Large Language Models Reflect Human Citation Patterns with a Heightened Citation Bias

Andres Algaba[1], Carmen Mazijn[1], Vincent Holst[1], Floriano Tori[1], Sylvia Wenmackers[1,2], Vincent Ginis[1,3]

[1]Data Analytics Lab, Vrije Universiteit Brussel, Belgium
[2]Centre for Logic and Philosophy of Science (CLPS), KU Leuven, Belgium
[3]School of Engineering and Applied Sciences, Harvard University, USA
Correspondence: andres.algaba@vub.be

Abstract

Citation practices are crucial in shaping the structure of scientific knowledge, yet they are often influenced by contemporary norms and biases. The emergence of Large Language Models (LLMs) introduces a new dynamic to these practices. Interestingly, the characteristics and potential biases of references recommended by LLMs that entirely rely on their parametric knowledge, and not on search or retrieval-augmented generation, remain unexplored. Here, we analyze these characteristics in an experiment using a dataset from AAAI, NeurIPS, ICML, and ICLR, published after GPT-4's knowledge cut-off date. In our experiment, LLMs are tasked with suggesting scholarly references for the anonymized in-text citations within these papers. Our findings reveal a remarkable similarity between human and LLM citation patterns, but with a more pronounced high citation bias, which persists even after controlling for publication year, title length, number of authors, and venue. The results hold for both GPT-4, and the more capable models GPT-4o and Claude 3.5 where the papers are part of the training data. Additionally, we observe a large consistency between the characteristics of LLM's existing and non-existent generated references, indicating the model's internalization of citation patterns. By analyzing citation graphs, we show that the references recommended are embedded in the relevant citation context, suggesting an even deeper conceptual internalization of the citation networks. While LLMs can aid in citation generation, they may also amplify existing biases, such as the Matthew effect, and introduce new ones, potentially skewing scientific knowledge dissemination.

1 Introduction

Large Language Models (LLMs) have revolutionized natural language understanding and generation, driving scientific research forward by assisting in all steps of the scientific process, ranging from identifying research gaps to accelerating complex data analysis (Boiko et al., 2023; Merchant et al., 2023; Romera-Paredes et al., 2024; Zheng et al., 2023).[1] One particularly interesting application is the generation of suggestions for appropriate scholarly references (Qureshi et al., 2023; Walters and Wilder, 2023). Yet, without the aid of web browsing or retrieval-augmented generation, these models rely entirely on their parametric knowledge encapsulated during their (pre-)training (Brown et al., 2020; Bubeck et al., 2023; Kaddour et al., 2023; Wei et al., 2022a). Our research focuses on this intrinsic citation behavior of GPT-4, exploring how the model recommends references based on its training data, and highlighting the potential biases that arise from this internalized knowledge (Acerbi and Stubbersfield, 2023; Manerba et al., 2023).

Biases in citation practices have long been a subject of scrutiny in the scientific community (Fortunato et al., 2018; Smith, 2012). Besides normative theory (Kaplan, 1965; Garfield, 1965), citations are well-known to be used for different motives, such as for instance rhetorical persuasion (Nigel Gilbert, 1977). On the other hand, the choice of citing work is also influenced by biases in the characteristics of the referenced work itself. In a seminal paper, Price (Price, 1976) demonstrated the "success breeds success" dynamic (cumulative advantage or preferential attachment). This dynamic underpins the "Matthew effect", in which highly cited papers accumulate even more citations (Wang, 2014). Beyond preferential attachment, other common biases include a preference for recent publications (Bornmann and Daniel, 2008), shorter titles (Letchford et al., 2015), high-profile publication venues (Lawrence, 2003). By examining how these biases manifest in LLM-generated references, we aim to uncover underlying patterns that could

[1]Data and code are available at `https://zenodo.org/records/11299894` and `https://github.com/AndresAlgaba/LLM_citation_patterns`

Findings of the Association for Computational Linguistics: NAACL 2025, pages 6844–6879

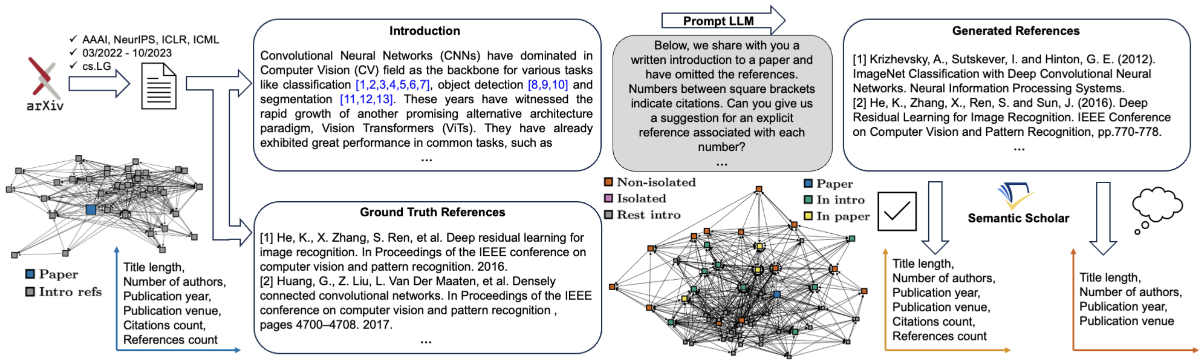

Figure 1: **Overview of our experiment evaluating the characteristics and biases of LLM generated references, when tasked to suggest references for anonymized in-text citations.** We collect 166 papers from the cs.LG category on arXiv which are published in the main tracks of AAAI, NeurIPS, ICML, and ICLR, and only appeared available online after GPT-4's knowledge cut-off date. We split the main content, which includes the author information, conference information, abstract, and introduction, from the ground truth references. GPT-4, GPT-4o and Claude 3.5 are prompted to generate suggestions of scholarly references for the anonymized in-text citations in the main content. We verify the existence of the generated references via Semantic Scholar and compare the characteristics, such as title length, publication year, venue, and number of authors, of the existing and non-existent generated references with the ground truth. For the existing generated references, we also compare additional characteristics, such as the number of citations and references, and analyze the properties of their citation networks.

amplify existing biases or introduce new ones, potentially reinforcing feedback loops.

In our experiment, we let GPT-4, GPT-4o and Claude 3.5 suggest scholarly references for anonymized in-text citations within a paper and compare the characteristics and citation networks of the LLM generated references against the ground truth. We provide a comprehensive analysis of 166 papers which are published in the main tracks of AAAI, NeurIPS, ICML, and ICLR, encompassing 3,066 references in total. All the papers are only first available online on arXiv after GPT-4-0613's knowledge cut-off date and belong to the cs.LG category. While this experimental setup may not fully reflect real-world usage of LLMs for citation generation, which often involves more interactivity and reliance on external data sources, it provides a controlled laboratory setting to assess the parametric knowledge and inherent biases of LLMs. Furthermore, our focused sample of papers ensures a homogeneous dataset, which allows us to minimize confounding factors that could arise from cross-disciplinary differences in citation practices.

Our setting differs from previous work which either let LLMs generate short papers or literature reviews, or is prompted for the most important papers on a certain topic (Walters and Wilder, 2023). We argue that these methods are more susceptible to the LLM's memorization capabilities (Chen et al., 2024; Kadavath et al., 2022). Moreover, the evaluation of the suggested references mostly focuses on their existence, bibliometric accuracy, or qualitative judgement by domain experts (Qureshi et al., 2023). Finally, another strand of the literature focuses on improving LLMs via search and retrieval-augmented generation (Lewis et al., 2020) or to reduce their hallucination rate via self-consistency (Agrawal et al., 2024) to enhance their capabilities in systematic literature reviews (Susnjak et al., 2024).

In our experiment, we find that GPT-4 exhibits strong preferences for highly cited papers, which persists even after controlling for multiple confounding factors such as publication year, title length, venue, and number of authors. Additionally, we observe a large consistency between GPT-4's existing and non-existent generated references, indicating the model's internalization of citation patterns. The same results hold for the more capable models GPT-4o and Claude 3.5 where the papers are part of the training data. By analyzing citation graphs, we show that the references recommended by GPT-4 are embedded in the relevant citation context, suggesting an even deeper conceptual internalization of the citation networks. While LLMs can aid in citation generation, our results underscore the need for identifying the model's biases and for developing balanced methods to interact with LLMs in general (Navigli et al., 2023).

2 Generating Citations with LLMs

Our data consists of 166 papers published at AAAI (25), NeurIPS (72), ICML (38), and ICLR (31) for a total of $3,066$ references. Our data collection process is depicted in Figure 1 (see Appendix A for more details) and begins by retrieving all the relevant papers from arXiv, focusing on those within the machine learning category (cs.LG) and posted between March 2022 and October 2023 (after GPT-4-0613's knowledge cut-off date). The papers are verified on Semantic Scholar where we store additional metadata, such as all the reference titles with corresponding Semantic Scholar IDs to construct the citation networks (see Appendix Table C3 for a full list of all the included papers).

We split the main content, which includes the author information, conference information, abstract, and introduction, from the ground truth references. Next, we prompt GPT-4, GPT-4o and Claude 3.5 to generate scholarly reference suggestions (see Appendix A for the prompts). We then post-process the responses to extract the title, venue, publication year, author names, and number of authors for each generated reference (see Appendix A for more details). To assess the robustness of this approach, we repeat this "vanilla" approach three (GPT-4o and Claude 3.5) to five (GPT-4) times.

A well-known issue in text generation by LLMs are hallucinations or confabulations, which refer to generated content that is nonsensical or untruthful in relation to certain sources, i.e., factual mistakes about historical events (Zhang et al., 2023). This is particularly problematic for the generation of scholarly references, as LLMs can fabricate references that do not exist or introduce subtle errors, making it impossible to retrieve the actual references (Walters and Wilder, 2023). There are two main approaches to verify the existence of LLM-generated references: one involves asking additional questions to the LLM to verify its self-consistency (Agrawal et al., 2024), and the second approach utilizes external databases to verify a reference's existence (Fabiano et al., 2024). In our experiment, we opt for the latter and determine via title and author names matching with Semantic Scholar entries whether the generated references exist (see Appendix A for more details). Finally, we also build on our "vanilla" approach, by introducing an "iterative" approach where we continue to prompt GPT-4 after having indicated which generated references do not exist and ask to replace those with existing ones (see Appendix A for more details). The previously existing generated and the newly generated references are then merged.

In Table 1, we report the GPT-4 summary statistics for each of the five vanilla (iterative) runs. On average, 65% (86%) of the generated references match with an entry in Semantic Scholar, while 13% (14%) and 17% (20%) of them appear in the introduction or paper itself, respectively. We further show that about 7% (7%) of the generated and ground truth references match pairwise and 13% (14%) if we only consider the uniquely identifiable references (i.e., omitting references included in *[4–8]* as there is no one-to-one correspondence) which indicates that GPT-4 has not memorized the references. In Appendix Table C1, we show that the average overlap between generated sets is 17%.

3 Reflecting Human Citation Patterns

Figure 2 displays the characteristics of the ground truth and GPT-4 generated references, and separately the characteristics of the generated references which match with a Semantic Scholar entry, and those which do not exist according to this database. Overall, we observe a remarkable similarity between human and LLM citation patterns and a large consistency between GPT-4's existing and non-existent generated references, indicating the model's internalization of citation patterns. All median differences between ground truth and (existing and non-existent) generated references shown are significant at the 1% level according to the pairwise two-sided Wilcoxon signed-rank test. In Appendix Figure B1, we also show that the newly generated papers from the "iterative" approach show nearly identical distributions.

The distributions of the title lengths show that existing generated reference titles tend to be the shortest, while non-existent generated reference titles are more similar in length to the ground truth, which indicates a learned pattern. Overall, the first effect dominates, so the average is skewed to shorter titles for generated references (Figure 2b). The temporal analysis reveals a similar pattern where non-existent generated references follow a distribution that is more similar to the ground truth than the existent ones (Figure 2c).

The distribution of the number of authors highlights a notable difference, with ground truth references typically involving three authors versus two for generated references, though the frequent use

Vanilla (Iterative)	Run 1	Run 2	Run 3	Run 4	Run 5
Existing generations in Semantic Scholar database (%)	64.3 (87.0)	63.3 (85.5)	62.8 (88.0)	64.2 (86.8)	67.6 (86.3)
Existing generations cited in the original paper (%)	17.5 (20.0)	17.1 (20.1)	15.7 (18.4)	16.8 (19.2)	18.0 (20.8)
Existing generations cited in the original intro (%)	13.4 (14.5)	13.2 (15.0)	12.2 (13.5)	12.9 (14.3)	13.9 (15.3)
Existing generations with a pairwise match (%) (for all references)	7.0 (7.1)	7.2 (7.3)	6.3 (6.6)	6.9 (7.0)	6.7 (7.1)
Existing generations with a pairwise match (%) (for uniquely identifiable references)	12.5 (12.5)	13.7 (14.1)	12.5 (12.9)	13.7 (13.7)	13.3 (14.0)

Table 1: **Summary statistics of GPT-4 generated references (in % with respect to total number of references).**

of "et al." in the generated references complicates exact author counts (Figure 2d). To further examine the potential impact of the "et al." problem, we only consider the existing generated references and their ground truth counterpart in Appendix Figure B2. There, we compare the characteristics of the references between two data sources, namely the original source (the paper or GPT generation) and the available information on Semantic Scholar. The similarity between the distributions of all characteristics shows that the data source has no impact and "et al." does not cause this observation.

The publication venue distributions show that for most venues the ground truth has the highest relative representation, followed closely by existing generated references, with non-existent generated references displaying the largest proportion of "Others" (Figure 2e). In Appendix Figure B3, we observe that the distributions of publication venues for both ground truth and generated references are very similar across the various conferences, i.e., AAAI, NeurIPS, ICML, and ICLR. The pairwise transition matrix from ground truth to generated publication venues at the reference level indicates a large overall agreement, but with a strong preference in GPT-4 generated references for arXiv, NeurIPS, and "Others" in the case of disagreement. The preference for NeurIPS may be due to the relatively large number of NeurIPS papers in our sample and the large share of arXiv and "Others" points to favoring a wider array of venues which may potentially dilute the perceived relevance of key conferences. Finally, the scatter plot affirms the strong pairwise correlation between the ground truth and generated references to the top conferences at the individual paper level.

Most prominently, we observe a significant citation bias in the existing generated references, which have a median citation count of $1,326$ higher than ground truth references (Figure 2f). The skewing of citation distributions caused by preferential attachment is very pronounced for the generated references. In Appendix Figure B6, we compare the characteristics for the corresponding ground truth references of existing and non-existent references, and for the existing references which also appear in the paper itself. We observe that the ground truth papers which correspond to existing references that appear in the paper itself have by far the most citations, followed by the existing references, and the ground truth papers corresponding to non-existent references have the lowest numbers of citations. These findings further indicate the tendency for GPT-4 to more easily generate references to highly cited papers. Finally, the distribution of references indicates that ground truth references cite slightly more papers than existing generated references (Figure 2g).

In Appendix Figures B4 and B5 and Table C2, we find similar results for three GPT-4o and Claude 3.5 runs, but with a higher existence rate which may be due to the models' capabilities or the papers being part of the training data.

4 Heightened Citation Bias

Figure 3 demonstrates that the citation bias observed in GPT-4 generated references is not merely a consequence of the recency of ground truth references. Specifically, the existing generated references show consistently higher citation counts com-

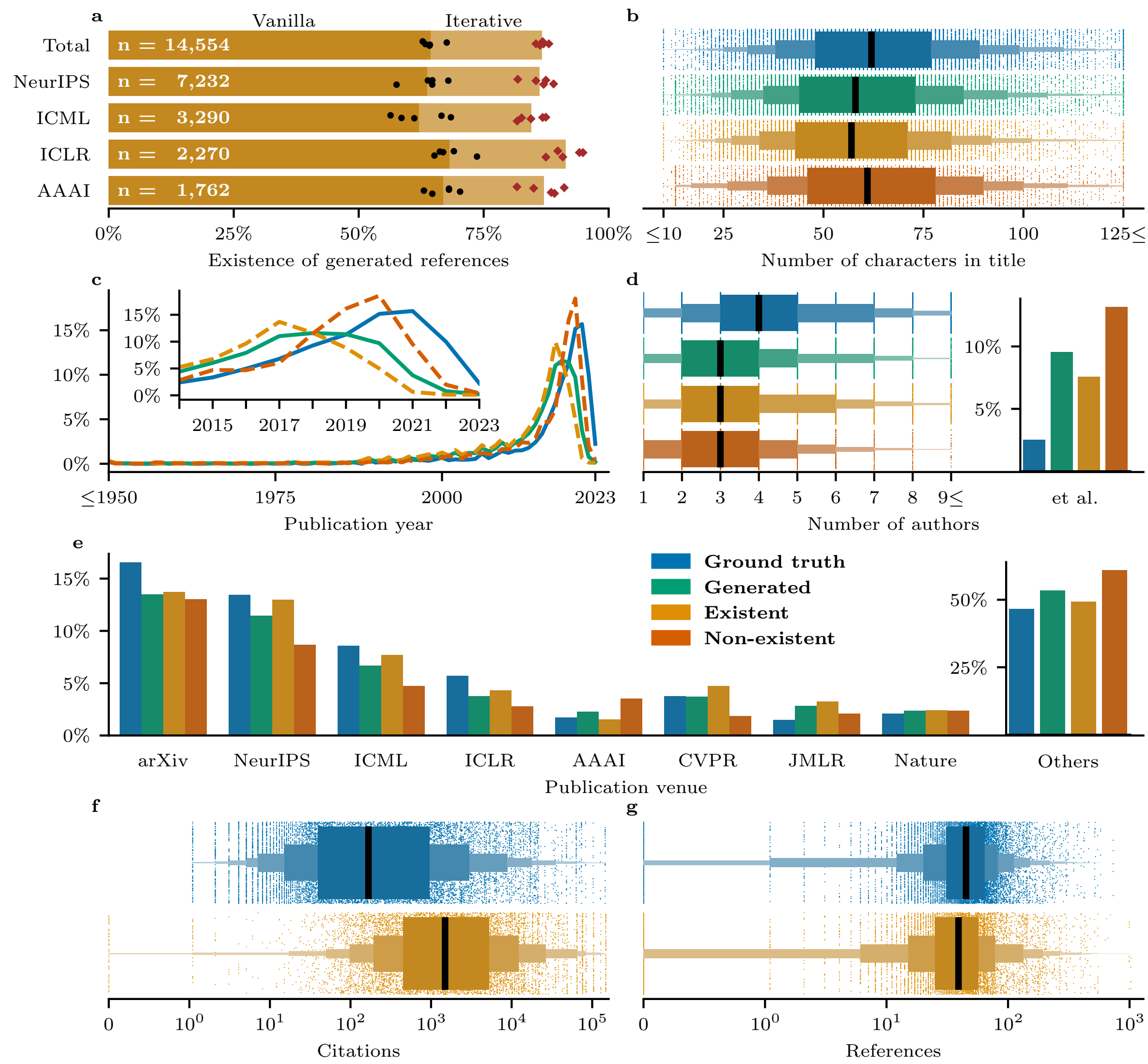

Figure 2: **Properties of the ground truth and GPT-4 generated introduction references for the vanilla strategy.** This figure displays the properties of the ground truth ($n = 14,554$, in blue) and GPT-4 generated references ($n = 14,554$, in green), further subdividing the generated references into existing ($n = 9,376$, in orange) and non-existent ($n = 5,178$, in red), from the original data sources of five runs for the vanilla strategy with GPT-4. **a,** The average percentage of existing generated references in total (64.4%) and for each publication venue under the vanilla and iterative strategy, with dots representing the percentage for each of the five runs with GPT-4. **b,** The distribution of the number of characters in the title shows some differences between the ground truth (median 62) and generated (median 58) references with the non-existent being slightly longer and with a larger variance compared to the existing generations. **c,** The distribution over time reveals that ground truth references are relatively more recent than generated references, with most references post-2010. The temporal distribution of the non-existent generated references aligns more with the ground truth than the existing generated references. **d,** The distribution of the number of authors demonstrates a disparity between the ground truth and generated references, having median values of three and two, respectively. However, GPT-4 more often generates “et al.” which does not allow for an exact computation, especially for the non-existent references. **e,** The distribution of publication venues shows that for most venues the ground truth has the highest relative representation, followed closely by existing references. The non-existent references deviate more from the ground truth as the proportion of “Others” is substantially larger. **f,** The distributions of citations for ground truth and existing generated references reveal a substantial citation bias in the generated references with a difference in median citations of $1,326$. **g,** Finally, the distribution of references shows that ground truth references cite slightly more papers than the existing generated references with a median difference in median references of 6.

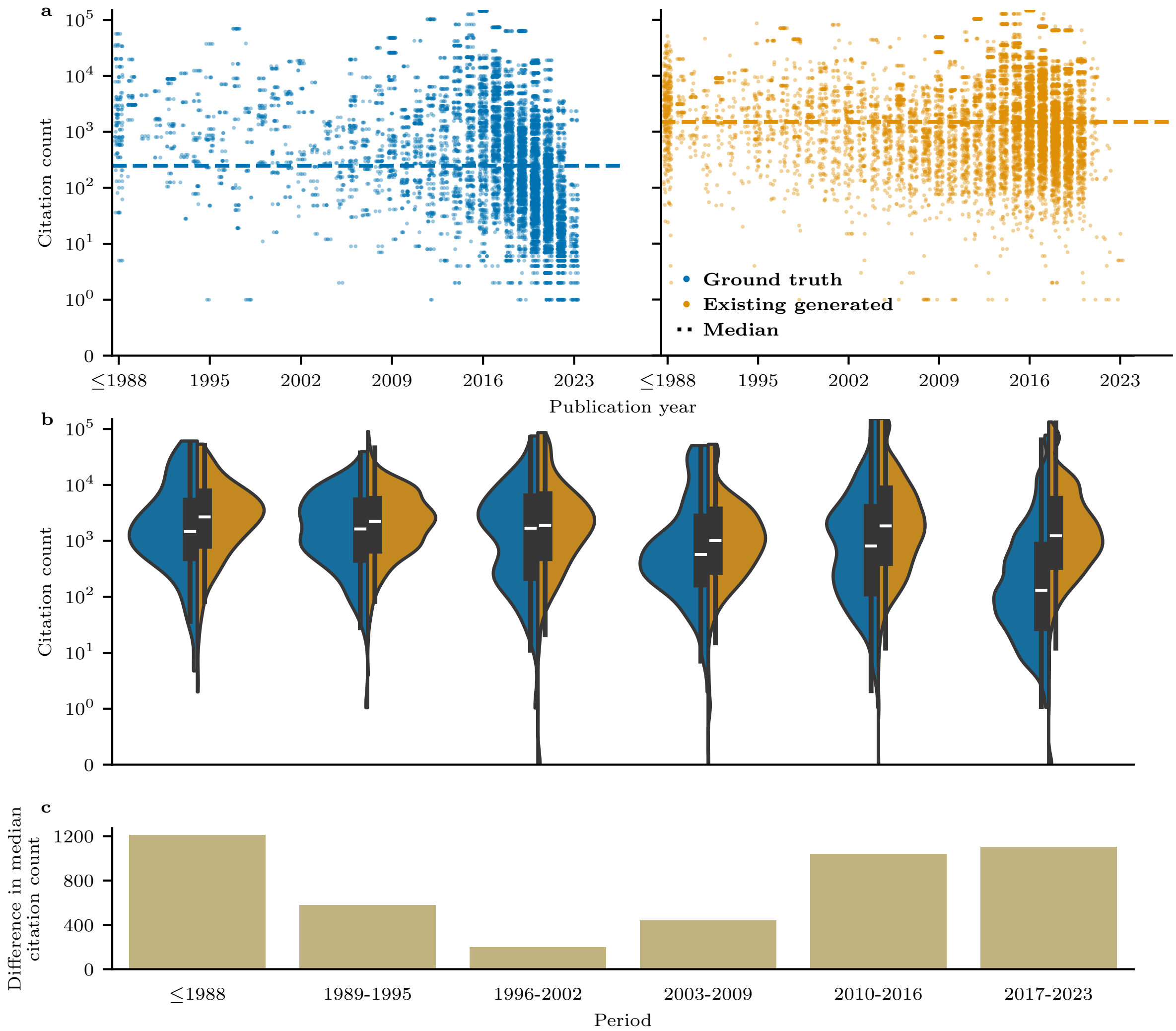

Figure 3: **The citation bias in existing GPT-4 generated references is not due to the recency of ground truth references.** This figure shows that the existing GPT-4 generated references ($n = 9,376$, in orange) consistently exhibit a higher citation count compared to their corresponding ground truth ($n = 9,376$, in blue) across subperiods. **a,** The citation counts across time for the ground truth and existing generated references reveal that the most recent references have a relatively low number of citations. The difference in median citations between the existing generated references and their corresponding ground truth references is 1,257. Since the ground truth references are relatively more recent compared to the existing generated references, we examine whether the observed citation bias is related to the recency of ground truth references. **b,** The distributions of citations by subperiod reveal that the existing generated references consistently exhibit a higher citation count than their corresponding ground truth counterparts. **c,** The difference in median citations is most pronounced in the early and late subperiods, i.e., $\leq$ 1988, 2010-2016, and 2017-2023.

pared to their ground truth counterparts across various subperiods. Figure 3a illustrates that ground truth references, particularly the most recent ones, tend to have lower citation counts. Despite the ground truth references being more recent on average, the citation counts of existing generated references remain significantly higher. Figure 3b further breaks down the citation distributions by subperiods, reaffirming that generated references consistently have higher citation counts than their corresponding ground truth references. Figure 3c highlights that this citation discrepancy is most pronounced in both the earliest (≤1988) and the most recent (2010-2016 and 2017-2023) subperiods, indicating that the citation bias persists across different time frames.

In Appendix Figure B7, we find that the heightened citation bias in generated references remains also after controlling for other possible confounding factors, such as title length, number of authors, and publication venue. In Appendix Figure B8 and Appendix Figure B9, we confirm that our findings are robust for the influential citation count which can be retrieved from Semantic Scholar (Valenzuela-Escarcega et al., 2015). This consistency across multiple factors underscores the inherent bias of LLMs towards generating references to highly cited papers, irrespective of other characteristics of the references.

5 LLMs and Human Citation Networks

Figure 4 displays the properties of the ground truth and GPT-4 generated citation networks. One of the primary purposes of this analysis is to provide an initial intuition about how plausible the generated references are and how easily they can be identified. Interestingly, the local citation networks of the generated references are strikingly similar to those of human-generated citations, indicating that they are not merely random selections of highly cited papers. While some systematic differences remain, the observed alignment suggests that GPT-4 has internalized key aspects of human citation behavior on the level of citation networks. In Figure 4a, we identify the focal paper (in blue), generated references that appear in the introduction (in green) or later in the paper (in yellow), generated references that are linked to ground truth or other generated references (in orange), generated references that are completely isolated (in purple), and ground truth references that are not cited by GPT-4 (in gray). The majority of generated references ($> 50\%$) is non-isolated, i.e., linked to the ground truth or generated references but not present in the focal paper itself, followed by a substantial amount of generated references appearing in the introduction and only a small fraction that do not appear in the introduction but still within the focal paper (Figure 4b). The remainder of generated references is completely isolated from the citation network. If GPT-4 did not pick up on human citation patterns, the generated citation network would resemble a random network containing only isolated citations. The heightened citation bias is also most pronounced for references that appear within the introduction or paper, with isolated generated references having the lowest number of citations (Figure 4c). This finding further indicates the tendency for GPT-4 to more easily identify and generate references to highly cited papers. The number of references is similar across all categories, except for the isolated generated references which have substantially less references (Figure 4f).

The normalized average clustering coefficients (Watts and Strogatz, 1998) of the ground truth (green and grey nodes) and the existing generated references (green, yellow, orange, and purple nodes) indicate that GPT-4's internalization of citation patterns extends to citation network properties (Figure 4d). This internalization is also reflected by the tight connection between the non-isolated generated and ground truth references. The connection appears on an individual level as measured by the Boolean edge density, as well as on the aggregate level as measured by the edge expansion. For instance, in the central graph shown in Figure 4a, a Boolean edge density of $\frac{2}{3}$ suggests one non-isolated generated reference links only within its group, while an edge expansion of $2\frac{1}{3}$ indicates strong connections between the other two non-isolated generated references and the actual ground truth references. So, we can exclude the possibility of GPT-4 generating suggestions of scholarly references that are connected to each other but move further and further away from actual content of the introduction. Regardless, three of the four categories (green, yellow and orange) are well embedded in the given citation context. It reflects how tight the connection between the non-isolated generations to the ground truth references is and the deeper conceptual internalization of the citation networks.

6 Discussion

We present an experiment to explore the intrinsic citation behavior of LLMs and their potential biases when generating scholarly references. Whereas, previous work focuses on LLMs generating short papers or literature reviews (Qureshi et al., 2023; Walters and Wilder, 2023), we let GPT-4, GPT-4o and Claude 3.5 generate suggestions of scholarly references for anonymized in-text citations. Importantly, we do not enhance the LLM through search and retrieval-augmented generation, but evaluate the model's internalization of citation patterns in its parametric knowledge obtained during training. While, our experimental setup may not fully reflect real-world usage of LLMs for citation generation, which often involves more interactivity and reliance on external data sources, it provides a controlled laboratory setting to assess the parametric knowledge and inherent biases of LLMs.

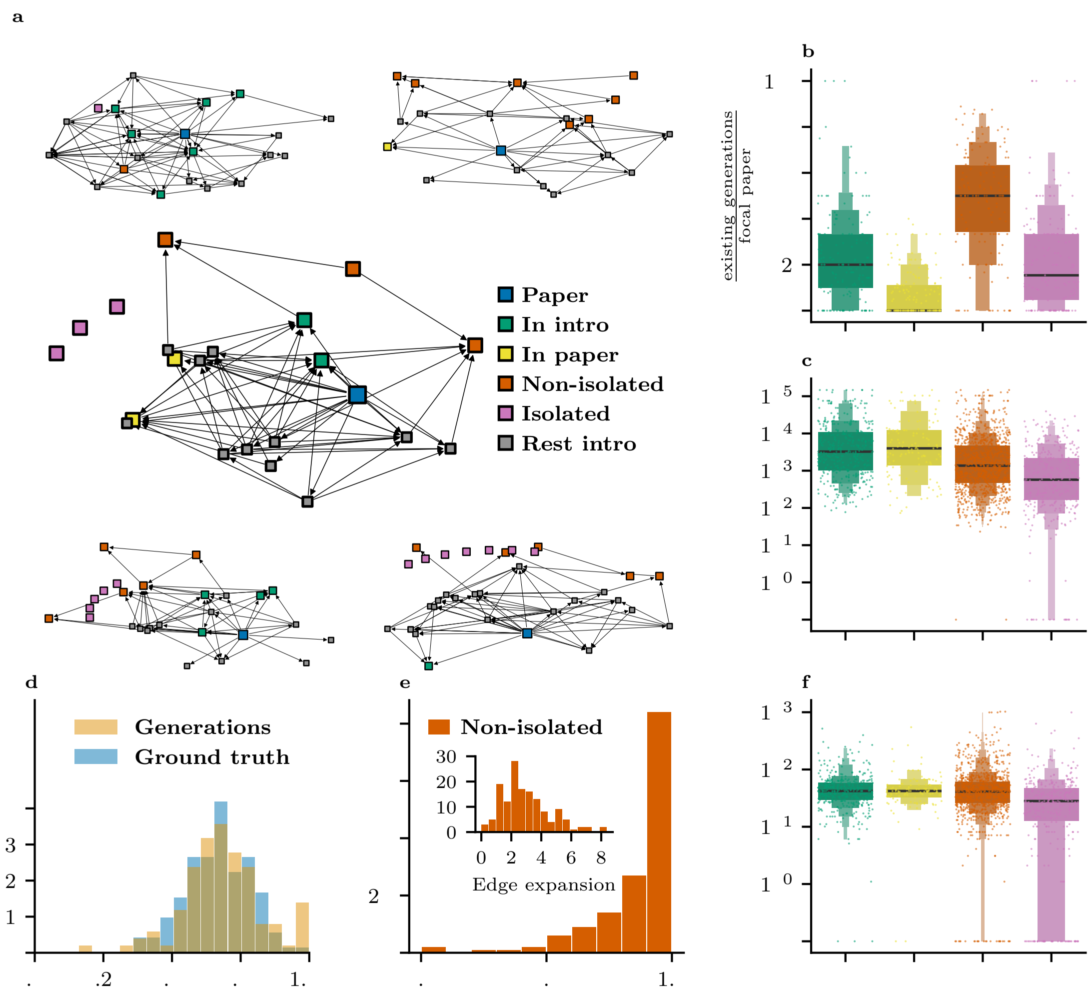

Figure 4: **The GPT-4 generated references display similar citation network properties as the ground truth references but with a heightened citation bias.** This figure displays how the existing GPT-4 generated references ($n = 2945$, first run of vanilla strategy) are embedded in the citation network of the focal papers ($m = 166$ in total). **a,** We depict the connections between the focal paper, the ground truth references, and the existing generated references by showing the underlying citation graphs. An arrow from A to B indicates that A cites B. We identify the focal paper (in blue), generated references that appear in the introduction (in green) or in the paper (in yellow), generated references that are linked to ground truth or other generated references (in orange), generated references that are completely isolated (in purple), and ground truth references that are not cited by GPT-4 (in gray). **b,** The majority of generated references does not appear in the introduction or paper itself, but is somehow connected to the ground truth references as only a small fraction of generated references is completely isolated. **c,** The heightened citation bias is most emphasized for generated references that appear in the introduction or the paper, with isolated generated references having the lowest number of citations. **d,** The normalized average clustering coefficients (Watts and Strogatz, 1998) of the ground truth (green and grey nodes) and the existing generated references (green, yellow, orange, and purple nodes) indicate that GPT-4's internalization of citation patterns extends to citation network properties. The clustering coefficient for a node A is given by $\frac{\#\text{triangles through A}}{\#\text{possible triangles through A}}$. The average is computed across the coefficients of all nodes in the respective graph (excl. nodes with coefficient zero) and indicates the tendency of the respective references to appear in clusters. **e,** The non-isolated generated references are tightly connected to the ground truth references, both on an individual level (Boolean edge density) as well as an aggregate level (edge expansion). The Boolean edge density is the fraction of non-isolated generations (orange nodes) that are connected to at least one ground truth reference (green and grey nodes) per focal paper. The edge expansion between those two sets is defined as the number of edges between the two sets divided by the smallest set size. **f,** The number of references is similar across all categories, except for the isolated generated references which have substantially less references.

Our findings are significant as they represent a first step towards understanding the real-world impact of LLMs in scientific research (Boiko et al., 2023; Lu et al., 2024; Zheng et al., 2023). By highlighting the heightened citation bias in LLM generated references, we demonstrate the models' tendency to favor highly cited papers, which could exacerbate existing biases in scientific discourse. This evaluation moves beyond more traditional LLM benchmarks (Srivastava et al., 2022), emphasizing the practical implications of deploying these models in academic contexts (Jimenez et al., 2024). The results suggest that while LLMs have the potential to streamline various aspects of research, careful consideration is needed to mitigate the amplification of biases, such as the "Matthew effect." If left unaddressed, these biases could reinforce citation inequalities, potentially disadvantaging emerging research and underrepresented scientific communities.

One plausible hypothesis for the heightened citation bias observed in LLMs is the increased frequency of citations to heavily cited papers within the model's training data. This prevalence makes these references more likely to be generated accurately and recognized as existent. Additionally, such biases may stem from generic training effects, where models preferentially learn patterns that are more common in the data, leading to biases towards shorter titles, more heavily cited, and slightly less recent works (Kandpal et al., 2023). These tendencies may persist despite improvements in data quantity or model sophistication as indicated by our experiments with GPT-4o and Claude 3.5. Future research could explore targeted debiasing techniques that explicitly encourage LLMs to generate a more balanced set of citations. Potential strategies include fair prompting methods that guide the model towards suggesting a diverse range of references (Ma et al., 2023), penalty-based training objectives that discourage over-reliance on highly cited works, and dynamic citation augmentation using controlled retrieval mechanisms that prioritize underrepresented but relevant research.

We develop and open-source an extensible, automated pipeline to systematically analyze the references generated by LLMs. Although our methodology is robust, it is not without limitations. The use of simple prompts and the zero-shot setting (Kojima et al., 2022) aims to minimize bias in the generation process, but this simplicity might not capture the full spectrum of potential LLM capabilities. There are numerous alternative approaches and prompt designs that future research can explore to enhance the accuracy and relevance of generated references (Wang et al., 2022; Wei et al., 2022b; Yao et al., 2024). However, our iterative approach indicates that biases remain inherent in these generations. Additionally, future research can also extend the experiment beyond our specific sample of papers and observe the impact of cross-disciplinary differences in citation practices.

Beyond the immediate implications for LLM-generated citations, our findings raise broader concerns about the potential feedback loops that AI-driven citation generation might introduce. As LLMs become increasingly integrated into scientific workflows, their tendency to favor highly cited papers could reinforce and accelerate existing citation disparities, leading to even greater concentration of attention on a small subset of already dominant papers. This could diminish the visibility of novel or less frequently cited works, potentially distorting the trajectory of scientific progress. While our study provides an initial framework for measuring these effects, a longitudinal approach involving real-world usage patterns and human-in-the-loop citation generation is needed to fully assess the downstream consequences.

In conclusion, while LLMs can significantly aid in citation generation, they also risk amplifying existing biases and introducing new ones, potentially skewing the structuring and the dissemination of scientific knowledge. Our study underscores the necessity for developing balanced methods to interact with LLMs, incorporating diverse datasets, and implementing bias mitigation strategies. Fair prompting techniques (Ma et al., 2023), for instance, can be employed to reduce bias, but continuous vigilance and methodological innovation are required to ensure that the integration of LLMs into academic workflows promotes accurate knowledge dissemination.

Limitations

We aim to assess, in a controlled laboratory setting, the parametric knowledge and inherent biases when generating reference suggestions with LLMs. While this approach may not fully capture real-world usage patterns (Skarlinski et al., 2024; Susnjak et al., 2024; Tilwani et al., 2024; Wu et al., 2024), it allows for a focused examination of LLMs' internal capabilities. The use of a vanilla prompt, chosen for its neutrality, potentially constrains the full spectrum of LLM capabilities. Our carefully curated sample of papers ensures dataset homogeneity but necessarily limits generalizability across disciplines (Radicchi et al., 2008; Wang et al., 2013). However, to our knowledge, our study presents one of the first systematic comparisons of human and LLM citation patterns.

The observed biases in LLM-generated citations are presented without implementing specific mitigation strategies, as our primary aim was to identify and characterize these biases. While we uncover significant patterns in LLM citation behavior, including a pronounced "Matthew effect" (Larivière and Gingras, 2010), the causal mechanisms remain an open question for future research. Our analysis provides a snapshot of current LLM capabilities, but does not capture the potential dynamic effects on the evolution of scientific knowledge networks (Price, 1965).

Our methodological choice to focus on LLMs' parametric knowledge, excluding retrieval-augmented generation, stems from the hypothesis that current external database-reliant methods may not substantially alter information dissemination patterns (Evans, 2008; Fortunato et al., 2018). This approach, while limiting direct comparisons with retrieval-based methods, offers insights into LLMs' potential as independent reasoning engines in scientific inquiry (Truhn et al., 2023).

Ethical Considerations

Our findings reveal potential risks of amplifying existing biases in scientific discourse, particularly the preferential attachment to highly-cited works. This could inadvertently reinforce the "Matthew effect" in science, potentially impeding the dissemination of novel ideas and diverse perspectives.

Our work contributes to the responsible development of AI in scientific applications by providing a rigorous analysis of LLM internal citation patterns. By elucidating these patterns, we aim to inform the scientific community about the potential influences of AI tools on research practices. We believe this awareness is crucial for developing strategies to mitigate biases and ensure that AI systems enhance rather than hinder the diversity and momentum of scientific progress.

Acknowledgments

Andres Algaba acknowledges a fellowship from the Research Foundation Flanders under Grant No.1286924N. Vincent Ginis acknowledges support from Research Foundation Flanders under Grant No.G032822N and G0K9322N. The resources and services used in this work were provided by the VSC (Flemish Supercomputer Center), funded by the Research Foundation - Flanders (FWO) and the Flemish Government.

References

Alberto Acerbi and Joseph M Stubbersfield. 2023. Large language models show human-like content biases in transmission chain experiments. *Proceedings of the National Academy of Sciences*, 120(44):e2313790120.

Ayush Agrawal, Mirac Suzgun, Lester Mackey, and Adam Kalai. 2024. Do language models know when they're hallucinating references? In *Findings of the Association for Computational Linguistics: EACL 2024*, pages 912–928, St. Julian's, Malta. Association for Computational Linguistics.

Daniil A Boiko, Robert MacKnight, and Gabe Gomes. 2023. Emergent autonomous scientific research capabilities of large language models. *arXiv preprint arXiv:2304.05332*.

Lutz Bornmann and Hans-Dieter Daniel. 2008. What do citation counts measure? a review of studies on citing behavior. *Journal of documentation*, 64(1):45–80.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. 2023. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*.

Jiawei Chen, Hongyu Lin, Xianpei Han, and Le Sun. 2024. Benchmarking large language models in retrieval-augmented generation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(16):17754–17762.

James A Evans. 2008. Electronic publication and the narrowing of science and scholarship. *science*, 321(5887):395–399.

Nicholas Fabiano, Arnav Gupta, Nishaant Bhambra, Brandon Luu, Stanley Wong, Muhammad Maaz, Jess G Fiedorowicz, Andrew L Smith, and Marco Solmi. 2024. How to optimize the systematic review process using ai tools. *JCPP Advances*, page e12234.

Santo Fortunato, Carl T Bergstrom, Katy Börner, James A Evans, Dirk Helbing, Staša Milojević, Alexander M Petersen, Filippo Radicchi, Roberta Sinatra, Brian Uzzi, et al. 2018. Science of science. *Science*, 359(6379):eaao0185.

Eugene Garfield. 1965. Can citation indexing be automated. In *Statistical association methods for mechanized documentation, symposium proceedings*, volume 269, pages 189–192. Citeseer.

Carlos E Jimenez, John Yang, Alexander Wettig, Shunyu Yao, Kexin Pei, Ofir Press, and Karthik R Narasimhan. 2024. SWE-bench: Can language models resolve real-world github issues? In *The Twelfth International Conference on Learning Representations*.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, et al. 2022. Language models (mostly) know what they know. *arXiv preprint arXiv:2207.05221*.

Jean Kaddour, Joshua Harris, Maximilian Mozes, Herbie Bradley, Roberta Raileanu, and Robert McHardy. 2023. Challenges and applications of large language models. *arXiv preprint arXiv:2307.10169*.

Nikhil Kandpal, Haikang Deng, Adam Roberts, Eric Wallace, and Colin Raffel. 2023. Large language models struggle to learn long-tail knowledge. In *International Conference on Machine Learning*, pages 15696–15707. PMLR.

Norman Kaplan. 1965. The norms of citation behavior: Prolegomena to the footnote. *American documentation*, 16(3):179–184.

Rodney Kinney, Chloe Anastasiades, Russell Authur, Iz Beltagy, Jonathan Bragg, Alexandra Buraczynski, Isabel Cachola, Stefan Candra, Yoganand Chandrasekhar, Arman Cohan, et al. 2023. The semantic scholar open data platform. *arXiv preprint arXiv:2301.10140*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35:22199–22213.

Vincent Larivière and Yves Gingras. 2010. The impact factor's matthew effect: A natural experiment in bibliometrics. *Journal of the American Society for Information Science and Technology*, 61(2):424–427.

Peter A Lawrence. 2003. The politics of publication. *Nature*, 422(6929):259–261.

Adrian Letchford, Helen Susannah Moat, and Tobias Preis. 2015. The advantage of short paper titles. *Royal Society open science*, 2(8):150266.

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

Chris Lu, Cong Lu, Robert Tjarko Lange, Jakob Foerster, Jeff Clune, and David Ha. 2024. The ai scientist: Towards fully automated open-ended scientific discovery. *arXiv preprint arXiv:2408.06292*.

Huan Ma, Changqing Zhang, Yatao Bian, Lemao Liu, Zhirui Zhang, Peilin Zhao, Shu Zhang, Huazhu Fu, Qinghua Hu, and Bingzhe Wu. 2023. Fairness-guided few-shot prompting for large language models. In *Advances in Neural Information Processing Systems*, volume 36, pages 43136–43155. Curran Associates, Inc.

Marta Marchiori Manerba, Karolina Stańczak, Riccardo Guidotti, and Isabelle Augenstein. 2023. Social bias probing: Fairness benchmarking for language models. *arXiv preprint arXiv:2311.09090*.

Amil Merchant, Simon Batzner, Samuel S Schoenholz, Muratahan Aykol, Gowoon Cheon, and Ekin Dogus Cubuk. 2023. Scaling deep learning for materials discovery. *Nature*, 624(7990):80–85.

Roberto Navigli, Simone Conia, and Björn Ross. 2023. Biases in large language models: origins, inventory, and discussion. *ACM Journal of Data and Information Quality*, 15(2):1–21.

G Nigel Gilbert. 1977. Referencing as persuasion. *Social studies of science*, 7(1):113–122.

Derek de Solla Price. 1976. A general theory of bibliometric and other cumulative advantage processes. *Journal of the American society for Information science*, 27(5):292–306.

Derek J De Solla Price. 1965. Networks of scientific papers: The pattern of bibliographic references indicates the nature of the scientific research front. *Science*, 149(3683):510–515.

Riaz Qureshi, Daniel Shaughnessy, Kayden AR Gill, Karen A Robinson, Tianjing Li, and Eitan Agai. 2023. Are chatgpt and large language models "the answer" to bringing us closer to systematic review automation? *Systematic Reviews*, 12(1):72.

Filippo Radicchi, Santo Fortunato, and Claudio Castellano. 2008. Universality of citation distributions: Toward an objective measure of scientific impact. *Proceedings of the National Academy of Sciences*, 105(45):17268–17272.

Bernardino Romera-Paredes, Mohammadamin Barekatain, Alexander Novikov, Matej Balog, M Pawan Kumar, Emilien Dupont, Francisco JR Ruiz, Jordan S Ellenberg, Pengming Wang, Omar Fawzi, et al. 2024. Mathematical discoveries from program search with large language models. *Nature*, 625(7995):468–475.

Michael D Skarlinski, Sam Cox, Jon M Laurent, James D Braza, Michaela Hinks, Michael J Hammerling, Manvitha Ponnapati, Samuel G Rodriques, and Andrew D White. 2024. Language agents achieve superhuman synthesis of scientific knowledge. *arXiv preprint arXiv:2409.13740*.

Derek R Smith. 2012. Impact factors, scientometrics and the history of citation-based research. *Scientometrics*, 92(2):419–427.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Teo Susnjak, Peter Hwang, Napoleon H Reyes, Andre LC Barczak, Timothy R McIntosh, and Surangika Ranathunga. 2024. Automating research synthesis with domain-specific large language model fine-tuning. *arXiv preprint arXiv:2404.08680*.

Deepa Tilwani, Yash Saxena, Ali Mohammadi, Edward Raff, Amit Sheth, Srinivasan Parthasarathy, and Manas Gaur. 2024. Reasons: A benchmark for retrieval and automated citations of scientific sentences using public and proprietary llms. *arXiv preprint arXiv:2405.02228*.

Daniel Truhn, Jorge S. Reis-Filho, and Jakob Nikolas Kather. 2023. Large language models should be used as scientific reasoning engines, not knowledge databases. *Nature Medicine*, 29(12):2983–2984.

Marco Antonio Valenzuela-Escarcega, Vu A. Ha, and Oren Etzioni. 2015. Identifying meaningful citations. In *AAAI Workshop: Scholarly Big Data*.

William H Walters and Esther Isabelle Wilder. 2023. Fabrication and errors in the bibliographic citations generated by chatgpt. *Scientific Reports*, 13(1):14045.

Dashun Wang, Chaoming Song, and Albert-László Barabási. 2013. Quantifying long-term scientific impact. *Science*, 342(6154):127–132.

Jian Wang. 2014. Unpacking the Matthew effect in citations. *Journal of Informetrics*, 8(2):329–339.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2022. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*.

Duncan J Watts and Steven H Strogatz. 1998. Collective dynamics of 'small-world'networks. *nature*, 393(6684):440–442.

Jason Wei, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama, Maarten Bosma, Denny Zhou, Donald Metzler, et al. 2022a. Emergent abilities of large language models. *arXiv preprint arXiv:2206.07682*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. 2022b. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837.

Kevin Wu, Eric Wu, Ally Cassasola, Angela Zhang, Kevin Wei, Teresa Nguyen, Sith Riantawan, Patricia Shi Riantawan, Daniel E Ho, and James Zou. 2024. How well do llms cite relevant medical references? an evaluation framework and analyses. *arXiv preprint arXiv:2402.02008*.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. 2024. Tree of thoughts: Deliberate problem solving with large language models. *Advances in Neural Information Processing Systems*, 36.

Yue Zhang, Yafu Li, Leyang Cui, Deng Cai, Lemao Liu, Tingchen Fu, Xinting Huang, Enbo Zhao, Yu Zhang, Yulong Chen, et al. 2023. Siren's song in the ai ocean: a survey on hallucination in large language models. *arXiv preprint arXiv:2309.01219*.

Yizhen Zheng, Huan Yee Koh, Jiaxin Ju, Anh TN Nguyen, Lauren T May, Geoffrey I Webb, and Shirui Pan. 2023. Large language models for scientific synthesis, inference and explanation. *arXiv preprint arXiv:2310.07984*.

Appendix

We detail our data processing in Appendix A and show supplementary figures and tables.

A Data

We describe the steps of our automated pipeline to retrieve all the necessary information for our analysis. Our data collection resulted in 166 papers published at AAAI (25), NeurIPS (72), ICML (38), and ICLR (31) for a total of 3,066 references (see Appendix Table C3 for a full list of included papers). The data collection pipeline uses GPT-4-0613 to postprocess parts of the data, which costs approximately 14 dollars for our experiment. Note that these steps only have to be carried out once for the data collection. However, steps 4 and 5 are also used to postprocess and enrich the information of the generated references and will need to be carried out for each run. The experiment was run on 4 November 2023 and each step was manually verified and tested. Besides using GPT-4-0613, we also ran steps 6 and 7 for GPT-4o-2024-05-13 and Claude-3-5-sonnet-20240620 on 27 July 2024.

Step 1. ArXiv We search for all papers on arXiv originally posted between 1 March 2022 and 31 October 2023 in the machine learning (cs.LG) category which refer to AAAI, NeurIPS, ICLR, or ICML in their journal reference. Note that we also verify whether we can use all these arXiv papers given their data licenses and attribute their participation in Appendix Table C3. We use keywords (i.e., workshop, tiny paper, 2020, 2021, track on datasets and benchmarks, and bridge) to remove papers that do not appear in the conference proceedings or earlier than 2022. We download and unzip the *tar.gz* file provided by the authors to arXiv and check whether the paper exists on Semantic Scholar via title matching. We store the title, ID, and date from arXiv and Semantic Scholar. Additionally, we store all the reference titles with their corresponding ID from Semantic Scholar (Kinney et al., 2023).

Step 2. Tex We check whether there is a main *tex* file in the unzipped paper folder by looking for a single file that contains \begin{document} and \end{document}. If we find a main *tex* file we start the cleaning process, otherwise, we exclude the paper from our analysis. The cleaning process consists of three steps. First, we remove everything except for the author information, conference information, abstract, introduction, and references. Second, we remove figures, tables, references to sections and appendices, ... Finally, we transform all citations to numbers between square brackets. After the cleaning, we look at whether there is a *bib* or *bbl* file available and compile the *tex* to *PDF*. If neither file is available or the paper has compilation errors, we exclude the paper from our analysis (Appendix Table C4). Note that a *bib* file allows for both *PDFLatex* and *bibtex* compilation, while only a *bbl* file does not allow for *bibtex* compilation. As a consequence papers with only a *bbl* file may potentially contain papers in their reference list that are not cited in the introduction of the paper. We solve this issue in the next step.

Step 3. PDF We transform the *PDF* to *txt* and split the main content of the paper (author information, conference information, abstract, and introduction) from the references. We then look for all in-text citations by using a regex pattern to capture numbers in between square brackets and match them with the reference list. This approach ensures that we only keep references that are cited in the introduction. We store the main content of the paper and the references cited in the introduction in separate *txt* files.

Step 4. Postprocessing A large number of variations and inconsistencies in the reference lists makes it difficult to structurally extract and analyze all the author information, title, publication venue, and year. We noticed that this behavior was even more outspoken in the LLM-generated references. Therefore, we examine the capabilities of GPT-4 to impose a structure on the reference list by postprocessing the data. We feed GPT-4 the reference list in *txt* accompanied by the default system message: "*You are a helpful assistant*" and the following postprocessing prompt:

Below, we share with you a list of references with their corresponding citation number between square brackets. Could you for each reference extract the authors, the number of authors, title, publication year, and publication venue? Please only return the extracted information in a markdown table with the citation number (without brackets), authors, number of authors, title, publication year, and publication venue as columns.
===
[LLM generated reference list]

We then store the markdown table in a *csv*. GPT-4 successfully structures the information and makes it more consistent, for example, by removing syllable hyphens. Sometimes a small hick-up is introduced (e.g., adding a final row with "..."), but these are manually solved in the verification process. Note that we also prompt for the number of authors. While we can easily compute the number of authors via the meta-data from Semantic Scholar, it allows us to verify the accuracy of GPT-4 on this task as we will use it later on to postprocess the generated references where a ground truth may be unavailable.

Step 5. Semantic Scholar We enrich the information from the introduction references by matching the extracted title from the *csv* file in the previous step with the reference titles that we extracted from Semantic Scholar in step 1. This approach provides an extra check that GPT-4 does not change the title information in Step 4. After matching, we can use the Semantic Scholar ID to retrieve the publication venue, year, authors, citation count, influential citation count, and reference count (Kinney et al., 2023). Additionally, we store the IDs of the papers to which the introduction references themselves refer.

Step 6. "Vanilla" prompting We prompt GPT-4-0613 with the main content, which includes the author information, conference information, abstract, and introduction, accompanied by the default system message: "*You are a helpful assistant*" and the following prompt:

Below, we share with you a written introduction to a paper and have omitted the references. Numbers between square brackets indicate citations. Can you give us a suggestion for an explicit reference associated with each number? Do not return anything except the citation number between square brackets and the corresponding reference.
===
[main content]

We then post-process GPT-4's response to extract the title, venue, publication year, author names, and number of authors for each generated reference using the same approach as described in step 4. We repeat this "vanilla" approach five times for all 166 papers.

Step 7. Existence check We determine whether the generated references exist via title and author names matching with Semantic Scholar entries (Kinney et al., 2023). We search Semantic Scholar for the three best matches based on the reference's title and then compute the title and author names similarity. For titles, we measure the similarity between the Semantic Scholar match and the generated reference by comparing the best matching substring. For authors, we compare them by splitting into tokens (words), removing duplicates, and then calculating the similarity based on the best partial match of the sets of tokens. In case of "et al.," we only consider the first author. The similarity is computed by character-level comparison. We determined the thresholds for the title and authors scores by manually labelling 100 matches as true or false and minimizing the false positive rate. We obtain on this sample an accuracy of 95% with 5 false positives, i.e. generated references falsely classified as non-existent.

Step 8. "Iterative" prompting We also build on our "vanilla" approach, by introducing an "iterative" approach where we prompt GPT-4-0613 with the main content accompanied by the default system message: "*You are a helpful assistant*" and the following prompt:

[vanilla prompt + LLM's response]
The following references associated with these citation numbers:
[numbers of non-existent generated references]
do not exist. Can you replace all these non-existent references with existing ones? Keep the other references as they are. Do not return anything except the citation number between square brackets and the corresponding reference.
===
[main content]

We again postprocess GPT-4's response using the same approach as described in steps 4, 5, and 7. The previously existing generated and the newly generated references are then merged.

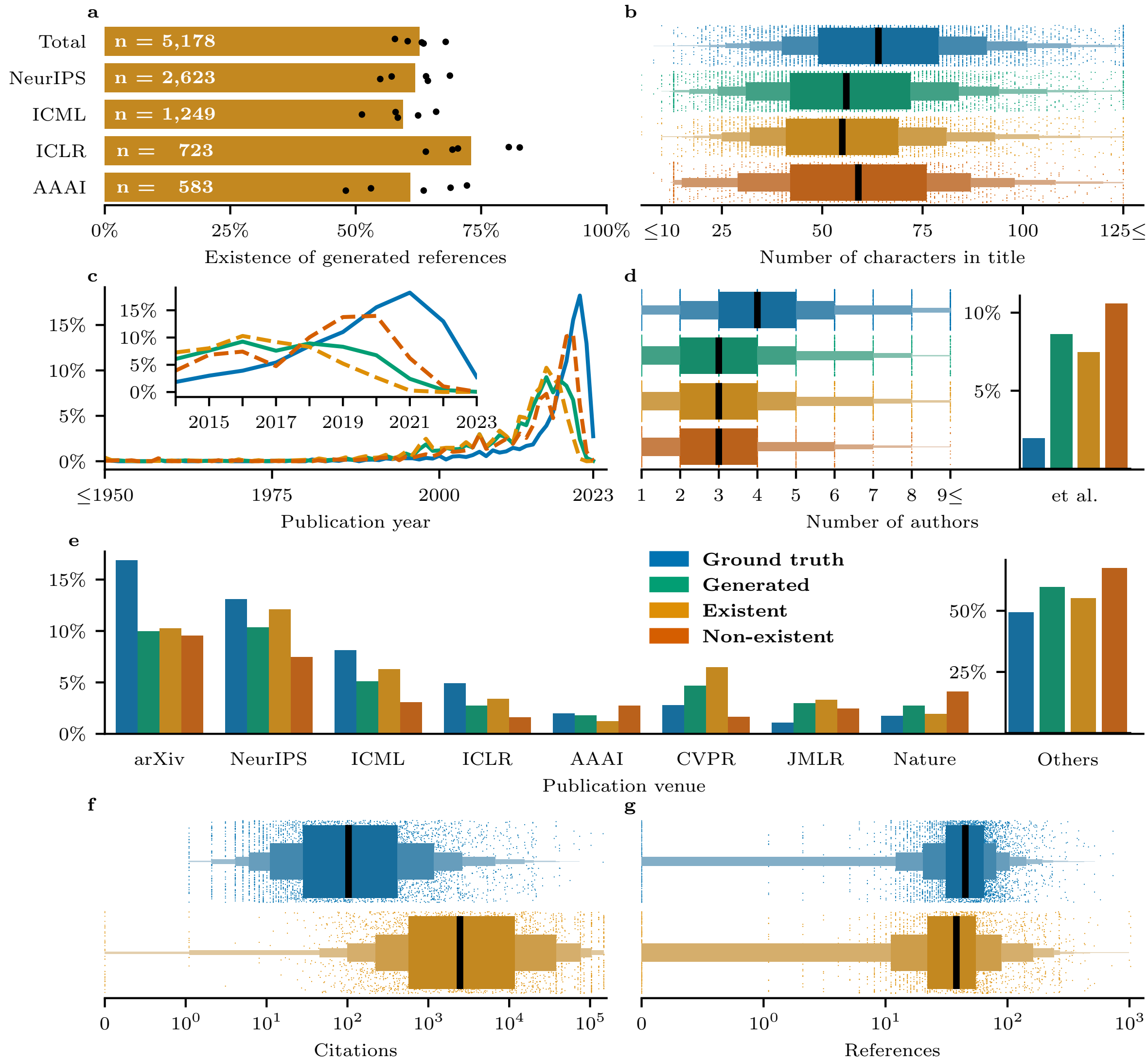

Figure B1: | **Properties of the ground truth and GPT-4 generated introduction references for the iterative strategy are consistent with the properties of the vanilla strategy.** This figure displays the properties of the ground truth ($n = 5,178$, in blue) and GPT-4 generated references ($n = 5,178$, in green), further subdividing the generated references into existing ($n = 3,244$, in orange) and non-existent categories ($n = 1,934$, in red), from the original data sources of five runs for the iterative strategy with GPT-4. Note that these are the references which are labelled “non-existent” in the vanilla strategy. **a**, **b**, **c**, **d**, **e**, **f** and **g,** The iterative results exhibit very similar properties to the vanilla results shown in Figure 2.

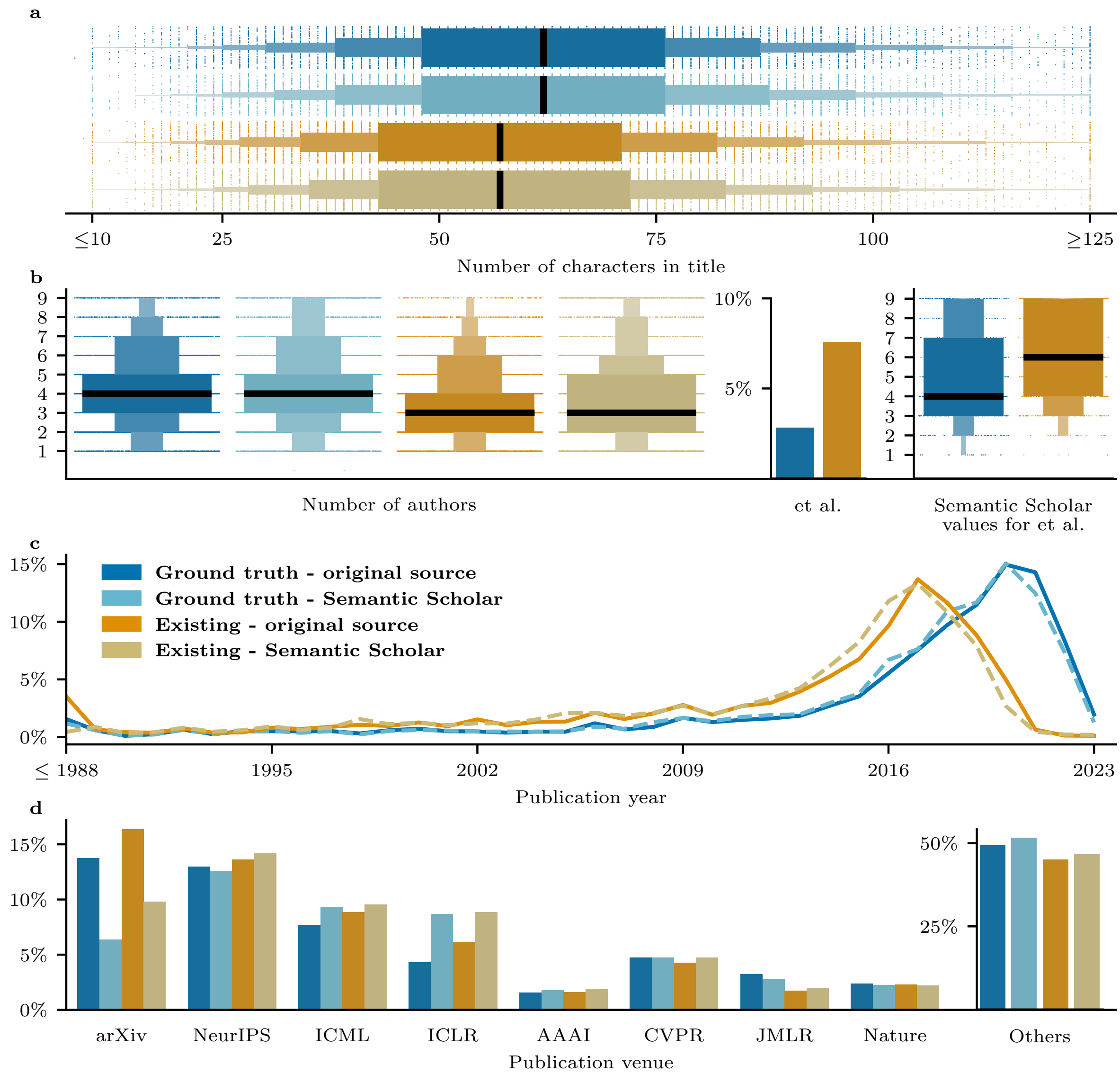

Figure B2: | **The properties of the existing GPT-4 generated references and their corresponding ground truth are consistent between the original data sources and Semantic Scholar data.** This figure compares the computation of the properties of the existing GPT-4 generated references ($n = 9.376$) and their corresponding ground truth ($n = 9.376$) between the data from the original sources (in dark blue and orange, as shown in Figure 2) and from Semantic Scholar (in light blue and orange). **a,** The distributions of the number of characters in title for the existing generated references and their corresponding ground truth are very similar between the data from the original sources and Semantic Scholar. **b,** There is a discrepancy between the data from the original sources and Semantic Scholar for the number of authors in the existing generated references due to the extensive use of "et al". This discrepancy results in a relatively larger portion of three authors or more, but does not change the previous conclusions. **c,** The distributions over time are very similar between the data from the original sources and Semantic Scholar. **d,** There is a discrepancy between the data from the original sources and Semantic Scholar for the publication venues. The discrepancy is consistent across the existing generated references and their corresponding ground truth as both have a lower number of arXiv papers and a larger number of ICLR papers.

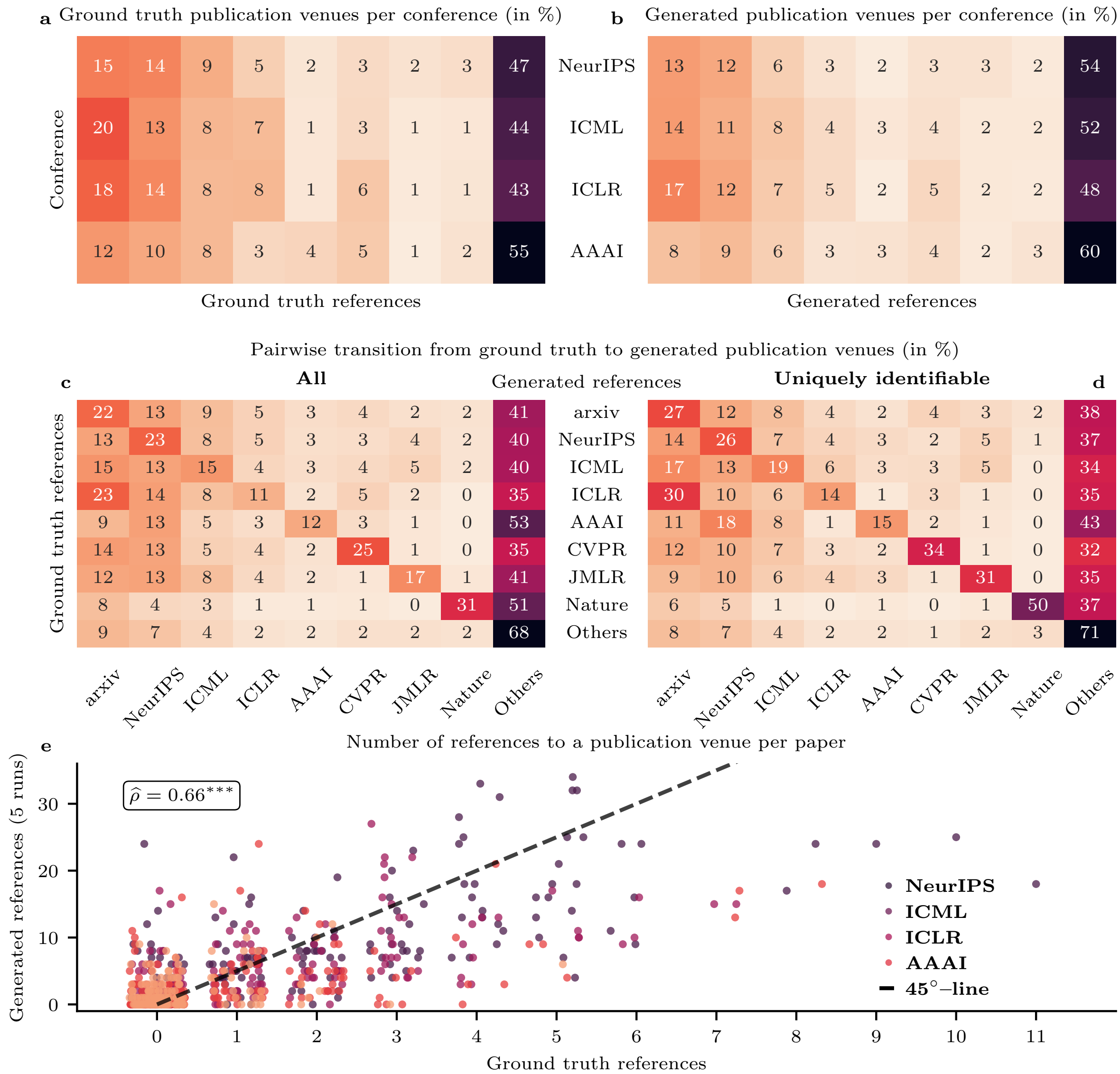

Figure B3: | **A high consistency in publication venue distributions between ground truth and GPT-4 generated references with a notable bias towards arXiv, NeurIPS, and "Others".** This figure displays the distributions and pairwise transition of publication venues for ground truth and GPT-4 generated references at the conference and individual paper and reference level. **a** and **b,** We observe that the distributions of publication venues for both ground truth and generated references are very similar across the various conferences, i.e., AAAI, NeurIPS, ICML, and ICLR. **c** and **d,** The pairwise transition matrix from ground truth to generated publication venues at the reference level indicates a large overall agreement, but with a strong preference in GPT-4 generated references for arXiv, NeurIPS, and "Others" in the case of disagreement. The preference for NeurIPS may be due to the relatively large number of NeurIPS papers in our sample and the large share of arXiv and "Others" points to favoring a wider array of venues which may potentially dilute the perceived relevance of key conferences. **e,** The scatter plot shows for each paper the number of ground truth references to one of the top conferences, i.e., AAAI, NeurIPS, ICML, and ICLR, and the corresponding number of generated references (×5 for five runs) which refer to the same conference.The strong pairwise correlation between the ground truth and generated references to the top conferences at the individual paper level affirms the high consistency in publication venue distributions between ground truth and GPT-4 generated references.

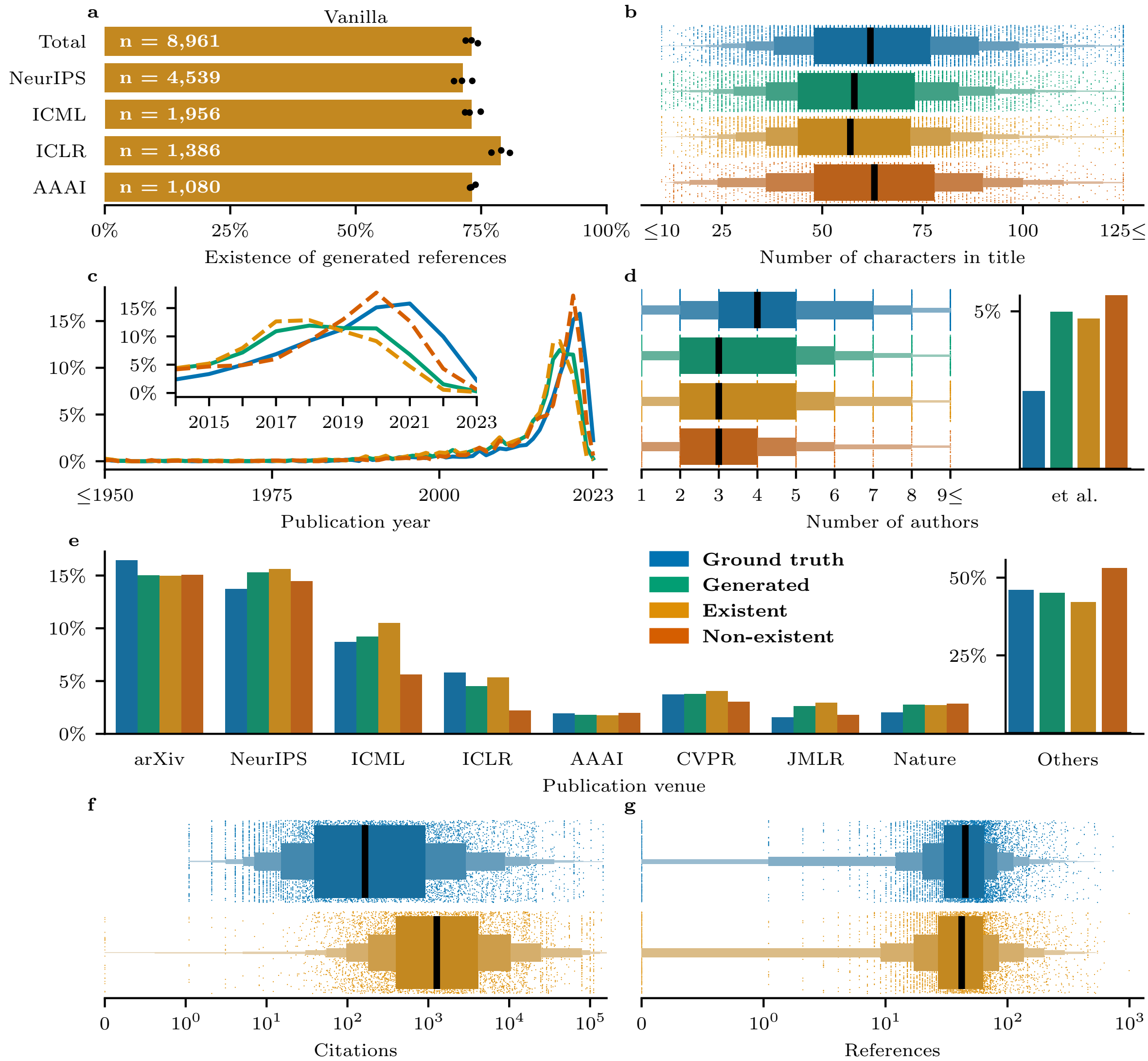

Figure B4: **Properties of the ground truth and GPT-4o generated introduction references are consistent with the properties of GPT-4.** This figure displays the properties of the ground truth ($n = 8,961$, in blue) and GPT-4o generated references ($n = 8,961$, in green), further subdividing the generated references into existing ($n = 6,552$, in orange) and non-existent categories ($n = 2,409$, in red), from the original data sources of three runs for the vanilla strategy with GPT-4o. **a**, **b**, **c**, **d**, **e**, **f** and **g,** The GPT-4o generated references exhibit very similar properties to the GPT-4 results shown in Figure 2, except for the existence rate which may be due to the papers now being part of the training data and the model's enhanced capabilities.

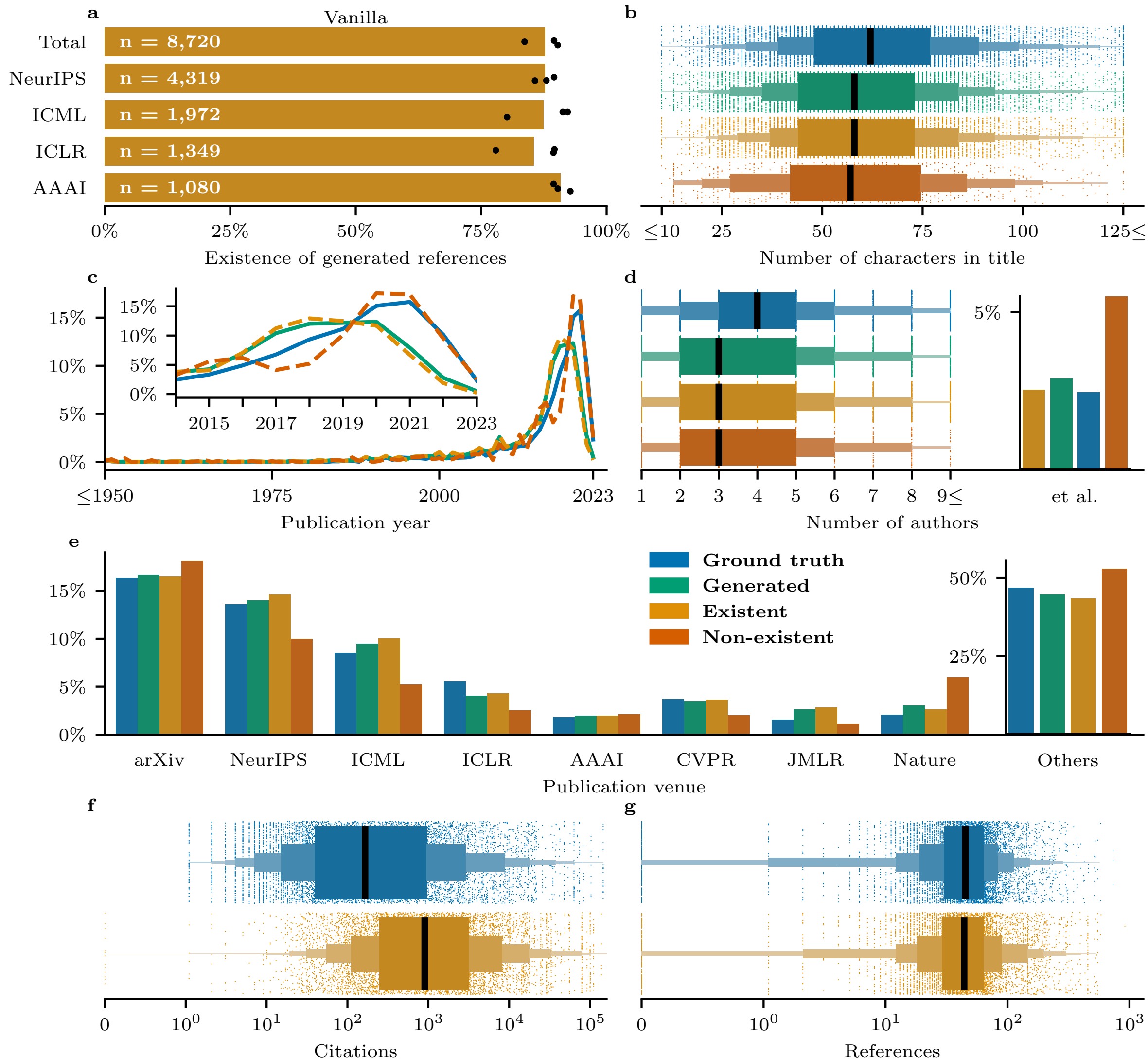

Figure B5: | **Properties of the ground truth and Claude 3.5 generated introduction references are consistent with the properties of GPT-4.** This figure displays the properties of the ground truth ($n = 2,893$, in blue) and Claude 3.5 generated references ($n = 2,893$, in green), further subdividing the generated references into existing ($n = 2,611$, in orange) and non-existent categories ($n = 282$, in red), from the original data sources of three runs for the vanilla strategy with GPT-4o. **a**, **b**, **c**, **d**, **e**, **f** and **g**, The Claude 3.5 generated references exhibit similar properties to the GPT-4 results shown in Figure 2, except for the existence rate which may be due to the papers now being part of the training data and the model's enhanced capabilities. Additionally, Claude 3.5, on average, generates non-existent references with shorter titles and proportionally published more in arXiv, Nature, and "others."

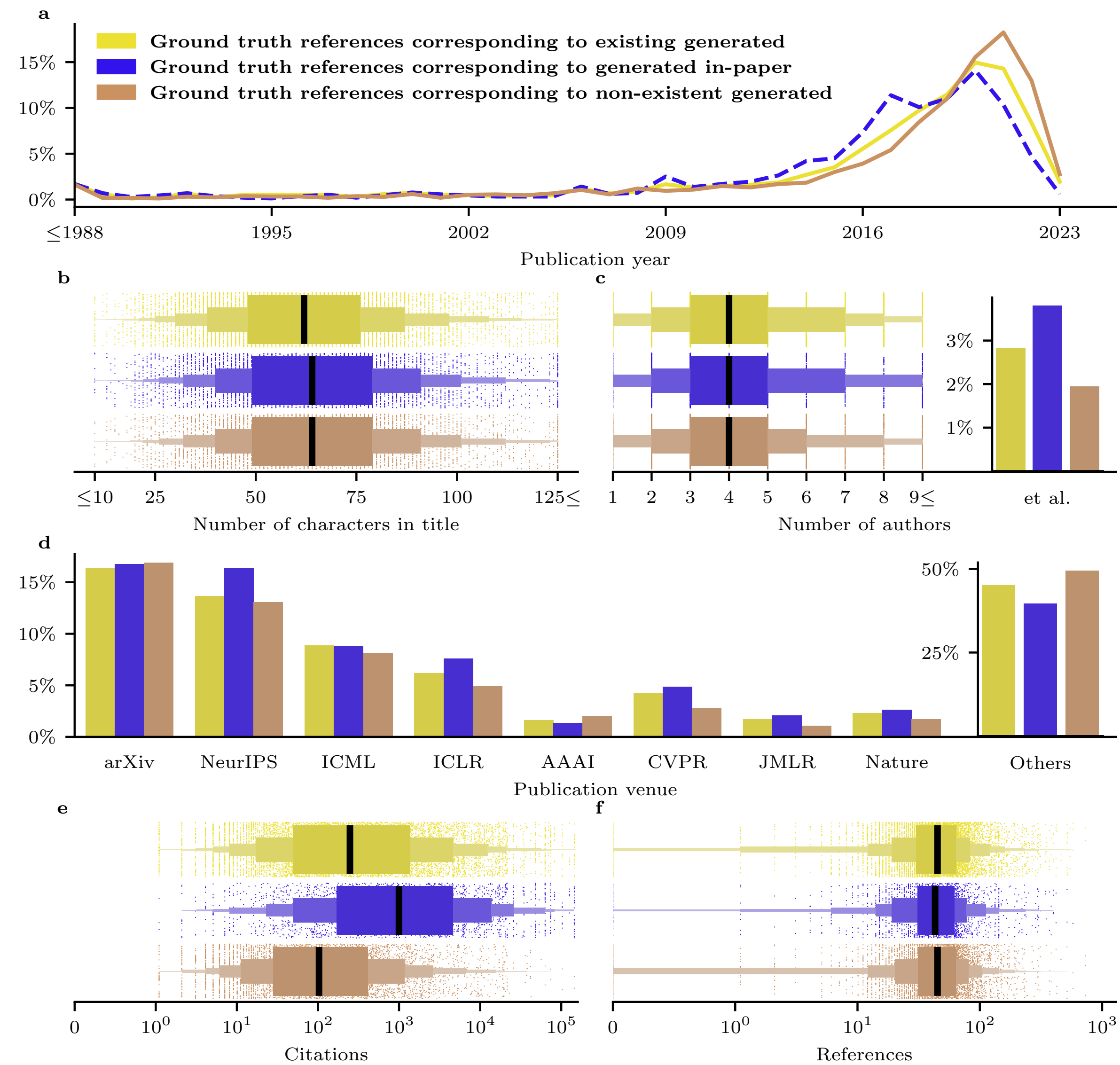

Figure B6: | **Ground truth papers which correspond to existing GPT-4 generated references that appear in the paper have substantially more citations.** This figure displays the properties of the ground truth references which correspond to existing GPT-4 generated references ($n = 9,376$, in yellow), the subset of existing generated references which appear in the paper itself ($n = 2,474$, in blue), and the non-existent generated references ($n = 5,178$, in green), from the original data sources of five runs for the vanilla strategy with GPT-4. **a**, **b**, **c**, **d**, **e** and **f,** The ground truth papers which correspond to existing references which appear in the paper have by far the most citations, followed by the existing references, and the ground truth papers corresponding to non-existent references have the lowest numbers of citations. These findings further indicate the tendency for LLMs to more easily generate references to highly cited papers. The distributions of all other characteristics are very similar.

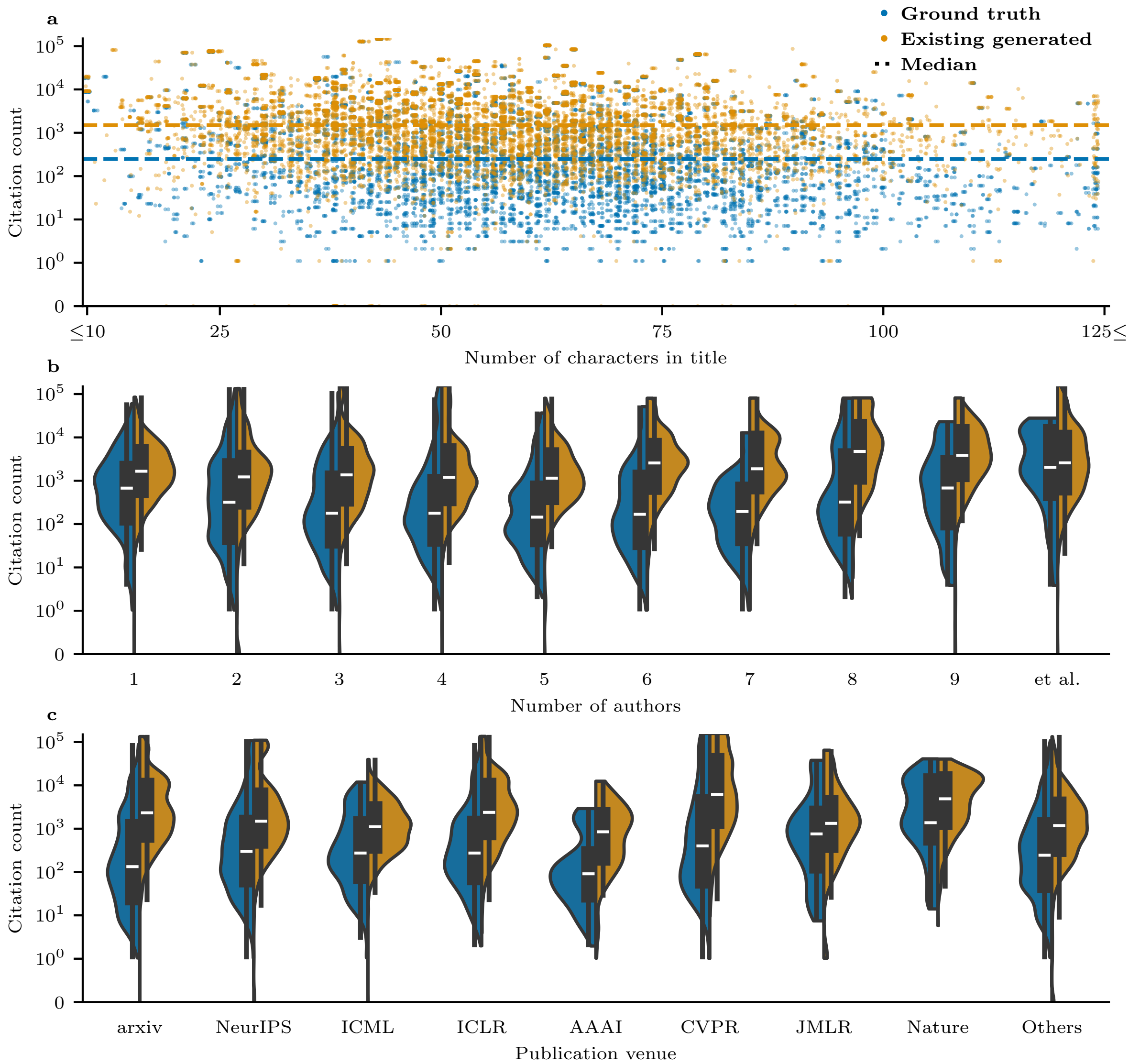

Figure B7: **The citation bias in existing GPT-4 generated references is consistent across title length, number of authors, and publication venue.** This figure shows that the existing GPT-4 generated references ($n = 9,376$, in orange) consistently exhibit a higher citation count compared to their corresponding ground truth ($n = 9,376$, in blue) across title length, number of authors, and publication venue. **a,** The citation counts across the number of characters in title reveals that the discrepancy in number of citations between the existing generated and ground truth references is consistent over various title lengths. **b** and **c,** The distributions of citation counts per number of authors and publication venues show that the existing generated references consistently exhibit a higher citation count than their corresponding ground truth counterparts.

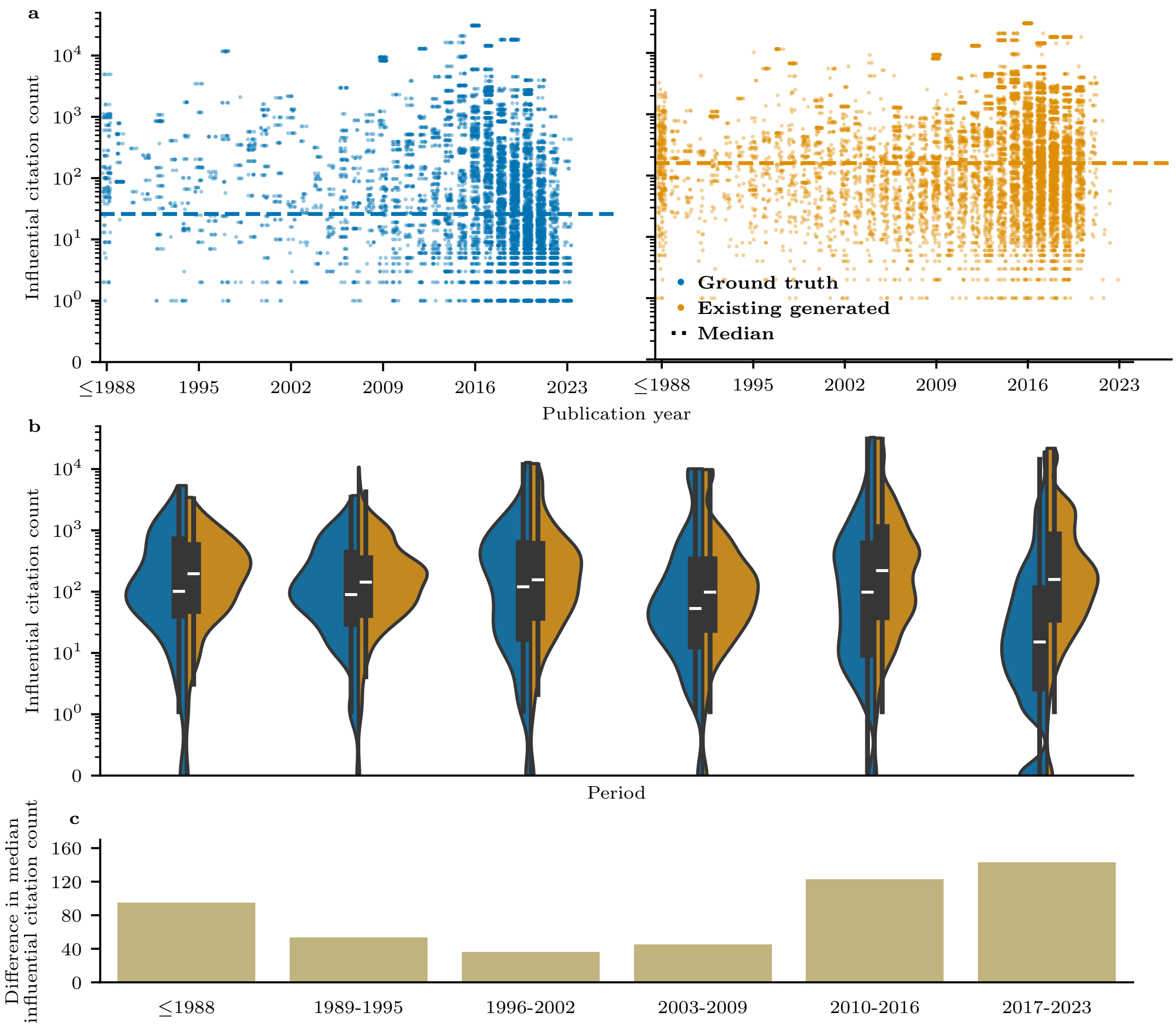

Figure B8: | **The influential citation bias in existing GPT-4 generated references is unrelated to the recency of ground truth references.** This figure shows that the existing GPT-4 generated references ($n = 9,376$, in orange) consistently exhibit a higher influential citation count compared to their corresponding ground truth ($n = 9,376$, in blue) across subperiods. **a**, **b** and **c,** Note that the influential citation count is retrieved from Semantic Scholar (Valenzuela-Escarcega et al., 2015).

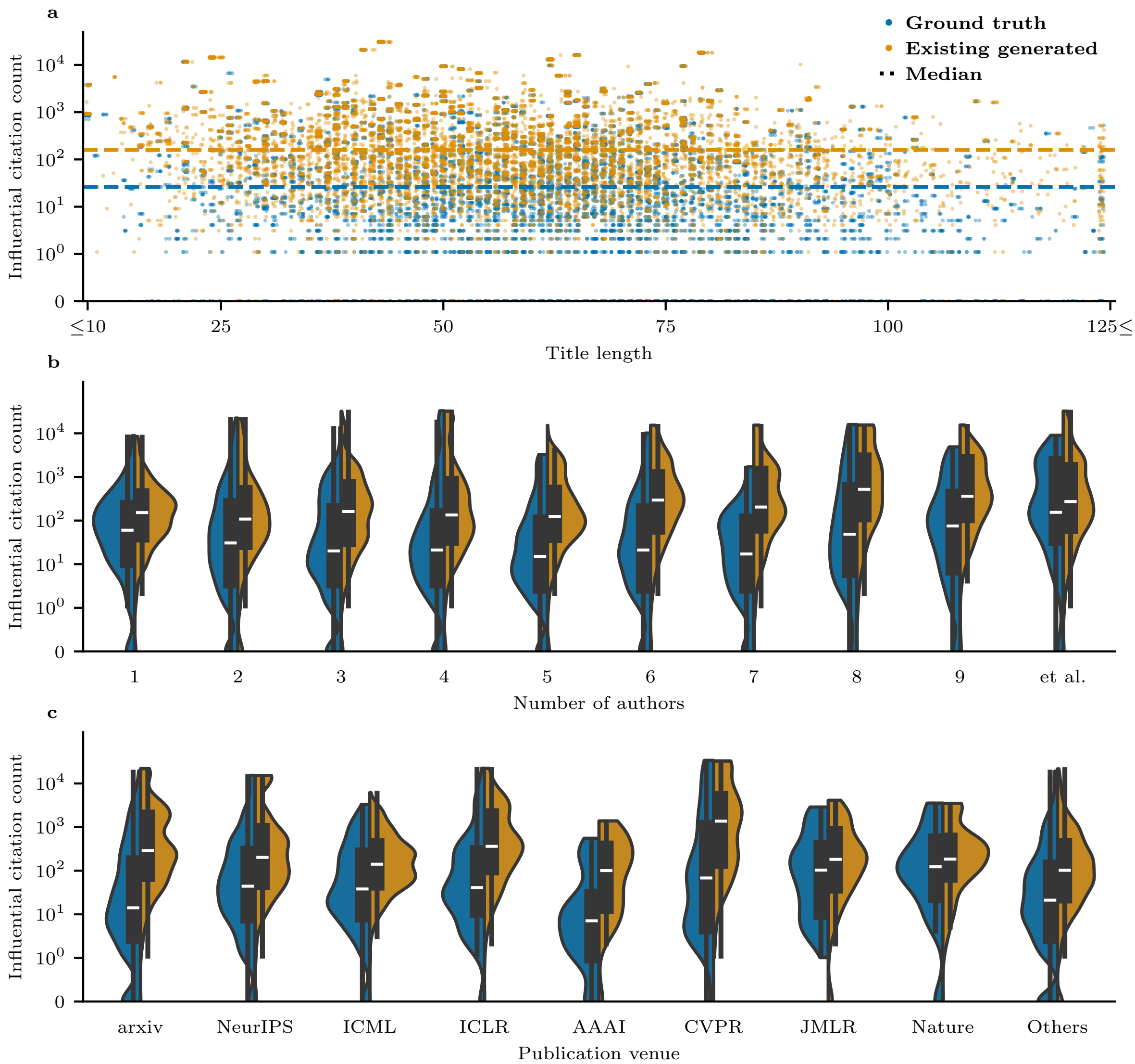

Figure B9: | **The infuential citation bias in existing GPT-4 generated references is unrelated to title length, number of authors, and publication venue.** This figure shows that the existing GPT-4 generated references ($n = 9,376$, in orange) consistently exhibit a higher influential citation count compared to their corresponding ground truth ($n = 9,376$, in blue) across title length, number of authors, and publication venue. **a**, **b** and **c,** Note that the influential citation count is retrieved from Semantic Scholar (Valenzuela-Escarcega et al., 2015).

Vanilla	Run 1	Run 2	Run 3	Run 4
Run 2	17.90			
Run 3	17.11	17.30		
Run 4	17.73	16.69	16.35	
Run 5	18.26	17.78	17.06	18.44

Table C1: **Overlap between generated sets of references of different runs by GPT-4.** We see on average a 17% overlap between different runs, which indicate that the models do not suffer from mode collapse (numbers in % with respect to total number of references).

Vanilla	GPT-4o			Claude 3.5		
	Run 1	Run 2	Run 3	Run 1	Run 2	Run 3
Existing generations in Semantic Scholar database (%)	74.34	73.10	71.92	90.25	89.53	83.66
Existing generations cited in the original paper (%)	32.16	33.28	33.71	42.08	41.50	39.26
Existing generations cited in the original intro (%)	24.15	26.23	25.81	34.23	33.53	31.85
Existing generations with a pairwise match (%) (for all references)	10.34	10.62	10.39	18.04	17.24	14.74
Existing generations with a pairwise match (%) (for uniquely identifiable references)	17.49	19.10	18.28	29.75	29.25	25.64

Table C2: **Summary statistics of generated references by GPT-4o and Claude 3.5.** (numbers in % with respect to total number of references).

Conference	Authors	Title
AAAI	Jakob Weissteiner, Jakob Heiss, Julien Siems, Sven Seuken	Bayesian Optimization-based Combinatorial Assignment
AAAI	Gobinda Saha, Kaushik Roy	Continual Learning with Scaled Gradient Projection
AAAI	Ruizhe Zheng, Jun Li, Yi Wang, Tian Luo, Yuguo Yu	ScatterFormer: Locally-Invariant Scattering Transformer for Patient-Independent Multispectral Detection of Epileptiform Discharges
AAAI	Sahil Manchanda, Sayan Ranu	Lifelong Learning for Neural powered Mixed Integer Programming
AAAI	Joris Guérin, Kevin Delmas, Raul Sena Ferreira, Jérémie Guiochet	Out-Of-Distribution Detection Is Not All You Need
AAAI	Taha Belkhouja, Yan Yan, Janardhan Rao Doppa	Training Robust Deep Models for Time-Series Domain: Novel Algorithms and Theoretical Analysis
AAAI	Minsoo Kang, Suhyun Kim	GuidedMixup: An Efficient Mixup Strategy Guided by Saliency Maps
AAAI	Su Kim, Dongha Lee, SeongKu Kang, Seonghyeon Lee, Hwanjo Yu	Learning Topology-Specific Experts for Molecular Property Prediction
AAAI	Daniel Silver, Tirthak Patel, Devesh Tiwari	QUILT: Effective Multi-Class Classification on Quantum Computers Using an Ensemble of Diverse Quantum Classifiers
AAAI	Kevin Osanlou, Jeremy Frank, Andrei Bursuc, Tristan Cazenave, Eric Jacopin, Christophe Guettier, J. Benton	Solving Disjunctive Temporal Networks with Uncertainty under Restricted Time-Based Controllability using Tree Search and Graph Neural Networks
AAAI	Joar Skalse, Alessandro Abate	Misspecification in Inverse Reinforcement Learning
AAAI	Edward Ayers, Jonathan Sadeghi, John Redford, Romain Mueller, Puneet K. Dokania	Query-based Hard-Image Retrieval for Object Detection at Test Time
AAAI	Shubham Gupta, Sahil Manchanda, Srikanta Bedathur, Sayan Ranu	TIGGER: Scalable Generative Modelling for Temporal Interaction Graphs
AAAI	Fanchen Bu, Dong Eui Chang	Feedback Gradient Descent: Efficient and Stable Optimization with Orthogonality for DNNs
AAAI	Shota Saito	Hypergraph Modeling via Spectral Embedding Connection: Hypergraph Cut, Weighted Kernel k-means, and Heat Kernel
AAAI	Haoran Luo, Haihong E, Ling Tan, Gengxian Zhou, Tianyu Yao, Kaiyang Wan	DHGE: Dual-View Hyper-Relational Knowledge Graph Embedding for Link Prediction and Entity Typing
AAAI	Yujin Kim, Dogyun Park, Dohee Kim, Suhyun Kim	NaturalInversion: Data-Free Image Synthesis Improving Real-World Consistency
AAAI	Tairan He, Weiye Zhao, Changliu Liu	AutoCost: Evolving Intrinsic Cost for Zero-violation Reinforcement Learning
AAAI	Shijie Liu, Andrew C. Cullen, Paul Montague, Sarah M. Erfani, Benjamin I. P. Rubinstein	Enhancing the Antidote: Improved Pointwise Certifications against Poisoning Attacks

Table C3: **Papers included in the analysis.**

Conference	Authors	Title
AAAI	Fan Zhou, Chen Pan, Lintao Ma, Yu Liu, Shiyu Wang, James Zhang, Xinxin Zhu, Xuanwei Hu, Yunhua Hu, Yangfei Zheng, Lei Lei, Yun Hu	SLOTH: Structured Learning and Task-based Optimization for Time Series Forecasting on Hierarchies
AAAI	Christopher W. F. Parsonson, Alexandre Laterre, Thomas D. Barrett	Reinforcement Learning for Branch-and-Bound Optimisation using Retrospective Trajectories
AAAI	Sourya Basu, Prasanna Sattigeri, Karthikeyan Natesan Ramamurthy, Vijil Chenthamarakshan, Kush R. Varshney, Lav R. Varshney, Payel Das	Equi-Tuning: Group Equivariant Fine-Tuning of Pretrained Models
AAAI	Harry Rubin-Falcone, Joyce Lee, Jenna Wiens	Forecasting with Sparse but Informative Variables: A Case Study in Predicting Blood Glucose
AAAI	Pierre Le Pelletier de Woillemont, Rémi Labory, Vincent Corruble	Automated Play-Testing Through RL Based Human-Like Play-Styles Generation
AAAI	Kai Klede, Leo Schwinn, Dario Zanca, Björn Eskofier	FastAMI – a Monte Carlo Approach to the Adjustment for Chance in Clustering Comparison Metrics
NeurIPS	Dhananjay Bhaskar, Kincaid MacDonald, Oluwadamilola Fasina, Dawson Thomas, Bastian Rieck, Ian Adelstein, Smita Krishnaswamy	Diffusion Curvature for Estimating Local Curvature in High Dimensional Data
NeurIPS	Shiro Takagi	On the Effect of Pre-training for Transformer in Different Modality on Offline Reinforcement Learning
NeurIPS	Yue Yu, Yuchen Zhuang, Jieyu Zhang, Yu Meng, Alexander Ratner, Ranjay Krishna, Jiaming Shen, Chao Zhang	Large Language Model as Attributed Training Data Generator: A Tale of Diversity and Bias
NeurIPS	Lingfeng Sun, Haichao Zhang, Wei Xu, Masayoshi Tomizuka	PaCo: Parameter-Compositional Multi-Task Reinforcement Learning
NeurIPS	Yang Yue, Rui Lu, Bingyi Kang, Shiji Song, Gao Huang	Understanding, Predicting and Better Resolving Q-Value Divergence in Offline-RL
NeurIPS	Jiaqi Leng, Yuxiang Peng, Yi-Ling Qiao, Ming Lin, Xiaodi Wu	Differentiable Analog Quantum Computing for Optimization and Control
NeurIPS	Kyriakos Flouris, Ender Konukoglu	Canonical normalizing flows for manifold learning
NeurIPS	Yuchen Bai, Jean-Baptiste Durand, Florence Forbes, Grégoire Vincent	Semantic segmentation of sparse irregular point clouds for leaf wood discrimination
NeurIPS	Lorenzo Giambagli, Lorenzo Buffoni, Lorenzo Chicchi, Duccio Fanelli	How a student becomes a teacher: learning and forgetting through Spectral methods
NeurIPS	Hanbyul Lee, Qifan Song, Jean Honorio	Support Recovery in Sparse PCA with Incomplete Data
NeurIPS	Zhang-Wei Hong, Aviral Kumar, Sathwik Karnik, Abhishek Bhandwaldar, Akash Srivastava, Joni Pajarinen, Romain Laroche, Abhishek Gupta, Pulkit Agrawal	Beyond Uniform Sampling: Offline Reinforcement Learning with Imbalanced Datasets
NeurIPS	Xiang Zhang, Ziyuan Zhao, Theodoros Tsiligkaridis, Marinka Zitnik	Self-Supervised Contrastive Pre-Training For Time Series via Time-Frequency Consistency

Table C3: **Papers included in the analysis.**

Conference	Authors	Title
NeurIPS	Antonin Schrab, Ilmun Kim, Benjamin Guedj, Arthur Gretton	Efficient Aggregated Kernel Tests using Incomplete U-statistics
NeurIPS	Wanyun Cui, Xingran Chen	Instance-based Learning for Knowledge Base Completion
NeurIPS	Aurelien Lucchi, Frank Proske, Antonio Orvieto, Francis Bach, Hans Kersting	On the Theoretical Properties of Noise Correlation in Stochastic Optimization
NeurIPS	Minsik Cho, Saurabh Adya, Devang Naik	PDP: Parameter-free Differentiable Pruning is All You Need
NeurIPS	Guangxi Li, Ruilin Ye, Xuanqiang Zhao, Xin Wang	Concentration of Data Encoding in Parameterized Quantum Circuits
NeurIPS	Xinrui Wang, Wenhai Wan, Chuanxin Geng, Shaoyuan LI, Songcan Chen	Beyond Myopia: Learning from Positive and Unlabeled Data through Holistic Predictive Trends
NeurIPS	Zihan Liu, Yun Luo, Lirong Wu, Zicheng Liu, Stan Z. Li	Towards Reasonable Budget Allocation in Untargeted Graph Structure Attacks via Gradient Debias
NeurIPS	Dingfan Chen, Raouf Kerkouche, Mario Fritz	Private Set Generation with Discriminative Information
NeurIPS	Zhan Yu, Hongshun Yao, Mujin Li, Xin Wang	Power and limitations of single-qubit native quantum neural networks
NeurIPS	Ibrahim Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, Lucas Beyer	Getting ViT in Shape: Scaling Laws for Compute-Optimal Model Design
NeurIPS	Manzil Zaheer, Kenneth Marino, Will Grathwohl, John Schultz, Wendy Shang, Sheila Babayan, Arun Ahuja, Ishita Dasgupta, Christine Kaeser-Chen, Rob Fergus	Learning to Navigate Wikipedia by Taking Random Walks
NeurIPS	Dohyun Kwon, Ying Fan, Kangwook Lee	Score-based Generative Modeling Secretly Minimizes the Wasserstein Distance
NeurIPS	Zhaoqi Li, Lillian Ratliff, Houssam Nassif, Kevin Jamieson, Lalit Jain	Instance-optimal PAC Algorithms for Contextual Bandits
NeurIPS	Masaki Adachi, Satoshi Hayakawa, Martin Jørgensen, Harald Oberhauser, Michael A. Osborne	Fast Bayesian Inference with Batch Bayesian Quadrature via Kernel Recombination
NeurIPS	Zhiqin Yang, Yonggang Zhang, Yu Zheng, Xinmei Tian, Hao Peng, Tongliang Liu, Bo Han	FedFed: Feature Distillation against Data Heterogeneity in Federated Learning
NeurIPS	Daniel Vial, Sujay Sanghavi, Sanjay Shakkottai, R. Srikant	Minimax Regret for Cascading Bandits
NeurIPS	Fabian Zaiser, Andrzej S. Murawski, Luke Ong	Exact Bayesian Inference on Discrete Models via Probability Generating Functions: A Probabilistic Programming Approach
NeurIPS	Cheng Chi, Amine Mohamed Aboussalah, Elias B. Khalil, Juyoung Wang, Zoha Sherkat-Masoumi	A Deep Reinforcement Learning Framework For Column Generation
NeurIPS	Mathieu Molina, Patrick Loiseau	Bounding and Approximating Intersectional Fairness through Marginal Fairness

Table C3: **Papers included in the analysis.**

Conference	Authors	Title
NeurIPS	Shuai Zhang, Hongkang Li, Meng Wang, Miao Liu, Pin-Yu Chen, Songtao Lu, Sijia Liu, Keerthiram Murugesan, Subhajit Chaudhury	On the Convergence and Sample Complexity Analysis of Deep Q-Networks with ε-Greedy Exploration
NeurIPS	Changlong Wu, Mohsen Heidari, Ananth Grama, Wojciech Szpankowski	Precise Regret Bounds for Log-loss via a Truncated Bayesian Algorithm
NeurIPS	Ching-Yao Chuang, Stefanie Jegelka	Tree Mover's Distance: Bridging Graph Metrics and Stability of Graph Neural Networks
NeurIPS	Felix Biggs, Antonin Schrab, Arthur Gretton	MMD-FUSE: Learning and Combining Kernels for Two-Sample Testing Without Data Splitting
NeurIPS	Thomas Fel, Victor Boutin, Mazda Moayeri, Rémi Cadène, Louis Bethune, Léo andéol, Mathieu Chalvidal, Thomas Serre	A Holistic Approach to Unifying Automatic Concept Extraction and Concept Importance Estimation
NeurIPS	Manel Baradad, Chun-Fu Chen, Jonas Wulff, Tongzhou Wang, Rogerio Feris, Antonio Torralba, Phillip Isola	Procedural Image Programs for Representation Learning
NeurIPS	Yang Ni	Bivariate Causal Discovery for Categorical Data via Classification with Optimal Label Permutation
NeurIPS	Gauthier Guinet, Saurabh Amin, Patrick Jaillet	Effective Dimension in Bandit Problems under Censorship
NeurIPS	Kyungmin Lee, Jinwoo Shin	RenyiCL: Contrastive Representation Learning with Skew Renyi Divergence
NeurIPS	Yihe Wang, Yu Han, Haishuai Wang, Xiang Zhang	Contrast Everything: A Hierarchical Contrastive Framework for Medical Time-Series
NeurIPS	Artyom Sorokin, Nazar Buzun, Leonid Pugachev, Mikhail Burtsev	Explain My Surprise: Learning Efficient Long-Term Memory by Predicting Uncertain Outcomes
NeurIPS	Yipeng Kang, Tonghan Wang, Xiaoran Wu, Qianlan Yang, Chongjie Zhang	Non-Linear Coordination Graphs
NeurIPS	Niv Giladi, Shahar Gottlieb, Moran Shkolnik, Asaf Karnieli, Ron Banner, Elad Hoffer, Kfir Yehuda Levy, Daniel Soudry	DropCompute: simple and more robust distributed synchronous training via compute variance reduction
NeurIPS	Jack Richter-Powell, Yaron Lipman, Ricky T. Q. Chen	Neural Conservation Laws: A Divergence-Free Perspective
NeurIPS	Peide Huang, Mengdi Xu, Jiacheng Zhu, Laixi Shi, Fei Fang, Ding Zhao	Curriculum Reinforcement Learning using Optimal Transport via Gradual Domain Adaptation
NeurIPS	Mark D. McDonnell, Dong Gong, Amin Parveneh, Ehsan Abbasnejad, Anton van den Hengel	RanPAC: Random Projections and Pre-trained Models for Continual Learning
NeurIPS	Haoyuan Sun, Kwangjun Ahn, Christos Thrampoulidis, Navid Azizan	Mirror Descent Maximizes Generalized Margin and Can Be Implemented Efficiently
NeurIPS	Rui M. Castro, Fredrik Hellström, Tim van Erven	Adaptive Selective Sampling for Online Prediction with Experts

Table C3: **Papers included in the analysis.**

Conference	Authors	Title
NeurIPS	Tonghan Wang, Paul Dütting, Dmitry Ivanov, Inbal Talgam-Cohen, David C. Parkes	Deep Contract Design via Discontinuous Networks
NeurIPS	Sourya Basu, Pulkit Katdare, Prasanna Sattigeri, Vijil Chenthamarakshan, Katherine Driggs-Campbell, Payel Das, Lav R. Varshney	Efficient Equivariant Transfer Learning from Pretrained Models
NeurIPS	Qianyi Li, Haim Sompolinsky	Globally Gated Deep Linear Networks
NeurIPS	Jonatha Anselmi, Bruno Gaujal, Louis-Sébastien Rebuffi	Reinforcement Learning in a Birth and Death Process: Breaking the Dependence on the State Space
NeurIPS	Zhanpeng Zhou, Yongyi Yang, Xiaojiang Yang, Junchi Yan, Wei Hu	Going Beyond Linear Mode Connectivity: The Layerwise Linear Feature Connectivity
NeurIPS	Jinyu Cai, Jicong Fan	Perturbation Learning Based Anomaly Detection
NeurIPS	Dan Zhao	Combining Explicit and Implicit Regularization for Efficient Learning in Deep Networks
NeurIPS	Leonard Papenmeier, Luigi Nardi, Matthias Poloczek	Increasing the Scope as You Learn: Adaptive Bayesian Optimization in Nested Subspaces
NeurIPS	Yang Song, Qiyu Kang, Sijie Wang, Zhao Kai, Wee Peng Tay	On the Robustness of Graph Neural Diffusion to Topology Perturbations
NeurIPS	Ibrahim Alabdulmohsin, Behnam Neyshabur, Xiaohua Zhai	Revisiting Neural Scaling Laws in Language and Vision
NeurIPS	Salva Rühling Cachay, Bo Zhao, Hailey Joren, Rose Yu	DYffusion: A Dynamics-informed Diffusion Model for Spatiotemporal Forecasting
NeurIPS	Indradyumna Roy, Soumen Chakrabarti, Abir De	Maximum Common Subgraph Guided Graph Retrieval: Late and Early Interaction Networks
NeurIPS	Divin Yan, Gengchen Wei, Chen Yang, Shengzhong Zhang, Zengfeng Huang	Rethinking Semi-Supervised Imbalanced Node Classification from Bias-Variance Decomposition
NeurIPS	Zhiying Lu, Hongtao Xie, Chuanbin Liu, Yongdong Zhang	Bridging the Gap Between Vision Transformers and Convolutional Neural Networks on Small Datasets
NeurIPS	Rémi Leluc, François Portier, Johan Segers, Aigerim Zhuman	A Quadrature Rule combining Control Variates and Adaptive Importance Sampling
NeurIPS	Kwangjun Ahn, Xiang Cheng, Hadi Daneshmand, Suvrit Sra	Transformers learn to implement preconditioned gradient descent for in-context learning
NeurIPS	Annie S. Chen, Archit Sharma, Sergey Levine, Chelsea Finn	You Only Live Once: Single-Life Reinforcement Learning
NeurIPS	Sen Lin, Daouda Sow, Kaiyi Ji, Yingbin Liang, Ness Shroff	Non-Convex Bilevel Optimization with Time-Varying Objective Functions
NeurIPS	Carl Hvarfner, Erik Hellsten, Frank Hutter, Luigi Nardi	Self-Correcting Bayesian Optimization through Bayesian Active Learning
NeurIPS	Abir De, Soumen Chakrabarti	Neural Estimation of Submodular Functions with Applications to Differentiable Subset Selection
NeurIPS	Ximing Lu, Sean Welleck, Jack Hessel, Liwei Jiang, Lianhui Qin, Peter West, Prithviraj Ammanabrolu, Yejin Choi	Quark: Controllable Text Generation with Reinforced Unlearning

Table C3: **Papers included in the analysis.**

Conference	Authors	Title
NeurIPS	Weirui Ye, Pieter Abbeel, Yang Gao	Spending Thinking Time Wisely: Accelerating MCTS with Virtual Expansions
NeurIPS	Axel Levy, Gordon Wetzstein, Julien Martel, Frederic Poitevin, Ellen D. Zhong	Amortized Inference for Heterogeneous Reconstruction in Cryo-EM
ICLR	Tim Pearce, Tabish Rashid, Anssi Kanervisto, Dave Bignell, Mingfei Sun, Raluca Georgescu, Sergio Valcarcel Macua, Shan Zheng Tan, Ida Momennejad, Katja Hofmann, Sam Devlin	Imitating Human Behaviour with Diffusion Models
ICLR	Yi Ren, Shangmin Guo, Wonho Bae, Danica J. Sutherland	How to prepare your task head for finetuning
ICLR	Kieran A. Murphy, Dani S. Bassett	Interpretability with full complexity by constraining feature information
ICLR	Julius Adebayo, Michael Muelly, Hal Abelson, Been Kim	Post hoc Explanations may be Ineffective for Detecting Unknown Spurious Correlation
ICLR	Roman Levin, Valeriia Cherepanova, Avi Schwarzschild, Arpit Bansal, C. Bayan Bruss, Tom Goldstein, Andrew Gordon Wilson, Micah Goldblum	Transfer Learning with Deep Tabular Models
ICLR	Aviv A. Rosenberg, Sanketh Vedula, Yaniv Romano, Alex M. Bronstein	Fast Nonlinear Vector Quantile Regression
ICLR	Edward De Brouwer, Rahul G. Krishnan	Anamnesic Neural Differential Equations with Orthogonal Polynomial Projections
ICLR	Ilya Trofimov, Daniil Cherniavskii, Eduard Tulchinskii, Nikita Balabin, Evgeny Burnaev, Serguei Barannikov	Learning Topology-Preserving Data Representations
ICLR	Trenton Bricken, Xander Davies, Deepak Singh, Dmitry Krotov, Gabriel Kreiman	Sparse Distributed Memory is a Continual Learner
ICLR	Steeven Janny, Aurélien Béneteau, Madiha Nadri, Julie Digne, Nicolas Thome, Christian Wolf	Eagle: Large-Scale Learning of Turbulent Fluid Dynamics with Mesh Transformers
ICLR	Clement Vignac, Igor Krawczuk, Antoine Siraudin, Bohan Wang, Volkan Cevher, Pascal Frossard	DiGress: Discrete Denoising diffusion for graph generation
ICLR	Xinting Hu, Yulei Niu, Chunyan Miao, Xian-Sheng Hua, Hanwang Zhang	On Non-Random Missing Labels in Semi-Supervised Learning
ICLR	Jiefeng Chen, Timothy Nguyen, Dilan Gorur, Arslan Chaudhry	Is forgetting less a good inductive bias for forward transfer?
ICLR	Matthew J. Tilley, Michelle Miller, David J. Freedman	Artificial Neuronal Ensembles with Learned Context Dependent Gating
ICLR	Zhang-Wei Hong, Tao Chen, Yen-Chen Lin, Joni Pajarinen, Pulkit Agrawal	Topological Experience Replay
ICLR	Lingkai Kong, Yuqing Wang, Molei Tao	Momentum Stiefel Optimizer, with Applications to Suitably-Orthogonal Attention, and Optimal Transport
ICLR	Zhang-Wei Hong, Pulkit Agrawal, Rémi Tachet des Combes, Romain Laroche	Harnessing Mixed Offline Reinforcement Learning Datasets via Trajectory Weighting

Table C3: **Papers included in the analysis.**

Conference	Authors	Title
ICLR	Meng Cao, Mehdi Fatemi, Jackie Chi Kit Cheung, Samira Shabanian	Systematic Rectification of Language Models via Dead-end Analysis
ICLR	Wentao Zhang, Yexin Wang, Zhenbang You, Meng Cao, Ping Huang, Jiulong Shan, Zhi Yang, Bin Cui	Information Gain Propagation: a new way to Graph Active Learning with Soft Labels
ICLR	Alexandre Perez-Lebel, Marine Le Morvan, Gaël Varoquaux	Beyond calibration: estimating the grouping loss of modern neural networks
ICLR	Jianwen Xie, Yaxuan Zhu, Jun Li, Ping Li	A Tale of Two Flows: Cooperative Learning of Langevin Flow and Normalizing Flow Toward Energy-Based Model
ICLR	Amrith Setlur, Don Dennis, Benjamin Eysenbach, Aditi Raghunathan, Chelsea Finn, Virginia Smith, Sergey Levine	Bitrate-Constrained DRO: Beyond Worst Case Robustness To Unknown Group Shifts
ICLR	Tim Z. Xiao, Robert Bamler	Trading Information between Latents in Hierarchical Variational Autoencoders
ICLR	Zixuan Ke, Yijia Shao, Haowei Lin, Tatsuya Konishi, Gyuhak Kim, Bing Liu	Continual Pre-training of Language Models
ICLR	Zhang-Wei Hong, Ge Yang, Pulkit Agrawal	Bilinear value networks
ICLR	Hanrong Ye, Dan Xu	Joint 2D-3D Multi-Task Learning on Cityscapes-3D: 3D Detection, Segmentation, and Depth Estimation
ICLR	Mohit Vaishnav, Thomas Serre	GAMR: A Guided Attention Model for (visual) Reasoning
ICLR	Noam Levi, Itay M. Bloch, Marat Freytsis, Tomer Volansky	Noise Injection Node Regularization for Robust Learning
ICLR	Paul F. Jaeger, Carsten T. Lüth, Lukas Klein, Till J. Bungert	A Call to Reflect on Evaluation Practices for Failure Detection in Image Classification
ICLR	Thomas M. Sutter, Laura Manduchi, Alain Ryser, Julia E. Vogt	Learning Group Importance using the Differentiable Hypergeometric Distribution
ICLR	AmirEhsan Khorashadizadeh, Anadi Chaman, Valentin Debarnot, Ivan Dokmanić	FunkNN: Neural Interpolation for Functional Generation
ICML	Ramki Gummadi, Saurabh Kumar, Junfeng Wen, Dale Schuurmans	A Parametric Class of Approximate Gradient Updates for Policy Optimization
ICML	Joshua P. Zitovsky, Daniel de Marchi, Rishabh Agarwal, Michael R. Kosorok	Revisiting Bellman Errors for Offline Model Selection
ICML	Jiayin Jin, Zeru Zhang, Yang Zhou, Lingfei Wu	Input-agnostic Certified Group Fairness via Gaussian Parameter Smoothing
ICML	Ching-Yao Chuang, Stefanie Jegelka, David Alvarez-Melis	InfoOT: Information Maximizing Optimal Transport
ICML	Ilgee Hong, Sen Na, Michael W. Mahoney, Mladen Kolar	Constrained Optimization via Exact Augmented Lagrangian and Randomized Iterative Sketching
ICML	Matthew Fahrbach, Adel Javanmard, Vahab Mirrokni, Pratik Worah	Learning Rate Schedules in the Presence of Distribution Shift
ICML	Samuele Marro, Michele Lombardi	Computational Asymmetries in Robust Classification

Table C3: **Papers included in the analysis.**

Conference	Authors	Title
ICML	Nicolas Chopin, Andras Fulop, Jeremy Heng, Alexandre H. Thiery	Computational Doob's h-transforms for Online Filtering of Discretely Observed Diffusions
ICML	Wentao Zhang, Zeang Sheng, Mingyu Yang, Yang Li, Yu Shen, Zhi Yang, Bin Cui	NAFS: A Simple yet Tough-to-beat Baseline for Graph Representation Learning
ICML	Disha Shrivastava, Hugo Larochelle, Daniel Tarlow	Repository-Level Prompt Generation for Large Language Models of Code
ICML	Chenlu Ye, Wei Xiong, Quanquan Gu, Tong Zhang	Corruption-Robust Algorithms with Uncertainty Weighting for Nonlinear Contextual Bandits and Markov Decision Processes
ICML	Anas Barakat, Ilyas Fatkhullin, Niao He	Reinforcement Learning with General Utilities: Simpler Variance Reduction and Large State-Action Space
ICML	Alberto Maria Metelli, Francesco Trovò, Matteo Pirola, Marcello Restelli	Stochastic Rising Bandits
ICML	Idan Shenfeld, Zhang-Wei Hong, Aviv Tamar, Pulkit Agrawal	TGRL: An Algorithm for Teacher Guided Reinforcement Learning
ICML	Benjamin Dupuis, George Deligiannidis, Umut Şimşekli	Generalization Bounds with Data-dependent Fractal Dimensions
ICML	Wanrong Zhang, Ruqi Zhang	DP-Fast MH: Private, Fast, and Accurate Metropolis-Hastings for Large-Scale Bayesian Inference
ICML	Zixuan Ni, Longhui Wei, Siliang Tang, Yueting Zhuang, Qi Tian	Continual Vision-Language Representation Learning with Off-Diagonal Information
ICML	Manuel Nonnenmacher, Lukas Oldenburg, Ingo Steinwart, David Reeb	Utilizing Expert Features for Contrastive Learning of Time-Series Representations
ICML	Shih-Yang Liu, Zechun Liu, Kwang-Ting Cheng	Oscillation-free Quantization for Low-bit Vision Transformers
ICML	Siqi Liu, Marc Lanctot, Luke Marris, Nicolas Heess	Simplex Neural Population Learning: Any-Mixture Bayes-Optimality in Symmetric Zero-sum Games
ICML	Guanghui Qin, Benjamin Van Durme	Nugget: Neural Agglomerative Embeddings of Text
ICML	Marc Härkönen, Markus Lange-Hegermann, Bogdan Raiţă	Gaussian Process Priors for Systems of Linear Partial Differential Equations with Constant Coefficients
ICML	Xiyao Wang, Wichayaporn Wongkamjan, Furong Huang	Live in the Moment: Learning Dynamics Model Adapted to Evolving Policy
ICML	Tanvir Islam, Peter Washington	Personalized Prediction of Recurrent Stress Events Using Self-Supervised Learning on Multimodal Time-Series Data
ICML	Krishna Pillutla, Kshitiz Malik, Abdelrahman Mohamed, Michael Rabbat, Maziar Sanjabi, Lin Xiao	Federated Learning with Partial Model Personalization
ICML	Jaesik Yoon, Yi-Fu Wu, Heechul Bae, Sungjin Ahn	An Investigation into Pre-Training Object-Centric Representations for Reinforcement Learning
ICML	Mehrdad Ghadiri, Matthew Fahrbach, Gang Fu, Vahab Mirrokni	Approximately Optimal Core Shapes for Tensor Decompositions

Table C3: **Papers included in the analysis.**

Conference	Authors	Title
ICML	Arpit Bansal, Ping-yeh Chiang, Michael Curry, Rajiv Jain, Curtis Wigington, Varun Manjunatha, John P Dickerson, Tom Goldstein	Certified Neural Network Watermarks with Randomized Smoothing
ICML	Mohamad Amin Mohamadi, Wonho Bae, Danica J. Sutherland	A Fast, Well-Founded Approximation to the Empirical Neural Tangent Kernel
ICML	Chuyang Ke, Jean Honorio	Exact Inference in High-order Structured Prediction
ICML	Wentao Zhang, Zheyu Lin, Yu Shen, Yang Li, Zhi Yang, Bin Cui	DFG-NAS: Deep and Flexible Graph Neural Architecture Search
ICML	Tongzhou Wang, Antonio Torralba, Phillip Isola, Amy Zhang	Optimal Goal-Reaching Reinforcement Learning via Quasimetric Learning
ICML	Yi-Fan Zhang, Xue Wang, Kexin Jin, Kun Yuan, Zhang Zhang, Liang Wang, Rong Jin, Tieniu Tan	AdaNPC: Exploring Non-Parametric Classifier for Test-Time Adaptation
ICML	Litian Liang, Yaosheng Xu, Stephen McAleer, Dailin Hu, Alexander Ihler, Pieter Abbeel, Roy Fox	Reducing Variance in Temporal-Difference Value Estimation via Ensemble of Deep Networks
ICML	Mohammed Nowaz Rabbani Chowdhury, Shuai Zhang, Meng Wang, Sijia Liu, Pin-Yu Chen	Patch-level Routing in Mixture-of-Experts is Provably Sample-efficient for Convolutional Neural Networks
ICML	Gal Leibovich, Guy Jacob, Or Avner, Gal Novik, Aviv Tamar	Learning Control by Iterative Inversion
ICML	Jiayin Jin, Jiaxiang Ren, Yang Zhou, Lingjuan Lyu, Ji Liu, Dejing Dou	Accelerated Federated Learning with Decoupled Adaptive Optimization
ICML	Krzysztof Choromanski, Arijit Sehanobish, Han Lin, Yunfan Zhao, Eli Berger, Tetiana Parshakova, Alvin Pan, David Watkins, Tianyi Zhang, Valerii Likhosherstov, Somnath Basu Roy Chowdhury, Avinava Dubey, Deepali Jain, Tamas Sarlos, Snigdha Chaturvedi, Adrian Weller	Efficient Graph Field Integrators Meet Point Clouds
ICML	Simone Parisi, Aravind Rajeswaran, Senthil Purushwalkam, Abhinav Gupta	The Unsurprising Effectiveness of Pre-Trained Vision Models for Control

Table C3: **Papers included in the analysis.**

Conference	Paper Title
AAAI	Neuro-symbolic Rule Learning in Real-world Classification Tasks Generalization Bounds for Inductive Matrix Completion in Low-noise Settings
NeurIPS	A General Framework for Robust G-Invariance in G-Equivariant Networks CLIFT: Analysing Natural Distribution Shift on Question Answering Models in Clinical Domain Partial Counterfactual Identification of Continuous Outcomes with a Curvature Sensitivity Model Attacks on Online Learners: a Teacher-Student Analysis Learning Feynman Diagrams using Graph Neural Networks Function Classes for Identifiable Nonlinear Independent Component Analysis Blackbox Attacks via Surrogate Ensemble Search Censored Quantile Regression Neural Networks for Distribution-Free Survival Analysis Semi-Discrete Normalizing Flows through Differentiable Tessellation Online Decision Mediation Exact Generalization Guarantees for (Regularized) Wasserstein Distributionally Robust Models Deep Learning with Kernels through RKHM and the Perron-Frobenius Operator Bridging RL Theory and Practice with the Effective Horizon Reliable learning in challenging environments
ICLR	Brain-like representational straightening of natural movies in robust feedforward neural networks Broken Neural Scaling Laws Parametrizing Product Shape Manifolds by Composite Networks Tier Balancing: Towards Dynamic Fairness over Underlying Causal Factors Guiding continuous operator learning through Physics-based boundary constraints Scaling Laws For Deep Learning Based Image Reconstruction Probabilistically Robust Recourse: Navigating the Trade-offs between Costs and Robustness in Algorithmic Recourse Domain Adaptation via Minimax Entropy for Real/Bogus Classification of Astronomical Alerts
ICML	Why Target Networks Stabilise Temporal Difference Methods Nonlinear Advantage: Trained Networks Might Not Be As Complex as You Think HyperImpute: Generalized Iterative Imputation with Automatic Model Selection

Table C4: **Papers excluded from the analysis.**

Note: The papers listed are excluded from the analysis due to tex compilation errors, such as bibtex errors.