Copyright Detection in Large Language Models: An Ethical Approach to Generative AI Development

David Szczecina
University of Waterloo
david.szczecina@uwaterloo.ca

Senan Gaffori
University of Waterloo
senan.gaffori@uwaterloo.ca

Edmond Li
University of Waterloo
e26li@uwaterloo.com

***Abstract*—The widespread use of Large Language Models (LLMs) raises critical concerns regarding the unauthorized inclusion of copyrighted content in training data. Existing detection frameworks, such as DE-COP, are computationally intensive, and largely inaccessible to independent creators. As legal scrutiny increases, there is a pressing need for a scalable, transparent, and user-friendly solution. This paper introduce an open-source copyright detection platform that enables content creators to verify whether their work was used in LLM training datasets. Our approach enhances existing methodologies by facilitating ease of use, improving similarity detection, optimizing dataset validation, and reducing computational overhead by 10-30% with efficient API calls. With an intuitive user interface and scalable backend, this framework contributes to increasing transparency in AI development and ethical compliance, facilitating the foundation for further research in responsible AI development and copyright enforcement.**

I. Introduction

A. Motivation

Large Language Models (LLMs) such as GPT-4 and Claude have revolutionized natural language processing, but also raise legal and ethical concerns about the unauthorized use of copyrighted content in training datasets [1]. Proprietary models often rely on large-scale web scraping [2], incorporating copyrighted material without clear consent mechanisms, compensation, and intellectual property protection [3].

A major concern is the lack of compensation for content creators whose work is used without permission. Legal frameworks for AI copyright enforcement are rapidly evolving, with landmark cases like New York Times v. OpenAI [4] bringing increased scrutiny to dataset curation. Transparency in AI training datasets is essential to ensure responsible and ethical development. Research indicates that as models increase in size, memorization tendencies become more pronounced, particularly in models exceeding 100 billion parameters [4], increasing the risk of unauthorized reproduction of copyrighted content.

Current detection methods, such as plagiarism checkers and statistical techniques, struggle to identify subtly paraphrased copyrighted content [2] [5]. While frameworks such as DE-COP offer promising approaches, they remain computationally expensive and complex; making them impractical for independent creators and smaller organizations. A scalable, cost-effective, and user-friendly solution is needed to verify whether copyrighted works have been used in LLM training datasets.

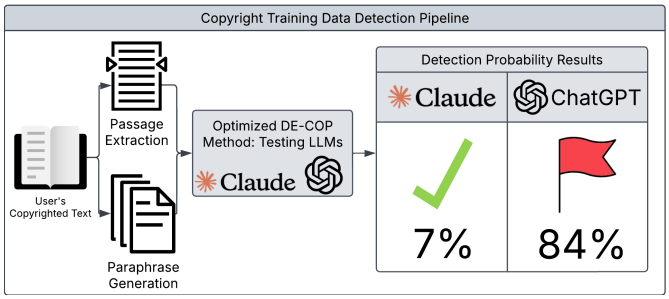

Fig. 1. Unique passages are extracted and paraphrased from a users content, next an LLM is prompted to determine the original passage. Final scores show the probability of the copyrighted content being used in training the LLM

B. Related Works

The detection of copyrighted content in LLM training datasets has been the subject of increasing research attention, particularly as legal and ethical concerns surrounding dataset curation intensify. While traditional plagiarism detection tools struggle to identify AI-generated reproductions of proprietary content [2], several machine learning-based approaches have been proposed to address this issue.

Membership inference attacks [6] analyze a model's confidence scores to determine whether a given text sample was likely included in the training data. Although effective in controlled experiments, this approach requires adversarial access to the model and often produces inconclusive results due to dataset augmentation and model fine-tuning techniques. Similarly, perplexity-based analysis is another detection approach by evaluating how confidently an LLM predicts a passage of text [7]. Low perplexity scores suggest memorization, however, this method struggles to distinguish between legally sourced and unauthorized content, making it unreliable for copyright enforcement. Another proposed approach is digital watermarking [8], where imperceptible markers are embedded into text data before model training. While useful for tracking known copyrighted works, watermarking is ineffective against existing datasets that were scraped from the web and fails to detect content that has been paraphrased or restructured.

A more recent approach, DE-COP: Detecting Copyrighted Content in Language Models Training Data, [2], introduces a method to determine whether a language model has memorized copyrighted content. Unlike statistical approaches, DE-COP introduces a multiple-choice question-answering framework,

where an LLM must distinguish an original verbatim passage from paraphrased alternatives. If a model consistently selects the correct passage, this suggests that the content was likely included in its training data. An overview of the DE-COP system is illustrated in Figure 2. Despite its advantages, DE-COP is computationally expensive, requiring approximately 590 seconds per book for open-source models (LLaMA-2 70B) [9] and 331 seconds on ChatGPT [2] [10]. Methods such as Min-K%-Prob [7], Prefix Probing [11] and Name Cloze Task [12] only required 13-17 seconds to perform the same task [2]. Additionally, the datasets presented in DE-COP were found to contain NULL values, errors message outputs, half finished sentences, and new paraphrases ranged from being 20% to 250% as long as the original passage [2]. DE-COP lacks robust features to handle these errors in its own dataset, and its evaluation metrics were based on questionable data, leaving lots of room for improvements.

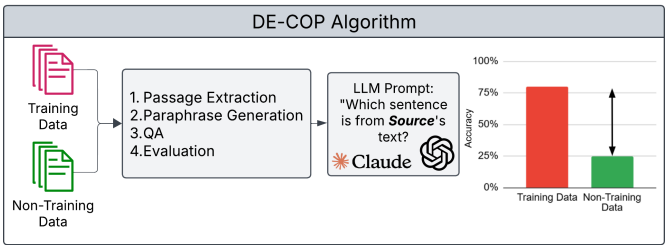

Fig. 2. DE-COP System Overview

While previous methods provide partial solutions to the problem of detecting copyrighted content in LLM training data, they often fall short in generalization and effectiveness. DE-COP introduced a black-box-compatible alternative that significantly improves detection accuracy [2]. However, optimizing its computational efficiency and reducing selection biases remains an open challenge for future work.

C. Problem Definition

Despite the concerns for copyright material in LLM training data, existing copyright detection methods remain insufficient and inaccessible. Traditional plagiarism detection tools struggle to identify paraphrased or subtly modified copyrighted content, making enforcement difficult [2]. Additionally, computationally intensive frameworks such as DE-COP, are impractical for independent content creators due to their technical complexity and high computational costs [13]. The absence of cost-effective and user-friendly solutions further limits the ability to verify whether copyrighted works have been used in LLM training. This paper introduces an open-source framework that enhances dataset validation, improving similarity detection, and optimizes computational efficiency. By significantly reducing processing costs while maintaining detection accuracy, our approach provides a scalable and accessible platform for copyright verification. This initiative ensures AI transparency, promotes fair compensation for content creators, and supports ethical AI development in the rapidly evolving landscape of generative AI.

II. Methodology

This project features a web-based UI where users can submit content for evaluation. The backend evaluation system, runs a multi-layered evaluation workflow integrating passage extraction, paraphrase generation, question-answering, multiple-choice evaluation, and statistical analysis to detect copyrighted content in LLM training data. A vector store maintains a record of previously evaluated content, allowing the system to check for duplicates, avoiding redundant evaluations, as well as search through past evaluations. Users can access a dashboard and analytics page to view evaluation histories and check accuracy metrics. An overview of the system architecture is illustrated in Figure 3.

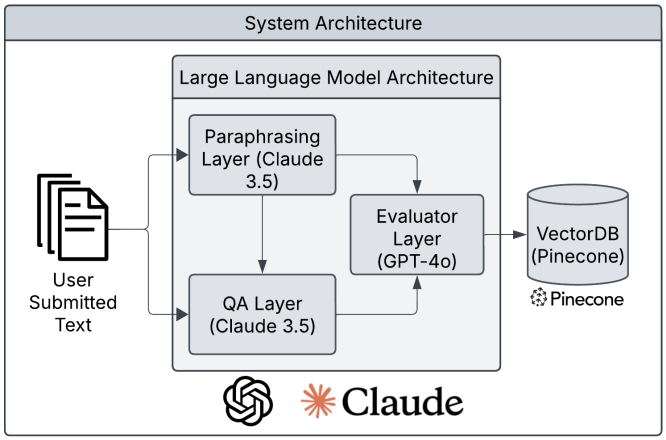

Fig. 3. System Architecture Diagram

A. Passage Extraction

Selecting highly unique passages enhances the accuracy and effectiveness of detecting copyrighted content in language model training data [14]. Unique passages minimize the risk of incorporating common phrases and generic text that may not sufficiently challenge the model's memorization capabilities [15]. To identify these passages, the BM25 algorithm [16] was employed to vectorize passages within the document and calculate similarity scores between them. By treating each passage as a query against the entire document, BM25 assigned scores based on term frequency and inverse document frequency. Passages with the lowest BM25 scores, indicating minimal similarity to other passages, were considered the most unique. These high-uniqueness passages were prioritized for use in the evaluation layer, providing a robust foundation for detecting memorized content.

B. Paraphrase Generation

The paraphrase generation layer is implemented using LangGraph's StateGraph to create a modular and dynamic workflow. This layer utilizes the Claude 3.5 Sonnet model [17] via the ChatAnthropic API with a temperature setting of 0.7, ensuring a balance between creativity and control in generating paraphrases. Unlike the original DE-COP approach, which applies standard paraphrasing prompts, our method

introduces specific paraphrasing strategies, including passive voice conversions, question-based restructuring, and language simplification. These templates promote greater diversity in paraphrases [18], enhancing the robustness of the evaluation by reducing model prediction patterns. Additionally, the implementation proposes XML formatting for paraphrases to support integration with instructional models, offering improved compatibility and structured data handling not present in the original DE-COP method.

C. Question-Answering

The QA layer is also built using LangGraph's StateGraph, facilitating an automated workflow that handles both "create" and "format" modes for generating evaluation questions. While the original DE-COP [2] method primarily focused on generating standardized multiple-choice questions, our implementation expands functionality by allowing the creation of custom questions that use exact text from the input content. The QA layer uses the ChatAnthropic model to generate questions in a structured JSON format, improving downstream processing and maintaining output consistency. This added flexibility enhances the evaluation's accuracy by testing the model's memorization across varied question formats, contributing to a more thorough assessment of the model's exposure to copyrighted content.

D. Multiple-Choice

The multiple-choice layer employs LangGraph to manage answer selection and evaluation workflows. In contrast to the original DE-COP's [2] exhaustive approach of generating all permutations of answer choices, our initial implementation used a simplified randomization strategy to prevent selection bias. However, we propose an enhancement that includes a dedicated permutation function to fully automate all possible answer orderings within LangGraph. The evaluation prompts are designed to elicit concise, formatted responses from the model, minimizing noise and ensuring clarity in the output. Incorporating full permutation handling would better strengthen the mitigation of selection biases in model responses.

E. Evaluation

The evaluation layer integrates multiple components, including paraphrase generation, question answering, multiple-choice testing, and statistical analysis, using GPT-4o [10] via LangGraph. Our implementation extends upon DE-COP's framework by incorporating advanced statistical methodologies such as receiver-operating characteristic (ROC) curve analysis, area under the curve (AUC) scoring, and hypothesis testing. This layer provides deeper insights into performance through robust statistical methods. A key enhancement over previous methodologies is the introduction of a permutation function which generates all answer permutations; mitigating selection biases in LLMs. The evaluation prompts guide the model through a structured evaluation process, emphasizing precise and formatted responses. These enhancements create a more modular and statistically robust framework, improving the accuracy and reliability of detecting copyrighted content in LLM training data.

F. Logging System and Similarity Search

To enable content tracking and retrieval, the system incorporates Pinecone, a serverless vector database. Documents are embedded using all-MiniLM-L6-v2 [19] (embedding model from HuggingFace) which offers a strong balance of embedding quality and efficiency. The model generates 384-dimensional embeddings to support fast and accurate approximate nearest neighbour (ANN) searches, while integrating seamlessly with Pinecone and LangGraph. Metadata attributes such as copyright ownership, evaluation timestamps, evaluation results, and content type, are stored directly in Pinecone as key-value pairs. Logging metedata enables quick access and tractability during content evaluations without requiring an external database. The ingestion pipeline is designed for single-document processing, embedding each submission and storing it with a unique identifier. To evaluate content, the system compares new submissions against stored vectors, retrieving the most similar documents and their metadata. This streamlined vectorized approach supports the goal of creating an open-source API that logs copyrighted content appearing in LLM training data; promoting transparency and accountability in AI development.

G. Data Processing Improvements

An analysis of DE-COP's dataset revealed several inconsistencies such as NULL values, API output errors, inconsistent formatting, and extreme variations in passage length [2]. These inconsistencies negatively impacted accuracy, skewed model predictions and increased token usage by up to 50%. To address this, a preprocessing pipeline was implemented using SBERT embeddings [20] and cosine similarity, ensuring that paraphrases retain semantic integrity, and any invalid passages are filtered out. Additionally, passage lengths are normalized to prevent instances where paraphrases are excessively short or long, improving paraphrase consistency. These enhancements eliminated inconsistencies in DE-COP's dataset [2], providing results that are more reproducible and statistically sound.

To further reduce API costs, the multiple-choice selection was expanded from three to four paraphrased options, reducing the probability of the original passage being randomly selected by 20%. By decreasing the probability of Type I error, the total number of passages requiring evaluation can be decreased without compromising the experiment's statistical power or significance. Consequently, this optimization substantially lowers overall API consumption by requiring less passages to be evaluated.

III. Results

Our proposed framework demonstrates significant improvements in detection accuracy, computational efficiency, and accessibility over existing methodologies. By providing our open-source solution as a hosted platform, we remove technical barriers, promoting ease of use and access to individual content creators.

The multi-layered workflow, which integrates passage extraction, paraphrase generation, question-answering, and multiple-choice testing, effectively differentiates between memorized

(copyrighted) and non-memorized text. By integrating a pre-screening pipeline using SBERT embeddings, cosine similarity, and normalized passage lengths, errors are caught and filtered out; enhancing reproducibility. The multiple-choice evaluation layer, with a streamlined randomization strategy and restructuring of question format reduced API consumption by 10-30%. Additionally, the Pinecone vector store enhances scalability and duplicate detection, avoiding redundant evaluations. These enhancements provide a scalable and practical solution that outperforms existing approaches, such as DE-COP, supporting ethical AI development and fair compensation for content creators.

IV. Conclusion

This paper introduces an open-source framework for detecting copyrighted content in LLM training datasets, addressing key limitations in accessibility, detection accuracy, and cost efficiency found in previous approaches such as DE-COP. By enhancing similarity detection, refining dataset validation, and optimizing computational efficiency, our system provides a scalable and accessible solution for copyright verification. Through our user-friendly interface, content creators can easily determine whether their work was appropriated for AI development, without a high technical barrier to entry. By promoting transparency and encouraging accountability, our system ultimately paves the way for ethical AI development.

V. Future Work

Future research may focus on developing methods for selective knowledge removal, such as *Unlearn* [21], to enable the erasure of copyrighted content from LLMs. This knowledge removal technique could be implemented for some of the standard LLM pretraining datasets such as *C4* [22] and *Pile* [23]. The legal implications of dataset memorization also warrant further investigation, particularly as AI copyright regulations continue to evolve. Additionally, expanding the scalability and adoption of our platform across different AI models and regulatory frameworks will be crucial for broader impact.

Acknowledgments

This research was enabled in part by funding, resources, and support provided by Wat.AI, the Waterloo AI Institute, and the Sedra Student Design Centre.

References

[1] J. Xu, S. Li, Z. Xu, and D. Zhang, "Do llms know to respect copyright notice?" *arXiv preprint arXiv:2411.01136*, 2024.

[2] A. V. Duarte, X. Zhao, A. L. Oliveira, and L. Li, "De-cop: Detecting copyrighted content in language models training data," *arXiv preprint arXiv:2402.09910*, 2024.

[3] J. Guo, Y. Li, R. Chen, Y. Wu, C. Liu, Y. Chen, and H. Huang, "Rag: Towards copyright protection for knowledge bases of retrieval-augmented language models," *OpenReview*, 2023.

[4] J. Freeman, C. Rippe, and E. Debenedetti, "Exploring memorization and copyright violation in frontier llms: A study of the new york times v. openai 2023 lawsuit," *arXiv preprint arXiv:2412.06370*, 2024.

[5] H. Shao, Z. Xu, S. Duan, and D. Zhang, "Measuring copyright risks of large language models via partial information probing," *arXiv preprint arXiv:2409.13831*, 2024.

[6] H. Tan, M. Duan, D. Liu, and L. Zhou, "Rethinking literary plagiarism in llms through the lens of copyright laws," in *Proceedings of the 16th Asian Conference on Machine Learning*, 2023.

[7] W. Shi, T. Ma, Y. Liu, and J. Zhang, "Detecting pretraining data from large language models," *arXiv preprint arXiv:2310.16789*, 2023.

[8] C. Kirchenbauer, J. Geiping, P. Carlini, and T. Jagielski, "Watermarking large language models," in *Proceedings of the International Conference on Machine Learning (ICML)*, 2023.

[9] H. Touvron, L. Martin, K. Stone, P. Albert, A. Almahairi, Y. Babaei, N. Bashlykov, S. Batra, P. Bhargava, S. Bhosale, D. Bikel, L. Blecher, C. C. Ferrer, M. Chen, G. Cucurull, D. Esiobu, J. Fernandes, J. Fu, W. Fu, B. Fuller, C. Gao, V. Goswami, N. Goyal, A. Hartshorn, S. Hosseini, R. Hou, H. Inan, M. Kardas, V. Kerkez, M. Khabsa, I. Kloumann, A. Korenev, P. S. Koura, M.-A. Lachaux, T. Lavril, J. Lee, D. Liskovich, Y. Lu, Y. Mao, X. Martinet, T. Mihaylov, P. Mishra, I. Molybog, Y. Nie, A. Poulton, J. Reizenstein, R. Rungta, K. Saladi, A. Schelten, R. Silva, E. M. Smith, R. Subramanian, X. E. Tan, B. Tang, R. Taylor, A. Williams, J. X. Kuan, P. Xu, Z. Yan, I. Zarov, Y. Zhang, A. Fan, M. Kambadur, S. Narang, A. Rodriguez, R. Stojnic, S. Edunov, and T. Scialom, "Llama 2: Open foundation and finetuned chat models," *arXiv preprint arXiv:2307.09288*, 2023.

[10] OpenAI, "Introducing chat-gpt," https://openai.com/blog/chatgpt, 2022, accessed: 2025-02-20.

[11] Z. Liu, A. Qiao, W. Neiswanger, H. Wang, B. Tan, T. Tao, J. Li, Y. Wang, S. Sun, O. Pangarkar, R. Fan, Y. Gu, V. Miller, Y. Zhuang, G. He, H. Li, F. Koto, L. Tang, N. Ranjan, Z. Shen, X. Ren, R. Iriondo, C. Mu, Z. Hu, M. Schulze, P. Nakov, T. Baldwin, and E. P. Xing, "Llm360: Towards fully transparent open-source llms," *arXiv preprint arXiv:2312.06550*, 2023.

[12] K. K. Chang, M. Cramer, S. Soni, and D. Bamman, "Speak, memory: An archaeology of books known to chatgpt/gpt4," *arXiv preprint arXiv:2305.00118*, 2023.

[13] S. Samsi, D. Zhao, J. McDonald, B. Li, A. Michaleas, M. Jones, W. Bergeron, J. Kepner, D. Tiwari, and V. Gadepally, "From words to watts: Benchmarking the energy costs of large language model inference," *arXiv preprint arXiv:2310.03003*, 2023, https://arxiv.org/pdf/2310.03003.

[14] A. de Wynter, X. Wang, A. Sokolov, Q. Gu, and S.-Q. Chen, "An evaluation on large language model outputs: Discourse and memorization," *arXiv preprint arXiv:2304.08637*, 2023.

[15] J. Lin and M. Ma, "A few brief notes on deepimpact, coil, and a conceptual framework for information retrieval techniques," *arXiv preprint arXiv:2106.14807*, 2021.

[16] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: Bm25 and beyond," *Foundations and Trends in Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

[17] Anthropic, "Claude 2," https://www.anthropic.com/news/claude-2, 2023.

[18] E. Bandel, R. Aharonov, M. Shmueli-Scheuer, I. Shnayderman, and N. Slonim, "Quality controlled paraphrase generation," in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2022, https://aclanthology.org/2022.acl-long.45/.

[19] N. Reimers and I. Gurevych, "Making monolingual sentence embeddings multilingual using knowledge distillation," in *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 2020, pp. 4512–4525.

[20] ——, "Sentence-bert: Sentence embeddings using siamese bert-networks," *arXiv preprint arXiv:1908.10084*, 2019.

[21] C. Eichler, T. Fischer, C. Sixt, and J. Yosinski, "Unlearn: Selective knowledge removal from large language models," *arXiv preprint arXiv:2408.04140*, 2024.

[22] C. Raffel, N. Shazeer, A. Roberts, K. Lee, S. Narang, M. Matena, and P. J. Liu, "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*, vol. 21, no. 140, pp. 1–67, 2020. [Online]. Available: https://jmlr.org/papers/v21/20-074.html

[23] G. Gao, S. Biderman, S. Black, L. Golding, H. He, and M. Shoeybi, "The pile: An 800gb dataset of diverse text for language modeling," https://arxiv.org/abs/2101.00027, 2020.