4 Versions

All That Glitters is Not Novel: Plagiarism in AI Generated Research

(/pdf?id=hC3ZW56NUo)

Tarun Gupta (/profile?id=~Tarun_Gupta4), Danish Pruthi (/profile?id=~Danish_Pruthi1)

15 Feb 2025 (modified: 15 Jan 2026) ACL ARR 2025 February Submission
February, Senior Area Chairs, Area Chairs, Reviewers, Authors
Revisions (/revisions?id=hC3ZW56NUo) BibTeX

Abstract:
Automating scientific research is considered the final frontier of science. Recently, several papers claim autonomous research agents can generate novel research ideas. Amidst the prevailing optimism, we document a critical concern: a considerable fraction of such research documents are smartly plagiarized. Unlike past efforts where experts evaluate the novelty and feasibility of research ideas, we request 13 experts to operate under a different situational logic: to identify similarities between LLM-generated research documents and existing work. Concerningly, the experts identify 24% of the 50 evaluated research documents to be either directly copied (with one-to-one methodological mapping), or significantly borrowed from existing work. These reported instances are cross-verified by authors of the source papers. Problematically, these LLM-generated research documents do not acknowledge original sources, and bypass inbuilt plagiarism detectors. Lastly, through controlled experiments we show that automated plagiarism detectors are inadequate at catching deliberately plagiarized ideas from an LLM. We recommend a careful assessment of LLM-generated research, and discuss the implications of our findings on research and academic publishing.

Paper Type: Long
Research Area: Resources and Evaluation
Research Area Keywords: evaluation methodologies, evaluation, corpus creation, benchmarking, human-centered evaluation
Contribution Types: NLP engineering experiment, Data analysis
Languages Studied: English
Reviewing No Volunteers Reason: All qualified authors are already involved in the reviewing process in some capacity (as Area Chairs, as Senior Area Chairs, etc.).
Reviewing Volunteers For Emergency Reviewing: N/A, no volunteers were provided in the previous question.
TLDR: Amidst prevailing optimism about the novelty of AI-generated research, our study finds that a considerable fraction of it is smartly plagiarized, bypassing inbuilt plagiarism checks and unsuspecting expert reviewers.
Reassignment Request Area Chair: This is not a resubmission
Reassignment Request Reviewers: This is not a resubmission
Preprint: yes
Preprint Status: We plan to release a non-anonymous preprint in the next two months (i.e., during the reviewing process).
Preferred Venue: ACL
Consent To Share Data: no
Consent To Share Submission Details: On behalf of all authors, we agree to the terms above to share our submission details.
A1 Limitations Section: This paper has a limitations section.
A2 Potential Risks: N/A
B Use Or Create Scientific Artifacts: Yes
B1 Cite Creators Of Artifacts: Yes
B1 Elaboration: Section 3, Section 6
B2 Discuss The License For Artifacts: N/A

B3 Artifact Use Consistent With Intended Use: N/A
B4 Data Contains Personally Identifying Info Or Offensive Content: N/A
B5 Documentation Of Artifacts: N/A
B6 Statistics For Data: Yes
B6 Elaboration: Section 4, Section 6
C Computational Experiments: Yes
C1 Model Size And Budget: N/A
C2 Experimental Setup And Hyperparameters: Yes
C2 Elaboration: Section 4
C3 Descriptive Statistics: N/A
C4 Parameters For Packages: N/A
D Human Subjects Including Annotators: Yes
D1 Instructions Given To Participants: Yes
D1 Elaboration: Appendix
D2 Recruitment And Payment: Yes
D2 Elaboration: Section 4
D3 Data Consent: N/A
D4 Ethics Review Board Approval: N/A
D5 Characteristics Of Annotators: N/A
E Ai Assistants In Research Or Writing: Yes
E1 Information About Use Of Ai Assistants: N/A
E1 Elaboration: In line with the ACL policies, AI assistance was merely used for minor paraphrasing of minor portions of the paper and for simple code refactoring.
Author Submission Checklist: yes
Association For Computational Linguistics - Blind Submission License Agreement: On behalf of all authors, I do not agree
Submission Number: 3418

Discussion (?id=hC3ZW56NUo#discussion)

Filter by reply type... | Filter by author... | Search keywords... | Sort: Newest First

Everyone | Submission3418... | Submission3418 Area... | Submission3418 Authors | Submission3418... | Program Chairs | Submission3418... | Submission3418... | Submission3418... | Submission3418... | Submission3418... | Submission3418...

16 / 16 replies shown

Meta Review of Submission3418 by Area Chair Z7J6

Meta Review by Area Chair Z7J6 09 Apr 2025, 16:05 (modified: 25 Apr 2025, 00:53)
Senior Area Chairs, Area Chairs, Authors, Reviewers Submitted, Program Chairs, Commitment Readers
Revisions (/revisions?id=5Oi1f8L3xt)

Metareview:
The paper covers the topic of plagiarism detection in AI generated research. The authors performed an annotation with 13 experts who assessed 50 research documents produced by AI, uncovering that between 24% (confirmed) and 36% (including suspected instances) contained substantial plagiarism. The work also highlights the limitations of current automated plagiarism detection tools in identifying these issues effectively.

Summary Of Reasons To Publish:
- The paper covers an interesting and important topic of AI generated plagiarism;
- A novel annotation for the setup topic is performed;
- The authors examine the LLM-detectors of plagiarism test dividing automated plagiarism detection into two stages and finds that locating similar source documents is the main bottleneck in the process.

Summary Of Suggested Revisions:

- The authors themselves together with reviewers acknowledge the limited annotation setup (at least, in terms of the amount of annotated articles);
- Thus, the annotation with a high probability includes bias;
- The experimental scope is a bit limited and can be extended to more models (ie, some opensource instances).
- The discussion on the results is also limited and can be extended diving into models performance details and error analysis.

Overall Assessment: 3 = Findings: I think this paper could be accepted to the Findings of the ACL.
Suggested Venues: *ACL Findings
Reported Issues: Yes, and I took them into account in my meta-review

Request regarding Reviewer qbHF's lack of response

Author-Editor Confidential Comment

by Authors (Danish Pruthi (/profile?id=~Danish_Pruthi1), Tarun Gupta (/profile?id=~Tarun_Gupta4))

05 Apr 2025, 20:35 Program Chairs, Senior Area Chairs, Area Chairs, Authors

Comment:
Dear Area Chair, Senior Area Chairs and other PC members,

We are writing to request your assistance in engaging with Reviewer qbHF. We believe that we have addressed their concerns by conducting two additional experiments they specifically requested:

1. Expanding beyond title-only analysis to include title+abstract comparisons across multiple LLMs (under "Title-only analysis and single-LLM in PCA projection study" in our general response), and
2. Adding a control group analysis using human-authored papers from the PeerRead dataset (under "Lack of control group" as a part of general response).

Despite these additional experiments and our detailed response to their other methodological questions, Reviewer qbHF has not acknowledged our response. We kindly request you to either remind the reviewer to read and assess our response or discount their concerns as we've addressed those through additional experiments.

General Response

Official Comment

by Authors (Danish Pruthi (/profile?id=~Danish_Pruthi1), Tarun Gupta (/profile?id=~Tarun_Gupta4))

31 Mar 2025, 15:52 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers, Reviewers Submitted, Authors, Commitment Readers

Revisions (/revisions?id=b5DtJk7VzK)

Comment:
We thank all reviewers for their thoughtful and detailed reviews. We appreciate that reviewers (a) recognize the value of our work in identifying a critical issue of plagiarism in AI-generated research (qbHF, pV7f, TvaC), (b) appreciate the thoroughness of our study design with expert evaluation and author verification (qbHF, cnQh), and (c) acknowledge the timeliness and importance of this topic (TvaC, cnQh).

Re: Confirmation bias

Reviewers qbHF and pV7f express concern that instructing experts to actively look for plagiarism may induce confirmation bias (which we also discuss in the Limitations section). However, we argue that this change in our instruction design is critical and key to our study. Perhaps, an analogy could be drawn to finding bugs in existing code. Deliberately looking for mistakes may result in annotators catching more bugs, which can later be objectively confirmed, therefore a presumption (about the existence of bugs in code) might offer a better estimate about the number of bugs in the code.

Similarly, by instructing experts to actively search for plagiarism, we create a different evaluation framework than prior work where experts evaluated LLM-generated research documents presuming no deliberate plagiarism. Without this change in perspective, even human experts miss potentially plagiarized content, as we discovered

when analyzing prior work (Si et al., 2024; Lu et al., 2024a). Once the plagiarism instances are flagged, we confirm those claims with the authors of source papers, which provide objectivity to our analysis. Moreover, we make our data available in an anonymous GitHub repository for verification by the reviewers and public.

Re: Title-only analysis and single-LLM in PCA projection study.

Reviewers qbHF and cnQh correctly note that our diversity analysis of LLM-generated research documents relies on a single LLM and only examined titles. We chose Claude 3.5 Sonnet because it generated the highest quality research content in prior work (Si et al., 2024; Lu et al., 2024a).

To address this concern, **we conduct new experiments** comparing concatenated titles and abstracts of LLM-generated papers and human-written papers, using two LLMs: Claude 3.5 Sonnet and GPT-4o (the two most commonly used LLMs in prior work). Similar to the title-only analysis, these new experiments suggest that LLM-generated research content remains easily distinguishable and less semantically diverse than human research.

LLM Model	Classification Accuracy (%) for Distinguishing LLM vs Human Content (title + abstract)	Diversity Measure: Cluster Spread Ratio (LLM/Human) (title + abstract)
Claude 3.5 Sonnet	98.6%	0.81
GPT-4o	97.6%	0.72

Re: Lack of control group

Reviewer qbHF raises a valid concern that our study lacked a control group of human-written papers for comparison of plagiarism rates. Given the labor-intensive nature of finding plagiarism in LLM-generated papers, we focused our resources on analyzing LLM-generated content. However, we agree that a control group is valuable, and as a quick proxy, we analyzed existing peer reviews from the PeerRead dataset for (accepted + rejected) human-written papers from NeurIPS 2017, ACL 2017, ICLR 2017, and CoNLL 2016. We use Claude 3.7 Sonnet to extract potential concerns of plagiarism from peer reviews. We emphasize that the **LLM is used only to extract information about reviewer claims, not to make judgments**. As shown below, the plagiarism rate in human-authored papers is significantly lower than the 24% we found in LLM-generated papers:

Conference	# Papers	Score 4 (%)	Score 5 (%)	Plagiarism rate (scores 4+) (%)
ACL 2017	123	0.8%	0%	0.8%
ICLR 2017	349	4.0%	2.3%	6.3%
ConLL 2016	19	5.3% (1/19)	0%	5.3%
NeurIPS 2017	499	1.8%	0%	1.8%

Furthermore, even if LLM plagiarism rates were comparable to human rates (the above experiments suggest otherwise), our findings are still critical as LLMs drastically lower the cost of plagiarizing content, which may result in increased submissions, and may overwhelm the existing review processes.

Methodological clarification for human study

During initial exploratory research, the first author assessed 10 exemplars from (Si et al., 2024; Lu et al., 2024a) and found substantial overlap with prior work in 6 of them, scoring them 4+ on the plagiarism rubric. We cross-checked all 6 of these claims with the source paper authors and adjusted scores accordingly. These adjusted scores are included in the 24% plagiarism rate. The first author is therefore considered one of the 13 experts in our human study.

Official Review of Submission3418 by Reviewer qbHF

Official Review by Reviewer qbHF 28 Mar 2025, 09:32 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer qbHF, Commitment Readers

Revisions (/revisions?id=eu6odsY9TT)

Paper Summary:

This paper investigates plagiarism in AI-generated research documents through expert evaluation and automated detection methods. The authors recruited 13 experts to evaluate 50 AI-generated research documents, finding that 24% (verified) to 36% (including unverified claims) contained significant plagiarism. The study also reveals that current automated plagiarism detection tools are inadequate at catching such cases. Additionally, the paper shows that AI-generated research ideas tend to be less diverse and follow more predictable patterns compared to human-written work.

Summary Of Strengths:

1. The paper identifies a critical issue in AI-generated research by systematically documenting plagiarism issues through expert evaluation and author verification.
2. The study design is thorough, using multiple approaches:

- Expert-led evaluation with clear scoring rubrics
- Verification from original paper authors
- Systematic evaluation of automated plagiarism detection tools
- Detailed case studies demonstrating sophisticated plagiarism

3. The paper provides comprehensive analysis of limitations in current plagiarism detection systems, testing multiple methods including SSAG, OpenScholar, and Turnitin.

Summary Of Weaknesses:

1. Lack of control group: The study doesn't include human-written papers in the expert evaluation process. This makes it impossible to establish whether the observed plagiarism rate is significantly different from human-written papers.
2. Limited scope of research agents:

- Heavy reliance on one research agent (40 out of 50 papers)
- The RAG-based methodology of the studied research agent may inherently lead to similarity with retrieved papers
- Results may not generalize to other research agents with different architectures

3. Potential methodological biases:

- The instruction to "presume plagiarism" may bias experts toward finding similarities
- The experts' ability to select which proposals to review could introduce selection bias

4. Limited analysis of diversity:

- PCA projection study uses only one LLM backbone
- Comparison may not be fair as human papers come from multiple authors
- Title-only analysis may not fully capture research diversity

Comments Suggestions And Typos:

Below are some suggestions:

1. Include human-written papers as a control group in the expert evaluation process
2. Expand the study to include more research agents and LLM architectures
3. Consider alternative metrics for measuring research diversity beyond title analysis
4. Design follow-up studies with more neutral expert instructions to reduce potential confirmation bias
5. Develop more sophisticated automated detection methods specifically designed for AI-generated content

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.

Soundness: 2.5

Excitement: 4.0 = Exciting: I would mention this paper to others and/or make an effort to attend its presentation in a conference.

Overall Assessment: 2 = Resubmit next cycle: I think this paper needs substantial revisions that can be completed by the next ARR cycle.

Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No

Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.

Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Response to Reviewer qbHF

Official Comment

by Authors (Danish Pruthi (/profile?id=~Danish_Pruthi1), Tarun Gupta (/profile?id=~Tarun_Gupta4))

31 Mar 2025, 15:55 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer qbHF, Commitment Readers

Revisions (/revisions?id=R67dvZh4ru)

Comment:
Thank you for your insightful comments! We address your concerns below:

Re: "Lack of control group"

Thanks for raising this valid concern! We conduct additional experiments to address this concern. Please see our general response on "Lack of control group".

Re: "Heavy reliance on one research agent (40 out of 50 papers)" "Results may not generalize to other research agents with different architectures"

The manual, labor-intensive nature of finding plagiarism in LLM-generated research and our limited expert resources allowed us to evaluate only one research agent. We chose Si et al.'s (2024) research agent for several reasons (discussed in detail in lines 276-293 of our paper): First, their method represents the kinds of prompt engineering approaches that are commonly used in prior work. Second, they perform the most comprehensive evaluation to date, finding that humans judge LLM-generated papers as more novel than human-written ones. Third, their minimal prompt engineering allows better assessment of LLMs' raw capabilities.

Re: "The RAG-based methodology of the studied research agent may inherently lead to similarity with retrieved papers"

Nearly all past efforts use RAG to find relevant literature, which is important for providing LLMs access to the latest developments. In the studied research agent (Si et al., 2024), the LLM is explicitly instructed not to copy or build upon retrieved papers. Further, claims about novelty of LLM-generated research are made in this context, therefore studying such agents felt appropriate. Focus of our work is on evaluating existing research agents rather than creating better ones.

Re: "The instruction to "presume plagiarism" may bias experts toward finding similarities"

Please see the general response related to confirmation biases.

Re: "The experts' ability to select which proposals to review could introduce selection bias"

We agree with this concern and acknowledge it in our Limitations section. Experts evaluate 3 out of 5 papers that are most aligned with their expertise. This approach was necessary due to the publication explosion in NLP and the many specialized subfields. Even with this flexibility, experts often spent over an hour reviewing a single LLM-generated paper.

Re: "PCA projection study uses only one LLM backbone" + "Title-only analysis may not fully capture research diversity"

We conduct additional experiments addressing this concern. Please see our general response on "Title-only analysis and single-LLM in PCA projection study".

Re: "Comparison may not be fair as human papers come from multiple authors"

In prior work, a single LLM's output is compared against papers from multiple authors. We control for topics by comparing LLM-generated and human-written papers on the same topics. Further controlling for authors would severely reduce our sample size and limit the scope of our analysis. Additionally, even individual researchers typically write papers with different sets of co-authors across their careers, making it impractical to isolate a single author's distinct research contribution for comparison with LLM outputs.

Re: "Develop more sophisticated automated detection methods specifically designed for AI-generated content"

Developing better automated detection methods is an important direction for future work but beyond the scope of our evaluation.

Follow up to Reviewer qbHF

Official Comment

by Authors (Danish Pruthi (/profile?id=~Danish_Pruthi1), Tarun Gupta (/profile?id=~Tarun_Gupta4))

03 Apr 2025, 15:21 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer qbHF, Commitment Readers

Revisions (/revisions?id=gcYcNUjR59)

Comment:

We truly appreciate the time and effort you have dedicated to reviewing our submission. We have replied to your questions and conducted additional experiments addressing methodology concerns highlighted as a weakness. Additionally, we have incorporated feedback from other reviewers, which may also help clarify any additional questions you might have.

Since the discussion phase ends soon, we wanted to follow up and would greatly value any further thoughts or concerns you might have so we can address them appropriately.

Thank you again for your time and commitment to the review process.

Official Review of Submission3418 by Reviewer pV7f

Official Review by Reviewer pV7f 24 Mar 2025, 21:11 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer pV7f, Commitment Readers

Revisions (/revisions?id=kXz36Q999c)

Paper Summary:

The paper presents a study of plagiarism in AI-generated research. It builds on top of (Si et al., 2024), where human experts found AI-generated research proposals on average more novel than human-generated ones. The current paper recruits human experts to thoroughly review 50 more research proposals (36 generated within this work using a search-augmented AI agent, 3 per 12 research topics, and 14 more proposals from previous work). The participants are instructed to actively look for non-cited overlaps with existing work, and they find them: out of the 50 proposals, 18 were found to borrow the proposed methods from a few existing works. A lot of overlaps appeared to be carefully masked, though. To evaluate existing automatic plagiarism detectors, the authors used a GPT-4o-based agent to generate 480 proposals with even more flagrant plagiarism, basing each proposal on a single paper. They used two LLMs (GPT-4 and Claude) to detect pagiarism in this dataset (with and without using an oracle knowledge of the source paper or the Semantic Scholar search system), as well as two dedicated search systems (OpenScholar and Turnitin). None of them reached sufficient accuracy without the oracle knowledge. Finally, the authors compare the distribution of titles of AI- and human-generated research proposals, and find

that AI-generated titles are less diverse and rather recognizeable. Overall, the paper warns that AI research agents tend to spontaneously plagiarize (and hide it well), and the existing plagiarism detection systems are inadequate against it.

Summary Of Strengths:

- The paper addresses a critical topic of AI-generated research and the potential contamination it brings into the academic world, and highlights two critical issue with it (presence of spontaneous plagiarism and its non-discoverability).
- The main finding is based on evaluation by human experts, which is supposed to be highly reliable.
- The paper demonstrates a successful attack on existing plagiarism detectors, which raises a call for developing more accurate ones.

Summary Of Weaknesses:

- As the authors themselves acknowledge, there results of human evaluation might contain confirmation bias. Inclusion of some proposals that are known to be original would strengthen the results.

Comments Suggestions And Typos:
No useful suggestions

Confidence: 2 = Willing to defend my evaluation, but it is fairly likely that I missed some details, didn't understand some central points, or can't be sure about the novelty of the work.
Soundness: 3 = Acceptable: This study provides sufficient support for its main claims. Some minor points may need extra support or details.
Excitement: 3 = Interesting: I might mention some points of this paper to others and/or attend its presentation in a conference if there's time.
Overall Assessment: 3 = Findings: I think this paper could be accepted to the Findings of the ACL.
Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 2 = Documentary: The new datasets will be useful to study or replicate the reported research, although for other purposes they may have limited interest or limited usability. (Still a positive rating)
Software: 2 = Documentary: The new software will be useful to study or replicate the reported research, although for other purposes it may have limited interest or limited usability. (Still a positive rating)
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Official Comment by Authors

Official Comment

by Authors (Danish Pruthi (/profile?id=~Danish_Pruthi1), Tarun Gupta (/profile?id=~Tarun_Gupta4))

03 Apr 2025, 15:25 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer pV7f, Commitment Readers

Revisions (/revisions?id=ujst4lfb4l)

Comment:
We truly appreciate the time and effort you have dedicated to reviewing our submission. We have replied to your questions. Additionally, we have incorporated feedback from other reviewers and conducted additional experiments, which may also help clarify any further questions you might have.

Since the discussion phase ends soon, we wanted to follow up and would greatly value any further thoughts or concerns you might have so we can address them appropriately.

Thank you again for your time and commitment to the review process.

Response to Reviewer pV7f

Official Comment

by Authors (Danish Pruthi (/profile?id=~Danish_Pruthi1), Tarun Gupta (/profile?id=~Tarun_Gupta4))

31 Mar 2025, 15:57 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer pV7f, Commitment Readers

Revisions (/revisions?id=81KTbRzgBA)

Comment:

Thank you for your review and insightful comments. We are glad to learn that you found that our study addresses critical challenges of AI-generated research. We address your concerns below:

Re: "As the authors themselves acknowledge, there results of human evaluation might contain confirmation bias"

Please see the general response regarding confirmation biases.

Re: "Inclusion of some proposals that are known to be original would strengthen the results."

This is an excellent suggestion! Our human study identified two LLM-generated research documents as completely novel (score of 1). We will include these in the Appendix.

Official Review of Submission3418 by Reviewer TvaC

Official Review by Reviewer TvaC 24 Mar 2025, 15:12 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer TvaC, Commitment Readers

Revisions (/revisions?id=5lKFLkEku8)

Paper Summary:

The paper is questioning the ability of current AI based systems to generated novel research ideas. They use a panel of experts to question the novelty of such system. Experts are working under the hypothesis that the proposed research question might mostly be "plagiarism of idea" of existing research. To evaluate novelty, the current pipeline of "AI scientist" like systems are relying on a step that check for novelty and should discard generated works that are too similar to existing works. The paper shows that systems used at this step are not effective enough to detect "plagiarism of idea".

Summary Of Strengths:

The topic of generating novel research ideas together with scientific papers has attract recently a lot of attention.

Results and experiments reported by this paper, despite weaknesses, are worth being presented and discussed.

Summary Of Weaknesses:

There are several imperfections both in form and content (see suggestions and typos). As an example the text mentions 12 experts (line 308) but figure 1 says 13. The scale used to measure plagiarism should have been discussed with regards to existing works on that topic.. Surely some previous work exists in accessing and measuring plagiarism of idea.

Comments Suggestions And Typos:

Introduction is quite long and mixes elements from various dimensions of the problem that are further detailed in other sections. This leads to text redundancy. Shortening this section to broadly sketch the approach, the main results and main conclusions would sharpen the message. Detailed descriptions and discussions should be kept for the following relevant sections. This would give space to discuss more in deep the "plagiarism" notion.

I think it would be worth reaffirming more clearly that plagiarism is not just copying and pasting text without acknowledgement. It might be worth discussing a bit more in deep "plagiarism of idea". If it might be acceptable for a human to "forget", "ignore" or "re-invent" existing work it is not acceptable to "steal" idea of others without crediting them. Up to which point current AI-systems are "re-using" existing works without acknowledging is a critical question.

Authors might be aware of the latest experiment where a generated paper was accepted in the ICLR workshop. They could add "Compositional regularization; unexpected obstacles in enhancing neural network generalization" as an example in appendix.

Detailed comments:

Line 70-73: add a few words to explain why these numbers? Line 92, 95 ... : please give both absolute numbers and %, reference to table 1 is unnecessary here (in it already referenced 2 line above) Line 103-106: reference to appendix (?) Related work section should include something about positioning the content of table 1 with respect to previous work on that topic. line 2056-256: it is not clear who did compute the 1%? maybe add a reference here? line 256-271: very surprising to have 4000-200 and 500-138, what can explain this? line 308: 12 experts... line 313-314: So 12 expert plus two or this two expert are taken from the same pool? line 319: please report absolute value and % 304-346: the definition and discussion an the scale should be factorized, it is currently spread along the paper 5.2 technical similarities should be further detailed and shows more explicitly why this could be considered as plagiarism. Is it always easy to differentiate methodological similarities? Figure 3 could go in appendix (?) to save space to discuss more indeed evidence of plagiarism and detailed further the proposed examples. The bibliography needs to be check carefully. For example, the usual form of an entry for a book list should be the following: Author(s) of the book. (Year of original publication) Title of the book. Reprint, Place of publication: Publisher, Year of reprint.
Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 3.5
Excitement: 5 = Highly Exciting: I would recommend this paper to others and/or attend its presentation in a conference.
Overall Assessment: 4.5 = Borderline Award
Best Paper Justification:
The topic of generating novel research ideas together with scientific papers has attract recently a lot of attention.

Results and experiments reported by this paper, despite weaknesses, are worth being presented and discussed.

Ethical Concerns:
There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 4 = They could mostly reproduce the results, but there may be some variation because of sample variance or minor variations in their interpretation of the protocol or method.
Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
Software: 4 = Useful: I would recommend the new software to other researchers or developers for their ongoing work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Response to Reviewer TvaC

Official Comment

by Authors (Danish Pruthi (/profile?id=~Danish_Pruthi1), Tarun Gupta (/profile?id=~Tarun_Gupta4))

31 Mar 2025, 15:58 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer TvaC, Commitment Readers

Revisions (/revisions?id=3qOOPgZ96f)

Comment:
Thanks for your detailed review, suggesting many improvements in the writing and also appreciating our research's timeliness and importance. We address your concerns below:

Re: "The scale used to measure plagiarism should have been discussed with regards to existing works on that topic"

To the best of our knowledge, no authoritative work exists that provides specific measurement rubrics or standardized scales for assessing plagiarism in research. We find common terms like "mosaic plagiarism" (corresponding to score 4 in our rubric), "self-plagiarism," and "accidental plagiarism." We will discuss these concepts in our updated manuscript.

Re: "Authors might be aware of the latest experiment where a generated paper was accepted in the ICLR workshop. They could add "Compositional regularization; unexpected obstacles in enhancing neural network generalization" as an example in appendix."

We are aware of this recent development and will discuss this instance.

Re: "very surprising to have 4000-200 and 500-138, what can explain this?"

As research agents generate more ideas, the percentage of unique ideas decreases and eventually plateaus, offering diminishing returns. This pattern is discussed in section 7.1 of Si et al. (2024). We will clarify this in our updated manuscript.

Re: "So 12 expert plus two or this two expert are taken from the same pool?"

We apologize for the confusion. Our study includes 13 experts total. Twelve experts evaluated 36 proposals (3 each). For the 14 exemplars from (Si et al., 2024; Lu et al., 2024a), one expert from the same pool worked along with one new expert.

Re: "Is it always easy to differentiate methodological similarities?"

No, it is challenging. Experts must carefully read both the AI-generated paper and candidate source papers to assess overlap. Our experts often spent over an hour processing a single LLM-generated paper.

Re: "I think it would be worth reaffirming more clearly that plagiarism is not just copying and pasting text without acknowledgement."

This is an excellent suggestion. We will modify our paper to clarify this important distinction.

Re: Other writing improvement suggestions

We appreciate your careful review of our writing and helpful suggestions! We will incorporate these improvements in our updated manuscript.

Official Review of Submission3418 by Reviewer cnQh

Official Review by Reviewer cnQh 18 Mar 2025, 15:48 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer cnQh, Commitment Readers

Revisions (/revisions?id=o2oGc2DhhJ)

Paper Summary:
This paper shows that many LLM-generated research proposals are effectively plagiarized, with experts verifying that up to 36% of 50 evaluated documents exhibit significant similarity to existing work. The authors employ a rigorous expert evaluation using a plagiarism scoring rubric to detect direct copying and systematic rewording of methodologies from prior research. Additionally, evaluations of automated plagiarism detection methods reveal that current tools fail to reliably identify such sophisticated plagiarism, raising serious concerns about the originality of LLM-generated research and automated plagiarism tools.

Summary Of Strengths:
· This paper challenges earlier claims that LLM-generated proposals are more novel than those written by humans by revealing that up to 36% of LLM-generated research ideas are skillfully plagiarized

· High plagiarism scores of LLM-generated research ideas are cross-verified by asking the original authors of the found source paper for confirmation. This human evaluation is valuable for the accuracy and relevance of the findings.

· The paper breaks down automated plagiarism detection into two steps and identifies that retrieving similar source papers is the bottleneck of automated plagiarism detection.

· The paper finds that LLM-generated outputs are less diverse and might follow predictable patterns, offering an outlook for future plagiarism detection methods.

· This paper is well written with main figures that effectively support its content, making the ideas and results easily accessible to readers

Summary Of Weaknesses:

If the methodology weaknesses are addressed, I am happy to adjust my score.

· It is well known that LLMs recycle ideas/papers from their training data. However, this line of research is not mentioned, and it is unclear how this paper differs from earlier papers that did similar investigations. You might want to consider these papers for example:

Carlini, N., Tramer, F., Wallace, E., Jagielski, M., Herbert-Voss, A., Lee, K., Roberts, A., Brown, T., Song, D., Erlingsson, U., et al. (2021, August). Extracting training data from large language models. In 30th USENIX Security Symposium (USENIX Security 21) (pp. 2633–2650).

Lee, J., Le, T., Chen, J., & Lee, D. (2023). Do language models plagiarize? In Proceedings of the ACM Web Conference 2023 (pp. 3637–3647).

· Only 36 of the 50 research proposals were rated by a single human expert, which may introduce bias, especially since experts must manually identify potential plagiarized source papers and rely on their familiarity with current literature. It is also unclear how the remaining 14 proposals were evaluated.

· While the paper acknowledges potential confirmation bias by instructing experts to actively search for plagiarism and relying on author verification to maintain objectivity, this verification process may still be affected by bias, as it depends on authors confirming plagiarism suspicions.

· Although the study shows that LLM-generated research titles are less diverse than human-generated ones, it would be beneficial to determine if this holds across titles generated from different LLMs and whether similar patterns emerge at the paragraph level (e.g., comparing abstracts).

Comments Suggestions And Typos:

· The paper adjusts plagiarism scores based on the responses from source paper authors, but it does not provide sufficient details on how frequently or substantially these scores were modified. Reporting these adjustments could serve as a measure of potential scoring bias.

· There is an inconsistency in the reported number of human experts: the abstract and Figure 1 mention 13 experts, whereas Section 4.1 refers to 12 experts.

Confidence: 4 = Quite sure. I tried to check the important points carefully. It's unlikely, though conceivable, that I missed something that should affect my ratings.
Soundness: 2.5
Excitement: 3.5
Overall Assessment: 3 = Findings: I think this paper could be accepted to the Findings of the ACL.
Ethical Concerns:

There are no concerns with this submission

Needs Ethics Review: No
Reproducibility: 3 = They could reproduce the results with some difficulty. The settings of parameters are underspecified or subjectively determined, and/or the training/evaluation data are not widely available.
Datasets: 4 = Useful: I would recommend the new datasets to other researchers or developers for their ongoing work.
Software: 3 = Potentially useful: Someone might find the new software useful for their work.
Knowledge Of Or Educated Guess At Author Identity: No
Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Knowledge Of Paper Source: N/A, I do not know anything about the paper from outside sources
Impact Of Knowledge Of Paper: N/A, I do not know anything about the paper from outside sources
Reviewer Certification: I certify that the review I entered accurately reflects my assessment of the work. If you used any type of automated tool to help you craft your review, I hereby certify that its use was restricted to improving grammar and style, and the substance of the review is either my own work or the work of an acknowledged secondary reviewer.

Response to Reviewer cnQh

Official Comment

by Authors (Danish Pruthi (/profile?id=~Danish_Pruthi1), Tarun Gupta (/profile?id=~Tarun_Gupta4))

31 Mar 2025, 16:01 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer cnQh, Commitment Readers

Revisions (/revisions?id=qL2vzcXR9l)

Comment:

Thanks for your thoughtful review and valuable suggestions. We appreciate your recognition of our paper's strengths in challenging existing claims about LLM-generated research novelty. We address your concerns below:

Re: "it is unclear how this paper differs from earlier papers that did similar investigations."

Our work differs markedly from both the mentioned papers:

Carlini et al.'s "Extracting training data from large language models" focuses on verbatim training data extraction using adversarial methods. In contrast, our study examines skillful reformulations of prior research—not verbatim copies—that evade current plagiarism detection tools and were missed by experts in previous studies.

Lee et al.'s "Do language models plagiarize?" studies plagiarism in the older GPT-2 model, requires access to pre-training data, and doesn't examine plagiarism in generated research specifically. Modern LLMs like Claude 3.5 Sonnet are much more sophisticated at paraphrasing content, and their training data are orders of magnitudes bigger and also not publicly available. These differences make our work distinct and complementary.

We will include this discussion in our updated manuscript's related work section.

Re: "Only 36 of the 50 research proposals were rated by a single human expert, which may introduce bias, especially since experts must manually identify potential plagiarized source papers and rely on their familiarity with current literature. It is also unclear how the remaining 14 proposals were evaluated."

For 36 proposals, 12 experts rated 3 proposals each (12 × 3 = 36) in their areas of expertise. Each expert invested considerable time (often over an hour per proposal) identifying candidate source papers and assessing similarities. With more experts evaluating each proposal, we might have found even more instances of plagiarism, suggesting our 24% plagiarism rate should be seen as a lower bound. The remaining 14 proposals were evaluated by 2 experts.

Re: "verification process may still be affected by bias, as it depends on authors confirming plagiarism suspicions."

Regarding source-paper author bias, we note that these authors are most familiar with their own work and its nuances, positioning them well to compare their work with LLM-generated research. We acknowledge that all humans have some bias based on their worldview—completely unbiased decision-making is impossible. However, asking someone other than the source-paper authors to make these assessments would be less optimal for two reasons: first, third parties would lack the deep understanding of the original work's methodology, contribution details, and research context that the original authors possess; second, hiring additional third-party domain experts for each specific topic would be prohibitively expensive and time-consuming given the diversity of research areas covered.

We believe our approach represents the best possible effort to produce accurate, reliable results given these practical constraints.

Re: "Although the study shows that LLM-generated research titles are less diverse than human-generated ones, it would be beneficial to determine if this holds across titles generated from different LLMs and whether similar patterns emerge at the paragraph level (e.g., comparing abstracts)."

We conduct additional experiments addressing this concern. Please see our general response on "Title-only analysis and single-LLM in PCA projection study"

Re: "Reporting these adjustments could serve as a measure of potential scoring bias."

This is an excellent suggestion! Our adjustment statistics show that compared to expert panel scores, source-paper authors reduced scores by 1 point in 33.3% of cases and by 2+ points in 22.2% of cases. This indicates that source-paper authors, with their intimate knowledge of their work, make more nuanced and often more conservative judgments about plagiarism. We will include this information in our updated manuscript.

Re: "There is an inconsistency in the reported number of human experts"

We apologize for the confusion. The correct number is 13 experts; we will fix this in the manuscript.

Follow up to Reviewer cnQh

Official Comment

by Authors (Danish Pruthi (/profile?id=~Danish_Pruthi1), Tarun Gupta (/profile?id=~Tarun_Gupta4))

03 Apr 2025, 15:22 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer cnQh, Commitment Readers

Revisions (/revisions?id=2r0MY2Hush)

Comment:

We truly appreciate the time and effort you have dedicated to reviewing our submission. We have replied to your questions and conducted additional experiments addressing methodology concerns highlighted as a weakness. Additionally, we have incorporated feedback from other reviewers, which may also help clarify any additional questions you might have.

Since the discussion phase ends soon, we wanted to follow up and would greatly value any further thoughts or concerns you might have so we can address them appropriately.

Thank you again for your time and commitment to the review process.

Response to Authors Response

Official Comment by Reviewer cnQh

03 Apr 2025, 20:34 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer cnQh, Commitment Readers

Revisions (/revisions?id=2agEv4JpN1)

Comment:

Thank you for your detailed response and for clearly differentiating your work from previous studies. While I appreciate these distinctions, I still have concerns about potential methodological biases. Over 50% of the scores were adjusted by 1 or 2 points by source-paper authors, which suggests that having 2–3 independent evaluations per proposal—or including a control group that compares human-written plagiarism with truly novel ideas—could strengthen the findings, even though I understand this approach is both costly and time-consuming.

Additionally, could you please clarify your approach to title diversity? Specifically, did you combine the generated abstracts from both GPT and Claude to enhance diversity, or were their diversities assessed separately?

Given these issues, I cannot raise my score at this time.

Replying to Response to Authors Response

Official Comment by Authors

Official Comment

by Authors (Danish Pruthi (/profile?id=~Danish_Pruthi1), Tarun Gupta (/profile?id=~Tarun_Gupta4))

03 Apr 2025, 21:33 (modified: 25 Apr 2025, 00:53)

Program Chairs, Senior Area Chairs, Area Chairs, Reviewers Submitted, Authors, Reviewer cnQh, Commitment Readers

Revisions (/revisions?id=wRiab10O2L)

Comment:

Thanks for your reply!

Re: Specifically, did you combine the generated abstracts from both GPT and Claude to enhance diversity, or were their diversities assessed separately?

We independently evaluated concatenated titles and abstracts from Claude 3.5 Sonnet and GPT-4o to determine if the lower diversity pattern is consistent across different models.

Re: or including a control group that compares human-written plagiarism with truly novel ideas

While our human study lacked a control group due to resource constraints, we have conducted an additional experiment analyzing peer reviews from the PeerRead dataset. As detailed in our General Response under "Lack of control group," this analysis showed plagiarism rates of 0.8-6.3% in human-authored papers across several conferences (NeurIPS 2017, ACL 2017, ICLR 2017, and CoNLL 2016), compared to the 24% we found in LLM-generated content. We hope this experiment helps contextualize our findings on LLM-generated research plagiarism.

Re: which suggests that having 2–3 independent evaluations per proposal could strengthen the findings

We agree that having multiple independent evaluations per proposal would have led to fewer corrections by source-paper authors and likely allowed us to identify more instances of plagiarism. This suggests that our final verified plagiarism rate of 24% should be seen as a lower bound.