Exploration of Large Language Model Usage in Computer Technology Patent Texts

Kyunam Park
School of Business and Technology Management
Korea Advanced Institute of Science and Technology
Daejeon, South Korea
erised10@kaist.ac.kr

Hangjung Zo
School of Business and Technology Management
Korea Advanced Institute of Science and Technology
Daejeon, South Korea
joezo@kaist.edu

***Abstract*—Using Large language models (LLMs) to augment human authors in writing professional texts leaves traceable evidence. By examining computer technology patents filed in the United States, this study investigates latent patterns that can be attributed to the widespread adoption of LLMs by finding word tokens that became more prevalent after the release of ChatGPT. After manually reviewing and excluding tokens attributed to inventions' technological elements, we found one token that has been used more frequently after the release of ChatGPT across all parts of patent documents. This research extends the study of LLM's widespread impact on academic manuscripts into examining technical documents where LLMs as an augmentation tool are actively explored. The results of this study show that LLM's impact is observable and quantifiable on the scale of a specific corpus, which has a significant implication for both researchers who study existing documents and practitioners who author new documents.**

***Index Terms*—large language model, patent, textual analysis**

I. Introduction

Generative artificial intelligence (GenAI) can improve users' quantitative and qualitative productivity in professional tasks [1]. However, several limitations of using genAI to augment writing texts have been pointed out, especially when drafting and reviewing technical documents [2], [3]. Therefore, contemporary efforts have been made to refine methodologies to detect, review, and audit documents for signs of genAI usage. [4] One approach within this field of study observes the impact of large language models (LLMs) on the scale of academic fields by examining changes in word frequencies among academic publications [5]. This study explores what features within professional documents are more likely to rely on the assistance of LLMs by observing US computer technology patents for statistical signs associated with AI-generated texts.

This research was supported by the MSIT (Ministry of Science and ICT), Korea, under the ITRC (Information Technology Research Center) support program (IITP-2024-2020-0-01787) supervised by the IITP (Institute of Information & Communications Technology Planning & Evaluation)

II. Literature Review

A. LLM usage, and detection of LLM usage

Contemporary LLMs are advanced enough to provide human professionals with summaries and categorizations of existing professional documents [6]. LLMs are expected to further augment human practitioners in analyzing and drafting [7] complex technical documents. LLMs, however, cannot discern the unusual from the typical in an explainable manner [2]. Moreover, it should not replace the process of authoring works that need to be original and accountable [3]. Accordingly, recent literature has investigated methods to detect individual documents written with an LLM's help [4] and the severity of misconduct within [3]. However, in quantitative analyses of many documents or exploratory studies for explainable patterns, approaches that do not have to review individual documents should be more useful and efficient.

B. Observing LLM adoption via keyword frequencies

LLMs use some words disproportionately more often in generating texts than human writers would do, and when enough proportion of documents were written with the help of LLMs, the difference becomes significant enough to change word frequencies of a corpus [5]. By exploring this pattern, we can study and quantify LLM's widespread impact over a collection of documents, which is not feasible when each document is examined individually. A study on academic manuscripts has shown that this approach can find what words have changed in their frequencies after the widespread adoption of LLMs and be further refined by discovering and cataloging additional keywords [3]. We apply this methodology to examine patent documents and find tokens associated with LLM augmentation.

III. Empirical Methodology and Analysis

From Patentsview we gathered documents of granted patents that were filed to the United States Patent and Trademark Office (USPTO) between January 1st, 2019, and August 31st, 2023[1]. To control the intensity of genAI

[1]The data was accessed from https://patentsview.org/download/data-download-tables on Aug. 23, 2024.

and LLM adoption between industries, we confine patents' technology field to computer technology. This was done by selecting patents whose first classification symbols were within either a) the G06 class sans the G06Q subclass, b) the G11C subclass, or c) the G10L subclass [8]. We then extract abstracts, claims, brief summary texts, and detail description texts from 155,889 computer technology patents to tokenize, stem, and eliminate stopwords [9]. For each token not included in computer technology classification titles[2], we compute their appearance ratio in a given month to find which word tokens became more frequent after ChatGPT's public release in November 2022.

IV. Results

Table 1. shows the number of tokens whose frequency increase lies within the top 0.1% of all measured tokens in a document section and how many overlap between two or more sections. Seven unique tokens have increased in frequency after November 2022 in all four parts of patent documents. These were, in alphabetical order, "baseboard," "bmc," "dedupl," "embodi," "fetch," "tier," and "wordlin." Among these, all but "embodi" can be attributed to industry-specific expressions: "Baseboard" and "bmc" to baseboard management controller (BMC), "tier" and "wordlin" to memory devices and memory management algorithms, "dedupl" to "deduplication" truncated by snowball stemming, and "fetch" to transactions of data within a system or a device. Fig. 1 compares the frequency of the token "embodi" with the token "fetch" within the dataset in the unit of months.

V. Conclusion

This study identified tokens that increased in frequency after the ChatGPT's deployment and reviewed each for their point of origin. It is most notable that in documents written to satisfy rigorous guidelines and legal regulations, a keyword that has increased in usage on a scale comparable to changes in technological trajectories can exist. On the other hand, the result shows fluctuations in the popularity of different subjects should have obscured more potential discoveries of LLM's latent effect on word frequency. Furthermore, limiting the dataset to granted patents leaves out more recent and still pending patent applications, where further salient examples could have been discovered from more recently authored documents.

These limitations notwithstanding, our study shows that it is possible to discover and measure how an industry sector's documents have changed after LLMs were introduced to augment practitioners. Future improvements can introduce ways to control the effects of technological trends so that even more changes can be found and reliably attributed to the introduction of LLMs. These findings may provide a broad overview of the technology field's change after adopting LLMs and help researchers in future studies choose which specific set of individual documents may be reviewed and analyzed more precisely. For using LLMs to augment writing to not hinder practitioners' ability to accurately differentiate each topic [2], practitioners should also be more vigilant for signs of LLMs suggesting certain words and phrases more often than deserved. With these results, our study expects to contribute to the literature of examining a corpus of technical and professional documents to study trends in a technology field.

[2]We assume words in classification titles to be instrumental in precise description of the invention, with or without using LLMs.

TABLE I
Token Counts by Category

Category	Tokens	Category	Tokens
Abstract (A)	42	C, D	48
Claims (C)	108	B, D	94
Brief summary (B)	256	C, B, D	28
Detail description (D)	10,229	A, B, D	9
A, C	20	A, C, D	10
A, B	17	A, C, B	9
A, D	14	A, C, B, D	7
C, B	45		

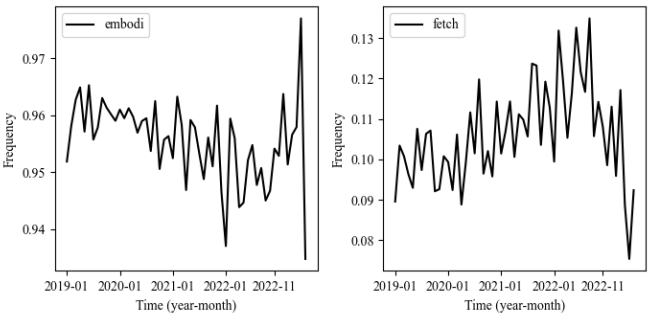

Fig. 1. Frequency of "embodi" (left) and "fetch" (right) per month

References

[1] S. Noy and W. Zhang, "Experimental evidence on the productivity effects of generative artificial intelligence," *Science*, vol. 381, no. 6654, pp. 187-192, 2023.

[2] C. Preiksaitis, C. A. Sinsky, and C. Rose, "ChatGPT is not the solution to physicians' documentation burden," *Nature Medicine*, vol. 29, no. 6, pp. 1296-1297, 2023.

[3] H. H. Thorp, "ChatGPT is fun, but not an author," *Science*, vol. 379, no. 6630, p. 313, 2023.

[4] A. Akram, "An Empirical Study of AI-Generated Text Detection Tools," *Advances in Machine Learning and Artificial Intelligence*, vol. 4, no. 2, pp. 44-55, 2023.

[5] A. Gray, "ChatGPT contamination: estimating the prevalence of LLMs in the scholarly literature," *arXiv preprint arXiv:2403.16887*, 2024.

[6] S. Pelaez, G. Verma, B. Ribeiro, and P. Shapira, "Large-scale text analysis using generative language models: A case study in discovering public value expressions in AI patents," *Quantitative Science Studies*, vol. 5, no. 1, pp. 153-169, 2024.

[7] J. S. Lee, "Evaluating generative patent language models," *World Patent Information*, vol. 72, p. 102173, 2023.

[8] U. Schmoch, "Concept of a technology classification for country comparisons," Final report to the World Intellectual Property Organisation (WIPO), WIPO, 2008.

[9] S. Arts, J. Hou, and J. C. Gomez, "Natural language processing to identify the creation and impact of new technologies in patent text: Code, data, and new measures," *Research Policy*, vol. 50, no. 2, p. 104144, 2021.