

UNIVERSITY OF SASKATCHEWAN

IWSC 2023 | 1 October 2023

Unveiling the Potential of Large Language Models in Generating Semantic and Cross Language Clones

By: **Palash R. Roy**, Ajmain I. Alam , Farouq Al-omari, Banani Roy, Chanchal K. Roy, Kevin A. Schneider



People's Choice Best Paper Award

iSE LAB
Interactive Software Engineering and Analytics Lab
UNIVERSITY OF SASKATCHEWAN

[SOAR]

UNIVERSITY OF SASKATCHEWAN
Software Research Lab



UNIVERSITY OF SASKATCHEWAN

Definition

Semantic Clones

```python
def factorial(n):
    x = n
    for i in range(1,n,1):
        x = x*(n-i)
    return x
```

```python
def factorial(n):
    if n <= 1:
        return 1
    else:
        return n * factorial(n-1)
```

Cross-language Clones

```java
// Java
public static boolean longerTF (boolean [] guess) {
    int variableTrue = 0;
    int variableFalse = 0;
    for (int x = 0;
    x < guess.length; x ++) {
        if (guess [x]) {
            variableTrue ++;
        } else {
            variableFalse ++;
        }
    }
    return variableTrue >= variableFalse;
}
```

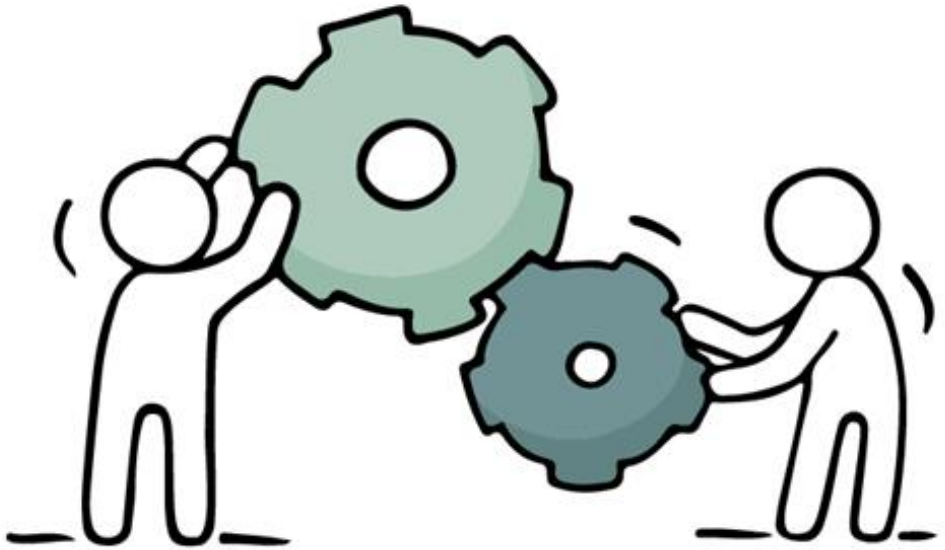
```python
// Python
def longerTF(guess):
    variableTrue = 0
    variableFalse = 0
    for x in range(len(guess)):
        if (guess[x]):
            variableTrue += 1
        else:
            variableFalse += 1
    return variableTrue >= variableFalse
```



UNIVERSITY OF SASKATCHEWAN

Motivation

- Inconsistent Changes.
- Facilitating Legacy Systems.
- Different Implementation.





UNIVERSITY OF SASKATCHEWAN

Research Question

- **(RQ1)** To what extent does GPT-3 demonstrate the capability to generate high-quality semantic clones?
- **(RQ2)** What is the efficacy of GPT-3 to generate cross-language clones (e.g., Java to Python)?



Methodology

- Followed GPTCloneBench Methodology ([ICSME 2023]).
- We randomly chose 15,000 semantic clone pairs and 12,000 cross-language clone pairs that were generated as part of the GPTCloneBench's prompt engineering/clone generation step.

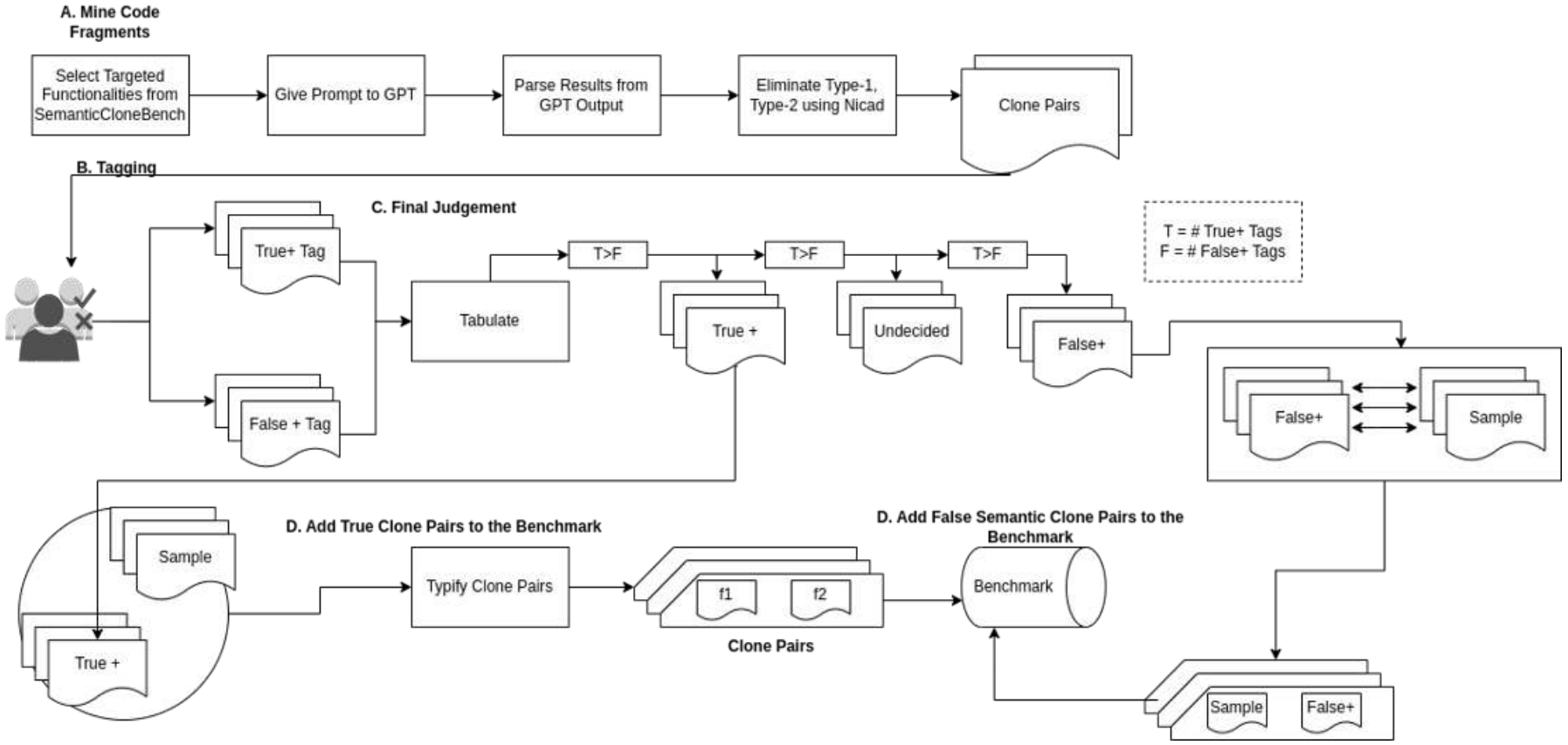


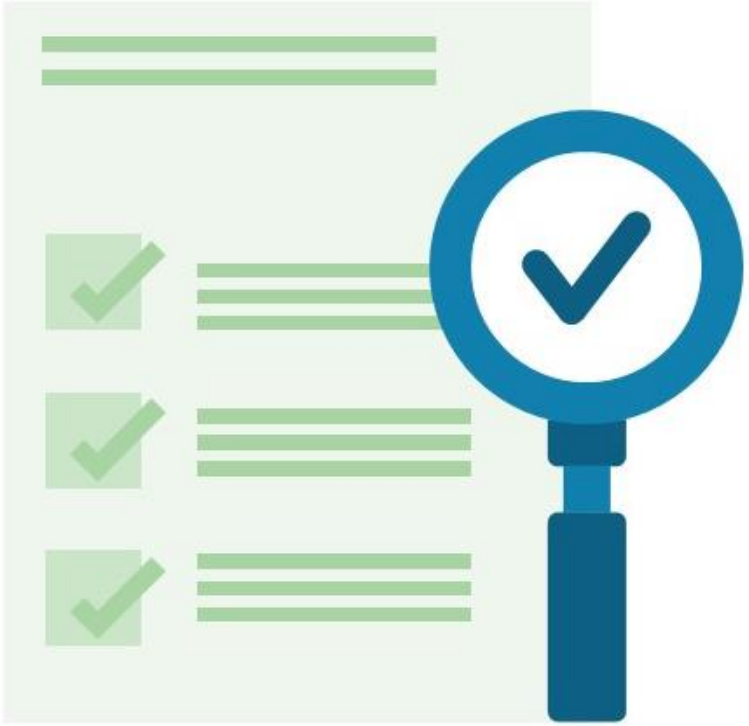
Image source: GPTCloneBench https://arxiv.org/pdf/2308.13963.pdf



UNIVERSITY OF SASKATCHEWAN

Manual Validation

- Two Rounds and One Discussion Session (For both semantic and cross-language clones)
- Nine Judges (6 undergrad research students and 3 Post-doctoral fellows).





Manual Validation (Round 1)

- Divided the semantic clone pairs into three groups after NiCad filtration.
- Each group consists of 2 undergraduate research students.
- Each member of the group validated each code individually.

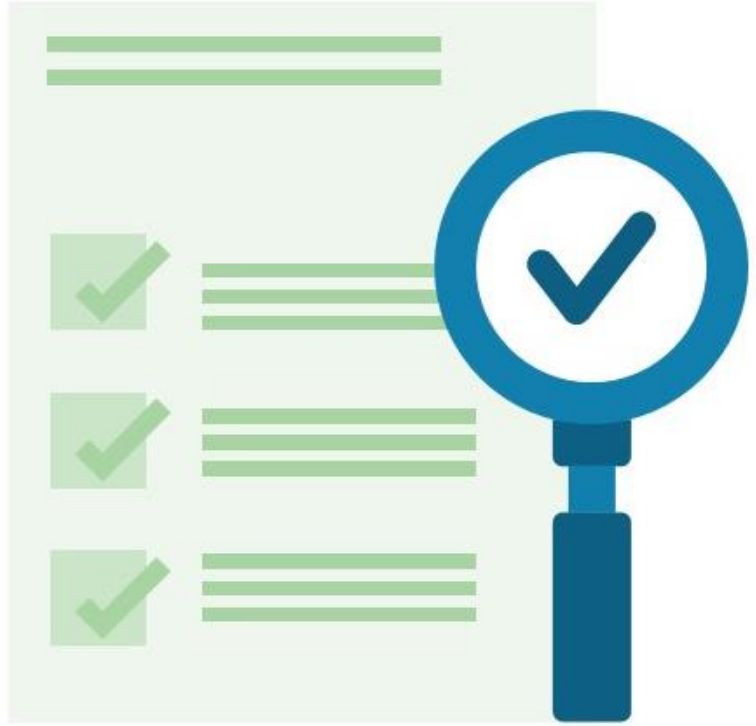
Clone Pair	Judge 1	Judge 2	Decision
Pair 1	True	True	Selected
Pair 2	False	False	Discarded
Pair 3	Undecided	Undecided	Going to Round 2
Pair 4	True	False	Going to Round 2
Pair 5	False	Undecided	Going to Round 2

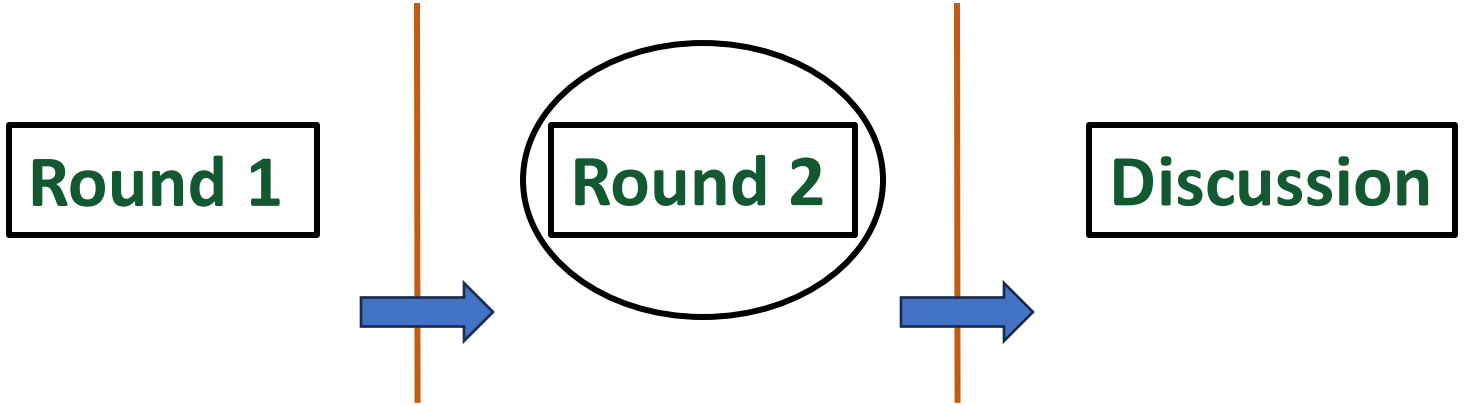
Table 1: Judges decision making rules for Round 1



Manual Validation

- Two Rounds and One Discussion Session (For both semantic and cross-language clones)
- Nine Judges (6 undergrad research students and 3 Post-doctoral fellows).





Manual Validation (Round 2)

- **Shuffled the undecided or conflicted pairs from Round 1 among the three group.**
- Each group consists of 2 undergraduate research students.
- We have made sure the groups are judging different codes compare to Round 1.

Clone Pair	Judge 1	Judge 2	Decision
Pair 1	True	True	Selected
Pair 2	False	False	Discarded
Pair 3	Undecided	Undecided	Going to Discussion
Pair 4	True	False	Going to Discussion
Pair 5	False	Undecided	Going to Discussion

Table 2: Judges decision making rules for Round 2



Cohen Kappa Interrater Agreement

- After completing both rounds, we have calculated Cohen Kappa Interrater Agreement among all the groups.



SEMANTIC CLONES COHEN-KAPPA INTERRATER AGREEMENT

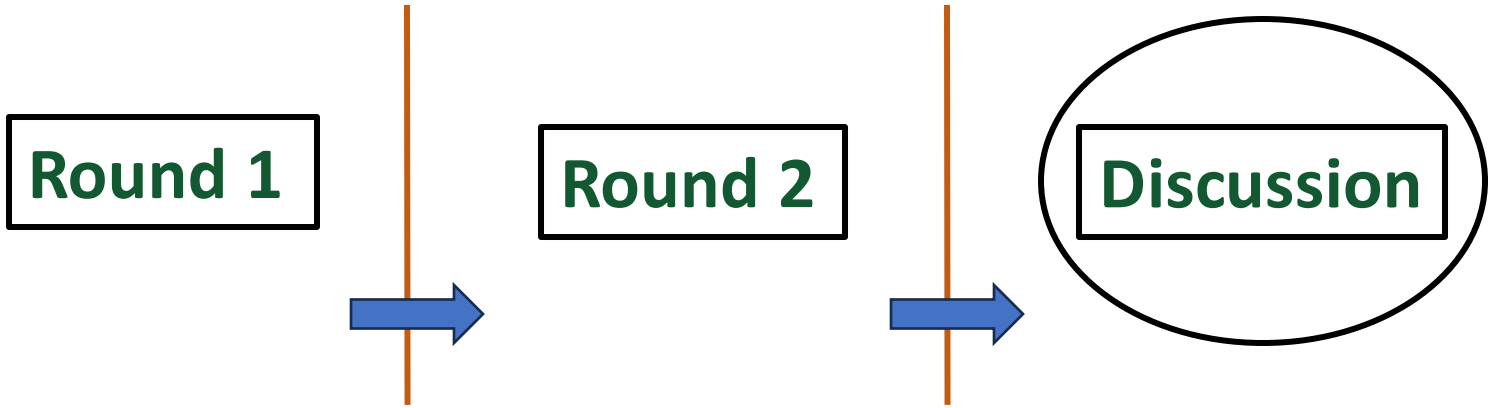
Round	Group	Interrater Agreement	Interpretation
Round 1	Group 1	0.700	Substantial Agreement
	Group 2	0.730	Substantial Agreement
	Group 3	0.77	Substantial Agreement
Round 2	Group 1	0.52	Moderate Agreement
	Group 2	0.58	Moderate Agreement
	Group 3	0.54	Moderate Agreement

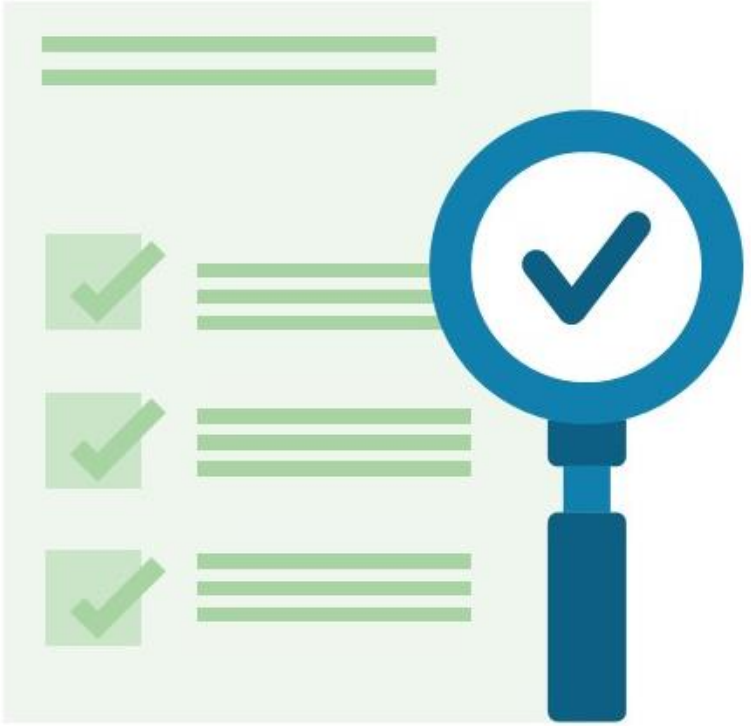
CROSS-LANGUAGE CLONES COHEN-KAPPA INTERRATER AGREEMENT

Round	Group	Interrater Agreement	Interpretation
Round 1	Group 1	0.69	Substantial Agreement
	Group 2	0.59	Moderate Agreement
	Group 3	0.73	Substantial Agreement
Round 2	Group 1	0.48	Moderate Agreement
	Group 2	0.56	Moderate Agreement
	Group 3	0.60	Substantial Agreement

Manual Validation

- Two Rounds and One Discussion Session (For both semantic and cross-language clones)
- Nine Judges (6 undergrad research students and 3 Post-doctoral fellows).





Manual Validation (Discussion)

- After 2 rounds, if there is still some clone pairs are remaining undecided or conflicted, 3 post-docs discussed among themselves to finally resolve all remaining pairs.

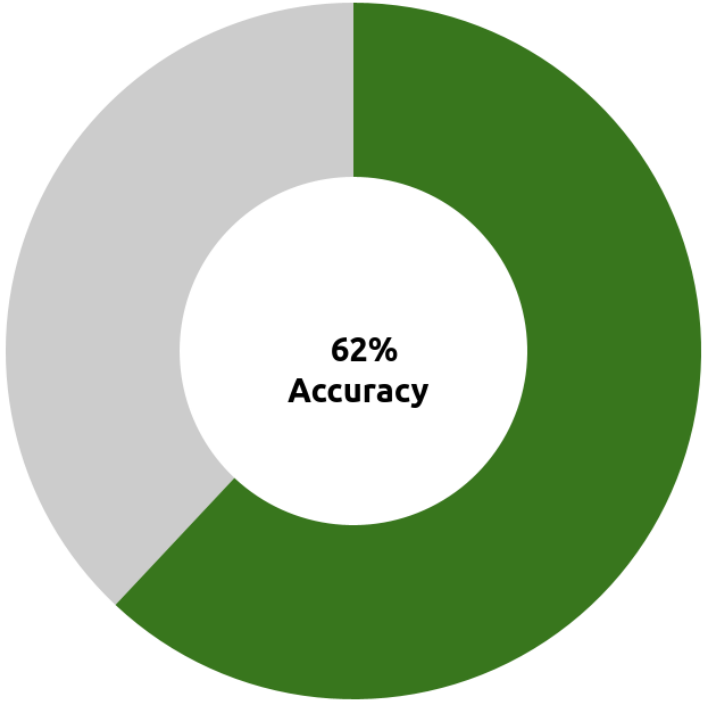
- *Finally, we have generated 9,321 semantic (from 15,000 clone pairs) and 10,950 cross-language (from 12,000 clone pairs) clone pairs.*



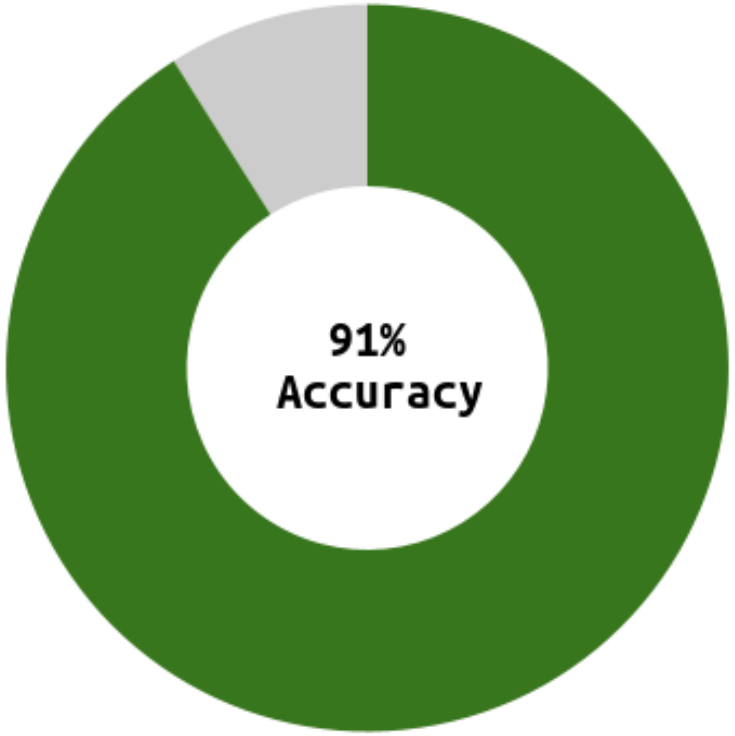
Results

$$Accuracy = \frac{Number\ of\ Validated\ True\ Clones}{Total\ Number\ of\ Randomly\ Selected\ Generated\ Clones} \quad (1)$$

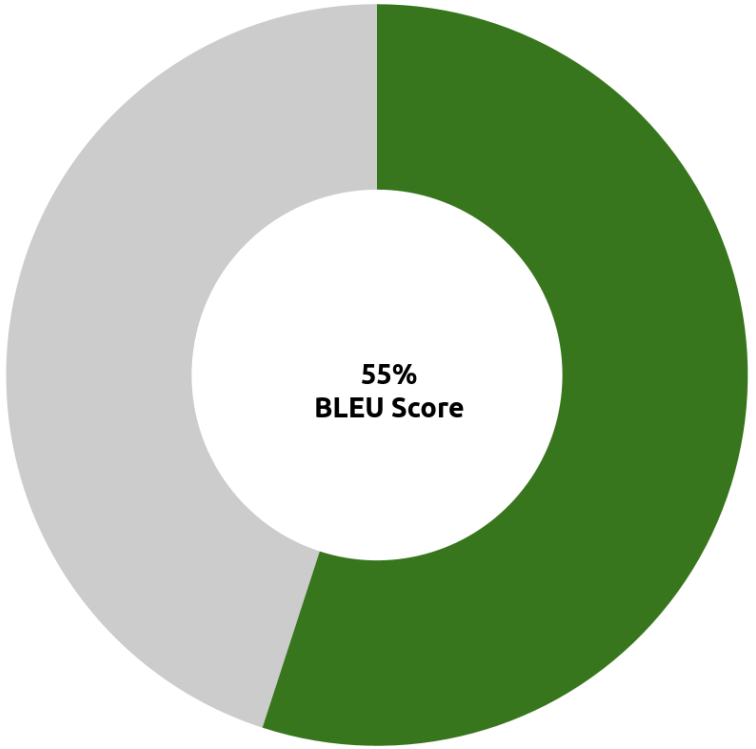
BLEU library:
https://www.nltk.org/_modules/nltk/translate/bleu_score.html



GPT semantic clones' generation accuracy



GPT cross-language clones' generation accuracy



GPT BLEU Score



UNIVERSITY OF SASKATCHEWAN

Discussion

(RQ1) To what extent does GPT-3 demonstrate the capability to generate high-quality semantic clones?

- Showcasing a notable degree of capability in generating high-quality semantic clones, as evidenced by an achieved accuracy of 62.14% with 0.55 BLEU score.

- High BLEU score can be interpreted that GPT generated codes are quality codes.

(RQ2) What is the efficacy of GPT-3 in accurately converting code snippets from one programming language to another?

- Demonstrating ability to comprehend the structural and syntactical intricacies inherent to different programming languages with accuracy of 91.25%.

Threats To validity

- Real vs Artificial Clone
- Biasness in Manual Evaluation





UNIVERSITY OF SASKATCHEWAN

Conclusion

- This research aims to introduce a transformative paradigm for code reuse, refactoring, migration and renovation.
- After thorough validation process, we have managed to get 9,321 true semantic clone pairs and 10,950 cross-language clone pairs after handling all the limitations of GPT-3.
- With a noteworthy accuracy rate of 62.14% and a 0.55 BLEU score, GPT-3 showcases its potential in accurately replicating semantic structures of a code.
- With an accuracy of 91.25% for cross-language clones, GPT shows prowess on generating cross-language clones.

References

- Roy, P. R., Alam, A. I., Al-omari, F., Roy, B., Roy, C. K., & Schneider, K. A. (2023). Unveiling the potential of large language models in generating semantic and cross-language clones. *arXiv preprint arXiv:2309.06424*.

- Alam, A. I., Roy, P. R., Al-Omari, F., Roy, C. K., Roy, B., & Schneider, K. A. (2023, October). GPTCloneBench: A comprehensive benchmark of semantic clones and cross-language clones using GPT-3 model and SemanticCloneBench. In *2023 IEEE International Conference on Software Maintenance and Evolution (ICSME)* (pp. 1-13). IEEE.



UNIVERSITY OF SASKATCHEWAN

Thank You

Acknowledgment

This work was supported in-part by the Natural Sciences and Engineering Research Council of Canada (NSERC) Discovery grants, NSERC USRAs, *the John R. Evans Leaders Fund (JELF) of the Canada Foundation for Innovation (CFI), and NSERC CREATE graduate program on Software Analytics Research (SOAR) grants.*



UNIVERSITY OF SASKATCHEWAN

Provocative Research Question

As GPT got a good accuracy on generating clones, can we utilize this process to generate new benchmark which can use to train/test Machine Learning models or evaluate existing clone detection tools?