The impact of Artificial Intelligence on Science

An analysis on how Artificial Intelligence may be standardizing and converging science, in its writing and topics addressed

André Costa Rodrigues

Dissertation written under the supervision of
Professor Miguel Godinho de Matos.

Dissertation submitted in partial fulfillment of the requirements for the MSc in Business Analytics at the Universidade Católica Portuguesa.

05/01/2026

The impact of Artificial Intelligence on Science

André Costa Rodrigues

05/01/2026

Abstract

In this thesis, we analyzed how Artificial Intelligence may be impacting academic articles, and two dimensions were considered: Writing and Topics Addressed. We applied a Latent Semantic Analysis (LSA) model to a dataset of academic articles published from 2010 to 2025 to compute their similarities, and used two other models to add robustness. Using this model, we observed what there appears to be some suggestive evidence that the use of AI tools may be, in a statistically significant way, causing an increase in writing and topic similarity, and changing the frequency of words used. We finalize our analysis by suggesting that in the field of Computer Science, those who are more prominent in using AI tools have had a significant increase in the similarity of their works in the writing dimension. We conclude with what we believe to be the most crucial next step for a more complete and conclusive study of the impacts of AI on Science.

Keywords: Artificial intelligence, Large Language Models, Latent semantic analysis, Writing, Topics, Fields of Study.

Acknowledgements

I would like to take this opportunity to express my gratitude to my family and friends, who provided continuous support during the creation of my master's thesis. Especially, I want to thank Professor Miguel Godinho de Matos for his guidance and constructive feedback.

The impact of Artificial Intelligence on Science

André Costa Rodrigues

05/01/2026

Abstract

Nesta tese, analisamos como a Intelingência Artificial pode estar a impactar os artigos académicos, considerando duas dimensões: a Escrita e os Tópicos Abordados. Aplicámos um modelo de Análise Semântica Latente (LSA) a um conjunto de artigos académicos publicados entre 2010 e 2025 para calcular as suas similariadades e utilizámos outros dois modelos para adicionar robustez aos resultados. Com este modelo, observamos indícios de que a utilização de ferramentas de IA pode estar, de forma estatisticamente significativa, a provocar um aumento da semelhança de escrita e de tópicos, bem como a alterar a frequência de palavras utilizadas. Terminamos a nossa análise sugerindo que, na área da Ciência da Computação, aqueles que utilizam ferramentas de IA com maior frequência apresentaram um aumento significativo da semelhança dos seus trabalhos na dimensão da escrita. Concluímos com aquele que acreditamos ser o próximo passo crucial para um estudo mais completo e conclusivo sobre os impactos da IA na Ciência.

Keywords: Inteligência artificial, Grandes modelos de linguagem, Análise semântica latente, Escrita, Tópicos, Áreas de estudo.

Agradecimentos

Gostaria de aproveitar esta oportunidade para expressar a minha gratidão à minha família e amigos, que me deram um apoio constante durante a elaboração da minha tese de mestrado. Em especial, agradeço ao Professor Miguel Godinho de Matos pela sua orientação e contributos construtivos.

Contents

1 Introduction

In the early days of Artificial Intelligence (AI) models, algorithms were typically limited to supervised and unsupervised learning, and many of these algorithms drew inspiration from biological organisms and nature physical properties as a means of establishing computational capabilities to solve data-intensive problems (Kar, 2016).

The primary drawback of these initial models was that they required structured data for both model construction and information processing. Algorithms such as neural networks, decision trees, random forests, support vector machine (SVM), and k-means clustering shared this problem and had their capabilities limited by the processing methods (Duan et al., 2019).

With various advances and new needs for algorithms with text mining and natural language processing (NLP) capabilities, algorithms that could run Bidirectional Encoder Representations from Transformers (BERT), Long Short-Term Memory (LSTM), and on unstructured data were developed (Guan et al., 2019; Kushwaha and Kar, 2024).

Concurrent with this evolution, the academic literature has also seen an increase in studies on the use of chatbots (Lokman and Ameedeen, 2018).

The recent launch of OpenAI's ChatGPT significantly extended some of the capabilities of these chatbots through the integration of deep learning and linguistic models based on the Generative Pre-trained Transformer (GPT) architecture (Radford et al., 2018).

The global use and adoption of ChatGPT has demonstrated a tremendous range of possible uses for this new technology, including software development and testing, poetry, essays, business letters, and contracts (Metz, 2023; Reed, 2022; Tung, 2023). In any case, this new technology has also provoked an adverse reaction and concerns about the difficulty of differentiating AI-generated texts within academic communities (Else, 2023), with some academics questioning how ownership of AI-generated texts should be attributed (Thorp, 2023). In 2023, prestigious journals, such as Nature and Science, made it clear that it was expressly forbidden to include AI and Large Language Models (LLMs) as authors of works (Lund and Naheem, 2024). The International Conference on Machine Learning (ICML) 2023, a major machine learning conference, has also prohibited the inclusion of text generated by LLMs such as ChatGPT in submitted papers, unless the generated text was used as part of the experimental analysis of the article (ICML, 2023).

As of 2025, in the latest report available from the National Bureau of Economic Research (NBER) on how ChatGPT is being used, a working paper by OpenAI's Economic Research team and Harvard economist David Deming, the number of weekly active users is predicted to be close to 700 million, about 10% of the entire global adult population. About 30% of the conversations in this chatbot are work-related, with two of the primary uses being text and idea generation (Chatterji et al., 2025).

In a recent work by Dwivedi et al. (2023), 43 professionals contributed their opinions and expectations regarding the use and implementation of AI across different sectors of activity, including Academia.

Marcello Mariani, a professor at Henley Business School (Oxford, UK), expresses some disappointment with the results obtained through AI, stating that they lack logical flow, are inaccurate in terms of factual accuracy and truth, are not critical in their data analysis, and are not novel. Mariani also states that the lack of originality in ChatGPT's outputs is even more pronounced in products from creative industries.

Michael Wade, a professor at IMD Business School (Lausanne, Switzerland), has a more positive view of AI's potential impact. He cites Drucker (1999) when stating that knowledge work, unlike physical work, is notoriously more difficult to study due to the difficulty of measuring inputs versus outputs. However, studies suggest that approximately 41% of knowledge workers' time is spent on discretionary activities that offer little personal satisfaction and could be handled competently by others, as noted by Birkinshaw (2013). Thus, the "others", typically seen as another person, could also become a technological solution, such as ChatGPT or other LLMs.

Complementing the previous opinion, Sven Laumer, professor at the School of Business, Economics and Society, Friedrich-Alexander University (Erlangen-Nuremberg, Germany), states that ChatGPT allows one to focus on what really matters in Academia: asking thought-provoking questions and pursuing research to find answers. This statement shifts the emphasis from writing summaries of findings to the findings themselves. Thus, researchers who adopt tools like ChatGPT should be able to explore more study topics and make tasks seen as repetitive and unsatisfactory more efficient.

This thesis, therefore, aims to explore how the use and adoption of the latest LLMs and generative AI models may affect the academic work environment. This thesis conducted an analysis of similarity scores assigned to the writing and the topics covered in each article, with the analysis being performed temporally and thematically across all works, between journal categories, between and within concepts.

2 Literature Review

In this section, we intend to showcase other works that have previously studied how AI tools may affect writing and creativity. Based on the literature presented, we intend to develop the methodology and analyzes necessary to examine these two dimensions, continuing the studies already conducted on the subject.

Given the recent emergence of LLMs and large-scale generative AI, few studies have been published in prestigious journals that analyze the consequences these tools may have on academic publications. Therefore, this thesis's literature review will be divided into two parts. In the first, we intend to present a collection of works that explore how AI may affect creativity and writing. Then, we will show works that already explore how AI may be affecting recently published academic works.

2.1 How AI is affecting creativity and writing

At the time of writing this thesis, there is still little, but controversial, evidence of how AI may be impacting creativity. On the contrary, the way it is affecting writing is more evident and factual.

Before the widespread adoption of LLMs, Arnold et al. (2020) found that "predictive text encourages predictable writing" in the context of single-word suggestions provided by smartphone keyboards. Subsequently, Kreminski et al. (2022) evaluated whether an AI-based poetry composition tool caused users to produce more or less similar poems over time as they continued to use the tool.

Two recent studies of LLM users provide direct evidence of a homogenization effect in LLM-supported writing. Padmakumar and He (2023) evaluate the baseline GPT-3 model (Brown et al., 2020) versus the instruction-finetuned InstructGPT variant (Ouyang et al., 2022) in the context of a short-form

argumentative essay writing task and find an homogenization effect of LLM assistance at both the lexical and content levels, but only for the instruction-finetuned LLM. Meanwhile, Doshi and Hauser (2024) evaluate GPT-4 in the context of short-form narrative fiction writing and observe a similar homogenization effect.

Beguš (2024) compares human-written short stories with GPT-generated short stories, noting that GPT-generated short stories are less diverse than human-written ones in structure but exhibit greater gender and sexuality diversity in character descriptions. Similarly, Chakrabarty et al. (2024) compares LLM-generated short stories with stories written by expert human writers and finds that human-written short stories substantially outperform LLM-generated short stories on all dimensions of creativity measured by the Torrance Tests of Creative Thinking (TTCT) (Torrance, 1966), including the originality dimension.

2.1.1 Idea Generation

Using the conceptualization of creativity used by Doshi and Hauser (2024), creativity is essentially based on two dimensions: novelty and usefulness (Amabile, 1982; Harvey and Berry, 2023), where novelty assesses the extent to which an idea departs from the status quo or expectations, and usefulness reflects the practicality and relevance of an idea.

In the work of Doshi and Hauser, a study was conducted with a sample of 500 participants, whose objective was to find statistical evidence for differences in the creativity of the participants, who were randomly divided into three groups (Human-only, Human with one GenAI idea, and Human with five GenAI ideas). It is possible to observe that in the exercise of creating stories, people with access to AI tend to write more enjoyable stories with plot twists, better written, and less boring.

From this work, it is also possible to conclude that having access to generative AI causally increases the average novelty and usefulness relative to human writers on their own.

On the other hand, one of the main conclusions is that, with generative AI, writers are individually better off, but collectively produce a narrower range of novel content. This downward spiral shows parallels to an emerging social dilemma (Hardin, 1968): If individual writers find out that their generative AI-inspired writing is evaluated as more creative, then they have an incentive to use generative AI more in the future, but by doing so, the collective novelty of stories may be reduced further.

To complement the previous study, Anderson et al. (2024) also examined the impact of generative AI on creativity, this time with 36 participants. This study, consistent with the homogenization hypothesis, finds that users tend to produce fewer semantically distinct ideas with ChatGPT than with alternative creativity support tools (CSTs). In this study, participants completed four divergent tasks based on an object from the TTCT. Collectively, the results of the study suggest that AI-driven homogenization stems from AI providing different users with similar ideas, rather than from increasing individual user-level fixation.

To assess originality, the primary method used was semantic similarity analysis via a sentence embedding model. Thus, regarding Group-Level Homogenization, when participants used ChatGPT, the ideas they produced were less divergent from the average embedding of all ideas generated for that task; Regarding fluency, participants generated about one additional idea when using ChatGPT, and as for flexibility, participants generated ideas that covered about 27% more categories when using ChatGPT compared to the number of categories covered when using other solutions.

Figure 1: Homogenization analysis. Oblique Strategies deck (CST) on the left and ChatGPT on the right (Anderson et al., 2024).

2.1.2 Neuroscience Evidence

From a neuroscience perspective, Kosmyna et al. (2025) work at MIT examines how the use of AI and LLMs may be impacting our work and thinking skills. When participants were divided into three groups (LLM group, Search Engine group, Brain-only group), the use of LLMs had a measurable impact on participants. While the benefits were initially apparent, as demonstrated over the course of 4 months, participants in the LLM group performed worse than their counterparts in the Brain-only group across all levels: neural, linguistic, and scoring. When applying cosine similarity, there is also evidence of a more "rippled" effect in LLM-written essays, showing a bigger similarity between the results of this group.

2.1.3 Workers Productivity and Quality

In a recent work by Dell'Acqua et al. (2023), a joint project between Harvard and Boston Consulting Group, it is shown that within the "jagged technological frontier", where some tasks are easily done by AI, AI can complement or even displace human work; others, though seemingly similar in difficulty level, are outside the current capability of AI. In those, the output is inaccurate, less valuable, and degrades human performance. The inside-the-frontier experiment focused on creative product innovation and development. On the other hand, tasks outside the frontier were tasks where consultants would excel, but AI would struggle without extensive guidance.

The results suggest that for tasks within the frontier, consultants with access to AI produced higher-quality outputs, were more complete, and achieved results faster, but their outputs were more similar. As for tasks outside the frontier, consultants with AI produce, on average, fewer correct outputs and take more time than consultants without AI.

The works presented previously are important for the construction of this thesis, as they show and suggest that, when AI tools are used, a convergence of the topics addressed can be observed collectively. This convergence stems, for example, from providing different users with similar ideas, rather than from increasing individual user-level fixation. We therefore study the assumption that the effects of these tools are also present in Science, with researchers who use said tools producing work that is more similar to each other.

2.2 How AI is affecting Academia

Given the difficulty of measuring AI's impact on creativity, much of the literature on AI's impact on Academia focuses on how texts are altered. This section aims to examine how Academia has been studied in the recent past, what methods have been applied, and how we can continue the studies already carried out.

2.2.1 AI usage by academics

In a study by Van Noorden and Perkel (2023) for the journal Nature, where 1,600 researchers were asked how they have been using AI in their work, 48% of them had directly developed or studied AI themselves, 30% had used AI for their research, and the remaining 22% did not use AI in their science. Of the total, when asked how they use AI at work, just over 30% said they use it to brainstorm ideas, and just under 30% said they use it to help write research manuscripts. One of the researchers, Johannes Niskanen (a physicist at the University of Turku in Finland), further states, "If we use AI to read and write articles, science will soon move from 'for humans by humans' to 'for machines by machines'".

2.2.2 LLM detectors

The work of Liang et al. (2024) highlights how AI may already be affecting the writing of academic papers. By analyzing nearly a million articles, using OpenAI's text-embedding-ada-002 model to create vector representations of each abstract and calculate their similarities, it is possible to observe a steady increase in LLM usage, with the most significant and fastest growth observed, in the figure 2, in Computer Science papers (up to 17.5%). In comparison, Mathematics papers and the Nature portfolio showed the least LLM modification (up to 6.3%). Liang also warns that community pressures may even motivate scholars to try to sound more similar, to assimilate to the "style" of text generated by an LLM. Alternatively, LLMs may be more commonly used in research areas where papers tend to be more similar to one another. Additionally, the rapid nature of LLM research and the associated pressure to publish quickly may encourage the use of LLM writing assistance (Foster et al., 2015).

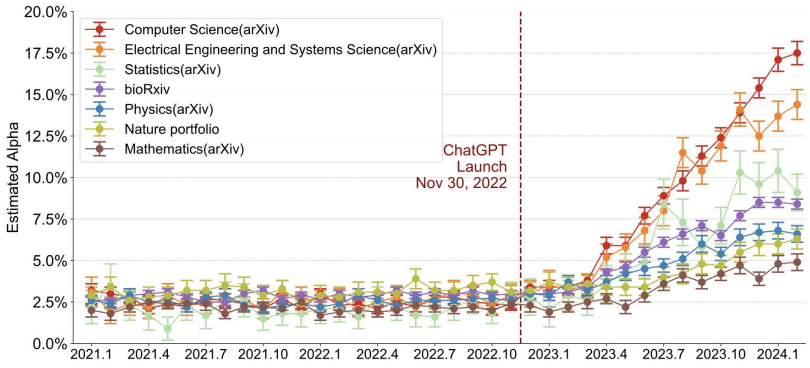


Figure 2: Estimate use of LLMs to write abstracts over time (with α representing the fraction of sentences estimated to have been substantially modified by LLM in abstracts) (Liang et al., 2024).

2.2.3 Word Frequency

In accordance with previous studies, Kobak et al. (2025) also finds evidence of a change in writing style within Academia, most notably in fields such as computation, environment, and healthcare. In his work, Kobak explores how the frequency of specific words changes across time, and although some points in time had an exponential increase in frequency of some words, such as 2020 with Covid, 2024 was the year with the most significant change in frequencies, with most of them being style-words instead of content-words, as was the case before, possible to observe in figure 3. Kobak's analysis of the excess frequency of such LLM-preferred style words suggests that at least 13.5% of 2024 PubMed abstracts were processed with LLMs.

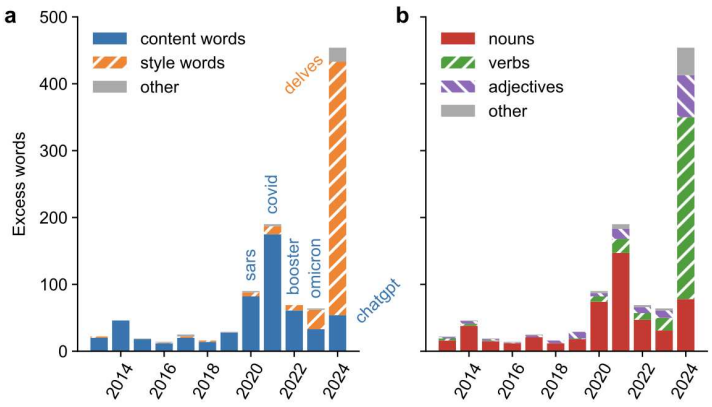


Figure 3: Excess frequency of types of words across PubMed abstracts (Kobak et al., 2025).

Kobak et al. also explores reasons for the change in frequencies across fields, countries, and journals, concluding that the high lower bound on LLM usage in computational fields (20%) could be due to computer science researchers being more familiar with and willing to adopt LLM technology. In non-English speaking countries, LLMs can also help authors with editing English texts, which could justify their extensive use. Finally, authors publishing in journals with expedited or simplified review processes might be using LLMs to write low-effort articles.

The works analyzed in this section suggest the growing use of AI tools within Science, both in idea generation and in writing. We assume, mainly from the last study analyzed here, that the identification of works done using AI must take into account the frequency of specific words.

3 Data and Descriptive Statistics

3.1 Data Collection

The data used in this thesis was collected through article requests from the OpenAlex repository. OpenAlex functions as a catalog of research publications, primarily from Microsoft Academic Graph (MAG) and Crossref, but also from other sources. The works requested are scholarly documents, such as journal articles, books, and theses. OpenAlex indexes over 240M works, with about 50,000 added daily.

Using OpenAlex's free API, 320,000 academic papers were collected (20,000 per year) between 2010 and 2025. The articles were selected based on the number of times they were cited by others; thus, the aim was to use the most relevant articles from each year rather than randomly selecting articles from less prestigious journals with less attention to quality.

After collecting the articles, those that lacked an abstract and those that were not in English were filtered out. Thus, the number of articles was reduced to 201,556 items.

3.2 Descriptive Statistics

The collected articles came from diverse fields of study and so it was necessary to identify each area for future analysis. Each article is assigned a set of concepts by OpenAlex, based on its title, abstract, and the title of its host venue, which is where you can find the best, closest to the version of record, copy of the work.

The tagging is done using an automated classifier that was trained on MAG's corpus. The concepts are assigned through scores, so the first concept is considered the most predominant in the article. Because there are a total of 65,000 possible concepts, the 19 root-level concepts were used (Political Science; Philosophy; Economics; Business; Psychology; Mathematics; Medicine; Biology; Computer Science; Geology; Chemistry; Art; Sociology; Engineering; Geography; History; Materials Science; Physics; Environmental Science). To assign each of the base concepts, the concept that first appeared in the list of concepts assigned to each article, and matched one of the root-level concepts, was then assigned as the main concept, as they were in order of score. If there was no root-level concept, the concept "other" was assigned.

Thus, it is possible to observe in figure 4 below the distribution of articles across the different areas of study, with some highly represented, such as Computer Science, Biology, Chemistry; while others rarely appear, such as Art or History.

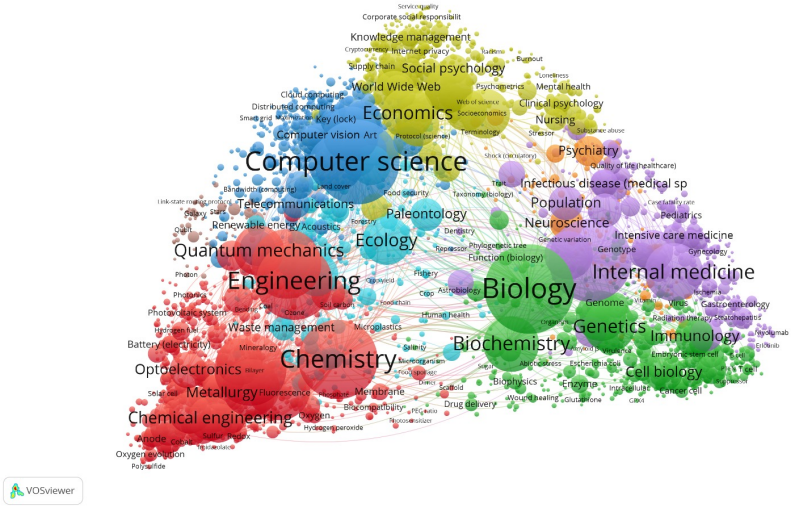


Figure 4: Map of the scientific landscape of works retrieved. With bubbles identifying concepts, their number of co-occurrences determining their size, and colors identifying each of the mathematically created clusters of concepts, by the VOSviewer tool (Waltman et al., 2010).

The order in which the articles were collected took into account the number of times each article was cited at the time of collection. Thus, the collected areas did not remain consistent in quantity over time, as shown in figure 5 below.

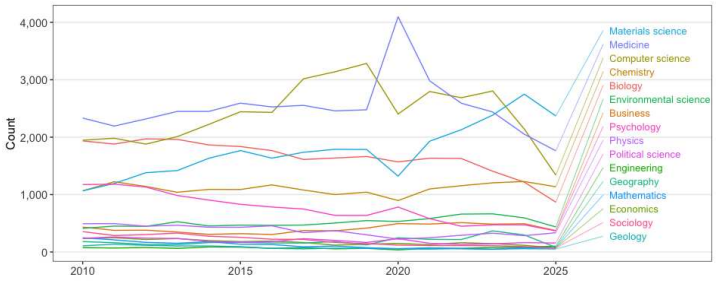


Figure 5: Works retrived over time, by main concept.

4 Methodology

In this thesis, the objective was to analyze how AI may be impacting academic publications, and two dimensions were considered: Writing and Topics Addressed.

For the analysis of the impact on writing, the abstracts of each article were considered, and we intend to study how the abstracts have changed with the appearance of AI tools, through the study of the similarity between them.

To analyze how topics addressed are changing, the distribution of topics assigned to each article was used. OpenAlex assigns each article to the topics that best suit it from a list of 4500 possible topics. Thus, in the domain of topics addressed, the potential standardization and convergence of topics across time within each area of study was analyzed.

4.1 Pre-Process of the Data

To study and compare abstracts across the different articles, all abstracts were normalized using the WordNet Lemmatizer by transforming the text into tokens. The tokenizer used transforms phrases into tokens by separating them into words, making all letters lowercase, removing URLs, email addresses, punctuation, digits, words with fewer than two letters, and stopwords (such as “the”, “a”, and “is”). Lastly, WordNet lemmatizer will transform words into their singular version (“dogs” into “dog”) if the word is found in WordNet. In this way, there are fewer rare tokens, such as emails, and fewer highly recurrent tokens, such as “We” or “the”, while maintaining the context of each abstract. Lastly, articles identified by concepts with little statistical representation (fewer than 30 appearances per year across several years), such as art or history, were also removed to reduce concepts with excessively high standard errors in their coefficients in further analysis and regressions.

4.2 Similarity between abstracts

To calculate the similarity between abstracts, a methodology applied by Larsen and Bong (2016) was used, which employed tools such as LSA and LDA to measure the similarity between IS (Information

Systems) articles. In addition to these two, we will also use SBERT (Reimers and Gurevych, 2019), a modification of the pretrained BERT network that uses siamese and triplet network structures, making it suitable for semantic similarity search. The similarity analysis is then performed year by year for each concept.

4.2.1 Latent Semantic Analysis

Latent Semantic Analysis (LSA) (Dumais (2004)) is a method for uncovering hidden meanings in texts. It examines how words appear in different documents and discovers patterns in their use. Instead of simply counting how many times words appear, LSA attempts to understand their context and relationships.

To apply this method to the dataset, firstly, the Term Frequency - Inverse Document Frequency (tf-idf) word count method is used, which, unlike the total word count, reduces the influence of words common to all documents, even after normalization, and gives greater importance to rarer words that may be related to the work's theme. In this count, only tokens that appeared in at least 10 works are taken into account to reduce computation time for infrequent words. One and two-gram tokens are considered to maintain some of the context of the abstracts; thus, "Artificial Intelligence" would be transformed into the tokens "Artificial", "Intelligence", and also into the bigram "Artificial Intelligence", which must be interpreted differently from the two individual tokens by the model.

To optimize the similarity computation process, an algorithm was developed to determine the optimal number of components/dimensions to consider for each year and concept. The code consists of a for loop with an interval of 10 to 1000 tokens (or the total number of tokens considered for a given concept), with a step size of 20 tokens. Thus, using the explained variance relation function, we can conclude that, for a given number of tokens, there is no increase in information for the LSA function. Figure 6 shows that, for Physics in 2010, the variance explained decreases as more components are added, while increasing the time required to execute the LSA code exponentially. Therefore, to optimize the number of components for detail assignment and account for the time needed to run each iteration of the model for the year and concept, the 80% information level was chosen as the most appropriate. In the example below, this implies that only 110 components need to be considered, rather than the 360 required to capture 100% of the information. The choice of this method is even more apparent and necessary when, for example, in 2010, the number of components needed for Medicine was 850, while for Engineering it was only 30. Assigning an equal number of components to all concepts would not only be arbitrary; it would also be inefficient. Still, it would introduce noise for concepts with fewer components and decrease precision for those that need more.

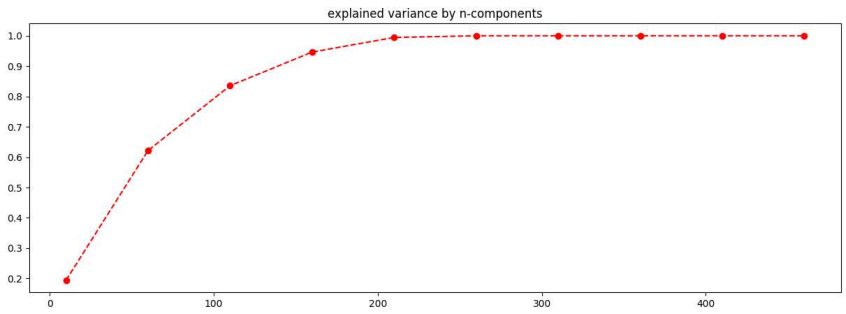


Figure 6: Explained Variance - Physics 2010

Using the chosen number of components and Singular Value Decomposition (SVD), a matrix is created whose number of columns is the minimum between the number of articles, the total number of tokens to be used, and the chosen dimensions to which the articles can belong.

With the SVD matrix, the cosine similarity between articles is then calculated, obtained according to the following formula:

$$\cos(\theta) = \frac{\mathbf{A} \cdot \mathbf{B}}{\|\mathbf{A}\|\|\mathbf{B}\|}$$

Applying the cosine similarity to each article, a similarity matrix is obtained for each pair of articles. Thus, a score is assigned to each article based on its average of cosine similarity scores, excluding its own score (which would be 1).

The use of LSA is mainly due to its advantages in finding and capturing hidden meanings in documents, as it goes beyond exact word matches and allows for Synonyms and Polysemy (the coexistence of many possible meanings for a word or phrase), so it can understand multiple meanings of words based on context. The use of 2 grams aims to address the disadvantage of not accounting for the order in which the tokens appear.

From the application of the LSA model to the dataset, the concepts that, on average, present the highest similarity score are Engineering (18.34), mostly from 2020, Geology (14.63), and Mathematics (13.20), while those that show the lowest score are Biology (4.34), Computer Science (4.19), and Medicine (3.87). It is also possible to observe in figure 7 the evolution of the scores assigned, with the latest concepts showing the greatest similarity scores being Geology and Geography, while Medicine and Material Science maintained the lowest scores across the 16 years.

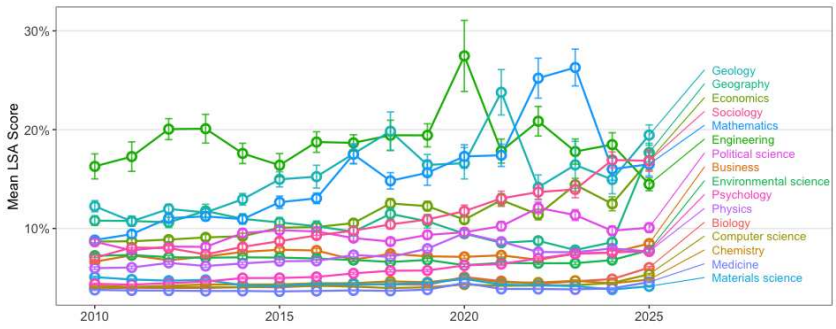


Figure 7: LSA mean scores over time for each concept for abstracts.

It is important to note that the assignment of similarity scores is highly penalizing for concepts with fewer annual observations. Belonging to one dimension will increase an article's average similarity value depending on the number of dimensions available. This penalization explains why concepts like "Engineering", with few observations and therefore fewer dimensions, show much more similarity among themselves, while concepts like "Medicine" show extremely low similarity.

4.2.2 Latent Dirichlet Allocation

Latent Dirichlet Allocation (LDA) (Pritchard et al. (2000)), the most widely applied topic modeling method, functions as an unsupervised probabilistic model. It assumes that similar documents will use

words in a similar way and, therefore, are likely to belong to the same topics. Each document is viewed as a mixture of topics, and a word distribution characterizes each topic.

Just as with the LSA, words that appear in a minimum of 10 of the works will also be used as tokens, but 2-grams will not be included.

For the model computation, because there is no mechanism for calculating the ideal number of components for each concept as efficient as the one used for the LSA model, we will use the same model, with the difference that it uses a tf-idf vectorizer that considers only 1-grams. In this way, we maintain some of the efficiency of not having the number of components entirely arbitrary.

The model is then trained on the created corpus and iterates through it 3 times to improve its learning. A matrix is created to assign the percentage of each document belonging to each learned topic.

For the similarity calculation, because the matrix represents the probability of each article belonging to each of the possible topics, the Jensen-Shannon spatial distance was used, as it already accounts for the assigned percentages, unlike cosine similarity. The Jensen-Shannon spatial distance formula is given by:

$$\sqrt{\frac{D(p \parallel m) + D(q \parallel m)}{2}}$$

Where m is the pointwise mean of p and q, and D is the Kullback-Leibler divergence.

From the application of the LDA model to the dataset, the concepts that on average present the highest similarity score are Engineering (61.23), Geology (60.14), and Mathematics (48.21), while those that present the lowest score are Materials science (14.10), Computer science (13.63), and Medicine (12.54). It is also possible to observe in figure 8 the evolution of the scores assigned, with the latest concepts with the highest similarity scores being Mathematics and Economics, while Medicine and Material Sciences had the lowest scores.

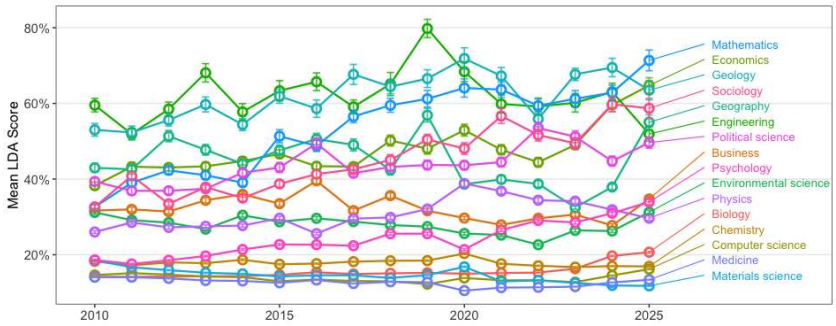


Figure 8: LDA mean scores over time for each concept for abstracts.

4.2.3 Sentence-BERT

Sentence-BERT (SBERT) (Reimers and Gurevych (2019)) is a modification of the BERT model that facilitates tasks such as semantic similarity search and unsupervised tasks such as clustering. SBERT uses Siamese and triple network structures to derive semantically meaningful sentence embeddings that can be compared using cosine similarity.

In this case, the "all-mpnet-base-v2" (Song et al. (2020)) model was used, a good general choice that maps sentences and paragraphs to a 768-dimensional dense vector space. Since the model has its own normalizer, the original abstract of each article was used instead of its normalized version. SBERT

adds a pooling operation to the output of BERT to derive a fixed-sized sentence embedding; through this output, it is then possible to apply cosine similarity between articles, as shown in the figure 9.

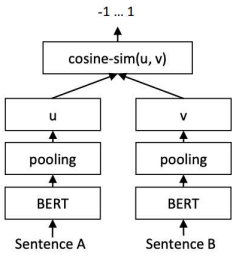


Figure 9: SBERT for similarity tasks (Reimers and Gurevych, 2019).

From the application of the SBERT model to the dataset, the concepts that, on average, present the highest similarity score are Geology (25.78), Materials science (25.41), and Sociology (24.21). In contrast, those that show the lowest score are Engineering (13.30), Computer Science (11.34), and Mathematics (9.53). It is also possible to observe in figure 10 the evolution of scores assigned, with the latest concepts with the highest similarity scores being Material Sciences and Chemistry, while Computer Science and Mathematics had the lowest scores.

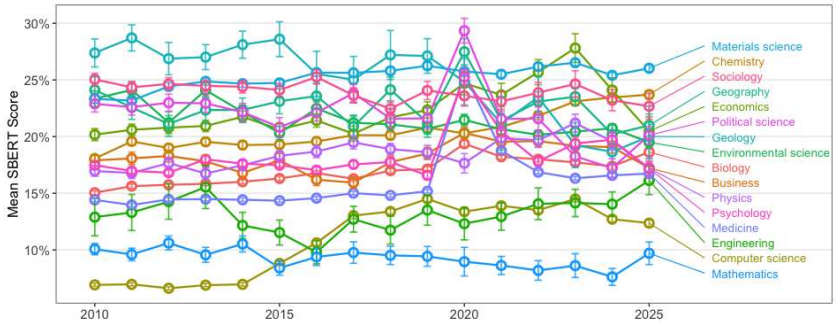
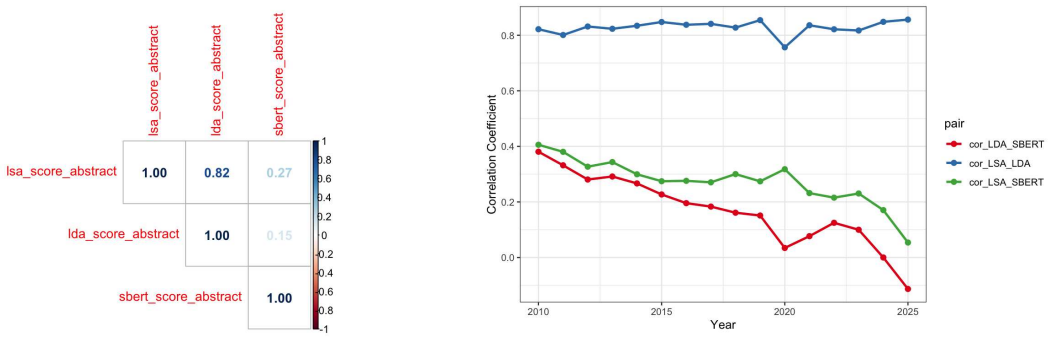


Figure 10: SBERT mean scores over time for each concept for abstracts.

4.2.4 Complementarity Between Models

As LSA was the method that obtained the best results in identifying the jingle fallacy, and the second best at identifying the jangle fallacy, by detecting similarity between items the closest to reality, in the study by Larsen and Bong (2016), it will be considered the representative method in assigning similarity scores on abstracts.

Based on the figures below, we observe that the LDA method maintains a score assignment highly correlated with that of the LSA model; however, the SBERT model diverges from both models over time, even reaching negative correlation with the LDA model.



(a) Correlation between tools (b) Cross-Sectional Correlation

Figure 11: Correlations (Abstracts)

When studying how much of each method is explained by the other two, we observe, through a regression of each of the models by the others, that the LSA model is explained (R^2) by 0.7, for the LDA it is 0.68, and for the SBERT it is 0.10.

Regarding the redundancy analysis, through the calculation of the Variance Inflation Factor (VIF), we obtain the values 3.33, 3.15, and 1.1 for the LSA, LDA, and SBERT models, respectively. Since all values are less than 5, we can assume low multicollinearity and that each measure is relatively unique.

Finally, Principal Component Analysis (PCA), a dimensionality reduction technique that helps us reduce the number of features in a dataset while keeping the most important information, was used, and as can be seen in figure 12, the SBERT model explains and explores a different dimension than those explored by other models. PCA uses linear algebra to transform data into new features called principal components. It finds these by calculating eigenvectors (directions) and eigenvalues (importance) of the covariance matrix. PCA then selects the principal components with the highest eigenvalues and projects the data onto them to simplify the dataset.

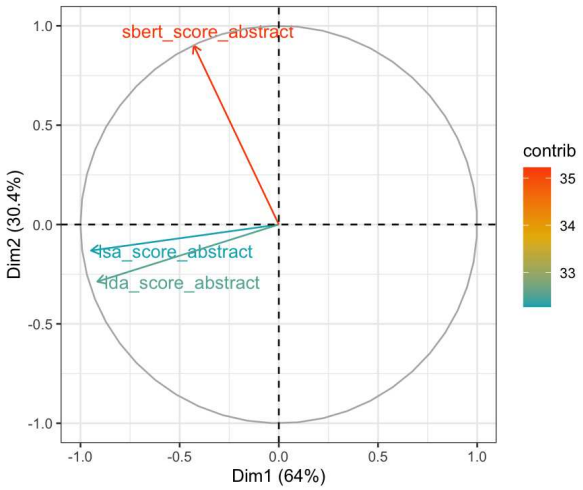


Figure 12: Principal Component Analysis between methods.

Thus, based on the analyses performed, with an average correlation of 0.4 between the models, a correlation that we consider moderate; an average unique variance of 50.8%, which indicates the

possibility of each model capturing unique information; a PC1 with an explained variance of 64%, from which we extract a main dimension, but where the others are also relevant; and an average VIF of 2.53, which indicates low redundancy between models, we conclude that the most appropriate approach is to maintain the three models.

LSA and LDA will essentially capture latent co-occurrence and topic-level semantic structure in the data. Both are rooted in analyzing word co-occurrence patterns and thematic latent spaces, which leads to substantial shared variance between them. In contrast, SBERT captures contextualized, sentence-level semantic meaning that is not well represented by traditional latent semantic or topic models, explaining why it shows distinct variance and low shared explanation with LSA and LDA. This pattern aligns with the understanding that LSA and LDA extract broad distributional topics and latent semantics, while SBERT produces dense, context-aware sentence embeddings derived from transformer architectures trained on semantic similarity tasks. (Dumais (2004), Pritchard et al. (2000), Reimers and Gurevych (2019))

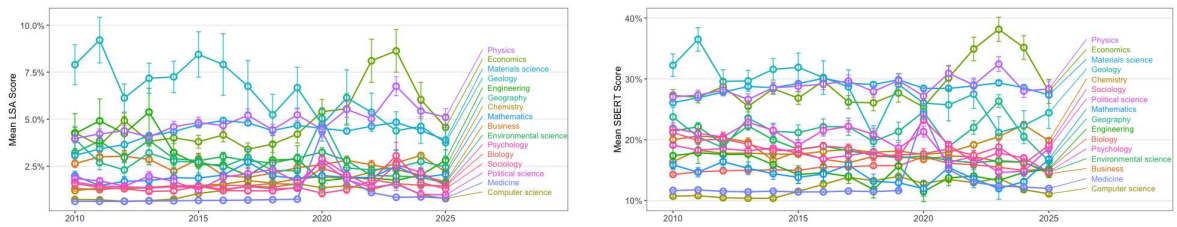
Therefore, the LDA model serves primarily as a robustness model, as it is highly correlated with the LSA model, while the SBERT model serves as a complementary model, explaining dimensions that the other models do not explain.

4.3 Similarity between topics

To calculate topic similarity, the same methodology used in the previous section was applied, using LSA and SBERT. The similarity analysis is again performed year by year for each concept.

For topic analysis, each topic assigned to each article is considered as a single token. OpenAlex assigned the topics, and each article can have between 1 and 3 topics, in order of what best represents each article. Furthermore, as with abstract analysis, the number of components is determined by iterating over each concept and year, and the choice of components ensures that at least 80% of each set of articles is explained. An additional change to the previously used model is that, because few topics are assigned to each article, tokens are considered regardless of how many times they appear in the created dictionary.

Once each model has been run across the topics of each article, it is possible to observe, in Figure 13, the evolution over time of the model's similarity scores. It is possible to observe the similarities between the LSA and SBERT models, which, despite the difference in the proportion in which similarity scores are given, both assign Physics, Material Science, and Economics as the concepts with the most similarity in 2025, with Geology being the concept with the highest average score over time in both models (6.74 and 25.78, respectively), mainly due to the early years of recording.



(a) LSA (b) SBERT

Figure 13: Mean scores over time for each concept for topics.

4.3.1 Multi Label Binarizer

The use of LDA to calculate topic similarity was not ideal, since LDA's purpose is precisely to assign topics to raw text; having the topics already defined does not allow the model to learn and assign realistic similarity scores to the different articles.

To add robustness and certainty to the scores of the other two models, we decided to implement a model that uses the MultiLabelBinarizer function (Pedregosa et al. (2011)). This model is ideal for the topic-similarity problem, as it expects pre-assigned topics as inputs and creates a direct binary representation of existing labels, without requiring learning, while being deterministic.

In figure 14, it is possible to observe that this model follows the same trend as the LSA and SBERT models.

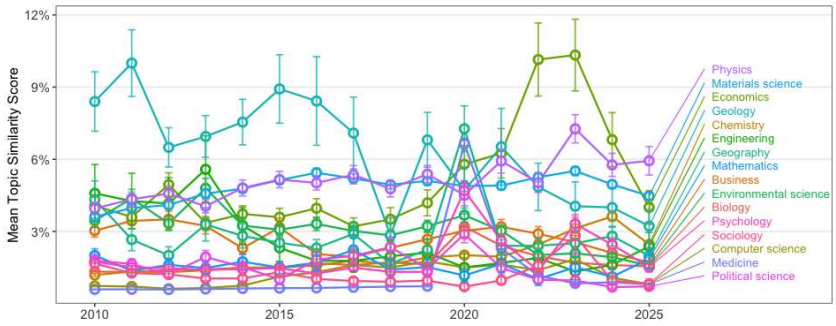
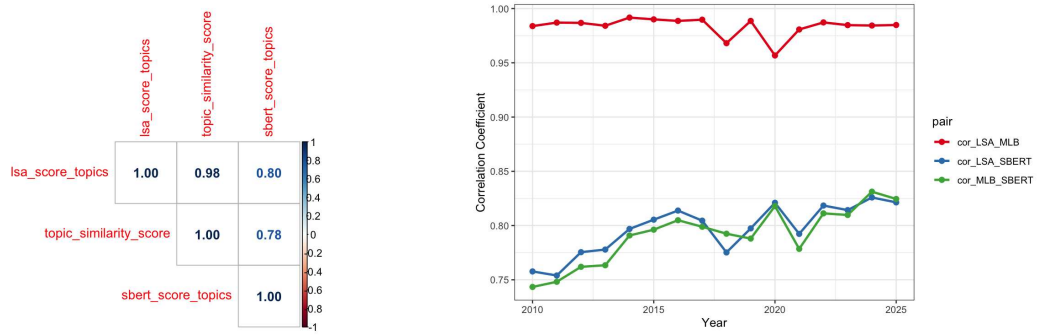


Figure 14: Multi Label Binarizer mean scores over time for each concept for topics.

4.3.2 Complementarity Between Models

As with the similarity models for abstracts, we will follow the assumption observed in the work of Larsen and Bong (2016) that the LSA model is the one that best detects similarity between items and therefore will be considered the representative model.

Unlike what was observed in the models applied to abstracts, figure 15 shows how the models are much more correlated with each other in assigning similarity scores.



(a) Correlation between tools

(b) Cross-Sectional Correlation

Figure 15: Correlations (Topics)

When the same analyses were performed, we found that the LSA model is explained by 0.954, the MLB model by 0.951, and the SBERT model by 0.638. In the redundancy analysis, the models obtained

VIFs of 21.81, 20.42, and 2.76 (LSA, MLB, and SBERT, respectively), indicating high multicollinearity between the LSA and MLB models. Finally, in the PCA analysis, one dimension is clearly dominant with a variance explained of 90.3%.

Thus, due to the high correlation, high average VIF, and the existence of a dominant PCA dimension, we used the LSA model as the representative model, but maintained the other two only for robustness and validation of the results obtained in the LSA model.

4.4 AI detector

Analyzing the levels of similarity in writing and topics addressed before and after 2022 is problematic, given that not all works were influenced by AI after November 2022 (the launch of ChatGPT), and more recent studies indicate that only around 15% of works have evidence of the use of such tools (Kobak et al. (2025)). It is excessive to group the post-2022 period as a whole and compare it with the past.

Therefore, it is necessary to identify which works may have used AI. For this purpose, unlike the various methodologies for such identification, such as the use of LLM detectors, implemented by Tang et al. (2024), keyword identification (Liang et al., 2024), or even by the frequency of words (Kobak et al., 2025), we applied a model created from scratch.

4.4.1 Methodology

To create this model, 1,000 papers were randomly selected from 2010 to 2022. Those works were sampled from different areas of study to prevent AI identification from being based on content and areas covered, rather than on writing style.

From these 1000 papers, the GPT-4o-mini model was prompted to rewrite each abstract (the prompt is shown in the Appendix). The choice of this model takes into account that, at the time this thesis was written, ChatGPT was the most widely used LLM, with its 4o model the one in use for the longest time. The use of its mini version is for monetary reasons.

From the two versions of each abstract, a dataset of 2000 entries is created, with each entry labeled as AI-generated or not. This dataset is then divided into training (1280 samples), validation (320 samples), and test (400 samples). The abstracts are then normalized to match those in the full dataset.

A simple multi-layer perceptron (MLP) was used as the model (Figure 16); it used the "multi-hot" encoding and 2-gram tokens to tokenize the abstracts. The 5000 most frequent tokens are considered, that then pass through a layer with 32 RELU functions, followed by a dropout layer for regularization, finally connected to a sigmoid layer that assigns a score to distinguish AI from non-AI.

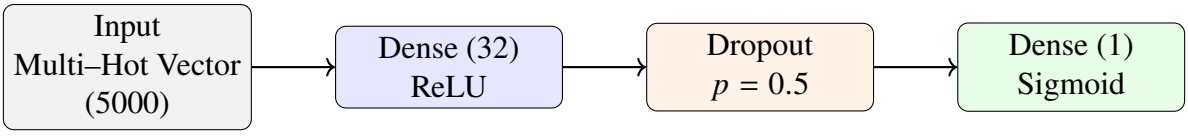


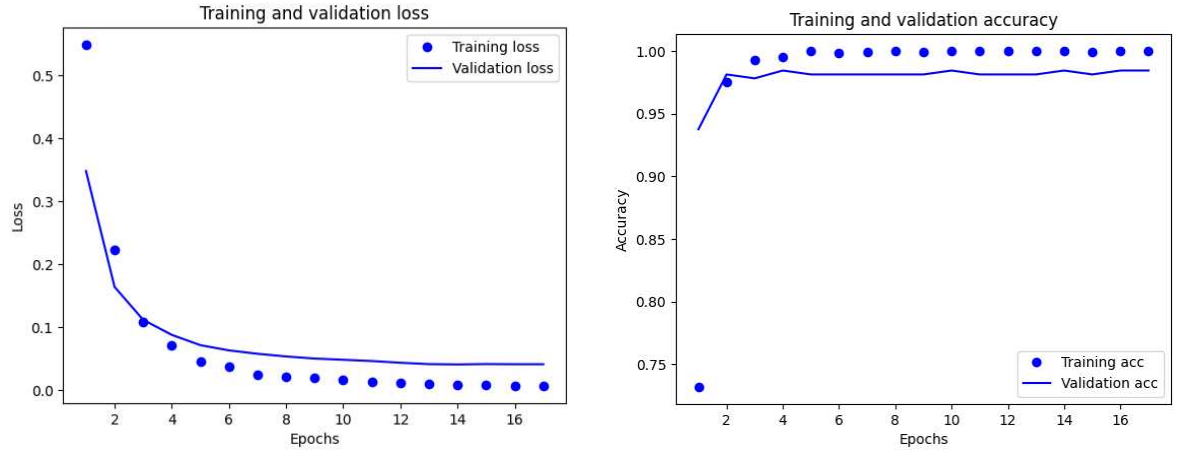
Figure 16: Model architecture.

The use of this model stems from the assumption that the identification between AI and human-made abstracts is primarily based on the words used in each abstract, as AI-generated abstracts should not and were not prompted to change the original content, only to rewrite it. Therefore, we believe it is

unnecessary to use more sophisticated models, such as LSTMs or Transformers, which take token order into account, unlike the MLP model, as they are not efficient for the task at hand.

4.4.2 Train and Validation

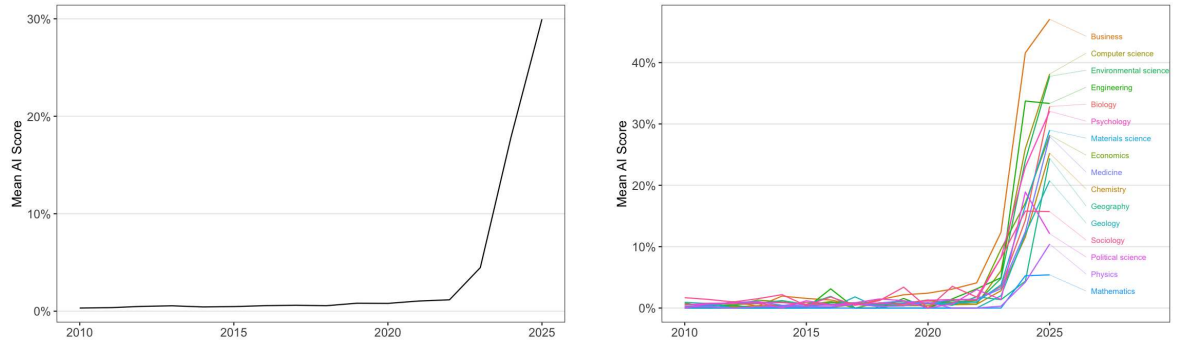
The model is then trained until the validation scores are three consecutive times lower than the previously obtained maximum. In this way, the model trains for 17 epochs and achieves a validation score of 98.44%, as shown in figure 17. Regarding out-of-sample validation, the model obtained a test score of 98.0%.



(a) Loss (b) Accuracy

Figure 17: Train and Validation of the model.

When applied to the total number of articles collected, figure 18 shows that the number of articles identified as having been influenced by AI (score greater than 0.5) is in line with expectations, with some insignificant false positives up to 2022, but followed by a sharp increase that indicates, with some error, that approximately 30% of the studies collected in 2025 have been influenced by AI. Of these, the areas with the highest proportion of abstracts identified as written with the help of AI are Business and Computer Science. On the other hand, Mathematics and Physics are the areas with the lowest proportions.



(a) Overall (b) Across Concepts

Figure 18: Train and Validation of the model.

5 Results

In this section, it is important to highlight a limitation of this thesis. Unlike the various studies analyzed in the literature review, in which mostly created specific AI-only and human-only groups, our analysis does not allow for the creation of the same type of groups beforehand. In the following subsections, we analyze the disparities in the similarity scores assigned to the works identified as AI and those who were not, but because the use of these tools was a choice of the authors, and our AI variable comes from a detector (which will not identify all AI works and will detect false positives), the assignment of this variable is not random, which does not allow us to conclude about the causality of AI impact on science. However, we can analyze the results obtained as indicative of possible evidence of the impact.

5.1 Change in Frequencies

A first observation that can be drawn, in line with the study by Kobak et al. (2025), is the difference in how abstracts are written when generated by AI, namely in the choice and frequency of word types.

When the sample for the model was created, with one abstract written by humans for each abstract written by AI, it became apparent that the frequency with which certain words appear changes. It should be emphasized that not only did some words become more frequent, such as study or research, but others lost frequency, such as data, model, or method. Even more important is the appearance of style words, such as significant, potential, or critical, while the most frequent words in human-made works are entirely content words. This finding is in line with the findings of Kobak et al. on how AI-generated texts are more likely to contain such word usage.

Table 1: Difference in frequencies between Human-made and AI

Token Human	Frequency	Token AI	Frequency
study	0.6044	study	0.8481
data	0.4555	finding	0.6243
patient	0.4485	significant	0.4895
model	0.4455	research	0.4195
cell	0.4195	analysis	0.3926
method	0.3766	potential	0.3716
research	0.3646	data	0.3616
result	0.3506	model	0.3556
new	0.3496	critical	0.3466
system	0.3466	method	0.3386

According to the model applied to all the works, Business was where the most significant growth in the use of AI tools to assist in developing abstracts was observed.

When observing the change in word frequency in figure 19, between 2020 and 2022, it is noted that the words that grew the most in their use did not increase as much as the words that grew the most from 2022 to 2024. As for the words themselves, between 2020 and 2022, the 10 words that grew the most - esg; urban; capability; emission; bank; investment; climate; green; performance; transformation - are all content words. On the other hand, among the words that grew the most from 2022 to 2024 are fostering,

enhancing, diverse, and explore; style words whose growth does not necessarily have an explanation because they are not associated with any phenomenon or event that occurred in 2024, which may be attributed to it being a choice of words made by AI.

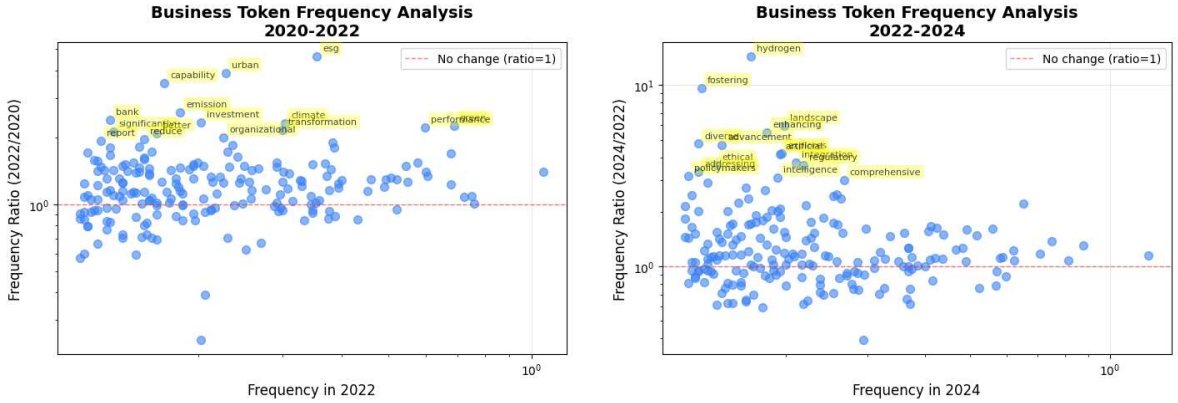


Figure 19: Change in word frequencies in Business.

On the other hand, when observing the difference in changes of word frequencies in the Physics area, which despite having a proportion of works identified as AI greater than mathematics, had more works collected over time, it is possible to observe in figure 20 that the words that grew the most in each of the moments (msub, mrow, xmlns, mml, cavity, flux, math, million, group, galaxy - in 2022; mtext, agn, inline, false, display, nircam, american, fluid, stretchy, ensuremath - in 2024) were all content words, and not style words.

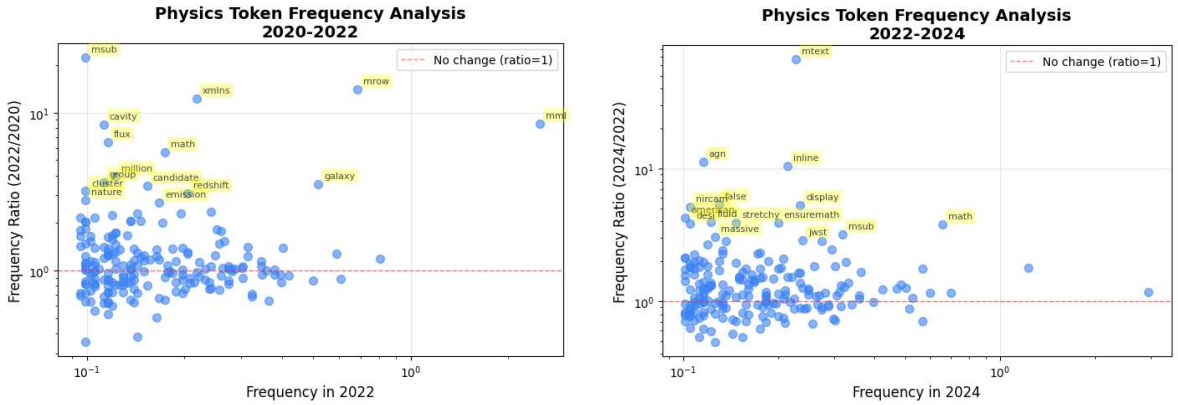


Figure 20: Change in word frequencies in Physics.

5.2 General/Descriptive Results

In the following subsection, as explained earlier, we present, in addition to the LSA method, the LDA and SBERT methods to add robustness and complementarity; however, as LSA has been suggested to be one of the most reliable methods for assigning similarity scores, the following analyses focus on this method.

Regarding the possible effects that AI has had on writing, in Figure 21 we see that in the years after the introduction of ChatGPT, in 2022, abstracts identified as having been written by AI have, on average, a higher similarity score than works identified as having been written by humans. Before 2022, the scores

show more dispersed results and wider confidence intervals, as the identification of works as having been written by AI up to that point was no more than 50 works per year, and they were false positives, because there was no such tool as ChatGPT, leading to high variance in similarity scores.

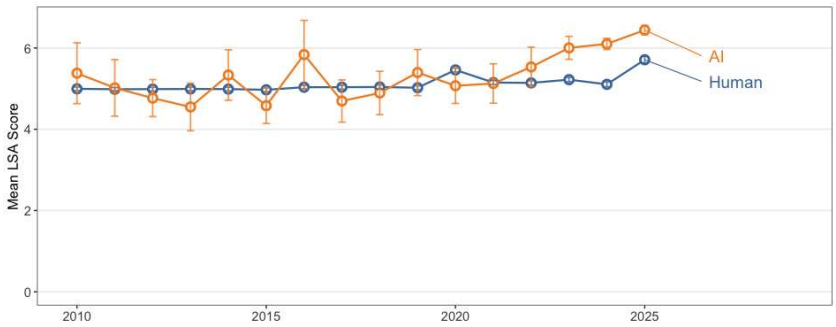


Figure 21: Mean scores over time for abstracts (LSA). Error bars indicate 95% confidence intervals.

Unlike what was observed in the possible effect of AI on writing, figure 22 shows that the effect that AI may have on topics addressed is the opposite, promoting less similar works for those who use AI in their studies.

Again, for the periods before 2022, the average similarity value between works identified as influenced by AI is highly variable and flexible over time. In any case, after 2022, it is possible to observe that the confidence intervals decrease and that the average score not only decreases but also falls below the average score for works written by humans.

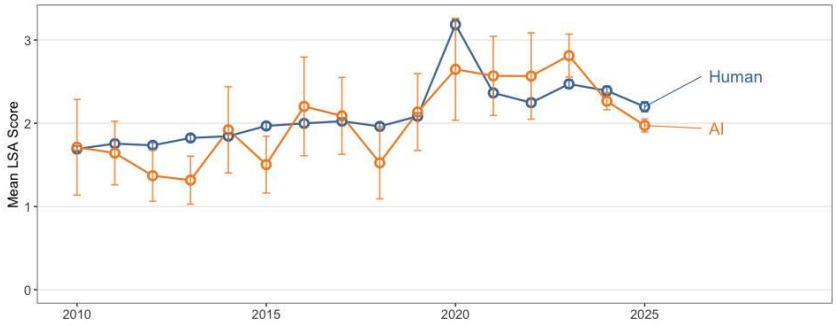


Figure 22: Mean scores over time for topics (LSA). Error bars indicate 95% confidence intervals.

To determine if the AI identification variable explains the similarity, a first instinct would be to use fractional regressions, considering the results obtained have scores ranging from 0 to 100. Still, as shown in figure 23, the distribution of LSA scores in abstracts is highly right-skewed, with a mean of 5.1, a median of 4.5, 95% of scores below 10% similarity, and no scores above 90% similarity. Previously, it was observed that the LDA method was the one that presented articles with the highest similarity among themselves, but even those exhibit a highly right-skewed distribution with zero works above 90% similarity. Thus, the use of fractional regressions, both probit and logit, was not necessarily ideal, as values close to zero, which is the case for most observations, will be pulled towards very negative values, which would make any analysis of the interaction between those classified as AI and the temporal variable insignificant.

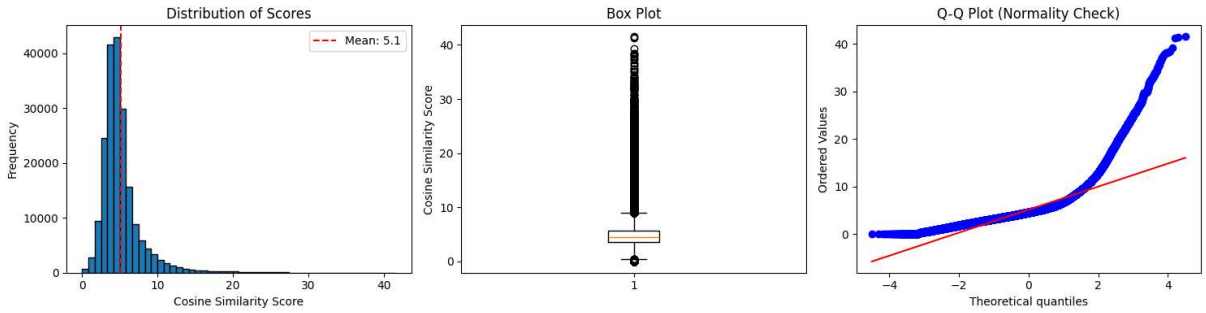


Figure 23: LSA scores for abstracts.

Therefore, the regression used was an Ordinary Least Squares (OLS) regression, as it presents no upper-boundary concerns, has a more straightforward interpretation, and is computationally faster. However, to solve the skewness problem, using a log-transformation would be preferable, as it would transform the distribution of the dependent variable into a near-normal distribution, as can be seen in figure 24. Unfortunately, with observations having similarity scores of 0, it is impossible to perform this analysis without removing these observations or adding something to the logarithmic transformation. Neither of these options is ideal; the first, by reducing our sample size, will increase the uncertainty of the coefficients and will not allow a complete analysis of the effects of AI on science. The second option, adding a small integer to the dependent variable so that it is not 0, is also not flawless, as previously explored works by Changyong et al. (2014) and Feng et al. (2013) show that adding an arbitrary constant to the logarithmic transformation impacts the result and coefficients of the regression.

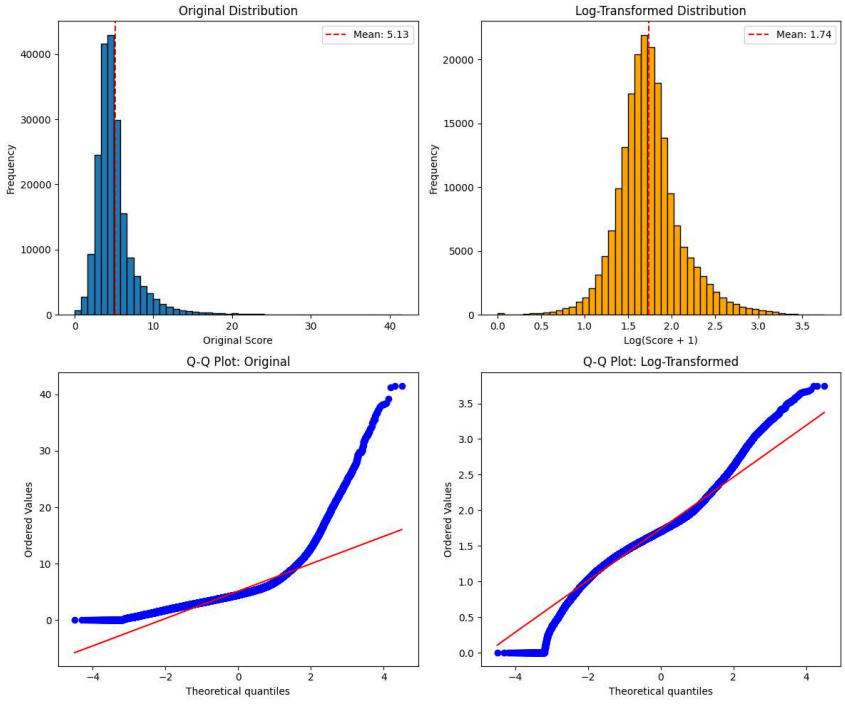


Figure 24: LSA scores - blue: unchanged; orange: log.

Thus, instead of using the logarithmic transformation, the transformation was done using the square root. Although not as efficient in reducing the effects of skewness, as can be seen in figure 25, it does not

have the impediment of scores equal to zero, as the square root of 0 is computable and is 0, and it reduces the proportion of results grouped in the same values.

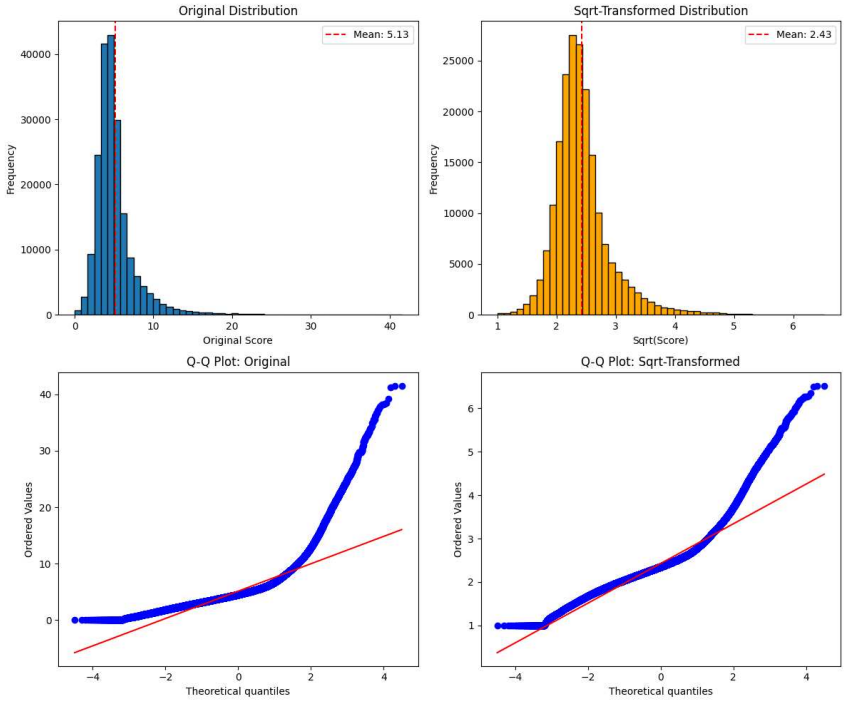


Figure 25: LSA scores - blue: unchanged; orange: log.

5.3 Regressions with Controls

To conduct a robust analysis, we considered the independent variables year, AfterLLM, main concept, journal's prestige, and AI as suggestive evidence of the similarity between articles (analysis of the variables main concept and journal's prestige can be found in the Appendix). Therefore, we do not group all articles published after 2022 together, but we acknowledge the possibility of year-on-year differences in similarity. We also recognize that the similarity is not equal across all areas of study, and that the addition of the main concept and journal's prestige variables provides a positive contribution to explaining the dependent variable. Thus, the similarity score is given by:

$$\sqrt{Similarity}:$$
$$\beta_0 + \beta_1 AI + \beta_2 AfterLLM + \beta_3 AI * AfterLLM + \beta_x Year_x + \beta_y MainConcept_y + \beta_z Prestige_z + \epsilon$$

Where AI represents the works identified as having been done by AI; AfterLLM all works after November 30, 2022, the date ChatGPT was launched; AI*AfterLLM the works that represent the interaction of both AI and AfterLLM; Year is a dummy variable for each of the years; Main concept represents each of the 16 main concepts considered for analysis; and Prestige a dummy variable that considers the H-index of each article.

Through the OLS regression, we observe in table 2 a suggestive effect of AI on similarity to be positively significant, at a confidence level of 1%, when considering the LSA model and the others. Once the dependent variable is in its square root version, the interpretation of the interaction coefficients is not

as intuitive as it would be in the original format, with the works identified as AI and after the introduction of AI, suggesting a significant increase in $\sqrt{Similarity}$ of 0.218 relative to the baseline for the LSA model.

Approximating the effect to the original units could be done by applying a first-order Taylor (delta-method), but that would only hold if the assumptions of zero mean error and homoskedasticity of the error were true (Greene (2003),Wooldridge (2010)). Because the variable AI is not random, we cannot assume that the expected error mean is zero, and consequently we won't apply this formula.

Table 2: Regression results for abstract similarity scores (W/ main concepts, prestige, years)

	Dependent variable:		
	LSA	LDA	SBERT
AI:AfterLLM	0.218***	0.214***	0.234***
	(0.012)	(0.016)	(0.027)
Constant	1.975***	3.911***	3.723***
	(0.005)	(0.006)	(0.010)
AI	Yes	Yes	Yes
AfterLLM	Yes	Yes	Yes
Year	Yes	Yes	Yes
Main Concept	Yes	Yes	Yes
Prestige	Yes	Yes	Yes
Observations	200,290	200,290	200,290
R^2	0.548	0.817	0.385
Adjusted R^2	0.548	0.817	0.384

Note: With dependent variable in sqrt(X) form; *p<0.1; **p<0.05; ***p<0.01

Performing the same analysis for topic address, we obtain the following table 3. We note that the interaction coefficient between the temporal variable and the AI variable is statistically significant only at the 10% confidence level for the LSA method, but significant at the 5% level for the other two. Therefore, based on the results obtained, we can suggest, consistent with the literature review, that collectively there is evidence of a possible convergence and standardization of the topics covered in the works using these tools.

Table 3: Regression results for topic similarity scores (W/ main concepts, prestige, years)

	Dependent variable:		
	LSA	MLB	SBERT
AI:AfterLLM	0.038*	0.058**	0.063**
	(0.021)	(0.026)	(0.029)
Constant	0.932***	0.910***	3.693***
	(0.008)	(0.009)	(0.011)
AI	Yes	Yes	Yes
AfterLLM	Yes	Yes	Yes
Year	Yes	Yes	Yes
Main Concept	Yes	Yes	Yes
Prestige	Yes	Yes	Yes
Observations	200,290	200,290	200,290
R^2	0.302	0.246	0.390
Adjusted R^2	0.302	0.246	0.390

Note: With dependent variable in sqrt(X) form; *p<0.1; **p<0.05; ***p<0.01

5.4 Deep Dive within Concepts

One question that may arise regarding this study is whether the impact we observe from these AI tools is due to the tools themselves or to the people. It is possible that the increases and decreases in similarity across the dimensions studied are not attributable to the AI tools, but instead to the different people who make up each area of study. Thus, in this section we propose an analysis where we select teams for two distinct areas (Computer Science - high AI proportion; and Physics - low AI proportion), for each area of study we compare a team of people who have been working together and where no work with AI has been identified, with a team from the same area of study with works identified as AI. Thus, by "freezing" the teams and constantly analyzing the same members of each group, we can observe if the impact on similarity stems from the use of these tools.

To create the teams, pairs were formed among authors who collaborated on works. The construction of the team clusters was then done with the help of the VOSviewer tool (Waltman et al., 2010), as it accounts for direct and indirect collaboration of authors. This tool will then create clusters of authors by maximizing the following function:

$$\hat{V}_{(x_1,\ldots,x_n)} = \frac{1}{2m} \sum_{i<j} \delta(x_i, x_j) w_{ij} \left(c_{ij} - \gamma \frac{c_i c_j}{2m} \right)$$

Where c_{ij} is given by the number of links between nodes i and j; c_i is the total number of links of node i; m denotes the total number of links in the network ($m = \frac{1}{2} \sum_i c_i$); γ is the resolution parameter (controls clustering granularity); $\delta(x_i, x_j)$ equals 1 if nodes $x_i = x_j$, and 0 otherwise; w_{ij} are weights given by: $w_{ij} = \frac{2m}{c_i c_j}$.

5.4.1 Computer Science

For the analysis of Computer Science teams, a group of 4,297 authors was selected as the non-AI team, including Jeffrey Pennington, Martin Abadi, and Alexey Dosovitskiy, with a total of 1,346 observations. The AI team was a combination of clusters with at least one paper identified as AI, including 11,875 authors, such as Kaiming He, Diederik P. Kingma, and Karen Simonyan, and had 14,936 observations.

Because the division between the teams does not prevent an author, identified as being in team AI, from contributing in a work of an author from the other team, and given the difficulty of then identifying to which team that work would belong to, all works that were duplicated (belonged to both teams) were removed, thus reducing the total number of works analyzed from 16,282 to 15,180.

In figure 26a, we observe that the AI team presents higher average similarity values in the abstracts than the non-AI team over all periods of time. However, we can also observe that after 2022, the difference between the two teams increases.

In figure 26b, the impact that these tools may be having on the topics covered is less conclusive. We observe that the AI team converges with the non-AI team, which, throughout the years before 2022, generally had fewer similar topics than the AI team. However, the AI team doesn't seem to have managed to surpass or even reach the same level of topic diversity as the team without AI. However, we can argue that, based on the convergence observed after 2022, the use of this tool may be leading to the exploration of more diverse topics within this team.

(a) Abstracts. (b) Topics.

Figure 26: Similarity Analysis Within Computer Science

Based on the table 4, we observe that the coefficient of the interaction between AI and the time variable is statistically significant at the 5% confidence level for the writing dimension, but not significant for the topic analysis. This analysis suggests that, in the field of Computer Science, teams of authors most likely to use AI tools are most likely producing increasingly similar work in terms of writing style, but not in terms of topics addressed.

Table 4: Regression results for Computer Science teams

	Dependent variable (LSA):	
	Abstract	Topics
TeamAI:AfterLLM	0.062^{**}	-0.009
	(0.029)	(0.050)
Constant	1.880^{***}	0.645^{***}
	(0.017)	(0.029)
TeamAI	Yes	Yes
AfterLLM	Yes	Yes
Year	Yes	Yes
Prestige	Yes	Yes
Observations	15,180	15,180
R^2	0.036	0.140
Adjusted R^2	0.034	0.139

Note: With dependent variable in sqrt(X) form; *p<0.1; **p<0.05; ***p<0.01

The conclusions drawn from the previous analysis are only valid if there is a parallel trend between the two teams in the years before 2022 and if the composition of the groups did not change, so that the significant coefficients obtained in the table above demonstrate that it is the use of these tools that suggests the impact on similarity.

To address the issue of changes in group members, during team formation, consideration was given to select only authors who had worked on a sufficient number of papers (by stipulating at least 100 co-authorships), so that the group would remain as identical as possible throughout the analysis.

Unfortunately, for the assumption of parallel trends, when performing the regression of similarity in abstracts using the LSA method, the interaction coefficient between Team AI and the years 2014, 17, 20, and 21 is statistically significant, while the regression for topic similarity gives significant interaction coefficients for all years from 2013 to 2023.

To study whether these coefficients are truly significant and impactful in the assumption of parallel trends, we propose to perform a joint significance test, where we group and study the hypothesis of assuming that all interaction coefficients up to 2022 are 0, and the same for the coefficients after 2022.

Thus, as can be seen in tables 5 and 6, we can assume that the significance of the interaction coefficients between the Team AI and the years before 2022 is 0, for the confidence level of 5%. On the other hand, we observe that the interaction coefficients for the subsequent years are considered significant for the 5% confidence level.

This analysis then allows us to suggest that there is evidence that tools such as AI are associated with an increase in the similarity in writing among authors more likely to use this type of tool.

Table 5: Joint Test of Interaction Effects: YearsBefore:TeamAI (Abstract)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	15157	1150.5				
2	15145	1149.0	12	1.5539	1.7069	0.05861*

Table 6: Joint Test of Interaction Effects: YearsAfter:TeamAI (Abstract)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	15148	1150.2				
2	15145	1149.0	3	1.29	5.668	0.00071***

The same cannot be observed for the impact analysis on the topics, as it is impossible to assume that the interaction coefficients are insignificant and equal to 0 for periods prior to 2022. However, the interaction coefficients between the years 2024 and 2025 and Team AI are not significant. For that reason, we shouldn't rule out the possibility of these tools, although not allowing their users, collectively, to conduct studies covering as many diverse topics as those who did not use such tools, to allow them to address more topics than they previously did.

Table 7: Joint Test of Interaction Effects: YearsBefore:TeamAI (Topics)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	15157	3379.8				
2	15145	3363.4	12	16.395	6.1519	6.557e-11***

Table 8: Joint Test of Interaction Effects: YearsAfter:TeamAI (Topics)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	15148	3366.3				
2	15145	3363.4	3	2.9036	4.3583	0.004489***

5.4.2 Physics

For the analysis of physics papers, because fewer papers were collected, the teams are more abstract and were created with less rigidity than the computer science teams.

In this area, the team without AI is composed of 12,732 authors (the high number of authors is due to the high number of authors credited in each paper, much higher than the average number of authors credited per paper in the area of computer science), among these authors are B.P. Abbott, N. Aghanim, and G. Aad, with a total of 2,053 papers. The AI team has 9,841 authors, including M. Zahid Hasan, Kin Fai Mak, and Xiao-Liang Qi, with 2,511 papers.

In the figures below, unlike what was observed in the computer science analysis, the AI team is converging with the no-AI team in terms of writing, while diverging in terms of topics approached. In the figure on the left, we observe that the AI team had, on average, greater similarity scores on abstracts during almost all the years before 2022, whereas with the emergence of these tools, the texts have become increasingly similar to the other team. Regarding the topics covered, we can also observe that before 2022, it was quite varied which team addressed the most similar topics, whereas after 2022, we observe that the AI team always addresses more similar topics. These two conclusions contradict the analysis previously done for the field of Computer Science.

It is worth noting that the analysis of these two teams is less clear than the analysis for Computer Science, considering that of the 2,511 papers in the AI team, only 16 were actually classified as AI. The creation of the AI team was based on the authors and co-authors who participated in these works, identified as such, but because there are very few of them, the existence of a false positive in the 16 leads to the addition of numerous authors who should be in the No-AI team, contaminating the analysis.

However, we will assume that such contamination did not occur and that the analysis is representative of what is happening in this area of science.

(a) Abstracts. (b) Topics.

Figure 27: Similarity Analysis Within Physics

In the table below, we can see, contrary to the analysis on the Computer Science teams, that the interaction coefficients between the time variable and the AI variable is now significant, for the 10% confidence level, for the topics addressed analysis, leading to the belief that the emergence and use of these tools did affect the team of people more prominent in using these tools in the field of Physics to produce work more related in the topics studied.

Table 9: Regression results for Physics teams

	Dependent variable (LSA):	
	Abstract	Topics
TeamAI:AfterLLM	−0.050	0.162*
	(0.043)	(0.093)
Constant	2.462***	2.164***
	(0.039)	(0.084)
TeamAI	Yes	Yes
AfterLLM	Yes	Yes
Year	Yes	Yes
Prestige	Yes	Yes
Observations	1,984	1,984
R^2	0.173	0.030
Adjusted R^2	0.164	0.019

Note: With dependent variable in sqrt(X) form; *p<0.1; **p<0.05; ***p<0.01

When the parallel trends hypothesis was studied, regarding the analysis of the abstracts, we can observe in tables 10 that we can assume the coefficients prior to 2022 as being 0, for a significance level of 5%. However, we cannot assume the opposite for the years after 2022, indicating possible evidence

on how these tools may actually be impacting the way of writing. This impact, based on the interaction coefficients, was mainly due to 2023, with the drop in the level of similarity in the No-AI team being more pronounced than the drop for the AI team. As this difference did not remain significant for the following years, we think it is best not to conclude that it was the introduction of this tool that caused the increase in similarity in the writing of the papers.

Table 10: Joint Test of Interaction Effects: YearsBefore:TeamAI (Abstract)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1961	196.31				
2	1949	195.24	12	1.0725	0.8922	0.5544

Table 11: Joint Test of Interaction Effects: YearsAfter:TeamAI (Abstract)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1952	196.47				
2	1949	195.24	3	1.2321	4.0999	0.006531***

Regarding the analysis of the topics, the tables below show that none of the interaction coefficients in the studied time intervals appear to be significant at the 5% level, so we cannot conclude whether there was an increase or decrease in the studied topics associated with the use of these artificial intelligence tools.

Table 12: Joint Test of Interaction Effects: YearsBefore:TeamAI (Topics)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1961	930.51				
2	1949	923.47	12	7.0342	1.2371	0.251

Table 13: Joint Test of Interaction Effects: YearsAfter:TeamAI (Topics)

	Res.Df	RSS	Df	Sum of Sq	F	Pr(>F)
1	1952	925.83				
2	1949	923.47	3	2.3562	1.6576	0.1742

6 Discussion

The results of this study suggest, based on our AI detector model, the emergence of academic papers and articles written, at least in part, with the help of AI tools and LLMs. It is also observable from the OLS regressions that the use of these tools is associated with the standardization and greater similarity in how papers are written and in the topics addressed.

Given the limitations we face in studying the causality of these tools on Science, as our variable of works identified as AI is not random, because it depends on each author's decision to use these tools, as well as the fact that it comes from a detector, we cannot conclude that the observed results are sufficient statistical evidence of any effect of the use of AI. However, the results obtained can serve as a suggestion of such an impact and be used to confirm the consistency of studies previously conducted on these topics.

Regarding the writing dimension, the results obtained are consistent with the work already carried out on this topic. The number of papers identified as having been influenced by AI tools in writing their articles, including the abstract, has increased, and these papers show greater similarity in the words they use. However, this growth in similarity is not necessarily negative, as the use of these tools can make the text more fluid, grammatically correct, and more appropriate in its word choice.

As for topics addressed, this thesis adds that these tools may have also had an impact on it, with the regression presenting a significant positive interaction coefficient for all the models. This is an addition

to previous studies on the impact of using these tools on Science, as well as corroboration from studies demonstrating the impact of using these tools on the convergence and standardization of topics covered.

It's also important to note that the article request process (the number of times each article was cited) prevents the sample used from containing works from "paper mills", organizations that do not have many criteria in the publication of their articles, and that consequently allow the use of these tools in a less limited way. Because of that, the results observed might have been different when considering those works.

Additionally, the works were identified as having AI influence solely based on their writing style; therefore, the number of articles where AI was only used for the creative process is not exact. We assume that those who used it to write the abstract may also have used it in the creative process, but we know nothing about its use solely for creativity and brainstorming purposes.

One final aspect to be considered is that, regardless of its impact on writing, this thesis presents evidence of an increasing use of AI tools in academic work. According to the study by Peters and Chin-Yee (2025), LLMs such as ChatGPT and DeepSeek, two of the most widely used LLMs in the world, generate article summaries that are more than 6 times as likely to contain generalized conclusions as the articles themselves. These generalizations sometimes oversimplify or even exaggerate scientific findings, leading to widespread misunderstanding, as the abstracts do not reflect the article's reality, and because the results are later presented as facts by LLMs that use them to learn/train or search the web.

7 Limitations

Some of the limitations encountered in this work include: a lack of articles; the models used for calculating similarity; the models used for AI identification; and the lack of available independent variables.

Regarding the number of articles collected, the collection of 20,000 articles annually, even with reductions due to a lack of abstracts or other factors that prevent analysis, is a sufficiently large sample that the observed results do not exhibit disproportionately large variances that would preclude analysis. The problem with requesting only 20,000 articles stems from the fact that the OpenAlex repository collects about 50,000 articles per day, on average. In this case, we have the limitation that this analysis only considers a considerably small sample of the universe of academic articles published annually, with the nuance of focusing on those that had the most significant impact each year (if we interpret the number of citations as overall impact). Thus, the conclusions drawn regarding the number of articles that may have used AI, the effect of this tool on writing, and changes in topics addressed may differ from reality when considering all articles in this time period. In any case, given the practical impossibility of collecting all available articles, considering OpenAlex API only allows the collection of 100,000 papers per day, and the computational impossibility of analyzing all articles published in this timeframe with the tools available, we intend with this thesis to present the possible consequences that these tools have been having on the publications that we consider are the most important in recent times.

Regarding the models used to calculate similarities, it would also be interesting to use other, more robust models that account for the order of tokens in sentences, or pre-trained language identification models that are more efficient than the SBERT model used. The non-use of these models is due to computational efficiency and economic reasons. On the one hand, we believe that the gains that would

be obtained from adding models with tools such as LSTM or Transformers would be less than the computational cost, i.e., time, spent on training and processing the articles. The models used took several hours to run, and the use of more complex tools would undoubtedly increase the total time exponentially, rendering the investigation inefficient within the defined time frame of this study. On the other hand, there are already tools with pre-trained models that are more efficient than the SBERT model used, but they come with the barrier of being paid for. Given the large number of articles to be analyzed and compared, the use of this type of tool is unfeasible without some grant or financial aid.

For the identification of AI articles, despite the high accuracy achieved in the validation and training processes, it must be admitted that there are still some doubts about how realistic the results obtained are. The model used to rewrite each abstract was GPT-4o-mini, a version chosen for its reduced price and because the GPT-4o model was the most widely used during the study period of this thesis. However, not all researchers who used tools like LLMs used this specific model, so it would be more realistic to train a model that identifies and assigns a score not only based on the GPT-4o-mini model, but also based on other widely used LLMs such as Claude, Gemini, or DeepSeek. Unfortunately, such a model is not included in this thesis due to time and financial constraints.

Finally, a limitation this thesis encountered was the lack of independent variables that could explain the different similarity scores and detach the "tool" from the "people" who used it. The work of Kobak et al. (2025) shows how the language of origin affects the similarity of studies. Unfortunately, the OpenAlex repository only allows retrieving the nationality of the institution where each article was published, and, in our opinion, this is too unreliable an assumption to be considered a fact, as many institutions work with researchers who are not from their country of origin.

8 Conclusion

When the world changes, our way of writing also changes. Historical events such as wars or revolutions result in changes in the frequency of words in text corpora (Bochkarev et al., 2014). We are now also witnessing a global event, the AI revolution, and similarly, we are also observing a change in the writing and frequency of words in text corpora.

This thesis aimed to continue and confirm previous studies on the evidence of AI use and impact on Science, while also attempting to complement those studies by adding the dimensions of its impact on writing and topics addressed. Given the limitations of this thesis, namely the impossibility of concluding a causal effect of the use of AI tools on Science, because the AI variable was not a random variable, we recommend the results obtained to be interpreted as suggestive and indicative of possible evidence, but we do not consider them sufficient to demonstrate the impact that these tools have been having on Science.

The change in writing style, as this thesis has shown, has been mainly felt in the way academic articles have been written in the periods following the introduction of tools such as ChatGPT, with the increase in style words, previously less used, compared to the extent to which they are now used.

Based on the robust analysis carried out, we observed that in both the writing and topics addressed dimensions, there appears to be some suggestive evidence that the use of AI tools may be, in a statistically significant way, causing an increase in writing similarity and topics covered. We complement our analysis

by adding two extra methods of assignment similarity scores that, consistent with the LSA method, also suggest the same impact on Science.

We concluded our analysis by attempting to distance the use of AI as much as possible from the people who use it, by studying two specific areas and creating teams that remain over time the same, with either users or non-users of AI. In this analysis, in the area of Computer Science, we observe a statistically significant coefficient, which serves as possible evidence of how the team of people most prone to using AI has disproportionately increased the similarity in their writing. On the other hand, in the field of Physics, we observe a statistically significant coefficient that suggests that the team most prone to use these tools may be converging in terms of topics being studied. Because this last analysis did not appear to be significant in the parallel trend hypothesis, we admit that those results may have been due to chance and need further investigation.

Having identified the limitations of this thesis, we conclude with what we believe to be the most crucial next step for a more complete and conclusive study of the impact of Artificial Intelligence on Science. That would be a more detailed exploration of the different areas of study. With this thesis, we noted the differences between areas of study and recognized the last analysis as the one that best demonstrates the possible impact of AI on science by attempting to separate tools from the people. However, adding more articles to be analyzed, more control variables, and studying other areas would make the evidence obtained in this thesis more factual and accurate.

A Appendix

A.1 AI prompt

Please rewrite the following academic abstract to meet modern top-tier journal standards.

The rewritten version should:

- Use clear, concise, and impactful language
- Follow a logical structure (background, methods, results, implications)
- Emphasize novelty and significance
- Use active voice where appropriate
- Be precise and specific about findings
- Maintain scientific rigor and accuracy

Please provide only the rewritten abstract, without any additional commentary.

A.2 Regression without controls

To study how significant the impact of AI is on writing, we initially assumed that similarity may depend on the use of these AI tools and that it can therefore be calculated according to the following formula:

$$\sqrt{Similarity} : \beta_0 + \beta_1 AI + \beta_2 AfterLLM + \beta_3 AI * AfterLLM + \epsilon$$

Where AI represents the works identified as having been done by AI, AfterLLM all works after November 30, 2022, the date ChatGPT was launched, and AI*AfterLLM the works that represent the interaction of both AI and AfterLLM.

Thus, through the OLS regression, we observe in table 14 a suggestive effect of AI on similarity to be positively significant, at a confidence level of 5%, when considering the LSA model. Once the dependent variable is in its square root version, it suggests a significant increase in $\sqrt{Similarity}$ of 0.198 relative to the baseline for the LSA model.

Table 14: Regression results for LSA, LDA, and SBERT models on abstracts

	Dependent variable:		
	LSA	LDA	SBERT
AI:AfterLLM	0.198***	0.180***	−0.032
	(0.018)	(0.036)	(0.034)
Constant	2.190***	4.220***	4.073***
	(0.001)	(0.003)	(0.002)
AI	Yes	Yes	Yes
AfterLLM	Yes	Yes	Yes
Observations	200,290	200,290	200,290
R^2	0.006	0.001	0.009
Adjusted R^2	0.006	0.001	0.009

Note: With dependent variables in their sqrt(x) form *p<0.1; **p<0.05; ***p<0.01

When applying the previously performed regression of the impact on writing onto the topics, we obtain the table 15 below, which demonstrates that for the LSA method is it possible to observe a statistically significant coefficient at the 5% level, showing how the use of AI tools after 2022 can be associated with studying more diverse themes than works done without the aid of these tools.

Table 15: Regression results for topic-based scores

	Dependent variable:		
	LSA	MLB	SBERT
AI:AfterLLm	−0.055**	−0.051*	−0.071*
	(0.025)	(0.029)	(0.037)
Constant	1.247***	1.249***	4.062***
	(0.002)	(0.002)	(0.003)
AI	Yes	Yes	Yes
AfterLLM	Yes	Yes	Yes
Observations	200,290	200,290	200,290
R^2	0.002	0.001	0.001
Adjusted R^2	0.002	0.001	0.001

Note: With dependent variable in sqrt(X) form *p<0.1; **p<0.05; ***p<0.01

A.3 Controls

In this section, we intend to demonstrate the relevance of control variables in explaining the similarity between the articles. Therefore, we will analyze the variables “main concept” and “journal prestige”.

A.3.1 Main Concept

We use this section to analyze how AI may be affecting each area of study. Previously, we observed that, on the one hand, the classification of works as having received AI assistance was not uniform across all areas. In some, such as Business, by 2025, around 40% of the works are considered to have had this type of assistance; in others, such as Mathematics or Physics, the proportion of works identified as such is much lower, but even these have around 5 to 10% of the works identified as AI. On the other hand, we observed in the methodology section that, when assigning similarity scores, it is not the areas with the most articles identified as AI that have the highest average similarity values, either in writing or in topics covered.

For this analysis, we simplify the study by selecting six areas out of the 16 available. These include Business, which had the highest proportion of AI; Physics, the second with the lowest AI ratio but more observations than Mathematics; and Computer Science, Medicine, Materials Science, and Chemistry, the four areas with the most collected works.

In the figure 28 below, we observe by field of study the regression coefficient of the interaction between AI and AfterLLM, as well as their standard errors. Consistent with the previous analysis, across all areas except Physics, the works identified as AI exhibit a statistically significant coefficient of similarity.

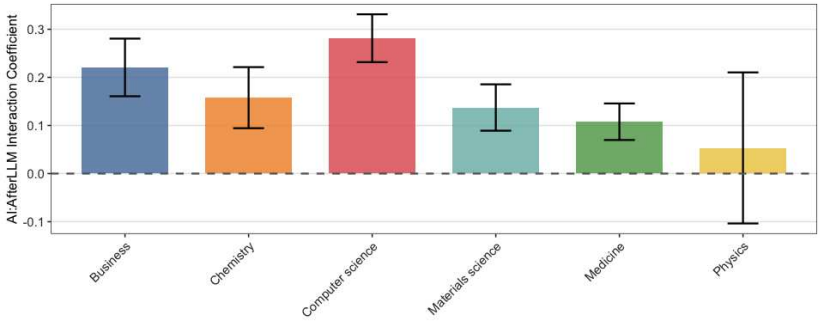


Figure 28: AI Effect After LLM Introduction by Field. (Abstract) (Error bars represent 95% confidence intervals. Dashed line at zero indicates no effect).

Regarding topic analysis, we would like to add that there is no statistically significant evidence that the use of these tools affect the topics covered in each field of study. However, Figure 29 shows, as indicated by the negative coefficients, that the use of these tools may be associated with some diversification in the topics covered, just not statistically significant enough.

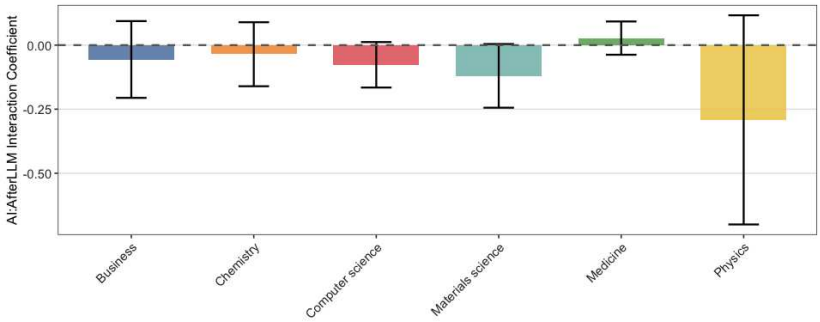


Figure 29: AI Effect After LLM Introduction by Field. (Topics) (Error bars represent 95% confidence intervals. Dashed line at zero indicates no effect).

A.3.2 Journal's Prestige

To define the prestige of each journal, we created a dataset containing every unique journal present in the dataset with all the works.

A first attempt to assign a prestige level was made using the ranking table created and edited by Scopus. The use of this ranking stems from its assignment of a prestige level (Q1-Q4) that takes into account each journal's area of publication. This standard ranking was not used because the works analyzed in this thesis were selected based on the number of citations, which implies that most are prestigious works. More specifically, about 40% of the journals were at level Q1, 40% were not at any level, and the rest were distributed between Q2 and Q3.

Thus, we created a ranking to assign the prestige levels. To do this, we used each journal's H-index, an index that combines the quality of the articles (as the number the times each article was cited) with the number of articles published, to assign a score. All those with a score above 300 in this index belong to the "top-tier" category, which includes journals such as Nature, Science, Management Science, and MIS Quarterly, and only 5% of journals are in this category. Journals with an H-index of at least 100 are placed in the "mid-tier" category, which includes about 34% of journals, and all others are placed in the "low-tier" category. All works that were not published through journals are then placed in the "other" category. It is important to note that this ranking is not similar to the rankings presented by institutions such as Scorpus; however, we understand prestige as the relationship between the quality and quantity of published work.

Regarding the writing dimension, it is possible to observe in figure 30 that the greater the prestige of the journal, the less likely it is to use vocabulary similar to that used by the rest of the work in the same area. Similarly, as the journal's prestige grows, the smaller the disparity between the work identified as done by AI and that identified as done by humans.

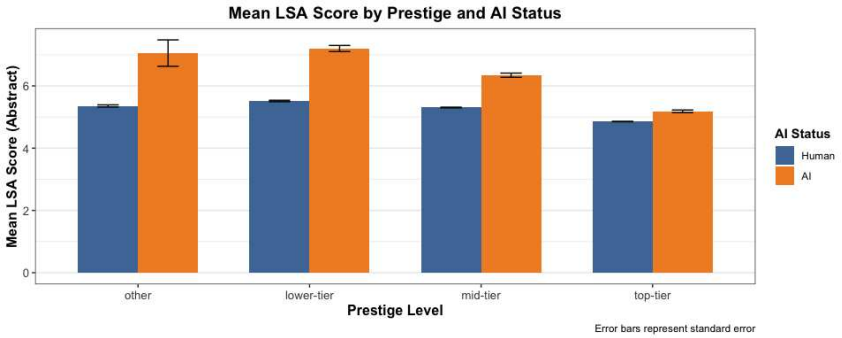


Figure 30

To study the possible impact of AI tools, we will use the same formula used previously, where similarity is dependent on AI identification and on being after 2022. We also added the main concept variable, as we consider it an explanatory variable that reduces the noise in the interaction between AI and After LLM. Again, the dependent variable is transformed to its square root to mitigate the effects of skewness. The formula is as follows:

$$\sqrt{Similarity} : \beta_0 + \beta_1 AI + \beta_2 AfterLLM + \beta_3 AI * AfterLLM + \beta_i MainConcept_i + \epsilon$$

As can be seen in Table 16, the suspicion that the impact of AI on journals is proportionally smaller the greater their prestige is confirmed. In any case, we find that even in top journals, the use of these tools is, on average, associated with an increase in similarity compared to papers that did not use this tool. In papers published in less prestigious journals, the increase in similarity is, on average, most likely greater.

Table 16: Regression results for journals

	Dependent variable: LSA Similarity (Abstracts)			
	Other	Low-tier	Mid-tier	Top-tier
AI:AfterLLM	0.370***	0.254***	0.187***	0.142***
	(0.099)	(0.043)	(0.020)	(0.017)
Constant	1.726***	2.043***	2.065***	2.044***
	(0.041)	(0.011)	(0.004)	(0.003)
AI	Yes	Yes	Yes	Yes
AfterLLM	Yes	Yes	Yes	Yes
Main Concept	Yes	Yes	Yes	Yes
Observations	8,615	22,650	66,713	102,312
R^2	0.560	0.594	0.604	0.462
Adjusted R^2	0.559	0.594	0.604	0.462

Note: With dependent variable in sqrt(X) form; *p<0.1; **p<0.05; ***p<0.01

Regarding the analysis of the impact on topics addressed, we observed the opposite trend, where top journals have, on average, increased similarity and repeated studied topics than the less prestige journals. We also observed that, contrary to the analysis of the effects on topics covered in the previous section, across all categories, the works identified as having used AI tools show a higher average similarity than those without these tools.

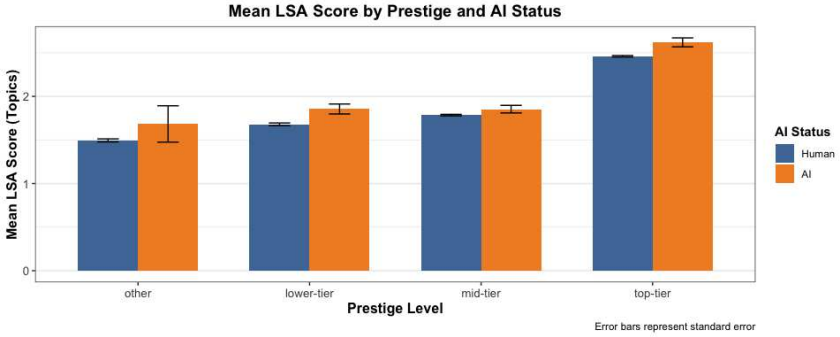


Figure 31

Applying the previous regression to topics analysis, we obtain the following table 17, which indicates that, across different prestige levels, the coefficient interaction of AI becomes non-significant for all types

of journals except mid-tier. In these, there is statistical evidence that works categorized as having AI assistance not only did not increase their topic similarity, but actually decreased it. For the remaining journal categories, despite the non-significant coefficients, all are negative (except "other", which has few observations), suggesting that, on average, these works address more diverse themes than those identified as not having used these AI tools.

Table 17: Regression results for journals

	Dependent variable: LSA Similarity (Topics)			
	Other	Low-tier	Mid-tier	Top-tier
AI:AfterLLM	0.125	−0.026	−0.102***	−0.047
	(0.142)	(0.069)	(0.034)	(0.029)
Constant	0.826***	1.089***	1.133***	1.176***
	(0.059)	(0.017)	(0.007)	(0.005)
AI	Yes	Yes	Yes	Yes
AfterLLM	Yes	Yes	Yes	Yes
Main Concept	Yes	Yes	Yes	Yes
Observations	8,615	22,650	66,713	102,312
R^2	0.095	0.139	0.201	0.351
Adjusted R^2	0.094	0.139	0.201	0.351

Note: With dependent variable in sqrt(X) form; *p<0.1; **p<0.05; ***p<0.01

References

Teresa M Amabile. Social psychology of creativity: A consensual assessment technique. *Journal of personality and social psychology*, 43(5):997, 1982.

Barrett R Anderson, Jash Hemant Shah, and Max Kreminski. Homogenization effects of large language models on human creative ideation. In *Proceedings of the 16th conference on creativity & cognition*, pages 413–425, 2024.

Kenneth C Arnold, Krysta Chauncey, and Krzysztof Z Gajos. Predictive text encourages predictable writing. In *Proceedings of the 25th International Conference on Intelligent User Interfaces*, pages 128–138, 2020.

Nina Beguš. Experimental narratives: A comparison of human crowdsourced storytelling and ai storytelling. *Humanities and Social Sciences Communications*, 11(1):1–22, 2024.

J. Birkinshaw, J. Cohen. Make time for work that matters. *Harvard Business Review, 91(9), 115–120.*, 2013.

Vladimir Bochkarev, Valery Solovyev, and Søren Wichmann. Universals versus historical contingencies in lexical evolution. *Journal of The Royal Society Interface*, 11(101):20140841, 2014.

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

Tuhin Chakrabarty, Philippe Laban, Divyansh Agarwal, Smaranda Muresan, and Chien-Sheng Wu. Art or artifice? large language models and the false promise of creativity. In *Proceedings of the 2024 CHI Conference on Human Factors in Computing Systems*, pages 1–34, 2024.

FENG Changyong, WANG Hongyue, LU Naiji, CHEN Tian, HE Hua, LU Ying, and M TU Xin. Log-transformation and its implications for data analysis. *Shanghai archives of psychiatry*, 26(2):105, 2014.

Aaron Chatterji, Thomas Cunningham, David J Deming, Zoe Hitzig, Christopher Ong, Carl Yan Shan, and Kevin Wadman. How people use chatgpt. Technical report, National Bureau of Economic Research, 2025.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan R Mollick, Hila Lifshitz-Assaf, Katherine Kellogg, Saran Rajendran, Lisa Krayer, François Candelon, and Karim R Lakhani. Navigating the jagged technological frontier: Field experimental evidence of the effects of ai on knowledge worker productivity and quality. *Harvard Business School Technology & Operations Mgt. Unit Working Paper*, (24-013), 2023.

Anil R Doshi and Oliver P Hauser. Generative ai enhances individual creativity but reduces the collective diversity of novel content. *Science advances*, 10(28):eadn5290, 2024.

P. F. Drucker. Knowledge-worker productivity: The biggest challenge. *California Management Review, 41(2), 79-94.*, 1999.

Yanqing Duan, John S Edwards, and Yogesh K Dwivedi. Artificial intelligence for decision making in the era of big data–evolution, challenges and research agenda. *International journal of information management*, 48:63–71, 2019.

Susan T Dumais. Latent semantic analysis. *Annual Review of Information Science and Technology (ARIST)*, 38:189–230, 2004.

Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. Opinion paper:"so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International journal of information management*, 71:102642, 2023.

Holly Else. Abstracts written by chatgpt fool scientists. *Nature*, 613:423, 2023.

Changyong Feng, Hongyue Wang, Naiji Lu, and Xin M Tu. Log transformation: application and interpretation in biomedical research. *Statistics in medicine*, 32(2):230–239, 2013.

Jacob G Foster, Andrey Rzhetsky, and James A Evans. Tradition and innovation in scientists' research strategies. *American sociological review*, 80(5):875–908, 2015.

William H Greene. *Econometric analysis*. Pearson education india, 2003.

Chaoyu Guan, Xiting Wang, Quanshi Zhang, Runjin Chen, Di He, and Xing Xie. Towards a deep and unified understanding of deep neural models in nlp. In *International conference on machine learning*, pages 2454–2463. PMLR, 2019.

Garrett Hardin. The tragedy of the commons. *Science*, 162(3859):1243–1248, 1968. doi: 10.1126/science.162.3859.1243.

Sarah Harvey and James W Berry. Toward a meta-theory of creativity forms: How novelty and usefulness shape creativity. *Academy of Management Review*, 48(3):504–529, 2023.

ICML. Clarification on large language model policy llm. In *https://icml.cc/Conferences/2023/llm-policy*, 2023.

Arpan Kumar Kar. Bio inspired computing–a review of algorithms and scope of applications. *Expert Systems with Applications*, 59:20–32, 2016.

Dmitry Kobak, Rita González-Márquez, Emőke-Ágnes Horvát, and Jan Lause. Delving into llm-assisted writing in biomedical publications through excess vocabulary. *Science Advances*, 11(27):eadt3813, 2025.

Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 4, 2025.

Max Kreminski, Isaac Karth, Michael Mateas, and Noah Wardrip-Fruin. Evaluating mixed-initiative creative interfaces via expressive range coverage analysis. In *IUI Workshops*, pages 34–45, 2022.

Amit Kumar Kushwaha and Arpan Kumar Kar. Markbot–a language model-driven chatbot for interactive marketing in post-modern world. *Information systems frontiers*, 26(3):857–874, 2024.

Kai R Larsen and Chih How Bong. A tool for addressing construct identity in literature reviews and meta-analyses. *Mis Quarterly*, 40(3):529–552, 2016.

Weixin Liang, Yaohui Zhang, Zhengxuan Wu, Haley Lepp, Wenlong Ji, Xuandong Zhao, Hancheng Cao, Sheng Liu, Siyu He, Zhi Huang, et al. Mapping the increasing use of llms in scientific papers. *arXiv preprint arXiv:2404.01268*, 2024.

Abbas Saliimi Lokman and Mohamed Ariff Ameedeen. Modern chatbot systems: A technical review. In *Proceedings of the future technologies conference*, pages 1012–1023. Springer, 2018.

Brady D Lund and KT Naheem. Can chatgpt be an author? a study of artificial intelligence authorship policies in top academic journals. *Learned Publishing*, 37(1):13–21, 2024.

A Metz. exciting ways to use chatgpt–from coding to poetry. techradar, 2023.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

Vishakh Padmakumar and He He. Does writing with language models reduce content diversity? *arXiv preprint arXiv:2309.05196*, 2023.

F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, J. Vanderplas, A. Passos, D. Cournapeau, M. Brucher, M. Perrot, and E. Duchesnay. Scikit-learn: Machine learning in Python. *Journal of Machine Learning Research*, 12: 2825–2830, 2011.

Uwe Peters and Benjamin Chin-Yee. Generalization bias in large language model summarization of scientific research. *Royal Society Open Science*, 12(4):241776, 2025.

Jonathan K Pritchard, Matthew Stephens, and Peter Donnelly. Inference of population structure using multilocus genotype data. *Genetics*, 155(2):945–959, 2000.

Alec Radford, Karthik Narasimhan, Tim Salimans, Ilya Sutskever, et al. Improving language understanding by generative pre-training. 2018.

L Reed. Chatgpt for automated testing: From conversation to code. *Sauce Labs*, 2022.

Nils Reimers and Iryna Gurevych. Sentence-bert: Sentence embeddings using siamese bert-networks. *arXiv preprint arXiv:1908.10084*, 2019.

Kaitao Song, Xu Tan, Tao Qin, Jianfeng Lu, and Tie-Yan Liu. Mpnet: Masked and permuted pre-training for language understanding. *Advances in neural information processing systems*, 33:16857–16867, 2020.

Ruixiang Tang, Yu-Neng Chuang, and Xia Hu. The science of detecting llm-generated text. *Communications of the ACM*, 67(4):50–59, 2024.

H Holden Thorp. Chatgpt is fun, but not an author. *Science*, 379(6630):313–313, 2023.

E Paul Torrance. Torrance tests of creative thinking. *Educational and psychological measurement*, 1966.

Liam Tung. Chatgpt can write code. now researchers say it's good at fixing bugs, too. zdnet, 2023.

Richard Van Noorden and Jeffrey M Perkel. Ai and science: what 1,600 researchers think. *Nature*, 621 (7980):672–675, 2023.

Ludo Waltman, Nees Jan Van Eck, and Ed CM Noyons. A unified approach to mapping and clustering of bibliometric networks. *Journal of informetrics*, 4(4):629–635, 2010.

Jeffrey M Wooldridge. *Econometric analysis of cross section and panel data*. MIT press, 2010.