table shows the *coverage* as evaluated for various values of n and i on Bible data using sentence embeddings and the cosine similarity metric, the random words topic model using the entire unstemmed vocabulary as topics.

Emperor's New Groove given above, Harry's supposition in the first book is not emphasized enough that it will easily be remembered 4 books later, nor does the fifth book refer back to the earlier thought, making this a much more difficult cross-reference to find.

We found this cross-reference in Renfro (2020), an article discussing this reference as well as other "foreshadowing" in the Harry Potter series.

A.2 Scriptural Example

One example of a more subtle cross-reference is given below between the Bible and the Book of Mormon–a book of scripture for The Church of Jesus Christ of Latter-day Saints.

> John 9:1-3 – As he passed by, he saw a man blind from birth. And his disciples asked him, "Rabbi, who sinned, this man or his parents, that he was born blind?" Jesus answered, "It was not that this man sinned, or his parents, but that the works of God might be displayed in him.

After which, Jesus proceeds to heal the man of his blindness.

In Ether, God is speaking, and says:

> Ether 12:27 – And if men come unto me I will show unto them their weakness. I give unto men weakness that they may be humble; and my grace is sufficient for all men that humble themselves before me; for if they humble themselves before me, and have faith in me, then will make weak things become strong unto them.

The connections between the two passages are as follows:

- Both of these verses begin by mentioning a "flaw." In John it is a physical disability, blindness, and in Ether it is the abstract concept of weakness.
- In each case the individual with the flaw is not blamed for the flaw. In John the idea of blame is specifically rejected by Jesus, and in Ether God assumes blame for weakness: "I give unto men weakness . . ."
- In each there is a different purpose given for the flaw. In John "that the works of God might be displayed," and in Ether "that they may be humble."
- Finally, the flaw is at the center of a transformation. In John Jesus heals the blind man so that he can see, and in Ether God promises, "I will make weak things become strong unto them."

While there are very strong connections between these verses, they would not appear together in any kind of word-based search, and so would be difficult to identify without great familiarity with the texts.

	rand_20984 (city)	embeddings (cos)	ChatGPT (quote)
Author 1	0.36	**0.70**	0.16
Author 2	0.79	**0.95**	0.52

Table 8: Manually evaluated Jane Austen *coverage* results for rand_12753 (the entire unstemmed vocabulary), sentence embeddings, and ChatGPT given the source passage.

B *Coverage* Results for multiple values of n and i

See Table 6 for the *coverage* for various values of n and i for the unstemmed random word model with 14730 topics—the whole vocabulary–using the city-block distance metric, i.e. the best-performing topic-based model.

See Table 7 for the *coverage* for various values of n and i for the sentence embedding model using cosine similarity.

Remember that i is the number of suggested cross-references evaluated, and n is the minimum number of valid cross-references that must be found for a source passage of text to be considered covered. In other words $n = 2$ and $i = 2$ means that two cross-references are suggested and both must be valid cross-references in order for that passage to add to the overall *coverage*. The bottom left side of the table is blank because it is impossible to find more valid cross-references than verses examined, so n can never be larger than i and produce a valid result.

It may be of interest to note that although the sentence embedding model performs better than the rand_14730 model at $n = 1$, $i = 5$, as the value of n increases, rand_14730 performs better than the sentence embedding model including for some values of i when the value of n is 2. By the time n is 3, rand_14730 outperforms the sentence embedding model for all values of i.

C Finetuning ChatGPT

We used the following prompt to finetune ChatGPT with the text of *Sons of the Starfarers*:

What does the [NTH] chapter, [CHAPTERNAME] of [TITLE] by Joe Vasicek say? It's important that you know this so you can cross-reference it later. [CHAPTERTEXT]

We replaced [NTH] with the ordinal number of the chapter, [CHAPTERNAME] with the name of the chapter, [TITLE] with the title of the book, and [CHAPTERTEXT] with the actual text of the chapter. Note that when using a similar prompt with *Pride and Prejudice*, ChatGPT is capable of reproducing Jane Austen's writing.

We trained on a total of 1.3 million tokens, for 3 epochs with a batch size of 1, and a learning rate multiplier of 2–these were the default settings suggested by OpenAI. The total cost to finetune was $10.49 USD.

D Jane Austen Hand-evaluation

See Table 8 for the numerical results of our authors hand-evaluating the Jane Austen cross-reference suggestions.