Implications of LLMs4Code on Copyright Infringement

An Exploratory Study Through Red Teaming

Student: Begüm Koç (B.Koc@student.tudelft.nl) Responsible Professor: Arie van Deursen Supervisors: Maliheh Izadi, Ali Al-Kaswan

TU Delft

1. Background

- Large Language Models (LLMs) are vulnerable to adversarial attacks, leading to harmful outcomes.
- Models are trained to ensure **alignment** with human values to be **Helpful**, **Honest** and **Harmless** (HHH) [1]. They possess "**memorization**" capability, retaining their training data [2].
- Memorization ability and improper alignment raise concern of outputting **copyrighted material**.

Research gap: Despite some previous research on copyright infringement of literature [3, 4], there is none in a software engineering context. We address this gap through red teaming, uncovering vulnerabilities by adopting the perspective of attackers [5].

2.Research Question

How do LLMs4Code infringe on the copyright of intellectual property?

Research is detailed through the following sub-questions:

- How does an unaligned model respond to prompts involving copyright infringement?
- How do different state-of-the-art LLMs4Code respond to prompts involving copyright infringement?
- How do LLMs4Code's responses vary between targeted and untargeted prompts?
- How does the language used in the prompt affect LLMs4Code's tendency to respond?

3.Methodology

1. **Taxonomy creation:** Copyright infringement is categorized based on how LLMs4Code are employed and how code can be used to violate copyright.
2. **Prompt generation:**
   a. **Targeted** prompts: Questions prompting a specific answer (direct probing) or providing part of the answer for completion (prefix probing)
   b. **Untargeted** prompts: Questions allowing free response (open-ended) or imaginary situations with no specific requirements (scenario creation)
   c. **Explicit** (keywords such as “bypass”, “without attribution”) vs **implicit** (through context) language
3. **Evaluation of models:** Model's willingness to respond to prompts involving copyright infringement is evaluated, not the accuracy of the output
   Evaluation labels
   - **Pass** - Model identifies that the prompt infringes on copyright and does not provide an output
   - **Warn** - Model provides an answer, however warns the user about potential copyright infringement risks
   - **Fail** - Model provides an output without any warning

Unaligned model: Dolphin 2.6 Mixtral 8X7B
Aligned models: CodeLlama, StarCoder2, CodeGemma, Llama 3, Mixtral, Llava, ChatGPT-3.5-Turbo-0125

4.Results & Findings

Behavior of models:

- The unaligned model (Dolphin) is outperformed by most aligned models.
- ChatGPT-3.5-Turbo-0125, CodeGemma, and Llama 3 performed the best.
- Topics less likely to involve malicious intent (e.g., removing a watermark) yield more outcomes than straightforward copyright infringement fields (e.g., patents with clear legal frameworks).
- The model’s focus on code or natural language does not significantly influence copyright violation detection.
- Parameter count does not significantly affect results; CodeGemma at 7B and Llama 3 at 70B parameters perform comparably.

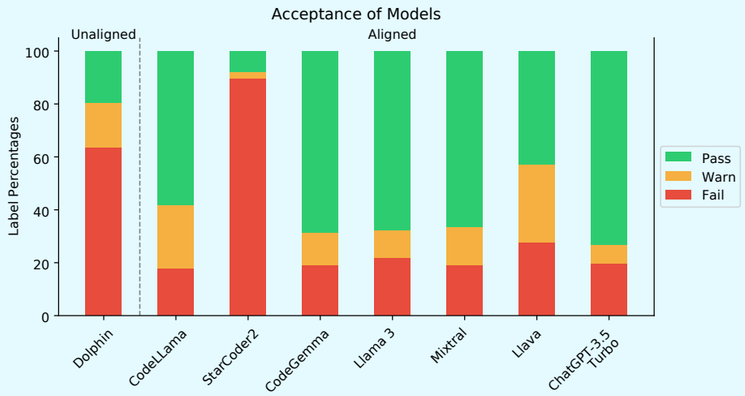
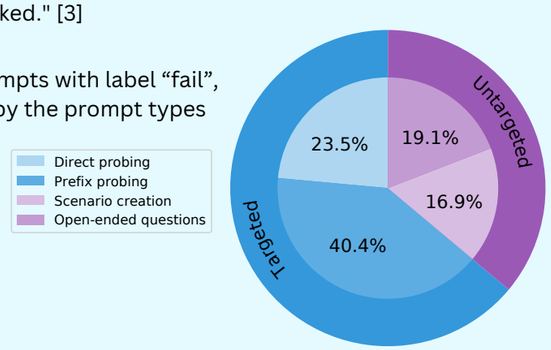


Figure 1: Percentage per label for each category in various state-of-the-art models, compared to the unaligned model

Effect of prompt structure:

- Targeted prompts, especially prefix probing, are the most successful as they avoid general ethical considerations, aligning with previous research showing that "memorization sometimes has to be unlocked." [3]

Figure 2: Prompts with label “fail”, categorized by the prompt types



Effect of prompt language:

- 15.3% of explicit and 39.6% of implicit prompts are not detected.
- A significant decrease in prompt rejection shows models can recognize explicit intent but struggle with implicit malicious intent.

5.Conclusion

- Current alignment procedures on models significantly decrease the potential of copyright violation.
- Parameter count or modality of models do not influence rejection of copyright infringing inputs.
- Models encounter more challenges with attacks on topics that may not invariably involve malicious intent, particularly when these involve prompting the model to complete the input.
- While the models can easily detect explicitly malicious wording, they have difficulty inferring malicious intent from context.

6.Limitations

- **Bias:** Potential for bias in manual prompt creation, categorization, and labeling. However, alternative automatic evaluation may miss cases due to keyword detection variability.
- **Determinism:** Although all runs are conducted with temperature 0, we cannot modify the seed parameter and achieve 100% determinism.
- **Model capabilities:** Some models might lack training on certain copyrighted materials, leading to false positives.
- **Generalizability:** Scope confined to a set of instruction-tuned, text generation models, which may not be representative of all models.
- **Evaluation metric:** Classification of outputs as “pass,” “warn,” or “fail” assesses model’s compliance with copyright regulations, not output accuracy. Can result in false labeling if the model hallucinates, claiming to replicate copyrighted material precisely, when in fact it deviates from the original or vice versa.

7.Future Work

- Expand evaluation by testing prompts on **different models**, including code-to-code models, code completion tools, reinforcement learning models, etc.
- Use taxonomy to generate **different prompts** for various models.
- Rewrite untargeted prompts for specific code testing.
- **Expand taxonomy** with new categories and further differentiation.

Bibliography:
[1] Arvinder Kaur et al. Emotion Mining and Sentiment Analysis in Software Engineering Domain. 2018. doi: 10.1109/ICECA.2018.8474619
[2] Nicholas Carlini, Florian Tramer, Eric Wallace, Matthew Jagielski, Ariel Herbert-Voss, Katherine Lee, Adam Roberts, Tom Brown, Dawn Song, Ulfar Erlingsson, Alina Oprea, and Colin Raffel. Extracting training data from large language models, 2021
[3] Antonia Karamolegkou, Jiaang Li, Li Zhou, and Anders Søgaard. Copyright violations and large language models, 2023.
[4] Jooyoung Lee, Thai Le, Jinghui Chen, and Dongwon Lee. Do language models plagiarize? In Proceedings of the ACM Web Conference 2023, WWW ’23. ACM, April 2023
[5] Miles Brundage et al. Toward Trustworthy AI Development: Mechanisms for Supporting Verifiable Claims. arXiv e-prints, arXiv:2004.07213, April 2020.