UNIVERSITY OF PADOVA

DEPARTMENT OF MATHEMATICS “TULLIO LEVI-CIVITA”

MASTER THESIS IN CYBERSECURITY

THE IMPACT OF LARGE LANGUAGE MODELS ON ACADEMIC WRITING

SUPERVISOR
PROF. ALESSANDRO GALEAZZI
UNIVERSITY OF PADOVA

MASTER CANDIDATE
SIMONA RACCUGLIA

ACADEMIC YEAR
2024/2025

STUDENT ID
2048659

Abstract

Large language models (LLMs) are increasingly used in academia for tasks such as drafting, editing, and summarizing. While these tools can improve productivity and accelerate the research and writing processes, their repercussions on academic writing conventions are a subject that requires further investigation. This work examines the impact of LLMs on academic writing by analyzing text similarity and linguistic trends in scientific research from 2020 to 2024, using several metrics to assess over 60,000 abstracts. Findings indicate that while LLMs have not yet led to a significant homogenization of academic abstracts in terms of lexicon and overall meaning, they are reshaping writing styles in distinct ways. Through the study of AI-generated revisions of existing abstracts, linguistic features associated with the "LLM writing style" were characterized, including a preference for conciseness, increased lexical diversity and complexity, and a more direct, immediate tone, often at the cost of readability and cohesion. Temporal analyses confirm most of these trends, highlighting the evolving role of LLMs in shaping the academic discourse, and raising important questions about originality, accessibility and the future of scientific research.

Contents

Listing of figures

Listing of tables

Listing of acronyms

AI …………… Artificial Intelligence

LLM …………Large Language Model

NLP ………… Natural Language Processing

ICLR ……….. International Conference on Learning Representations

NeurIPS ……. Conference and Workshop on Neural Information

EMNLP …….. Empirical Methods in Natural Language Processing

TF-IDF ………Term Frequency-Inverse Document Frequency

BERT ……….. Bidirectional Encoder Representations from Transformers

S-BERT ………Sentence-BERT

TTR …………. Type-Token Ratio

1 Introduction

From the *Maschinenmensch* in *Metropolis,* to *HAL 9000* in *2001: A Space Odyssey* and *Samantha* in *Her,* humanity has always been fascinated by the idea of consciousness, and the thinking and feeling machine. In science fiction, artificial intelligence (AI) is often depicted as either the helpful virtual assistant or, more famously, as the evil robot turning against its creators and plotting mankind's inevitable downfall. While real-world AI may not be as sentient and futuristic (or murderous) as its fictional counterparts, its rapid evolution and consequent integration into everyday life have sparked countless discussions about its potential risks, limitations and overall repercussions on society.

Today, the conversation on AI is mostly centered on language models (LLMs), deep learning models that specialize in natural language processing (NLP). The vast majority of LLMs are based on the Transformer architecture, which uses a self-attention mechanism to understand the context of each word and its relationship to the others in the text, regardless of position. This allows them to generate original text by predicting one word at a time with high accuracy [1]. Trained on massive datasets containing billions of words, these systems can go beyond simple text generation. State-of-the-art models like OpenAI's ChatGPT and Google's Gemini interact conversationally with the user and are fine-tuned with Reinforcement Learning from Human Feedback (RLHF) to produce human-like responses [2]. They can be used for tasks such as summarizing or translating text, writing and debugging code, question answering, creative storytelling, and more, making them invaluable tools in many domains, including academia.

The breakthrough moment for LLMs came with the release of ChatGPT on November 30, 2022. Before that, most people's experience with AI was limited to the use of virtual assistants like Siri and Alexa, playing against bots in video games or using machine translation software. ChatGPT's unprecedented ability to produce coherent, detailed and human-sounding responses to prompts sparked an AI boom, capturing public attention and leading Silicon Valley companies to start investing in competing products [3]. Google quickly advanced its own conversational AI with Bard (now known as Gemini), while other companies such as Anthropic and Meta launched models like Claude and LLaMA, respectively. Today, this dynamic competition continues to accelerate the adoption of LLMs across multiple sectors.

In recent years, the automation of many processes through LLMs has significantly boosted productivity and efficiency across various fields [4]. In content creation, AI is used to draft and edit articles, blog entries and social media posts, suggest creative directions and writing prompts, or even generate original story ideas. Legal and medical professionals employ AI to analyze large amounts of data and obtain useful insights for legal research or assistance in patient diagnosis. In software development, tools like GitHub Copilot assist programmers by suggesting code snippets and debugging errors in real time [5]. These diverse applications highlight the growing impact of LLMs on society, but they also raise important concerns about misinformation, copyright, human creativity, and job displacement [6].

Specifically, the adoption of AI in academia has the potential of improving both teaching and learning, as well as transforming the way researchers interact with and interpret data. ChatGPT can create customized learning plans and study materials, including quizzes, exercises and presentations. AI tools can increase student engagement and motivation through gamification elements such as badges and leaderboards [7]. They can assist teachers in explaining complex concepts, grading homework or providing students constructive feedback [5, 7]. Platforms like Semantic Scholar use AI to help researchers sift through large amounts of academic literature by extracting meaning and identifying connections from within papers

[8]. Overall, these technologies are facilitating access to information, helping students that struggle with traditional learning methods, and speeding up the research process.

However, the widespread use of such tools does not come without concerns. One major issue is “AI hallucinations”, that is AI-generated responses that contain incorrect or misleading information. OpenAI itself lists the production of “plausible-sounding but incorrect or nonsensical answers” as one of ChatGPT’s limitations [2], and some argue that since hallucinations stem from the mathematical and logical structure of LLMs, it's impossible to fully eliminate them [9]. The issue of false scientific references is particularly common, with a 2024 study finding that chatbots make citation errors between 30% and 90% of the time, often misattributing papers' titles, first authors, or publication years [10]. Researchers and students must carefully verify the correctness of every claim before deciding to incorporate AI-generated insights into their work or general understanding of a topic. Moreover, the alarmingly common issue of scientific fraud, where researchers falsify results and fabricate data to gain funding and prestige, may be further exacerbated by AI. With its ability to generate convincing but fraudulent scientific data, AI makes it even more challenging for peer reviewers to identify and prevent misconduct [11].

Another potential cause of misinformation comes from the models’ training data. Since LLMs generate responses based on patterns in the data they were trained on, they can propagate the biases inherent in such data, or even introduce new ones. This can result in the reinforcement of specific cultural or linguistic perspectives, as well as gender or racial stereotypes [5]. For example, a recent study showed that GPT-simulated physicians making life-or-death decisions in resource-limited settings exhibit significant bias, favoring patients of similar race, gender, age, political affiliation, and sexual orientation [12]. Furthermore, AI can be easily exploited by malicious actors to facilitate the generation and dissemination of fake news through text, images, audio, and video [13] for political propaganda and misinformation campaigns.

Another point of contention is the authorship problem. Since LLMs can be used to generate research ideas, analyze data and even write entire essays and papers, questions arise about plagiarism and the extent to which AI-assisted work can be considered truly original [5]. Additionally, LLMs’ capability of answering complex questions with clarity, coherence and in-depth coverage of a topic could be exploited by students to cheat on homeworks and exams, particularly online ones [14]. Universities are tackling this problem in different ways, with some allowing AI-generated content with proper attribution, and others outright banning its use. In response to this, AI-detection software like GPTZero and Turnitin's AI detection tools have been developed to identify AI's contributions in a text. However, these tools are far from perfect, with high rates of false positives, particularly for neurodivergent writers and those who speak English as a second language [15].

Perhaps the most concerning consequence of LLM use in academia is the potential for over-reliance on AI-generated content. While LLMs can reduce cognitive load by making research and essay writing more manageable, studies indicate that students who rely heavily on these tools are less likely to engage deeply with the learning material, hindering their ability to process and critically analyze information [16]. Additionally, as more people turn to AI for their academic work, diversity of thought and originality in writing may decrease in favor of efficiency and productivity. This concern extends beyond academia, with overdependence on AI raising similar fears regarding the loss of human touch in art and literature [17].

Overall, despite the undeniable appeal of personalized learning, enhanced research capabilities, and increased efficiency, it is crucial not to overlook the potential downsides. Integrating LLMs into academia must be approached with caution, addressing serious concerns such as academic integrity, the spread of misinformation, and the loss of critical thinking to ensure responsible and effective use.

As large language models have already begun to reshape the academic landscape, much remains to be understood about their effects on writing itself. Previous studies have identified key characteristics of AI-generated academic text, including the disproportionate use of certain adverbs and adjectives [18], as well as shifts in word frequency trends [19]. Tools like ChatGPT tend to favor declarative sentences, active voice, and the simple present tense when generating academic prose. AI-generated writing is also characterized by frequent use of jargon, technical terms, and abbreviations without explicit definitions, along with high keyword and lexical density [20]. Additionally, OpenAI has noted that ChatGPT often exhibits excessive verbosity and a tendency to overuse specific phrases [2].

This research investigates the impact of LLMs on academic writing by analyzing different linguistic features that contribute to shaping writing style, as well as text similarity, in abstracts from research papers published between 2020 and 2024 in the open-access repository arXiv, or presented at one of three leading AI conferences (ICLR, NeurIPS, and EMNLP) during the same period. To assess these changes, the writing style of recent academic publications was compared to patterns observed in AI-generated text. Specifically, the study examines how LLMs affect revisions of human-written abstracts, and identifies similarities between linguistic trends in published research and tendencies seen in AI-assisted editing. These findings add to the ongoing discussion on LLMs' role in academic work, aiming to foster a more nuanced understanding of its benefits and drawbacks.

The rest of this work is organized as follows. Chapter 2 provides an overview of previous research on the impact of LLMs on society, with a particular focus on academia. It also reviews studies on the characteristics of LLMs' writing style, and approaches to compute text similarity. Chapter 3 describes the data collected and the pre-processing performed to prepare it for analysis. Chapter 4 outlines the selected metrics to evaluate text similarity and writing style, along with their implementation. Chapter 5 presents the experiments conducted on the data and its AI-generated revisions, followed by a discussion of the results. Finally, Chapter 6 concludes this work with a summary of findings, key results, and potential directions for future research.

2

Related Works

This chapter reviews past studies relevant to the topics covered in this work, including concerns about LLM use, particularly in academia, characteristics of LLMs' writing style, and the issue of text similarity.

2.1 Impact of LLMs on society and academia

The impact of LLMs on different facets of society has been extensively investigated by researchers. While much of the discussion centers on ChatGPT due to its popularity and widespread adoption, these considerations likely apply to similar tools as well.

In *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope* (2023) [5], the author P. P. Ray explores potential applications of ChatGPT across various domains, while also highlighting its problems. One of the major issues is reliability: the model sometimes generates misleading or inaccurate information which can be problematic, particularly in critical fields like healthcare, law, and finance. Since ChatGPT is trained on large datasets that reflect human prejudices, its outputs can be skewed or discriminatory, raising ethical concerns as well as questions about accountability for AI-generated mistakes. Additionally, over-reliance on AI could diminish users' critical thinking and problem-solving skills.

The model's "black box" nature means that its decision-making process is not transparent, which complicates the task of diagnosing errors or understanding how it generated certain answers. Other limitations include computational costs and scalability: training and deploying large AI models like ChatGPT require significant resources, raising concerns for the environment. The use of ChatGPT comes with safety risks, as it can be used to generate harmful content including hate speech and fake news, privacy concerns associated with the potential misuse of the data users share with ChatGPT, as well as concerns about intellectual property and authorship of AI-generated content.

In *Exploring ChatGPT and its impact on society* (2024) [6], Haque and Li further examine some of these problems and add to the debate with a discussion on the effects of ChatGPT on the job market. The ability to automate tasks such as text summarization, translation, content creation and question answering poses a risk of job displacement, particularly in industries like customer service, data entry, and technical support. Workers with technical and AI-related skills may gain a competitive advantage, leading to a rise in economic inequality. However, ChatGPT's integration into various industries has the potential to also create new job opportunities, including roles such as AI trainers, ethicists, content moderators, developers, and consultants.

The increasing reliance on ChatGPT also raises concerns about social isolation: excessive dependence on AI for information and entertainment may reduce human interaction. In some cases, users may develop an over-reliance on ChatGPT for social and emotional support, which can negatively impact their real-world relationships, physical activity and overall health.

The paper *Chatting about ChatGPT: How AI and GPT May Impact Academia and Libraries* (2023) [21] by Brady D. Lund and Ting Wang, explores the technological foundations, capabilities, and potential implications of ChatGPT, particularly within academic and library settings. The study includes a direct interaction with ChatGPT, questioning it about potential applications in academia and libraries. ChatGPT is presented as a powerful tool for research and education, assisting with literature reviews, summaries, data analysis, and language translation. However, its use raises ethical and privacy concerns, particularly regarding bias, misinformation, and data security. In libraries, ChatGPT could enhance search and discovery, metadata generation, and content creation, but it also presents risks related to intellectual property, user privacy, and the perpetuation of biases found in training data. While the technology has the potential to transform academia and libraries by improving efficiency and accessibility, it also requires responsible use to mitigate ethical risks and integration that enhances human expertise rather than replacing it.

ChatGPT's critical thinking skills were examined in *ChatGPT: The End of Online Exam Integrity?* (2022) [14] by Teo Susnjak. In this study, ChatGPT was tasked with generating complex questions for undergraduate students across various disciplines, providing answers, and critically evaluating them. The responses demonstrated clarity, precision, and comprehensive coverage of topics while maintaining logical coherence despite their length. According to the author, ChatGPT's ability to critique its own answers, identifying strengths and weaknesses and suggesting improvements, indicates genuine critical thinking rather than simple memorization and rephrasing. This makes ChatGPT a powerful tool that students could exploit for academic dishonesty, particularly in online exams, where its responses are virtually indistinguishable from those of a human.

In *Generative AI and the Automating of Academia* (2023) [22], Watermeyer et al. explore how the rise of generative artificial intelligence (GAI) tools, particularly LLMs like ChatGPT, impacts academic work within the higher education system. A survey of 284 UK academics revealed a near-even split between those using and not using GAI, though a majority acknowledged that these tools are altering their work practices. While some academics use AI to delegate administrative tasks and editing, enhancing productivity and enabling them to meet institutional demands more efficiently, others perceive it as a threat to academic integrity and quality, and worry that this newfound efficiency will lead to greater expectations for output, intensifying workload. The study highlights concerns that GAI reinforces academia's obsession with metrics, deepens existing status hierarchies, and aggravates inequalities. It warns that uncritical adoption of GAI risks further alienating academics from their intellectual work, supporting a system that values quantity over quality.

2.2 LLMs’ writing style

The stylistic features of AI-generated writing are another subject explored by previous research. Tools like ChatGPT are capable of adapting their writing style based on context and specific requirements. These chatbots can mimic different tones and formats, producing formal responses for essays while adopting a more relaxed and conversational style in casual interactions. This adaptability enables LLMs to serve various applications, from academic writing and professional reports to informal conversations and creative content.

In *Do Artificial Intelligence Chatbots Have a Writing Style? An Investigation into the Stylistic Features of ChatGPT-4* (2023) [20] by AlAfnan and MohdZuki, the authors analyze the stylistic features of ChatGPT-4’s responses to prompts with a focus on word choice, sentence length, paragraph length, tense usage, mood, voice, and pronoun selection. The study reveals variations across different types of writing. In case study-based responses, paragraphs were generally short, with active, imperative sentences, reflecting a direct communication style. These responses primarily employed the second-person pronoun “you” and the possessive “your.” Keyword and lexical density were relatively low, with limited repetition and an average level of lexical diversity, resulting in a high reading ease. For business correspondence, ChatGPT-4 generated short paragraphs using predominantly declarative sentences in simple present and simple past tenses. These responses were structured in active voice and primarily written in the third person. The study observed that business correspondence responses used subject-specific technical terms without offering definitions and exhibited high keyword and lexical density, and low lexical diversity, making them less readable compared to other forms of writing. In academic writing, ChatGPT-4 generated longer paragraphs with declarative sentences. The writing employed multiple tenses, predominantly in active voice. Pronoun usage was limited, with “they” and “their” used mainly for comparison and general references. ChatGPT-4’s academic writing style incorporated abbreviations and technical terminology without explanation and keyword and lexical density were high, while lexical diversity was low, leading to an average level of reading ease.

Both *Is ChatGPT Transforming Academics' Writing Style?* (2024) [19] by Geng and Trotta and *Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews* (2024) [18] by Liang et al. observe a shift in the frequency of specific terms in research papers, which can be attributed to the introduction of large language models in academic writing practices. Liang et al., after comparing term frequency in peer reviews from AI conferences between 2020 and 2024, noted a significant increase in the use of adjectives such as "commendable," "meticulous," and "intricate." They also compiled lists of adjectives and adverbs disproportionately used by AI. On the other hand, Geng and Trotta detected and analyzed ChatGPT's impact in arXiv abstracts through a statistical analysis of word frequency. The changes in word frequency after ChatGPT's release align with predictions made from simulations of ChatGPT's influence on users' prompts for revisions or editing. Furthermore, the study showed that the category most affected by AI use is, predictably, Computer Science, while mathematicians appear less inclined to adopt such technologies.

2.3 Text similarity

All these considerations informed the analysis of the impact of LLMs in academia in this work, which is particularly focused on text similarity and writing style.

The problem of text similarity is fundamental to many applications in natural language processing. *A Survey of Text Similarity Approaches* (2013) [23] by Gomaa and Fahmy provides a comprehensive review of various techniques used to measure text similarity, categorizing them into four main approaches. String-based similarity methods compare text at the string level, disregarding meaning. Corpus-based similarity determines similarity based on information retrieved from large corpora. Knowledge-based similarity relies on semantic networks, such as WordNet, to compute similarity. Finally, hybrid similarity methods integrate two or more of these approaches to enhance accuracy and robustness.

A more modern perspective is given by *Evolution of Semantic Similarity—A Survey* (2020) [24] by Chandrasekaran and Mago, which centers on semantic similarity. This survey adopts a similar classification (corpus-based, knowledge-based and hybrid methods) but puts more emphasis on recent approaches, describing text embedding techniques such as Word2Vec and GloVe, as well as the application of deep learning models like BERT to the text similarity problem.

3

Data & Pre-Processing

This chapter describes the two datasets used to assess the effects of LLM adoption in academic writing. It outlines the data collection process and the pre-processing steps taken to remove noise and ensure the data's suitability for analysis. Additionally, a section of this chapter defines two smaller datasets, which were derived from the original ones. These subsets were chosen due to their higher likelihood of reflecting the effects of LLM usage, as indicated by the increased presence of terms that previous research has linked to AI-generated content.

3.1 Datasets and pre-processing

The data collection process for this study aimed to compile a comprehensive set of academic texts that would allow a thorough analysis of the impact of LLMs on academia in terms of text similarity and writing style. Abstracts from research papers (dated from 2020 to 2024) were chosen as they tend to have a standardized structure, making it easier to identify patterns and differences between human-written and AI-generated text. Their brevity ensures they're compatible with models like S-BERT, which facilitate text embedding while preserving semantic meaning, but often come with token limits. Additionally, since academics primarily use LLMs for tasks such as editing, summarizing, and refining language, the abstract, which should capture the essence of a paper, is likely where AI's influence on tone, quality, and style is most pronounced.

Two main datasets were created for this study. The first one, which will be referred to as the *conferences dataset* from now on, consist of abstracts from papers presented at three leading conferences on artificial intelligence and machine learning: the *International Conference on Learning Representations (ICLR)* [25]*, the Conference and Workshop on Neural Information Processing Systems (NeurIPS)* [26], *and Empirical Methods in Natural Language Processing (EMNLP)* [27]. These conferences were selected based on a study by Liang et al. [18], which found that between 6.5% and 16.9% of peer-review submissions to these venues may have been substantially modified by LLMs beyond basic proofreading. The dataset includes abstracts from the past five editions of these conferences (2020-2024), along with each paper's title, a list of authors, the conference it belongs to, and the year of publication. The total number of abstracts in this dataset amounts to 41,129, with the following table summarizing the number of abstracts per year and conference.

	2020	**2021**	**2022**	**2023**	**2024**
ICLR	2592	3006	3414	4946	7377
NeurIPS	1893	2326	2823	3522	4526
EMNLP	751	847	825	1047	1234

Table 3.1: Number of abstracts per year and conference.

Interestingly, there is a noticeable increase in the number of papers presented at these AI-focused conferences, especially in more recent years. This trend may reflect an early consequence of LLM adoption in academia: greater efficiency in the research process leading to higher productivity and an increased volume of papers. Additionally, the surge in interest in artificial intelligence, particularly following the launch of ChatGPT in 2022, could have contributed to this growth.

The second dataset, referred to as the *arXiv dataset,* comprises abstracts from papers hosted on the open-access repository *arXiv* [28]. Papers on *arXiv* are classified into different sub-categories based on their topic, with some papers assigned to multiple sub-categories. For this research, papers were classified according to the broader category of their first-listed subcategory, and only the papers focusing on the fields of *Computer Science, Mathematics* and *Physics* were selected, for a total of 23,341 abstracts. As with the conferences dataset, abstracts span from 2020 to 2024, and are accompanied by the paper's title, a list of authors, the assigned category, and the year of publication. The following table summarizes the number of abstracts per year and category.

	2020	2021	2022	2023	2024
Computer Science	2455	2679	2495	2512	1604
Mathematics	1114	1067	1009	893	549
Physics	1557	1379	1552	1483	993

Table 3.2: Number of abstracts per year and arXiv category.

For pre-processing, abstracts exceeding 512 tokens were removed to prevent truncation during the embedding process with S-BERT, ensuring that the overall meaning of the text is preserved. For most metrics, the abstracts were analyzed in their original form. However, embeddings like TF-IDF required additional pre-processing steps, including converting the abstracts to lowercase, removing numbers, punctuation, and stopwords, and applying lemmatization to reduce words to their root form.

3.2 Adjectives and adverbs disproportionately used by AI

Two additional datasets were created based on lists of the top 100 adjectives and top 100 adverbs disproportionately used by AI (Liang et al., 2024 [18]). The assumption is that abstracts containing a higher percentage of these terms are more likely to have been written with the aid of LLMs, therefore an analysis of the chosen metrics performed on such abstracts should result in a stronger signal of LLMs' impact on academic writing. To test this, for each year and conference/category, the 100 abstracts with the highest percentages of these terms were selected, obtaining two smaller datasets containing 1500 abstracts each.

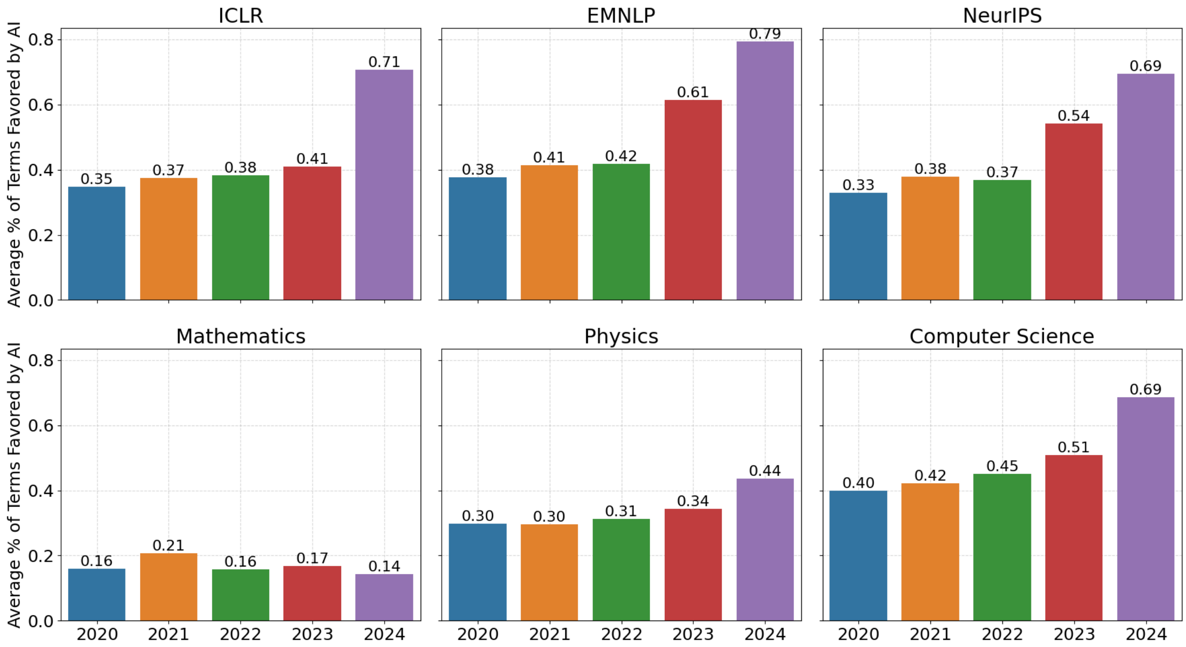

Figure 3.1: Average percentage of adverbs and adjectives favored by AI over time.

Furthermore, tracking the percentage of these adjectives and adverbs over time reveals a notable and consistent increase, particularly in 2023 and 2024, across all categories except Mathematics, as shown by Figure 3.1. This suggests a stronger adoption of LLMs in research writing over the past two years. These findings also align with Geng and Trotta (2024) [19], whose frequency-based analysis of arXiv papers indicated that the Mathematics field was less affected by the diffusion of LLMs, while Computer Science was affected the most.

4

Metrics & Methods

This chapter provides a detailed overview of the metrics chosen to evaluate the impact of LLMs on academic writing, with a focus on text similarity and writing style. It defines each metric, explains why it was selected for the analysis, and discusses how it contributes to shaping different stylistic aspects of a text, particularly in research paper abstracts. Additionally, the chapter outlines the implementation of these metrics, including the specific computational methods and tools used. Understanding these factors will serve as the foundation for gaining deeper insights into how LLM-generated text differs from human writing, and the extent to which LLMs can influence stylistic elements in scientific publications.

4.1 Text similarity

In Natural Language Processing (NLP), text similarity refers to the measurement of how similar or different two pieces of text are. This concept has a wide range of applications, including plagiarism detection, question answering, and machine translation. Text similarity can be categorized into two main types:

- Lexical similarity: Measures how close two texts are at the surface level, focusing on common words or characters while ignoring their meaning.
- Semantic similarity: Assesses how similar two texts are in terms of meaning, regardless of the specific words used.

This research aims to determine whether the widespread adoption of LLMs in academia for tasks such as drafting, editing, or generating abstracts affects their similarity to one another. The concern behind this investigation is the potential homogenization of academic writing, which could result in a loss of diversity and originality. LLMs like ChatGPT may favor certain terms, phrases [2], or syntactic structures, leading to a more standardized language. This tendency might cause different abstracts to appear very similar, especially if they belong to papers sharing the same topic or research area. It's important to note, however, that standardization is not always negative. In fact, it could serve as a means to make research more accessible, particularly for non-native English speakers. By analyzing both lexical and semantic similarity, this study investigates whether abstracts written after the recent AI boom show a trend of increased uniformity in vocabulary and overall meaning.

Text similarity is a fundamental concept in many NLP applications, and as such, various approaches to the problem have been developed. Typically, text data is embedded into high-dimensional vectors, and the similarity between these vectors is then computed using metrics such as cosine similarity or Euclidean distance.

Cosine similarity is a technique used to measure how similar two vectors are by calculating the cosine of the angle between them. It can range from -1 to 1, with -1 indicating completely dissimilar vectors and 1 indicating identical vectors. This metric is widely used because it is not affected by the length of the vectors, making it robust across a variety of text lengths and structures. In contrast, Euclidean distance is computed as the square root of the sum of squared differences between corresponding elements of the vectors, with larger distances indicating lower similarity. This metric is sensitive to vector length, meaning that if a term appears more frequently in one text, it might mean it's more important or simply that the text is longer. While both metrics have their uses, this research adopts cosine similarity as the primary measure for text similarity due to its effectiveness in normalizing for vector length and its consistent performance across various types of text.

To represent text as vectors, various word and sentence embeddings are commonly used in NLP. In this work, TF-IDF was chosen for lexical similarity, due to its emphasis on word frequency and its low computational demands, making it a practical option for this analysis. As for semantic similarity, S-BERT, a modification of BERT, was selected as it represents the state of the art in capturing complex semantic relationships between words and phrases, providing embeddings that offer a deeper understanding of text beyond surface-level word matching.

4.1.1 Lexical similarity

A common technique for computing lexical similarity involves embedding pre-processed text, where stopwords are removed and words are lemmatized, using TF-IDF (Term Frequency-Inverse Document Frequency). This method transforms the text into a high-dimensional vector, where each dimension represents a word and is assigned a TF-IDF score, calculated as follows:

$$TF - IDF(t, d, D) = TF(t, d) \times IDF(t, D)$$

Where:

$$TF(t, d) = \left(\frac{\text{Number of times t appears in d}}{\text{Total number of terms in d}} \right)$$

$$IDF(t, D) = \log \left(\frac{\text{Total number of documents in D}}{\text{Number of documents containing t}} \right)$$

In this formula, *D* represents the entire corpus of texts, *d* is a specific document within the corpus, and *t* is a term in the document. The *TF* component assigns higher scores to terms that appear more frequently within a document, while the *IDF* component assigns higher scores to terms that are less common across the entire corpus [29]. By combining them, TF-IDF highlights important terms in a document while penalizing common, uninformative words, such as articles, pronouns, and conjunctions.

While TF-IDF does not capture word semantics as effectively as models like BERT, it remains widely used due to its simplicity and efficiency, especially in scenarios with limited computational resources or smaller datasets. For this study, the texts were embedded using TF-IDF with the Python library `scikit-learn`'s `TfidfVectorizer`, and the resulting scores were computed using the `cosine_similarity` function [30].

4.1.2 Semantic similarity

Semantic similarity measures how close two pieces of text are in terms of meaning, no matter which specific words are used. This approach is essential for understanding the relationships between texts in a more profound way, as it goes beyond the surface-level word matching seen in lexical similarity. While lexical similarity focuses on the presence of common words or characters, semantic similarity evaluates how similar the underlying concepts or meanings are between two texts.

There are several methods for embedding text while preserving its overall meaning, ranging from traditional models like Word2Vec and GloVe, to more advanced Transformer-based ones. Each of these models contributes to understanding the semantics of a text, but Transformer-based models like BERT have become the standard due to their ability to capture complex, nuanced relationships in language [31]. These models excel at understanding word meanings in context, effectively addressing challenges such as polysemy (where a word has multiple meanings) and synonyms (where different words have the same meaning). By considering the surrounding context, they generate richer, more nuanced representations of words. However, generating these meaningful embeddings comes at the cost of significantly more computational power compared to traditional models. To measure semantic similarity between abstracts, this research employs S-BERT (Sentence-BERT) [32], an extension of the Transformer-based model BERT.

BERT (Bidirectional Encoder Representations from Transformers) uses the encoder component of a Transformer architecture to generate attention-based word vectors. One of BERT's key strengths is its bidirectional nature, meaning it considers the context of a word from both the left and right sides within a sentence. BERT was pre-trained on a massive corpus of 3.3 billion words. The training process is designed to prevent the model from simply memorizing words through two tasks:

- Masked Language Modeling (MLM), where the model predicts randomly masked words in a sentence.
- Next Sentence Prediction (NSP), where the model determines whether two sentences are consecutive or not.

These tasks allow BERT to learn the relationships between words and sentences in a deeper, more context-aware manner [31].

S-BERT is a modified version of BERT designed specifically for efficiently and accurately computing semantic similarity scores between sentences. Unlike traditional models that focus on individual word embeddings, S-BERT uses Siamese networks and is trained on pairs of sentences with labeled similarity scores. This architecture generates fixed-sized sentence embeddings that place semantically similar sentences closer together in the vector space. This makes S-BERT particularly well-suited for tasks like semantic textual similarity, where understanding the meaning behind entire sentences is key [32].

Unlike lexical similarity methods, which are typically based on word-level comparisons and often exclude stopwords and punctuation, embedding techniques using pre-trained models like S-BERT work better when applied directly to the original text. These models leverage the full context, including elements like stopwords and punctuation, to gain a more complete understanding of the text and produce more meaningful representations.

For this study, S-BERT was implemented using the Python module `sentence-transformers`, with the specific model `paraphrase-MiniLM-L6-v2` [33] chosen for its balance between speed and accuracy in generating sentence embeddings. Once again, the similarity scores were computed using the `cosine_similarity` [30] function.

4.2 Writing Style

Previous studies have shown that LLMs exhibit a distinct writing style [19, 20], indicating that their influence on academic writing may go beyond simply increasing text similarity. Writing style, at its core, is the way ideas are conveyed through language. It goes beyond grammatical correctness, encompassing word choice, sentence and paragraph structure, and more. The purpose of writing style is not simply to adhere to conventions but to communicate a message in a manner that is clear, engaging, and convincing.

In literature and communication, writing styles generally fall into four main categories: persuasive, narrative, expository, and descriptive. Persuasive writing aims to convince the reader of a particular position or opinion through evidence. Narrative writing tells a story, bringing characters, settings, and conflicts to life. Expository writing is used to inform an audience by presenting facts, statistics, and analyses in a clear and straightforward manner. Descriptive writing, on the other hand, uses sensory details and literary techniques to paint a vivid image in the reader's mind [34].

To gain a comprehensive understanding of how LLMs' unique writing style can influence academic writing, whether through drafting, editing or text generation, this study examines several key metrics, divided into broader categories. Lexical features include basic statistics, such as the number and average length of words, and the prevalence of different parts of speech, as well as vocabulary richness metrics and an analysis of word frequency. Syntactic features also include basic statistics, such as the number and average length of sentences, along with measures of syntactic complexity and referential cohesion. The cognitive complexity of abstracts is measured through readability scores and entropy, while an analysis of verbs' voices, tenses and moods provides insights on tone and effectiveness of a text.

4.2.1 Lexical features

Words are the building blocks of any text, and the choice and usage of these words are crucial in shaping a piece of writing. For this reason, the first category of writing style metrics this research focuses on includes word-level statistics, as well as measures of lexical diversity and repetitiveness.

Basic lexical statistics

Statistical metrics can be used to analyze the structural features of text, offering an overview of its composition. These include measures such as the total number of words and word length, measured here in syllables per word.

The length of an abstract is a key factor in determining the level of detail. A longer abstract allows for a more comprehensive summary of the content of a paper, providing sufficient background, methodology, and a preview of the research results. This gives readers a clear sense of what to expect in the full text. On the other hand, excessively long abstracts might risk becoming too information-dense and potentially overwhelm the reader. Another important indicator is word length. The use of longer, more complex words, often signals a higher level of technicality or formality in the writing. In academic abstracts, the use of polysyllabic words can reflect the adoption of specialized vocabulary, contributing to the professional tone of the text [35]. However, while the use of such vocabulary may make the abstract appear more authoritative or scholarly, it can also affect readability, especially for a broader audience.

The prevalence of different parts of speech, whether noun, verb, adjective, adverb, or connective, plays a significant role in how information and ideas are being presented. The careful selection and arrangement of words contribute to the overall flow of a text, ensuring that complex information is communicated clearly while maintaining the necessary level of formality and precision.

Nouns represent people, objects, concepts, and ideas. Verbs are central to expressing action or states, while adjectives provide descriptive details that modify nouns, adding specificity and nuance. Adverbs modify verbs, adjectives, or other adverbs, offering additional details about how, when, where, or to what degree an action occurs. Finally, connectives serve a vital role in linking ideas and shaping the logical flow of sentences. They help establish relationships between different parts of the text, indicating contrasts, showing cause and effect, or signaling the sequence of actions [36].

In academic writing, using precise and specific nouns is essential for clarity, as they define the key elements being discussed. However, excessive use of overly technical nouns can make a text too information-dense and difficult to understand, particularly if these terms are not defined or explained clearly. Verbs and adjectives also contribute to efficiently and precisely communicating ideas, but overusing adjectives can clutter the text and distract the reader from the main points. Adverbs can convey subtler meanings but excessive use, especially when they don't add critical information, can make the writing feel repetitive or imprecise. The use of connectives strengthens the logical expression of sentences, adds depth and organization to the presentation of research, but may, at times, increase textual complexity if overused [35]. Effective writing requires conciseness; avoiding unnecessary words while maintaining sufficient detail ensures that each word contributes to the text meaningfully [37].

These indicators not only offer insights into the structure and tone of the abstract but also highlight how various parts of a text contribute to its overall strength. The analysis of such indicators was performed using the Python library `spaCy` [38], with model `en_core_web_lg`, chosen for its balance between speed and accuracy. SpaCy is a powerful and widely used tool for natural language processing, designed to facilitate various linguistic tasks, including tokenization, part-of-speech tagging, dependency parsing, and syntactic analysis [39].

Vocabulary richness

Vocabulary richness refers to the variety and complexity of words used in a text. A higher vocabulary richness suggests a more sophisticated or technical style, while a lower richness may indicate simplicity or repetition.

A key measure of vocabulary richness is lexical diversity, which quantifies the proportion of unique words in a text. One common way to assess lexical diversity is the Type-Token Ratio (TTR), calculated as the ratio between the number of unique words (types) and the total number of words in a text (tokens):

$$TTR = \left(\frac{\text{Number of types}}{\text{Number of tokens}} \right)$$

The scores range from 0 to 1, with higher TTRs indicating greater lexical diversity, often found in more complex texts. Conversely, a lower TTR suggests simpler or more repetitive writing. One important consideration to make is that TTR can be sensitive to text length: shorter texts tend to have higher TTRs since they offer fewer opportunities for word repetition [40]. However, since this study focuses on abstracts, which generally have a consistent length, TTR remains a valid metric for analysis.

Another measure of vocabulary richness is Yule's K-Complexity, which has the advantage of being nearly independent of text length [41]. It is calculated using the formula:

$$K = 10^4 \times \left[-\frac{1}{N} + \sum_{i=1}^{V} V(i, N) \left(\frac{i}{N} \right)^2 \right]$$

where:

- N is the total number of words in the text,
- V is the number of unique words, and
- $V(i, N)$ represents the number of words that appear i times in the text.

This measure is lower-bounded by 0, which occurs when all words in the text are unique ($V(1,N)$ = N). However, it does not have an upper bound. In general, higher K values indicate lower vocabulary richness, meaning the text is more repetitive [42].

For this research, both TTR and K-Complexity were implemented using the Python library `LexicalRichness` [43].

Additionally, this study investigates repetitiveness as another aspect of vocabulary richness, using gzip complexity. This measure is defined as:

$$g = \frac{s_{raw} - s_{compressed}}{s_{raw}}$$

where *raw* is the size of the raw text, and *compressed* is the size of the text after compression using *gzip*. If a text is highly repetitive, g is close to 1, since the compressed version is significantly smaller than the raw text, while lower values of g indicate less repetition and greater vocabulary richness. It's important to note, however, that g can be negative for very short texts, where compression may actually increase the file size [44].

Word frequency

Often accompanying the study of vocabulary richness, word frequency provides valuable insights into the structure of languages and has a significant impact on the writing style of a text. In natural language, word frequency typically follows Zipf's Law, an empirical principle that describes an inverse relationship between word rank and frequency in a given text or corpus [45]. In simple terms, the most frequent word (often a function word like "the" or "and") will appear roughly twice as often as the second-most frequent word, three times as often as the third-most frequent word, and so on. This creates a statistical pattern where the majority of words are used only once or very infrequently. This distribution is commonly observed in both written and spoken language, and reflects the principle of least effort in human communication [46].

However, word frequencies are not stationary, meaning that the probability of using a word depends on various factors, such as the topic of discussion [47]. Studies have shown that a word's features and the context in which it is used contribute to its frequency, particularly in relation to the desire to convey information or align with a specific social group [48]. Additionally, frequency can be influenced by social, technological, and political factors [49].

Word frequency is also historically linked to the concept of readability. Thorndike's *Teacher's Word Book* (1921), the first extensive listing of English words based on frequency, provided teachers with a way to measure the difficulty of words and texts, laying the foundation for much of the readability research that followed. The book was based on the idea that the more frequently a word is used, the more familiar it is, making it easier to understand and use [50]. The fundamental role of word frequency in readability is evident: humans not only use certain words much more often than others, but they also recognize these frequent words more quickly, favor them, and learn them more easily [51]. As a result, the degree to which the word frequency of a text follows Zipf's Law directly impacts writing style, particularly in terms of lexical diversity and readability. In casual or conversational writing, there is often a higher repetitiveness of common, everyday words that dominate the frequency distribution. These frequent words, such as pronouns, prepositions, and conjunctions, make up a large portion of the text, making it easier to follow, as readers encounter familiar terms more often. In contrast, formal or academic writing tends to include a higher proportion of varied and specialized vocabulary. Consequently, the word frequency curves for such texts usually show a more even distribution, with a greater presence of lower-frequency terms compared to informal writing.

This research not only analyzes the general word frequency of the collected abstracts, but also tracks the percentage of adjectives and verbs most frequently used by AI, based on the lists proposed in Liang et al. (2024) [18]. These linguistic features serve as a primary indicator of the impact of large language models on academic writing, and may provide a way to identify abstracts that have been drafted, edited, or entirely generated by AI.

4.2.2 Syntactic features

Sentences organize words to express a complete thought, often in the form of a statement, question, command, wish or exclamation. The second category of writing style metrics in this research focuses on sentence-level statistics, measures of syntactic complexity, and cohesion across sentences.

Basic syntactic statistics

Statistical measures can analyze the syntactic dimension in addition to the lexical one, including the number of sentences and sentence length, measured here by the number of words per sentence. These features reflect the overall organization and pacing of writing, significantly influencing tone, clarity, and effectiveness in academic texts.

The considerations previously made about abstract length naturally apply to the number of sentences as well. Additionally, sentence length can play a crucial role in defining the tone of an abstract. Longer sentences tend to suggest a more formal, academic style, as they are commonly associated with complex ideas and detailed explanations. This style is typical in scientific and technical writing, where the information density is high and precision is necessary. On the other hand, shorter sentences typically result in a more direct and conversational tone, which can make the text feel more accessible but may risk oversimplification.

Syntactic complexity

Syntactic complexity refers to the grammatical sophistication of sentences, encompassing various features of sentence construction that contribute to how difficult or easy a text is to understand. It is commonly assessed using a range of metrics that capture different aspects of sentence structure, such as length of production unit, overall sentence complexity, amount of subordination, and amount of coordination [52].

To better understand these concepts, it's essential to clarify the role of clauses in syntax. A clause is a group of words that typically contains both a subject and a predicate (the verb), forming a unit of meaning within a sentence. An independent clause can stand alone as a complete sentence, expressing a complete thought, while a dependent clause provides additional information but cannot function independently as a sentence. Subordination refers to the relationship between an independent clause and its associated dependent clauses, while coordination involves joining two or more independent clauses of equal importance using coordinating conjunctions (for, and, nor, but, or, yet, so) [36]. Furthermore, the combination of a main clause and its dependent clauses is defined as a T-unit, which is the smallest grouping of words that can be classified as a sentence [53].

For the abstracts analyzed in this research, the length of the production unit is measured by the previously mentioned average sentence length. The overall sentence complexity is represented by the average clause density, which indicates how many clauses (independent and dependent) are packed into each sentence. The degree of subordination is given by the average number of dependent clauses per T-unit, while the amount of coordination is indicated by the average number of T-units per sentence.

Typically, abstracts are concise summaries of research papers and their structural complexity often increases due to the need to condense a large amount of information into a small piece of text, packing multiple ideas into each sentence. The selected metrics can indicate whether an abstract is syntactically dense even when it contains only a few sentences: for example, it might have a low sentence count but feature long sentences and a high ratio of dependent to independent clauses, making its structure particularly dense and complex. This, combined with high clause density, suggests a writing style that conveys substantial information through relatively few, yet syntactically sophisticated, sentences.

The analysis of the syntactic complexity of the abstracts in this study was conducted using the Python libraries `spaCy` (with model `en_core_web_lg`) and `syllapy` [38, 54].

Referential Cohesion

Cohesion measures how well-connected different parts of a text are in terms of meaning, based on the consistent and strategic use of grammatical or lexical devices that link concepts together. One way to tie two concepts together within a text is through reference, where words have to be interpreted semantically not in their own right, but in relation to the concepts they refer to [55]. Strong referential cohesion enhances clarity, helping readers understand the relationships between ideas and ensuring smooth transitions between sentences, paragraphs, or sections.

In scientific writing, referential cohesion is essential for clarity and readability. It allows the readers to easily track and relate various concepts and findings, which is particularly important in abstracts, where information needs to be condensed into a short yet coherent summary. The cohesion in such texts is often measured by examining repeated references to concepts, objects, or entities across sentences, which helps ensure the text is unified and the ideas are logically connected. Some key measures of referential cohesion include noun overlap, argument overlap and stem overlap.

- Noun overlap occurs when consecutive sentences share some of the same nouns, reinforcing conceptual continuity.
- Argument overlap tracks whether consecutive sentences contain nouns with the same lemma, indicating a shared reference even if different word forms are used.

- Stem overlap assesses whether consecutive sentences share any lemmas, ensuring consistent reference to key concepts without unnecessary repetition [56].

This study measures referential cohesion by calculating these metrics as percentages of the total number of consecutive sentence pairs. The analysis was conducted using the Python library `spaCy` [38] with model `en_core_web_lg`.

4.2.3 Cognitive complexity

Cognitive complexity refers to the mental effort needed to process and understand a text. It is influenced by factors such as sentence structure, word choice, information density, and referential cohesion. This research examines cognitive complexity through two key metrics: readability scores and entropy.

Readability

Readability evaluates how easily a text can be understood based on its writing style [57], or the degree to which a specific group of people finds a particular text engaging and understandable [58], highlighting the connection between the text and its readers. Commonly, readability is assessed using formulas that take into account factors like word difficulty, syllable count, sentence length and complexity, and overall text structure. These formulas offer several advantages, including simplicity, availability, and the ability to estimate difficulty levels without prior reading. This helps writers tailor their work to their audience, saving time and effort. However, they also have limitations, such as their inability to measure actual comprehension and the inconsistencies that arise when different formulas yield varying results for the same text [59]. This study focuses on three commonly used metrics to evaluate readability: the Flesch Reading Ease Score, the Gunning Fog Index, and the SMOG Index.

Score	US School level	Interpretation
100.0-90.0	5th grade	Very easy to read. Understood by approximately 93% of US adults.
90.0-80.0	6th grade	Easy to read. Understood by approximately 91% of US adults.
80.0-70.0	7th grade	Fairly easy to read. Understood by approximately 88% of US adults.
70.0-60.0	8th and 9th grade	Standard English. Understood by approximately 83% of US adults.
60.0-50.0	10th to 12th grade	Fairly difficult to read. Understood by approximately 54% of US adults.
40.0-30.0	College	Difficult to read. Understood by approximately 33% of US adults.
30.0-0.0	College graduate	Very difficult to read. Understood by approximately 4.5% of US adults.

Table 4.1: Flesch Reading Ease Score.

The Flesch Reading Ease Score [60] is the most widely used readability formula. It calculates readability based on average sentence length, measured as the number of words per sentence, and the average number of syllables per word, using the following formula:

$$FRE = 206.835 - 1.015 \times \left(\frac{\text{Total words}}{\text{Total sentences}}\right) - 86.4 \times \left(\frac{\text{Total syllables}}{\text{Total words}}\right)$$

Scores generally range from 0 to 100, with higher scores indicating greater readability. Standard passages typically fall within the 60–70 range. Table 4.1 provides an interpretation of these scores [61].

Index	**US School level**	**Reading ease**
17	College graduate	Difficult
16	College senior	Difficult
15	College junior	Difficult
14	College sophomore	Difficult
13	College freshman	Difficult
12	High school senior	Average
11	High school junior	Average
10	High school sophomore	Average
9	High school freshman	Average
8	Eighth grade	Easy
7	Seventh grade	Easy
6	Sixth grade	Easy

Table 4.2: Gunning Fog Index.

The Gunning Fog Index [62], known for its ease of use, estimates the years of formal education required to understand a piece of text on a first reading. It takes into account average sentence length and the percentage of complex words, where complex words are defined as those with three or more syllables. The formula is as follows:

$$GF = 0.4 \times \left(\frac{\text{Total words}}{\text{Total sentences}} + \frac{\text{Complex words}}{\text{Total words}} \times 100 \right)$$

A score of 9-12 suggests that the text is comprehensible to high school students, while scores of 17 or higher indicate a college-level difficulty.

Finally, the SMOG Index (Simple Measure of Gobbledygook) [58] is a readability metric developed as a more accurate and more easily calculated substitute for the Gunning Fog Index. It is calculated using the following formula:

$$SMOG = 1.043 \times \sqrt{\text{Number of polysyllables} \times \frac{30}{\text{Number of sentences}}} + 3.1291$$

This index focuses on the number of polysyllabic words (words with three or more syllables) in a text, making it particularly effective for evaluating the readability of academic and professional writing. Higher SMOG scores indicate more complex texts that require advanced reading skills.

For this study, readability scores were computed using the Python library `textstat` [63], which automates the calculation of various readability metrics, including the ones just described.

Entropy

Entropy is a statistical measure that quantifies the unpredictability or randomness of a text, drawing from information theory to assess how evenly characters are distributed throughout a piece of writing. Texts with higher entropy are often longer, they feature a broad range of vocabulary, varied sentence structures, and topics covered, resulting in less predictability and higher information density. Conversely, lower entropy suggests a more formulaic or predictable pattern in word selection, where certain words or phrases are repeated more frequently, leading to less variation and lower information density.

Entropy is often calculated based on the frequency distribution of characters in a text, and is usually expressed in bits. The formula for entropy is typically:

$$H = -\sum_{i=1}^{n} p_i \log_2 (p_i)$$

where p_i is the probability of the *i*-th character occurring in the text, and n is the total number of unique characters [64]. By analyzing entropy, it's possible to gain insights into the stylistic features of a text, particularly in terms of its unpredictability and information content.

4.2.4 Verb forms

The fourth and final category of metrics focuses on verbs and how the choice of specific tenses, moods, and voices serves as a strong indicator of writing style. For this reason, an analysis of the percentages of these different verb forms was included in this study to better understand how LLMs influence academic writing. The choice of verbs significantly affects tone, clarity, formality, and the overall effectiveness of academic communication, making it a crucial aspect to examine when assessing how LLMs generate or modify academic content. For example, a frequent use of the past tense may indicate a historical or narrative style, while the present tense is commonly employed in academic or journalistic writing due to its immediacy. A verb's voice impacts the directness of the text, and the choice between indicative, subjunctive, or imperative moods influences tone and reader engagement. By evaluating verb tense, mood, and voice, it is possible to gain insights into the linguistic patterns produced by LLMs and how they differ from those of human writing.

The use of tense in any type of writing shapes the temporal framework of the text. While some languages do not have tenses, many others primarily feature three main tenses: present, past, and future. These tenses play a crucial role in establishing the timing of events and actions within a narrative, allowing readers to understand when occurrences take place and how they relate to one another.

- Present Tense: This tense imparts a direct and active tone to writing, conveying a sense of immediacy and relevance. It is often employed to describe general truths, established facts, or ongoing actions. Therefore, in academic writing, it is ideal for presenting current knowledge, theories, or results.
- Past Tense: The past tense is typically used for reporting completed actions or findings. In research contexts, it is the most common tense for describing methods, results, and experiments that have already taken place, setting a more definitive tone.

- Future Tense: Future tense is used to discuss predictions, proposals, or potential future research. It introduces an anticipatory tone, focusing on upcoming developments or hypotheses that are yet to be explored [65].

The mood of a verb influences how the writer conveys certainty, suggestion, or possibility, allowing speakers to express their attitude toward what they are saying. While some languages have as many as sixteen moods, the primary moods in English are indicative, imperative, and subjunctive.

- Indicative Mood: The indicative mood is the most common in academic writing, used to present facts, describe actions, or make statements about reality. It is the default mood for straightforward, factual communication, helping to assert conclusions or observations.
- Imperative Mood: The imperative mood is used to issue commands, suggestions, or instructions. It is direct and often found in procedural or instructional writing, such as in methods sections or guidelines.
- Subjunctive Mood: The subjunctive mood is used for expressing hypothetical situations, recommendations, or desires. It is often used in academic writing when discussing potential outcomes, proposing ideas, or expressing uncertainty [66].

The voice of a verb describes the relationship between the verb itself and its arguments, the subject and the object. The choice between active and passive voice has a significant impact on the tone and emphasis of the writing.

- Active Voice: In the active voice, the subject performs the action. Active voice is typically more concise, direct, and easier for readers to follow. It is commonly used in academic writing to highlight the person or entity responsible for an action or finding.
- Passive Voice: Passive voice focuses on the action itself. In academic writing, passive constructions are often used to emphasize the process, results, or object of study, rather than the researchers or authors. It provides a more impersonal, objective tone, but excessive use can lead to vagueness or overly complex sentence structures [66].

The verb analysis in this research was performed with the aid of Python library `spaCy` [38] with model `en_core_web_lg`. Specifically, spaCy's part-of-speech tags, dependency parsing, and morphological features were combined with common heuristics to classify verbs by tense, voice, and mood. While these heuristic methods are widely used and offer a practical solution for verb analysis, they may not capture all nuances and variability of language use, potentially leading to inaccuracies. Future research should explore more sophisticated or machine learning–based approaches to further assess the impact of LLMs on verb form choices in academic writing.

5
Analysis & Results

This chapter provides a detailed description of the experiments conducted to assess the impact of LLMs on academic writing. The experiments evaluate the set of metrics introduced in the previous chapter, with a focus on how they change across different years or AI-generated revisions. The results are thoroughly analyzed, highlighting their significance and implications for the future of academic writing.

5.1 LLM-generated revisions of existing abstracts

The first experiment had the objective of understanding how LLMs impact academic writing, based on the assumption that most researchers and academics who use LLMs for assistance primarily do so for editing. To facilitate comparison, only abstracts from 2020 and 2024 were selected from both datasets. The abstracts from 2020 were selected because they were written during a time when LLMs were not yet common, making them more likely to be written exclusively by humans. In contrast, the abstracts from 2024 have a higher probability of being AI-assisted, given the widespread adoption of LLMs in academia following the launch of ChatGPT in 2022.

Detecting Interactions from Neural Networks via Topological Analysis (2020) by Liu et al.	**SwapTransformer: Highway Overtaking Tactical Planner Model via Imitation Learning on OSHA Dataset (2024) by Shamsoshoara at al.**
Detecting statistical interactions between input features is a crucial and challenging task. Recent advances demonstrate that it is possible to extract learned interactions from trained neural networks. It has also been observed that, in neural networks, any interacting features must follow a strongly weighted connection to common hidden units. Motivated by the observation, in this paper, we propose to investigate the interaction detection problem from a novel topological perspective by analyzing the connectivity in neural networks. Specially, we propose a new measure for quantifying interaction strength, based upon the well-received theory of persistent homology. Based on this measure, a Persistence Interaction Dection (PID) algorithm is developed to efficiently detect interactions. Our proposed algorithm is evaluated across a number of interaction detection tasks on several synthetic and real-world datasets with different hyperparameters. Experimental results validate that the PID algorithm outperforms the state-of-the-art baselines.	This paper investigates the high-level decision-making problem in highway scenarios regarding lane changing and over-taking other slower vehicles. In particular, this paper aims to improve the Travel Assist feature for automatic overtaking and lane changes on highways. About 9 million samples including lane images and other dynamic objects are collected in simulation. This data; Overtaking on Simulated HighwAys (OSHA) dataset is released to tackle this challenge. To solve this problem, an architecture called SwapTransformer is designed and implemented as an imitation learning approach on the OSHA dataset. Moreover, auxiliary tasks such as future points and car distance network predictions are proposed to aid the model in better understanding the surrounding environment. The performance of the proposed solution is compared with a multi-layer perceptron (MLP) and multi-head self-attention networks as baselines in a simulation environment. We also demonstrate the performance of the model with and without auxiliary tasks. All models are evaluated based on different metrics such as time to finish each lap, number of overtakes, and speed difference with speed limit. The evaluation shows that the SwapTransformer model outperforms other models in different traffic densities in the inference phase.
Identifying statistical interactions among input features is crucial yet challenging. While recent work extracts interactions from trained neural networks, we propose a novel topological approach. Observing that interacting features exhibit strongly weighted connections to shared hidden units, we leverage persistent homology to quantify interaction strength. This yields a new algorithm, Persistence Interaction Detection (PID), which efficiently identifies interactions. We evaluate PID on diverse synthetic and real-world datasets, demonstrating superior performance to state-of-the-art methods across various hyperparameter settings.	This paper addresses high-level decision-making for automated highway lane changing and overtaking. We introduce SwapTransformer, a novel imitation learning architecture trained on the OSHA dataset, a newly released collection of 9 million simulated highway driving samples including lane images and dynamic object data. To enhance environmental understanding, SwapTransformer incorporates auxiliary tasks predicting future vehicle positions and distances. We evaluate SwapTransformer against MLP and multi-head self-attention baselines using metrics such as lap time, number of overtakes, and speed deviation from the limit, demonstrating superior performance across varying traffic densities.
Identifying statistical feature interactions is crucial but challenging. Existing methods often rely on post-hoc analysis of trained neural networks. We introduce Persistence Interaction Detection (PID), a novel topological approach that leverages persistent homology to directly quantify interaction strength from the network's hidden layer connectivity. PID efficiently identifies interactions by analyzing the weights of connections to shared hidden units. Experiments on diverse synthetic and real-world datasets demonstrate PID's superior performance compared to state-of-the-art methods across a range of hyperparameters.	This paper presents SwapTransformer, a novel imitation learning architecture for automated highway lane changing and overtaking. Trained on the newly released OSHA dataset (9 million simulated driving samples), SwapTransformer leverages auxiliary tasks predicting future vehicle positions and distances to enhance environmental awareness. Evaluation against MLP and multi-head self-attention baselines, using metrics such as lap time, overtaking frequency, and speed adherence, shows SwapTransformer's superior performance across diverse traffic conditions.
Identifying statistically significant feature interactions is crucial yet challenging. Existing methods often rely on post-hoc analysis of trained neural networks, lacking direct quantification of interaction strength. We introduce Persistence Interaction Detection (PID), a novel topological method that directly quantifies feature interaction strength from the connectivity of a neural network's hidden layers using persistent homology. PID efficiently identifies interactions by analyzing weights of connections to shared hidden units. Experiments on diverse synthetic and real-world datasets demonstrate PID's superior performance compared to state-of-the-art methods, robustly across a range of hyperparameters.	This paper introduces SwapTransformer, a novel imitation learning architecture for autonomous highway lane changing and overtaking. Trained on the extensive OSHA dataset (9 million simulated driving samples), SwapTransformer incorporates auxiliary tasks predicting future vehicle positions and distances to improve its understanding of the driving environment. Comparative evaluations against MLP and multi-head self-attention baselines demonstrate SwapTransformer's superior performance in terms of lap time, overtaking frequency, and speed adherence across varied traffic conditions.

Table 5.1: Abstract samples from the datasets.

The selected abstracts were processed using Google's Gemini, specifically the `gemini-`

`1.5-flash` [67] model, with the prompt: "Revise this abstract." Each abstract underwent three revisions using the same prompt, and all revisions were saved alongside the original abstracts for comparison.

A preliminary review of the abstracts and their revisions, based on a sample reading, revealed several patterns. The revised abstracts are noticeably shorter than the originals, reflecting Gemini's preference for conciseness. The model tends to introduce key concepts earlier, remove unnecessary words, combine sentences to create more compact paragraphs, and adopt a more direct writing style (e.g. "an architecture [...] is designed" versus "We introduce [...] an architecture"). Additionally, the second and third revisions are often very similar, differing only in minor word substitutions, suggesting that most changes occur in the first revision. Table 5.1 shows two examples from 2020 and 2024, illustrating these trends.

Following this preliminary analysis, the metrics discussed in the previous chapter were calculated for each abstract and its revisions. The results across revisions were compared using the Wasserstein distance, a measure of the minimum "cost" required to turn one distribution into another. Specifically, for each metric, WD1 represents the distance between the original abstracts' distribution and the first revisions' distribution, WD2 measures the distance between the first revisions' distribution and the second revisions' distribution, and WD3 reflects the distance between the second revisions' distribution and the third revisions' distribution. Overall, the computed metrics largely confirm the trends identified in the initial analysis.

The first metric analyzed is similarity, both lexical and semantic, computed between consecutive revision steps for each abstract. Specifically, similarity was measured between original abstracts and first revisions, first and second revisions, and second and third revisions. As a result, only WD1 and WD2 are applicable for these metrics. WD1 represents the distance between the distributions of similarity scores computed between the original abstracts and first revisions, and those computed between the first and second revisions. Similarly, WD2 represents the distance between the distributions of similarity scores computed between the first and second revisions, and those computed between the second and third revisions.

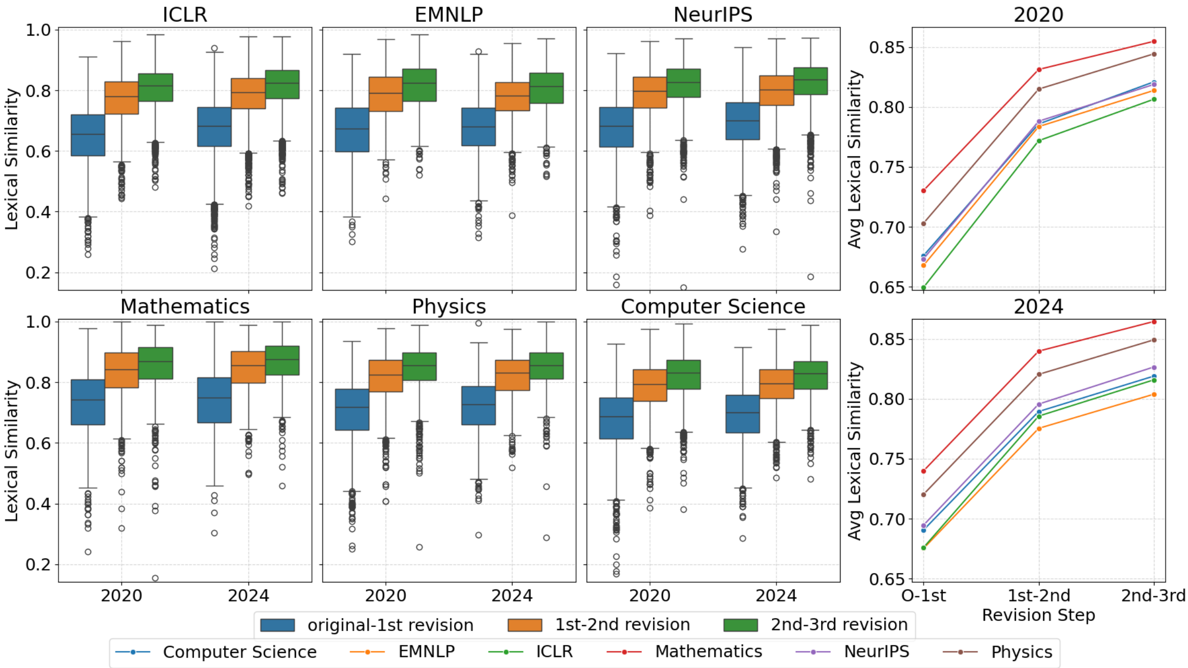

Figure 5.1: Lexical similarity across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	0.1224	0.1157	0.1147	0.1098	0.1009	0.1119
WD1 (2024)	0.1096	0.1002	0.101	0.0988	0.0999	0.1001
WD2 (2020)	0.0346	0.0302	0.0309	0.0351	0.0245	0.0296
WD2 (2024)	0.0302	0.0285	0.031	0.0295	0.0249	0.0295

Table 5.2: Wasserstein distances for lexical similarity across revisions.

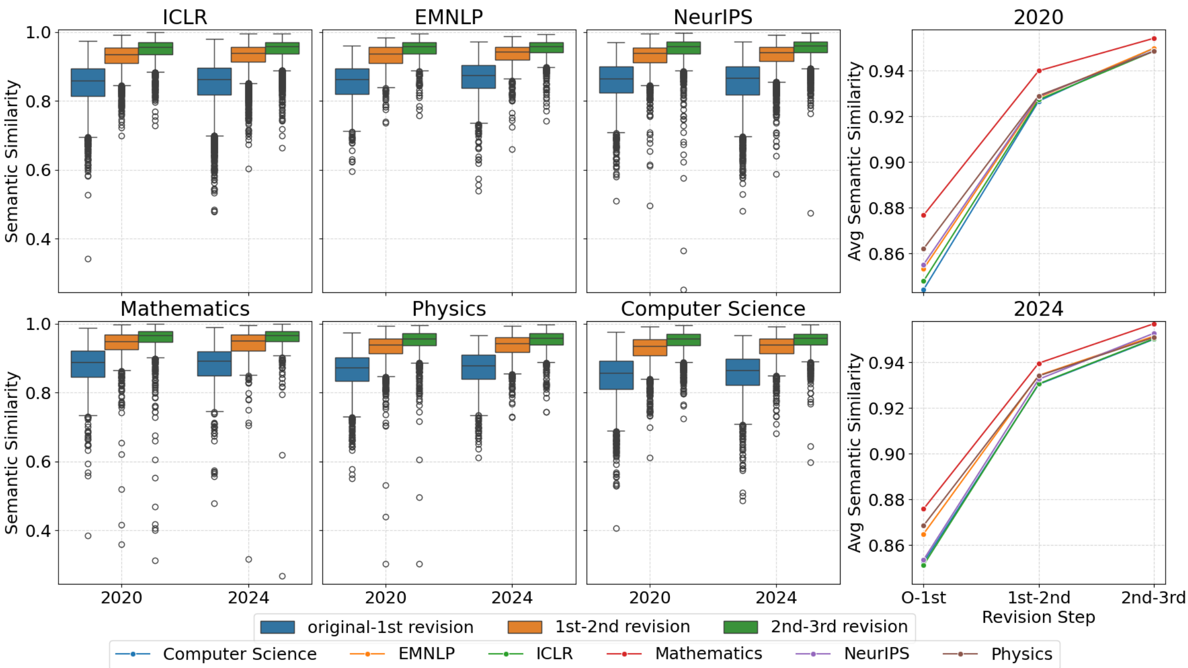

Figure 5.2: Semantic similarity across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	0.0795	0.075	0.0736	0.0826	0.0636	0.0676
WD1 (2024)	0.0794	0.0694	0.0791	0.0779	0.0644	0.0655
WD2 (2020)	0.0212	0.0216	0.0206	0.0232	0.0166	0.0195
WD2 (2024)	0.0197	0.0171	0.0202	0.0199	0.0178	0.0169

Table 5.3: Wasserstein distances for semantic similarity across revisions.

Most similarity scores between original abstracts and their first revision fall within the range of 0.6–0.8 for lexical similarity and a higher 0.8–0.9 for semantic similarity. Both types of similarity increase across revisions, indicating that revisions are more similar to each other than they are to the original abstracts. For both metrics, WD1 is significantly higher than WD2, indicating that most of the changes occur with the first revision. Moreover, WD1 is generally higher in 2020 than 2024. Combined with the fact that average similarity scores tend to be higher in 2024, this suggests that recent abstracts are written in a style that more closely resembles the "LLM writing style," therefore requiring fewer refinements, which could be a sign of LLM use during their production.

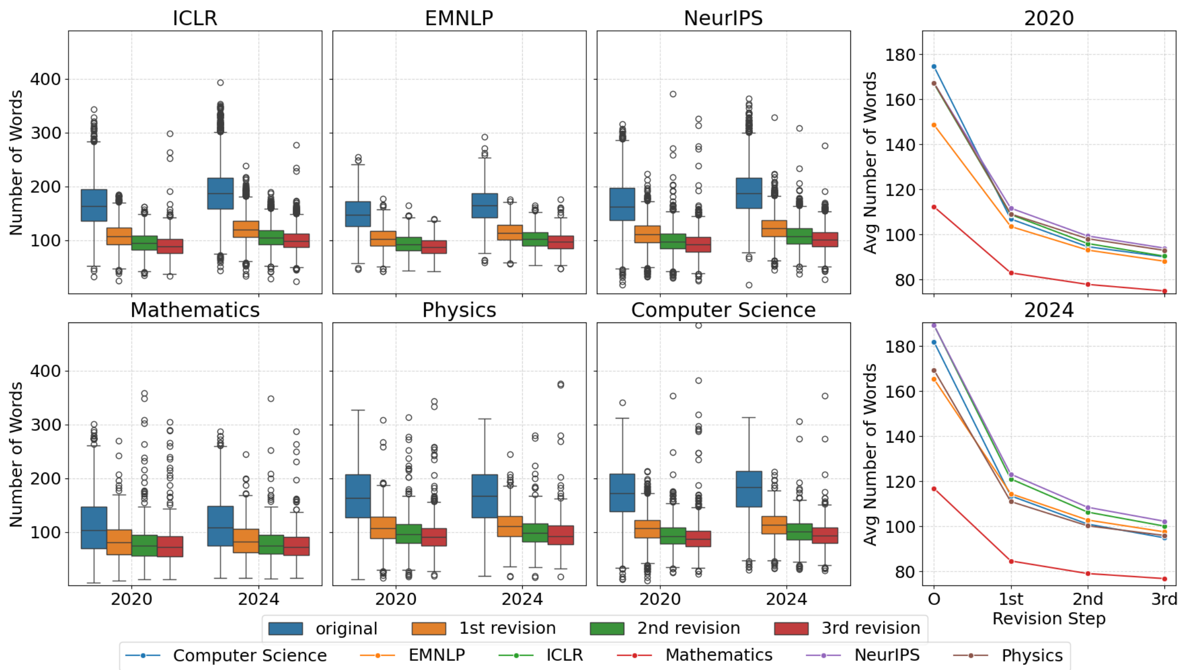

Figure 5.3: Number of words across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	58.2002	45.1784	55.6197	67.7849	29.4578	58.341
WD1 (2024)	68.3538	50.9408	66.2067	68.4613	32.255	58.4914
WD2 (2020)	12.8561	10.4714	12.6603	12.5365	6.6373	11.3911
WD2 (2024)	14.7349	11.6759	14.6947	12.7319	6.3333	11.0524
WD3 (2020)	5.9286	4.9667	5.5774	5.433	3.1373	5.4399
WD3 (2024)	6.2919	5.2942	6.2032	6.1764	2.6812	4.9789

Table 5.4: Wasserstein distances for number of words across revisions.

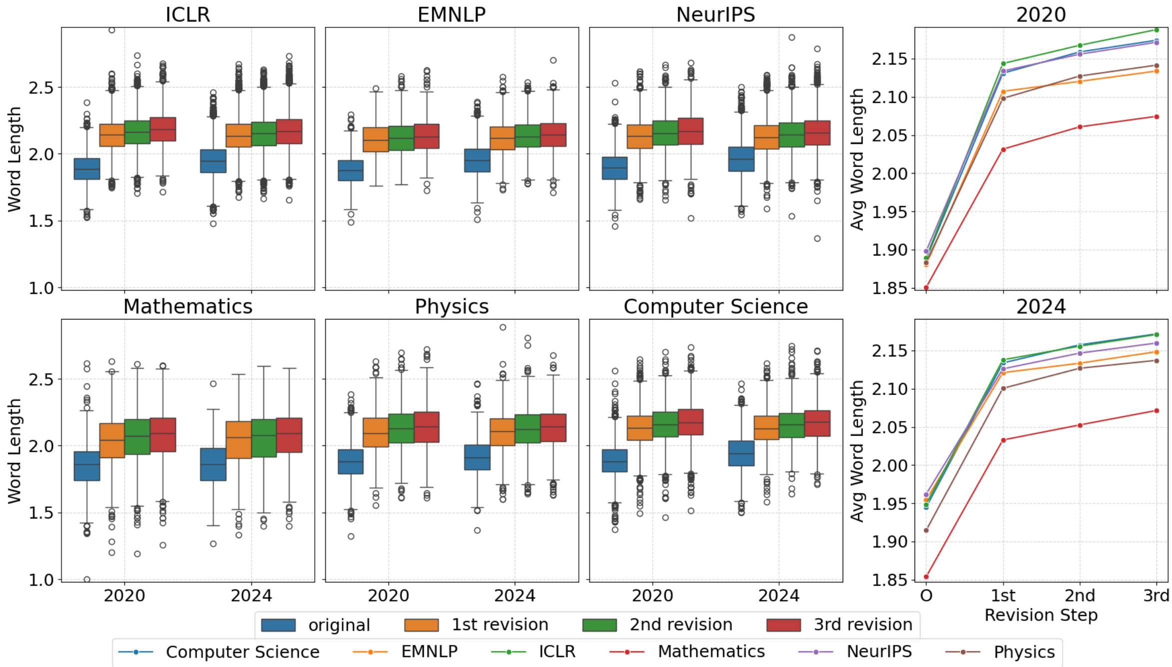

Figure 5.4: Average word length across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	0.2541	0.227	0.2358	0.2433	0.1811	0.2151
WD1 (2024)	0.1892	0.1665	0.164	0.189	0.1791	0.1859
WD2 (2020)	0.024	0.014	0.022	0.0278	0.0294	0.0291
WD2 (2024)	0.0179	0.013	0.0206	0.0237	0.0203	0.0264
WD3 (2020)	0.0206	0.0137	0.0159	0.0154	0.0138	0.0144
WD3 (2024)	0.0152	0.0153	0.0134	0.0148	0.0191	0.0113

Table 5.5: Wasserstein distances for word length across revisions.

The analysis of basic lexical features shows that most abstracts range from 100 to 200 words, with an average word length of 1.75–2 syllables. Both metrics tend to exhibit higher averages in 2024 than in 2020. Across revisions, word count decreases while word length increases, suggesting Gemini favors conciseness and possibly more complex or technical language. WD1 is significantly higher than WD2 and WD3 for both metrics, indicating that changes are most pronounced in the initial revision, while later revisions result only in minor modifications. In 2024, WD1 is larger than in 2020 for word count, whereas word length follows the opposite trend.

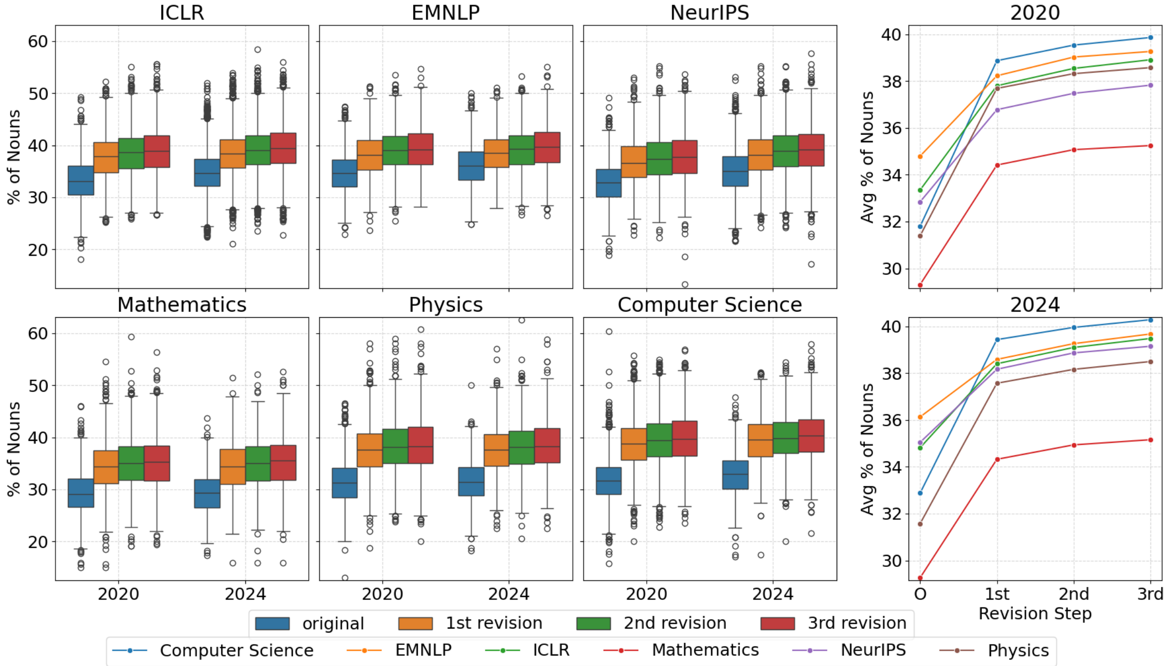

Figure 5.5: Percentage of nouns across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	4.4514	3.428	3.9337	7.0756	5.1024	6.292
WD1 (2024)	3.5854	2.456	3.1189	6.5522	5.0616	5.9921
WD2 (2020)	0.7395	0.7989	0.7041	0.6714	0.6595	0.6335
WD2 (2024)	0.6934	0.6766	0.7031	0.5222	0.6397	0.597
WD3 (2020)	0.3718	0.2574	0.3757	0.3246	0.2301	0.2789
WD3 (2024)	0.3864	0.4314	0.2922	0.336	0.3175	0.3574

Table 5.6: Wasserstein distances for percentage of nouns across revisions.

Regarding the prevalence of different parts of speech, nouns constitute 30–40% of total words in most abstracts, verbs range from 7 to 15%, adjectives from 9 to 15%, adverbs from 2 to 5%, and connectives from 2.5 to 5.5%. Average values are generally higher in 2024 than in 2020, with the percentage of connectives being the exception. While revisions increase the use of nouns, verbs, and adjectives, adverbs and connectives tend to decrease. This shift suggests that revisions are more information-dense and content-focused, with precise terminology possibly replacing adverbial modifiers. However, the reduction in connectives may result in a less cohesive structure, making logical relationships between ideas less explicit. For these metrics, the greatest shifts happen in the first revision as indicated by WD1 being significantly greater than WD2 and WD3. Additionally, WD1 is mostly higher in 2020 than in 2024 for all metrics except the percentage of adverbs.

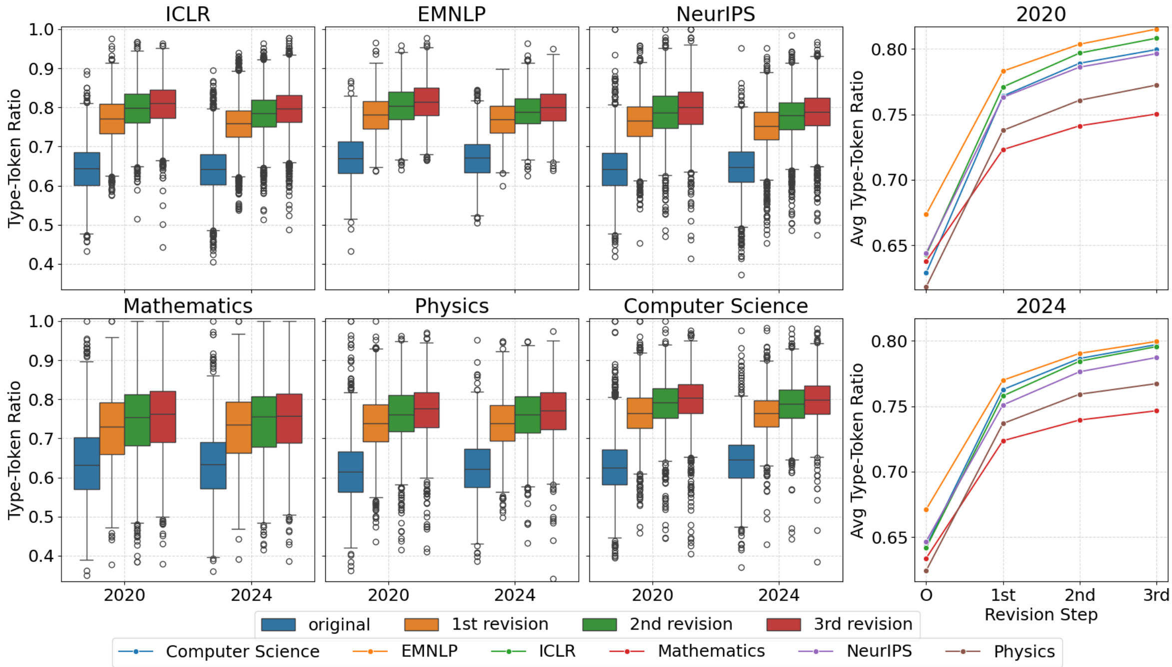

Figure 5.6: TTR across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	0.1281	0.1094	0.119	0.1352	0.0853	0.1196
WD1 (2024)	0.116	0.0987	0.1041	0.1196	0.0899	0.1124
WD2 (2020)	0.0259	0.0206	0.0233	0.0253	0.0189	0.0234
WD2 (2024)	0.0265	0.0205	0.0254	0.0239	0.0175	0.0228
WD3 (2020)	0.0116	0.0115	0.0105	0.0109	0.009	0.0116
WD3 (2024)	0.0112	0.0091	0.0109	0.0107	0.0073	0.0086

Table 5.7: Wasserstein distances for TTR across revisions.

Vocabulary richness increases across revisions according to all computed metrics, indicating that Gemini favors a more varied lexicon with fewer repetitions. Most abstracts have a TTR in the 0.55–0.7 range, Yule's K-complexity between 100 and 200, and gzip complexity between 0.4 and 0.5. The average TTR values for original abstracts do not exhibit a clear increasing or decreasing trend over the years, whereas average K-complexity is higher in 2020, and gzip complexity is higher in 2024. For all metrics, WD1 is much greater than WD2 and WD3, reinforcing once again that the first revision accounts for the majority of the changes. WD1 is higher in 2020 than in 2024, except for Mathematics, suggesting that some abstracts in 2024 may have been generated or edited with LLMs.

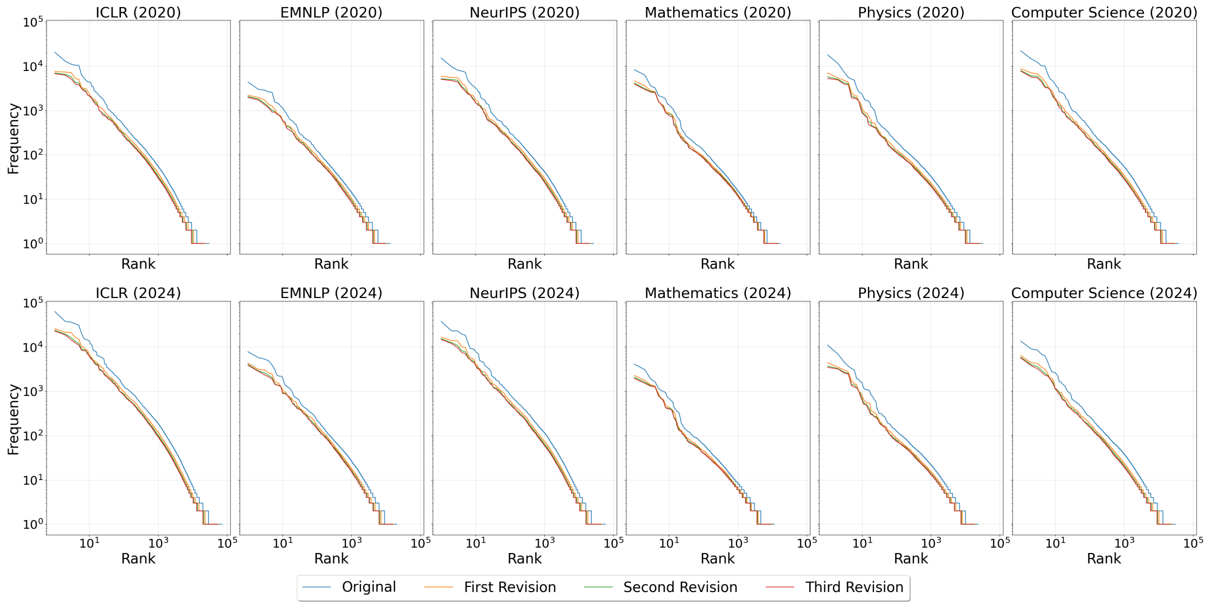

Figure 5.7: Word frequency curves across revisions.

Word frequency analysis reveals that the most common words in the original abstracts are function words such as “the”, “and”, “to”, “for”, which is expected. However, as revisions progress, there is a noticeable shift toward more content-specific terms like “paper” and “novel.” In nearly all cases, the word frequency curve flattens across revisions, indicating that Gemini encourages diversity in language, potentially incorporating more technical terms while reducing reliance on the most frequent function words. Moreover, in the arXiv dataset the slope of 2024 abstracts’ curves tends to be less steep than in 2020, further suggesting that the writing of some of these abstracts may have been assisted by AI.

The percentage of adjectives and adverbs from Liang et al. (2024)’s lists ranges from 0 to 1% in the original abstracts, with 2024 exhibiting higher average values in all categories except Mathematics. This percentage increases across revisions, supporting the notion that AI models like Gemini have a preference for these terms. As with other linguistic features, WD1 is greater than other distances, highlighting that most changes are introduced with the first revision. This distance is higher in 2020 for all categories except Mathematics, suggesting again that some abstracts in 2024 may have been written or refined with LLMs.

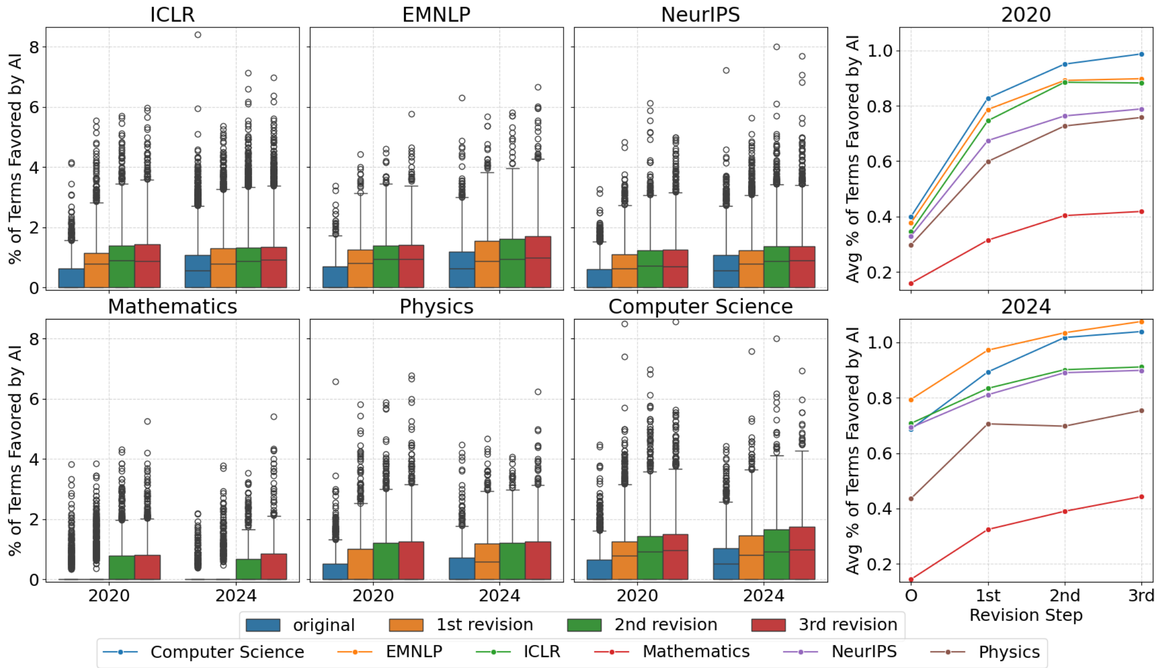

Figure 5.8: Percentage of terms favored by AI across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	0.3995	0.4095	0.3449	0.4279	0.1554	0.3022
WD1 (2024)	0.1573	0.2008	0.1462	0.2271	0.18	0.2698
WD2 (2020)	0.1391	0.1049	0.0891	0.1249	0.0888	0.1298
WD2 (2024)	0.0875	0.0818	0.0997	0.1245	0.0684	0.0328
WD3 (2020)	0.0619	0.0535	0.0373	0.0492	0.0243	0.036
WD3 (2024)	0.0269	0.0646	0.0285	0.0428	0.0532	0.0564

Table 5.8: Wasserstein distances for percentage of terms favored by AI across revisions.

The analysis of basic syntactic features shows that original abstracts typically consist of 5 to 10 sentences, each containing an average of 20 to 30 words. The number of sentences tends to be higher in 2024 than in 2020, while sentence length generally follows the opposite trend. However, the revising process leads to a decrease in both metrics, indicating once again Gemini's inclination towards shorter, more concise abstracts. The first revision contributes the most to overall modifications, as proved by WD1 being significantly higher than WD2 and WD3. Additionally, WD1 is more pronounced in 2024 for the number of sentences and in 2020 for sentence length.

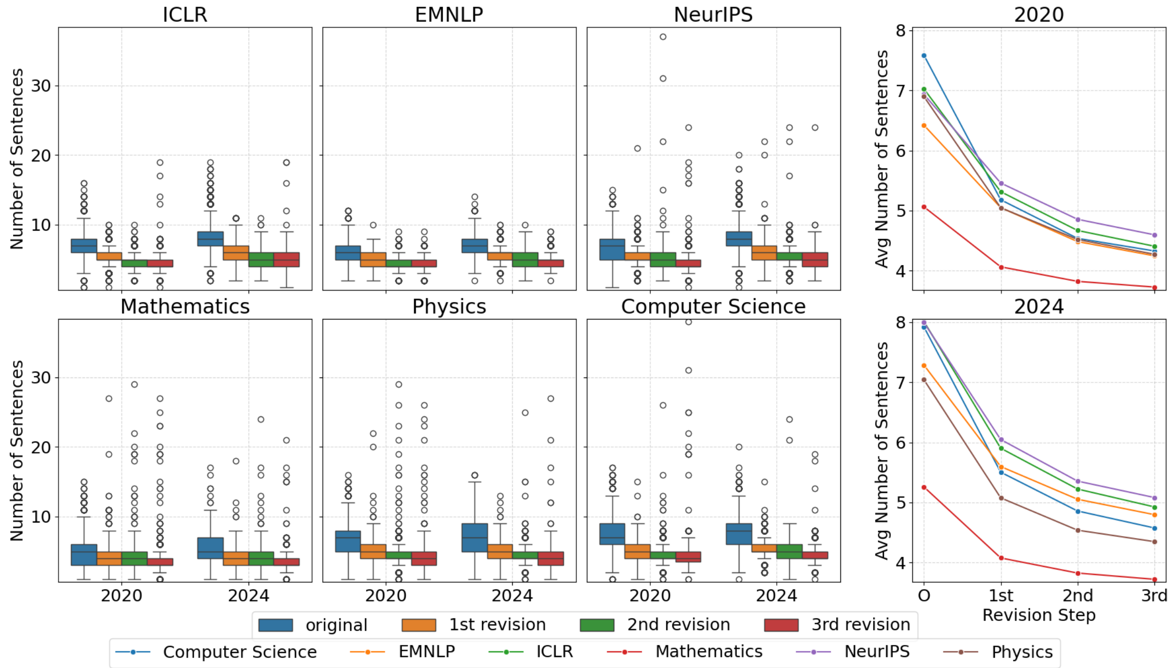

Figure 5.9: Number of sentences across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	1.7141	1.3728	1.486	2.4069	1.0269	1.8651
WD1 (2024)	2.1099	1.6896	1.9518	2.4196	1.1858	1.9688
WD2 (2020)	0.6462	0.5646	0.6619	0.6656	0.351	0.6333
WD2 (2024)	0.679	0.5446	0.6954	0.669	0.3716	0.5821
WD3 (2020)	0.2843	0.237	0.2736	0.2994	0.158	0.2486
WD3 (2024)	0.305	0.2561	0.2734	0.2911	0.1512	0.2286

Table 5.9: Wasserstein distances for number of sentences across revisions.

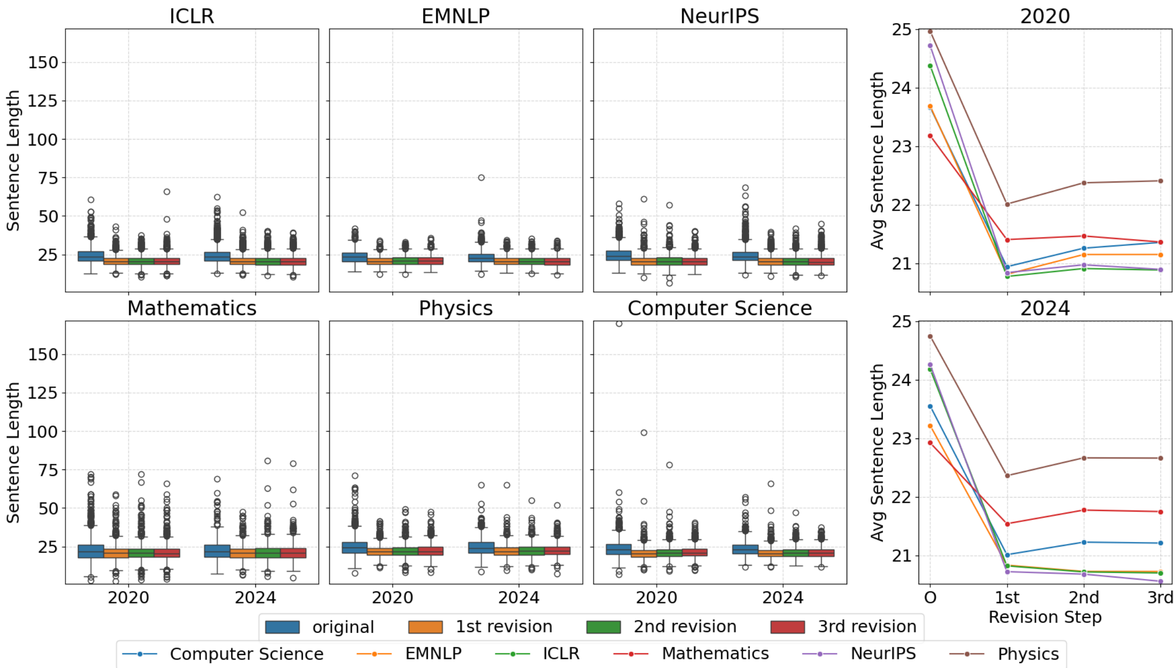

Figure 5.10: Average sentence length across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	3.5999	2.8732	3.8791	2.7505	2.156	2.9796
WD1 (2024)	3.3597	2.385	3.535	2.5499	1.6434	2.3867
WD2 (2020)	0.1836	0.3554	0.2117	0.3503	0.2206	0.3926
WD2 (2024)	0.1326	0.1449	0.1351	0.2921	0.3655	0.3404
WD3 (2020)	0.1143	0.1271	0.112	0.1956	0.1771	0.1707
WD3 (2024)	0.0471	0.1408	0.1458	0.1759	0.2296	0.1894

Table 5.10: Wasserstein distances for average sentence length across revisions.

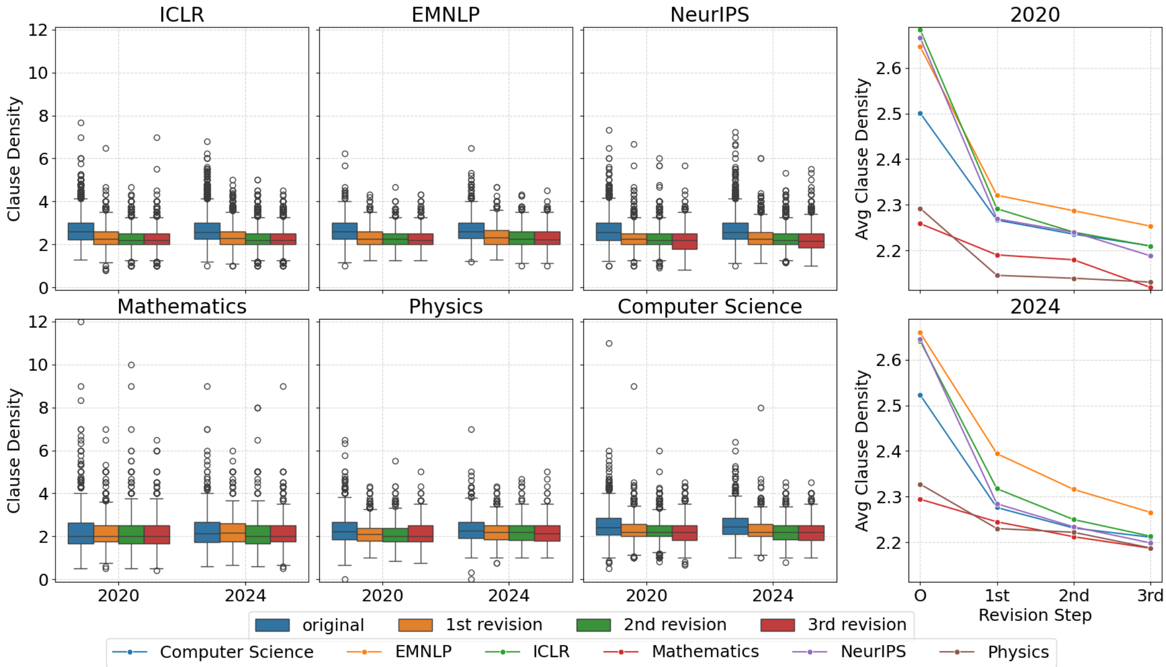

Figure 5.11: Average clause density across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	0.3929	0.3275	0.3975	0.2361	0.1679	0.1538
WD1 (2024)	0.3242	0.2661	0.3616	0.2495	0.1114	0.1074
WD2 (2020)	0.0519	0.0366	0.0371	0.0324	0.0448	0.0199
WD2 (2024)	0.0677	0.0784	0.0512	0.0461	0.071	0.0231
WD3 (2020)	0.0337	0.0446	0.0516	0.0326	0.0606	0.0211
WD3 (2024)	0.0365	0.0506	0.037	0.0227	0.0552	0.0394

Table 5.11: Wasserstein distances for average clause density across revisions.

Regarding syntactic complexity, original abstracts predominantly display an average clause density of 2–3 clauses per sentence, average T-unit density of 1-1.5 and 0.5-1.5 dependent clauses per T-unit. Average values for clause density and amount of subordination per T-unit tend to be higher in 2024, while T-unit density follows the opposite trend. All three metrics decrease across revisions, reflecting Gemini's tendency to shorten abstracts and simplify sentence structure. WD1 is larger than subsequent distances, proving again that most changes occur in the first revision. Furthermore, 2020 exhibits generally higher WD1 values for all metrics except dependent clauses per T-units, suggesting possible AI intervention in some abstracts from 2024.

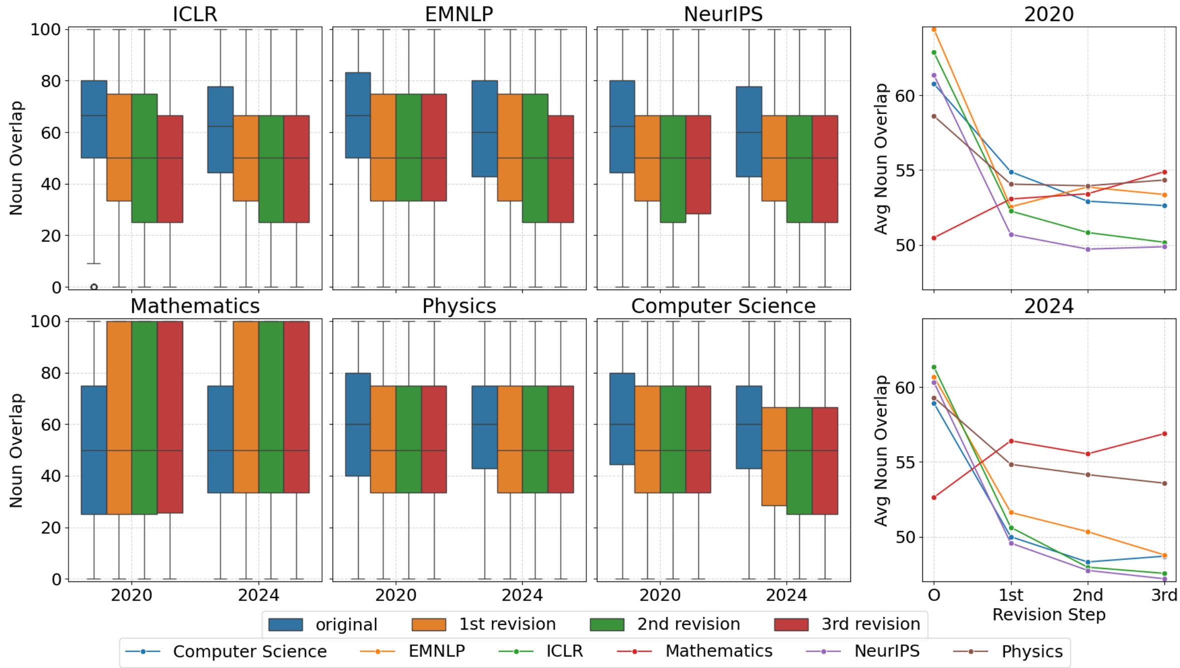

Figure 5.12: Noun overlap across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	10.6139	11.8833	10.6668	6.7024	3.4986	5.4534
WD1 (2024)	10.7308	9.2315	10.759	9.2671	5.2123	6.5189
WD2 (2020)	2.859	3.0086	3.3324	3.5896	1.265	2.1148
WD2 (2024)	3.0522	1.8313	2.3296	2.7422	2.3458	1.8371
WD3 (2020)	1.0264	1.8019	0.586	1.2988	1.487	1.3913
WD3 (2024)	0.8314	1.5828	1.0985	0.9966	1.3656	1.3041

Table 5.12: Wasserstein distances for noun overlap across revisions.

In the majority of original abstracts, 40-80% of consecutive sentences show some noun overlap, 50-85% argument overlap and almost all consecutive sentences have some lemmas in common. On average, these values tend to be higher in 2020 than in 2024, except for stem overlap. Noun and argument overlap decrease across revisions, while stem overlap shows a slight increase. This suggests an overall decline in referential cohesion, which may indicate that revisions favor conciseness at the expense of explicit connections between ideas, making the text harder to understand. The most substantial changes occur during the first revision, as indicated by WD1 being significantly higher than WD2 and WD3. In 2024, WD1 is generally larger than it is in 2020 for noun and argument overlap, while stem overlap follows the opposite trend. For all metrics, the Mathematics category often behaves as an outlier, exhibiting distinct trends from other categories or showing significantly different Wasserstein distances.

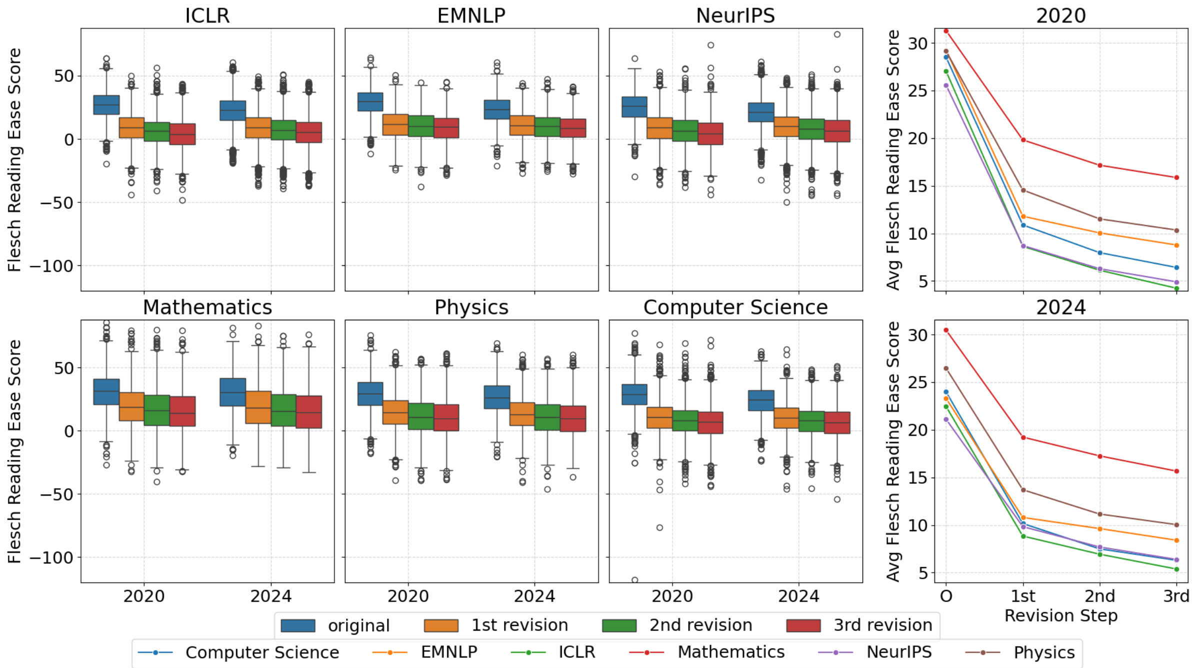

Figure 5.13: Flesch Reading Ease Score across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	18.3941	17.4474	16.8467	17.6464	11.4738	14.5866
WD1 (2024)	13.6029	12.5022	11.2775	13.8224	11.2444	12.7623
WD2 (2020)	2.5232	1.7528	2.4425	2.9479	2.6711	3.0309
WD2 (2024)	1.9042	1.1824	2.1399	2.6745	1.9993	2.5653
WD3 (2020)	1.8889	1.3237	1.4038	1.5607	1.3077	1.2399
WD3 (2024)	1.5588	1.221	1.331	1.2336	1.6323	1.209

Table 5.13: Wasserstein distances for Flesch Reading Ease Score across revisions.

All readability scores indicate a decline in readability across revisions. The Flesch Reading Ease Score for original abstracts typically ranges between 20 and 40, while the Gunning Fog Index and the SMOG Index fall between 15 and 17. Average values for the Flesch Reading Ease Score are higher in 2020, suggesting greater readability, whereas the Fog and SMOG Indices are higher in 2024, indicating increased complexity. WD1 is larger than WD2 and WD3, confirming most alterations are introduced with the first revision. This distance is more pronounced in 2020 than in 2024, reinforcing the likelihood of LLM involvement in more recent abstracts, though Mathematics shows some exceptions.

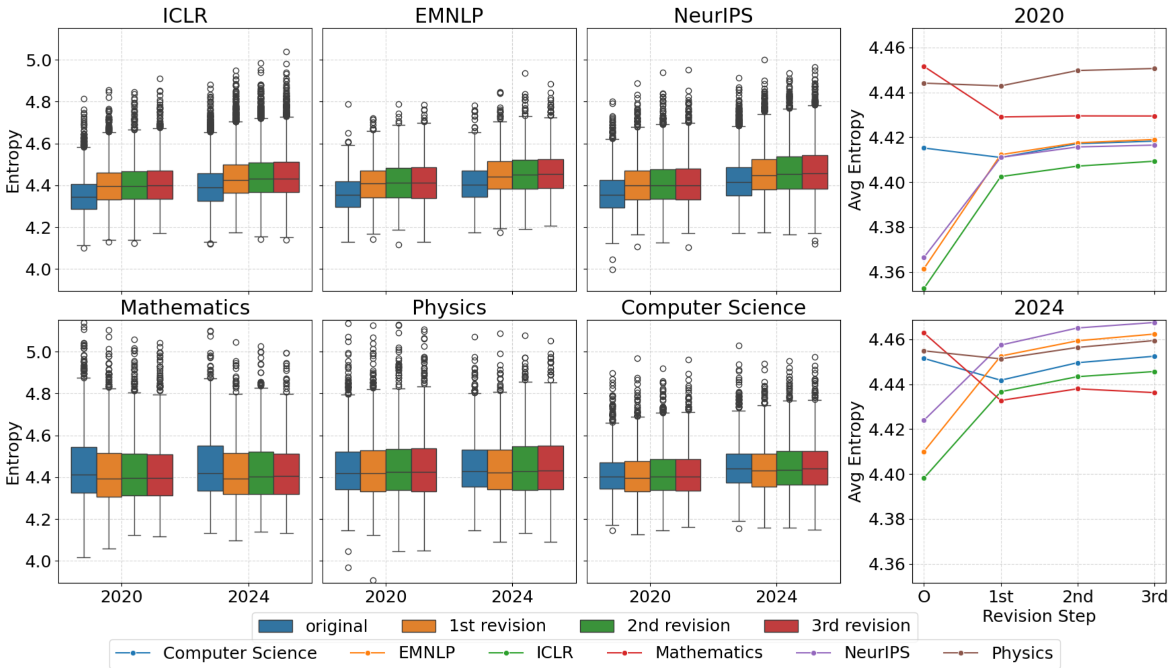

Figure 5.14: Entropy across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	0.0498	0.051	0.0446	0.0087	0.0231	0.0075
WD1 (2024)	0.0385	0.0425	0.0335	0.0124	0.0301	0.0104
WD2 (2020)	0.0048	0.0057	0.005	0.0065	0.0031	0.0071
WD2 (2024)	0.0068	0.0079	0.0077	0.0079	0.0063	0.0069
WD3 (2020)	0.003	0.0039	0.0026	0.0021	0.0039	0.0031
WD3 (2024)	0.0027	0.0039	0.003	0.0039	0.0051	0.0043

Table 5.14: Wasserstein distances for entropy across revisions.

Entropy values in original abstracts range from 4.3 to 4.5 bits per character, with higher averages in 2024. Across revisions, entropy exhibits a slight increase, except for Mathematics. Higher entropy indicates greater unpredictability and information density, which can make texts more difficult to process. The most significant changes occur with the first revision, as proven by WD1 being higher than WD2 and WD3. WD1 is greater in 2020 for the conference dataset and in 2024 for the arXiv dataset, suggesting varying levels of AI influence across different sources.

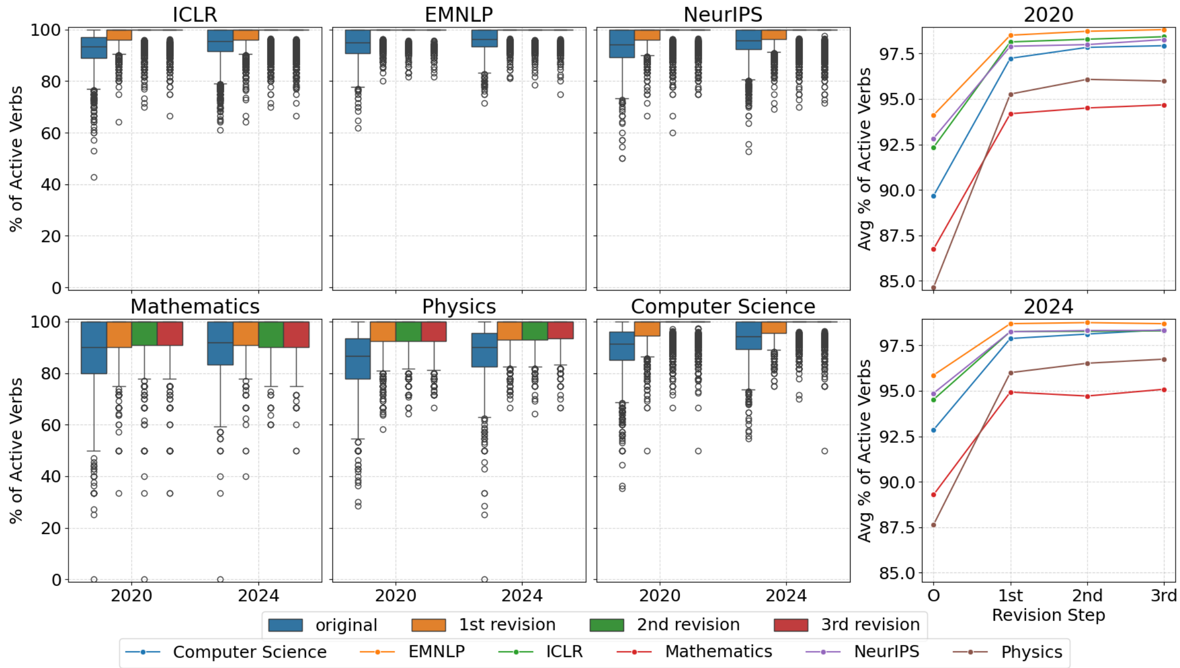

Figure 5.15: Percentage of active verbs across revisions.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD1 (2020)	5.8022	4.3958	5.0658	7.5633	7.4501	10.6341
WD1 (2024)	3.7403	2.8299	3.3821	5.0237	5.6493	8.3666
WD2 (2020)	0.2888	0.2595	0.2089	0.6013	0.5616	0.8104
WD2 (2024)	0.1933	0.1313	0.2115	0.2845	0.5866	0.5255
WD3 (2020)	0.187	0.115	0.2784	0.1636	0.3748	0.1694
WD3 (2024)	0.1603	0.1132	0.1191	0.33	0.4487	0.3422

Table 5.15: Wasserstein distances for percentage of active verbs across revisions.

Active verbs comprise 80–100% of total verbs in most abstracts, while passive verbs account for 0–20%. In terms of tense distribution, 25–45% of verbs appear in the present tense, 15–40% in the past, and only a very small percentage in the future. Imperative verbs are almost absent, with most verbs in the indicative mood (95–100%), while subjunctives make up a smaller 0–8%. On average, active and indicative verbs are more prevalent in 2024, whereas passive, imperative, subjunctive, and past-tense verbs appear more frequently in 2020. Across revisions, active, indicative and present-tense verbs increase while passive, subjunctive, past and future-tense verbs decrease, reflecting a shift toward a more direct and engaging tone, as well as a preference for clarity, conciseness, and assertiveness, making the text more compelling and easier to follow. As with other linguistic features, the most significant changes occur in the first revision. WD1 is significantly higher than other distances, and most of these metrics and categories show larger distances in 2020 than 2024, indicating LLMs may have played a role in shaping more recent abstracts.

Overall, this analysis of LLM-generated revisions of research paper abstracts reveals important insights regarding the possible impact of AI on academic writing. First revisions consistently introduce the majority of changes, with subsequent revisions contributing with only minor refinements. Notably, many of the metrics observed show that abstracts from 2024 require fewer modifications than those from 2020, suggesting that they are already closer to the writing style favored by AI models. This strongly indicates that many 2024 abstracts were either edited by LLMs before submission or written with LLM assistance from the start. In contrast, 2020 abstracts, likely written by humans without AI support, exhibit more substantial transformations in their first revision as they are adjusted to match the typical LLM style.

The characteristics of these revisions shed light on the writing tendencies of LLMs like Gemini. Through revisions, abstracts become more concise, with a decrease in word count, sentence length, and syntactic complexity. At the same time, lexical diversity increases, with abstracts exhibiting less word repetitions. This, coupled with the increase of content-dense words like nouns, verbs and adjectives, suggest that LLMs prioritize efficiency and precision, reducing redundancy and enhancing specificity.

However, these revisions also reveal a trade-off: while LLMs try to simplify writing, they also tend to reduce referential cohesion by decreasing noun and argument overlap. Additionally, a decline in connectives and adverbs results in less explicit logical relationships between ideas. Combined with an increase in entropy and greater lexical complexity and diversity, this likely contributes to the observed drop in readability scores. Although the text becomes more information-dense and varied in vocabulary, it may also become more challenging to process.

Another notable shift is the change in verb usage. Revisions increase the use of active, indicative, and present-tense verbs while reducing passive constructions and the adoption of other moods and tenses. This suggests a preference for more direct, engaging, and immediate language.

Taken together, these trends indicate that LLM-assisted writing is shaping the evolution of academic style, favoring concise, lexically diverse, and structurally efficient texts. However, this shift may come at the cost of readability and cohesion.

5.2 LLM-generated abstracts

The second experiment builds on the findings of the first one, leveraging the observed trend that 2024 abstracts require less editing to achieve an LLM-like writing style. Specifically, most metrics indicate that the distributions for original abstracts are more similar to their first revision in 2024 rather than in 2020, suggesting that some recent abstracts are already written in a style more consistent with LLM-generated text. To further validate this consistency, a different approach was developed.

D1 is defined as the absolute value of the difference between the metric computed for the first revision and the original abstract, while D2 represents the absolute value of the difference between the metric computed for the second and first revision. The previous analysis confirms that D1 is larger than D2 for the vast majority of abstracts, as the most substantial changes almost always occur in the first revision, while subsequent revisions introduce only minor improvements. Given this pattern, D1 - D2 serves as an effective measure of how much an abstract changes across revisions. If an abstract was written with LLM assistance, its D1 - D2 values are expected to align more closely with those of 2024 abstracts rather than 2020 abstracts, indicating that fewer edits are needed to reach an LLM-like style.

To test this hypothesis, Gemini was tasked with generating 100 abstracts using the prompt: “Write an abstract for an academic paper on a random IT topic.” All the metrics were then evaluated for these AI-generated abstracts, and their D1 - D2 values were compared to those of 2020 and 2024 abstracts. The results show that the D1 - D2 distributions of AI-generated abstracts tend to be closer to those of 2024 abstracts than to those from 2020. Notably, the only exceptions are the number of sentences and percentage of adverbs, which did not exhibit the original trend of 2024 abstracts being more similar to their first revisions than 2020 abstracts. Furthermore, the distances between the D1 - D2 distributions for 2020 and 2024 abstracts are, in most cases, smaller than those between 2024 and AI-generated abstracts. This suggests that while recent abstracts are adopting a more LLM-like style, there are still notable differences when compared to fully AI-generated abstracts. It also accounts for the fact that only some of the 2024 abstracts are likely written with LLMs.

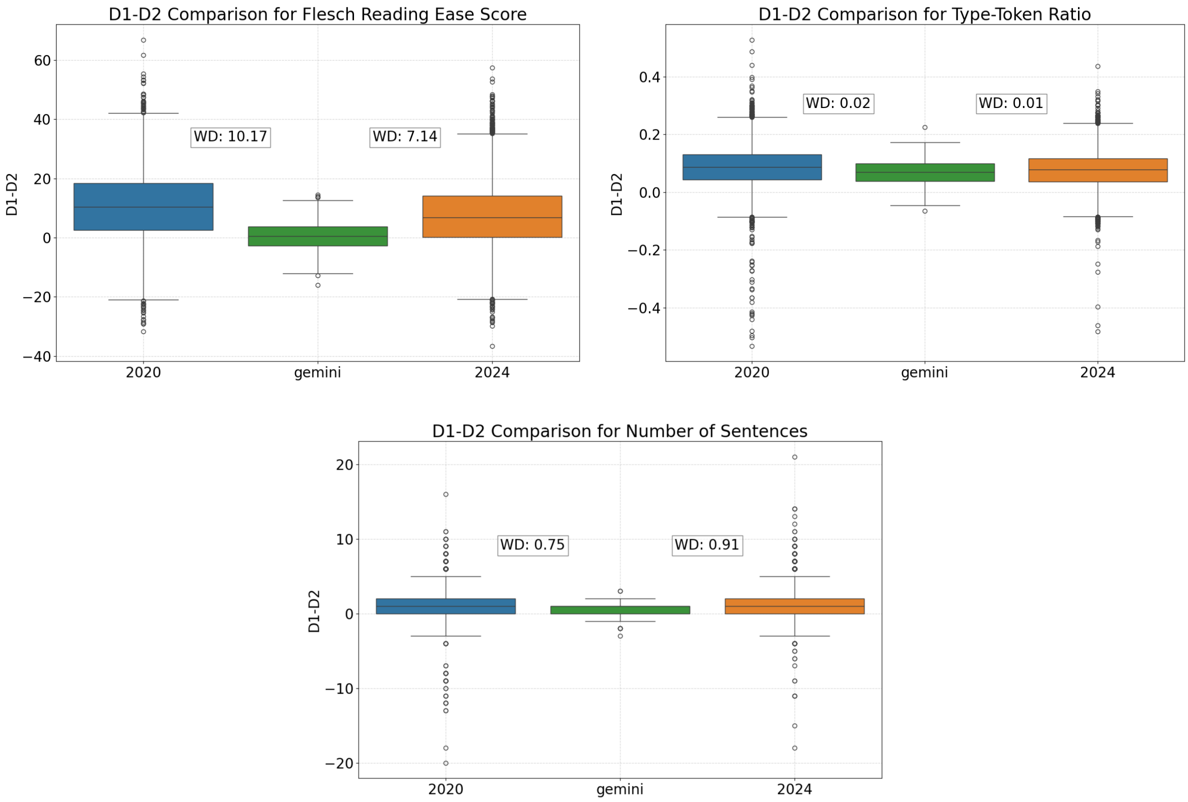

Figure 5.16: D1-D2 Comparisons.

5.3 Analysis of abstracts over the years

Lastly, text similarity and writing style metrics were analyzed for original abstracts from 2020 to 2024 to determine whether the patterns observed in Gemini's revisions during the first experiment also manifest as broader trends over the years. If similar patterns emerge, especially after 2022, the year of ChatGPT's launch and consequent widespread diffusion of LLMs, it would suggest that LLM integration in academia has had a measurable impact on writing styles. It would also indicate that academic writing is gradually shifting toward a more LLM-like style, possibly leading to greater uniformity and standardization between texts. While this shift may enhance conciseness, information density, and lexical richness, it could also reduce readability, cohesion, and stylistic diversity, as authors increasingly conform to the linguistic preferences inherent to LLM-generated text. In the long run, this convergence could reshape scientific writing, making it more formulaic and potentially damaging the variety and originality of individual expression in academia.

For this analysis, lexical and semantic similarity were computed comparing each abstract from a specific year and conference or arXiv category with all the others in the same year and conference/category, to check if the abstracts in the past years have become more similar to each other in terms of specific words used and overall meaning. Additionally, the analysis was repeated for both metrics, focusing specifically on the two smaller datasets made of the top 100 abstracts per year and conference/category with the highest percentage of adjectives and adverbs favored by AI. This was done to assess whether certain trends are more or less pronounced in abstracts that have a higher likelihood of being written with LLMs.

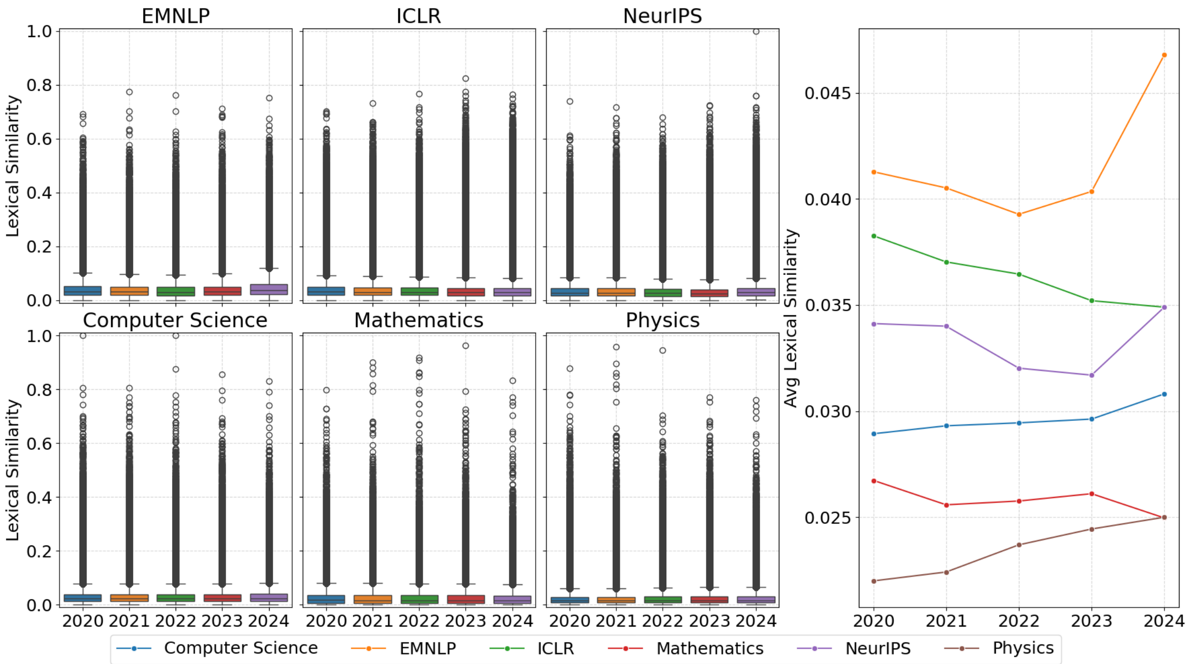

Figure 5.17: Lexical similarity over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.0012	0.0009	0.0005	0.0004	0.0012	0.0004
WD 2021-2022	0.0008	0.0013	0.002	0.0002	0.0002	0.0013
WD 2022-2023	0.0013	0.0014	0.0003	0.0002	0.0004	0.0007
WD 2023-2024	0.0004	0.0065	0.0032	0.0012	0.0012	0.0006

Table 5.16: Wasserstein distances for lexical similarity over the years.

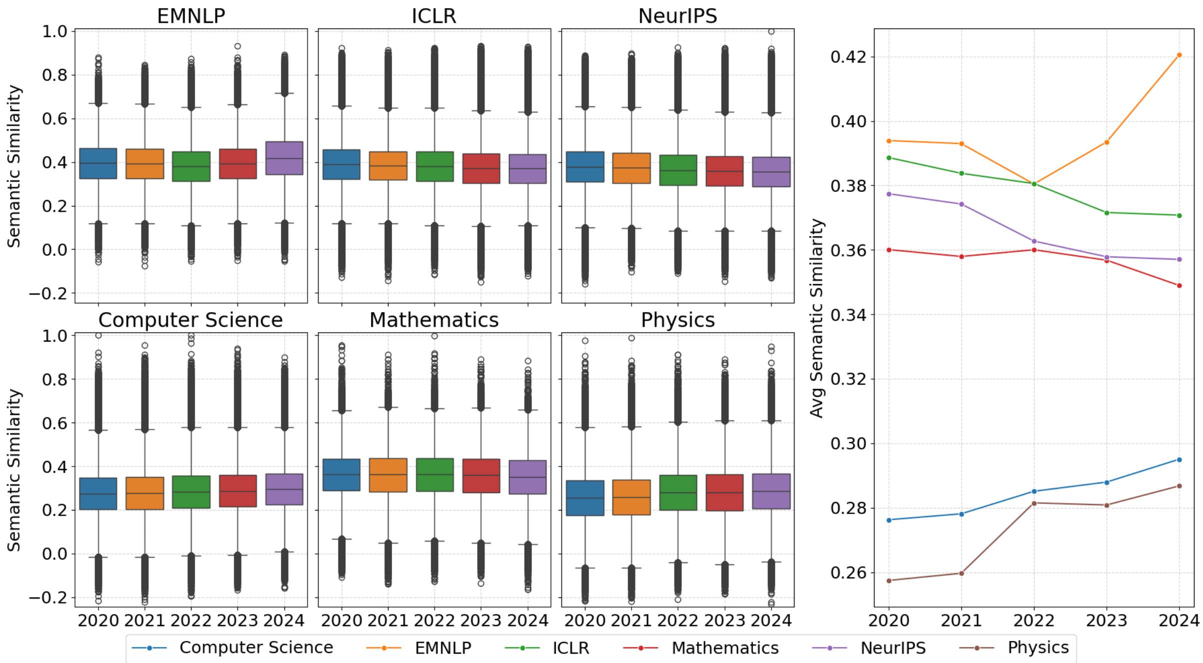

Figure 5.18: Semantic similarity over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.0052	0.0011	0.0035	0.0018	0.0048	0.0023
WD 2021-2022	0.0032	0.0126	0.0115	0.007	0.0025	0.0218
WD 2022-2023	0.009	0.013	0.0049	0.0029	0.0034	0.0024
WD 2023-2024	0.0017	0.0271	0.0022	0.0071	0.0078	0.0059

Table 5.17: Wasserstein distances for semantic similarity over the years.

The results from the conference dataset show that, after 2022, lexical similarity slightly increased in two out of three conferences, while semantic similarity increased in only one. The arXiv dataset reveals a similar trend, with semantic similarity increasing in two out of three categories. When restricting the analysis to abstracts with higher percentages of AI-preferred adjectives and adverbs, lexical similarity increased in all three conferences and arXiv categories, while semantic similarity increased in all categories except Mathematics.

However, these similarity scores are generally low, mostly ranging well below 0.1 for lexical similarity and between 0.2 and 0.4 for semantic similarity, with changes over the years being even smaller. These results suggest LLM adoption might've had a slight impact on abstract uniformity in terms of lexicon and meaning, but more research is required to make stronger claims. Notably, a potential factor influencing the low similarity scores obtained could be the subdivision of the abstracts by conferences or broad arXiv categories. The wide range of topics covered within these categories may have reduced the overall similarity between abstracts, and possibly hidden stronger trends.

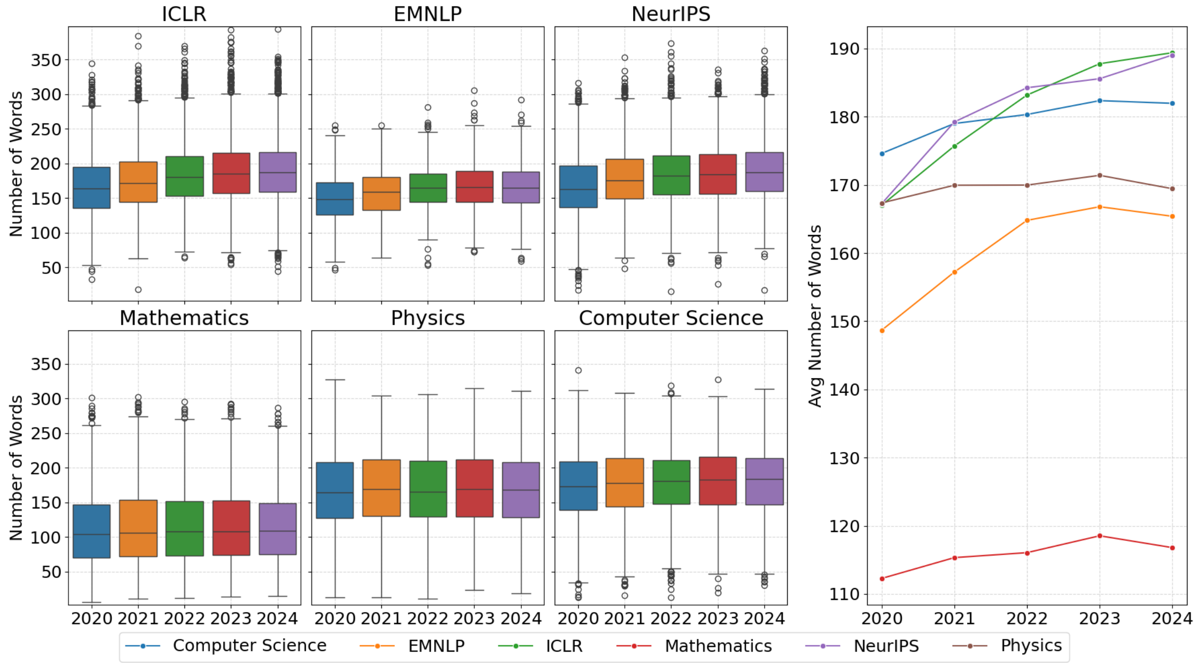

Figure 5.19: Number of words over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	8.7044	8.5324	11.9956	4.5168	3.1641	3.3025
WD 2021-2022	7.4903	7.7202	5.1285	2.4069	2.1995	2.6575
WD 2022-2023	4.6414	2.3392	2.0344	2.41	3.0737	2.4753
WD 2023-2024	1.9385	1.4805	3.479	1.7273	2.8021	2.4852

Table 5.18: Wasserstein distances for number of words over the years.

As for basic lexical features, the number of words in most abstracts ranges between 100 and 225, with an average word length of 1.75–2 syllables. Both metrics increased over the years, suggesting a preference for longer abstracts and potentially more complex, technical vocabulary. This trend contrasts with earlier findings indicating that LLMs favor conciseness. However, measuring the Wasserstein distance between distributions for consecutive years reveals that categories more likely to adopt LLMs because of their closer relation to AI (e.g the conferences dataset and Computer Science category) exhibit larger shifts between 2020 and 2022. From 2022 onward, these shifts decrease, whereas more traditional fields like Mathematics and Physics show steady, incremental changes. This pattern suggests that while abstracts have generally grown longer, LLM adoption after 2022 may have slowed this trend, encouraging more concise writing. Additionally, restricting the analysis to abstracts with the highest percentages of AI-favored adjectives and adverbs still shows an increase in abstract length, though the average word count is slightly lower.

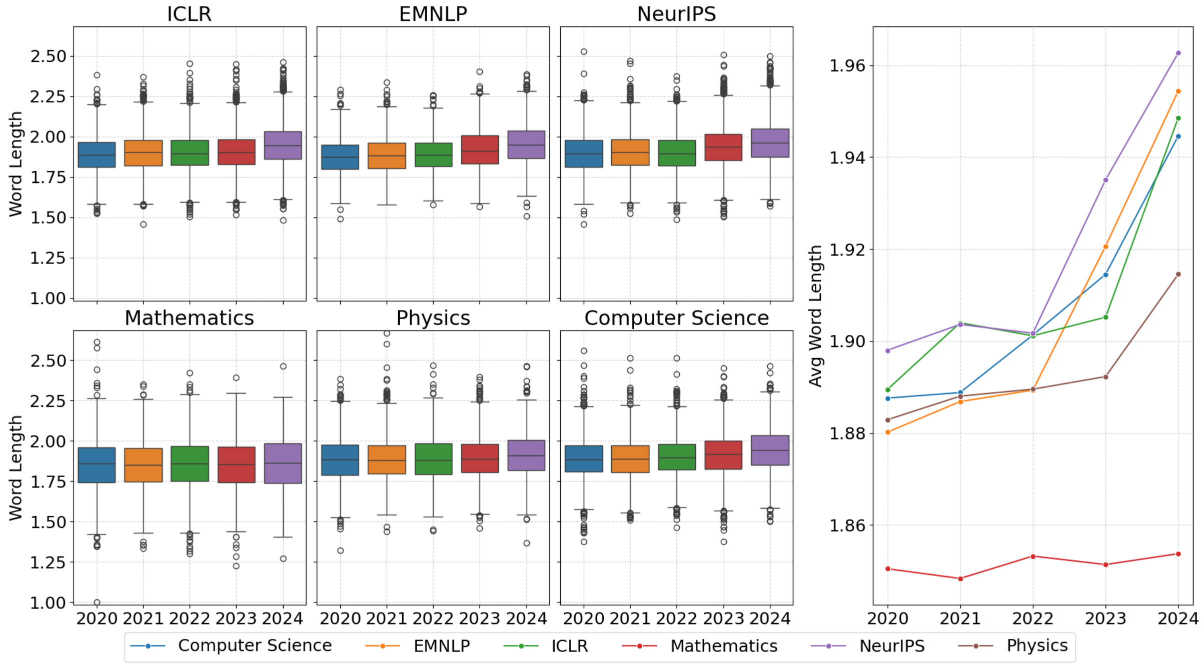

Figure 5.20: Average word length over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.0146	0.0076	0.008	0.0043	0.0077	0.0064
WD 2021-2022	0.0049	0.0043	0.0049	0.0126	0.0086	0.0054
WD 2022-2023	0.0041	0.0314	0.0334	0.014	0.0069	0.0052
WD 2023-2024	0.0434	0.034	0.0277	0.0301	0.0144	0.023

Table 5.19: Wasserstein distances for average word length over the years.

A similar pattern emerges for word length: the overall trend aligns with previous findings that LLMs favor longer, more complex words, and on top of that the conference dataset and Computer Science category display more pronounced shifts after 2022. In contrast, Mathematics and Physics show smaller or delayed changes, implying that LLM adoption may have contributed to the use of longer words. Notably, the trend toward increased word length persists even when focusing only on abstracts with a higher concentration of AI-favored terms, with average word length being slightly higher.

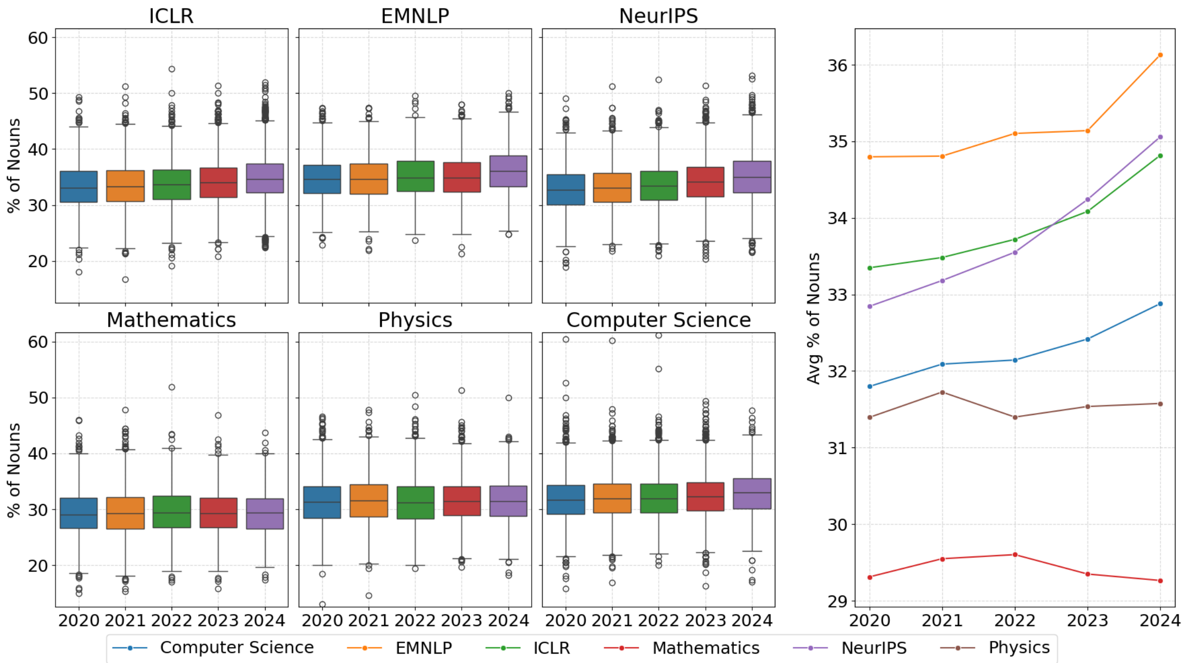

Figure 5.21: Percentage of nouns over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.1603	0.1692	0.339	0.3107	0.3186	0.3412
WD 2021-2022	0.2641	0.3328	0.374	0.1032	0.2781	0.3922
WD 2022-2023	0.3674	0.2068	0.6905	0.3208	0.2696	0.2452
WD 2023-2024	0.7337	0.9916	0.8193	0.5188	0.1854	0.1487

Table 5.20: Wasserstein distances for average percentage of nouns over the years.

Regarding the prevalence of different parts of speech, in most abstracts nouns make up 30–40% of total words, verbs 8–15%, adjectives 10–15%, adverbs 2–5%, and connectives 2.5–5.5%. Over the years, these trends partially align with findings from the analysis of AI-generated revisions (with Mathematics sometimes acting as an outlier), showing an increase in nouns, verbs, and adjectives, while adverbs and connectives exhibit no distinct trend. This could mean that more recent abstracts are written in a compact, information-dense, and efficient writing style.

The analysis of abstracts with the highest percentages of AI-favored adjectives and adverbs reveals trends that mirror those of AI-generated revisions, including a decrease in adverbs and connectives. Additionally, metrics such as the percentage of nouns and verbs exhibit higher Wasserstein distances between consecutive years after 2022, particularly between 2023 and 2024. This effect is especially evident in fields more likely to have adopted LLMs, suggesting that LLM usage may have contributed to the increased use of such parts of speech.

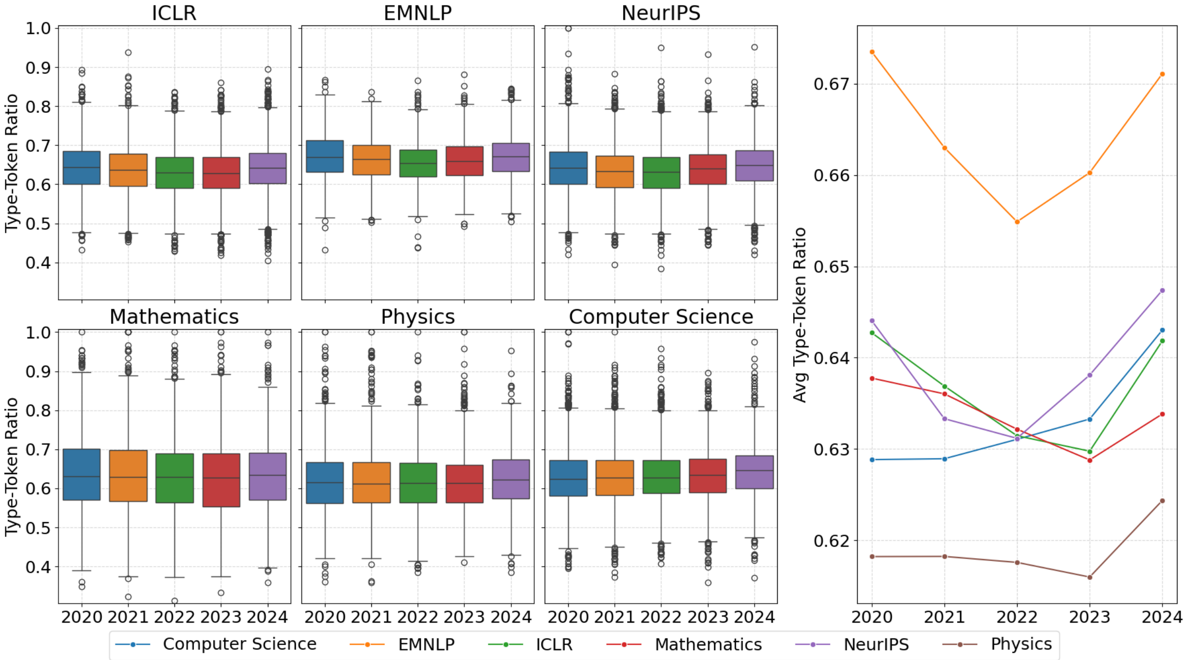

Figure 5.22: TTR over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.0059	0.0108	0.0108	0.0021	0.0043	0.0022
WD 2021-2022	0.006	0.0084	0.0024	0.0027	0.0051	0.0036
WD 2022-2023	0.0018	0.0057	0.0074	0.0037	0.0051	0.0029
WD 2023-2024	0.0121	0.011	0.0093	0.0098	0.0079	0.0099

Table 5.21: Wasserstein distances for TTR over the years.

With respect to vocabulary richness, most abstracts exhibit a TTR between 0.6 and 0.7, Yule's K-complexity between 100 and 200, and gzip complexity between 0.4 and 0.5. All three metrics indicate an increase in vocabulary richness after 2022, though gzip complexity shows more fluctuations, particularly for Physics and Mathematics. This aligns with previous results finding that LLM revisions lead to greater lexical diversity and a less repetitive vocabulary. The analysis of abstracts containing more terms preferred by AI largely supports this trend, with higher averages for TTR, and lower averages for the other two metrics. Furthermore, TTR and K-complexity also display larger Wasserstein distances between yearly distributions after 2022. This suggests that LLM adoption may be contributing to the increasing complexity and richness of vocabulary in academic writing.

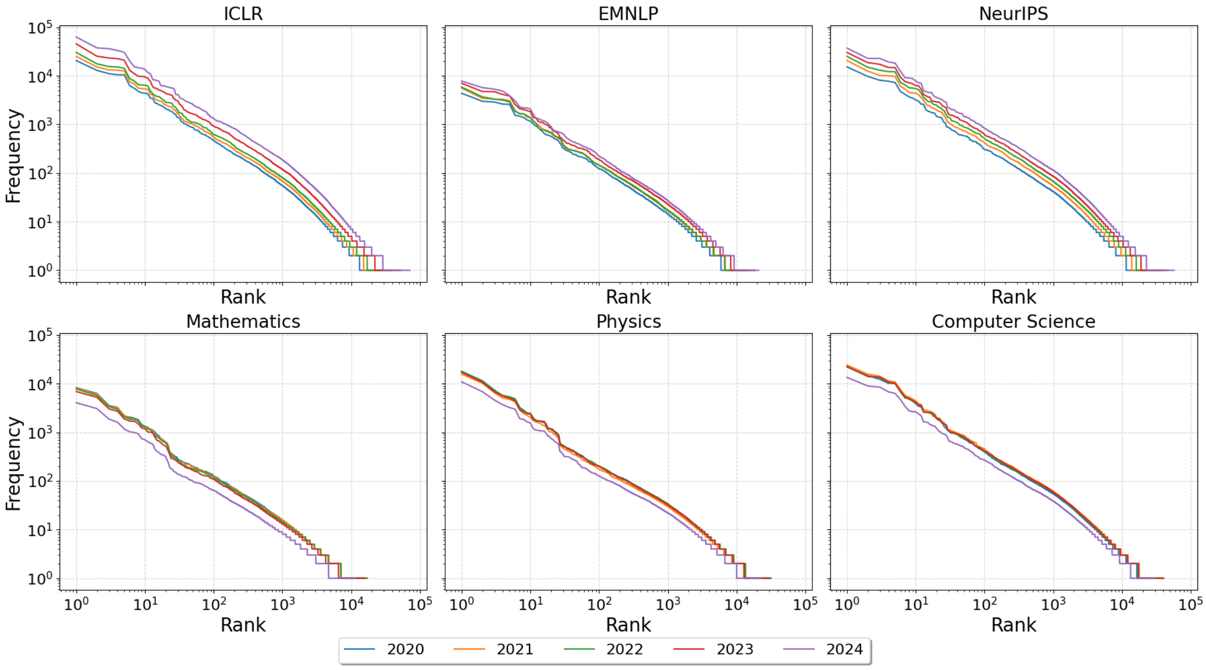

Figure 5.23: Word frequency curves over the years.

The analysis of word frequency across the years reveals that the most frequent words remain common function words (the, of, and, to, a, in, is, for, that), with "we" becoming more prominent after 2022, suggesting a shift toward a more collaborative writing style. Overall, curve slopes remain stable over the years, though small changes in the arXiv dataset indicate increasing lexical diversity, aligning with previous findings on LLM-generated revisions. Interestingly, EMNLP shows the opposite trend, suggesting a growing dominance of high-frequency words. When restricting the analysis to abstracts with the highest percentages of AI-favored terms, the slopes tend to flatten, but no clear pattern emerges, with some categories showing an increase in lexical diversity, while others remain stable or decline.

For most abstracts, AI-favored adjectives and adverbs make up only 0–1% of the total words, with the highest percentages reaching 3–4%, particularly in recent years. Their usage has increased over time, aligning with previous findings, and Wasserstein distances between consecutive yearly distributions become more pronounced after 2022. This effect is especially noticeable in fields more closely related to AI, further suggesting that LLMs might've contributed to accelerating the trend and may inherently favor these types of adjectives and adverbs. Once again, Mathematics acts as an outlier, exhibiting a decrease in the prevalence of such terms.

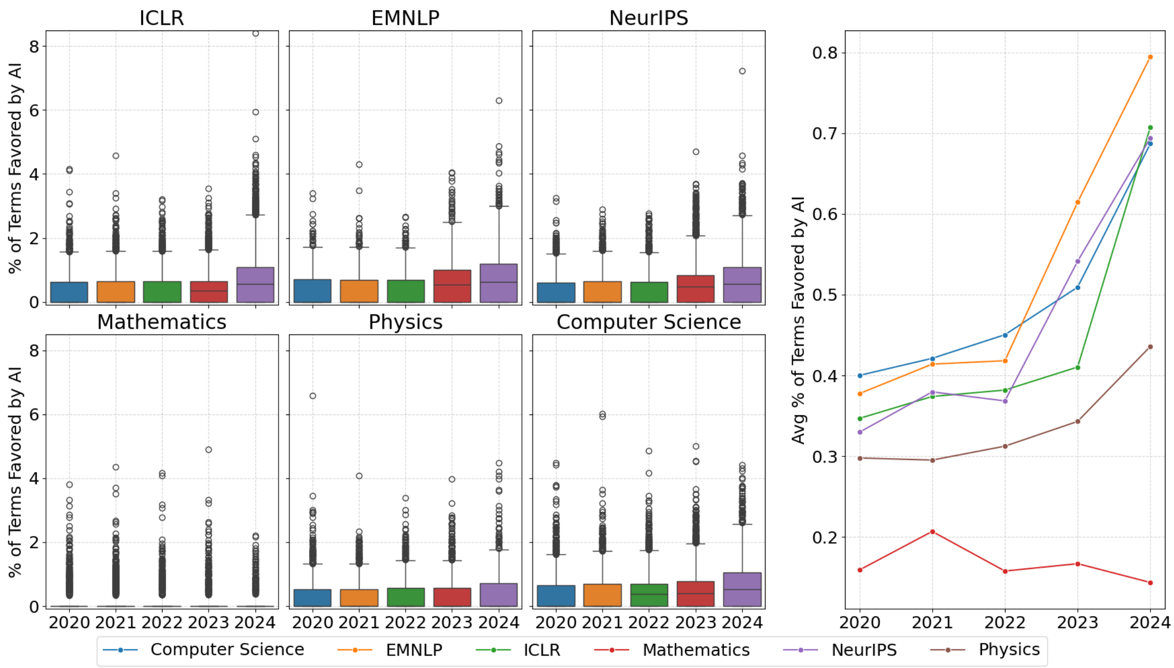

Figure 5.24: Percentage of terms favored by AI over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.0324	0.0394	0.0511	0.0284	0.0492	0.0148
WD 2021-2022	0.0116	0.0265	0.0131	0.0316	0.0512	0.0207
WD 2022-2023	0.0291	0.1963	0.1729	0.0589	0.0156	0.0307
WD 2023-2024	0.2971	0.1801	0.1525	0.1783	0.0233	0.0928

Table 5.22: Wasserstein distances for percentage of terms favored by AI over the years.

Moving on to the syntactic features, abstracts typically consist of 5 to 10 sentences with an average length of 20–30 words. Over time, the number of sentences increased, while sentence length mostly decreased. These trends partly contrast with the findings on LLMs revisions, which showed that LLMs favor shorter abstracts with fewer, smaller sentences. However, lower Wasserstein distances after 2022 for the sentence count, combined with the fact that abstracts with higher percentages of AI-favored terms reflect the same trend as the main datasets but exhibit mostly lower averages, suggest that LLM adoption may have slowed down the shift toward longer abstracts.

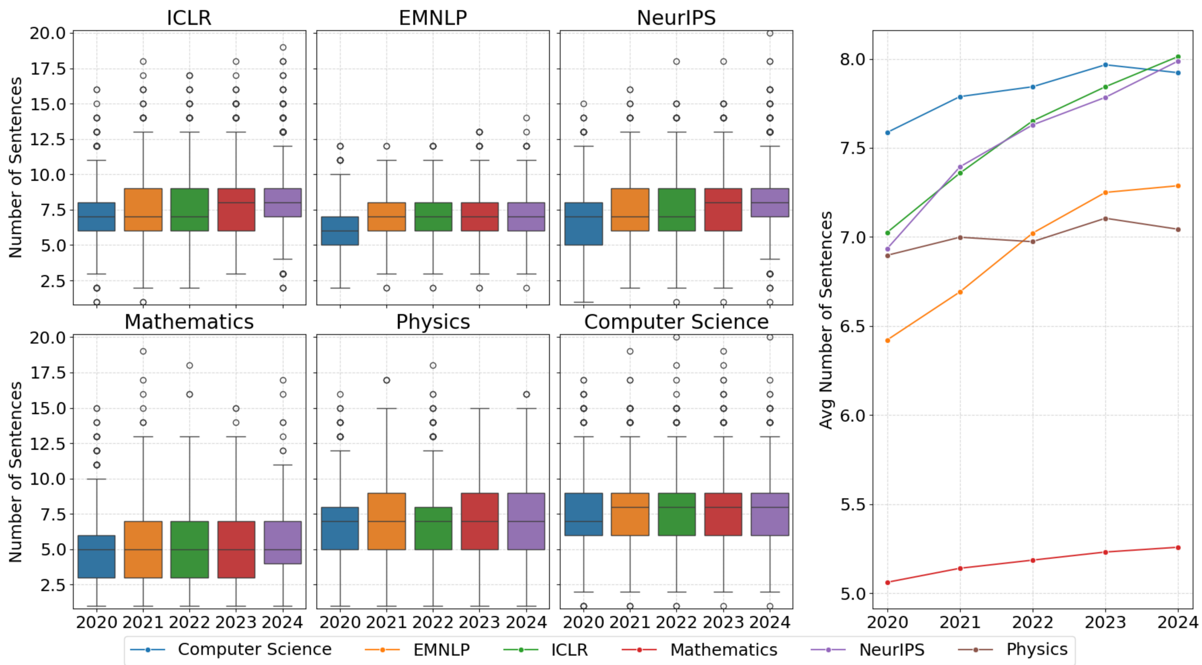

Figure 5.25: Number of sentences over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.3328	0.2757	0.4593	0.205	0.1107	0.1191
WD 2021-2022	0.2938	0.3301	0.2378	0.1145	0.1005	0.1071
WD 2022-2023	0.198	0.2378	0.1564	0.1293	0.0791	0.1927
WD 2023-2024	0.1709	0.0843	0.2041	0.0856	0.1539	0.1004

Table 5.23: Wasserstein distances for number of sentences over the years.

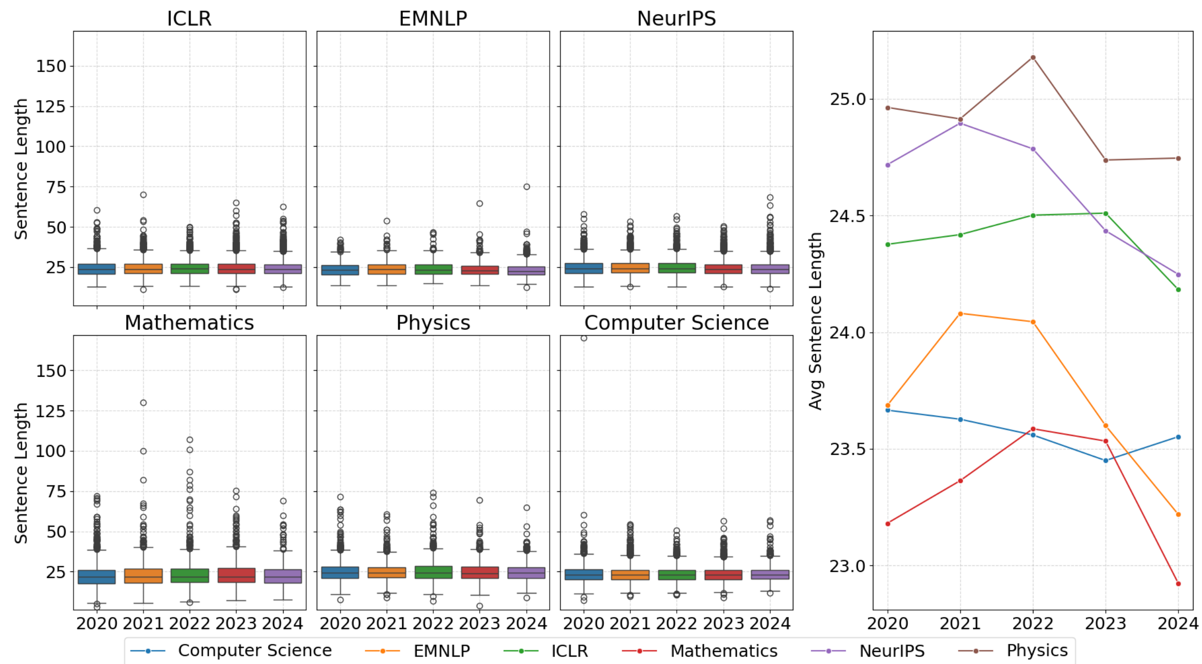

Figure 5.26: Average sentence length over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.18	0.411	0.2516	0.2601	0.6665	0.3528
WD 2021-2022	0.1886	0.2497	0.1826	0.2241	0.4142	0.4229
WD 2022-2023	0.1141	0.4907	0.3753	0.215	0.3911	0.4488
WD 2023-2024	0.3296	0.4259	0.2541	0.1792	0.6959	0.254

Table 5.24: Wasserstein distances for average sentence length over the years.

According to the analysis of syntactic complexity, abstracts typically exhibit an average clause density of 2-3 clauses per sentence, average T-unit density of 1–1.5, and average number of dependent clauses per T-unit between 1 and 1.5. Over time, clause density and T-unit density generally decreased, while the ratio of dependent clauses increased. These results are partially in contrast with the findings from the experiments with LLM revisions, which showed a decrease for all syntactic complexity metrics and a preference for simpler, less convoluted constructions. However, despite presenting the same trends as the main datasets, the datasets of abstracts with greater percentages of adjectives and adverbs favored by AI mostly exhibit lower averages for the ratio of dependent clauses, possibly suggesting that LLM contribution in writing some of these abstracts may have slowed down the increasing trend.

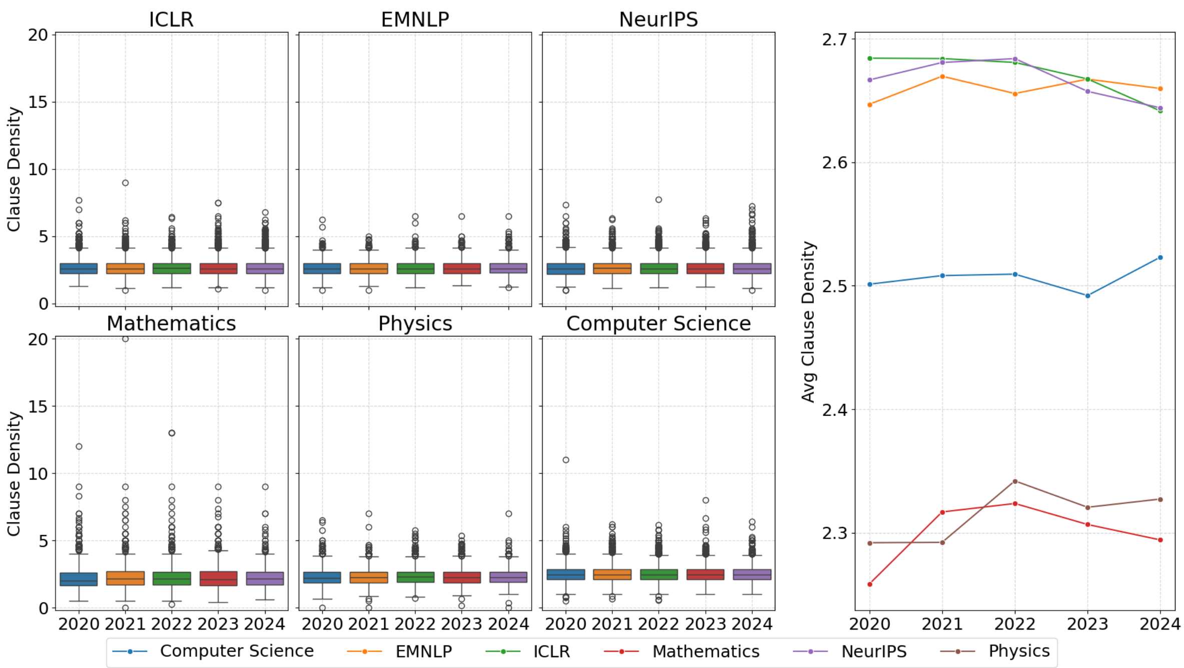

Figure 5.27: Average clause density over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.0301	0.0402	0.0322	0.0168	0.0954	0.0254
WD 2021-2022	0.0204	0.0314	0.0178	0.0155	0.0543	0.0524
WD 2022-2023	0.0206	0.0209	0.0329	0.0316	0.0607	0.0247
WD 2023-2024	0.0284	0.0424	0.0299	0.0335	0.0586	0.0425

Table 5.25: Wasserstein distances for average clause density over the years.

As for referential cohesion, in most abstracts 40-80% of adjacent sentences exhibit noun overlap, 50-85% show argument overlap, and almost all stem overlap. Over time, all types of overlap mostly decreased, especially in 2024. The analysis of abstracts with higher proportions of AI-favored terms aligns with these trends, even exhibiting lower averages for the decreasing metrics. These findings are largely coherent with the results from the analysis of AI-generated revisions, and suggest a shift toward a less cohesive style, which favors conciseness over explicitly connected concepts.

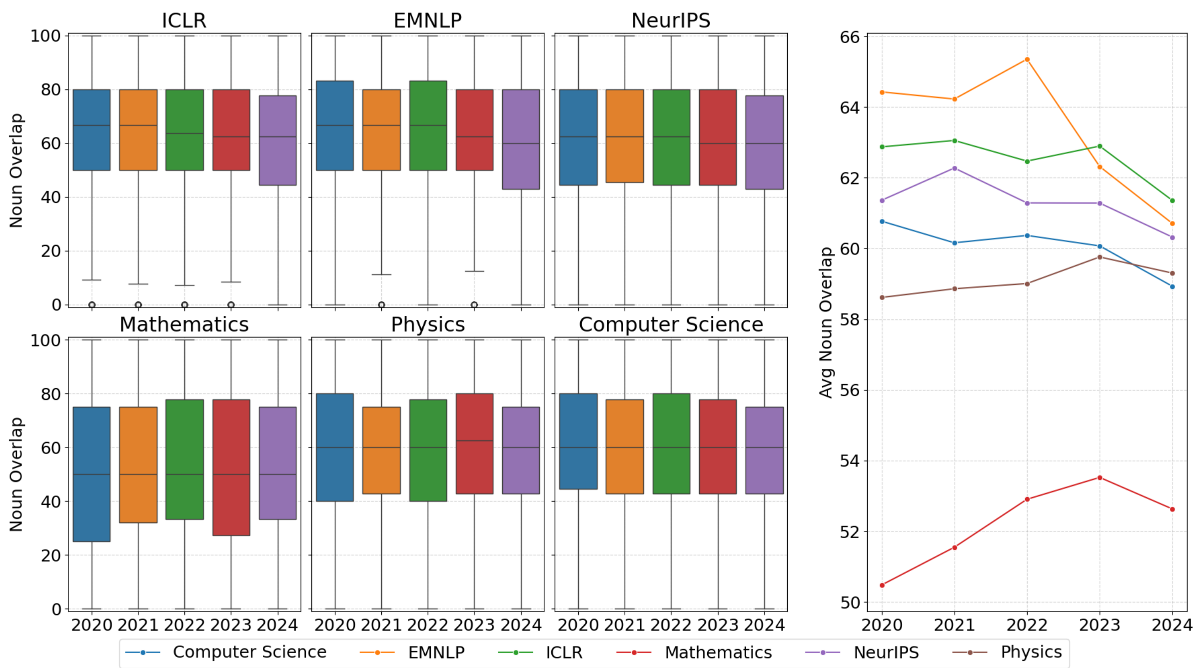

Figure 5.28: Noun overlap over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.6358	1.0835	1.0419	0.6179	1.3456	1.5552
WD 2021-2022	0.7247	1.1423	1.066	0.4096	1.3595	1.1491
WD 2022-2023	0.6425	3.059	0.3375	0.6134	0.9696	1.7373
WD 2023-2024	1.5369	1.6109	1.0604	1.2153	2.4351	1.6006

Table 5.26: Wasserstein distances for noun overlap over the years.

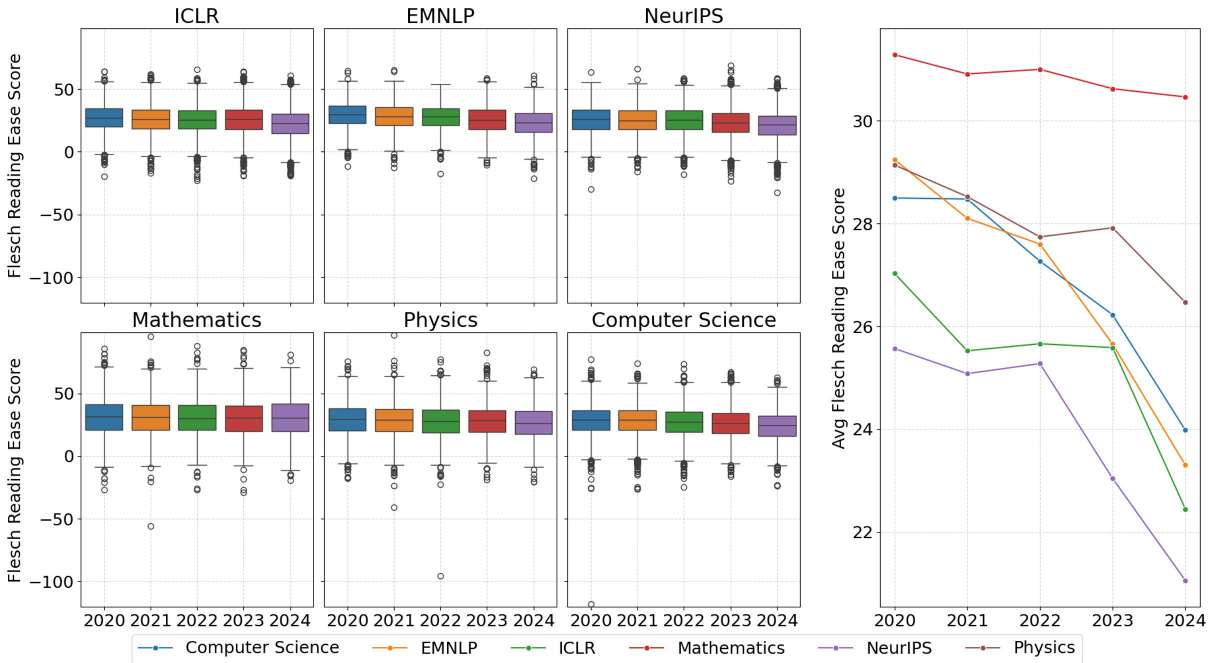

Figure 5.29: Flesch Reading Ease Score over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	1.5166	1.2194	0.6515	0.2407	0.8183	0.7712
WD 2021-2022	0.2302	0.5972	0.3443	1.2364	0.8572	0.8474
WD 2022-2023	0.219	2.1682	2.2656	1.1351	0.7556	0.6464
WD 2023-2024	3.142	2.36	1.9849	2.2467	1.3746	1.4784

Table 5.27: Wasserstein distances for Flesch Reading Ease Score over the years.

Readability declined over the years across all three metrics analyzed, suggesting a shift toward abstracts that are more difficult to understand. This aligns with previous findings showing that LLM revisions tend to reduce readability. Flesch Reading Ease Scores mainly range from 20 to 40 and have decreased across years, while the Gunning Fog and SMOG Indices range from 15 to 17 and have increased. These trends are even more pronounced in abstracts with higher proportions of AI-favored adjectives and adverbs, where average values are often significantly lower for the Flesch Reading Ease Scores and higher for the other two metrics. Additionally, Wasserstein distances tend to be larger for all three metrics in the most recent years, especially in 2024, suggesting LLM adoption may have accelerated the decrease in readability.

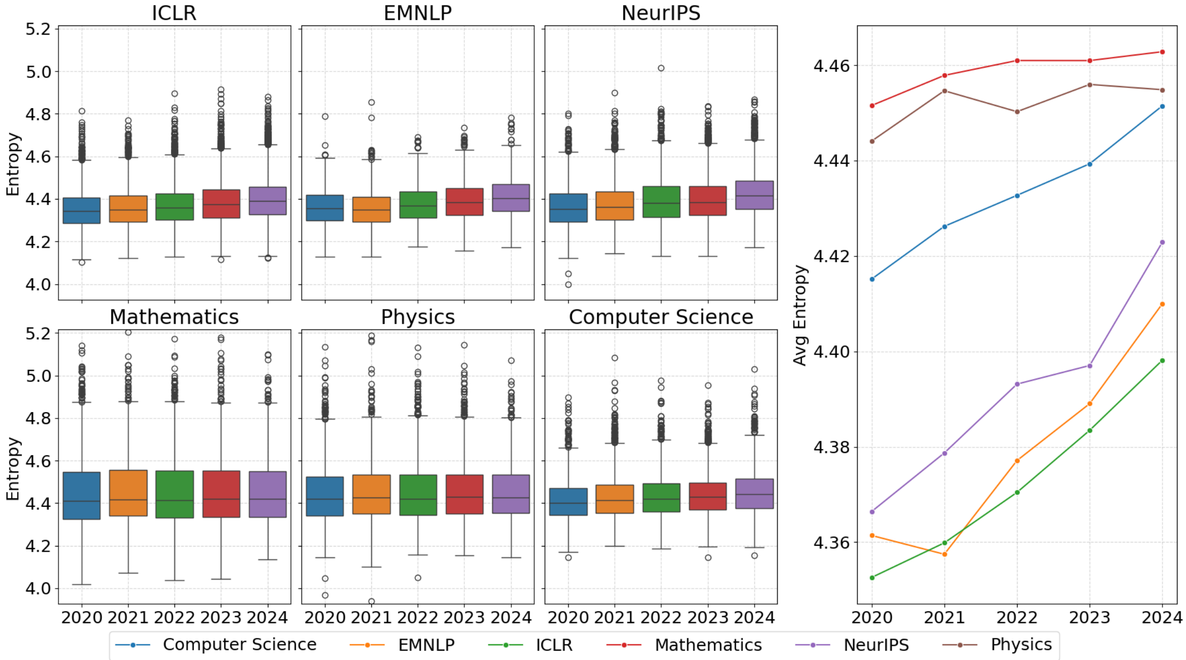

Figure 5.30: Entropy over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.0074	0.008	0.0123	0.0111	0.0104	0.0112
WD 2021-2022	0.0106	0.0203	0.0145	0.0073	0.0079	0.0076
WD 2022-2023	0.013	0.0128	0.0062	0.0071	0.0072	0.0086
WD 2023-2024	0.0149	0.021	0.0258	0.0122	0.006	0.0041

Table 5.28: Wasserstein distances for entropy over the years.

Most abstracts have an entropy between 4.3 and 4.5 bits per character, with values slightly growing over the years. This trend aligns with findings from LLM revision experiments, which suggest that LLMs increase entropy, editing abstracts to be more information-dense and unpredictable. The same pattern is observed in abstracts with higher proportions of AI-favored adjectives and adverbs.

The trends in verb mood, tense, and voice choices observed in LLM-revised abstracts largely appear in the temporal analysis as well. In most abstracts, active verbs account for 80–100% of total verbs, while passive verbs make up 0–20%. Present tense comprises 20–45% of verbs, past tense 15–40%, and future tense a very small percentage. The indicative mood is the most dominant, representing 95–100% of total verbs, while subjunctives range from 0–8%, and imperatives are almost absent. Over the years, passive voice, past and future tense, and the imperative and subjunctive moods have declined, especially in 2024, while active voice and the indicative mood have increased, leading to a more direct, immediate style with fewer complex passive constructions. These trends are particularly pronounced in abstracts with higher proportions of AI-favored adjectives and adverbs, where increasing metrics tend to have even higher averages, while decreasing ones show lower values. Moreover, Wasserstein distances are often greater after 2022, further suggesting that AI adoption in academia may have contributed to these shifts.

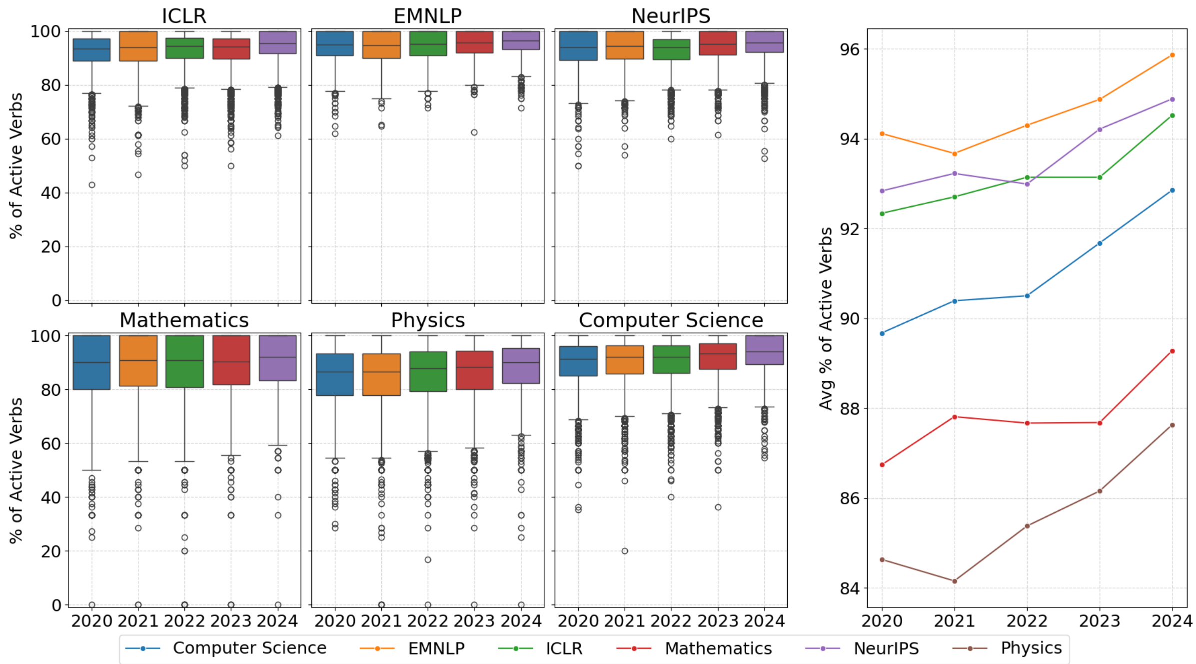

Figure 5.31: Percentage of active verbs over the years.

	ICLR	EMNLP	NeurIPS	Computer Science	Mathematics	Physics
WD 2020-2021	0.3732	0.6427	0.419	0.7294	1.1227	0.5151
WD 2021-2022	0.4626	0.6311	0.2699	0.2251	0.4195	1.2219
WD 2022-2023	0.0855	0.592	1.224	1.1776	0.6704	0.7811
WD 2023-2024	1.3822	1.0062	0.6845	1.1785	1.6038	1.5663

Table 5.29: Wasserstein distances for percentage of active verbs over the years.

Overall, this analysis of research paper abstracts published between 2020 and 2024 suggests that while the evidence that lexical and semantic similarity among academic texts has increased with the adoption of LLMs in academia is not particularly strong, significant linguistic shifts, many of which align with trends observed in LLM-generated revisions, indicate that AI may be influencing changes in academic writing styles.

Over the years, abstracts have grown longer in both word and sentence count. The increased prevalence of nouns, verbs, and adjectives, along with higher entropy and a preference for more varied and complex vocabulary, suggests a shift toward a more efficient, precise, and information-dense style. More recent abstracts feature shorter sentences and a tendency toward simpler syntax and straightforward active-voice constructions.

Despite this, readability has declined, likely due to the adoption of a denser, possibly more technical language and the reduced use of referential cohesion to explicitly link ideas together. Additionally, shifts in verb form preferences indicate a move toward a more direct, engaging, and immediate tone.

Notably, metrics that exhibit trends diverging from the results of previous experiments on LLM-generated revisions, often show a decrease in Wasserstein distances between consecutive yearly distributions after 2022, especially in the conferences dataset and the Computer Science category. This suggests that while LLM adoption may not have reversed existing trends, it may have slowed them down. Moreover, a notable pattern is that the Mathematics category often behaves as an outlier across many metrics, displaying trends that diverge from those of other categories. This aligns with Geng and Trotta (2024) [19], who observed that among arXiv categories, Mathematics appeared to be the least affected by the adoption of LLMs.

The results of this analysis confirm that LLM integration in academia has had a measurable impact on academic writing, with research paper abstracts gradually adopting certain LLM-like characteristics. These include a more direct tone, simpler constructions, and a richer, content-dense vocabulary, though often at the expense of readability and cohesion. As AI tools become more prominent in scientific research, future studies should examine how these stylistic shifts influence comprehension, engagement, and the overall accessibility of scholarly communication.

6 Conclusions

This study provided a comprehensive analysis of the impact of LLMs on academic writing by examining text similarity and linguistic trends in research paper abstracts from 2020 to 2024. The findings suggest that while LLMs might not have made abstracts significantly more lexically or semantically uniform, they are shaping academic writing styles in noticeable ways.

To better understand how LLMs shape writing and what linguistic features characterize AI-generated text, an analysis was conducted on LLM revisions of existing abstracts. The results revealed that during the revision process, LLMs tend to favor conciseness, a more diverse and content-dense vocabulary, and a direct and immediate tone, reducing syntactic complexity and passive constructions. At the same time, however, they also decrease readability and referential cohesion by adopting a more sophisticated vocabulary and reducing connectives, making logical relationships between ideas less explicit. Most of the modifications made by an LLM occur in the first revision, with subsequent revisions introducing only minor refinements. More importantly, abstracts from 2024 tend to require fewer modifications to match the "LLM writing style" compared to those from 2020, implying that many may already be written or edited by LLMs before submission. To further validate this observation, an LLM was tasked with generating and iteratively revising academic abstracts on AI-related topics. The comparison showed that AI-generated abstracts more closely resemble those from 2024 than those from 2020, reinforcing the hypothesis that the extensive use of LLMs had a relevant impact on academic writing.

The broader analysis of abstracts over time confirms many of the trends observed in LLM-revised texts. Abstracts have become longer, more lexically diverse, and information-dense, with increased entropy and a greater prevalence of content-heavy words like nouns and adjectives. Syntactic complexity has generally decreased, leading to the prevalence of simpler constructions, but readability has also declined, possibly due to the growth of variety and complexity in lexicon, and the reduced use of cohesive devices. While many of these shifts align with LLMs' preferences, a few of them don't. For example, metrics such as sentence and word count deviate from patterns observed in LLM-generated revisions, suggesting that factors beyond AI adoption may have played a larger role in shaping the evolution of some academic writing conventions. However, the decline in Wasserstein distances between yearly distributions after 2022, particularly for those categories that are more closely related to AI, suggests that while LLMs may not be completely reversing linguistic trends, they may be slowing them down.

These findings raise important questions about the role of LLMs in academia. One major concern is the potential loss of originality, as standardization in writing styles could lead to more uniform, less distinctive academic texts. The growing adoption of LLMs might cause a homogenization of academic writing, reducing stylistic diversity and individuality. Furthermore, the decline in readability and cohesion suggests that while LLMs may increase efficiency in scientific research, it might come at the cost of clarity, potentially making research papers more difficult to understand, especially for broader audiences.

However, standardization is not necessarily negative. In scientific fields, consistency in writing styles can enhance clarity, making research findings easier to interpret and compare across studies. Moreover, for non-native English speakers, LLMs can serve as valuable tools for improving grammar, structure, and vocabulary, helping to reduce linguistic barriers in academia. Balancing the benefits of standardization with the need to preserve individual expression and intellectual diversity in academic writing is a challenge that requires thoughtful solutions.

Future research should explore these dynamics further. Expanding the analysis to additional datasets, such as research papers from various disciplines or news articles, could help determine whether these trends are domain-specific or more widespread, as well as the extent to which LLMs are influencing different sectors. Additional metrics like sentiment analysis could help assess how LLMs impact framing in scholarly communication. Moreover, further investigations into semantic similarity could provide clearer insights into whether AI-assisted writing is making research papers not just stylistically similar, but also more conceptually uniform. Lastly, a deeper examination of how LLM-generated text influences peer review, citation patterns, and reader comprehension could shed light on the long term implications of AI in academic publishing.

As LLMs continue to shape academic writing, understanding their impact is crucial for maintaining both the accessibility and diversity of academic discourse. Careful integration of AI tools, coupled with awareness of their limitations, will be essential in ensuring that advances in technology contribute to, rather than hinder, the progress of knowledge dissemination.

References

[1] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., Polosukhin, I. (2017). *Attention Is All You Need.* ArXiv. https://doi.org/10.48550/arXiv.1706.03762.

[2] OpenAI. (2023). *Introducing ChatGPT.* https://openai.com/index/chatgpt/.

[3] Weise, K., Metz, C., Grant, N., Isaac, M. (2023). *Inside the A.I. Arms Race That Changed Silicon Valley Forever.* The New York Times. https://www.nytimes.com/2023/12/05/technology/ ai-chatgpt-google-meta.html.

[4] Noy, S., Zhang, W. (2023). *Experimental evidence on the productivity effects of generative artificial intelligence.* Science. Vol 381, Issue 6654 pp. 187-192. DOI: 10.1126/science.adh2586.

[5] Ray, P. P. (2023). *ChatGPT: A comprehensive review on background, applications, key challenges, bias, ethics, limitations and future scope*. Internet of Things and Cyber-Physical Systems. https://doi.org/10.1016/j.iotcps.2023.04.003.

[6] Haque, A., Li S. (2024). *Exploring ChatGPT and its impact on society.* Springer. https://doi.org/10.1007/s43681-024-00435-4.

[7] Pinzolits R. (2023). *AI in academia: An overview of selected tools and their areas of application*. MAP Education and Humanities, 4, 37-50. https://doi.org/10.53880/2744-2373.2023.4.37.

[8] Semantic Scholar (2025). *About Semantic Scholar.* https://www.semanticscholar.org/about.

[9] Banerjee S., Agarwal A., Singla S. (2024). *LLMs Will Always Hallucinate, and We Need to Live With This*. arXiv. https://doi.org/10.48550/arXiv.2409.05746.

[10] Chelli M., Descamps J., Lavoué V., Trojani C., Azar M., Deckert M., Raynier J., Clowez G., Boileau P., Ruetsch-Chelli C. (2024). *Hallucination Rates and Reference Accuracy of ChatGPT and Bard for Systematic Reviews: Comparative Analysis.* J Med Internet Res 2024;26:e53164. DOI: 10.2196/53164.

[11] Barnett, A. (2023). *Scientific fraud is rising, and automated systems won't stop it. We need research detectives*. The Conversation. https://theconversation.com/scientific-fraud-is-rising-and-automated-systems-wont-stop-it-we-need-research-detectives-206235.

[12] Ayoub N. F., Balakrishnan, K., Ayoub, M. S., Barrett, T. F., David, A. P., Gray, S. T. (2024). *Inherent Bias in Large Language Models: A Random Sampling Analysis*. Mayo Clinic Proceedings: Digital Health, Vol. 2, Issue 2, p177-288. DOI: 10.1016/j.mcpdig.2024.03.003.

[13] Barman, D., Guo, Z., Conlan, O. (2024). *The Dark Side of Language Models: Exploring the Potential of LLMs in Multimedia Disinformation Generation and Dissemination.* Machine Learning with Applications Volume 16, June 2024, 100545. https://doi.org/10.1016/j.mlwa.2024.100545.

[14] Susnjak, T. (2022). *ChatGPT: The End of Online Exam Integrity?* arXiv. https://doi.org/10.48550/arXiv.2212.09292.

[15] Gegg-Harrison, W., Quarterman, C. (2024). *AI Detection's High False Positive Rates and the Psychological and Material Impacts on Students.* Academic Integrity in the Age of Artificial Intelligence (pp.199-219). DOI:10.4018/979-8-3693-0240-8.ch011.

[16] Stadler, M., Bannert, M., Sailer, M. (2024). *Cognitive ease at a cost: LLMs reduce mental effort but compromise depth in student scientific inquiry.* Computers in Human

Behavior, Volume 160, November 2024, 108386. https://doi.org/10.1016/j.chb.2024.108386.

[17] Bellaiche, L., Shahi, R., Turpin, M.H. et al. (2023). *Humans versus AI: whether and why we prefer human-created compared to AI-created artwork*. Cogn. Research 8, 42 (2023). https://doi.org/10.1186/s41235-023-00499-6.

[18] Liang, W., Izzo, Z., Zhang, Y., Lepp, H., Cao, H., Zhao, X., Chen, L., Ye, H., Sheng, L., Huang, Z., McFarland, D. A., Zou, J. Y. (2024). *Monitoring AI-Modified Content at Scale: A Case Study on the Impact of ChatGPT on AI Conference Peer Reviews*. arXiv. https://doi.org/10.48550/arXiv.2403.07183.

[19] Geng, M., Trotta, R. (2024). *Is ChatGPT Transforming Academics' Writing Style?* arXiv. https://doi.org/10.48550/arXiv.2404.08627.

[20] AlAfnan, M. A. & MohdZuki, S. F. (2023). *Do Artificial Intelligence Chatbots Have a Writing Style? An Investigation into the Stylistic Features of ChatGPT-4.* Journal of Artificial Intelligence and Technology, 2023, 3, 85-94. https://doi.org/10.37965/jait.2023.0267.

[21] Lund, B. D. & Wang, T. (2023). *Chatting about ChatGPT: how may AI and GPT impact academia and libraries?* Library Hi Tech News Number 3, 2023, pp. 26-29. DOI 10.1108/LHTN-01-2023-0009.

[22] Watermeyer, R., Phipps, L., Lanclos, D., Knight, C. (2023) *Generative AI and the Automating of Academia.* Postdigital Science and Education (2024) 6:446–466 https://doi.org/10.1007/s42438-023-00440-6.

[23] Gomaa, W. H., & Fahmy, A. A. (2013). *A survey of text similarity approaches*. International Journal of Computer Applications, 68 (13), 13-18. DOI:10.5120/11638-7118.

[24] Chandrasekaran, D., & Mago, V. (2021). *Evolution Of Semantic Similarity - A Survey*. ACM Computing Surveys 54 (2), 1-37. https://doi.org/10.1145/3440755.

[25] González-Márquez, R. & Kobak, D. (2024). *Learning representations of learning representations*. arXiv. https://doi.org/10.48550/arXiv.2404.08403.

[26] NeurIPS (2025). *NeurIPS Proceedings.* https://papers.nips.cc/.

[27] ACL (2025). *ACL Anthology.* https://aclanthology.org/volumes/.

[28] Kaggle (2024). *arXiv Dataset. https://www.kaggle.com/datasets/Cornell-University/arxiv/.*

[29] Spärck Jones, K. (1972). *A Statistical Interpretation of Term Specificity and Its Application in Retrieval.* Journal of Documentation. 28 (1): 11–21. DOI:10.1108/eb026526.

[30] https://scikit-learn.org/stable/index.html.

[31] Devlin, J., Chang, M. W., Lee, K., & Toutanova, K. (2018). *BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding*. arXiv. https://doi.org/10.48550/arXiv.1810.04805.

[32] Reimers, N., & Gurevych, I. (2019). *Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks*. arXiv. https://doi.org/10.48550/arXiv.1908.10084.

[33] https://huggingface.co/sentence-transformers/paraphrase-MiniLM-L6-v2.

[34] Perry, D. (2020). *Writing Styles.* University of Rhode Island Graduate Writing Center. https://web.uri.edu/graduate-writing-center/writing-styles/.

[35] Hu, H., Wang, D., Deng, S. (2021). *Analysis of the scientific literature's abstract writing style and citations.* Online Information Review Vol. 45 No. 7, 2021 pp. 1290-1305. DOI 10.1108/OIR-05-2020-0188.

[36] Cambridge University Press. *English grammar reference*. Cambridge Dictionary.

https://dictionary.cambridge.org/grammar.

[37] Strunk, W., & White, E. B. (1999). *The Elements of Style* (4th ed.). Pearson.

[38] https://spacy.io/.

[39] Neumann, M., King, D., Beltagy, I., Ammar W. (2019). *ScispaCy: Fast and Robust Models for Biomedical Natural Language Processing.* ArXiv. https://doi.org/10.18653/v1/W19-5034.

[40] Torruella, J., Capsada, R. (2013). *Lexical Statistics and Tipological Structures:A Measure of Lexical Richness*. Procedia-Social and Behavioral Sciences, 95, 447-454. https://doi.org/10.1016/j.sbspro.2013.10.668.

[41] Tweedie, F. J. & Bayeen, R. H. (1998). *How Variable May a Constant be? Measures of Lexical Richness in Perspective.* Computers and the Humanities 32: 323–352, 1998. https://doi.org/10.1023/A:1001749303137.

[42] Yule, C. U. (2014). *The Statistical Study of Literary Vocabulary*. Cambridge University Press.

[43] https://pypi.org/project/lexicalrichness/.

[44] Di Marco, N., Loru, E., Bonetti, A., Serra, A. O. G., Cinelli, M., Quattrociocchi, W. (2024). *Patterns of linguistic simplification on social media platforms over time.* PNAS 2024 Vol. 121 No. 50 e2412105121. https://doi.org/10.1073/pnas.2412105121.

[45] Zipf, G.K. (1935). *The Psychobiology of Language*. New York, NY: Houghton-Mifflin.

[46] Zipf, G. K. (1949). *Human behavior and the principle of least effort*. Addison-Wesley.

[47] Piantadosi S. T. (2014). *Zipf's word frequency law in natural language: a critical review*

and future directions. Psychonomic bulletin & review, 21(5), 1112–1130. https://doi.org/10.3758/s13423-014-0585-6.

[48] Altmann, E. G., Pierrehumbert, J. B., Motter, A. E. (2011) *Niche as a Determinant of Word Fate in Online Groups.* PLOS. https://doi.org/10.1371/journal.pone.0019009.

[49] Petersen, A., Tenenbaum, J., Havlin, S. et al. (2012). *Statistical Laws Governing Fluctuations in Word Use from Word Birth to Word Death.* Sci Rep 2, 313. https://doi.org/10.1038/srep00313.

[50] Dubay, W. H. (2004). *The Principles of Readability*. CA. 92627949. 631-3309.

[51] Klare, G. R. (1968). The Role of Word Frequency in Readability. Elementary English, 45(1), 12–22. http://www.jstor.org/stable/41387663.

[52] Lu, X. (2010). *Automatic analysis of syntactic complexity in second language writing.* International Journal of Corpus Linguistics, 15(4), 474-496. https://doi.org/10.1075/ijcl.15.4.02lu.

[53] Hunt, K. (1965). *Grammatical structures written at three grade levels*. NCTE Research Report No. 3. Champaign, IL, USA: NCTE.

[54] https://github.com/mholtzscher/syllapy.

[55] Halliday, M.A.K, Ruqayia H. (1976): *Cohesion in English*. London: Longman.

[56] Graesser, A.C., McNamara, D.S., Louwerse, M.M., Cai, Z. (2004). *Coh-Metrix: Analysis of text on cohesion and language*. Behavior Research Methods, Instruments, & Computers 36, 193–202. https://doi.org/10.3758/BF03195564.

[57] Klare, G. R. (1963). *The measurement of readability*. Ames, Iowa: Iowa State University Press.

[58] McLaughlin, G. H. (1969). *SMOG grading - a new readability formula*. Journal of reading 22:639-646.

[59] Zamanian, M. & Heydari, P. (2012). *Readability of Texts: State of the Art.* Theory and Practice in Language Studies, Vol. 2, No. 1, pp. 43-53, January 2012. DOI:10.4304/tpls.2.1.43-53.

[60] Flesch, R. (1948). *A new readability yardstick.* Journal of Applied Psychology 32:221-233.

[61] Flesch, R. (1949). *The art of readable writing*. New York: Harper.

[62] Gunning, R. (1952). *The technique of clear writing*. McGraw-Hill.

[63] https://pypi.org/project/textstat/.

[64] Shannon, C. E. (1948). *A Mathematical Theory of Communication*. Bell System Technical Journal. 27 (3): 379–423. DOI:10.1002/j.1538-7305.1948.tb01338.x.

[65] Bryson, S. (2023). V*erb Tenses in Academic Writing | Rules, Differences & Examples.* Scribbr. https://www.scribbr.com/verbs/tenses/.

[66] *English Grammar and Usage Unit 9 – Mood and Voice*. Fiveable. https://library.fiveable.me/english-grammar-usage/unit-9.

[67] https://ai.google.dev/gemini-api/docs/models/gemini#gemini-1.5-flash.

Acknowledgements

I would like to thank my supervisor, Prof. Alessandro Galeazzi, for the opportunity to work on this project and for his guidance throughout the research process, as well as my family and friends for their support over the years.