Evidencing LLM Misuse: A Hands-on Forensic Tutorial on Copyright Infringement and Plagiarism Detection

Denghui Zhang*
dzhang42@stevens.edu
Stevens Institute of Technology
Hoboken, NJ, USA

Guangwei Zhang
kwongwai@19pine.ai
Pine AI
Singapore, Singapore

Dongwon Lee
dongwon@psu.edu
The Pennsylvania State University
University Park, PA, USA

Abstract

Large Language Models (LLMs) introduce serious risks of content misuse, spanning copyright infringement in the legal domain and plagiarism in the ethical and academic domain. Although prior work has studied these risks, researchers and practitioners still need practical ways to audit, interpret, and evidence them. This tutorial presents a unified forensic perspective on LLM content misuse. First, we introduce **Copyright Detective**, an interactive forensic system for detecting, analyzing, and visualizing potential copyright leakage in LLM outputs. Participants will learn how inference-time scaling reveals sporadic memorization under output uncertainty, and how persuasive jailbreak probing can serve as defensive red teaming for examining alignment-suppressed leakage. Second, we introduce **LLM Plagiarism Detection**, covering verbatim copying, paraphrased reuse, and idea-level appropriation. We will discuss, and where appropriate demonstrate, how candidate source retrieval and text alignment support plagiarism analysis, while highlighting factors such as model size, decoding strategies, and fine-tuning corpus similarity. By combining hands-on copyright-risk auditing with a flexible plagiarism module, this tutorial equips attendees with practical and conceptual tools for auditing black-box models, interpreting similarity evidence, and understanding content misuse beyond simple text matching.

Keywords

Large Language Models; Copyright Infringement; Plagiarism Detection; Forensic Analysis; Red Teaming

ACM Reference Format:
Denghui Zhang, Guangwei Zhang, and Dongwon Lee. 2026. Evidencing LLM Misuse: A Hands-on Forensic Tutorial on Copyright Infringement and Plagiarism Detection. In *Proceedings of the 32nd ACM SIGKDD Conference on Knowledge Discovery and Data Mining V.2 (KDD 2026), August 9–13, 2026, Jeju Island, Republic of Korea.* ACM, New York, NY, USA, 2 pages. https://doi.org/10.1145/3770855.3816473

1 Introduction

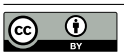
LLMs possess the capability for creative reasoning, yet simultaneously retain the risk of regurgitation of their training corpora [5], precipitating two distinct but critical crises regarding content misuse: *Copyright Infringement* [1, 6, 8], a legal violation involving the unauthorized reproduction of protected expression, and *Plagiarism* [2, 3], an ethical violation involving the uncredited appropriation of words, structures, or core ideas. Addressing these risks requires more than theoretical discussion; it demands robust, practical forensic and conceptual capabilities. Models have demonstrated the vulnerability to reproduce copyrighted text verbatim, raising significant legal concerns, while simultaneously acting as sophisticated engines for plagiarism that can paraphrase content to evade traditional detection. However, auditing these risks is non-trivial due to practical challenges such as *output uncertainty* where risks "flicker" across generations and *alignment suppression* where safety fine-tuning masks latent memorization. This tutorial bridges the gap between theoretical risk assessment and practical auditing. We organize the tutorial around two complementary perspectives on content misuse in LLMs:

- **Copyright Detective** [9]: An integrated forensic system for detecting, analyzing, and visualizing potential copyright risks in LLM outputs, including verbatim memorization and paraphrase-level leakage. We will demonstrate how inference-time scaling helps reveal sporadic leakage under output uncertainty, and how persuasive jailbreak probing [4] can be used as a defensive red-teaming method to examine alignment-suppressed memorization.
- **Plagiarism Detection** [3]: A module introducing the detection and analysis of plagiarism in LLM-generated text, covering verbatim copying, paraphrased reuse, and idea-level appropriation. We will discuss, and where appropriate demonstrate, how automatic plagiarism detection can combine candidate source retrieval with text alignment to identify suspicious overlaps, while highlighting key factors such as model size, decoding strategies, and fine-tuning corpus similarity that shape plagiarism behavior.

2 Tutorial Outline

(1) **Forensic Foundations of LLM Content Misuse (45 min).** We introduce copyright infringement and plagiarism as two distinct forms of LLM content misuse, covering verbatim memorization, paraphrase-level leakage, and idea-level reuse. We also discuss key auditing challenges, including output uncertainty, sporadic leakage, alignment suppression.
(2) **Break (15 min).**
(3) **Hands-on Copyright Auditing with Copyright Detective (45 min).** We present Copyright Detective as a unified system for black-box copyright auditing. Participants will test text- and document-level memorization, apply inference-time scaling, craft persuasive prompts, and interpret visual evidence through similarity metrics.

*Corresponding Tutor: dzhang42@stevens.edu



KDD 2026, Jeju Island, Republic of Korea.

ACM ISBN 979-8-4007-2259-2/2026/08
https://doi.org/10.1145/3770855.3816473

(4) **Break (15 min).**
(5) **Hands-on Plagiarism Detection and Forensic Reporting (45 min).** We introduce an automated pipeline for detecting plagiarism in LLM outputs. Participants will retrieve candidate sources, run text alignment, analyze verbatim, paraphrase, and idea-level plagiarism, and organize findings into forensic audit reports.
(6) **Wrap-up and Open Challenges (15 min).** We summarize the auditing workflows and discuss open challenges, including the boundary between technical evidence and legal or ethical judgment, as well as future directions for forensic analysis of Generative AI systems.

3 Audience, Prerequisites, and Participation

This tutorial targets 50-100 participants from machine learning, legal technology, education, and responsible-AI communities. It is intended for researchers and practitioners interested in AI safety, model evaluation and copyright compliance. The tutorial requires only a basic conceptual understanding of LLMs and current copyright and plagiarism risks, with no advanced programming background expected.

Audience participation is central to the tutorial. Participants will use a Streamlit-based interface and accompanying tutorial materials for *Copyright Detective* to conduct real-time copyright-risk audits. The hands-on session will cover content recall testing, document-level memorization analysis, inference-time scaling, and persuasive jailbreak probing as defensive red teaming. Participants will compare model outputs with user-provided reference passages, interpret similarity metrics and visual evidence panels, and discuss how technical findings can be organized into reproducible forensic evidence. For the plagiarism component, the tutorial will introduce how LLM-generated text may exhibit plagiarism beyond verbatim copying, including paraphrased reuse and idea-level appropriation. Participants will examine representative examples of LLM plagiarism, discuss how candidate source retrieval and text alignment can support detection, and reflect on the broader implications for academic integrity and responsible AI deployment.

4 Resources and Materials

- **Participants:** Laptop with a web browser.
- **Infrastructure:** We will provide access to the tutorial web application[1] and supporting code repository[2].

5 Relation to Prior Tutorials

This tutorial has not been presented elsewhere in this integrated hands-on form. A related NAACL 2025 tutorial, *LLMs and Copyright Risks: Benchmarks and Mitigation Approaches* [7], covered copyright risk assessment and mitigation from a broader lecture-style perspective. In contrast, this tutorial is system-focused and hands-on: it emphasizes practical forensic auditing, integrates *Copyright Detective* [9] with a conceptual overview of LLM plagiarism risks and detection methods, expanding the scope from copyright memorization to broader plagiarism evidence.

[1] https://copyright-detective.streamlit.app/
[2] https://github.com/Brit7777/LM-plagiarism

6 Societal Impact

This tutorial contributes to Responsible AI by serving three practical needs: helping authors and legal professionals document potential copyright leakage, supporting AI developers in red-teaming deployed models, and providing educators and the public with accessible ways to understand copyright and plagiarism risks in Generative AI. Through reproducible, interactive, and black-box auditing workflows, it promotes transparency, academic integrity, and legally aware LLM deployment.

7 Tutors and In-person Presenters

Dr. Denghui Zhang is an Assistant Professor at Stevens Institute of Technology. He studies the interplay between LLMs/GenAI, legal and socio-ethical issues. He is the corresponding author of seminal works on LLM copyright evaluation.
Guangwei Zhang is an Engineer at Pine AI. His expertise spans speech algorithms and autonomous agents. His research primarily focuses on the safety and trustworthiness of LLMs, with a specific emphasis on copyright protection.
Dr. Dongwon Lee is a Professor at The Pennsylvania State University. His research focuses on the trustworthiness of AI, specifically plagiarism detection and copyright issues. He is the corresponding author of seminal works on LLM plagiarism.

References

[1] Kent K. Chang, Mackenzie Cramer, Sandeep Soni, and David Bamman. 2023. Speak, Memory: An Archaeology of Books Known to ChatGPT/GPT-4. arXiv:2305.00118 [cs.CL] https://arxiv.org/abs/2305.00118
[2] Jooyoung Lee, Toshini Agrawal, Adaku Uchendu, Thai Le, Jinghui Chen, and Dongwon Lee. 2025. PlagBench: Exploring the Duality of Large Language Models in Plagiarism Generation and Detection. In *Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics (NAACL)*. Albuquerque, New Mexico, 7519–7534.
[3] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. 2023. Do Language Models Plagiarize?. In *Proceedings of the ACM Web Conference 2023* (Austin, TX, USA) *(WWW '23)*. Association for Computing Machinery, New York, NY, USA, 3637–3647. doi:10.1145/3543507.3583199
[4] Jikai Long, Ming Liu, Xiusi Chen, Jialiang Xu, Shenglan Li, Zhaozhuo Xu, and Denghui Zhang. 2025. Profiling LLM's Copyright Infringement Risks under Adversarial Persuasive Prompting. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 15799–15823. doi:10.18653/v1/2025.findings-emnlp.855
[5] Weijia Shi, Anirudh Ajith, Mengzhou Xia, Yangsibo Huang, Daogao Liu, Terra Blevins, Danqi Chen, and Luke Zettlemoyer. 2024. Detecting Pretraining Data from Large Language Models. In *International Conference on Learning Representations*.
[6] Jialiang Xu, Shenglan Li, Zhaozhuo Xu, and Denghui Zhang. 2024. Do LLMs Know to Respect Copyright Notice?. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen (Eds.). Association for Computational Linguistics, Miami, Florida, USA, 20604–20619. doi:10.18653/v1/2024.emnlp-main.1147
[7] Denghui Zhang, Zhaozhuo Xu, and Weijie Zhao. 2025. LLMs and Copyright Risks: Benchmarks and Mitigation Approaches. In *Proceedings of the 2025 Annual Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 5: Tutorial Abstracts)*, Maria Lomeli, Swabha Swayamdipta, and Rui Zhang (Eds.). Association for Computational Linguistics, Albuquerque, New Mexico, 44–50. doi:10.18653/v1/2025.naacl-tutorial.7
[8] Guangwei Zhang, Qisheng Su, Jiateng Liu, Cheng Qian, Yanzhou Pan, Yanjie Fu, and Denghui Zhang. 2025. ISACL: Internal State Analyzer for Copyrighted Training Data Leakage. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng (Eds.). Association for Computational Linguistics, Suzhou, China, 10786–10807. doi:10.18653/v1/2025.findings-emnlp.571
[9] Guangwei Zhang, Jianing Zhu, Cheng Qian, Neil Gong, Rada Mihalcea, Zhaozhuo Xu, Jingrui He, Jiaqi Ma, Yun Huang, Chaowei Xiao, Bo Li, Ahmed Abbasi, Dongwon Lee, Heng Ji, and Denghui Zhang. 2026. Copyright Detective: A Forensic System to Evidence LLMs Flickering Copyright Leakage Risks. arXiv:2602.05252 [cs.CL] https://arxiv.org/abs/2602.05252