ALFREDO: AGENTIC LLM-BASED FRAMEWORK FOR CODE DEOBFUSCATION

Members:
Ching Yuhui Natalie (River Valley High School)
Sophie Tung Xuan Ying (Nanyang Girls' High School)

Mentors:
Siow Jing Kai, Phyllis Poh Hui-Li (DSO National Laboratories)

BACKGROUND

Malware employs a variety of obfuscation techniques to **evade identification or analysis**, making code deobfuscation a significant challenge in cybersecurity.

Limitations of current (methodical) deobfuscation tools:

1) Manual and tedious

2) Limited to one transformation

Recent studies conclude that **Large Language Models (LLMs)** are a promising approach to deobfuscation tasks, but are not yet ready to be used independently [1].

Novel agentic approaches can enhance LLM's reliability for technical tasks by helping them:

1) Leverage compilation tools

2) Dissect and organise tasks

3) Exercise control over decision-making

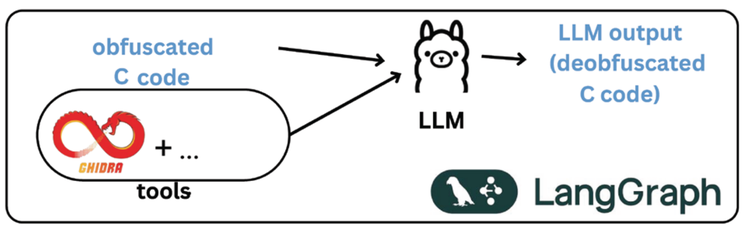
Research Objective

We hypothesise that with the help of agentic frameworks, LLMs can demonstrate potential in tackling code deobfuscation problems, posing a viable alternative to methodical deobfuscators.

METHODOLOGY

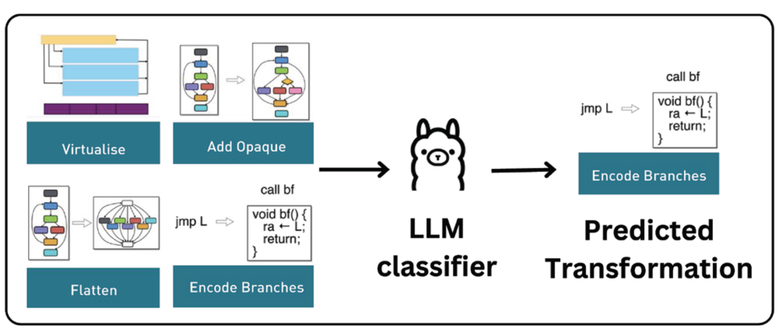
Agentic AI system

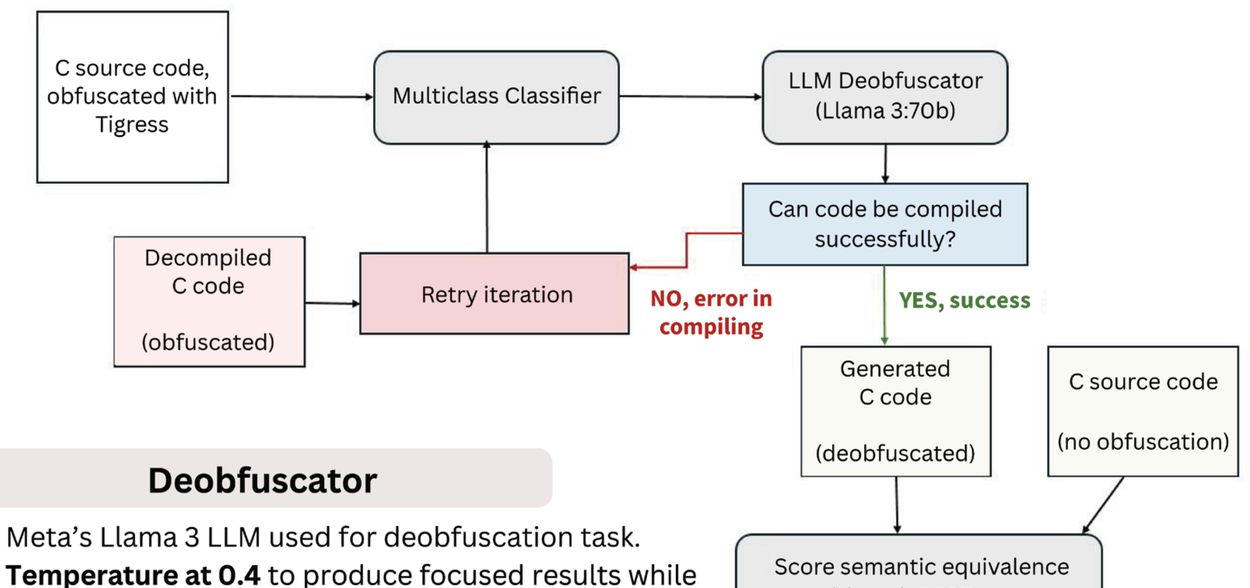
Agentic AI system using **LangGraph**, allowing the model to utilise Ghidra Headless. The agentic system is extensible, for easy addition of other analysis tools.



Multiclass Classifier

Obfuscating transformation is identified by the multiclass classifier from 4 known transformations.





Deobfuscator

- Meta's Llama 3 LLM used for deobfuscation task.
- **Temperature at 0.4** to produce focused results while encouraging model creativity, to explore alternative solutions rather than repetitive and rigid ones.

Iterative Process

- An iterative process **combats compilation errors**, improving model performance.
- When compilation fails, deobfuscation process is repeated with **additional context** - pseudo C generated from Ghidra Headless.
- On successful compilation, **codeBERT score** is evaluated to test deobfuscation accuracy.

RESULTS & DISCUSSION

		MODeflattener	deflat	ALFREDO
Average F-measure	Flattening	0.93	0.95	0.80
	Add Opaque	-	-	0.78
	Encode Branches	-	-	0.84
	Virtualisation	-	-	0.66
	Overall	-	-	0.77
Typical runtime (sec)		0.09	3.43	9.42~52.8

RQ1: How accurate is ALFREDO in recovering the original C code?

- **High F-measure** of 0.77, but lower than methodical deobfuscators.
- Longer runtime, likely due to the nature of LangGraph and the iterative process.

RQ2: How does the specific transformation applied affect ALFREDO?

- Virtualised code had the lowest F-measure, aligning with existing knowledge that Virtualisation is challenging to deobfuscate even by manual approaches.

RQ3: How does the multiclass classifier affect deobfuscation ability?

- Average F-measure remained the same without the classifier, showing no apparent decline in performance.
- We hypothesise that other techniques such as Support Vector Machines and Random Forest are preferable to LLMs for identifying transformations, potentially benefitting deobfuscation accuracy.

Societal Impact

ALFREDO is **more flexible** in deobfuscating code of various transformations, and may perform better than methodical deobfuscators in real-world contexts, where more than one transformation is usually used to interfere with traditional approaches.

Additionally, agentic systems **allow support with analysis tools** like taint analysis, maximising technical competence of LLM-based deobfuscators while retaining a flexible, creative decision-making approach.

Our work demonstrates the potential of agentic LLM applications in an entirely new field, being to the best of our knowledge, the **first existing work** to investigate the usage of agentic LLMs in code deobfuscation.

Future Work

1. Integrate **more sophisticated analysis tools** into agentic system, such as control flow graphs or symbolic execution
2. Explore **other machine learning techniques** for the classifier
3. **Fine tune** the deobfuscator for specific transformations

References

[1] Constantinos Patsakis, Fran Casino, Nikolaos Lykousas, Assessing LLMs in malicious code deobfuscation of real-world malware campaigns, https://doi.org/10.1016/j.eswa.2024.124912.