Controllable Citation Sentence Generation with Language Models

Nianlong Gu
Linguistic Research Infrastructure,
University of Zurich
nianlong.gu@uzh.ch

Richard H.R. Hahnloser
Institute of Neuroinformatics,
University of Zurich and ETH Zurich
rich@ini.ethz.ch

Abstract

Citation generation aims to generate a citation sentence that refers to a chosen paper in the context of a manuscript. However, a rigid citation generation process is at odds with an author's desire to control specific attributes, such as 1) the citation intent, e.g., either introducing background information or comparing results, and 2) keywords that should appear in the citation text. To provide these degrees of controllability during citation generation, we propose to integrate the manuscript context, the context of the referenced paper, and the desired control attributes into a structured template and use it to fine-tune a language model (LM) via next-token prediction. We then utilize Proximal Policy Optimization to directly optimize the LM in favor of a high score of our proposed controllability metric. The proposed workflow harmoniously combines citation attribute suggestion and conditional citation generation into one LM, allowing for better user control. [1]

1 Introduction

A common practice in scientific writing is to cite and discuss relevant papers that support an argument, provide background information, or compare results (Penders, 2018). Recent studies aim to facilitate this citation process by using neural networks to generate a citation sentence based on the context of the manuscript and the paper to be cited. These efforts focused primarily on developing a sequence-to-sequence pipeline that works in a fully automated, uncontrolled manner, leaving little room for users to control the generation process. We believe control is desirable because authors often have a clear motivation before writing a citation sentence. For example, they may have a specific *intent* to cite, such as comparing results or presenting background information, or they may have *keywords* in mind to appear in the citation sentence. When the generated citation does not match an author's motivation, the author may wish to change the generation by specifying certain attributes, such as citation intent or keywords.

This study aims to develop a citation generation model that is controllable, such that users can alter the citation intent or topic by explicitly providing the citation intent and keywords. Our proposed method involves the following two phases:

Supervised fine-tuning. We design a structured prompt template that systematically and consecutively incorporates contextual information, citation attributes (intent and keywords), and the citation sentence into a sequence of tokens, and fine-tune an LM via next-token prediction. Through supervised fine-tuning, the LM learns to generate citation sentences not only based on the manuscript/cited paper's context but also conditioned on the citation attributes, thus allowing flexible control of generation by altering the citation attributes (Keskar et al., 2019).

Controllability enhancement via reinforcement learning. We propose measuring the controllability of a citation generation system from multiple aspects with the following metrics: i. Intent Alignment Score (IAS), which measures whether the intent of the generated citation sentence matches the given control intent; ii. Keyword Recall (KR), which measures the recall of the control keywords in the generated citation; iii. Fluency Score (FS); and iv. ROUGE-F1 (Lin, 2004) score of the generated sentence compared with the groud-truth citation. These controllability evaluation metrics provide a further opportunity to guide the training of our system by using them to estimate a reward function for Proximal Policy Optimization (PPO) (Schulman et al., 2017). This allows us to explore the effect of using feedback from our chosen metrics to improve our system's controllability.

Our contributions are summarized as follows:

- We present a novel strategy that unifies cita-

[1] Our code and data are available at https://github.com/nianlonggu/LMCiteGen

Proceedings of the Fourth Workshop on Scholarly Document Processing (SDP 2024), pages 22–37

tion attributes and citation sentence generation within one language model, thereby enabling user control of the citation generation process.

- We evaluate the control exerted by the various attributes, employing a multi-metric system that includes an intent matching score, keyword recall, ROUGE score, and a reference-less fluency score, and we use these controllability metrics as a reward for Proximal Policy Optimization (PPO) to effectively improve the controllability of our model following its initial supervised training.

- We curate a comprehensive dataset, parsing both contextual text and citation attributes, to offer a valuable resource for future controllable citation generation research.

2 Related Work

Previous work in citation generation, including Xing et al. (2020); Ge et al. (2021); Wang et al. (2022), approached the task as a sequence-to-sequence translation problem, utilizing recurrent networks and knowledge graph enhancements. Nikiforovskaya et al. (2020) proposed a BERT-based extractive summarizer (Liu, 2019) that produces a paper review by extracting one sentence from each of the related papers. Chen and Zhuge (2019) proposed automatically generating a related work section by extracting information on how papers in the reference list have been cited in previous articles. Xing et al. (2020) developed an RNN-based pointer generator network that can copy words from the manuscript and the abstract of the cited paper based on cross-attention. Ge et al. (2021) further extended this work by enhancing citation generation using information from the citation graph. Despite their advances, these approaches do not fully address the complexities of citation sentence generation. Our research highlights the importance of user-specified attributes and illustrates the limitations of large language models in reliably inferring key attributes such as topic keywords. We propose that this semantic gap necessitates a shift towards models that take explicit user control into account in citation generation.

Recent research Jung et al. (2022); Wu et al. (2021); Yu et al. (2022) has also explored controlled citation generation, though the focus is often limited to controlling a single attribute like citation intent. Our work extends this concept to a broader range of attributes, introducing methods to suggest potential attributes and balance the automation and controllability of citation generation. We further differentiate our research by conducting extensive experiments with cutting-edge language models and investigating the augmentation of controllability via reinforcement learning, contributing to a more comprehensive understanding of controlled citation generation. In addition, we go beyond prompt-based approaches (Yang et al., 2022) by investigating enhancing the controllability of our citation generation system with reinforcement learning.

3 Method

We first describe our supervised method for fine-tuning an LM, and then we describe how we measure the control of the generated sentence by the attributes. Finally, we introduce strategies for enhancing controllability through reinforcement learning. Our method is illustrated in Figure 1.

3.1 LM Supervised Fine-tuning

We fine-tune a language model (LM) to generate a citation sentence $\mathcal{S}$ given the context $\mathcal{C}$ and citation (control) attributes $\mathcal{A}$. The context, denoted as $\mathcal{C}$, consists of information from three distinct sources: 1) the local manuscript context, including up to N_s sentences from the same section (we use $N_s = 5$) that precede the citation sentence to be generated; 2) the manuscript's title and abstract, serving as the global context; and 3) the title and abstract of the paper to be cited, providing an external context.

The citation attributes denoted $\mathcal{A}$, include the combined citation intent and keywords. Following the framework proposed in Cohan et al. (2019), we define three categories of citation intents: 1) *background*: The citation offers context or background information about a pertinent problem, concept, method, or topic. 2) *method*: The citation refers to a specific method, tool, approach, or dataset in the cited paper. 3) *result*: The citation contrasts or compares the results or findings of the manuscript with those in the referenced paper. In terms of keywords, we define keyword attributes as one or two noun phrases extracted from the target citation sentence that bears semantic similarity to the context of either the manuscript or the cited paper.

The training objective is to optimize the LM to maximize the log-likelihood of generating both the citation attributes $\mathcal{A}$ and the citation sentence $\mathcal{S}$

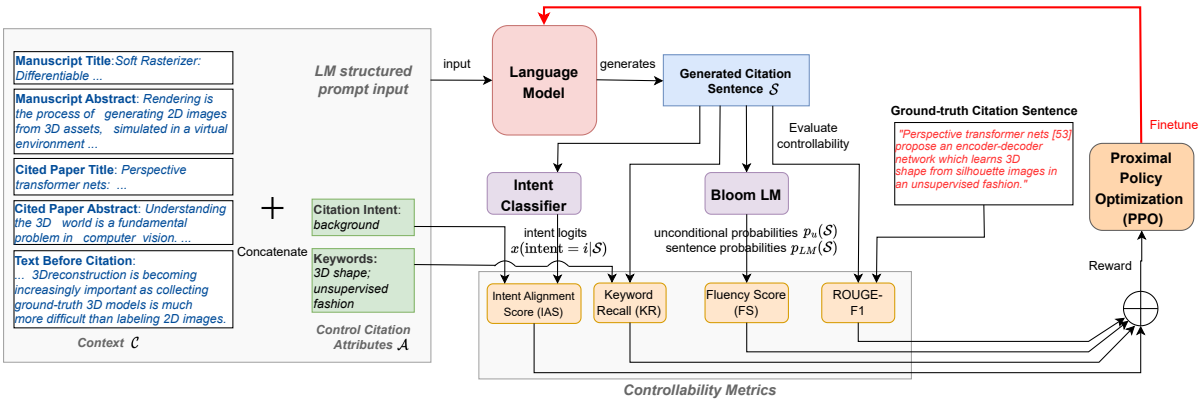

Figure 1: A schematic representation of the workflow for generating citation sentences using a Language Model, with subsequent controllability evaluation, and optimization using Proximal Policy Optimization (PPO) with the summed controllability metrics as a reward.

given the context $\mathcal{C}$. Let a_i and s_i be the tokens in $\mathcal{A}$ and $\mathcal{S}$, with total tokens $|A|$ and $|S|$ respectively, the training objective can be expressed as:

$$\begin{aligned}\hat{\theta} &= \arg\max_{\theta} \log p(\mathcal{S}, \mathcal{A}|\mathcal{C}) \\ &= \arg\max_{\theta} \left[\log p(\mathcal{A}|\mathcal{C}) + \log p(\mathcal{S}|\mathcal{A}, \mathcal{C})\right] \\ &= \arg\max_{\theta} \left[\sum_{i=1}^{|\mathcal{A}|} \log p(a_i|a_1 \ldots a_{i-1}, \mathcal{C}) + \right. \\ &\qquad \left. \sum_{i=1}^{|\mathcal{S}|} \log p(s_i|s_1 \ldots s_{i-1}, \mathcal{A}, \mathcal{C})\right] \end{aligned} \quad (1)$$

Such an objective allows us to fine-tune a single LM for two purposes: 1) inferring citation attributes $\mathcal{A}$ given the context $\mathcal{C}$, and 2) crafting citation sentences $\mathcal{S}$ based on the context $\mathcal{C}$ and control attributes $\mathcal{A}$. This dual-purpose training strategy enables the fine-tuned LM to switch dynamically between controlled and uncontrolled modes: In the **controlled mode**, the LM assimilates the context and the user-specified control attributes to generate the citation sentence. In the **uncontrolled mode**, the LM initiates by automatically inferring possible citation attributes based solely on the context, and uses these inferred attributes to guide itself when generating the citation sentence. This flexibility also allows us to compare the performance of our citation generation model between uncontrolled and controlled modes in a fair manner, since we can compare the performance on the same model just with different working modes.

Structured Input Template for Supervised Fine-tuning. We design an input template to help the LM differentiate between input sources. The template arranges different input components into a unique prompt, including the cited paper's global context, the manuscript's local and global contexts, and the citation attributes, as in Raffel et al. (2020); Keskar et al. (2019); Taori et al. (2023). The final input text is structured as follows:

```
###Manuscript Title: XXX XXX
###Manuscript Abstract: XXX XXX
###Cited Paper Title: XXX XXX
###Cited Paper Abstract: XXX XXX
###Text Before Citation:
        XXX XXX (manuscript local context)
###Citation Intent: XXX (one word from 'background', 'method' and 'result')
###Keywords: XXX; XXX (keywords relevant to the citation, separated by '; ')
###Citation: XXX XXX (target citation sentence)
```

For decoder-only LMs like GPT-NEO (Black et al., 2021), the fine-tuning phase uses the generated prompt as input for next-token prediction, masking context-related tokens (from "`###Manuscript Title: XXX`" to "`###Text Before Citation: XXX`") to confine prediction loss within citation attributes $\mathcal{A}$ and citation sentence $\mathcal{S}$, as per Equation (1). In the inference stage, the LM is fed with the template up to "`###Citation Intent:`", "`###Keywords:`", and "`###Citation:`" when the task is to decode the citation intent, keywords, and the citation sentence, respectively. In contrast, for encoder-decoder LMs like BART (Lewis et al., 2020), context-related tokens are the encoder input during both fine-tuning and inference. During fine-tuning, the LM learns to decode tokens within $\mathcal{A}$ and $\mathcal{S}$. At inference, the

decoder generates citation attributes and sentences based on the provided prompt.

3.2 Controllability Evaluation Metrics

We deem good controllability by control attributes can be reflected in the following aspects:

1. The generated citation sentence matches the given intent. For example, given a control intent "background", the model will generate a sentence that introduces the background of the cited paper.

2. The generated citation sentence contains the given control keywords.

3. The generated sentence is fluent so that the control keywords are embedded into the sentence in a logically coherent way.

4. The generated sentence is content-wise related to the cited paper and fits well with the context of the manuscript.

Along this vein, apart from the human evaluation (Section 6.4), we propose the following four automatic metrics, each corresponding to one aspect:

1. **Intent Alignment Score (IAS)** evaluates the alignment between the generated citation sentence and the specified citation intent. Suppose the control intent is i (one of three possible intents: '*background*', '*method*', and '*result*') and LM generates a citation sentence $\mathcal{S}$, we use SciBERT (Beltagy et al., 2019) to process the citation sentence (preceded by a "[CLS]" token) and compute the last hidden state of "[CLS]", which we input to the intent scoring head, a fully connected two-layer network to yields the intent logits $x(\text{intent} = i|\mathcal{S})$ for three possible intents. The intent alignment score $\text{IAS}(\mathcal{S})$ is given by the probability of the intent i by applying softmax to the logits:

$$\text{IAS}(\mathcal{S}) = \frac{\exp\left(x(\text{intent} = i|\mathcal{S})\right)}{\sum_{k\in\text{all intents}} \exp\left(x(\text{intent} = k|\mathcal{S})\right)} \tag{2}$$

The intent scorer is trained using the SciCite dataset (Cohan et al., 2019) containing human-annotated intents. Details on training and evaluating are further described in Appendix A.

2. **Keyword Recall (KR)** is a metric that assesses the presence of provided keywords in the generated citation sentence. A higher value signifies that the generated sentence includes the given keywords, indicating good control over keyword incorporation. Given a generated citation sentence $\mathcal{S}$, the keyword recall $\text{KR}(\mathcal{S})$ is calculated using ROUGE-L recall to compare the keyword attribute (or a concatenated string of multiple keywords, if applicable) with $\mathcal{S}$.

3. **Fluency Score (FS)** evaluates the fluency of the generated citation sentence $\mathcal{S}$. In addition to keyword recall, the fluency metric ensures that the citation generator incorporates the keyword attributes naturally and logically without compromising fluency. Drawing inspiration from Kann et al. (2018), we employ SLOR, a language model-based fluency metric, and we normalize this score using a sigmoid function. $\text{FS}(\mathcal{S})$ is calculated as follows:

$$\begin{aligned}\text{FS}(\mathcal{S}) &= \frac{1}{1+e^{-(\text{SLOR}(\mathcal{S})-\eta)}} \\ \text{SLOR}(\mathcal{S}) &= \frac{1}{|\mathcal{S}|}(\log(p_{LM}(\mathcal{S})) - \log(p_u(\mathcal{S})))\end{aligned} \tag{3}$$

In this equation, $|\mathcal{S}|$ denotes the number of tokens in sentence $\mathcal{S}$, $p_{LM}(\mathcal{S})$ represents the probability of generating the sentence with a pre-trained language model, and $p_u(\mathcal{S})$ is the product of the unconditional probabilities of all tokens in the sentence. The unconditional probability of a token refers to the probability of that token being generated as the first token in a sentence. As $\text{SLOR}(\mathcal{S})$ does not naturally range between [0,1], we apply an offset η and a sigmoid function to normalize the score. The offset η is introduced to the SLOR scores before sigmoid normalization, allowing greater distinguishability between fluency scores for fluent and less fluent sentences. We set the offset η to 4 based on empirical observations. We utilize the 560m-Bloom (Scao et al., 2023) language model to calculate the SLOR because its expansive vocabulary of 250k tokens helps to avoid over-segmenting words into subtokens.

4. **ROUGE-F1** measures the textual alignment between the generated citation sentence and the ground truth. A high score is desired, as it implies that the produced sentence is informative and contextually fitting. We used ROUGE-1,2,L F1 scores also to validate the effectiveness of citation attributes in controlling generation.

3.3 Controllability Enhancement via PPO

To enhance the controllability of our citation generator, we opted for Proximal Policy Optimization (PPO) (Schulman et al., 2017; Ramamurthy et al., 2023) due to its capability to use controllability

metrics (IAS, KR, FS, ROUGE) as rewards, facilitating a direct optimization of the LM. While cross-entropy loss during supervised fine-tuning does provide some controllability, it does not consistently ensure that specific controllability criteria are satisfied. To illustrate, consider the ground-truth sentence centered on keywords "**L1; sparse parameters**": "*L1 regularization leads to sparse parameters after training.*" Two possible generated sentences are:

A. "*L1 optimization leads to dense parameters after training.*" and

B. "*We show that L1 regularization results in sparse parameters in the model's learned weights.*"

Despite sentence A seemingly being favored by cross-entropy loss, it misinterprets the keyword "sparse" and thus alters the sentence's meaning completely. Conversely, sentence B aptly captures the essence, emphasizing keyword recall. This underscores the limitation of relying solely on cross-entropy for optimal controllability. PPO's adaptability, allowing for controllability metrics as rewards, ensures the language model's outputs are both accurate and controllable, offering a more nuanced and flexible optimization strategy than other prevalent methods like adjusting beam search or 'bag of words' techniques (Pascual et al., 2020; Dathathri et al., 2020).

We used the parameters of the supervised fine-tuned model to initialize the LM. During the PPO training process, given the context and citation attributes used as input, the LM generated a batch of citation sentences, which we refer to as "episodes". These episodes were evaluated using the reward:

$$R = \frac{1}{4}\Big(\text{IAS}(\mathcal{S})+\text{KR}(\mathcal{S})+\text{FS}(\mathcal{S})+\text{RS}(\mathcal{S})\Big), \quad (4)$$

where $\text{RS}(\mathcal{S})$ is the sum of ROUGE-1/2/L F1 scores, allowing its magnitude to be comparable with the other metrics. Importance sampling was used during the optimization step, with mini-batches of episodes sampled along with their associated rewards, to estimate the expected return under the new policy using data collected with the old policy, aiming to minimize divergence from the previous policy while improving the expected return. We implemented the PPO training using the Transformer Reinforcement Learning (TRL) framework (von Werra et al., 2020).

Information	Training	Validation	Test
# samples	233,616	1,299	1,080
# citing papers	120,425	1,175	1,005
# cited papers	69,664	998	846

Table 1: The statistics of our dataset.

4 Dataset

From a subset of arXiv computer science papers, we extracted triplets consisting of the citing paper (treated as the manuscript), the citation sentence within it, and the corresponding cited paper. The necessary components for the input context, including the local and global contexts of the manuscript, as well as the global context of the cited paper, were then extracted from these triplets.

Regarding the citation attributes, we used the SciBERT-based intent scorer, as outlined in Section 3.2, to predict the most probable citation intent for each citation sentence. To obtain the keywords attribute, we extracted noun phrases from each citation sentence and ranked them based on the cosine similarity between the keyword's Sentence-BERT (version "all-mpnet-base-v2", Apache License 2.0) (Reimers and Gurevych, 2019) embeddings and the embeddings of both the manuscript and the cited paper. This approach allowed us to retrieve up to two keywords per citation sentence.

To create the training, validation, and test sets, we used a chronological split. All the citing papers used in the training set were published before March 1st, 2023, while those in the validation and test sets were published after this date. This strategy prevents any unfair advantage for the tested language models (Section 5) by ensuring they have not previously encountered the same papers and their associated citation sentences in the validation and test sets during pretraining. The statistics of our dataset are shown in Table 1.

5 Experiment

We experimented with the encoder-decoder model BART (Lewis et al., 2020) and decoder-only models including GPT-Neo (Black et al., 2021), Galactica (Taylor et al., 2022), and LLaMa (Touvron et al., 2023), which varied in size from 125M to 7B parameters. All models were pretrained with corpora before March 1st, 2023. During supervised fine-tuning, the smaller models (125M to 1.3B parameters) were fine-tuned in float16 preci-

sion with a learning rate of 1e-5. For larger models like Galactica-6.7B and LLaMa-7B, we utilized INT4 low-precision quantization (Dettmers et al., 2023) and Low-Rank Adaptation (LoRA) (Hu et al., 2021) to reduce the GPU memory footprint. We set the LoRA parameters with the rank $r = 16$ and the scaling factor $\alpha = 32$. For these larger models, we used a higher learning rate of 1e-4, in line with the guidance from Taori et al. (2023); von Werra et al. (2020). All models were optimized with the AdamW optimizer, using default betas (0.9, 0.999), weight decay 0.05, and cosine learning rate decay. Models were fine-tuned for 5k steps with a batch size of 128 with a maximum token length of 1024. We selected the best model checkpoint based on the validation set loss every 1k steps.

To enhance controllability, we subjected Galactica and LLaMa to additional fine-tuning using Proximal Policy Optimization (PPO), as described in Section 3.3. During this process, the language models generated batches of citation sentences (termed 'episodes') for subsequent mini-batch-wise backpropagation. We set the learning rate to 1.41e-5 for all PPO fine-tuning. The PPO-finetuned models are named Galactica-125M-PPO, Galactica-6.7B-PPO, and LLaMa-7B-PPO, respectively (Table 3). More PPO hyperparameter details and hardware requirements are described in Appendix B.

In addition to fine-tuning models, we leveraged GPT-3.5-turbo-0301[2], the backbone of ChatGPT, through prompt engineering. The aim was to compare our fine-tuned models' performance against a large language model with carefully crafted prompts. We designed specific prompts (detailed in Appendix C) for the ChatGPT API to generate citation sentences. We conducted this in controlled (providing context and citation attributes) and uncontrolled (providing only context) modes.

6 Results and Discussion

We aimed to investigate several key questions: 1) whether and to what extent controllability offers advantages in citation generation, 2) whether PPO helps to improve controllability, and 3) how the model size and the nature of pretraining tasks influence overall performance. In addition, we conducted a human evaluation to compare the controllability of our fine-tuned citation generator against GPT-3.5. This evaluation provided insights beyond the automatic metrics discussed in Section 3.2.

[2] https://platform.openai.com/docs/models/overview

Model	Intent precision	Keyword ROUGE-1 (%)		
		precision	recall	F1
BART-base-140M	0.6083	22.05	16.70	17.62
BART-large-400M	<u>0.6454</u>	24.92	18.47	19.68
GPT-Neo-125M	0.5861	21.10	16.36	17.13
GPT-Neo-1.3B	0.6352	28.00	23.18	23.58
Galactica-125M	0.6204	26.15	21.86	22.11
Galactica-1.3B	**0.6602**	**29.89**	**25.53**	**25.86**
Galactica-6.7B	0.6380	<u>29.49</u>	<u>24.78</u>	<u>25.10</u>
LLaMa-7B	0.6352	28.13	22.78	23.40

Table 2: Performance of LMs on citation attribute inference given context. We present ROUGE-1 metrics for keyword predictions in relation to the ground-truth keywords of the target citation sentence, alongside the precision of intent prediction, represented as the proportion of correctly inferred intents in the test set. The top scores are bolded, and the next best are underlined.

6.1 Controllability Is Necessary

We assessed LMs in three modes: 1) uncontrolled mode, where the LMs utilize only the given context to infer citation attributes, and then use the inferred attributes to guide themselves when generating citation sentences; 2) intent-controlled mode, where we provide the gold citation intent to control generation while the keywords are still model-inferred; and 3) intent- and keywords-controlled mode, where the LMs are given all relevant input: context, citation intent, and keywords, from which the LMs generate citation sentences taking into account all available information.

Given only the context as input (same as the setting in Xing et al. (2020); Ge et al. (2021)), we observed limited success of LMs in matching inferred attributes with ground-truth attributes extracted from the target citation sentence (Table 2). Even the best-performing model, Galactica-1.3B, achieves an intent prediction precision of just 0.66 and a keyword ROUGE-1 F1 of only 0.26. Consequently, misinterpreted citation attributes can lead to off-topic generated citation sentences, as reflected by low ROUGE scores achieved by all LMs in the uncontrolled mode (refer to Table 3).

Intriguingly, a marked improvement in ROUGE scores was already observed when LMs generated citations merely with gold citation intent in the intent-controlled mode (Table 3). Even though the citation intent is not an explicit part of the citation sentence, its presence guides LMs to generate sentences that are more aligned with the tar-

Model	Uncontrolled generation			Intent-controlled generation			Intent- and keywords-controlled generation					
	R1	R2	RL	R1	R2	RL	R1	R2	RL	IAS	KR	FS
BART-base-140M	25.49	4.26	18.28	26.05	4.52	18.71	31.63	8.79	22.74	0.6789	0.6444	0.7156
BART-large-400M	27.39	5.67	19.85	27.90	6.00	20.17	32.33	9.12	23.20	0.6521	0.5877	0.7510
GPT-Neo-125M	23.54	3.67	17.58	23.62	3.69	17.59	30.48	9.44	22.83	0.6252	0.6793	0.7996
GPT-Neo-1.3B	28.48	6.12	20.78	29.04	6.39	21.28	36.26	13.48	26.81	0.7018	0.7936	0.7595
Galactica-125M	28.03	5.77	20.23	28.70	6.27	20.96	35.67	13.07	26.50	0.7037	0.7914	0.7540
Galactica-1.3B	30.07	7.34	22.06	30.66	7.62	22.64	38.06	15.21	28.50	0.6925	0.8299	0.7399
Galactica-6.7B	**30.61**	**7.97**	**22.59**	30.89	8.03	**22.87**	38.29	15.58	28.70	0.6734	0.8150	0.7468
LLaMa-7B	30.19	7.28	22.13	30.49	7.46	22.32	37.71	14.80	28.30	0.6688	0.8380	0.7584
Galactica-125M-PPO	–	–	–	28.81	6.12	20.97	36.49	13.55	27.09	**0.7273**	0.8313	0.7651
Galactica-6.7B-PPO	–	–	–	**31.00**	**8.16**	22.85	**38.49**	**15.81**	**28.98**	0.6740	0.8334	0.7519
LLaMa-7B-PPO	–	–	–	30.64	7.64	22.51	37.72	14.83	28.31	0.6769	**0.8430**	0.7591
GPT-3.5-turbo	23.04	3.88	14.93	23.92	3.61	15.66	29.10	8.11	18.97	0.5716	0.8420	**0.8493**

Table 3: Performance comparison of various language models (LMs) in citation generation across three operational modes: uncontrolled, intent-controlled, and intent- and keywords-controlled. The table lists ROUGE F1 scores (R1, R2, RL) in percentages, as well as Intent Alignment Score (IAS), Keyword Recall (KR), and Fluency Score (FS), with higher scores indicating superior performance. IAS, KR, and FS definitions are provided in Section 3.2.

get citations, as evidenced by enhanced ROUGE scores. This enhancement is further amplified when ground-truth keywords are provided along with the citation intent, with some models witnessing a doubling in ROUGE-2 F1 scores.

Case Study. We further demonstrated the effectiveness of controllability with a case study in Table 4. In the uncontrolled mode, the generated citation sentence was a background sentence and semantically mismatched with the gold citation. By just providing the citation intent "result", the generated sentence presented the results accurately that matched well with the gold citation. Finally, introducing the control keyword "policy iteration-based algorithms" further contributed to accurately including the keyword in the generated citation sentence, yielding the highest ROUGE-L F1. Our results underscore the importance of citation attribute controllability in citation generation.

6.2 PPO Enhances LM Controllability

We compared the LMs' performance in the intent-controlled and intent- and keywords-controlled modes before and after PPO. We excluded the uncontrolled mode from this comparison, as we assume that post-PPO, the LMs consistently operate in a controlled mode where citation attributes are given. Comparing Galactica-125M-PPO with Galactica-125M, Galactica-6.7B-PPO with Galactica-6.7B, and LLaMa-7B-PPO with LLaMa-7B, we observed a consistent and marked improvement in the IAS, KR, FS, and ROUGE F1 scores, especially for the Galactica-6.7B-PPO model, which achieved clear improvement in all metrics and the best performance in terms of ROUGE-F1. This result underscores the efficacy of PPO in directing the language model to adhere to the specifications of the provided attributes more effectively. Consequently, the generated citations better align with the specified intent, more comprehensively incorporate attribute keywords, and exhibit improved fluency, suggesting the effectiveness of PPO in enhancing the controllability of the LM citation generator.

6.3 Model Size and Pretraining Matter

A noticeable trend of improved performance accompanies the increase in language model (LM) size, underlining the role of model scale in citation generation tasks. Additionally, notable variations in performance exist among LMs of identical sizes. Galactica consistently outperforms its counterparts, while GPT-Neo tends to underperform. Despite both models being based on the Transformer's decoder, the performance disparity can be ascribed to their distinct pretraining datasets: GPT-Neo is pretrained on a diverse corpus with only a minor portion of scientific literature, whereas Galactica's pretraining is on a large corpus of scientific texts. This result indicates that the specificity of pretraining data can significantly enhance citation generation performance, underlining the critical influence of pretraining corpora on performance.

cited title: Reinforcement Learning: A Survey

cited abstract: This paper surveys the field of reinforcement learning from a computer-science perspective. It is written to be accessible to researchers familiar with machine learning. Both the historical basis of the field and a broad ...

text before citation: ... the agent is trained against multiple environments simultaneously and gains shared experience, leading to faster convergence and improved performance. Besides, NNDP approach is value iteration–based RL algorithm, whereas our approach is policy iteration–based RL algorithm. ***»generate a citation sentence HERE«***

Ground truth citation sentence: In general, the policy iteration-based algorithms converge faster than value iteration-based algorithms #REFR, which is another reason for the superior performance of our approach. Our experimental results further support our arguments.

Generated citation sentences of different modes:

Mode: Uncontrolled, **Intent**: N/A, **Keywords**: N/A, **ROUGE-L F1**: 24.14
Value iteration-based RL algorithm learns the value function of each state, whereas policy iteration-based RL algorithm learns the policy of each state #REFR.

Mode: Intent-controlled, **Intent**: result, **Keywords**: N/A, **ROUGE-L F1**: 34.62
It has been shown that policy iteration-based RL algorithm converges faster than value iteration-based RL algorithm #REFR.

Mode: Intent- and keywords-controlled, **Intent**: result, **Keywords**: policy iteration-based algorithms, **ROUGE-L F1**: 50.00
Policy iteration-based algorithms are known to converge faster than value iteration-based algorithms #REFR.

Table 4: A case study shows that citation sentences generated by Galactica-6.7B-PPO are guided by the citation intent and the keywords.

Metric	User Preference (%)		
	GPT-3.5-turbo	Neutral	Galactica-6.7B-PPO
Intent Alignment	17.1	57.1	**25.7**
Keyword Recall	15.2	67.6	**17.1**
Fluency	**34.3**	50.5	15.2
Similarity to GT	24.8	30.5	**44.8**

Table 5: Percentage distribution of user preferences for citation sentences generated by GPT-3.5-turbo and Galactica-6.7B-PPO across four criteria. “Neutral” indicates an equal preference for both sentences. Values in bold denote the model with a higher preference.

6.4 Human Evaluation

Intriguing disparities emerge in our comparative analysis of GPT-3.5-turbo and the fine-tuned LM Galactica-6.7B-PPO (Table 3). Despite the second-best KR score and the best FS, GPT-3.5-turbo’s performance falls short in ROUGE-F1 and IAS metrics. Such a disparity in automatic controllability metrics poses challenges in comparing the performance and controllability between models. To this end, a deeper understanding of these discrepancies was sought through a user study involving a subset of 105 examples randomly sampled from the test set. For each example, two citation sentences were presented - one generated by Galactica-6.7B-PPO, the other by GPT-3.5-turbo with the structured prompt (Appendix C), and the presentation order was randomized to avoid bias. Four voluntary participants (PhD candidates with NLP research background, which matches well with the domain of our test dataset) were asked to express their preference using a four-criterion scale (intent alignment, keyword recall, fluency, and similarity to the ground truth), mirroring our automatic metrics (IAS, KR, FS, and ROUGE). A "no preference" option was also provided (Figure 4) in case participants have equal preference for both sentences.

The results from the user study (Table 5) aligned with our automatic evaluations (Table 3). GPT-3.5-turbo-generated sentences, while preferred for fluency, often diverged significantly from the ground-truth citation sentences. This suggests that GPT-3.5-turbo struggles to generate contextually accurate citation sentences using our prompt template despite its prompting capabilities. Thus, high-capacity models like GPT-3.5-turbo may require further prompt refinement or few-shot tuning to enhance citation generation performance.

7 Conclusion

In this study, we introduced a controllable citation generation framework that leverages language models, highlighting the importance of user-specified attributes in the generation process. We emphasized the necessity for attribute control, underlining the complexities of citation generation, and explored the potential of enhancing controllability through Proximal Policy Optimization (PPO). Our experiments affirmed that large language models pretrained on scientific corpora are essential for citation generation, with the fine-tuned model showing advantages over GPT-3.5-turbo in both automatic metrics and human evaluations.

Acknowledgements

We acknowledge support from the Swiss National Science Foundation (grant 31003A_182638) and the NCCR Evolving Language, Swiss National Science Foundation Agreement No. 51NF40_180888. We also thank the anonymous reviewers for their useful comments.

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.

Sid Black, Leo Gao, Phil Wang, Connor Leahy, and Stella Biderman. 2021. GPT-Neo: Large Scale Autoregressive Language Modeling with Mesh-Tensorflow. If you use this software, please cite it using these metadata.

Jingqiang Chen and Hai Zhuge. 2019. Automatic generation of related work through summarizing citations. *Concurrency and Computation: Practice and Experience*, 31(3):e4261. Number: 3 _eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/cpe.4261.

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.

Sumanth Dathathri, Andrea Madotto, Janice Lan, Jane Hung, Eric Frank, Piero Molino, Jason Yosinski, and Rosanne Liu. 2020. Plug and Play Language Models: A Simple Approach to Controlled Text Generation. *arXiv:1912.02164 [cs]*. ArXiv: 1912.02164.

Tim Dettmers, Artidoro Pagnoni, Ari Holtzman, and Luke Zettlemoyer. 2023. Qlora: Efficient finetuning of quantized llms.

Yubin Ge, Ly Dinh, Xiaofeng Liu, Jinsong Su, Ziyao Lu, Ante Wang, and Jana Diesner. 2021. BACO: A Background Knowledge- and Content-Based Framework for Citing Sentence Generation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1466–1478, Online. Association for Computational Linguistics.

Edward J. Hu, Yelong Shen, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, and Weizhu Chen. 2021. Lora: Low-rank adaptation of large language models.

Shing-Yun Jung, Ting-Han Lin, Chia-Hung Liao, Shyan-Ming Yuan, and Chuen-Tsai Sun. 2022. Intent-controllable citation text generation. *Mathematics*, 10(10):1763.

David Jurgens, Srijan Kumar, Raine Hoover, Dan McFarland, and Dan Jurafsky. 2018. Measuring the evolution of a scientific field through citation frames. *Transactions of the Association for Computational Linguistics*, 6:391–406.

Katharina Kann, Sascha Rothe, and Katja Filippova. 2018. Sentence-level fluency evaluation: References help, but can be spared! In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 313–323, Brussels, Belgium. Association for Computational Linguistics.

Nitish Shirish Keskar, Bryan McCann, Lav R. Varshney, Caiming Xiong, and Richard Socher. 2019. Ctrl: A conditional transformer language model for controllable generation.

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.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Yang Liu. 2019. Fine-tune bert for extractive summarization. *ArXiv*.

Anna Nikiforovskaya, Nikolai Kapralov, Anna Vlasova, Oleg Shpynov, and Aleksei Shpilman. 2020. Automatic generation of reviews of scientific papers. In *2020 19th IEEE International Conference on Machine Learning and Applications (ICMLA)*, pages 314–319.

Damian Pascual, Beni Egressy, Florian Bolli, and Roger Wattenhofer. 2020. Directed beam search: Plug-and-play lexically constrained language generation. *CoRR*, abs/2012.15416.

Bart Penders. 2018. Ten simple rules for responsible referencing. *PLOS Computational Biology*, 14(4):e1006036.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text

transformer. *Journal of Machine Learning Research*, 21(140):1–67.

Rajkumar Ramamurthy, Prithviraj Ammanabrolu, Kianté Brantley, Jack Hessel, Rafet Sifa, Christian Bauckhage, Hannaneh Hajishirzi, and Yejin Choi. 2023. Is reinforcement learning (not) for natural language processing: Benchmarks, baselines, and building blocks for natural language policy optimization.

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*. Association for Computational Linguistics.

BigScience Workshop: Teven Le Scao, Angela Fan, and et al. 2023. Bloom: A 176b-parameter open-access multilingual language model.

John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. Proximal policy optimization algorithms.

Rohan Taori, Ishaan Gulrajani, Tianyi Zhang, Yann Dubois, Xuechen Li, Carlos Guestrin, Percy Liang, and Tatsunori B. Hashimoto. 2023. Stanford alpaca: An instruction-following llama model. `https://github.com/tatsu-lab/stanford_alpaca`.

Ross Taylor, Marcin Kardas, Guillem Cucurull, Thomas Scialom, Anthony Hartshorn, Elvis Saravia, Andrew Poulton, Viktor Kerkez, and Robert Stojnic. 2022. Galactica: A large language model for science.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, Aurelien Rodriguez, Armand Joulin, Edouard Grave, and Guillaume Lample. 2023. Llama: Open and efficient foundation language models.

Leandro von Werra, Younes Belkada, Lewis Tunstall, Edward Beeching, Tristan Thrush, and Nathan Lambert. 2020. Trl: Transformer reinforcement learning. `https://github.com/lvwerra/trl`.

Yifan Wang, Yiping Song, Shuai Li, Chaoran Cheng, Wei Ju, Ming Zhang, and Sheng Wang. 2022. Disencite: Graph-based disentangled representation learning for context-specific citation generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 36(10):11449–11458.

Jia-Yan Wu, Alexander Te-Wei Shieh, Shih-Ju Hsu, and Yun-Nung Chen. 2021. Towards generating citation sentences for multiple references with intent control.

Xinyu Xing, Xiaosheng Fan, and Xiaojun Wan. 2020. Automatic Generation of Citation Texts in Scholarly Papers: A Pilot Study. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 6181–6190, Online. Association for Computational Linguistics.

Intent Category (# samples)	Background (1,014)	Method (613)	Result (260)	Average (Macro)
Jurgens et al. (2018)	84.7	74.7	78.2	79.2
Cohan et al. (2019)	87.8	84.9	79.5	84.0
SciBERT+scaffolds (Ours)	**89.1**	**87.1**	**84.0**	**86.7**

Table 6: The F1 scores on three citation intent categories and the average (macro) F1, tested on the SciCite dataset created by Cohan et al. (2019).

Kexin Yang, Dayiheng Liu, Wenqiang Lei, Baosong Yang, Mingfeng Xue, Boxing Chen, and Jun Xie. 2022. Tailor: A prompt-based approach to attribute-based controlled text generation.

Mengxia Yu, Wenhao Yu, Lingbo Tong, and Meng Jiang. 2022. Scientific comparative argument generation.

A Training and Evaluation of SciBERT-based Intent Scorer

To train the SciBERT-based intent scorer, we employ a multitask training strategy (Cohan et al., 2019) with the main task being citation intent classification, which aims to minimize the cross-entropy loss:

$$\mathcal{L} = -\log \frac{\exp\left(x_{\text{intent}}(i_{\text{true}})\right)}{\sum_{i \in \text{all intents}} \exp\left(x_{\text{intent}}(i)\right)} \quad (5)$$

Additionally, we incorporate two auxiliary classification tasks (scaffolds) (Cohan et al., 2019) to enhance the main task's performance: 1) classifying the title of the section (from 5 normalized section titles: Introduction, Related Work, Method, Experiments, Conclusion) in which the cited sentence appears, and 2) determining the citation worthiness of the sentence. We utilize a separate functional head, a two-layer fully connected network for each auxiliary task, with the SciBERT-encoded "CLS" hidden states as input. The training loss is a weighted sum of the three cross-entropy losses (Equation (5)), with weights of 1.0 for the main task, and 0.05 and 0.01 for the first and second auxiliary tasks, respectively. The accuracy of the intent scorer is shown in Table 6.

B Hardware Requirements and PPO Training Hyperparameters

We used 4x NVIDIA A100 80 GPUs for training and a single NVIDIA RTX A6000 for inference. During PPO training, we adjusted the batch and mini-batch sizes according to the model size and architecture. Specifically, we used (256, 16) for Galactica-125M, (256, 4) for Galactica-6.7B, and

(32, 2) for LLaMa-7B, and we ran the PPO steps until no further improvement in reward.

C Prompt Templates for Querying GPT-3.5-turbo API

The prompt templates utilized to query the GPT-3.5-turbo (version 0301) API are presented in this section. As illustrated in Listing 1, the template for uncontrolled citation generation creates a citation sentence solely based on the context. In contrast, Listing 2 demonstrates the template for the controlled citation generation mode, where citation attributes are included alongside the context to guide the generation process. Before sending the query message to the API, we need to provide the actual content of the parameters that are listed at the beginning of the templates, such as the cited paper's title and abstract, the manuscript's title and abstract, and the local context (the text before the target citation sentence in the manuscript), and specify the citation intent and keywords in the controlled generation mode. During the API call to GPT-3.5-turbo, we configured the maximum token limit to 2k and set the generation temperature at 0.1.

D Influence of Beam Size on Citation Generation Performance

During the citation generation process with Language Models (LMs), we conducted experiments with multiple beam sizes: 1, 2, 4, and 8. The performance metrics derived from the validation set suggest that the use of a beam size of 1 yields the most consistent results. Interestingly, larger beam sizes do not enhance the performance, but negatively influence the ROUGE scores (Table 7). These results could hint that larger beam sizes introduce a broader diversity in the generated text, which may affect its precision. For instance, it could impact the accurate generation of specific topic keywords and affect the control over the generation process. In a scenario where a keyword citation attribute is provided, LMs operating with a larger beam size may opt to generate a citation sentence encompassing synonyms of the keywords rather than the keyword itself, which could consequently lead to a decrease in ROUGE scores. As a result, in our experiments, we defaulted to a beam size of 1. We will delve deeper into this phenomenon in future research.

E Citation Generation Examples

We showcase instances (Figure 2 and 3) of citation sentences produced by Galactica-6.7B-PPO in three modes: 1) Uncontrolled, 2) Intent-Controlled, and 3) Intent- and Keywords-Controlled. Notably, Figure 2 is the complete example of the case study in Table 4. With the provision of accurate citation intent alone, the language model citation generator can align well with the stated intent, thereby enhancing ROUGE F1 scores. The addition of keyword attributes further boosts these scores.

F Web Interface for Human Evaluation

We have designed a web interface using Streamlit[3] to assess the citation sentences produced by Galactica-6.7B-PPO and GPT-3.5-turbo. As shown in Figure 4, this user-friendly platform enables participants to peruse context details, citation attributes, the original citation, and the model-generated sentences with ease. Upon rating the generated sentences based on the four provided criteria, participants can submit their evaluation by clicking on the "Submit" button, thereby preserving the data. A "Skip" button is also available, allowing participants to bypass any examples that fall outside their area of expertise and proceed to the next one.

[3] https://streamlit.io/

```
cited_paper_title = "****Cited paper's title****"
cited_paper_abstract = "****Cited paper's abstract****"
manuscript_title = "****Manuscript's title****"
manuscript_abstract = "****Manuscript's abstract****"
manuscript_local_text_before_citation = "****The sentences before the target
    citation sentence in the manuscript (local context)****"

messages = [
    {
        "role": "system",
        "content": "You are a scientific writing assistant. Your task is to infer
    the citation intent and relevant keywords based on the provided context, and
    generate a citation sentence for a given manuscript. The citation sentence
    should seamlessly follow the local context, reflect the inferred citation intent
    , and incorporate the inferred keywords."
    },
    {
        "role": "user",
        "content": f"""
The authors need to cite the reference paper:

Title: {cited_paper_title}
Abstract: {cited_paper_abstract}

, in the manuscript with the global context:

Title: {manuscript_title}
Abstract: {manuscript_abstract}

, immediately following the provided local context:

{manuscript_local_text_before_citation}

Your task:
Please generate a citation sentence that cites the reference paper and seamlessly
    follows the local context. The citation sentence should implicitly reflect one
    of the following citation intents and incorporate relevant keywords:

1) Background: The citation provides background information or additional context
    about a relevant problem, concept, approach, or topic.
2) Method: The citation refers to the use of a specific method, tool, approach, or
    dataset from the reference paper.
3) Result: The citation compares or contrasts the results or findings of the
    manuscript with those in the reference paper.

Requirements:
1. Insert the citation marker "#REFR" at the position in the sentence where the
    reference paper should be cited.
2. Put the citation marker "#REFR" correctly in the generated citation sentence. The
     citation marker should replace the entire in-text citation (e.g., authors and
    year of publication), should not be enclosed in any brackets, and should be
    placed within the sentence before the ending punctuation.

Please return only the generated citation sentence.
            """
    },
]
```

Listing 1: The presented template was employed for querying the GPT-3.5-turbo (version 0301) API instructing it to generate a citation sentence given the context which comprises the title and abstract of the cited paper the title and abstract of the manuscript and the local context (sentences before the target citation in the manuscript). The generation was executed in an uncontrolled mode without the control of explicit citation attributes.

```
cited_paper_title = "****Cited paper's title****"
cited_paper_abstract = "****Cited paper's abstract****"
manuscript_title = "****Manuscript's title****"
manuscript_abstract = "****Manuscript's abstract****"
manuscript_local_text_before_citation = "****The sentences before the target
    citation sentence in the manuscript (local context)****"
citation_intent = "****Specified citation intent****"
keywords = "****Specified keywords****"

messages = [
    {
        "role": "system",
        "content": "You are a scientific writing assistant. Your task is to generate
     citation sentences for a given manuscript, following detailed instructions.
    These instructions involve taking into account the context, desired citation
    intent, specific keywords in your responses."
    },
    {
        "role": "user",
        "content": f"""
The authors need to cite the reference paper:

    Title: {cited_paper_title}
    Abstract: {cited_paper_abstract}

, in the manuscript with the global context:

    Title: {manuscript_title}
    Abstract: {manuscript_abstract}

, immediately following the provided local context:

    {manuscript_local_text_before_citation}

Your task:
Please generate one citation sentence that cites the reference paper, seamlessly
    follows the local context, reflects the specified citation intent, and
    incorporates the specified keywords.

Requirements:
1. The generated citation sentence should reflect the citation intent: {
    citation_intent}. Citation intents include:
    1) background: The citation provides background information or additional
    context about a relevant problem, concept, approach, or topic.
    2) method: The citation refers to the use of a specific method, tool, approach,
    or dataset from the reference paper.
    3) result: The citation compares or contrasts the results or findings of the
    manuscript with those in the reference paper.
2. The generated citation sentence should contain the specified keywords: {keywords
    }. All the provided keywords should be used. If no keywords are specified,
    please infer one or two keywords by yourself and generate the citation sentence
    based on them.
3. Insert the citation marker "#REFR" at the position in the sentence where the
    reference paper should be cited.
4. Put the citation marker "#REFR" correctly in the generated citation sentence. The
     citation marker should replace the entire in-text citation (e.g., authors and
    year of publication), should not be enclosed in any brackets, and should be
    placed within the sentence before the ending punctuation.

Please return only the generated citation sentence.
            """
    },
]
```

Listing 2: The presented template was employed for querying the GPT-3.5-turbo (version 0301) API instructing it to generate a citation sentence given the context which comprises the title and abstract of the cited paper the title and abstract of the manuscript and the local context (sentences before the target citation in the manuscript). The generation was also controlled by the specified citation attributes including citation intent and keywords.

Cited Paper:

Title: Reinforcement Learning: A Survey

Abstract: This paper surveys the field of reinforcement learning from a computer-science perspective. It is written to be accessible to researchers familiar with machine learning. Both the historical basis of the field and a broad selection of current work are summarized. Reinforcement learning is the problem faced by an agent that learns behavior through trial-and-error interactions with a dynamic environment. The work described here has a resemblance to work in psychology, but differs considerably in the details and in the use of the word ``reinforcement.'' The paper discusses central issues of reinforcement learning, including trading off exploration and exploitation, establishing the foundations of the field via Markov decision theory, learning from delayed reinforcement, constructing empirical models to accelerate learning, making use of generalization and hierarchy, and coping with hidden state. It concludes with a survey of some implemented systems and an assessment of the practical utility of current methods for reinforcement learning.

Manuscript:

Title: Evolving Reinforcement Learning Environment To Minimize Learner'S Achievable Reward: An Application On Hardening Active Directory Systems

Abstract: We study a Stackelberg game between one attacker and one defender in a configurable environment. The defender picks a specific environment configuration. The attacker observes the configuration and attacks via Reinforcement Learning (RL trained against the observed environment). The defender's goal is to find the environment with minimum achievable reward for the attacker. We apply Evolutionary Diversity Optimization (EDO) to generate diverse population of environments for training. Environments with clearly high rewards are killed off and replaced by new offsprings to avoid wasting training time. Diversity not only improves training quality but also fits well with our RL scenario: RL agents tend to improve gradually, so a slightly worse environment earlier on may become better later. We demonstrate the effectiveness of our approach by focusing on a specific application, Active Directory (AD). AD is the default security management system for Windows domain networks. AD environment describes an attack graph, where nodes represent computers/accounts/etc., and edges represent accesses. The attacker aims to find the best attack path to reach the highest-privilege node. The defender can change the graph by removing a limited number of edges (revoke accesses). Our approach generates better defensive plans than the existing approach and scales better.

Text Before Citation: NNDP attacker's policy #OTHEREFR trains the model against one defensive plan at a time, due to which it forgets the previous plan. This way, it keeps learning and forgetting the plans. However, we train our RL based attacker's policy against multiple defensive plans at a time, due to which it learns shared experience and performs better. For RL agent, diverse environment configurations are only different in the "opening games", whereas the "end games" or "mid games" are likely to be similar across different environments. The similarity in later stages can be utilized in parallel training, where the agent is trained against multiple environments simultaneously and gains shared experience, leading to faster convergence and improved performance. Besides, NNDP approach is value iteration-based RL algorithm, whereas our approach is policy iteration-based RL algorithm.

Ground-truth Citation Sentence:

In general, the policy iteration-based algorithms converge faster than value iteration-based algorithms #REFR , which is another reason for the superior performance of our approach. Our experimental results further support our arguments.

Generated Citation Sentences:

Uncontrolled Mode:

Citation Intent: N/A

Keywords: N/A

Generation: Value iteration-based RL algorithm learns the value function of each state, whereas policy iteration-based RL algorithm learns the policy of each state #REFR.

ROUGE F1: R1: 34.48 R2: 17.86 RL: 24.14

Intent-controlled Mode:

Citation Intent: result

Keywords: N/A

Generation: It has been shown that policy iteration-based RL algorithm converges faster than value iteration-based RL algorithm #REFR.

ROUGE F1: R1: 34.62 R2: 24.00 RL: 34.62

Intent- and keywords-controlled Mode:

Citation Intent: result

Keywords: policy iteration-based algorithms

Generation: Policy iteration-based algorithms are known to converge faster than value iteration-based algorithms #REFR.

ROUGE F1: R1: 50.00 R2: 43.48 RL: 50.00

Figure 2: Example citation sentences generated by Galactica-6.7B-PPO under the uncontrolled mode, the intent-controlled mode, and the intent- and keywords-controlled mode.

Model	Uncontrolled generation			Intent-controlled generation			Intent- and keywords-controlled generation					
	R1	R2	RL	R1	R2	RL	R1	R2	RL	IAS	KR	FS
Galactica-125M-beam1	27.93	6.00	20.39	28.67	6.41	21.09	35.85	13.44	26.88	0.7128	0.7667	0.7539
Galactica-125M-beam2	27.26	5.80	19.61	28.00	6.26	20.24	36.00	13.81	26.68	0.6946	0.7865	0.7526
Galactica-125M-beam4	27.15	6.01	19.50	27.69	6.29	19.98	35.44	13.59	26.13	0.6872	0.7656	0.7466
Galactica-125M-beam8	26.47	6.03	18.87	26.91	6.38	19.44	34.99	13.67	25.91	0.6724	0.7400	0.7425

Table 7: Performance of the supervised fine-tuned Galactica-125M model on the validation set, utilizing various beam sizes for inference.

Cited Paper:

Title: Optimal Scheduling For Discounted Age Penalty Minimization In Multi-Loop Networked Control

Abstract: Age-of-information (AoI) is a metric quantifying information freshness at the receiver. Since AoI combines packet generation frequency, packet loss, and delay into a single metric, it has received a lot of research attention as an interface between communication network and application. In this work, we apply AoI to the problem of wireless scheduling for multi-loop networked control systems (NCS), i.e., feedback control loops closed over a shared wireless network. We model the scheduling problem as a Markov decision process (MDP) with AoI as its observable states and derive a relation of control system error and AoI. We further derive a stationary scheduling policy to minimize control error over an infinite horizon. We show that our scheduler outperforms the state-of-the-art scheduling policies for NCS. To the best of our knowledge, this is the first work proposing an AoI-based wireless scheduling policy that minimizes the control error over an infinite horizon for multi-loop NCS.

Manuscript:

Title: Large Population Games On Constrained Unreliable Networks

Abstract: This paper studies an N-agent cost-coupled game where the agents are connected via an unreliable capacity constrained network. Each agent receives state information over that network which loses packets with probability p. A Base station (BS) actively schedules agent communications over the network by minimizing a weighted Age of Information (WAoI) based cost function under a capacity limit C < N on the number of transmission attempts at each instant. Under a standard information structure, we show that the problem can be decoupled into a scheduling problem for the BS and a game problem for the N agents. Since the scheduling problem is an NP hard combinatorics problem, we propose an approximately optimal solution which approaches the optimal solution as N → ∞. In the process, we also provide some insights on the case without channel erasure. Next, to solve the large population game problem, we use the mean-field game framework to compute an approximate decentralized Nash equilibrium. Finally, we validate the theoretical results using a numerical example.

Text Before Citation: Further, we improve upon the convergence guarantees in #OTHEREFR for the case where the network is erasure free by proposing a novel scheduling policy. Finally, we employ this policy to construct an approximate Nash solution for the finite-agent consensus problem. In literature, the early works #OTHEREFR , #OTHEREFR have dealt with an optimal control problem with unreliable communication, albeit, for a single agent system and an unconstrained network under the TCP and the UDP communication protocols. The work #OTHEREFR extends the setting to multi-agent games; however, the considered network is unconstrained. In order to measure timeliness in communication networks, age of information (AoI) has been introduced as a potential metric.

Ground-truth Citation Sentence:

In the context of networked feedback systems, the AoI-based policies have been proposed for solving resource allocation and enduser uncertainty reduction problems as in #REFR .

Generated Citation Sentences:

Uncontrolled Mode:

Citation Intent: N/A

Keywords: N/A

Generation: The work #REFR studies a wireless scheduling problem for multi-loop networked control systems (NCS), i.e., feedback control loops closed over a shared wireless network.

ROUGE F1: R1: 23.08 R2: 0.00 RL: 11.54

Intent-controlled Mode:

Citation Intent: method

Keywords: N/A

Generation: The authors in #REFR have used AoI to solve a wireless scheduling problem for multi-loop networked control systems.

ROUGE F1: R1: 35.56 R2: 4.65 RL: 13.33

Intent- and keywords-controlled Mode:

Citation Intent: method

Keywords: networked feedback systems

Generation: The work #REFR proposes an optimal scheduling policy for a multi-loop networked feedback systems based on AoI.

ROUGE F1: R1: 36.36 R2: 9.52 RL: 22.73

Figure 3: Example citation sentences generated by Galactica-6.7B-PPO.

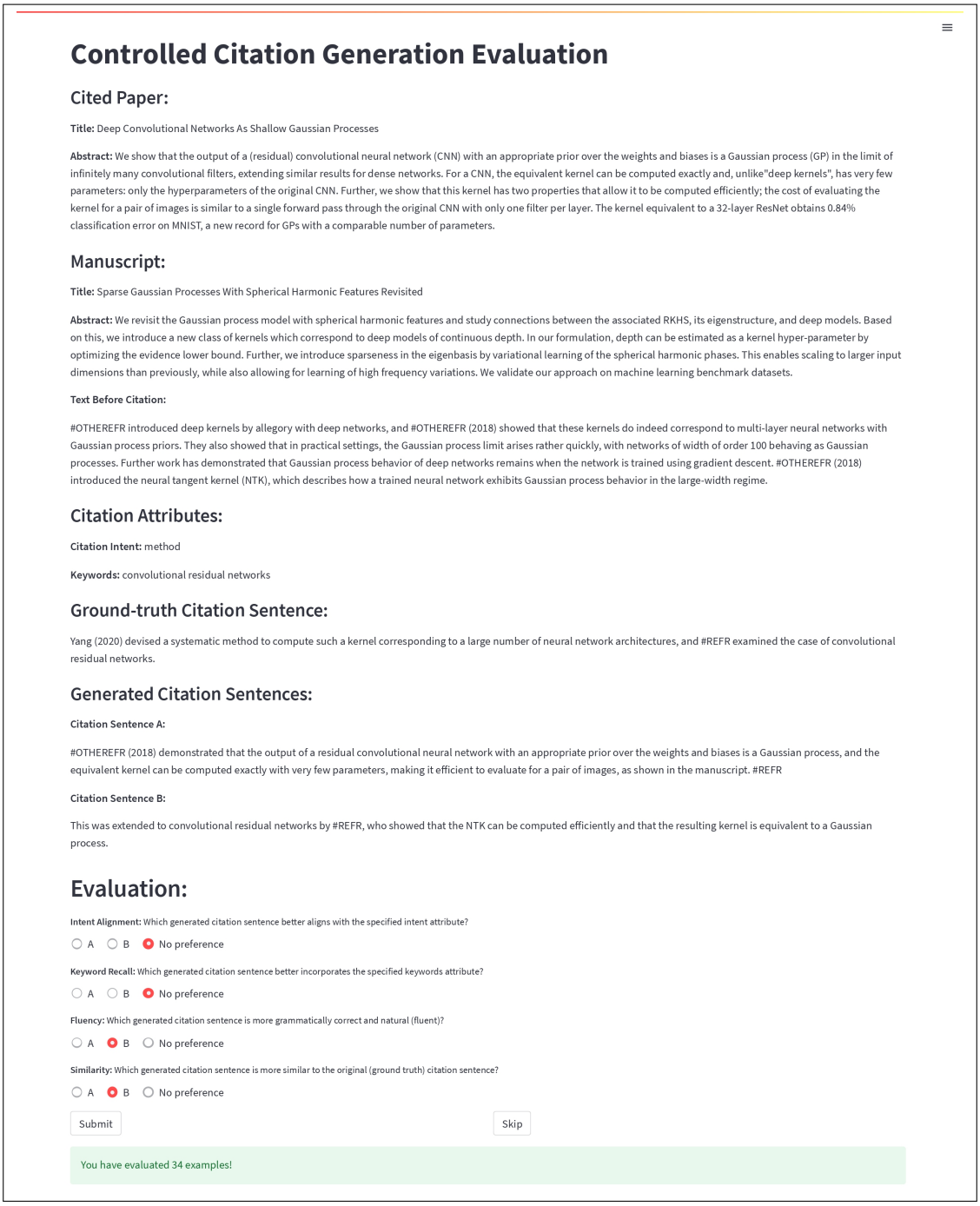

Figure 4: The user interface for the side-by-side comparison of citation sentences generated by Galactica-6.7B-PPO and GPT-3.5-turbo.