

# Evaluating Large Language Models for Code Translation: Effects of Prompt Language and Prompt Design

Aamer Aljagthami, Mohammed Banabila, Musab Alshehri, Mohammed Kabini, and Mohammad D. Alahmadi

Department of Software Engineering, College of Computer Science and Engineering, University of Jeddah  
Jeddah, Saudi Arabia

{aaljagthami0002, mbanabeelah, malshehri0625, mkabini, mdalahmadi}@uj.edu.sa

## ABSTRACT

Large language models (LLMs) have shown promise for automated source-code translation, a capability critical to software migration, maintenance, and interoperability. Yet comparative evidence on how model choice, prompt design, and prompt language shape translation quality across multiple programming languages remains limited. This study conducts a systematic empirical assessment of state-of-the-art LLMs for code translation among C++, Java, Python, and C#, alongside a traditional baseline (TransCoder). Using BLEU and CodeBLEU, we quantify syntactic fidelity and structural correctness under two prompt styles (concise instruction and detailed specification) and two prompt languages (English and Arabic), with direction-aware evaluation across language pairs. Experiments show that detailed prompts deliver consistent gains across models and translation directions, and English prompts outperform Arabic by 13–15%. The top-performing model attains the highest CodeBLEU on challenging pairs such as Java→C# and Python→C++. Our evaluation shows that each LLM outperforms TransCoder across the benchmark. These results demonstrate the value of careful prompt engineering and prompt language choice, and provide practical guidance for software modernization and cross-language interoperability.

## KEYWORDS

Code Translation, LLMs, BLEU, CodeBLEU.

## ACM Reference Format:

Aamer Aljagthami, Mohammed Banabila, Musab Alshehri, Mohammed Kabini, and Mohammad D. Alahmadi. 2025. Evaluating Large Language Models for Code Translation: Effects of Prompt Language and Prompt Design. In *Proceedings of Make sure to enter the correct conference title from your rights confirmation email (Conference acronym 'XX)*. ACM, New York, NY, USA, 5 pages. <https://doi.org/XXXXXX.XXXXXXX>

## 1 INTRODUCTION

Large language models (LLMs) have shown promise for automated source code translation, a capability critical to software migration, maintenance, and interoperability [1]. Translating across languages

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.

*Conference acronym 'XX, June 03–05, 2018, Woodstock, NY*

© 2025 Copyright held by the owner/author(s). Publication rights licensed to ACM.

ACM ISBN 978-x-xxxx-xxxx-x/YY/MM...\$15.00

<https://doi.org/XXXXXX.XXXXXXX>

can reduce technical debt, enable reuse, and support polyglot development workflows [2]. At the same time, code translation remains challenging due to differences in type systems, idioms, libraries, and runtime models [3].

Despite rapid progress, comparative evidence on LLM-based code translation is still limited. Prior studies often vary in datasets, metrics, and evaluation setups, which hinders fair comparison [4]. The roles of prompt design and prompt language are not well quantified across models and directions. Direction-aware analyses are also scarce, even though translation difficulty can differ markedly from source to target [5]. Baseline comparisons to traditional systems such as TransCoder are likewise inconsistent [4].

This paper conducts a systematic empirical assessment of LLMs for code translation among C++, Java, Python, and C#, with a side-by-side comparison to a traditional baseline under a unified setup. We measure performance with BLEU and CodeBLEU to capture both surface correspondence and structural correctness [6]. We study two prompt styles (concise instruction and detailed specification) and two prompt languages (English and Arabic), and we use direction-aware evaluation across language pairs. Our experiments show three consistent patterns: detailed prompts yield measurable gains across models and directions, English prompts outperform Arabic by roughly 13–15% on average, and each LLM surpasses the baseline under identical evaluation settings. These results demonstrate the value of careful prompt engineering and prompt language choice, and they provide actionable guidance for software modernization and cross-language interoperability.

Our contributions are summarized as follows:

- We conduct a systematic empirical assessment of LLMs for code translation among C++, Java, Python, and C#, with a side-by-side comparison to a traditional baseline under identical evaluation settings.
- We study the effects of prompt design by contrasting concise instructions with detailed task specifications, and we quantify the impact of prompt language by evaluating English and Arabic variants.
- We provide direction-aware analyses across language pairs to reveal asymmetries that affect translation difficulty and model behavior.
- We report comprehensive results with BLEU and CodeBLEU, and derive practical recommendations for deploying LLMs in automated code migration workflows, and provide a replication package<sup>1</sup>

<sup>1</sup><https://doi.org/10.5281/zenodo.17065727>

Our experiments indicate three consistent patterns. First, detailed prompts deliver measurable gains across models and translation directions. Second, English prompts outperform Arabic by approximately 13–15% on average. Third, all evaluated LLMs exceed the TransCoder baseline under the same conditions, with the strongest CodeBLEU observed on challenging directions such as Java→C# and Python→C++. These findings demonstrate the value of careful prompt engineering and prompt language choice, and they offer actionable guidance for practitioners seeking reliable, high-quality code translation at scale.

## 2 RELATED WORK

### 2.1 Classical and Neural Code Translation

Early systems approached code translation with rule-based compilers and compiler-driven rewriting, which ensured syntactic fidelity but struggled with idioms, library mappings, and semantic gaps across ecosystems [7]. Neural machine translation reframed the task as sequence-to-sequence learning over tokens or subwords, sometimes augmented with abstract syntax trees or grammar constraints to improve structural validity [8]. Large-scale pretraining for code, as in PLBART and CodeT5, improved downstream translation and repair by leveraging mined repositories and task-adaptive finetuning [9]. TransCoder demonstrated that denoising objectives and mined parallel data provide strong baselines, although performance varied by direction and often required post-processing [10].

### 2.2 LLM-based Translation, Prompting, and Comparability

Recent large language models trained on mixed natural language and code have advanced program synthesis, editing, and translation without task-specific finetuning [11] [12]. Prompt design has emerged as a central factor, including explicit constraints, structured I/O formats, and function-signature hints [13]. Multilingual prompting raises additional questions about the effect of prompt language on model behavior, yet controlled comparisons remain limited in code translation [14]. Our study addresses these gaps by using a unified evaluation protocol across four languages (C++, Java, Python, C#), by contrasting concise versus detailed prompts and English versus Arabic prompts, and by reporting direction-aware analyses that expose asymmetries between source and target.

Compared with prior work, we provide a head-to-head comparison of contemporary LLMs against a traditional baseline under identical settings, isolate the effects of prompt style and prompt language, and present results stratified by translation direction. This design improves comparability and yields practical guidance for software modernization and cross-language interoperability.

## 3 EMPIRICAL STUDY

This section states the research questions, describes the datasets and tasks, and details the models, prompting protocol, metrics, and procedure used in our evaluation. The goal is a controlled and comparable assessment of LLM-based code translation for software modernization and cross-language interoperability.

### 3.1 Research Questions

We organize the study around three questions:

*RQ<sub>1</sub>* *Model choice*: Which state-of-the-art LLM achieves the highest translation quality across directed language pairs, and how does it compare with a traditional baseline under identical settings?

*RQ<sub>2</sub>* *Prompt factors*: How do prompt style (concise instruction versus detailed specification) and prompt language (English versus Arabic) affect translation quality across models and directions?

*RQ<sub>3</sub>* *Directionality*: How does performance vary by translation direction among C++, Java, Python, and C#, and what asymmetries emerge between source and target languages?

### 3.2 Datasets

The dataset we used for our experiment comes from the MUST (Multilingual Code Snippets Training for Program Translation) dataset [15], a publicly available benchmark designed for assessing code translation models across multiple programming languages. This dataset comprises parallel implementations (same functions in the same files for the various languages) of functions in C++, Java, Python, and C#, ensuring direct comparability between equivalent code snippets in different languages.

For this study, 24 functions (6 per language) were randomly sampled from the dataset to maintain diversity and mitigate selection bias. The way to randomly sample the dataset was to use a random number generator between 1 and 2702 to choose between the 2702 entries. Entries where the code did not include all of the four languages specified in the study were ignored. Each function was paired with its corresponding translations in the other languages, serving as ground-truth references for evaluation. For illustration, Listings 1 and 2 show a Python–Java pair that implements the same recursive check for identical linked lists; the Java version separates the recursion into a helper method to follow standard object-oriented conventions.

```

1 def areIdentical(a, b):
2
3     if (a == None and b == None):
4         return True
5
6     if (a != None and b != None):
7         return ((a.data == b.data) and
8                 areIdentical(a.next, b.next))
9
10    return False

```

**Listing 1: Python example**

```

1 boolean areIdenticalRecur(Node a, Node b)
2 {
3
4     if (a == null && b == null)
5         return true;
6
7     if (a != null && b != null)
8         return (a.data == b.data) &&
9                 areIdenticalRecur(a.next, b.next);
10
11    return false;
12 }

```

```

14 boolean areIdentical(LinkedList listb)
15 {
16     return areIdenticalRecur(this.head, listb.head);
17 }

```

**Listing 2: Java example**

### 3.3 Models and Baseline

We compare four state-of-the-art LLMs for code translation—GPT-4o (OpenAI), Gemini 2.0 (Google), DeepSeek-V3 (DeepSeek AI), and Claude 3.7 (Anthropic)—in a zero-shot setting (no fine-tuning), and we include TransCoder as a traditional baseline [15].

### 3.4 Prompt Design

We study two prompt styles and two prompt languages.

*Concise prompts.* For each directed pair we issue a short instruction. Examples: C++ → Java: “Please translate this code from C++ to Java.”; C++ → Python: “Please translate this code from C++ to Python.”; C++ → C#: “Please translate this code from C++ to C#.” Similarly for Java, Python, and C# as sources: “Please translate this code from X to Y.” We instantiate each instruction in English and in Arabic.

We batch prompts using a CSV with four input columns: *Code\_ID*, *Source\_Language*, *Target\_Language*, *Source\_Code*. The model is instructed to translate from the source to the target language under the following requirements: (1) preserve logic, algorithm, and computational flow; (2) maintain variable names where syntactically valid; (3) keep function or method structure and naming conventions adapted to the target language; (4) preserve comments and their relative positions; (5) maintain control-flow structure; (6) use equivalent data structures and types; (7) ensure identical outputs for identical inputs; (8) follow target-language naming conventions; (9) include necessary imports or includes; (10) preserve code organization and hierarchy. The model is asked to return only code unless otherwise requested. The output is written to a CSV with six columns: *Code\_ID*, *Source\_Language*, *Target\_Language*, *Source\_Code*, *Translated\_Code*, *Status*. When rules (2) and (8) conflicted, models were expected to prioritize target language conventions while preserving semantic meaning.

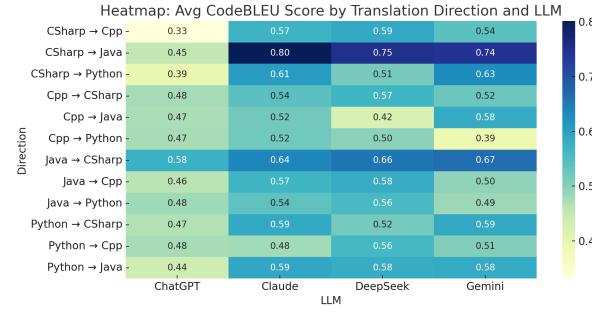
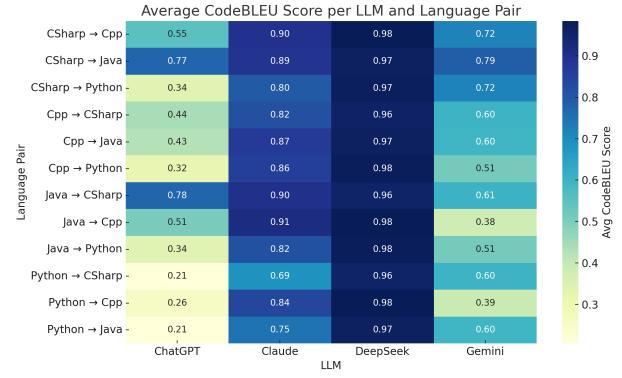
### 3.5 Evaluation Metrics

We report BLEU and CodeBLEU to capture surface correspondence and structural similarity, respectively [6]. BLEU measures n-gram overlap with references. CodeBLEU augments n-gram matching with syntax- and data-flow-aware components. We present direction-aware scores as well as macro-averages to surface asymmetries between source and target languages.

## 4 RESULTS AND DISCUSSION

### 4.1 RQ<sub>1</sub>: Model Choice

DeepSeek and Claude achieve the strongest overall translation quality across directions. DeepSeek leads on challenging pairs such as Java→C# and Python→C++ in the CodeBLEU heatmaps (Figs. 1, 2). GPT-4o is competitive on several pairs but trails the top two on stricter targets; Gemini generally ranks lower. **Table 1** (BLEU) reinforces this ordering and shows a wide gap between LLMs and

**Figure 1: CodeBLEU by model and direction using the simple prompt.****Figure 2: CodeBLEU by model and direction using the detailed prompt.**

traditional baselines across most directions. As shown in Fig. 3, the per-model BLEU and CodeBLEU under the simple prompt follow the same ordering observed in the heatmaps.

### 4.2 RQ<sub>2</sub>: Prompt Factors (Style and Language)

Prompt structure and prompt language both affect outcomes. Comparing the simple vs. detailed prompts (Figs. 1, 2), the detailed specification yields consistent gains across models and directions, amplifying DeepSeek’s lead over GPT-4o and Gemini. Prompt language also matters: English prompts outperform Arabic by about 13–15% in CodeBLEU, with similar trends in BLEU (Fig. 4).

### 4.3 RQ<sub>3</sub>: Directionality

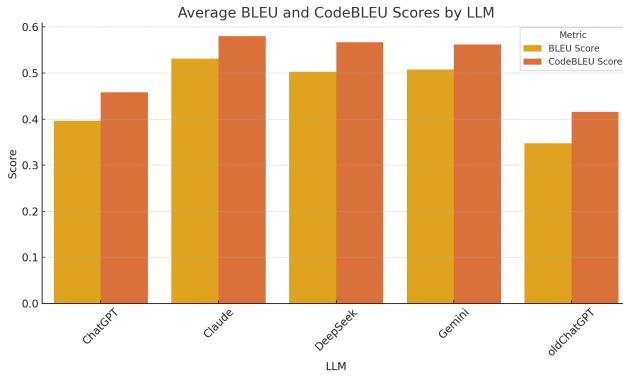
Performance is direction-dependent. Under the detailed prompt (Fig. 5), we observe asymmetries such as C#→Java > Java→C#, and Python→Java > Java→Python across most models. Directions into stricter type systems (e.g., · →C# or · →C++) are typically harder than into Python.

## 5 CONCLUSION

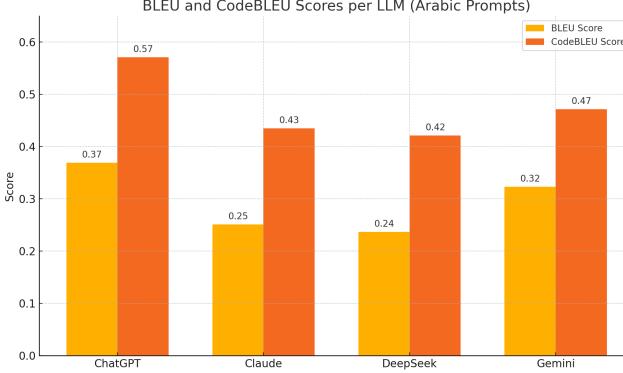
This study evaluated state-of-the-art LLMs for code translation among C++, Java, Python, and C#, using a MuST-based test set and measuring quality with BLEU and CodeBLEU under two prompt styles and two prompt languages. Three findings stand out. First, model choice matters: DeepSeek and Claude deliver the strongest

Source	LLM	Java-Py	Py-Java	Java-C++	C++-Java	Java-C#	C#-Java	Py-C++	C++-Py	Py-C#	C#-Py	C++-C#	C#-C++
Baseline	Naive Copy	34.56	34.27	66.53	66.57	77.15	77.23	36.58	36.58	35.69	35.76	67.22	67.16
Baseline	Transformer	31.22	38.15	44.38	43.93	47.34	45.6	37.42	33.9	36.91	32.64	45.32	42.65
Baseline	Transformer+MuST	40.9	43.97	58.35	54.61	73.7	71.68	42.86	39.06	43.42	42.34	57.84	57.49
Baseline	CodeBERT	38.7	41.35	65.48	53.47	85.46	82.45	43.96	38.37	46.4	41.1	63.01	67.17
Baseline	CodeBERT+MuST	55.5	57.66	81.09	78.69	90.47	86.76	58.91	55.98	59.13	55.45	79.05	81.54
Baseline	TransCoder	24.98	21.98	30.09	30.42	44.85	29.4	23.03	23.52	40.4	18.81	41.91	25.3
Baseline	TransCoder+MuST	60.73	65.53	87.09	81.64	91.74	27.7	68.7	62.92	66.52	16.88	82.4	29.44
LLM	Claude	81.68	74.81	91.39	87.22	90.17	89.17	84.13	86.00	68.82	79.66	82.04	89.64
LLM	Gemini	51.09	59.62	38.48	59.62	60.67	79.12	39.25	51.09	60.21	71.67	60.21	72.00
LLM	ChatGPT	33.75	20.74	50.59	42.54	77.60	77.39	25.59	31.83	20.61	33.71	44.25	55.40
LLM	DeepSeek	97.99	96.69	98.37	96.69	95.76	96.69	98.14	97.99	95.76	97.42	96.00	98.37

**Table 1: BLEU Scores for LLMs compared with models from the Multilingual Code Snippets Training for Program Translation study.**

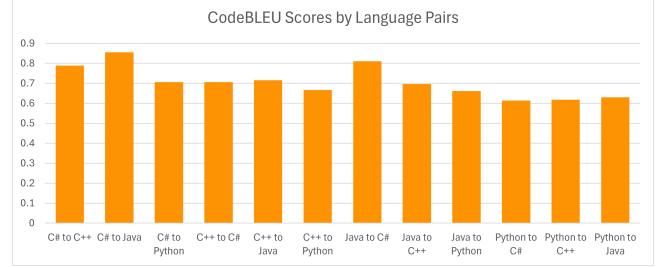


**Figure 3: BLEU and CodeBLEU per LLM under the simple prompt.**



**Figure 4: English vs. Arabic prompts (BLEU and CodeBLEU).**

overall performance, with DeepSeek leading on challenging directions such as Java→C# and Python→C++. GPT-4o is competitive on several pairs, while Gemini generally ranks lower. Second, prompt factors are influential: detailed task specifications yield reliable gains over concise instructions, and English prompts outperform Arabic by about 13–15%. Third, translation is direction-dependent: we observe stable asymmetries (for example, C#→Java exceeds Java→C#, and Python→Java exceeds Java→Python), with targets that enforce stricter type systems typically harder.



**Figure 5: Direction-wise CodeBLEU under the detailed prompt.**

These results provide practical guidance for software modernization and cross-language interoperability. Practitioners should prefer detailed specifications over minimal instructions, use English prompts when feasible, report direction-aware scores rather than only averages, and select models with attention to target-language strictness. Future work will explore scaling the benchmark, adding execution-based checks and human assessment, broadening language coverage, and developing prompt designs or tuning strategies that improve performance with non-English prompts.

## REFERENCES

- [1] Eniser, Hasan Ferit and Zhang, Hanliang and David, Cristina and Wang, Meng and Christakis, Maria and Paulsen, Brandon and Dodds, Joey and Kroening, Daniel, "Towards translating real-world code with LLMs: A study of translating to Rust," *arXiv preprint arXiv:2405.11514*, 2024.
- [2] Niephaus, Fabio and Felgentreff, Tim and Pape, Tobias and Hirschfeld, Robert and Taeumel, Marcel, "Live multi-language development and runtime environments," *arXiv preprint arXiv:1803.10200*, 2018.
- [3] Zhang, Hanliang and David, Cristina and Wang, Meng and Paulsen, Brandon and Kroening, Daniel, "Scalable, validated code translation of entire projects using large language models," *Proceedings of the ACM on Programming Languages*, vol. 9, no. PLDI, pp. 1616–1641, 2025.
- [4] Jiao, Mingsheng and Yu, Tingrui and Li, Xuan and Qiu, Guanjie and Gu, Xiaodong and Shen, Beijun, "On the evaluation of neural code translation: Taxonomy and benchmark," in *2023 38th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pp. 1529–1541, 2023.
- [5] Tao, Qingxian and Yu, Tingrui and Gu, Xiaodong and Shen, Beijun, "Unraveling the Potential of Large Language Models in Code Translation: How Far Are We?," *arXiv preprint arXiv:2410.09812*, 2024.
- [6] Ren, Shuo and Guo, Daya and Lu, Shuai and Zhou, Long and Liu, Shujie and Tang, Duyu and Sundaresan, Neel and Zhou, Ming and Blanco, Ambrosio and Ma, Shuai, "Codebleu: a method for automatic evaluation of code synthesis," *arXiv preprint arXiv:2009.10297*, 2020.
- [7] Aggarwal, Karan and Salameh, Mohammad and Hindle, Abram, "Using machine translation for converting python 2 to python 3 code," *PeerJ PrePrints*, 2015.

- [8] Ahmed, Amr and Hanneman, Greg, "Syntax-based statistical machine translation: A review," *Computational Linguistics*, 2005.
- [9] Wang, Yue and Wang, Weishi and Joty, Shafiq and Hoi, Steven CH, "Codet5: Identifier-aware unified pre-trained encoder-decoder models for code understanding and generation," *arXiv preprint arXiv:2109.00859*, 2021.
- [10] Roziere, Baptiste and Lachaux, Marie-Anne and Chanussot, Lowik and Lample, Guillaume, "Unsupervised translation of programming languages," *Advances in neural information processing systems*, vol. 33, pp. 20601–20611, 2020.
- [11] Sagodi, Zoltán and Síkér, István and Ferenc, Rudolf, "Methodology for code synthesis evaluation of LLMs presented by a case study of ChatGPT and copilot," *Ieee Access*, vol. 12, pp. 72303–72316, 2024.
- [12] Yin, Xin and Ni, Chao and Nguyen, Tien N and Wang, Shaohua and Yang, Xiaohu, "Rectifier: Code translation with corrector via llms," *arXiv preprint arXiv:2407.07472*, 2024.
- [13] Fu, Yingjie and Li, Bozhou and Li, Linyi and Zhang, Wentao and Xie, Tao, "The first prompt counts the most! an evaluation of large language models on iterative example-based code generation," *Proceedings of the ACM on Software Engineering*, vol. 2, no. ISSTA, pp. 1583–1606, 2025.
- [14] Ghosh, Akash and Datta, Debayan and Saha, Sriparna and Agarwal, Chirag, "The multilingual mind: A survey of multilingual reasoning in language models," *arXiv preprint arXiv:2502.09457*, 2025.
- [15] Zhu, Ming and Suresh, Karthik and Reddy, Chandan K, "Multilingual code snippets training for program translation," in *Proceedings of the AAAI conference on artificial intelligence*, pp. 11783–11790, 2022.