# Large Language Models for Code Analysis: Do LLMs Really Do Their Job?

Chongzhou Fang, Ning Miao, Shaurya Srivastav, Jialin Liu, Ruoyu Zhang, Ruijie Fang,
Asmita Asmita, Ryan Tsang, Najmeh Nazari, Han Wang and Houman Homayoun

Aug. 14, 2024


# Outline

- Introduction
- Experiment Settings
- Results & Findings
- Conclusion

# Introduction

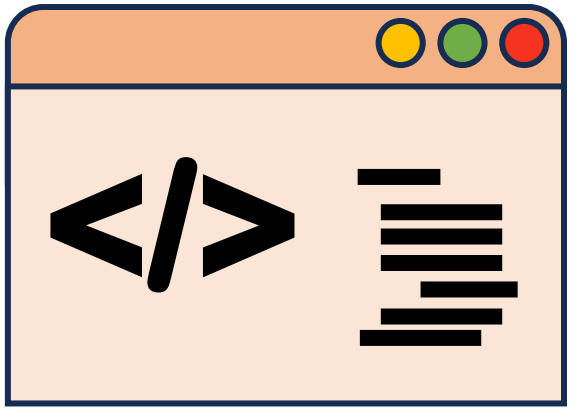
# Background

- Code Analysis
- Code Obfuscation

```
1    function hi() {
2        console.log("Hello World!");
3    }
4    hi();
```

```
1    function _0x1ec3(){var _0x3ed452=['259790KgLPlj','297688NTFutg','35ACWDkX','
         145716kEyGyf','18SFCPKB','1701952aKOEga','192jjwxUU','5LPjNwr','142417
         rtWDUq','Hello\x20World!','121610lhBPGW','2032200UghFpX','5nCOmEq','log'
         ];_0x1ec3=function(){return _0x3ed452;};return _0x1ec3();}(function(
         _0x22b342,_0x360ffb){var _0x5047be=_0xfb3c,_0x4c7c5c=_0x22b342();while
         (!![]){try{var _0x40c3be=parseInt(_0x5047be(0x90))/0x1*(-parseInt(
         _0x5047be(0x8e))/0x2)+-parseInt(_0x5047be(0x95))/0x3+parseInt(_0x5047be(0
         x97))/0x4*(parseInt(_0x5047be(0x99))/0x5)+parseInt(_0x5047be(0x8f))/0x6+
         parseInt(_0x5047be(0x94))/0x7*(parseInt(_0x5047be(0x93))/0x8)+parseInt(
         _0x5047be(0x96))/0x9*(-parseInt(_0x5047be(0x92))/0xa)+parseInt(_0x5047be
         (0x9a))/0xb*(-parseInt(_0x5047be(0x98))/0xc);if(_0x40c3be===_0x360ffb)
         break;else _0x4c7c5c['push'](_0x4c7c5c['shift']());}catch(_0x33f4b4){
         _0x4c7c5c['push'](_0x4c7c5c['shift']());}}}(_0x1ec3,0x52a68));function
         _0xfb3c(_0x257a0b,_0x17c420){var _0x1ec321=_0x1ec3();return _0xfb3c=
         function(_0xfb3ca7,_0x44b6b2){_0xfb3ca7=_0xfb3ca7-0x8d;var _0x34ca8b=
         _0x1ec321[_0xfb3ca7];return _0x34ca8b;},_0xfb3c(_0x257a0b,_0x17c420);}
         function hi(){var _0x2da467=_0xfb3c;console[_0x2da467(0x91)](_0x2da467(0
         x8d));}hi();
```

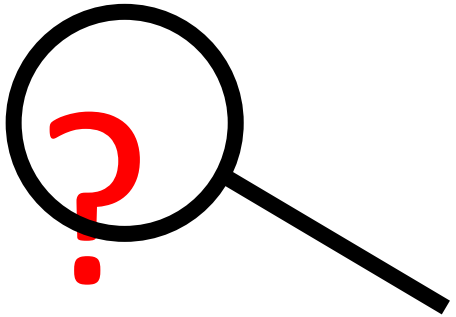
# Background

# Experiment Settings

# Research Questions

- **RQ1:** Do LLMs understand code?
- **RQ2**: Can LLMs understand obfuscated code?



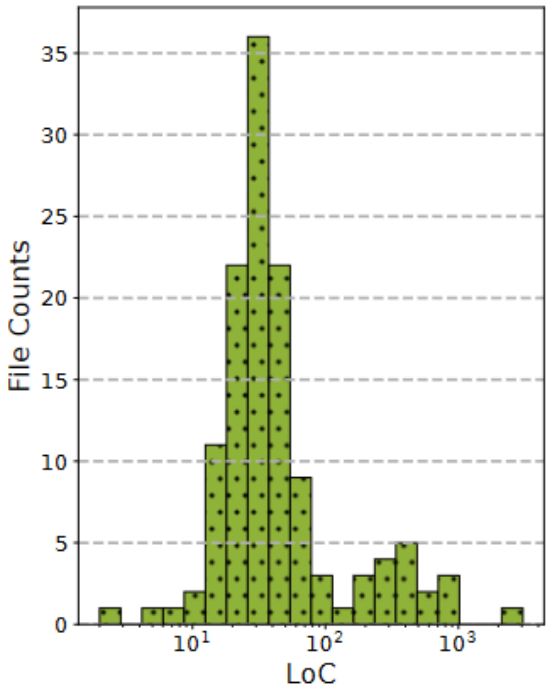
# LLM Selection

- GPT-3.5-turbo
- GPT-4
- LLaMA-2-13B
- Code-LLaMA-2-13B-Instruct
- StarChat-Beta

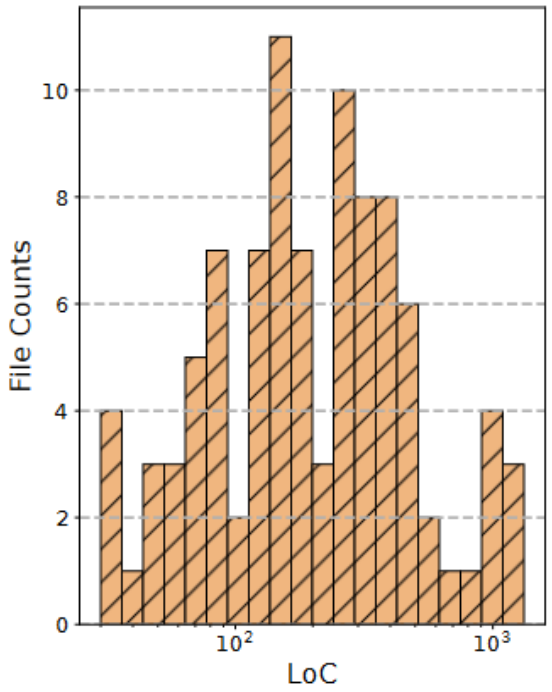
# Datasets

- Non-Obfuscated Code Dataset
  - C
    - Popular benchmarks
    - POJ-104
  - JavaScript
    - Octane 2.0
    - A list of practical JavaScript applications
  - Python
    - CodeSearchNet

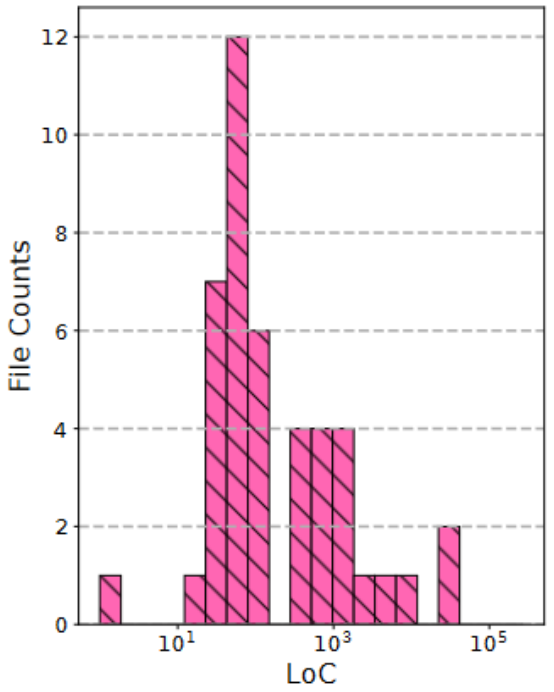
# Datasets



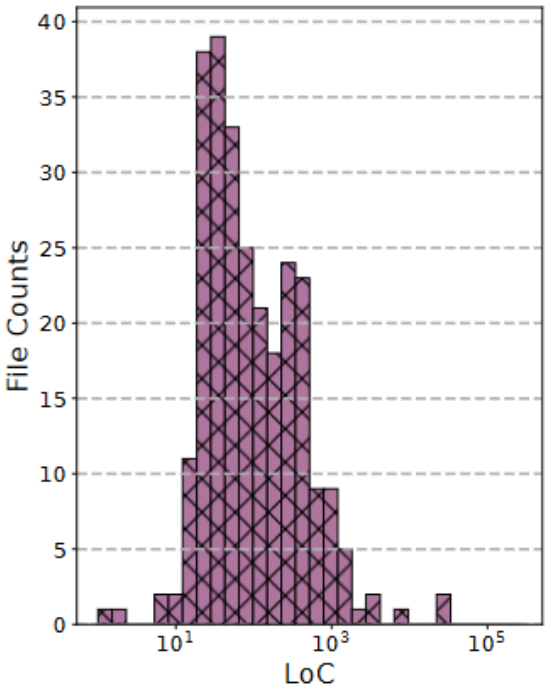

(a) C.




(b) Python.




(c) JavaScript.




(d) Overall.

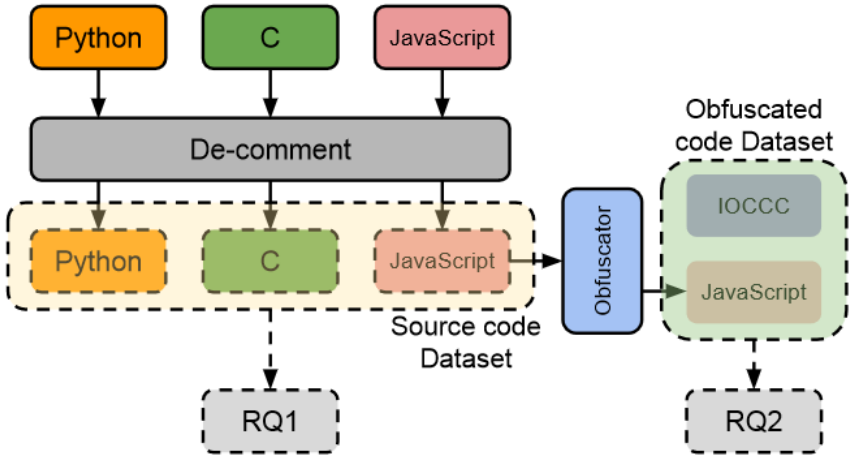
# Datasets

- Non-Obfuscated Code Dataset
  - C
    - Popular benchmarks
    - POJ-104
  - JavaScript
    - Octane 2.0
    - A list of practical JavaScript applications
  - Python
    - CodeSearchNet

- Obfuscated Code Dataset
  - Based on JavaScript Non-Obf Code
    - Default (DE)
    - Dead Code Injection (DCI)
    - Control Flow Flattening (CFF)
    - Split String (SS)
    - Wobfuscator (WSM) [1]
  - International Obfuscated C Code Contest (IOCCC)

[1] Alan Romano, Daniel Lehmann, Michael Pradel, and Weihang Wang. Wobfuscator: Obfuscating javascript malware via opportunistic translation to webassembly. In 2022 IEEE Symposium on Security and Privacy (SP), pages 1574–1589. IEEE, 2022

# Measurement Methods



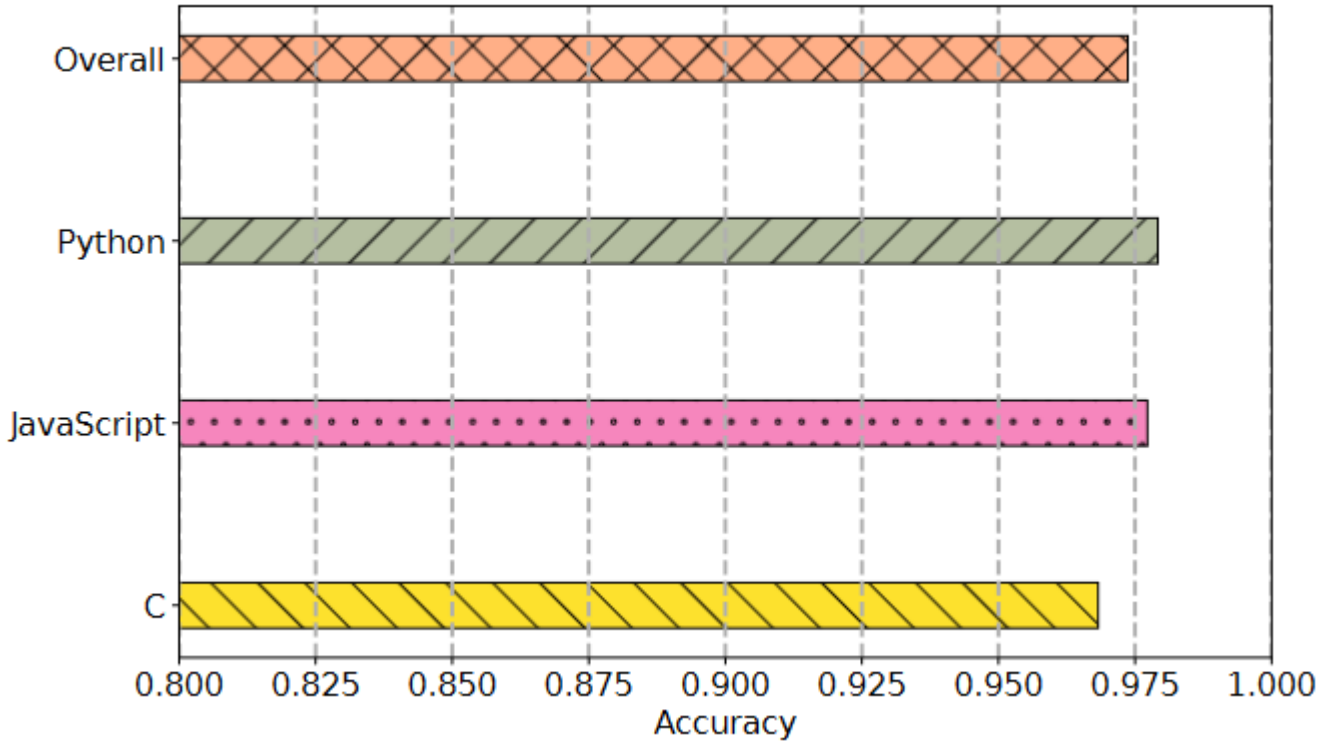

- Ground Truth

- Comparison Metrics
  - Cosine Similarity
  - Bert-Based Semantic Similarity [2]
  - ChatGPT-Based Similarity

[2] Semantic-text-similarity. https://github.com/AndriyMulyar/semantic-text-similarity

# Results & Findings

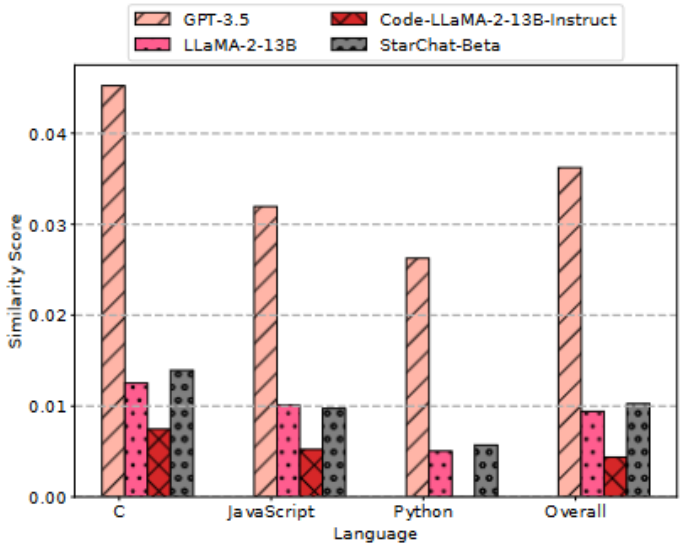
# Results: Non-Obfuscated Code Dataset



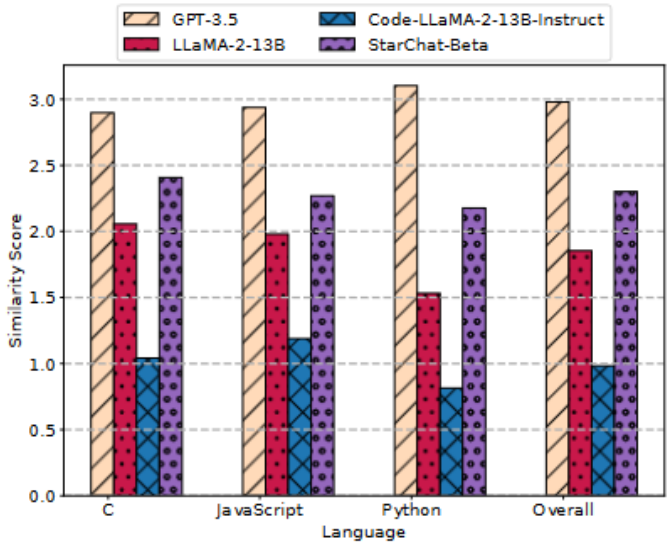

**Finding:** GPT utilizes information provided in identifier names to assist code analysis.

**Finding:** GPT-4 occasionally makes wrong associations.

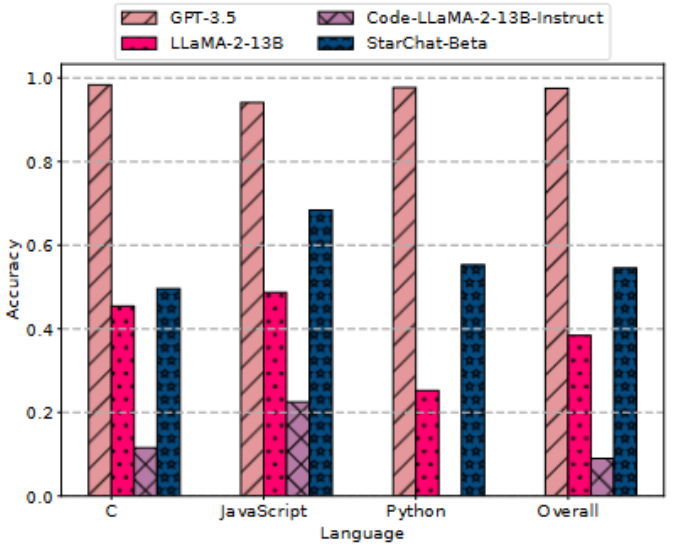
# Results: Non-Obfuscated Code Dataset




(a) Cosine similarity score.



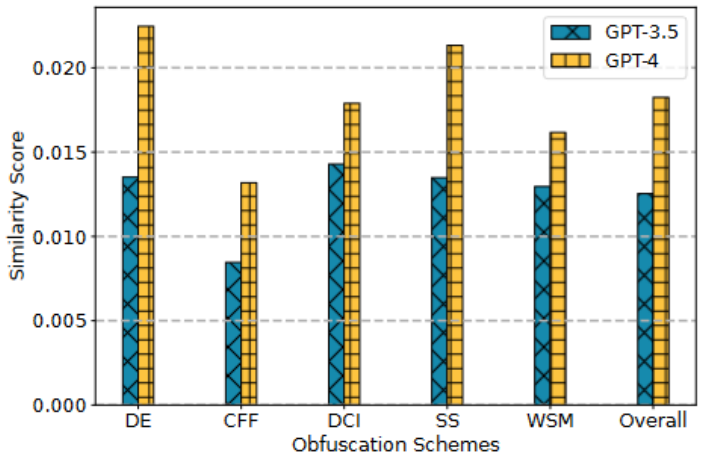

(b) Bert-based semantic similarity score.



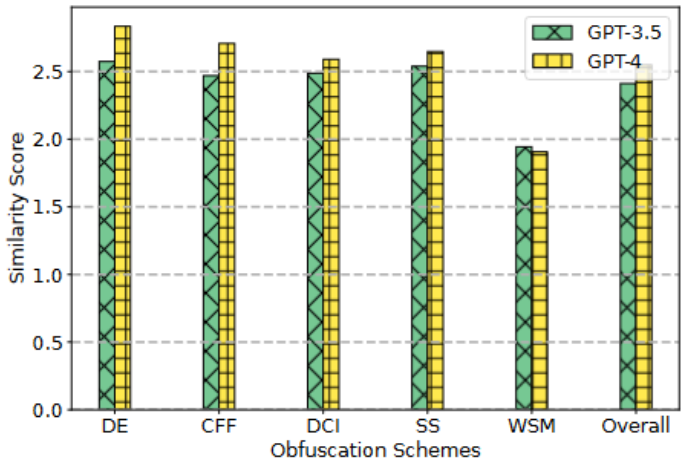

(c) ChatGPT measured accuracy results.

**Finding:** Smaller models are unable to reliably generate consistent paragraphs of code analysis results.

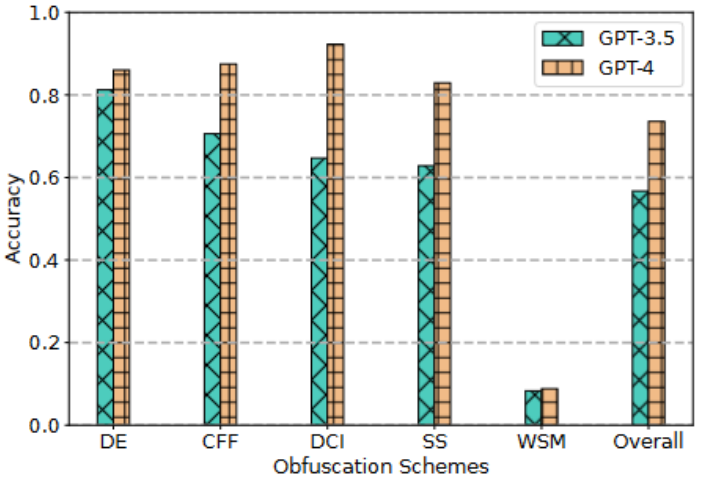
# Results: Obfuscated Code Dataset




(a) Cosine similarity score.




(b) Bert-based semantic similarity score.

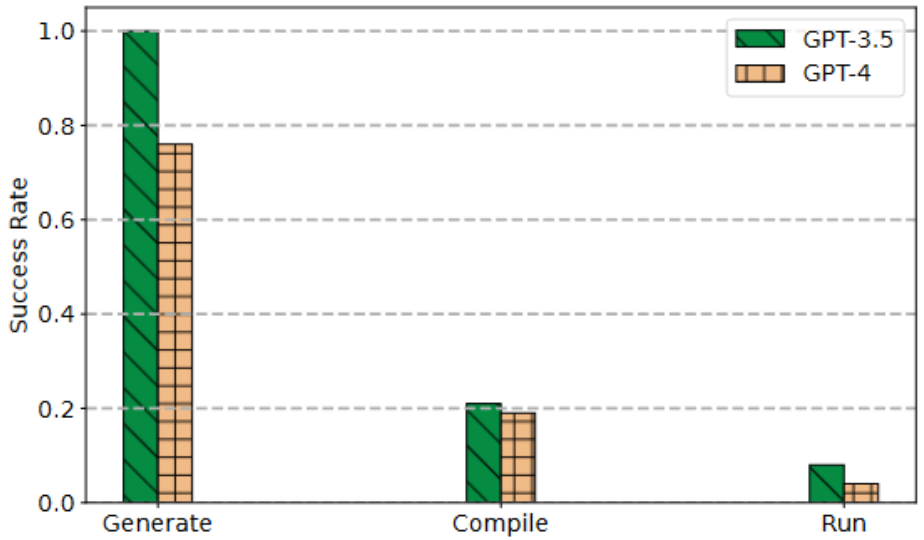



(c) ChatGPT measured accuracy results.

**Finding:** Basic obfuscation techniques only slightly influence the ability of GPT models to perform code analysis.

**Finding:** LLMs are not able to decipher obfuscated code generated by Wobfuscator.

# Results: Obfuscated Code Dataset




**Finding:** All models fall short of generating compilable and runnable de-obfuscated code.

**Finding:** GPT-4 generates code with higher readability.

# Conclusion

# Conclusion

- A thorough evaluation of code analysis capabilities of popular LLMs
- Limitations of LLM
- Online Appendix: https://github.com/aseec-lab/llms-for-code-analysis

# Q&A