CiteBART: Learning to Generate Citations for Local Citation Recommendation

Anonymous ACL submission

Abstract

Local citation recommendation (LCR) suggests a set of papers for a citation placeholder within a given context. This paper introduces CiteBART, citation-specific pre-training within an encoder-decoder architecture, where author-date citation tokens are masked to learn to reconstruct them to fulfill LCR. The global version (CiteBART-Global) extends the local context with the citing paper's title and abstract to enrich the learning signal. CiteBART-Global achieves state-of-the-art performance on LCR benchmarks except for the FullTextPeerRead dataset, which is quite small to see the advantage of generative pre-training. The effect is significant in the larger benchmarks, e.g., Refseer and ArXiv., with the Refseer pre-trained model emerging as the best-performing model. We perform comprehensive experiments, including an ablation study, a qualitative analysis, and a taxonomy of hallucinations with detailed statistics. Our analyses confirm that CiteBART-Global has a cross-dataset generalization capability; the macro hallucination rate (MaHR) at the top-3 predictions is 4%, and when the ground-truth is in the top-k prediction list, the hallucination tendency in the other predictions drops significantly. We publicly share our code[1] to support reproducibility[2].

1 Introduction

Citations are essential building blocks in scientific writing. Their accurate placements indicate quality as one requires to put the current study in the context of the existing work from different aspects, such as background information, method, and result comparison (Cohan et al., 2019).

Citation prediction is a two-step process where the former focuses on where in the sentence to place the citation (Buscaldi et al., 2024), while the latter (citation recommendation) obtains a set of candidate papers once there is a specified citation placeholder in a given context. In this study, we focus on the latter, referred to as the Local Citation Recommendation (LCR). LCR functions as a citation suggestion mechanism for local textual contexts that are presumed to contain citations. The suggestions can be considered additional reading material alongside the targeted paper, corresponding to the ground-truth citation.

LCR has been addressed in a few works. BERT-GCN (Jeong et al., 2020) utilizes a feedforward neural network to combine local citation context representations using BERT with citation encodings through Graph Convolutional Neural Networks (GCN). The most recent solutions to the problem adopt a two-step process that consists of pre-fetching and re-ranking. DualEnh (Medić and Snajder, 2020) enhances a local citation context with the citing article's title and abstract and uses this enhanced context as the query vector to retrieve the most similar candidate articles using their titles and abstracts. It performs the ranking through BiLSTM representations of inputs with attention layers on top. HAtten (Gu et al., 2022) initially pre-fetches a set of papers similarly. Afterward, it re-ranks the selected candidate papers using a fine-tuned SciBERT (Beltagy et al., 2019) model where the input is the query text concatenated with a candidate paper's title and abstract.

Fierro et al. (2024) support information-seeking using query-focused summarization, responding to user queries by answers with source attributions. The ALCE benchmark (Gao et al., 2023) collects a diverse set of questions and retrieved passages to support answer generation with appropriate citations. CiteBART is different from these works as it aims to fill in a citation placeholder, not targeting retrieval-based summaries with citations.

CiteBART-Base learns through the masked cita-

[1] https://anonymous.4open.science/r/CitationRecommendation-5FA1

[2] We will also share the base and global datasets and pre-trained models through a link later.

Table 1: An example for input and target formats for pre-training and evaluation with CiteBART. Due to space constraints, we present the contexts and abstracts in an abbreviated form.

Strategy	Input	Target
Base	... error rate of 5.8% and a word error rate of 28.7%, which are on par with previous reported results **<mask>** . Unlike prior work, we do not use a language model during decoding and ...	**Yao and Zweig, 2015**
Global	... error rate of 5.8% and a word error rate of 28.7%, which are on par with previous reported results **<mask>** . Unlike prior work, we do not use a language model during decoding and ... *</s>* Deep Voice: Real-time Neural Text-to-Speech *</s>* We present Deep Voice, a production-quality text-to-speech system constructed entirely from deep neural ...	**Yao and Zweig, 2015**

tion context. In CiteBART-Global, we extend the masked context with the citing paper's global information, e.g., title and abstract (Table 1). Inspiring from pre-training under the REALM framework (Guu et al., 2020), we append this global information to the local context, allowing backpropagation through the global information to learn associations with the pool of papers from the corpus.

CiteBART achieves superior performance without relying on a pre-fetch and re-rank pipeline. It is an end-to-end learning system. Unlike the previous works, we do not exploit the global information (titles and abstracts) of target papers to make the recommendation. CiteBART-Global learns solely from the relation of citing papers' global information with local citation contexts.

We summarize our contributions as follows:

- We propose an end-to-end learning system, CiteBART, with custom citation masking for LCR.
- CiteBART-Global achieves state-of-the-art performance on LCR benchmarks except for the FullTextPeerRead dataset, which is quite small to see the advantage of generative pre-training. The effect is significant in the larger benchmarks, e.g., Refseer and ArXiv. CiteBART-Base is still a strong baseline.
- We provide a qualitative analysis to gain insight into the working of the approach, including the cross-dataset generalization capability.
- We provide a taxonomy of hallucinated citations and report macro hallucination rates (MaHR) for them.
- Our ablation study confirms the central role of local citation contexts in the learning process. It also shows the effectiveness of the Global training scheme over Base.

2 Related Work

BERT (Devlin et al., 2019) is an encoder-only pre-training model that adopts the Masked Language Modeling (MLM) objective. MLM masks tokens in a uniformly random fashion and predicts them, allowing the generation of learning signals bidirectionally. Some BERT variants were released to meet the requirements for masking a group of tokens. SpanBERT (Joshi et al., 2020) builds on this objective by masking random contiguous text spans. In the same direction, PMI-Masking (Levine et al., 2021) masks word n-grams based on their PMI (Pointwise Mutual Information) scores. Pre-training encoder decoders, e.g., BART (Lewis et al., 2020), combine the strengths of bidirectional learning of encoders with the autoregressive nature of decoders, capturing the local patterns of tokens within their generative capabilities.

Similar to citation recommendation, the recent work of Luo et al. (2023) predicts provisions of the U.S. Code by pretraining RoBERTa (Liu et al., 2019) and LegalBERT (Chalkidis et al., 2020) on the curated dataset (PACER (Luo et al., 2023)) of the US federal court documents where each provision source text is given with its associated target citation. SciBERT (Beltagy et al., 2019) performs pretraining exclusively on scientific texts to learn global representations for scientific papers. SPECTER (Cohan et al., 2020) learns citation-aware global representations for scientific papers using a citation-based pretraining objective. SPECTER-produced representations introduced remarkable results in the paper classification and global citation recommendation tasks.

LCR has four benchmark datasets for evaluation. BERT-GCN (Jeong et al., 2020) introduced the FullTextPeerRead dataset, extended from the original PeerRead (Kang et al., 2018). Throughout this paper, we refer to the FullTextPeerRead dataset as PeerRead for brevity. An additional dataset is ACL-ARC (Bird et al., 2008), derived from the ACL Anthology Reference Corpus. We run our experiments on its ACL-200 subcategory, analogous to DualEnh (Medić and Snajder, 2020) and HAtten (Gu et al., 2022). Finally, Refseer (Huang et al., 2015) and ArXiv (Gu et al., 2022) are the largest

benchmarks for this task.

BERT-GCN (Jeong et al., 2020) combines two encoders for citation recommendation. The first encoder generates local context embeddings using BERT, while the second one creates the graph embeddings of citation networks. DualEnh (Medić and Snajder, 2020) trains a Bi-LSTM model to relate a target paper to its candidate papers. The target paper provides a context with a citation placeholder, and the model utilizes the titles and abstracts of candidate papers to calculate the semantic similarity scores. HAtten (Gu et al., 2022) uses a Hierarchical Attention Text Encoder and SciBERT-based Re-ranking scheme for LCR. It starts by prefetching potential candidate papers from a pool of citations. In the re-ranking phase, the authors assign scores to candidate papers using a SciBERT model with a classification layer on top. HAtten achieves state-of-the-art results on all of the benchmark datasets.

Lastly, GM-s2orc-H (Buscaldi et al., 2024) proposes approaches for predicting where in the context to place the citation. Although their results are not directly comparable to CiteBART due to differences in task objectives, their findings highlight the advantages of generative models in citation-related tasks.

3 Methodology

We propose CiteBART, a novel pre-training strategy designed to predict citations within the contexts of scientific papers. We mask placeholder tokens, which replace ground-truth citations in the parenthetical author-date style, for the continual pre-training of a vanilla BART-base to generate the correct parenthetical author-date citation for a given context.

3.1 Custom BART Pre-training for LCR

BART (Lewis et al., 2020) is a sequence-to-sequence model with an encoder and a decoder. It introduces a set of document corruption (denoising) schemes and then optimizes a reconstruction loss, the cross-entropy between the original document and the decoder's outputs. The denoising transformations that are applied to the encoder during pre-training are as follows: Random token masking (similar to BERT), token deletion, text infilling (span masking with span lengths drawn from a Poisson distribution ($\lambda = 3$)), sentence permutation, and document rotation with a randomly selected token leading the document.

We propose a citation learning strategy using BART. BART employs MLM similar to BERT. Additionally, to effectively reconstruct the masked contexts, it masks a span of k tokens with a single mask. In return, it can predict multiple tokens for a single mask. Thus, CiteBART can generate complex parenthetical author-date citations after custom pre-training for citation tokens without requiring further architectural modifications.

We propose two training schemes for our approach: CiteBART-Base and CiteBART-Global. In CiteBART-Base, the model gets the masked context with the ground-truth citation as input. This setting tests the model's performance in a local context-only situation (Table 1). With the underlying idea that good citation recommendation requires relating local citation contexts with the citing papers' global information, such as titles and abstracts, we devised an innovative way to accomplish it. Inspiring from pre-training under the REALM framework (Guu et al., 2020), in CiteBART-Global, we append the citing paper's title and abstract to the local context, allowing backpropagation through the global information that considers the pool of papers from the corpus. Specifically, we used the "</s>" token designated by the pre-trained BART-base model as the separator.

Table 2: Statistics of LCR benchmarks.

Dataset Name	ACL-200	PeerRead	RefSeer	Arxiv
Train Size	30,390	9,363	3,521,582	2,988,030
Validation Size	9,381	492	124,911	112,779
Test Size	9,585	6,184	126,593	104,401
# of Papers	19,776	4,837	624,957	1,661,201
Publication Years	2009-2015	2007-2017	-2014	1991-2020

3.2 Dataset Preprocessing

We conduct our experiments on the existing citation recommendation benchmarks of ACL-200, PeerRead, RefSeer, and ArXiv. Table 2 presents the statistics of these datasets. They provide citation contexts from various articles where all contexts have a target citation in the middle. The context sizes are in terms of characters, which causes some incomplete words at the start and end of the contexts.

The datasets originally include a "TARGETCIT" marker as a placeholder for citations within each context. We replaced these markers with "<mask>" tokens to align with our pretraining process. Additionally, to ensure CiteBART focuses

solely on predicting target citations, we removed any non-target citations from all four datasets.

We encountered some issues during the preprocessing of ACL-200 and RefSeer. First, they include local contexts with author name conflicts in the citation tokens. For example, the "Petrović et al., 2010" citation token was incorrectly written as "Petrovic et al., 2010" in the target citation column of ACL-200. Another problem is the incorrect ordering of two-author citations. For instance, the local citation context provides the citation "Rivera and Zeinalian, 2016"; the paper metadata includes "Zeinalian and Rivera, 2016". There are also a few cases of incorrect citations. Moreover, there are some contexts with empty author names. We removed all these cases from the aforementioned datasets to ensure consistency.

After the preprocessing, we worked with the train and test sets. As CiteBART involves continual pre-training, we perform it on the training partition and evaluate the performance on the test partition. Table 3 shows the final statistics of our preprocessed datasets[3] including the training and test partition sizes for all the benchmarks.

Table 3: Statistics of the preprocessed datasets.

Dataset Name	ACL-200	PeerRead	RefSeer	Arxiv
# of local contexts	63,365	16,669	3,739,189	3,205,210
Size of the training split	50,692	13,335	2,991,351	2,564,168
Size of the test split	12,673	3,334	747,838	641,042
# of removed contexts	403	0	39,577	0
# of unique citations	5,266	2,043	351,896	368,284

4 Experiments

We conducted our experiments on devices with NVIDIA RTX6000 Ada GPU and NVIDIA V100 GPU[4]. The following hyperparameters were utilized in all our experiments. The number of epochs was set to 15, as the change in loss values between epochs became negligibly small beyond this point. Only the PeerRead Global dataset has been trained for 30 epochs since the generative model requires longer training for the relatively smaller PeerRead dataset. We employed a learning rate of $2e-5$ and an attention dropout rate of 0.12. Given that BART is a generative model, we adjusted its generation parameters to produce outputs that align with our requirements. Specifically, we utilized the grouped beam search with 20 beams and applied a diversity penalty of 1.5 to generate more diverse results. The maximum number of generated tokens was 25 since the generated citations should not exceed it. Apart from these specific modifications, we did not alter the architecture of the BART model.

4.1 Results

We report our results using Recall@10 (R@10) and Mean Reciprocal Rank (MRR)[5] and compare with the state-of-the-art approaches in Table 4[6]. As can be seen from the table, CiteBART-Global outperforms others on the existing benchmarks except for the smallest PeerRead dataset, while the base scheme is still a strong baseline, surpassing BERT-GCN on PeerRead, DualEnh, and HAtten on Refseer. The table includes the best-reported results of HAtten with $2k$ pre-fetched candidates. As for DualEnh (Medić and Snajder, 2020), we chose their superior "DualEnh-ws" model for the comparison. BERT-GCN's (Jeong et al., 2020) results are available only on the PeerRead dataset.

As shown in Table 4, CiteBART-Global demonstrates its advantage over HAtten on Refseer most since Refseer includes more training contexts compared to ArXiv. Given that CiteBART is a generative model, access to a larger training set contributes to its improved results.

To observe the citation prediction capabilities of CiteBART in detail, we performed a qualitative analysis.[7] Furthermore, we provide another qualitative analysis on the performance of LLMs in LCR.[8].

4.2 Ablation Study

We conducted an ablation study to show different components' contributions to the overall results. The analysis was carried out on the ACL-200 dataset. Table 5 shows the results for CiteBART with a model pre-trained on the ACL-200 Global dataset in 15 epochs.

The first three experiments test the contribution of the local context, title, and abstract to the overall performance. First, we remove the local context to see the performance due to the global information-only training (#1 in Table 5). We discard the title and abstract in the second and third configurations

[3]Please find information on token limits in Appendix A.

[4]Please find information on training and evaluation times in Appendix B.

[5]The metric definitions are given in Appendix C.

[6]We share our Exact Match (EM) scores in Appendix D.

[7]Please find our qualitative analysis in Appendix E.

[8]Please find our additional analysis in Appendix F

Table 4: Comparison with state-of-the-art on LCR benchmarks.

Model	ACL-200		PeerRead		Refseer		Arxiv	
	R@10	MRR	R@10	MRR	R@10	MRR	R@10	MRR
BERT-GCN[a]	-	-	0.529	0.418	-	-	-	-
DualEnh-ws	0.703	0.366	-	-	0.534	0.280	-	-
HAtten	0.633	-[c]	**0.757**	-[c]	0.454[b]	-[c]	0.439[b]	-[c]
CiteBART-Base	0.686	0.504	0.570	0.424	0.606	0.449	0.355	0.240
CiteBART-Global	**0.739**	**0.513**	0.669	**0.502**	**0.652**	**0.479**	**0.502**	**0.305**

[a] BERT-GCN performs evaluation by excluding the papers cited less than five times in each dataset.
[b] The reported results are based on a $10k$ subset of the test set.
[c] The authors did not report their MRR scores.

(#2 and #3 in Table 5). The results show that excluding the local context brings about a sharp reduction in the performance metrics (a drop from 0.739 to 0.588 in Recall@10), confirming its decisive role in generating citations. On the other hand, removals of title or abstract do not lead to a statistically significant decrease in performance.

In the fourth ablation study, we further expand the global information with the cited paper's title and abstract during pre-training (#4 in Table 5). The evaluation stays the same, feeding the local context with the citing paper's title and abstract during inference. Contrary to expectations, adding the ground-truth paper's global information during pre-training does not help; the model falls in its performance. This failure may be explained by the model learning to associate the citation token with the global information of both the citing and cited article in the training phase. However, lacking the cited paper's global information in the test phase confuses the model's predictions.

The previous studies (Medić and Snajder (2020), Gu et al. (2022)) utilize an all-including training and inference configuration where citing and cited paper's global information is concatenated with the local citation context. Their pre-fetch and re-ranking pipeline is well-suited to this setup and benefits from it as the inference step also allows incorporating the cited paper's title and abstract, which is not the case in a learning approach like ours'. CiteBART-Global outperforms these models without relying on global information about the cited papers, representing a more ideal scenario for the LCR task.

4.3 Taxonomy and Measurement of Hallucinated Citations

CiteBART, similar to other generative models, is prone to hallucination, occasionally producing citations that do not correspond to any real work. A generated citation is classified as hallucination if it is not present in the citation list of the dataset including the input context. Hallucinations in CiteBART are typically entity-error hallucinations or fabrications.

To measure the degree of hallucinations in LLM-generated responses, Li et al. (2024) propose two metrics, MaHR (macro hallucination rate) and MiHR (micro hallucination rate), respectively. While MaHR calculates the proportion of hallucinatory responses in all the responses (Equation 1), MiHR gives the average rate of hallucinations within each response (Equation 2).

$$MaHR = \frac{Count(hallucinatory\ responses)}{n} \tag{1}$$

$$MiHR = \frac{1}{n}\sum_{i=1}^{n}\frac{Count(hallucinatory\ facts)}{Count(all\ facts\ in\ r_i)} \tag{2}$$

In LCR, MaHR represents the proportion of hallucinated citations across all generated citations. As the task is evaluated with top-k predictions for each test instance, the total number of responses becomes $k*n$ where n is the number of test instances. Thus, MaHR is the fraction of hallucinated citations among $k*n$ responses (Equation 3). MiHR, on the other hand, measures the average hallucination rate in individual contexts. In LCR, as each context gets top-k predictions, the number of facts in each response is fixed with k (the denominator in MiHR), which makes MaHR and MiHR produce identical results.

Table 5: Ablation study results on ACL-200 Global dataset under four different configurations.

	Approach	Training Input	Recall@10	EM	MRR
	Base	Context	0.686	**0.422**	0.504
	Global	Context + Citing Title & Abstract	**0.739**	0.417	**0.513**
1	No context	Citing Title & Abstract	0.588	0.205	0.311
2	No title	Context + Citing Abstract	0.731	0.415	0.509
3	No abstract	Context + Citing Title	0.712	0.396	0.490
4	All-including	Context + Citing Title & Abstract + Cited Title & Abstract	0.111	0.039	0.056

$$MaHR = \frac{Count(hallucinated\ citations)}{k * n}$$
$$MiHR = \frac{1}{n}\sum_{i=1}^{n}\frac{Count(hallucinated\ citations\ in\ context_i)}{k} \quad (3)$$

In addition to MaHR (or MiHR), we propose the following metrics to pinpoint hallucination behavior. Each metric targets a type of hallucination we categorized by examining **hallucinations versus ground truth** citations for given contexts.

- **Incorrect year (all-names-GT):** The generated citation fully matches the author(s) in the ground truth citation while failing to match the publication year.
- **Partially correct author list (one-name-GT):** One of the two author names is correct, and the generated year may or may not be correct in these cases.
- **Correct year with incorrect authors (year-GT):** Some hallucinations match the year of the ground truth citation, even if the author names are incorrect.
- **wrong-format:** If the generated citation's format does not conform to the parenthetical author-date citation style, it is considered a wrong-format hallucination.
- **other-hal:** There is no overlap with any part of GT in these hallucinations.

Additionally, we term the aggregation of the hallucinations corresponding to partially correct responses MaHR-partial and we relate MaHR with MaHR-partial using Equation 4.

$$MaHR\text{-}partial = all\text{-}names\text{-}GT + one\text{-}name\text{-}GT + year\text{-}GT$$
$$MaHR = MaHR\text{-}partial + wrong\text{-}format + other\text{-}hal \quad (4)$$

Table 6 presents the results of the hallucination metrics for the CiteBART-Global models. To observe the effect of the k value, we performed each analysis with top-3, top-5, and top-10 generated predictions, respectively. The results conclude that MaHR-partial accounts for almost half of the hallucinations in the top 3 predictions, which implies that when the model is forced to make fewer predictions, its hallucinations do not deviate much from the ground truth. The proportion gradually diminishes in the top-5 and top-10 predictions. Interestingly, on Refseer and Arxiv Global, the incorrect year (all-names-GT) hallucination, which is the closest to the ground truth, decreases with increasing k values. In overall performance, the ACL-200 Global dataset gives the lowest hallucination rates all over the k values. Arxiv Global is the second best, with very close scores to ACL-200 Global.

Table 7 reports the values of some extended metrics built upon MaHR:

- **top-k-match-MaHR**: This metric considers hallucinated predictions only when one of the other predictions in the same top-k group matches the ground truth (GT).
- **exact-match-MaHR:** This metric is similar to top-k-match-MaHR but specifically focuses on the cases where the exact match occurs.

These metrics approach the problem differently by examining the hallucination tendency when the model can hit the ground truth citation in its top-k predictions. In other words, the research question is whether the model suffers less from the hallucination given the correct prediction in the top-k list (when the model knows the answer). The results confirm this hypothesis as top-k-match-MaHR and exact-match-MaHR are different from MaHR in a statistically significant way with $p < 0.001$. Furthermore, Arxiv Global is the best model to mitigate hallucinations when it hits the ground truth, outperforming others in the hallucination rates.

4.4 Qualitative Analysis on Hallucinations

In this section, we provide additional examples to illustrate the types of hallucinations (Table 8). The first example shows an ideal scenario with no hallucinations in the top-10 prediction list. The other examples, except the last, depict different types of hallucinations. The last example show-

Table 6: Results for proposed hallucination metrics on Global datasets for top-3, top-5, and top-10 predictions. Metric values are shown as percentages (%). The best values are shown with **bold**.

Metrics	ACL-200			PeerRead			Refseer			Arxiv		
	Top-3	Top-5	Top-10	Top-3	Top-5	Top-10	Top-3	Top-5	Top-10	Top-3	Top-5	Top-10
all-names-GT	0.63	0.54	0.68	0.63	0.59	0.85	1.21	1.07	1.01	1.01	0.85	0.80
one-name-GT	0.24	0.29	0.44	1.31	1.06	1.21	0.63	0.75	0.82	0.43	0.56	0.65
year-GT	1.03	1.56	2.48	1.72	3.08	5.95	0.50	0.84	1.48	0.55	0.99	1.94
MaHR-partial	**1.89**	**2.39**	3.60	3.66	4.73	8.01	2.34	2.66	**3.31**	1.99	2.40	3.39
wrong-format	0.02	0.02	0.08	0.00	0.07	0.28	2.18e-5	4.97e-5	0.01	1.35e-5	2.12e-5	0.01
other-hal	2.20	4.00	9.02	4.36	7.26	15.02	2.94	5.25	10.05	2.66	4.74	9.95
MaHR	**4.12**	**6.42**	**12.69**	8.02	12.06	23.31	5.28	7.91	13.37	4.64	7.14	13.35

Table 7: Results for extended MaHR metrics on Global datasets for top-3, top-5, and top-10 predictions. Metric values are shown as percentages (%). The best values are shown with **bold**.

Metrics	ACL-200			PeerRead			Refseer			Arxiv		
	Top-3	Top-5	Top-10	Top-3	Top-5	Top-10	Top-3	Top-5	Top-10	Top-3	Top-5	Top-10
MaHR	**4.12**	**6.42**	**12.69**	8.02	12.06	23.31	5.28	7.91	13.37	4.64	7.14	13.35
top-k-match-MaHR	2.54	4.55	9.69	5.05	8.15	16.53	2.76	4.74	8.60	**1.79**	**3.12**	**6.28**
exact-match-MaHR	2.40	3.81	7.08	4.76	6.94	12.42	2.46	3.97	6.58	**1.48**	**2.33**	**3.95**

cases the cross-dataset generalization capability of CiteBART. Due to space limitations, contexts and abstracts have been abbreviated. Hallucinated predictions are designated with the * symbol. Correctly predicted citations are displayed in bold.

5 Discussion and Conclusion

CiteBART is distinctive as it performs LCR by end-to-end learning. Unlike the pre-fetch and re-rank pipelines, it does not exploit titles and abstracts of the cited papers for inference. In CiteBART-Base, we rely solely on local citation contexts, while CiteBART-Global incorporates the citing paper's global information to make predictions. CiteBART-Global achieves state-of-the-art performance on LCR benchmarks except for PeerRead, which is quite small to see the advantage of generative pre-training.

We comment on the pros of using BART over encoder-based pre-training models such as RoBERTa. BART's MLM objective is flexible and allows the masking of all the tokens in the parenthetical author-date style. RoBERTa cannot add citation tokens to its vocabulary by its MLM. Moreover, constraining predictions to citation tokens for RoBERTa is not straightforward. While BART is prone to hallucination, its capabilities significantly enhance LCR performance.

Furthermore, our comprehensive hallucination analysis sheds light on the hallucination behavior, MaHR-partial taking up significant proportions (almost half of the hallucinations in the top 3 predictions), which implies that all the hallucinations should not be rejected beforehand but show signs of promising generalization capabilities as MaHR-partial is the aggregation of partially correct hallucinations that are correct in all the author names, single author names, and year, respectively. The hallucinations that are (partially) correct in the author names may be useful for finding suggested reading material along with the ground truth paper as they reveal relevant authors. Another finding is that when the prediction is successful in the top-k list, the hallucination tendency in the other predictions drops significantly, the Arxiv Global trained model being the most advantageous, highlighting that the largest model also shows good traits in mitigating hallucinations.

As shown in our ablation study, extending the local citation context with both the citing and cited paper's title and abstract during the continual pre-training does not produce a better result, which can be evaluated counter-intuitive as one has all the information to learn a citation relationship. The missing global information for the cited paper in the test phase complicates finding out the associated citation token.

For future work, we plan to investigate further the all-including configuration in the ablation study. More sophisticated masking strategies besides citation token masking may be helpful. Additionally, we should investigate the potential solutions to the citation-specific hallucinations and tackle a way to reduce the number of hallucinated recommendations in the top k.

Table 8: Examples of hallucination categories. **(a)** No hallucination in any of the top-10 predictions. **(b)** Hallucinated publication years in the fourth, sixth, seventh, and ninth predictions. **(c)** Hallucinated author name in the sixth prediction. Fabricated author list in the ninth prediction. **(d)** Hallucinated author name in the fifth prediction. (A typo in the first author's name). **(e)** Hallucinated author name in the sixth prediction (A single letter as the first author name). **(f)** CiteBART predicts a citation that has the same author name as the ground truth while in a different citation format and publication year. Unlike the other examples, the model's pretraining dataset is different from the dataset associated with the given context.

	Context	Ground Truth	Pretraining Dataset of the Model	Dataset of the Example	Predicted Citations
(a)	... exploits similarity on the target side in another language by extracting source phrases that share common translations **<mask>** , but recent approaches have combined this approach with source phrase ... </s> Example-based Paraphrasing for Improved Phrase-Based Statistical Machine Translation </s> In this article, an original view on how to improve phrase translation estimates is proposed. This proposal is ...	Bannard and Callison-Burch, 2005	ACL-200 Global	ACL-200 Global	1. Callison-Burch et al., 2006 2. Koehn et al., 2003 3. Irvine and Callison-Burch, 2014 4. **Bannard and Callison-Burch, 2005** 5. Quirk et al., 2004 6. Mirkin et al., 2009 7. Irvine and Callison-Burch, 2013 8. Koehn and Knight, 2002 9. Schroeder et al., 2009 10. Koehn and Knight, 2003
(b)	... supertags, the supertagger re-analyses the sentence with a more relaxed beam (adaptive supertagging). A* Parsing **<mask>** a) introduce A* parsing for PCFGs. The parser maintains a chart and an agenda, which is a priority queue of ... </s> A* CCG Parsing with a Supertag-factored Model </s> We introduce a new CCG parsing model which is factored on lexical category assignments. Parsing is then simply a deterministic ...	Klein and Manning, 2003	ACL-200 Global	ACL-200 Global	1. **Klein and Manning, 2003** 2. Auli and Lopez, 2011 3. Ait-Mokhtar and Chanod, 1997 4. Ait-Mokhtar and Chanod, 2005 * 5. Pauls et al., 2009 6. Pauls et al., 2006 * 7. Ait-Mokhtar and Chanod, 2006 * 8. Och, 2003 9. Aitouni et al., 2006 * 10. Clark and Curran, 2004
(c)	... Google Analogy Test Set, which contains 14 types of relations with a varying number of instances per relation **<mask>**, the gger Analogy Test Set , which contains 40 relations with 50 instances per relation, and the ffVec Test Set ... </s> Probabilistic Relation Induction in Vector Space Embeddings </s> Word embeddings have been found to capture a surprisingly rich amount of syntactic and semantic knowledge. However, it is not ...	Mikolov et al., 2013	PeerRead Global	PeerRead Global	1. Vylomova et al., 2015 2. Valenzuela-escárcega et al., 2015 3. Abadi et al., 2016 4. Heinsohn, 2013 5. Holzmann and Risse, 2017 6. Valenzuela-escárárcega et al., 2015 * 7. Davies et al., 2015 * 8. Dinu et al., 2014 9. Holzmann and Riedl, 2016 * 10. Gaunt et al., 2016
(d)	... produces a false positive rate of 0.0027, as noted above, but in a situation where 3 key-value items were being stored per n-gram on average, this error rate would in fact require a storage cost of 60 bits per original n-gram. 2.2.2 Bloomier Filters More recently, **<mask>** have proposed an approach to storing large language models which is based on the Bloomier Filter technique of OTHERCIT. Bloomier Filters generalize the Bloom Filter to allow values ...	Talbot and Brants, 2008	ACL-200 Base	ACL-200 Base	1. **Talbot and Brants, 2008** 2. Talbot and Osborne, 2007 3. Lavoie and Rambow, 1997 4. Pennacchiotti and Pantel, 2009 5. MTalbot and Brants, 2008 * 6. Galanis and Androutsopoulos, 2010 7. Lavoie and Rambow, 2009 * 8. Pennacchiotti and Pantel, 2006 9. Mintz et al., 2009 10. Talbot et al., 2011
(e)	... signature generators can be mislead into generating bad signatures; specifically higher false negative rates. Shield **<mask>** , Vigilante, DACODA, and our own work, all attempt to work around such problems by directly deriving ... </s> A lightweight end-to-end system for defending against fast worms </s> The vulnerabilities which plague computers cause endless grief to users. Slammer compromised millions of hosts in minutes; a hit-list worm ...	Wang et al., 2004	Refseer Global	Refseer Global	1. **Wang et al., 2004** 2. Cui et al., 2007 3. Brumley et al., 2006 4. Brumley et al., 2004 * 5. Dasgupta et al., 2004 6. W et al., 2004 * 7. Shavitt and Tankel, 2003 8. Shavitt and Tanenbaum, 2005 * 9. Daswani and S, 2007 * 10. Chen and Wagner, 2007
(f)	... tab while waiting for the original one to load, i.e., tab switching. More recently, a Web navigation study by **<mask>** found their participants using multiple windows frequently, enabling them to compare search results ... </s> Parallel Browsing Behavior on the Web </s> Parallel browsing describes a behavior where users visit Web pages in multiple concurrent threads. Web browsers explicitly support this by providing tabs. Although parallel browsing ...	Weinreich, 2006	ACL-200 Global	Refseer Global	1. Weinreich et al., 2008 2. Nakagawa and Uchimoto, 2007 * 3. Weinreich et al., 2010 * 4. Navigli and Crisafulli, 2010 5. Nakashole et al., 2012 * 6. Webber et al., 2003 * 7. Lin and Bilmes, 2011 8. Resnik and Smith, 2003 9. Stoica and Hearst, 2004 * 10. Lin and Bilmes, 2008 *

Limitations

We recognize the following limitations in this study. First, CiteBART addresses the task of LCR, and given context with a citation placeholder, it predicts the best candidates for the placeholder. As a citation placeholder indicates that the context is worth citation, CiteBART builds upon the assumption of the citation worthiness of a local context.

Second, CiteBART necessitates pre-training on a specific dataset to recommend citations from the pool of papers in it. Thus, it may omit to cite some work or authors if they are not included in its training corpus. However, unlike the past works, as CiteBART is generative, it can recommend unseen papers, hallucinating. Although the fabricated citations in the top k predictions show that they capture the author names of the ground-truth citations, hallucination is still a problem.

There can be a bias towards citing papers as CiteBART learns from both local context and citing papers. Leveraging all the parts of a citation relationship, citing paper, local context, and cited paper should provide a more balanced learning process once it can be made learning. We leave this possibility for future exploration.

Ethics Statement

CiteBART is a tool to support the scientific community in paper writing; it in no way replaces a researcher or alternates the thoughtful process of choosing the most appropriate references to cite in a local context.

References

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. SciBERT: A pretrained language model for scientific text. 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 3615–3620, Hong Kong, China. Association for Computational Linguistics.

Steven Bird, Robert Dale, Bonnie Dorr, Bryan Gibson, Mark Joseph, Min-Yen Kan, Dongwon Lee, Brett Powley, Dragomir Radev, and Yee Fan Tan. 2008. The ACL Anthology reference corpus: A reference dataset for bibliographic research in computational linguistics. In *Proceedings of the Sixth International Conference on Language Resources and Evaluation (LREC'08)*, Marrakech, Morocco. European Language Resources Association (ELRA).

Davide Buscaldi, Danilo Dessì, Enrico Motta, Marco Murgia, Francesco Osborne, and Diego Recupero. 2024. Citation prediction by leveraging transformers and natural language processing heuristics. *Information Processing & Management*, 61:103583.

Ilias Chalkidis, Manos Fergadiotis, Prodromos Malakasiotis, Nikolaos Aletras, and Ion Androutsopoulos. 2020. LEGAL-BERT: The muppets straight out of law school. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 2898–2904, Online. Association for Computational Linguistics.

Arman Cohan, Waleed Ammar, Madeleine van Zuylen, and Field Cady. 2019. Structural scaffolds for citation intent classification in scientific publications. 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 3586–3596, Minneapolis, Minnesota. Association for Computational Linguistics.

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.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. 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 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Constanza Fierro, Reinald Kim Amplayo, Fantine Huot, Nicola De Cao, Joshua Maynez, Shashi Narayan, and Mirella Lapata. 2024. Learning to plan and generate text with citations. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11397–11417, Bangkok, Thailand. Association for Computational Linguistics.

Tianyu Gao, Howard Yen, Jiatong Yu, and Danqi Chen. 2023. Enabling large language models to generate text with citations. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6465–6488, Singapore. Association for Computational Linguistics.

Nianlong Gu, Yingqiang Gao, and Richard H. R. Hahnloser. 2022. Local citation recommendation with hierarchical-attention text encoder and scibert-based reranking. In *Advances in Information Retrieval*, pages 274–288, Cham. Springer International Publishing.

Kelvin Guu, Kenton Lee, Zora Tung, Panupong Pasupat, and Ming-Wei Chang. 2020. Realm: retrieval-augmented language model pre-training. In *Proceedings of the 37th International Conference on Machine Learning*, ICML'20. JMLR.org.

Wenyi Huang, Zhaohui Wu, Chen Liang, Prasenjit Mitra, and C. Lee Giles. 2015. A neural probabilistic model for context based citation recommendation. In *Proceedings of the Twenty-Ninth AAAI Conference on Artificial Intelligence*, AAAI'15, page 2404–2410. AAAI Press.

Chanwoo Jeong, Sion Jang, Eunjeong Park, and Sungchul Choi. 2020. A context-aware citation recommendation model with bert and graph convolutional networks. *Scientometrics*, 124.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Dongyeop Kang, Waleed Ammar, Bhavana Dalvi, Madeleine van Zuylen, Sebastian Kohlmeier, Eduard Hovy, and Roy Schwartz. 2018. A dataset of peer reviews (PeerRead): Collection, insights and NLP applications. 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 1647–1661, New Orleans, Louisiana. Association for Computational Linguistics.

Yoav Levine, Barak Lenz, Opher Lieber, Omri Abend, Kevin Leyton-Brown, Moshe Tennenholtz, and Yoav Shoham. 2021. Pmi-masking: Principled masking of correlated spans. In *9th International Conference on Learning Representations, ICLR 2021, Virtual Event, Austria, May 3-7, 2021*. OpenReview.net.

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

Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. 2024. The dawn after the dark: An empirical study on factuality hallucination in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand. Association for Computational Linguistics.

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*, abs/1907.11692.

Chu Fei Luo, Rohan Bhambhoria, Samuel Dahan, and Xiaodan Zhu. 2023. Prototype-based interpretability for legal citation prediction. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 4883–4898, Toronto, Canada. Association for Computational Linguistics.

Zoran Medić and Jan Snajder. 2020. Improved local citation recommendation based on context enhanced with global information. In *Proceedings of the First Workshop on Scholarly Document Processing*, pages 97–103, Online. Association for Computational Linguistics.

A Token Limits

Before pre-training with citation objectives, we ensured that each context has its "<mask>" token in its middle position after tokenization. Another critical aspect was the determination of correct lengths for citation contexts. We limited citation contexts in each dataset to an optimal number of tokens to avoid increasing time and memory costs. An exploratory analysis of context lengths shows that the contexts of ACL-200 and Peerread are significantly longer than those of the other datasets. After tokenization, we observed that $200 - 400$ tokens were optimal for all base datasets. This limit allows sufficiently long contexts without a need for excessive amounts of padding tokens. As an exception, ACL-200 has 607 contexts that exceed the 400 limit. We have shortened them to the 400 token limit as they correspond to a small proportion of the whole number of contexts and also because the number of discarded tokens is negligible.

Table 9: Maximum token limits for the preprocessed datasets.

Dataset Name	Base Token Limit	Global Token Limit
ACL-200	400	350
FullTextPeerRead	400	350
Refseer	200	350
Arxiv	300	350

For each global dataset, we chose the token limit as 350. Since abstracts require a higher number of tokens, we limited the local context sizes to 100 for the global versions of the datasets. We also ensured that there are 50 tokens each on the left and right sides of the <mask> tokens. We used a token limit of 200 for abstracts for all datasets since most abstracts can fit into it. Table 9 shows the maximum token limits for both the base and global training schemes.

B Training and Evaluation Times

We conducted our experiments on devices with NVIDIA RTX6000 Ada GPU and NVIDIA V100 GPU for Global and Base datasets, respectively. For global datasets, the pre-training for Peerread and ACL-200 lasts for 2 and 6 hours, respectively. The larger datasets, Arxiv and Refseer, take up to $8-9$ days since they have similar sizes. For base datasets, the training for the smaller datasets, Peerread and ACL-200, lasts for 8 and 20 hours, respectively. The larger datasets, Arxiv and Refseer, take up to 14-15 days. However, we believe these relatively longer times are the result of training on the device with NVIDIA V100 GPU.

Our evaluation of the corresponding test sets takes considerable time since generating the top 10 predictions for each example is resource-intensive. Especially with our limited hardware resources, acquiring the results on the larger datasets takes up to 2 days. The smaller datasets require less time, 20 minutes for Peerread and 2 hours for ACL-200. We performed our evaluations on the device with NVIDIA RTX6000 Ada GPU.

The issue of slow evaluation for larger datasets is not exclusive to our work. Gu et al. (2022) reported their results using only a smaller subsection ($10K$) of the test sets due to long evaluation times.

C Metric Definitions

To evaluate CiteBART, we used the Recall@10, Exact Match and Mean Reciprocal Rank metrics. The past works on citation recommendation have generally used Recall@10 and Mean Reciprocal Rank as evaluation metrics.

Recall@10 is the ratio of the correctly predicted items in the top k recommendations. The benchmark datasets have only one actual target for each context. Therefore, recall@10 measures whether the target citation matches any recommendations in top k.

Exact match (EM) calculates whether the first prediction of the model is the same as the target citation. It is the same as accuracy since there is only one ground-truth citation for each context.

Mean Reciprocal Rank (MRR) considers the position of the ground-truth label in a top-k ranked recommendation list. It is the mean of the reciprocal rank of the correctly recommended citation in the recommendation list. Thus, in Equation 1, U corresponds to the total number of contexts in the dataset (test set size), and i is the position of the ground-truth citation for context u in the top-k results. We used k as 10 in our experiments.

$$MRR = \frac{1}{U}\sum_{u=1}^{U}\frac{1}{rank_i} \tag{5}$$

D Exact Match Scores

Table 10 presents the exact match (EM) scores of CiteBART. While previous studies did not report EM scores, we consider this metric valuable for assessing the model's ability to generate the correct citation on its first attempt. As shown in the table, CiteBART successfully predicts the correct citation directly for a substantial portion of the benchmark datasets.

Table 10: Exact Match (EM) score of CiteBART on LCR benchmarks.

Model	ACL-200 EM	PeerRead EM	Refseer EM	Arxiv EM
CiteBART-Base	**0.422**	0.363	0.382	0.184
CiteBART-Global	0.417	**0.430**	**0.404**	**0.230**

E Qualitative Analysis

To provide insights into the working of CiteBART, we present some top 10 prediction examples. We analyze four different scenarios shown in Table 11. Since CiteBART is a generative model, it is prone to hallucination. In the examples, the hallucinated predictions are designated with the * symbol. Correctly predicted citations are displayed in bold.

We first present an example context that is tested on a model pre-trained on the PeerRead Base dataset. It belongs to the test set of PeerRead Base and receives top 10 citation predictions for the mask. As demonstrated below, the model fails to predict the correct citation in the top 10 predictions. Actually, the ground-truth citation is the $18th$ entry in the ranked prediction list.

In a deeper analysis of the recommended citations for the first example, we bring up their connections with the ground-truth citation. The ground truth citation, "Hu et al., 2015", focuses on sentence-level semantics using convolutional neural networks (CNNs) with an application in dialogue generation. Similarly, the second prediction, "Vinyals and Le, 2015" leverages the sequential structure of sentences in dialogue systems. The fourth prediction, "Serban et al., 2015", also aims to model the hierarchical structure of sentences

Table 11: Four example top-10 citation predictions using CiteBART. Due to space limitations, contexts and abstracts have been abbreviated. The hallucinated predictions are designated with the * symbol.

#	Context	Ground Truth	Pretraining Dataset of the Model	Dataset of the Example	Predicted Citations
1	... Twitter. Previously, a series of NLP tasks have tried to utilize the social annotations like followers , emoticons and responses **<mask>** etc. two kinds of common social labels, i.e., hyper-links and hashtags are leveraged for ...	Hu et al., 2015	PeerRead Base	PeerRead Base	1. Shang et al., 2015 2. Vinyals and Le, 2015 3. Baqapuri, 2015 4. Serban et al., 2015 5. Sordoni et al., 2015 6. Tan et al., 2015 7. Tan et al., 2014 8. Yin and Schutze, 2015 * 9. Dhingra et al., 2016 10. Tan et al., 2016
2	... Twitter. Previously, a series of NLP tasks have tried to utilize the social annotations like followers , emoticons and responses **<mask>** etc. two kinds of common social labels, i.e., hyper-links and hashtags are leveraged for ... </s> TGSum: Build Tweet Guided Multi-Document Summarization Dataset </s> The development of summarization research has been significantly hampered by the ...	Hu et al., 2015	PeerRead Global	PeerRead Global	**1. Hu et al., 2015** 2. Vinyals and Le, 2015 3. Bing et al., 2015 4. Tan et al., 2014 5. Dhingra et al., 2016 6. Xiao and Cho, 2016 7. Qu and Hovy, 2016 * 8. Bing et al., 2014 * 9. Lei et al., 2015 10. Qu and Zuidema, 2015 *
3	... in some latent space. There are many ways to structure G. The DCGAN **<mask>** uses fractionally-strided convolutions to upsample images instead of ... </s> Gang of GANs: Generative Adversarial Networks with Maximum Margin Ranking </s> Traditional generative adversarial networks (GAN) and many of its variants are trained by minimizing the KL or JS-divergence loss ...	Radford et al., 2015	ACL-200 Global	PeerRead Global	1. Kalchbrenner et al., 2014 2. Kalchbrenner and Blunsom, 2013 3. Sha and Pereira, 2003 4. Mikheev et al., 2013 * 5. Finkel et al., 2008 6. Mikheev et al., 1999 7. Gimpel and Smith, 2012 8. Kim et al., 2014 9. Blitzer et al., 2006 10. Henderson, 2004
4	... models to autoregressive models and stochastic variations of neural networks. Among them **<mask>** developed an approach for training a generative model with variational inference by performing ... </s> Learning to Generate Chairs, Tables and Cars with Convolutional Networks </s> We train a generative convolutional neural network which is able to generate images of objects given object type, viewpoint ...	Rezende et al., 2014	Arxiv Global	PeerRead Global	1. **Rezende et al., 2014** 2. Kusner and Hern'andez-lobato, 2016 3. Gregor et al., 2015 4. Mnih and Gregor, 2014 5. Doersch, 2016 6. Kusner and Hern'andez-lobato, 2015 * 7. Ioffe and Szegedy, 2015 8. Lamb et al., 2016 9. Salimans and Kingma, 2016 10. Salimans and Knowles, 2012

(utterances) for building an end-to-end dialogue system. The first prediction, "Shang et al., 2015," is still concerned with capturing sentence connections for a generative motivation. However, the primary reason for its top placement should be related to its experiments on Twitter data since the term Twitter appears in the local citation context. Analogously, the predictions $3, 5, 7$, and 9 utilize Twitter as the data source. Lastly, the model may have proposed the entries 6 and 10 due to their overlaps in authors' names with 7.

The second example has the same context as the first one, but this time, the citing paper's global information (title and abstract) is attached to it. Moreover, the model pre-trained on the PeerRead Global dataset makes the prediction, returning the ground truth citation in the first index. One can observe that the citations "Vinyals and Le, 2015", "Tan et al., 2015", and "Dhingra et al., 2016" still appear in the top-10 prediction list. There are also some hallucinated responses. The newly recommended "Bing et al., 2015" in the third position is also relevant since it tackles constructing sentences from fine-grained textual units.

The third example highlights our model's cross-dataset generalization capability. We input a context from the PeerRead Global dataset into a model pre-trained on ACL-200 Global. The model fails to predict the correct citation as it is missing in the training dataset. Its predictions are NLP papers since ACL-200 is an NLP corpus. On the other hand, PeerRead includes both vision and text papers. The ground-truth citation, "Radford et al., 2015," focuses on image classification using CNNs,

emphasizing unsupervised learning. Our analysis reveals that multiple predicted citations, among the top ten, are relevant to the ground-truth citation. For example, the papers in predictions 1 and 2 also employ CNNs but with a focus on sentence modeling. The papers from predictions 3 and 5 are about conditional random fields (CRFs). While their primary research areas differ significantly from the ground truth, terms such as 'conditional' and 'random' frequently appear in the ground truth paper. Moreover, the paper in Prediction 7 closely aligns with the ground-truth paper by strongly emphasizing unsupervised learning.

The fourth example emphasizes our model's cross-dataset generalization capability from a different perspective. In this example, a model pretrained on the Arxiv Global dataset manages to correctly predict the ground truth citation for a context from the PeerRead Global dataset. Upon closer inspection, we observed that this citation exists in both datasets but with different contexts. CiteBART-Global can predict the correct ground truth citation for an unseen context, leveraging another context citing the same reference.

F Qualitative Analysis on Large Language Models' Performances in LCR

We conducted experiments on a Large Language Model (LLM) to evaluate its performance in local citation recommendation. We prompted the open-source "Llama-2-70b-chat" model for our trials. In each prompt, we first list a set of citation tokens (200, due to the limits of chat windows) from our dataset, followed by a few examples of masked contexts with the corresponding ground truth mask values. Subsequently, we ask the model to fill in the mask for a new context by selecting a citation from the initially provided list.

We present four examples in Figures 1 and 2 to illustrate the workings of the base and global pre-training schemes, respectively. Due to space constraints, we partially display the list of citations, example contexts, and citing abstracts in the prompts. Each example consists of three parts: the prompt, the LLM's answer, and the ground truth value of the masked citation token provided at the end of the prompt.

Figure 1 includes a correct prediction in Part (a) and an incorrect one in (b). Indeed, the correct prediction is the only successful example in several trials using the base approach. The model responds to the prompt by "Shwartz et al., 2016" explaining its choice. On the other hand, the model fills in the mask by "Bahdanau et al., 2016" in Part (b), where "Bluche, 2016" is expected. Its reasoning sheds light on its wrong choice as it strongly associates the term "attention-based mechanisms" in the local context with Bahdanau et al.'s seminal paper on attention-based sequence modeling.

In Figure 2, Part (a) presents a successful example based on the global dataset where the prompt includes the citing paper's title and abstract with the local citation context. The LLM generates the correct citation without an explanation, unlike other predictions. The second example in Part (b) belongs to an incorrect prediction, yet the LLM makes a plausible choice here, judging from its grounding. We can conclude from the observed behavior that LLMs need custom pre-training for the citation tokens to perform well in the task of local citation recommendation.

Our further trials with LLMs demonstrate that they tend not to restrict their predictions to the provided list of citations but to recommend the best choice based on their prior knowledge. They also exhibit a known deficiency. They sometimes ask for confirmation when they provide an answer, and even if you confirm, they lean towards changing the answer. In conclusion, they suffer from hallucinations.

PROMPT:

Here is a list of citations:
"Stalzer, 2017"
"Langford et al., 2008"
"Zhao et al., 2015"
…

There are example context windows taken from scientific papers below. Each of them has a <mask> inside. Actual values of masks have been provided as well. They are filled using one the citations listed above.

An example context window: "… compared with hand-crafted networks.The recent work MetaQNN <mask> and ral Architecture Search with reinforcement learning reported surprising …"
Actual value of the mask: "Baker et al., 2016"
…

Now, find the following context window's actual mask using one of the citations listed above:

An example context window: "… relation between the words . Recently, presented xNET, an extension of HypeNET <mask>. This method integrates both path-based and …"
What is the actual value of the mask?

ANSWER:

The actual value of the mask in the given context window is "**Shwartz et al., 2016**".

Here's how you can identify the correct citation: The context window mentions "xNET" and "HypeNET", which are both neural network models used for semantic relation classification. The authors mention that the approach presented in the paper integrates both path-based and distributional information for semantic relation classification. The paper was published recently, relative to the publication date of the paper containing the context window. All of these clues point to the citation being "Shwartz et al., 2016", which is indeed the correct answer.

Ground Truth Citation: Shwartz et al., 2016

(a) Base approach - Successful example

PROMPT:

Here is a list of citations:
"Stalzer, 2017"
"Langford et al., 2008"
"Zhao et al., 2015"
…

There are example context windows taken from scientific papers below. Each of them has a <mask> inside. Actual values of masks have been provided as well. They are filled using one the citations listed above.

An example context window: "… compared with hand-crafted networks.The recent work MetaQNN <mask> and ral Architecture Search with reinforcement learning reported surprising …"
Actual value of the mask: "Baker et al., 2016"
…

Now, find the following context window's actual mask using one of the citations listed above:

An example context window: "… follow-up work solves this problem with attention based mechanisms <mask> , this will be investigated in future work. The recognizer is trained on both …"
What is the actual value of the mask?

ANSWER:

The actual value of the mask in the given context window is likely to be "**Bahdanau et al., 2016**". This citation is mentioned in the list of references provided earlier and fits the context of the sentence where it mentions "attention-based mechanisms". Therefore, the complete reference for the context window would be:

"… follow-up work solves this problem with attention-based mechanisms Bahdanau et al., 2016, this will be investigated in future work. The recognizer is trained on both …"

Ground Truth Citation: Bluche, 2016

(b) Base approach - Unsuccessful example

Figure 1: Prompt examples on a Large Language Model for Base dataset.

PROMPT:

Here is a list of citations:
"Mirowski et al., 2016"
"Lowe et al., 2017"
"Walsh, 2014"
…

There are example context windows taken from scientific papers below. Each of them has a <mask> inside. Alongside the context windows, the titles and abstracts of citing papers, which contain the context windows, are provided. Actual values of masks have been provided as well. They are filled using one the citations listed above.

An example context window, citing title and abstract: "... that it searches for critical points rather than local optima <mask> . Therefore, it is more common to use natural gradient or updates... </s> A Kronecker-factored approximate Fisher matrix for convolution layers </s> Second-order optimization methods such as natural gradient descent have the potential to speed up ..."
Actual value of the mask: "Pascanu et al., 2014"
…

Now, find the following context window's actual mask using one of the citations listed above:

An example context window, citing title and abstract: "… for generating specific responses, inspired by the semi-supervised deep generative model <mask> . The principle idea is to generate the next response … </s> A Conditional Variational Framework for Dialog Generation </s> Deep latent variable models have been shown to facilitate the response generation for open-domain dialog …"
What is the actual value of the mask?

ANSWER:

The actual value of the mask is "**Kingma et al., 2014**".

Ground Truth Citation: Kingma et al., 2014

(a) Global approach - Successful example

PROMPT:

Here is a list of citations:
"Mirowski et al., 2016"
"Lowe et al., 2017"
"Walsh, 2014"
…

There are example context windows taken from scientific papers below. Each of them has a <mask> inside. Alongside the context windows, the titles and abstracts of citing papers, which contain the context windows, are provided. Actual values of masks have been provided as well. They are filled using one the citations listed above.

An example context window, citing title and abstract: "... that it searches for critical points rather than local optima <mask> . Therefore, it is more common to use natural gradient or updates... </s> A Kronecker-factored approximate Fisher matrix for convolution layers </s> Second-order optimization methods such as natural gradient descent have the potential to speed up ..."
Actual value of the mask: "Pascanu et al., 2014"
…

Now, find the following context window's actual mask using one of the citations listed above:

An example context window, citing title and abstract: " ... stacked Denoising Auto-encoders to extract meaningful representations. <mask> main adaptation from multiple sources. For domain adaptation ... </s> Knowledge Adaptation: Teaching to Adapt </s> Domain adaptation is crucial in many real-world applications where the distribution of the training data differs from the distribution of ..."
What is the actual value of the mask?

ANSWER:
Based on the context window, I'm going to take a guess that the actual value of the mask is "**Nguyen et al., 2016**". My reasoning is as follows:
The context window mentions "domain adaptation" and "learning domain-invariant representations", which suggests that the paper is related to domain adaptation.
The sentence with the mask mentions "main adaptation from multiple sources", which suggests that the paper is discussing a method for adapting to multiple sources or domains.
The citation "Nguyen et al., 2016" is listed in the original list of citations, and the title of the paper is not provided, but the abstract mentions "domain adaptation" and "multiple sources", which matches the context of the context window.

Ground Truth Citation: Chen et al., 2012

(b) Global approach - Unsuccessful example

Figure 2: Prompt examples on a Large Language Model for Global dataset.