

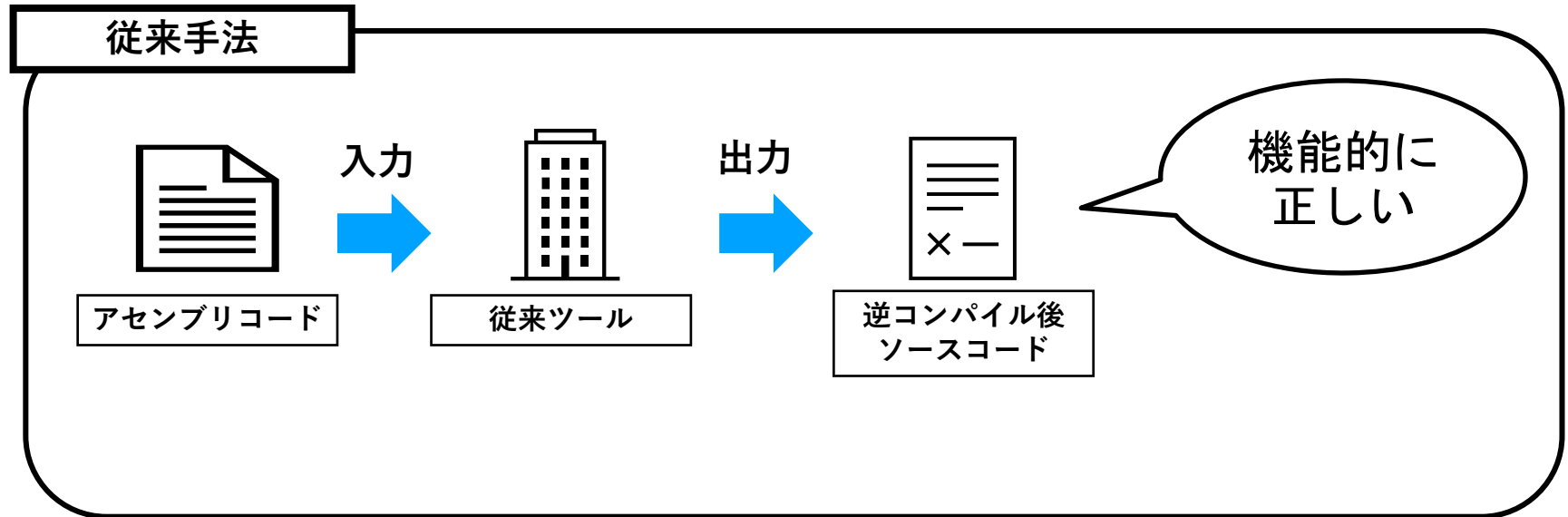
LLMs as Idiomatic Decompilers: Recovering High-Level Code from x86-64 Assembly for Dart

Raafat Abualazm
*Faculty of Engineering
Email:
raafat.202210476@eng-
st.cu.edu.eg

Ayman Abo Elhassan
Cairo University, Giza, Egypt
Email:
ayman.abo.elmaaty@eng.cu.edu.eg

SANAR 2026
学生番号 : 60276216
発表者: 福井 遥陽

背景:従来のデコンパイラ



一方で

- ×人間が読みづらい
- ×プログラムの意味的情報が失われる

背景:近年の研究

Transformerベースのdecompileモデル
例) SLade, Nova, LLM4Decompile

コードの読みやすさ↑
コードが実行可能か↑

→C言語コードのみ対象

言語固有のイディオムを持つ現代言語 (Dart, Swift)
への適用に限界がある

目的

「x86-64アセンブリから, Dart/Swiftらしい自然なコードを生成できるか」

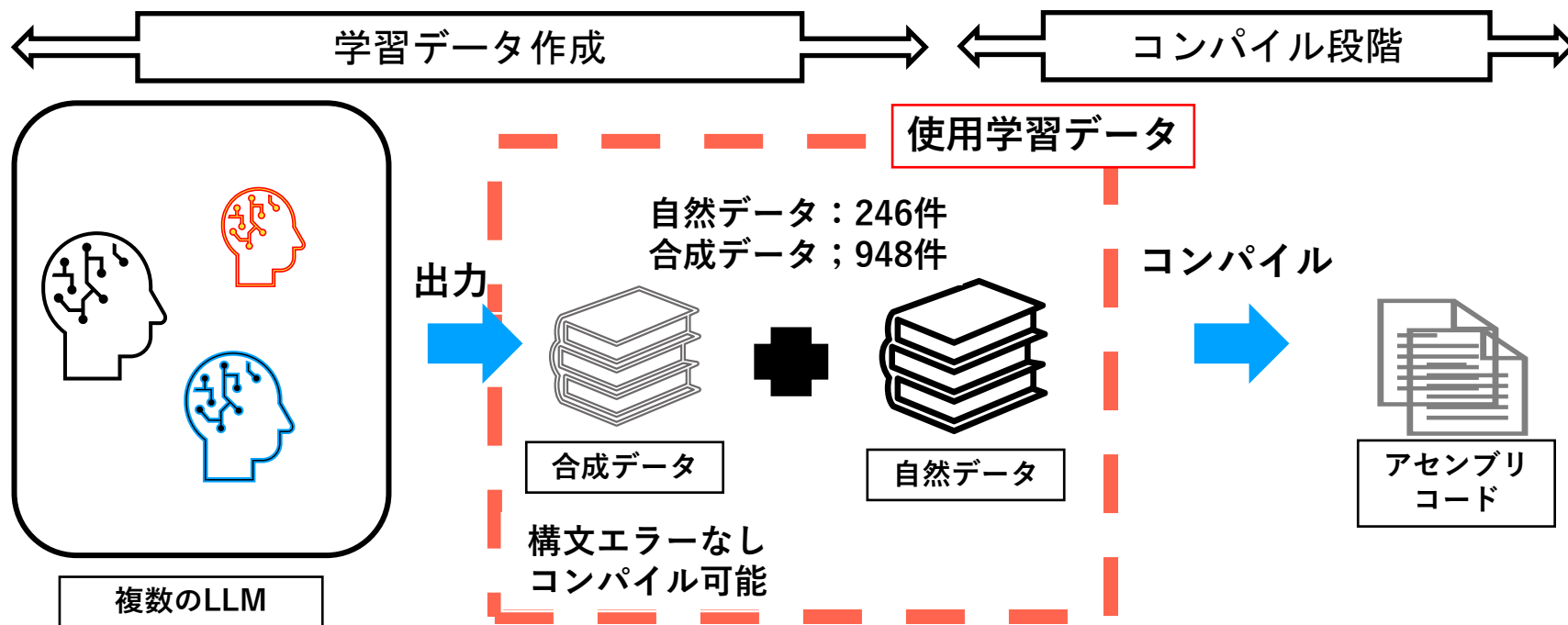
を小規模で容量制約のあるLLMを用いて検証.

研究課題

- Dartの学習データが少ない
 - Dartの合成データを追加
 - 類似言語のSwiftを追加

実験: データ準備

合成データ+自然データ



Swiftデータ+自然データ



実験

使用モデル

- Qwen3-4B-Thinking-2507
- DeepSeek-R1-0528-Qwen3-8B

→両モデルをLoRA+DoRAでファインチューニングした者を採用

評価指標

- CodeBLEU
 - 文字列の類似性 + 構文議の類似性 + 変数利用の類似性
- 実行可能性
 - 上位 k 個の内いくつかのコードがコンパイルを通るか

実験結果

モデル	CodeBLEU	compile@5
Qwen3-4B(base)	66.1	64.7%
Qwen(Dart)	71.3	79.4%
Qwen(Dart+swift)	69.2	64.7%

※全34個の関数を実施

- 実行可能性において64.7%から79.4%に上がっているが、統計的優位はなかった。
- CodeBLEUにおいて、4BモデルではSwiftの追加は数値向上の妨げとなった。（8Bモデルでは、Dartのみのファインチューニングモデルに対して8.7ポイントの向上を示した）

考察

合成データVS類似言語学習

- 4Bモデルでは、合成データによる学習モデルの方がSwift学習モデルより高いCodeBLEUを記録した。
 - 容量が小さいため、二つの言語パターンを学習しきれていない？
 - 二つの言語には異なる部分が多い。（メモリ管理、文法、標準ライブラリ等）その二つの特徴を分離できていない？