CodeSynapse

- Manan Patel, Zac Perry, Shayana Shrestha, Eric Vaughan

THE UNIVERSITY OF
TENNESSEE

Team members

Manan Patel
mpatel65@vols.utk.edu

Zac Perry
zperry4@vols.utk.edu

Shayana Shrestha
sshres25@vols.utk.edu

Eric Vaughan
evaugha3@vols.utk.edu

Objective

- Test how well the Llama-3.2/3B, Deepseek-coder/6.7B, Phi/2.7B LLM models can translate code between Python, C++, and Java

- Check how accurate the models are when translating between languages that use different styles of programming (like object-oriented or imperative)

- Use existing dataset that already has code translations between Python, C++, and Java

- Build a website where users can enter code and choose the language they want it translated into, and the LLM will do the translation

Related Work

CodeBLEU (Ren et al., 2020)

- Syntax-aware metric for evaluating code generation
- Captures structure, data flow, and keywords better than BLEU
- *Used in our evaluation pipeline*

Unraveling LLMs in Code Translation (2024)

- Benchmarks multiple LLMs on cross-language translation tasks
- Shows smaller models (3B–7B) can be effective with proper tuning
- *Supports our model selection*

Methodology

- **Dataset Used**
  - **XLCoST :** Extract subsets relevant to **Python, C++, and Java.**
    - 100 samples from each language

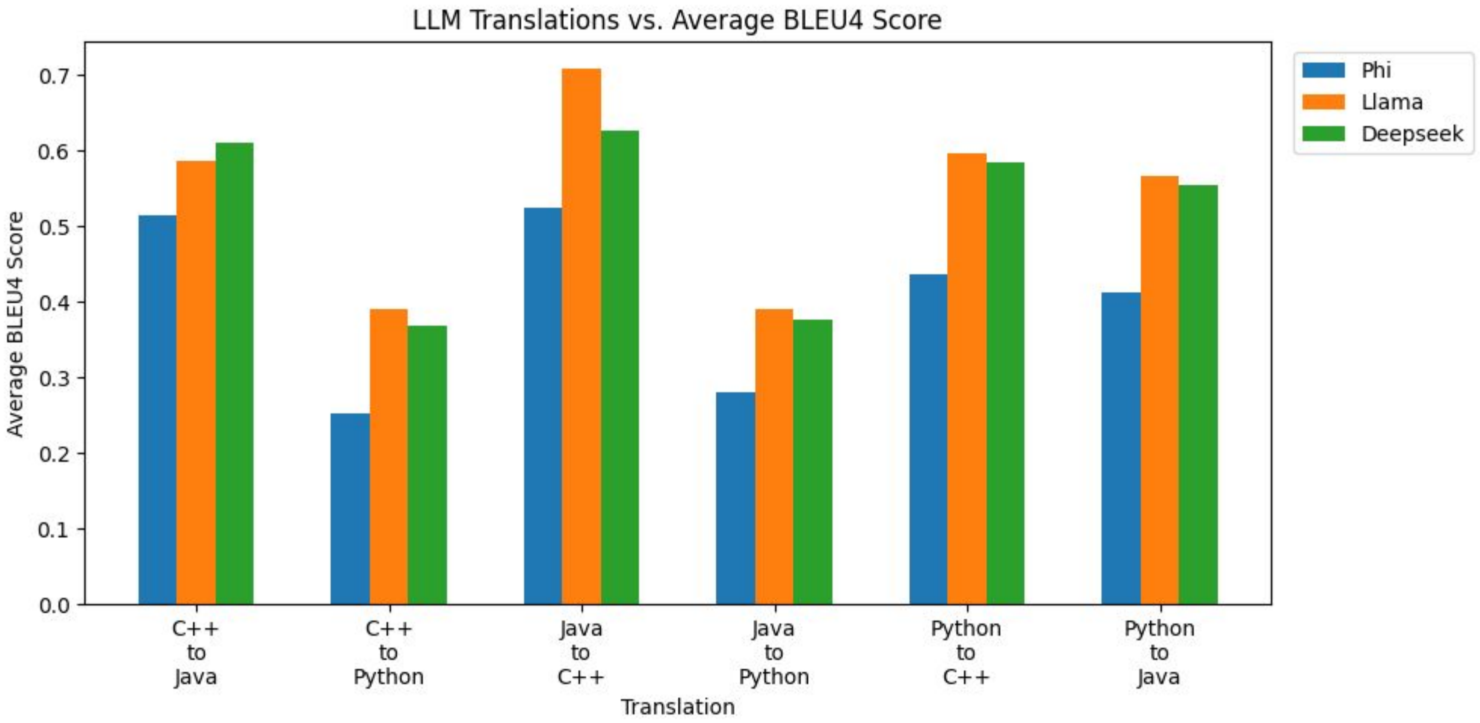
- **LLMs for Evaluation**
  - Llama-3.2/3B
  - Deepseek-coder/6.7B
  - Phi/2.7B

Methodology

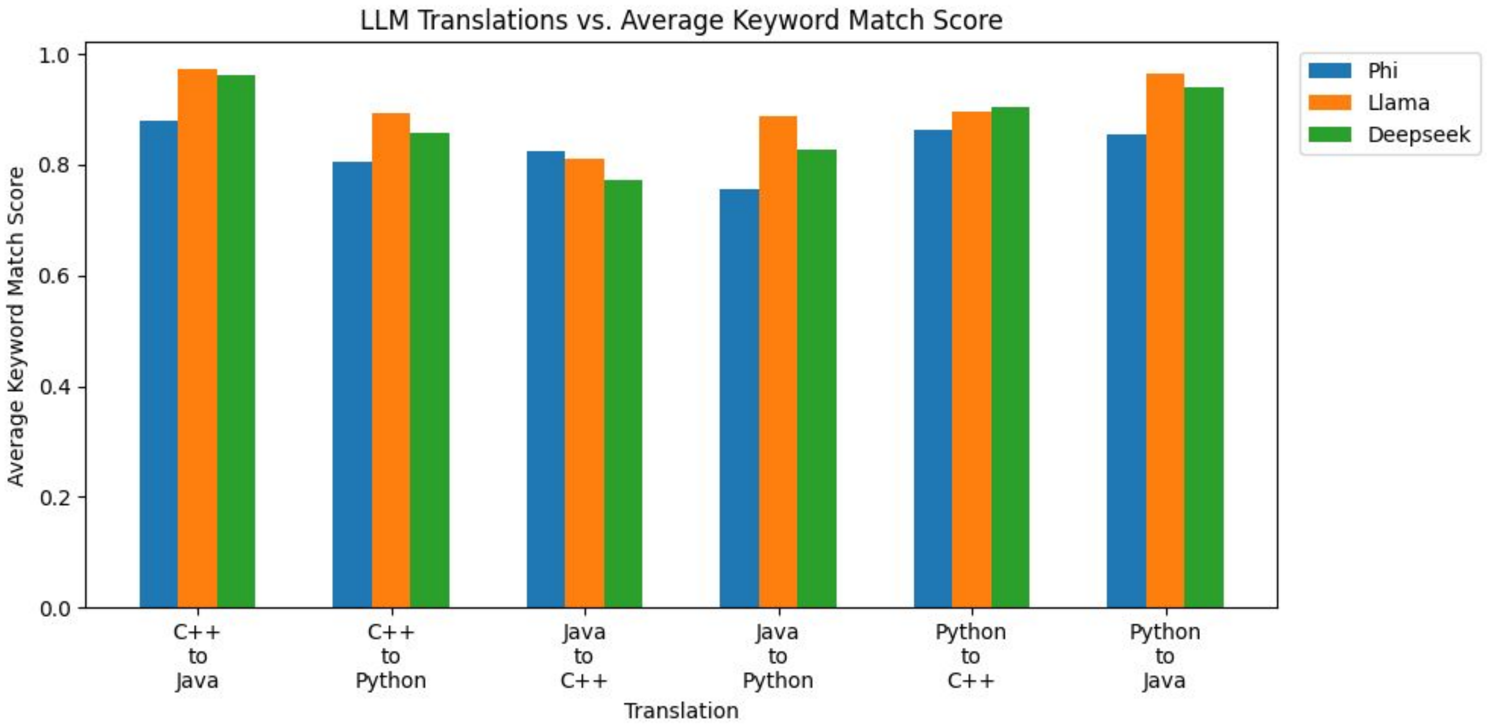
- **Evaluation Strategy**
  - Each LLM translates a shared set of code snippets between the three selected languages
  - CodeBleu, Bleu metric with different weights, keyword match

- **Final Product**
  - A web-based tool where users:
    - Submit a code snippet with source and target language (Python, C++, or Java)
    - Receive the translated code using the LLM that performs best for that language pair

Results - BLEU4

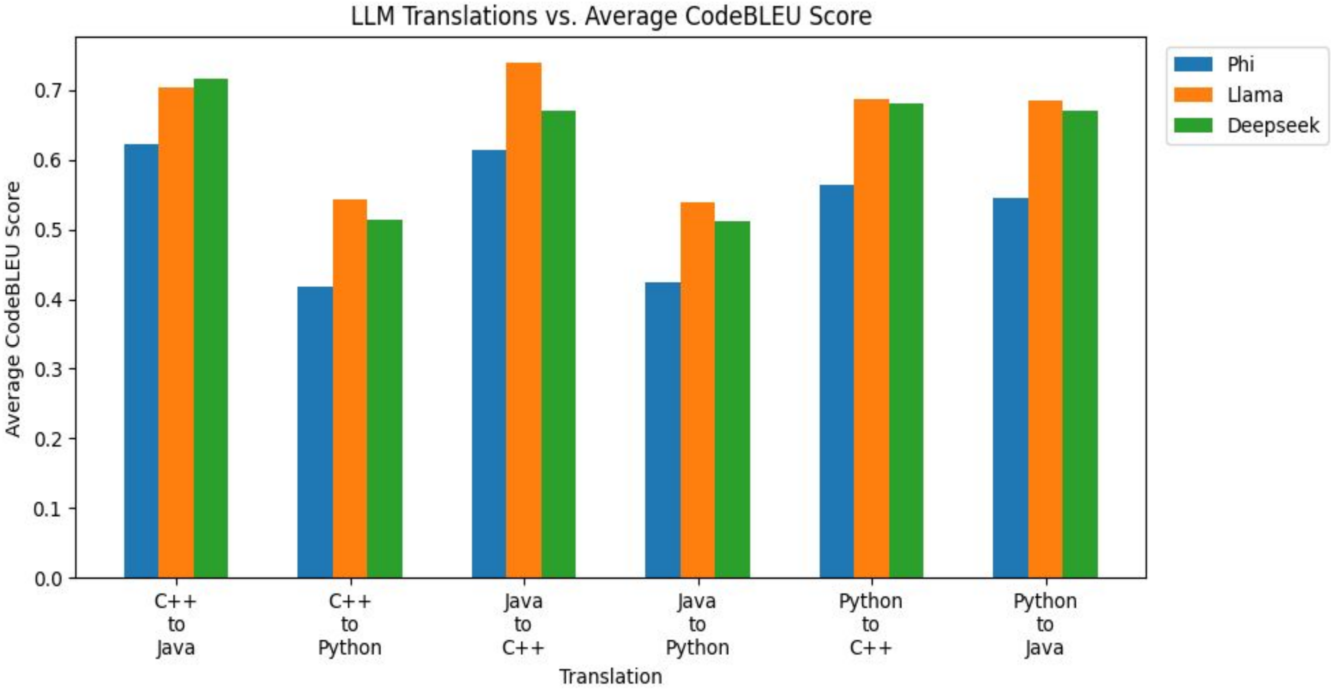


Results - Keyword match



Results - CodeBLEU

CodeBLEU = (0.7 x BLEU4) + (0.3 x Keyword match)



Conclusion

- Llama is the best overall LLM at code translations
  - Best at all translations except for C++ to Java

- No clear correlation between model size and translation ability
  - Llama mostly outperformed Deepseek

- Smaller LLMs, when well-tuned, can rival or outperform larger models in specific code tasks

- Metrics like CodeBLEU are more suitable than BLEU because they account for syntax and semantics

Limitations/Future Work

- Extend the study to evaluate more translation pairs for more languages
  - JavaScript/TypeScript, Go, Rust, etc.
- Evaluate more models
  - More GPT-based models, Claude 3.7-Sonnet
- Limitations:
  - Hardware, GPU, etc.
  - Money

References

- https://github.com/reddy-lab-code-research/XLCoST
- https://arxiv.org/abs/2009.10297
- https://arxiv.org/abs/2410.09812

Demo Time!!!

Thank you!