Program Decomposition and Translation with Static Analysis

Ali Reza Ibrahimzada

University of Illinois Urbana-Champaign

In Proceedings of 46th IEEE/ACM International Conference on Software Engineering Companion (ICSE '24 Companion)

April 14-20, 2024, Lisbon, Portugal

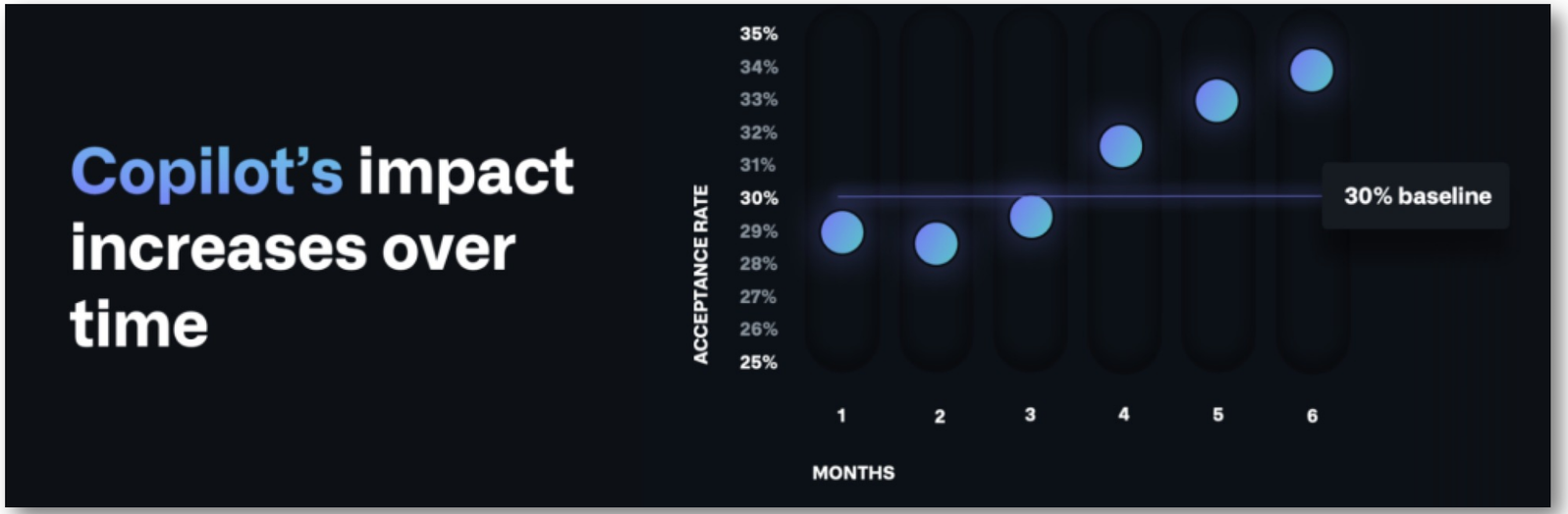
UNIVERSITY OF ILLINOIS URBANA-CHAMPAIGN

INTELLIGENT Code Analysis Testing

LLMs for Code Generation & Understanding

- LLMs are used for assisting developers
  - **2021 & 2022:** Codex, CodeT5, GitHub Copilot , AlphaCode, etc.
  - **2023 & 2024:** GPT-4, Gemini , CodeLlama , Claude 3, etc.

Today, GitHub Copilot has been activated by more than one million developers and adopted by over 20,000 organizations. It has generated over three billion accepted lines of code, and is the world's most widely adopted AI developer tool.



https://github.blog/2023-06-27-the-economic-impact-of-the-ai-powered-developer-lifecycle-and-lessons-from-github-copilot/

What is Code Translation?

Code translation converts source code from one programming language to another

Java Code

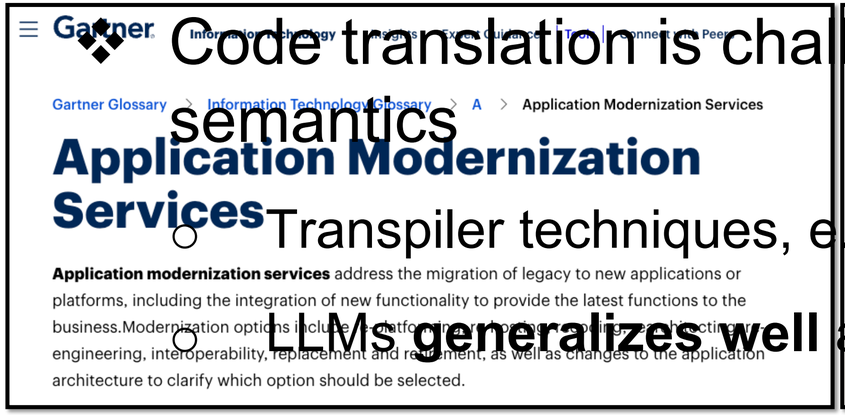
```java
public class Main {
    public static void main(String[] args) {
        int max = 5;
        for (int i = 0; i < max; i++) {
            System.out.println(i);
        }
    }
}
```

Python Code

```python
max = 5
for i in range(max):
    print(i)
```

Why Code Translation?

- Use cases of code translation:
  - Application modernization
  - Migrating legacy software to cloud-native applications
- Code translation is challenging as it requires understanding both syntax and semantics
  - Transpiler techniques, e.g., C2Rust, CxGO, Java2C#
  - LLMs **generalizes well** and generates more **natural code**
- Existing learning-based techniques use small models (**<1B** parameter) and fail to achieve high accuracy



Effectiveness of LLMs in Code Translation

Crafted benchmark datasets and real-life projects

Dataset	Source Language	Source Samples	# Tests	Target Language	#Translations
CodeNet[1]	C	200	200	C++,Go,Java,Python	800
	C++	200	200	C,Go,Java,Python	800
	Go	200	200	C,C++,Java,Python	800
	Java	200	200	C,C++,Go,Python	800
	Python	200	200	C,C++,Go,Java	800
Total / Average (CodeNet)	-	1,000	1,000	-	4,000
AVATAR[2]	Java	249	6,255	C,C++,Go,Python	996
	Python	250		C,C++,Go,Java	1,000
Total / Average (AVATAR)	-	499	6,255	-	1,996
Evalplus[3]	Python	164	2,682	Java	164
Commons-CLI[4]	Java	22	310	Python	22
Click[5]	Python	15	611	Java	15
Total / Average (All)	-	1,700	10,858	-	6,197

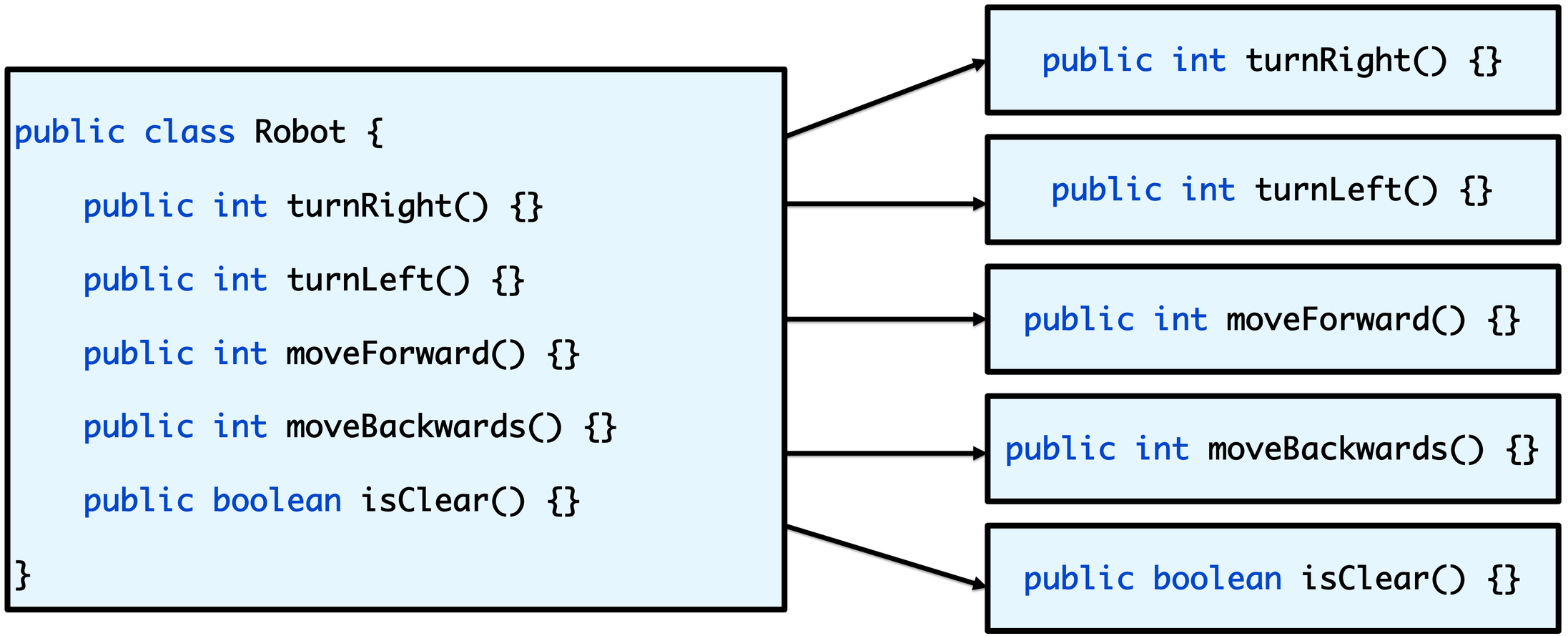
[1]https://github.com/IBM/Project_CodeNet [2]https://github.com/wasiahmad/AVATAR [3]https://github.com/evalplus/evalplus
[4]https://github.com/apache/commons-cli [5]https://click.palletsprojects.com/en/8.1.x/

Effectiveness of LLMs in Code Translation

How do state-of-the-art general and code LLMs perform in code translation across various benchmarks?

Dataset	Source Language	% Successful Translations						
		CodeGen	CodeGeeX	StarCoder	GPT-4	Llama 2	TB-Airoboros	TB-Vicuna
CodeNet	C	23.4%	14.9%	42.0%	83.0%	14.9%	18.8%	4.4%
	C++	14.0%	3.6%	39.1%	80.0%	9.5%	8.3%	3.4%
	Go	14.3%	5.9%	42.0%	85.5%	16.9%	6.6%	0.9%
	Java	21.3%	10.3%	30.3%	81.3%	13.9%	6.5%	0.1%
	Python	17.5%	7.3%	33.3%	79.9%	11.0%	6.5%	1.0%
Total / Average (CodeNet)	-	18.1%	8.4%	37.3%	82.0%	13.2%	9.3%	2.0%
AVATAR	Java	8.1%	1.8%	11.9%	70.8%	1.8%	5.0%	0.0%
	Python	3.8%	1.6%	14.2%	52.2%	4.7%	0.9%	0.9%
Total / Average (AVATAR)	-	5.9%	1.7%	13.0%	61.5%	3.2%	3.0%	0.4%
Evalplus	Python	16.5%	3.7%	22.0%	79.3%	1.2%	14.0%	7.9%
Commons-CLI	Java	0.0%	0.0%	0.0%	13.6%	0.0%	0.0%	0.0%
Click	Python	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%	0.0%
Total / Average (All)	-	8.1%	2.8%	14.5%	47.3%	3.5%	5.3%	2.1%

Program Decomposition



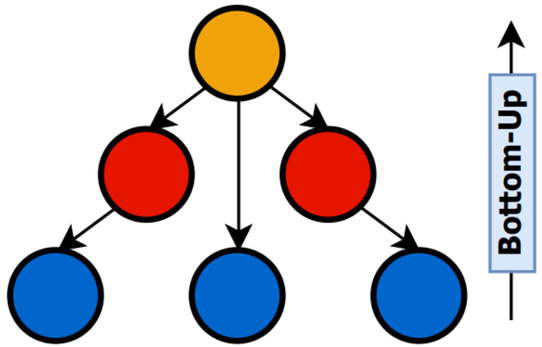
Method-level Program Decomposition

20 real-world Java projects with **60K** methods

Projects	% Files >2K Tokens	# Methods	Avg. Tokens / Method	% Methods >2K Tokens	% 2K Context
bcel	11.29%	4,094	70.42	0.15%	3.44%
beanutils	29.84%	2,675	107.09	0.07%	5.23%
cli	30.77%	582	97.91	0.17%	4.78%
codec	48.30%	1,788	189.29	0.84%	9.24%
collections	19.34%	6,354	74.37	0.02%	3.63%
csv	27.08%	871	102.53	0.11%	5.01%
daemon	27.78%	60	108.63	0.00%	5.30%
dbcp	38.52%	3,622	63.02	0.03%	3.08%
dbutils	13.54%	869	61.44	0.00%	3.00%
fileupload	16.67%	401	77.8	0.00%	3.80%
geometry	39.13%	6,615	124.93	0.03%	6.10%
imaging	14.78%	2,530	143.71	0.20%	7.02%
io	22.07%	5,957	77.94	0.07%	3.81%
jexl	25.70%	3,967	109.37	0.20%	5.34%
lang	40.34%	9,134	103.33	0.12%	5.05%
net	23.83%	2,023	98.22	0.15%	4.80%
pool	22.68%	1,377	94.13	0.00%	4.60%
rng	36.60%	3,245	139.69	0.52%	6.82%
text	28.32%	2,712	99.85	0.04%	4.88%
validator	38.00%	1,181	147.42	0.17%	7.20%
Average	**27.73%**	**3,002.85**	**104.55**	**0.14%**	**5.11%**

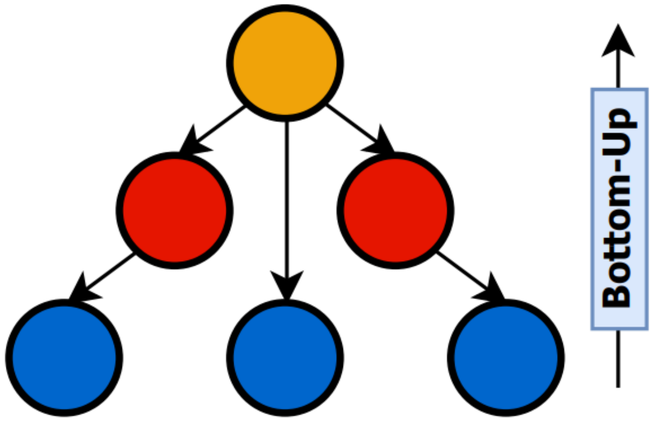
Real-world Project Translation

Commons-Cli translation enabled w/ program decomposition



Decomposition Technique	# Source Files	# Out-of-Context Inputs	% Context Occupied
No Decomposition	22	8	36%
Method-level Decomposition	22	0	3%

Prompt Engineering




```
You are an expert $SRC_LANG and $TRG_LANG programmer.

$ICL_EXAMPLE

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

Translate the following $SRC_LANG method to $TRG_LANG like the
example above:

$SRC_CODE

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

The following fields(s) are used in the method body and are already
translated:

$DEPENDENT_FIELDS

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

The following method(s) are called in the method body and are
already translated:

$DEPENDENT_METHODS

- - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - - -

$TRG_CODE: // Generated Code
```

Thank You!

SRC Paper

ICSE Paper

Leaderboard

Code