

LLMを用いたDecompile

発表者: 福井 遥陽

背景

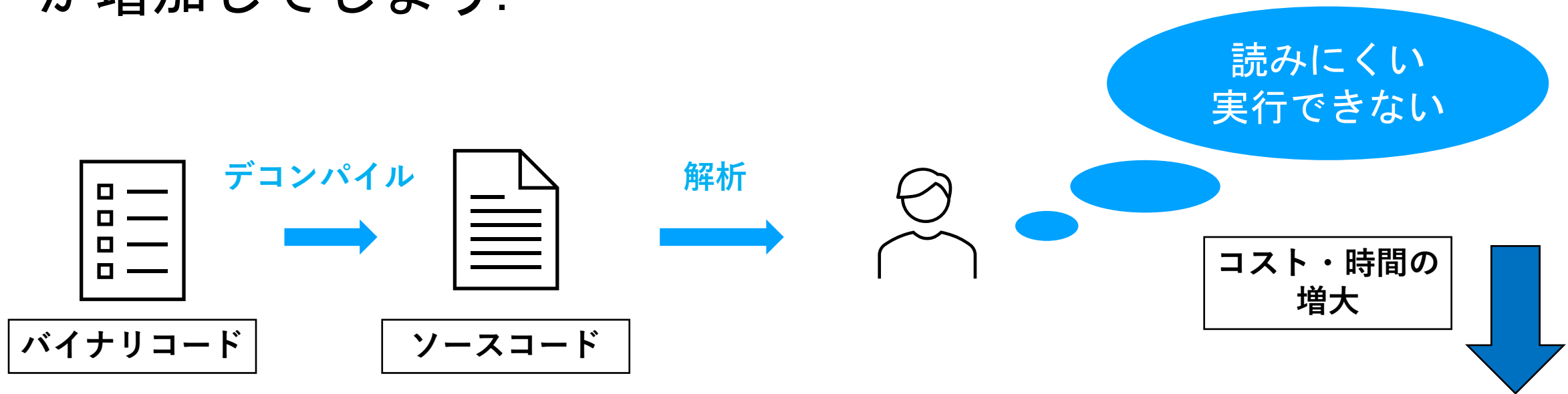
デコンパイルとは

- | バイナリ形式のコードから高級言語(C,Python,java)を得る手段.
 - | リバースエンジニアリングの分野においてデコンパイルを用いたマルウェア解析が行われている.
- デコンパイルには専用のツール(Ghidra,IDA Pro)を用いる.

課題

現状，開発されているツールや手法ではバイナリコードから，完全なソースコードを復元するのは困難.

特に，可読性や再実行可能性の観点から解析にかかるコストが増加してしまう.



先行研究

LLM4Decompile

バイナリコードを出力として直接LLMにデコンパイルさせる手法

バイナリコードをツールを使ってデコンパイルし，そこから得たコードをLLMで改良する手法

課題点

この手法では平均50%ほどの再実行可能性を示したが，可読性の観点では実際にスコアとしては出されていない。

学習データの不足により，大規模な学習が出来ていない。

研究目的

デコンパイルツールの出力をLLMを用いて改善する手法は数多く存在する一方で、直接LLMを用いてデコンパイルする手法はほとんどない。

→そのため、デコンパイルに使うLLMを複数用いて、どのLLMがデコンパイルに適しているかの研究がない。

LLMごとのコンパイル比較や大規模なデータセットを使ったファインチューニング、LLMにデコンパイルさせてその出力コードにLLMを使って整形するなど

研究目的

最終的には、これらのアプローチを行って再実行可能性や可読性の観点で、既存のツールの出力にLLMで成形する既存手法より優位性を示したい。

また、そこから分類タスクにも応用できるLLMを使用して脆弱性予測も行えるようにしたい。

今後の方針

- デコンパイルの主目的はマルウェアなどの脆弱性検出にあることを念頭に置いて、解決する課題点を選定.
 - デコンパイルにLLMと既存のツールのどちらを使用するか、また両方を使用するかを検討.
- 今の段階では、精度の向上などの目的よりも手法を試してみたい気持ちの方が強いいため、方針を定めてアプローチを絞る必要がある.