

Exploring the Possibilities of Applying Transfer Learning Methods for Natural Language Processing in Software Development

Wei Ding, 05.10.2020, Master Thesis

Advisor: Ahmed Elnaggar

Chair of Software Engineering for Business Information Systems (sebis)

Faculty of Informatics

Technische Universität München

www.matthes.in.tum.de

- Motivation and Introduction
- Research Questions
- Model Architecture
- Tasks and Datasets
- Schedule
- Result – Single Task Learning
- References

- **Motivation and Introduction**
- Research Questions
- Model Architecture
- Tasks and Datasets
- Schedule
- Result – Single Task Learning
- References

- Software Development in NLP
 - Software code is a language

- NLU and NLG applications:
 - Analytics Dashboards
 - Chatbot
 - Content Creation

- Visions of NLP in the Software Development world
 - Better readability of the code
 - Easier to compare and evaluate the code
 - Smoother developing process

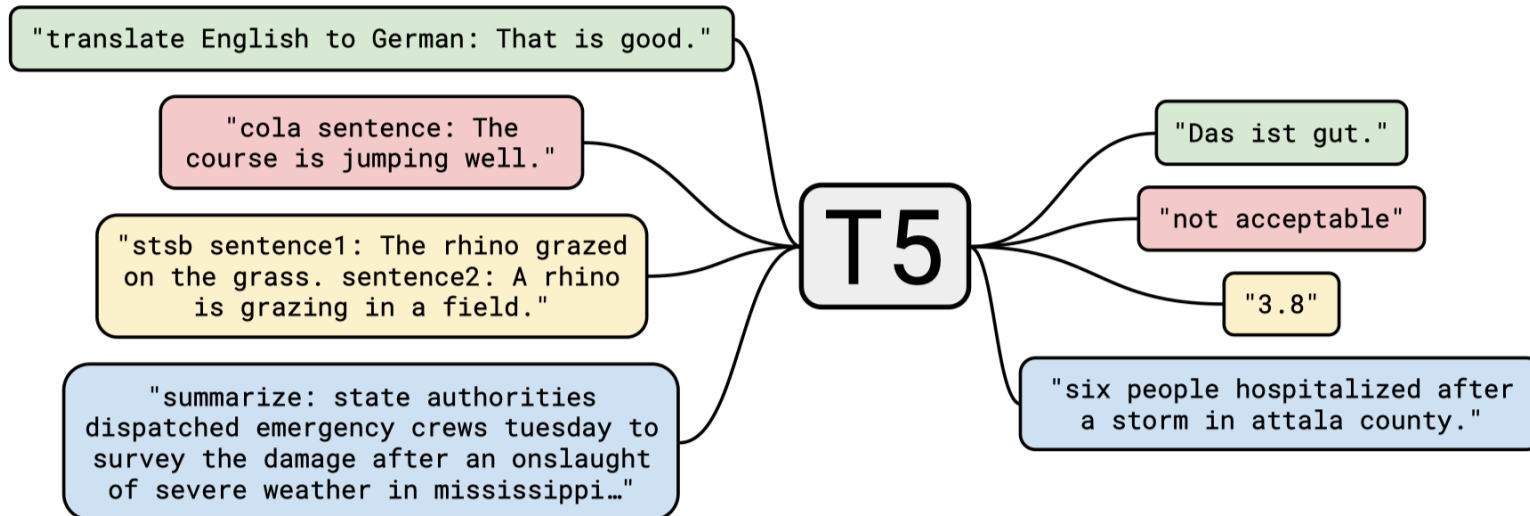
- Recent success of Transfer Learning



- Text-To-Text Transfer Transformer

- Advantages of T5

- same model, loss function, and hyperparameters on *any* NLP task



- Motivation and Introduction
- **Research Questions**
- Model Architecture
- Tasks and Datasets
- Schedule
- Result – Single Task Learning
- References



What kind of transfer learning techniques would work best on software development domain?



How would transfer learning improve the performance comparing with only training on the labeled data alone?

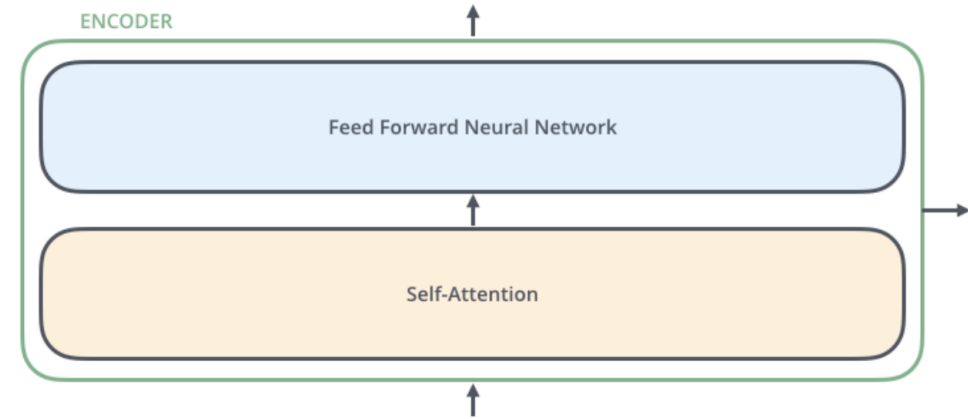
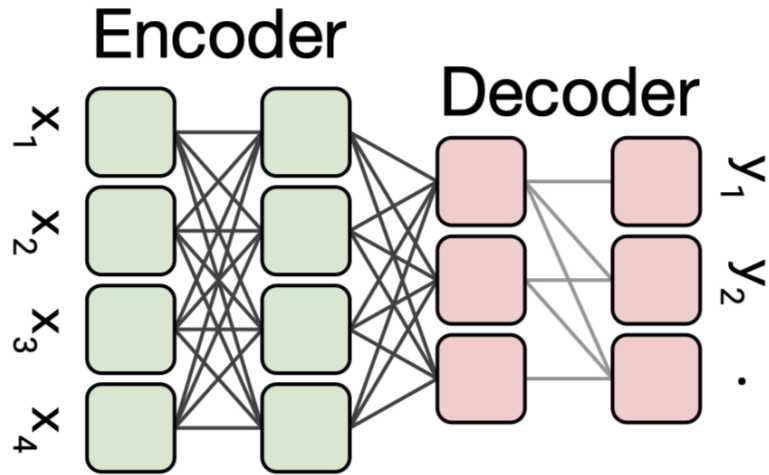


Would transfer learning perform better than multi-task learning for the similar tasks?

- Motivation and Introduction
- Research Questions
- **Model Architecture**
- Tasks and Datasets
- Schedule
- Result – Single Task Learning
- References

Model Architecture

- Encoder-decoder Models



- Fill-in-the-blank-style denoising pretraining objectives

Input
Thank you <X> me to your party <Y> week .

Target
<X> for inviting <Y> last <Z>

- Motivation and Introduction
- Research Questions
- Model Architecture
- **Tasks and Datasets**
- Schedule
- Result – Single Task Learning
- References

- Fine-tune tasks
 1. Code documentation generation
 - Code Language: Python, Java, Go, Php, Ruby, Javascript
 - Code Source: Github
 - Data Example:

Input: Code Function/Method

```
def parse_query_param(url, param):  
    try:  
        return parse.parse_qs(parse.urlparse(url)  
                                .query)[param][0]  
    except:  
        return None
```

Target: Natural Language Documentation Text

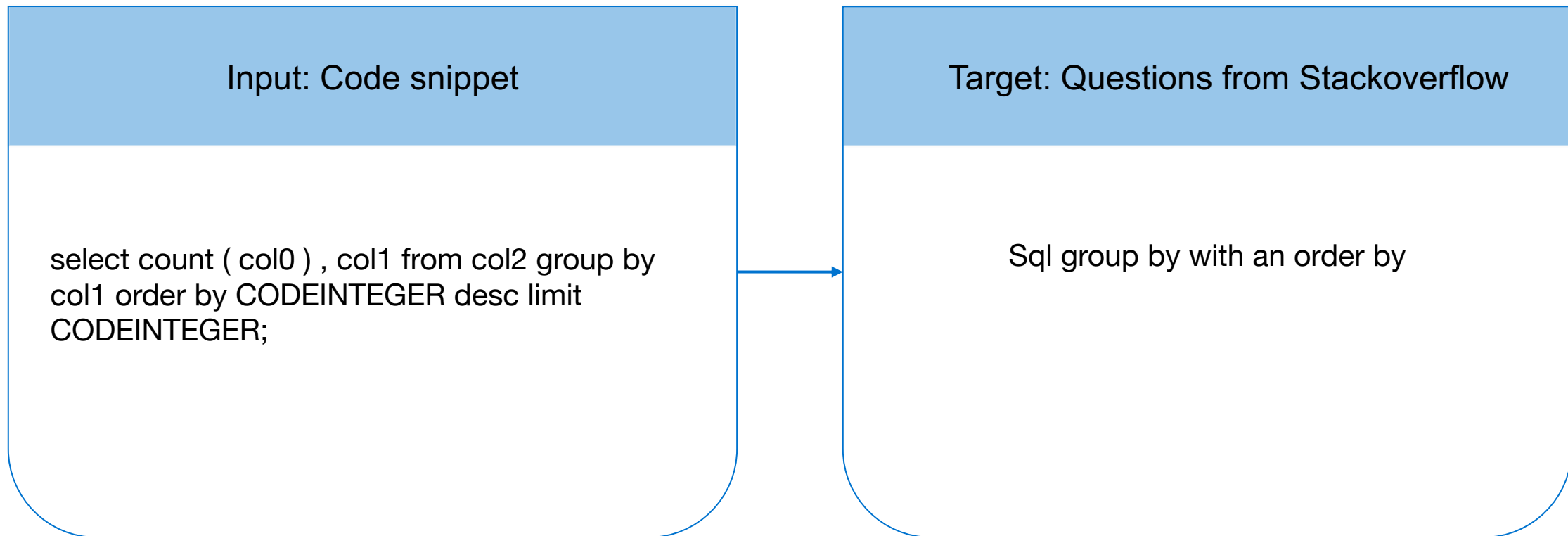
Parses the query string of a URL and returns the value of a parameter.

Args:

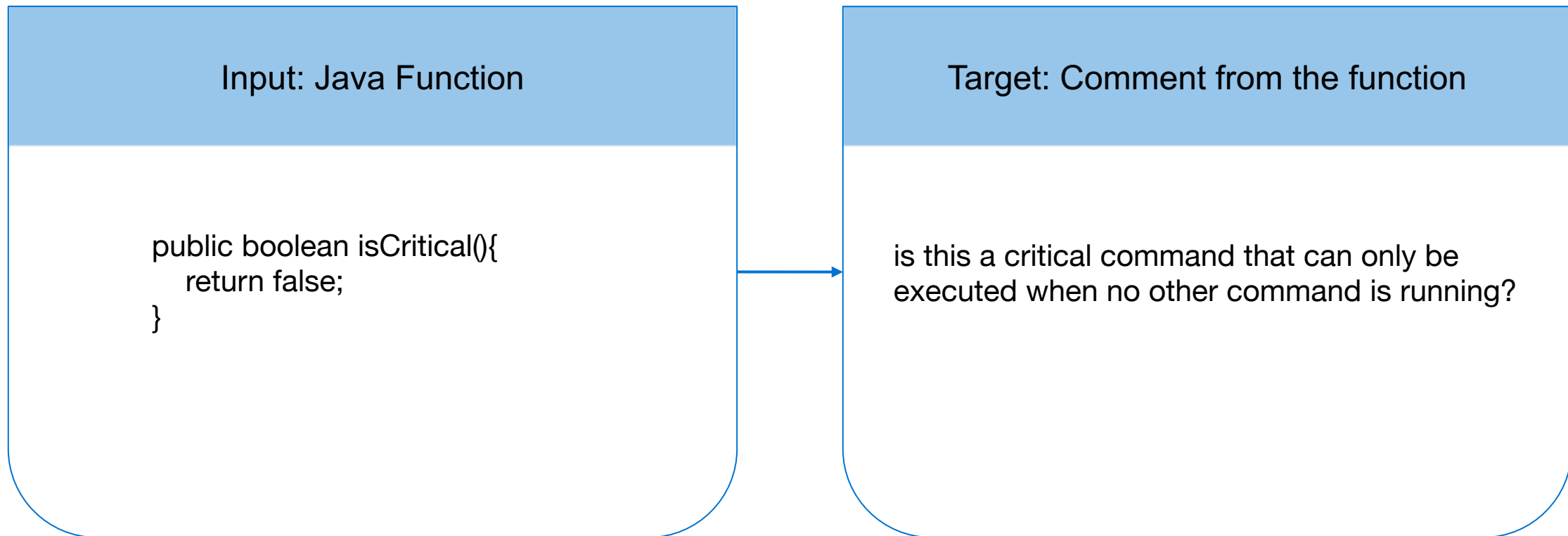
- url: A URL.
- param: A string representing the name of the parameter.

Returns: The value of the parameter.

- Fine-tune tasks
 2. Source summarization
 - Code Language: SQL, CSharp, Python
 - Code Source: StackOverflow
 - Data Example:

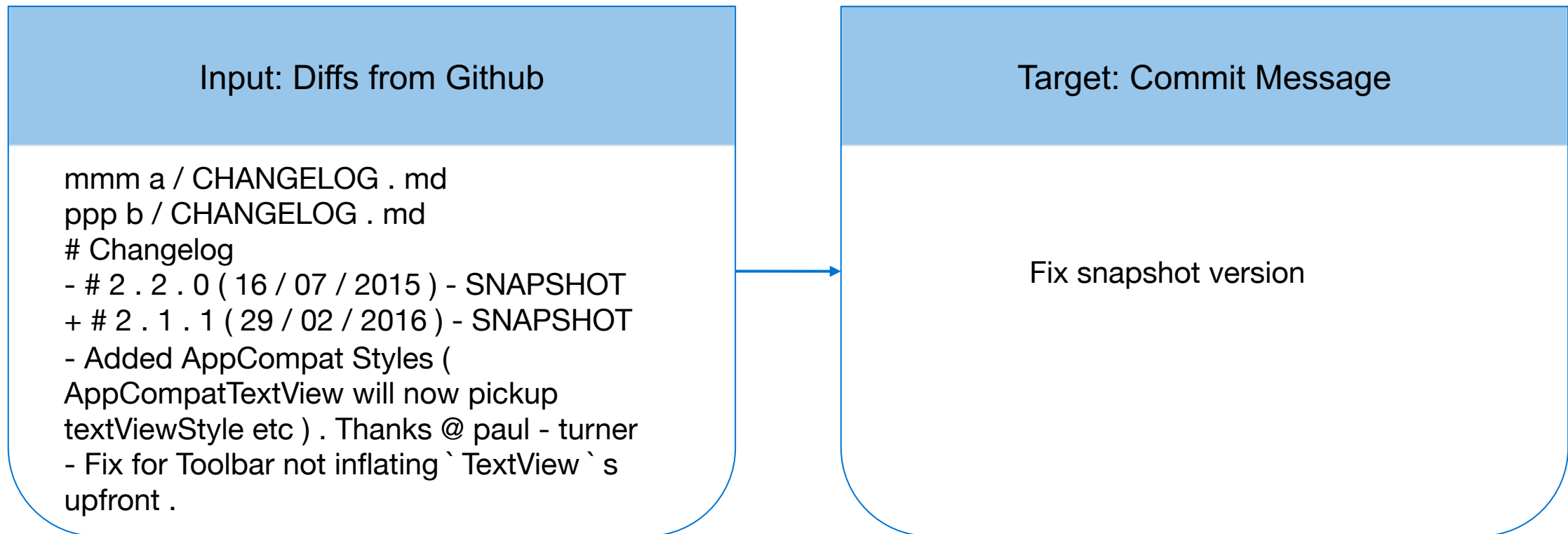


- Fine-tune tasks
 3. Code comment generation
 - Code Language: Java
 - Code Source: Github
 - Data Example:



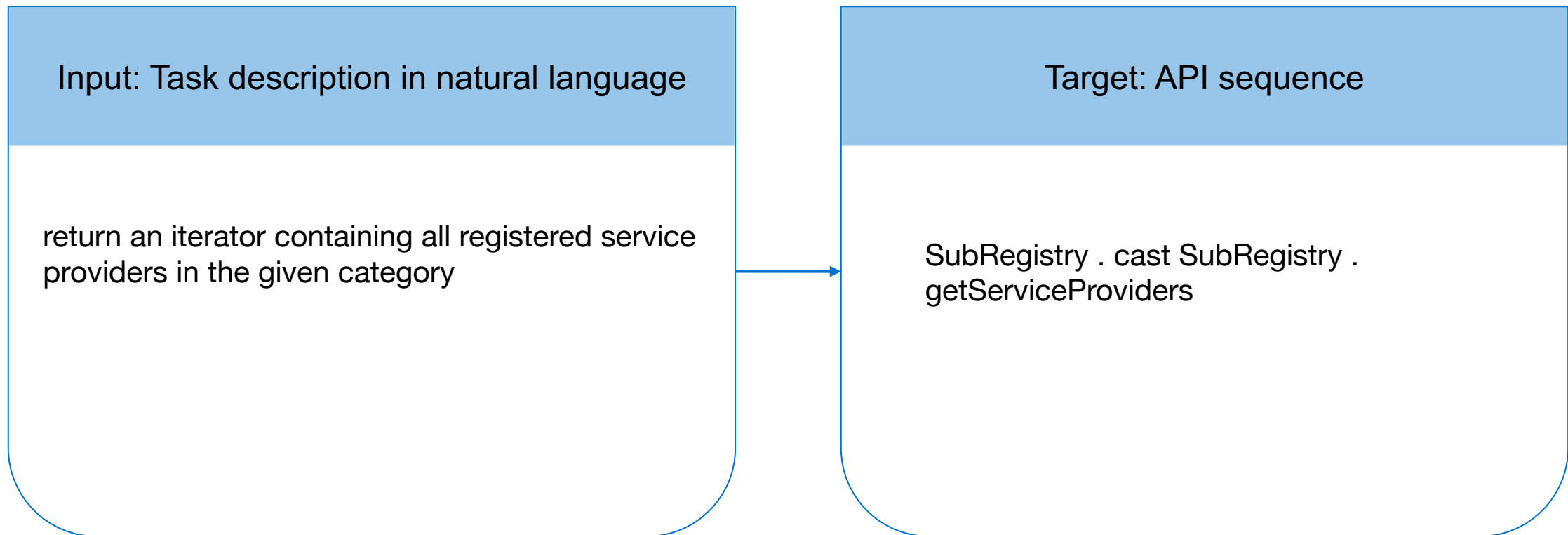
Tasks and Datasets

- Fine-tune tasks
 4. Commit message generation
 - Code Language: Java
 - Code Source: Github
 - Data Example:



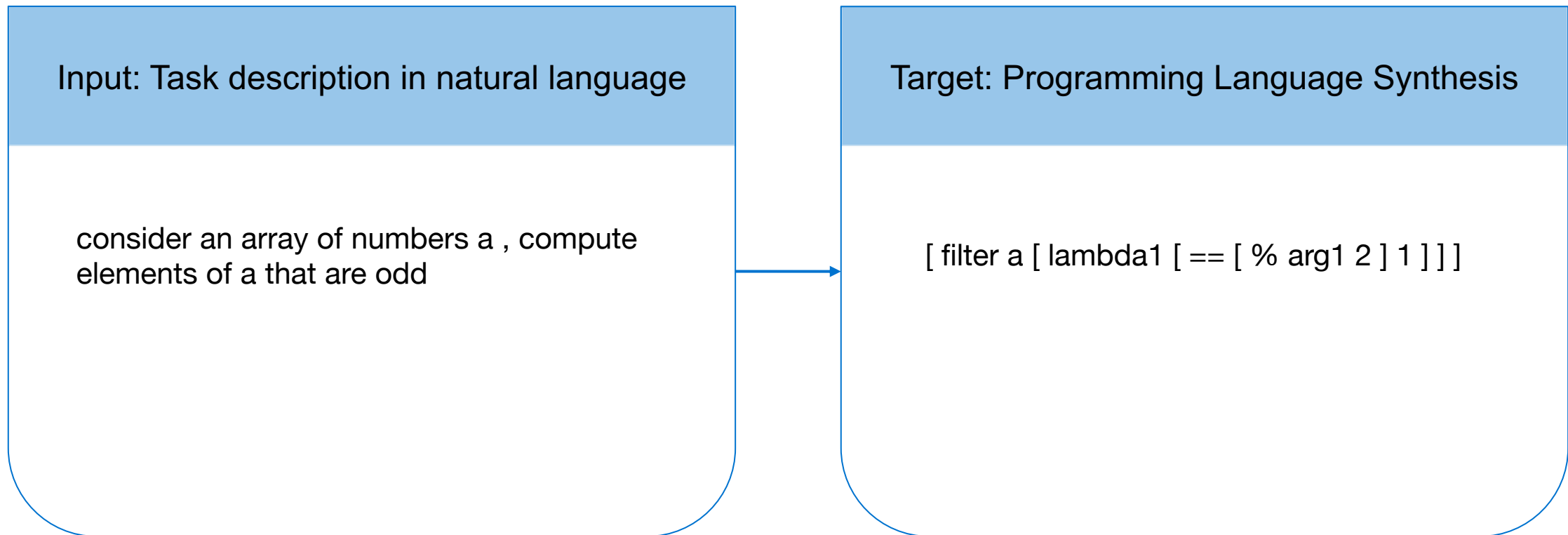
Tasks and Datasets

- Fine-tune tasks
 - 5. Api sequence recommendation
 - Code Language: Java
 - Code Source: Github
 - Data Example:



Tasks and Datasets

- Fine-tune tasks
 6. Programming Language and Synthesis:
 - Code Language: LISP
 - Code Source: Computer Science Student Homework
 - Data Example:

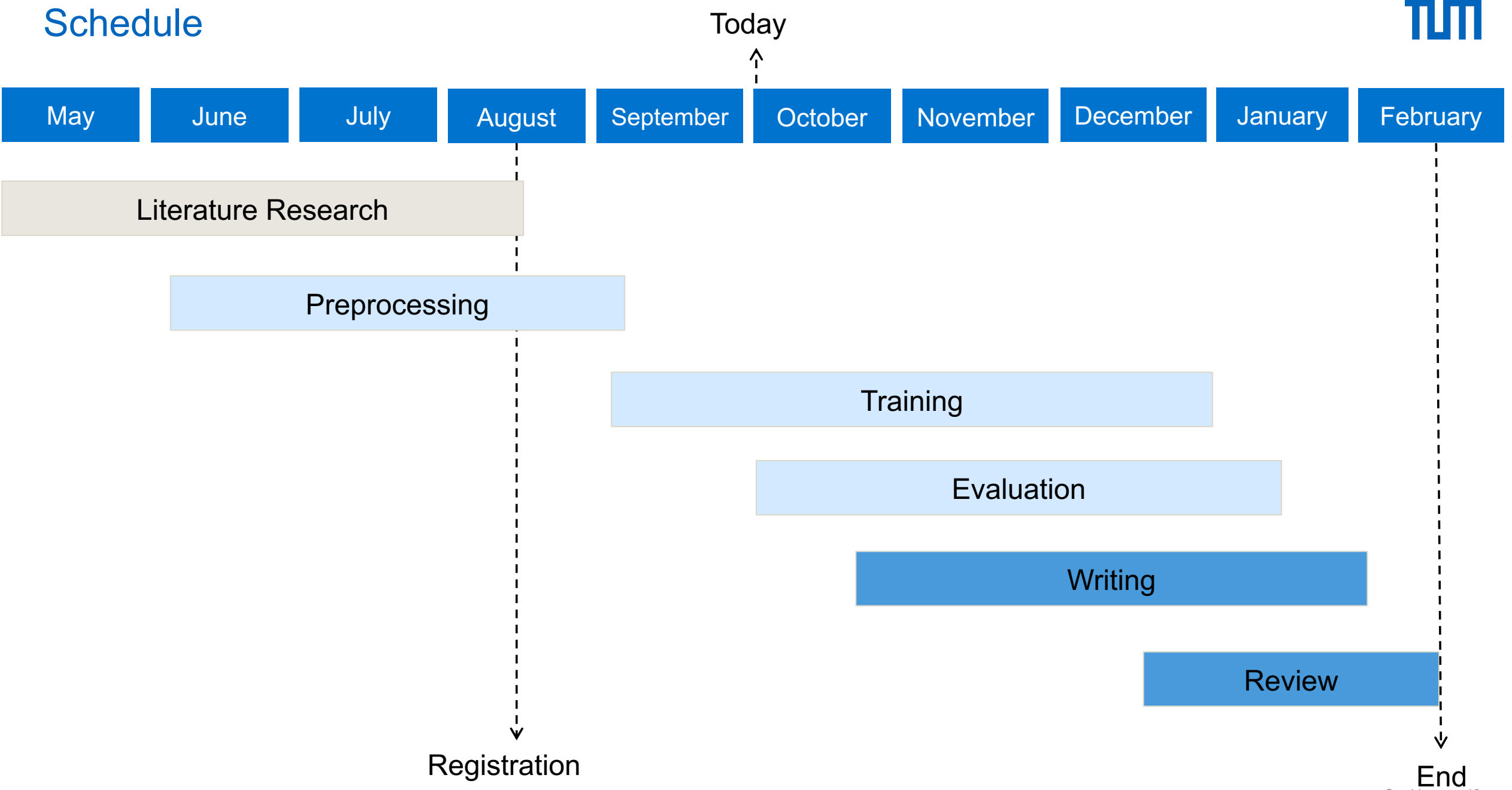


Tasks and Datasets

Language	Code Documentation Generation		Source Summarization		Code Comment Generation		Commit Message Generation		API Sequence Recommendation		Programming Language and Synthesis		Unsupervised
	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train	Test	Train
Java	164,923	10,955			468,000	58,638	26,208	3,000	7,475,850	10,000			1,747,022
Python	251,820	14,918	12,004	2,783									801,697
JavaScript	58,025	3,291											1,699,847
Go	167,288	8,122											378,979
Ruby	24,927	1,261											106,655
Php	241,241	14,014											260,508
C#			52,943	6,629									469,038
SQL			25,671	3,340									133,191
LISP											79,214	9,967	156,739
English													30,913,716

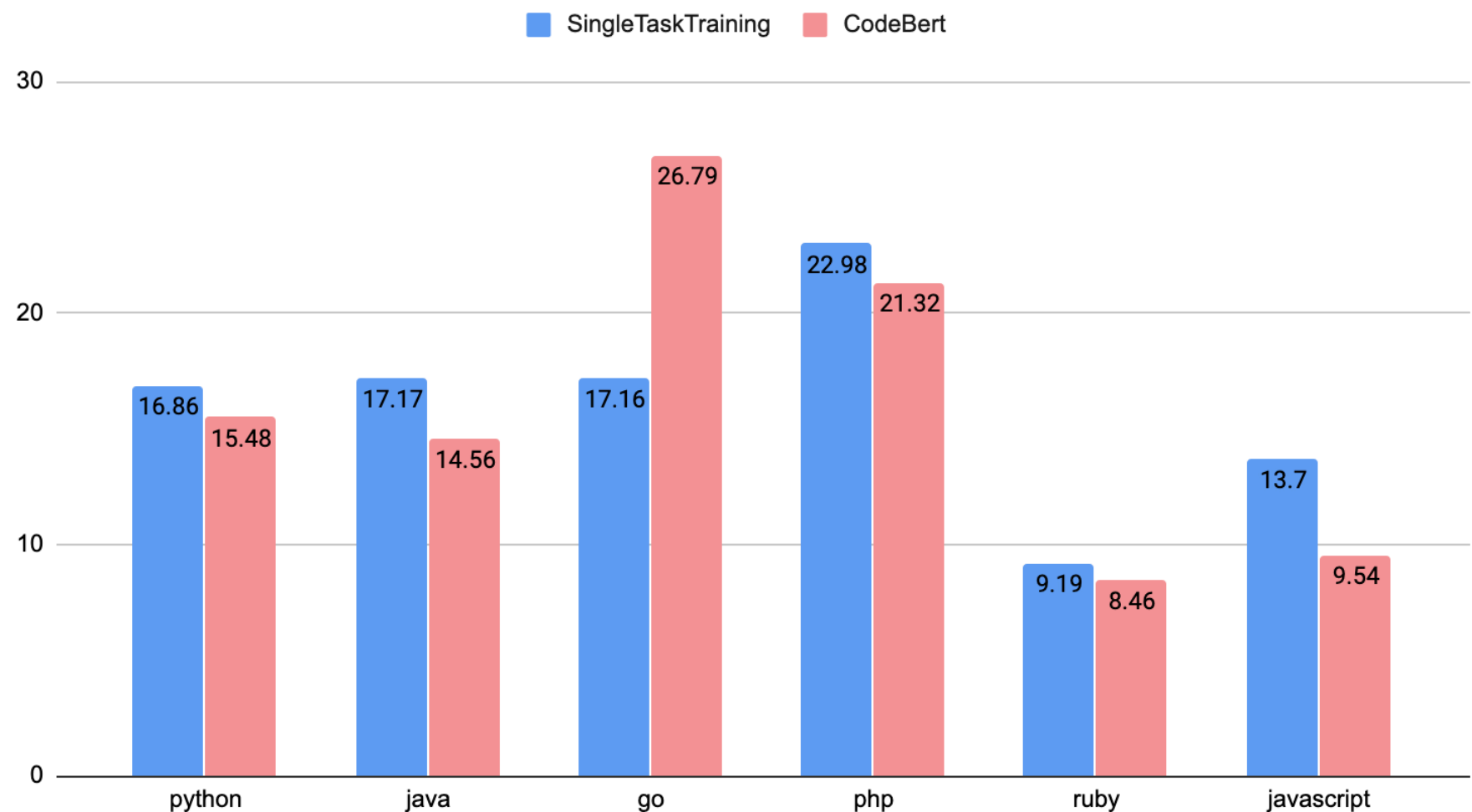
- Motivation and Introduction
- Research Questions
- Model Architecture
- Tasks and Datasets
- **Schedule**
- Result – Single Task Learning
- References

Schedule

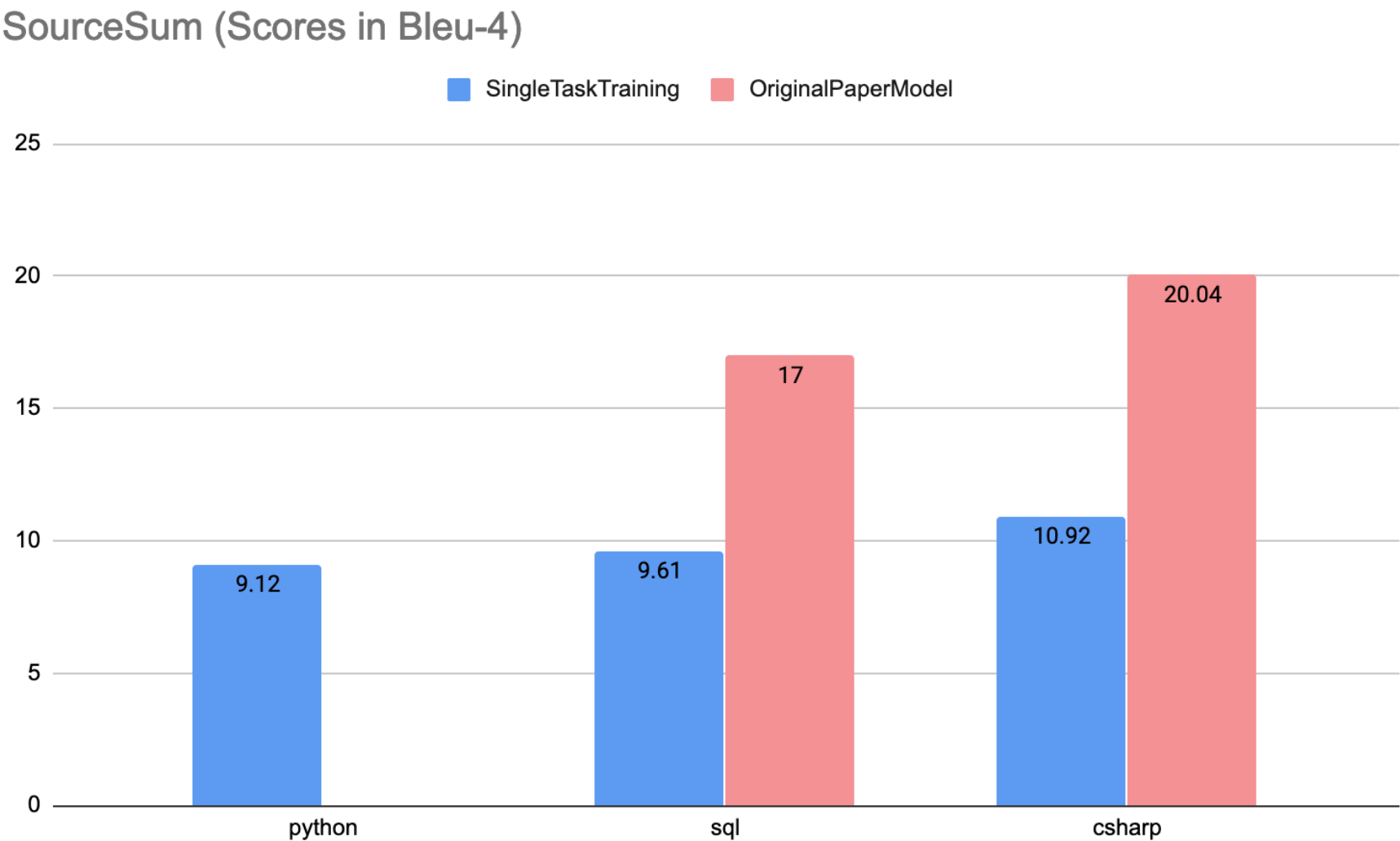


- Motivation and Introduction
- Research Questions
- Model Architecture
- Tasks and Datasets
- Schedule
- **Result – Single Task Learning**
- References

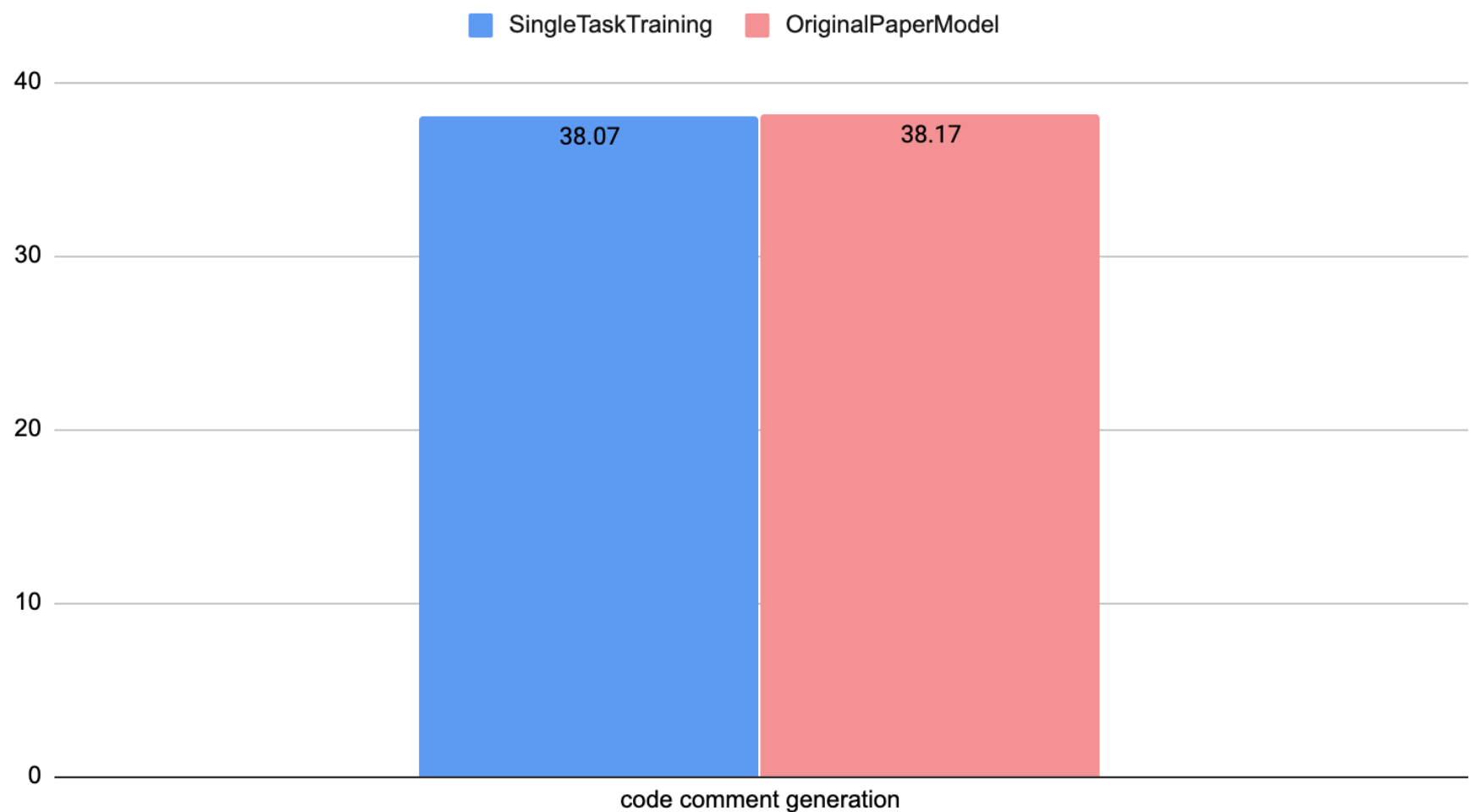
Function Documentation Generation (Scores in Bleu-4)



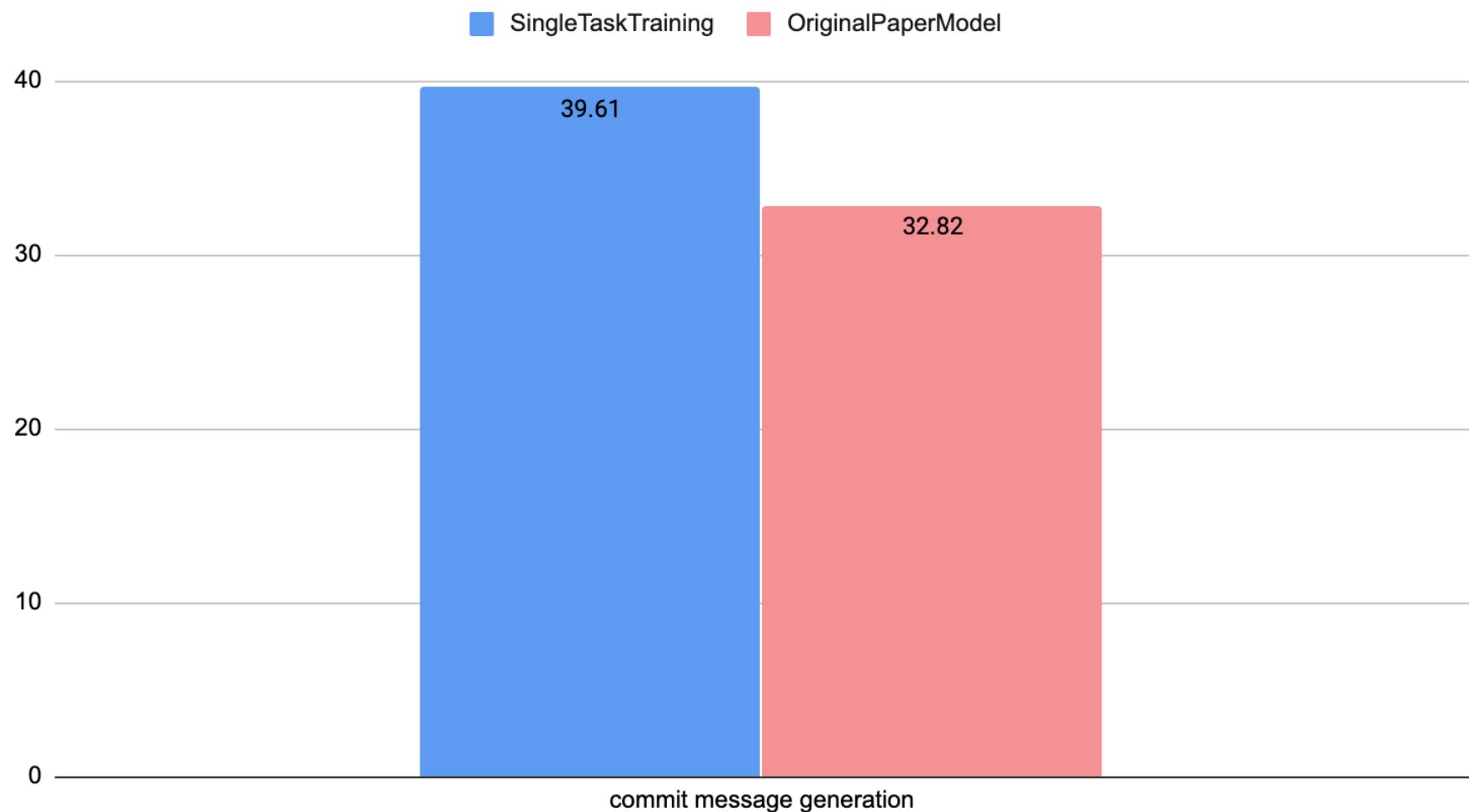
Result – Single Task Learning



Code Comment Generation (Scores in Bleu-4)

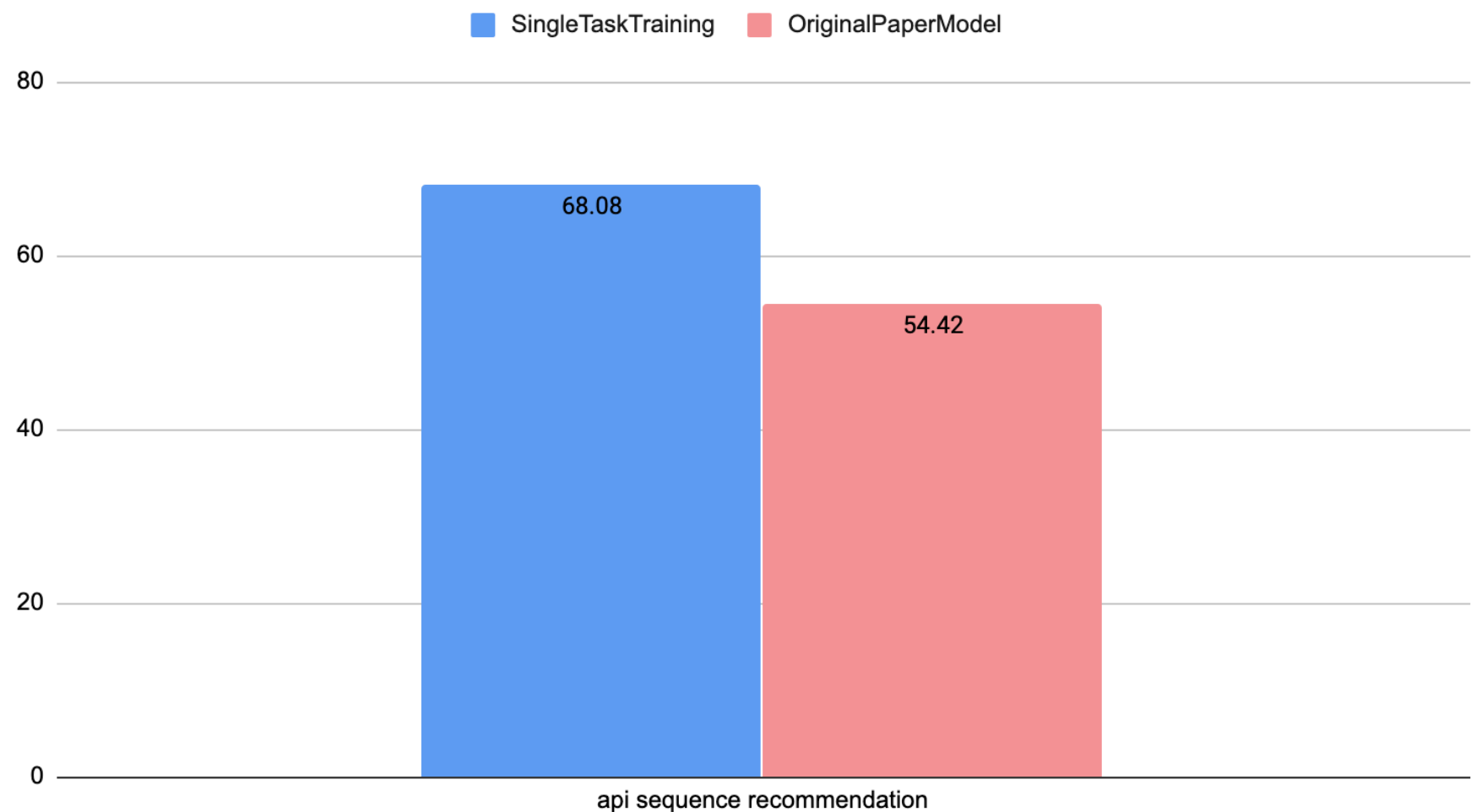


Commit Message Generation (Scores in Bleu-4)

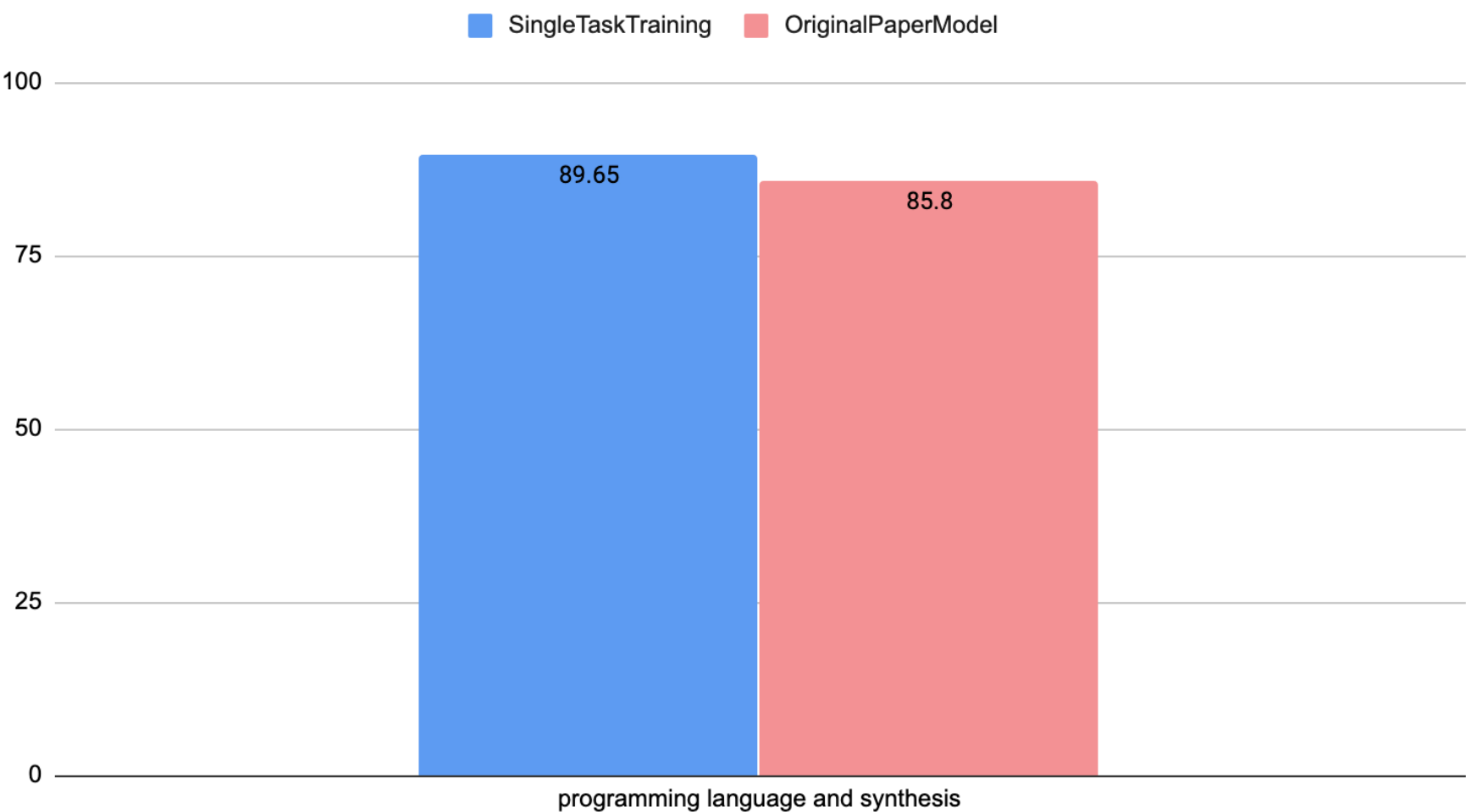


Result – Single Task Learning

Api Sequence Recommendation (Scores in Bleu-4)



Programming Language and Synthesis (Scores in Accuracy)



- Motivation and Introduction
- Research Questions
- Model Architecture
- Tasks and Datasets
- Schedule
- Result – Single Task Learning
- **References**

- SEVERINI, SILVIA. "Multi-task Deep Learning in the Software Development domain." (2019)
- Raffel, Colin, et al. "Exploring the limits of transfer learning with a unified text-to-text transformer." *arXiv preprint arXiv:1910.10683* (2019)
- Feng, Zhangyin, et al. "Codebert: A pre-trained model for programming and natural languages." *arXiv preprint arXiv:2002.08155* (2020)
- Husain, Hamel, et al. "Codesearchnet challenge: Evaluating the state of semantic code search." *arXiv preprint arXiv:1909.09436* (2019)
- Hu, Xing, et al. "Deep code comment generation." *2018 IEEE/ACM 26th International Conference on Program Comprehension (ICPC)*. IEEE, 2018
- Jiang, Siyuan, and Collin McMillan. "Towards automatic generation of short summaries of commits." *2017 IEEE/ACM 25th International Conference on Program Comprehension (ICPC)*. IEEE, 2017
- Gu, Xiaodong, et al. "Deep API learning." *Proceedings of the 2016 24th ACM SIGSOFT International Symposium on Foundations of Software Engineering*. 2016
- Polosukhin, Illia, and Alexander Skidanov. "Neural program search: Solving programming tasks from description and examples." *arXiv preprint arXiv:1802.04335* (2018)
- Iyer, Srinivasan, et al. "Summarizing source code using a neural attention model." *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*. 2016
- <http://jalammar.github.io/illustrated-transformer/>



Thanks!
Questions?

Technische Universität München
Faculty of Informatics
Chair of Software Engineering for Business
Information Systems

Boltzmannstraße 3
85748 Garching bei München

Tel +49.89.289. 17132
Fax +49.89.289.17136

matthes@in.tum.de
www.matthes.in.tum.de



Tasks and Datasets

- Pre-train tasks

Language	Code Search Net Corpus	1 Billion Word Corpus	150K Python Dataset	SQL Corpus	Java from PGA	C# from PGA
Java	1,026,898				720,124	
Python	652,583		149,114			
JavaScript	1,699,847					
Go	378,979					
Ruby	106,655					
Php	260,508					
C#						469,038
SQL				133,191		
English		30,913,716				