

SPECTRA: Enhancing the Code Translation Ability of Language Models by Generating Multi-Modal Specifications

Vikram Nitin¹, Rahul Krishna², Baishakhi Ray¹,

¹Columbia University ²IBM T.J. Watson Research Center

Abstract

Large language models (LLMs) are increasingly being used for the task of automated code translation, which has important real-world applications. However, most existing approaches use only the source code of a program as an input to an LLM, and do not consider the different kinds of specifications that can be extracted from a program. In this paper, we propose SPECTRA, a multi-stage approach that uses a novel self-consistency filter to first generate high-quality static specifications, test cases, and natural language descriptions from a given program, and then uses these along with the source code to improve the quality of LLM-generated translations. We evaluate SPECTRA on three code translation tasks - C to Rust, C to Go, and JavaScript to TypeScript - and show that it can enhance the performance of six popular LLMs on these tasks by up to 10 percentage points and a relative improvement of 26%. Our research suggests that generating high-quality specifications could be a promising and efficient way to improve the performance of LLMs for code translation. We make our code and data available¹, anonymized for review.

1 Introduction

Code translation involves transforming code from one programming language into functionally equivalent code in another language. This task is crucial because new programming languages (e.g., Rust, Go, TypeScript) are continuously developed to overcome the limitations of older ones (e.g., C, JavaScript). Maintaining code written in outdated languages can be costly, susceptible to security vulnerabilities, and challenging to enhance—the “technical debt” associated with maintaining legacy code costs the USA over a trillion dollars annually (Mims, 2024). Consequently, there is a pressing need to translate legacy code into modern programming languages.

¹<https://github.com/spectra822/emnlp24>

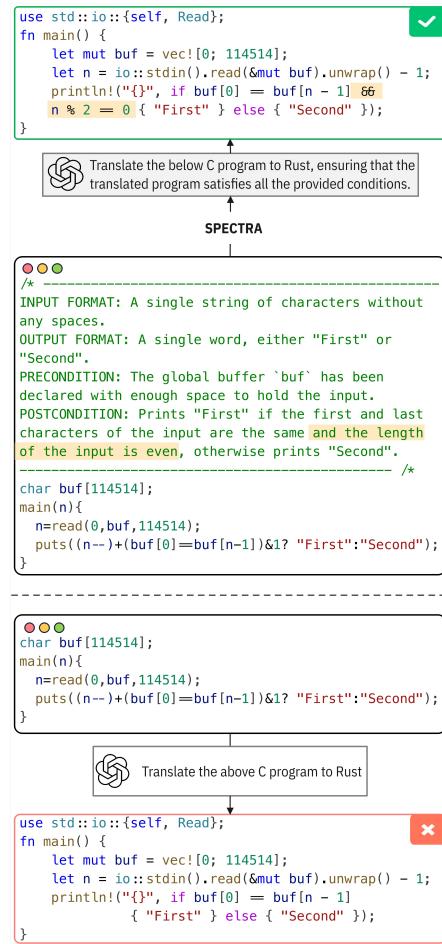


Figure 1: Correct program specifications in the prompt can help LLMs in program translations (top figure), whereas translating with no specifications provided (bottom figure) can cause erroneous translations (as shown by incorrect translation due to a missing conditional).

Traditionally, translating code between languages has been accomplished with compilers (Galois, 2018; Feldman, 1990). These are language-specific rule-based systems that use intricately designed algorithms and heuristics to produce code in the target language that is guaranteed to be functionally identical to the code in the source language. Although technically correct, this code is often non-

idiomatic, *i.e.*, very different from typical human-written code. For instance, it may have unusual control flow, use uninformative identifier names, call foreign functions from the original source language, and often contain same problems of the original legacy code. This can be hard to read and maintain, and often defeats the purpose of translating the code in the first place.

Over the past few years, Large Language Models (LLMs) have been employed to tackle a range of code-related tasks, including code generation, code summarization, bug fixing, and test case generation (Li et al., 2022; Ahmad et al., 2022; Gao et al., 2023). An expanding body of research (Roziere et al., 2020; Ahmad et al., 2022; Roziere et al., 2021; Pan et al., 2024; Yang et al., 2024) focuses on leveraging LLMs for automatic code translation between programming languages. Unlike transpilers, LLMs generate idiomatic code that is more readable, though they offer fewer guarantees of correctness by construction (see Figure 1).

In this paper, we aim to enhance the code translation capabilities of LLMs by integrating the principles of transpilers into the LLM framework. Transpilers are known for their ability to ensure functional correctness, meaning that the translated code maintains the *functional specifications* of the code written in source language without introducing semantic errors. To leverage this advantage within LLMs, we expose the functional specifications of the source code within the prompt provided to the LLM while translating. By explicitly including these specifications in the prompt, we guide the LLM to adhere to the original functional requirements, thus aiming to combine the readability and idiomatic quality of LLM-generated code with the correctness traditionally associated with transpilers. This method enhances the accuracy and utility of LLMs in code translation tasks, providing a robust solution that balances readability and correctness.

To this end, we propose SPECTRA, an approach to enhance the code translation ability of LLMs using multi-modal specifications that capture a program’s functionality. A program’s functionality can be represented in several ways. *Static* specifications express relationships between inputs and outputs that hold true for all inputs/outputs. Functionality can also be represented through a subset of *input/output* behavior (I/O specifications), which correspond to specific instances of inputs and outputs to the program (as opposed to all I/Os mentioned in the previous case). Additionally, natural language

descriptions, though often less precise and more ambiguous, provide valuable information about a program’s functionality. In the remainder of this paper, we leverage these three modalities—static, dynamic, and natural language specifications—to capture a program’s functionality.

However, inferring specifications of a program is challenging. Traditional methods like Daikon (Perkins and Ernst, 2004) have limitations: they often fail to scale across different programming languages and the specifications they produce are typically written in some domain specific languages (DSL). These DSLs are not easily understood by LLMs, especially if an LLM has not been the DSL during training. This disconnect poses a significant challenge when trying to utilize LLMs for specification generation. To address this issue, here we leverage a SOTA LLM to generate specs of different modalities of the program written in source language. This approach mirrors using an LLM as an annotator, but with a critical enhancement to ensure the quality and validity of the generated specifications. For I/O spec, we simply run the program to check its validity. For other two types of spec, after the LLM generates the specifications, we ask it to regenerate the original source code based on these specifications. If the regenerated code matches the original implementation, we take this as an indication that the generated specifications are valid and accurately capture the program’s functionality. This validation step is inspired by the self-consistency evaluation method (Min et al., 2023). By ensuring that the LLM can reproduce the original source code from the generated specifications, we increase confidence in the accuracy and usefulness of these specifications. Next, we provide each of these specifications one at a time to an LLM prompt along with the program’s source code, and generate multiple candidate translations in the target language.

We evaluate SPECTRA on three code translation tasks - converting **C** to **Rust**, **C** to **Go**, and **JavaScript** to **TypeScript**. There is a lot of interest within the software engineering community in converting between these pairs of languages (Google, 2021a,b), and thus these tasks have direct real-world applicability. We find that SPECTRA is able to enhance the performance of 6 popular LLMs on these tasks by up to **26%** compared to a uni-modal baseline.

To summarize, our contributions are as follows:

1. We propose a novel approach based on self-consistency to generate specifications of a program in the form of static specifications, test cases, and natural language descriptions.
2. We integrate these self-consistent specifications with the source code and propose SPECTRA, an approach that utilizes multi-modal specifications to improve the quality of LLM-generated code translations.
3. We evaluate SPECTRA on three program translation tasks and six popular open source and proprietary LLMs of various sizes. The findings of this paper indicate that SPECTRA improves the performance of baseline models by up to 26% (relative) and 10% (absolute).

2 Methodology

In this section, we introduce SPECTRA, a methodology for translating a program (\mathcal{S}) from a source language to a target language using a large language model (\mathbb{M}). This process involves conditioning the model on one of several specification modalities (π). The validity of the translated program is assessed by an evaluator (\mathcal{E}), which typically consists of a ground-truth test case. This test case runs both the source and target versions of the program with the same input and expects identical output.

2.1 Specification Generation

The first stage of SPECTRA focuses on the generating three types of specifications using a large language model (here, GPT4o): a) static (§2.1.1), b) input-output (§2.1.2), and c) descriptions (§2.1.3).

2.1.1 Static Specifications

Static specifications are a structured representation of the behaviour of section of the program. We use an LLM (\mathbb{M}') and a prompt composing function (λ_{stat}) to take a given program (\mathcal{S}) and generate k (here, $k = 3$) candidate static specifications (π_{stat}^i) for the given program. This can be formulated as:

$$\Pi_{\text{stat}} = \mathbb{M}' (\lambda_{\text{stat}}(\mathcal{S})) = \{\pi_{\text{stat}}^1, \pi_{\text{stat}}^2, \dots, \pi_{\text{stat}}^k\} \quad (1)$$

For full example prompts, refer to the supplementary material. The generated static specifications have the following components:

- ◊ Input Format: The initial user input required for the program to execute, *e.g., a single string of characters without any spaces.*

- ◊ Output Format: The result or data produced by the program after execution, *e.g., a single word, either "First" or "Second".*
- ◊ Pre-condition: The conditions that must be true before the function is executed. *e.g., the buffer buf must be declared and large enough to store the input string.*
- ◊ Post-condition: The conditions that must be true after the function has executed. *e.g., Prints "First" if the first and last characters of the input are the same and the length of input is even, otherwise prints "Second".*

2.1.2 Input-output Specifications

Input-output specifications represent the expected behavior (*i.e., output*) of a section of a program given a specific input. As with the static specification, k input-output specifications ($\pi_{I/O}^k$) are generated using an LLM \mathbb{M}' for the given program (see Fig. 3). We make use of another prompt composing function ($\lambda_{I/O}$) to take a program and create a text prompt containing instructions to generate test inputs (x_i) and corresponding outputs (y_i). In other words, $\pi_{I/O}^i = (x_i, y_i)$. In summary, the input-output specification generation process may be formulated as:

$$\Pi_{I/O} = \mathbb{M}' (\lambda_{I/O}(\mathcal{S})) = \{\pi_{I/O}^1, \pi_{I/O}^2, \dots, \pi_{I/O}^k\} \quad (2)$$

2.1.3 Generating Descriptions

Program descriptions are an informal and a free-form textual summary of the code to be translated. The procedure to generate these descriptions follows a similar process to that of generating static specifications (as described in §2.1.1). The distinction lies within the prompt-composing function λ_{desc} which takes the given program (\mathcal{S}) and instructs an LLM \mathbb{M}' to generate k the descriptions $\Pi_{\text{desc}} = \{\pi_{\text{desc}}^1, \pi_{\text{desc}}^2, \dots, \pi_{\text{desc}}^k\}$. This may be represented mathematically as follows:

$$\Pi_{\text{desc}} = \mathbb{M}' (\lambda_{\text{desc}}(\mathcal{S})) = \{\pi_{\text{desc}}^1, \pi_{\text{desc}}^2, \dots, \pi_{\text{desc}}^k\} \quad (3)$$

2.2 Specification validation

Specifications generated in the previous stage may be incomplete and/or incorrect. In order to best assist a language model in generating correct code translations, it would be most beneficial to provide complete and verifiable specification. Such specifications are termed as being *self-consistent*. The objective of this stage, is to retain only *self-consistent* specifications from the candidate specifications.

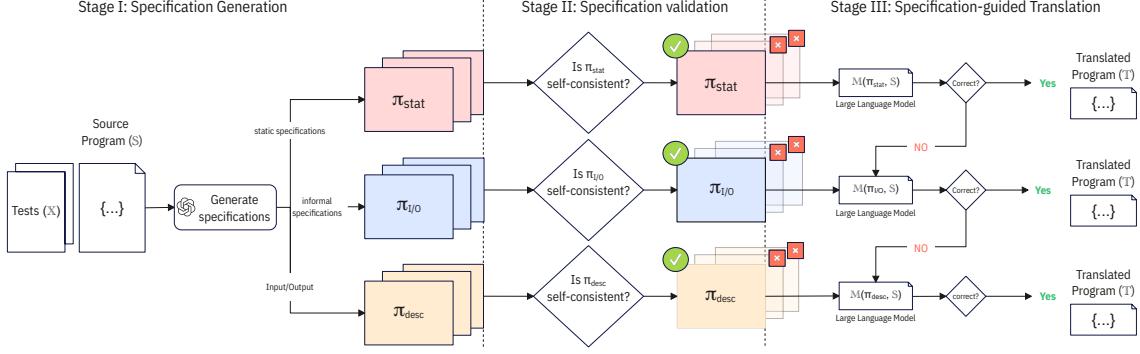


Figure 2: An overview of SPECTRA.

```

char buf[114514];
----- [Input and output specification] -----
Input: abcde
Output: False
-----
main(n) {
    n=read(0,buf,114514);
    puts((buf[0]==buf[n-1]) & 1 ? True : False + n--);
}

```

Figure 3: An input-output specification for a sample C program. For full example prompts, refer to the supplementary material.

For the set of static specifications (Π_{stat}) and descriptions (Π_{desc}), we use the same language model (\mathbb{M}') to regenerate the *original* source code (\mathcal{S}) with a prompt-composition function ($\lambda_{codegen}$). This produces a variant of the source code ($\mathcal{S}_{stat|desc}^i$). That is,

$$\mathcal{S}_{stat|desc}^i = \mathbb{M}' \left(\lambda_{codegen}(\pi_{stat|desc}^i) \right) \forall i \in \{1, \dots, k\} \quad (4)$$

We define an evaluator $eval(\cdot)$ to compare the equivalence (based on an evaluation criteria ε) of the variants of the source code ($\mathcal{S}_{stat|desc}^i$) that were generated using $\pi_{stat|desc}^i$ with original source code S . We may represent this as follows:

$$eval(\mathcal{S}_{stat|desc}^i, S, \varepsilon) = \begin{cases} \text{True} & \text{if } \mathcal{S}_{stat|desc}^i \equiv_\varepsilon S \\ \text{False} & \text{otherwise} \end{cases} \quad (5)$$

Our design allows for a variety of comparator functions to be employed. These may be heuristics that measure code similarity (such as BM25 (Robertson and Walker, 1994)), or distinct test cases that exercise the program in a specific manner and expect a pre-determined output. This work employs the latter approach in that we used existing test input and expected output pairs (denoted

by $(X, Y) = \{(x_1, y_1), (x_2, y_2), \dots, (x_n, y_n)\}$) as evaluators. Specifically, our evaluation function would assess if the regenerated code \mathcal{S}_{stat}^i behaves the same as the original source code S for all tests.

All the static specifications and descriptions for which the above evaluation fails are discarded.

$$\Pi_{stat|desc} = \{\pi_{stat|desc}^i \mid eval(\mathcal{S}_{stat|desc}^i, S, \varepsilon) = \text{True}\} \quad (6)$$

For the input-output specifications ($\Pi_{I/O}$), we exercise the input program S with each $\pi_{I/O}^i = \{x'_i, y'_i\}$ and discard ones that are inconsistent.

$$\Pi_{I/O} = \{\pi_{I/O}^i \mid S(x'_i) = y'_i\} \quad (7)$$

Here, $S(x'_i)$ denotes the output of running the program S using x'_i as input. Note, in cases where discarding inconsistent $\pi_{I/O}$ results in an empty specification set ($\Pi_{I/O} = \emptyset$), we transform the specification to keep the generated input x'_i and the *actual* program outputs $S(x'_i)$.

2.3 Specification-guided Translation

Having generated and validated the specifications π_{stat} , $\pi_{I/O}$, and π_{desc} , we now pick one specification at random for each type (say) π_{stat}^i , $\pi_{I/O}^i$, and π_{desc}^i from Π_{stat} , $\Pi_{I/O}$, and Π_{desc} respectively. Next we use these specifications sequentially one at a time along with S to generate a prompt with a prompt-composer $\lambda_{spec}(S, \pi^i)$. This prompt is then used to instruct a language model \mathbb{M} to generate translations (\mathcal{T}) such that: $\mathcal{T} = \mathbb{M}(\lambda_{spec}(S, \pi^i))$.

In this work, we use an ordered sequence of translations starting with the static specification (π_{stat}). If this approach fails, we then use the input-output specifications ($\pi_{I/O}$), and, if needed, the natural language descriptions (π_{desc}). The sequence

of specifications is determined by their formality and completeness, starting with the most structured (π_{stat}) and ending with the least formal (π_{desc}). In the rare case that all of these fail, we fall back on the “vanilla” translation with no specifications.

3 Experimental Setup

Datasets. We evaluate our approach on a subset of the CodeNet dataset (Puri et al., 2021). CodeNet contains 4053 competitive coding problems, and over 13 million code submissions in a variety of languages including C and JavaScript. These submissions are licensed for public use and redistribution, and anonymized to protect the identity of the authors. We first filter out the problems which don’t have test cases. At random, we select 300 C solutions and 300 JavaScript solutions from the remaining problems, making sure that we don’t pick more than one solution for the same problem. These 300 C and 300 JavaScript programs, each with accompanying test cases, comprise our evaluation dataset.

Models. We evaluate on six large language models: (a) 4 proprietary LLMs: gpt-4o and gpt-3.5-turbo from OpenAI, claude-3-opus from Anthropic; gemini-1.0-pro from Google; and (b) 2 open-source LLMs: Deepseek coder (33B), and Granite (34B). We accessed these models through their respective web APIs. To generate multiple candidate specifications, we use a temperature of 0.6. To generate a single candidate translation, we use greedy decoding with temperature 0, and if we need multiple candidate translations, we use a temperature of 0.3.

Evaluation. We compile each program into an executable, and run the executable with the provided test input. If the output matches the provided test output, then we mark this as correct. If the program doesn’t compile, or exits with an error code, or if the output doesn’t match, then we mark this as incorrect. We create a list of specifications in this order - $[\pi_{stat}, \pi_{I/O}, \pi_{desc}, \text{none}]$. If we were unable to generate one modality of specification, then we omit it from the list.

To evaluate translations, we categorize translations into steps. In step 1, we generate translations with just the *first* specification from the above list and measure the translation accuracy. Next, for step 2, we generate a pair of translations using the first *two* specifications from the list, and measure translation accuracy for *either* of the these candidate transla-

Lang	Total	Static	I/O	Desc
C	300	164 (54.6%)	293 (97.7%)	121 (40.9%)
JS	300	188 (62.7%)	298 (99.3%)	137 (45.7%)

Table 1: The number of source programs for which we are able to generate a successful specification.

tions. Lastly, for step 3, we generate using the first *three* specifications from the list and measure translation accuracy for *either* of the these candidate translations. In order to establish a baseline, we generate three translations *without* specifications and measure pass@1 for stage 1, pass@2 for stage 2, and pass@3 for stage 3.

4 Results

4.1 Evaluating the Generated Specifications

We would like to determine whether our specification generation process is able to generate high-quality self-consistent specifications. To do this, we check how many programs in our benchmark dataset we are able to generate self-consistent specifications for. For each program, we generate up to $k = 10$ candidate tests, and up to $k = 6$ candidate static specifications and descriptions. In all the cases, we stop as soon as we find a self-consistent specification. We use gpt-4o as our LLM for both the “forward” specification generation as well as the “reverse” code generation process.

The number of self-consistent specifications generated by gpt-4o of each category is shown in Table 1. We also measured the generated test coverage for C programs using the gcov tool. The average coverage was **91.5%**, which is an indicator that the tests are of high quality. For the generated descriptions and static specifications, the fact that we are able to recover a functioning program from most of them is a strong validation of their quality. We also manually inspected several of them and confirmed that they are a good representation of the program’s behavior.

4.2 Impact of Individual Specification Modalities

First, we evaluate the efficacy of each specification modality individually. For each of the 300 C and JavaScript programs, we attempt to perform C to Rust, C to Go, and JavaScript to TypeScript translation using gpt-4o. While generating translations, we provide each specification modality individually

Table 2: Comparing SPECTRA to the baseline for C to Rust and C to Go translations. The left half of each table shows the absolute number of correct translations out of 300 problems. The absolute number for SPECTRA is highlighted in orange. Cells highlighted in blue show higher percentage of correct translations compared to the baseline, and cells highlighted in pink show a lower percentage. We see that SPECTRA outperforms the baseline in most cases.

		C TO RUST						C TO GO					
		No. correct			improvement %			No. correct			improvement %		
		pass@1	pass@2	pass@3	pass@1	pass@2	pass@3	pass@1	pass@2	pass@3	pass@1	pass@2	pass@3
GPT4o	BASELINE	159	173	183	.	.	.	180	195	203	.	.	.
	SPECTRA	170	203	206	7%	17%	13%	182	212	217	1%	9%	7%
Claude	BASELINE	141	151	165	.	.	.	193	209	220	.	.	.
	SPECTRA	142	167	177	1%	11%	7%	184	211	218	-5%	1%	-1%
GPT3.5	BASELINE	103	120	129	.	.	.	142	151	159	.	.	.
	SPECTRA	102	134	143	-1%	12%	11%	156	175	175	10%	16%	10%
Gemini	BASELINE	43	53	57	.	.	.	51	62	68	.	.	.
	SPECTRA	53	64	68	23%	21%	19%	49	78	83	-4%	26%	22%
Granite	BASELINE	54	68	78	.	.	.	78	98	114	.	.	.
	SPECTRA	50	71	80	-7%	4%	3%	80	101	110	3%	3%	-4%
Deepseek	BASELINE	37	53	62	.	.	.	135	154	159	.	.	.
	SPECTRA	42	65	73	14%	23%	18%	133	161	165	-1%	5%	4%

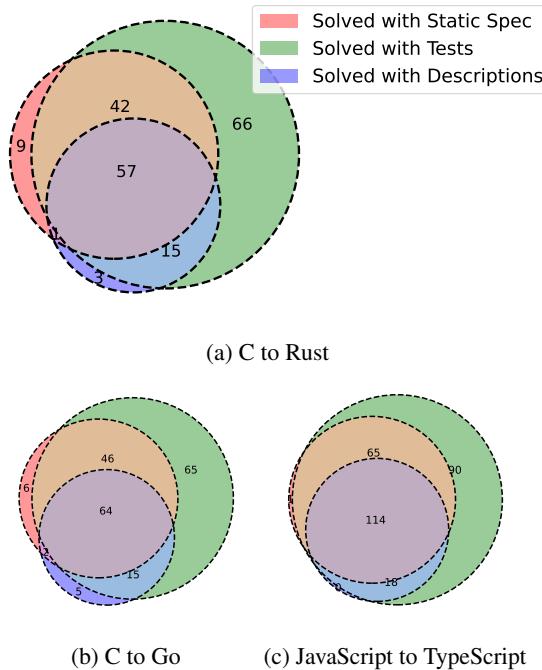


Figure 4: Our different specification modalities are complimentary. This diagram shows how many out of the 300 translations are solved using each individual specification modality. All results are with gpt-4o.

along with the program and obtain one candidate translation per modality.

Our hypothesis was that different modalities would act in a complementary fashion, whereby each modality might provide some extra information for translation that the other modalities do not provide. To investigate this, we evaluate all the translations and plot a Venn Diagram in Figure 4 to visualize the contributions of each individual specification modality. While there exists some overlap, the different modalities are indeed complementary to one another.

4.3 Impact of Multiple Specification Modalities

To evaluate the impact of multiple modalities, we measure the number of accurate translations produced SPECTRA (at various steps) against the baseline (where the prompt contains no additional specs) at various $pass@k$. Note that each pass at k corresponds to the equivalent step- k , *i.e.*, $pass@1$ corresponds to step 1, and so on. We also measure % improvement over the baseline at various steps/ $pass@k$'s. In this comparison, a positive percentage score indicates that, compared to the baseline, more source programs are accurately translated by SPECTRA. Our findings are summarized in Tables 2 and 3. The tabulated findings may be

Table 3: Comparing SPECTRA to the baseline for Javascript to Typescript. SPECTRA absolute numbers are in orange, percentage increases are in blue, percentage decreases are in pink. SPECTRA does improve performance slightly, but gains are limited.

		Pass@1	Pass@2	Pass@3	Pass@1	Pass@2	Pass@3
GPT4o	Baseline	282	286	288			
	SPECTRA	284	291	291	1%	2%	1%
Claude	Baseline	284	287	290			
	SPECTRA	279	289	290	-2%	1%	0%
GPT3.5	Baseline	268	274	276			
	SPECTRA	267	279	280	-0.3%	2%	1%
Gemini	Baseline	261	268	269			
	SPECTRA	260	275	277	-0.1%	3%	3%
Granite	Baseline	229	240	246			
	SPECTRA	194	236	245	-15%	-2%	0%
Deepseek	Baseline	273	280	281	.	.	.
	SPECTRA	263	282	285	-4%	1%	1%

summarized as below:

- *SPECTRA produces more correct translations compared to baselines:* SPECTRA generally shows improvement over the baseline. The largest relative gains are on the C to Rust translation task, and in particular, Gemini shows relative improvements of between **19%** and **23%**. The best-performing base models are GPT4o and Claude, but they too are able to benefit from using SPECTRA.
- *The largest relative increases generally occur for models that have the lowest baseline performances:* For example, Deepseek and Gemini are the two worst models at the C to Rust task, and their “improvement %” of up to 23% are the highest of all models on this task. Conversely, Deepseek’s baseline performance is much better on C to Go, and its performance gain with SpecTra is low. However, Granite is an exception to this trend. Its baseline performance is poor, but it is unable to effectively use our specifications for any of the tasks.
- *SPECTRA also helps improve Javascript to TypeScript translation albeit with considerably less pronounced gains compared to the other two translation tasks:* Translation from Javascript to TypeScript (Table 3) indicate an similar trend the other two translation pairs in that SPECTRA generally improves results over the baseline, with notable percentage improvements for GPT4o and Gemini. However, overall improvements are far less pronounced compared to other translation tasks. We conjecture that this is due to the high degree of

Table 4: Assessing the improvements of providing multiple specification modalities. The columns compare the relative % improvements in correct translation by using only *one* kind of specification to using either of *two* kinds of specifications (step 2 vs. step 1), as well as to using either of *three* as opposed to just *one* (step 3 vs. step 1). Cells highlighted in blue show higher improvements compared to the baseline.

		step2 vs. step1	Step3 vs. step1	step2 vs. step1	step3 vs. step1	step2 vs. step1	step3 vs. step1
		C to Rust	C to Go	JS to TS			
GPT4o	BASELINE	9	15	8	13	1	2
	SPECTRA	19	21	16	19	2	2
Claude	BASELINE	7	17	8	14	1	2
	SPECTRA	18	25	15	18	4	4
GPT3.5	BASELINE	17	25	6	12	2	3
	SPECTRA	31	40	12	12	4	5
Gemini	BASELINE	23	33	22	33	3	3
	SPECTRA	21	28	59	69	6	7
granite	BASELINE	26	44	26	46	5	7
	SPECTRA	42	60	26	38	22	26
Deepseek	BASELINE	43	68	14	18	3	3
	SPECTRA	55	74	21	24	7	8

semantic and syntactic similarity between the two languages. JavaScript and TypeScript share many features and structures, and the translation process is inherently simpler and less reliant on detailed specifications to achieve correctness. This similarity reduces the need for extensive specification guidance, as the language model can more easily infer the correct translations based on existing structural and semantic parallels. Further, providing additional specifications in fact can be detrimental to some smaller models like Granite and Deepseek. For instance, the step-1 accuracy for Granite on JavaScript to TypeScript translation is 15% worse than the baseline model without specifications.

Lastly, in order to assess our hypothesis that providing additional modalities of specification when the static specification is insufficient for correct translation can benefit translation, we compare the following in Table 4:

- The percentage (%) additional correct translation for SPECTRA by using either 2 or 3 different specification modalities individually (*i.e.*, step 2 and step 3 in Table 4) compared to only one single specification modality (step-1).
- The percentage (%) additional correct translation for baseline model at Pass@3 and Pass@2 versus

the same baseline model at Pass@1. We note that the baseline uses no specifications for translation.

The findings summarized in Table 4 demonstrate that incorporating additional specification modalities when static specifications are inadequate can significantly enhance translation accuracy. In all the cases, additional steps increase the number of correct translations (as evidenced by positive and higher percentages shown in blue). We also note that Step 3 vs. Step 1 is greater than Step 2 vs. Step 1, indicating that even informal descriptions can help increase the number of correct translations. Lastly, it is worth noting that although there are marginal (albeit positive) improvements for Javascript to Typescript, models with a lower baseline accuracy like Granite see considerable improvements with providing additional specification modalities - a 22% improvement when using up to two kinds of specifications and 26% when using up to 3 kinds of specifications.

5 Related Work

5.1 Code Translation with LLMs

Roziere et al. (2020) proposed TransCoder, an early attempt to tackle code translation using transformer models. They proposed Back Translation, a fully unsupervised approach. (Ahmad et al., 2022) refined Back Translation with the addition of a code summarization task. In contrast to unsupervised learning, supervised learning requires aligned parallel data across pairs of programming languages, which is challenging to obtain at scale. (Roziere et al., 2021) address this challenge by generating synthetic aligned data using another language model and filtering out incorrect translations using generated unit tests. The last few years have seen the development of large foundation models for code and natural language, that are pretrained on massive amounts of data with an autoregressive objective function. These can perform a variety of code-related tasks, including translation. (Pan et al., 2024) and (Yang et al., 2024) perform thorough empirical evaluations of the translation ability of these models.

5.2 Generating Program Specifications with LLMs

The three kinds of program specifications that we consider in this paper are test cases, natural language descriptions, and static specifications. There have been a few recent papers that use LLMs to

generate test cases for competitive coding problems. EvalPlus (Liu et al., 2023) uses ChatGPT to generate test cases, and mutates them in order to rigorously test LLM-generated code solutions. CodeT (Chen et al., 2022) uses LLM-generated test cases to filter out code solutions from a list of LLM-generated candidate solutions. Recently, Yang et al. (2024) proposed to generate test cases, provide them along with the code to an LLM for translation, and use test feedback to iteratively repair the generated translation.

Separately, there has been a long line of work on using transformer models for generating natural language descriptions of code (Ahmad et al., 2022; Gao et al., 2023; Gong et al., 2022; Gao and Lyu, 2022; Wu et al., 2021; Tang et al., 2021). There has been comparatively less research into generating static specifications. Endres et al. (2023) transform natural language descriptions of functions into assert statements and use these assertions to detect incorrect code.

Parsel (Zelikman et al., 2023) is a recent approach that generates and utilizes multiple kinds of specifications. Starting with a problem description, Parsel first generates a program sketch consisting of function-level descriptions and test cases, and then generates code to complete the sketch.

6 Conclusion

In this paper, we have presented SPECTRA, an approach to 1) generate self-consistent multi-modal specifications, and 2) use these specifications to improve the program translation performance of large language models. We show relative performance gains of up to 26% across multiple models and languages pairs. Our results indicate the possibility for further research into generating high-quality program specifications as a cheap and efficient way to improve the performance of real-world LLMs on code translation. Future work could explore generating specifications in formal language or assert statements, which can then be automatically cross-verified against the source code. Another promising direction is to use these specifications for downstream tasks other than translation.

7 Ethical Impact and Potential Risks

The inherent risk associated with using LLMs is that the code they produce is not guaranteed to be correct. Even though the generated code may pass some given test cases, there may be other corner

cases and unexpected behavior that are not exposed. If code translation with LLMs starts being used at a large scale without adequate testing, it could create an ecosystem of untested, unproven, potentially insecure code. Although our paper generates different kinds of specifications, it does not cross-verify the final translations against these specifications. We encourage the development of specification generation and formal verification methods for LLM-generated programs as a means to counter potential security vulnerabilities.

8 Limitations

One limitation of our approach for generating static specifications is that these specifications are in natural language and cannot be precisely validated against the source program. A self-consistent static specification, while *likely* to be correct, is not *guaranteed* to be correct.

Our paper deals only with code translation in a competitive coding setting, where we have relatively simple standalone files which take text input from STDIN and write text to STDOUT. However, real-world legacy code is far more complex, often involving multiple files with hundreds of lines of code each, and complex inter-dependencies. Testing such programs is also non-trivial, and may involve, say, simulating a server and querying it with HTTP requests, or setting up a mock database to simulate reads and writes. Performing code translation in this kind of setting is an important challenge and a natural direction to extend our work.

References

- Wasi Uddin Ahmad, Saikat Chakraborty, Baishakhi Ray, and Kai-Wei Chang. 2022. Summarize and generate to back-translate: Unsupervised translation of programming languages. *arXiv preprint arXiv:2205.11116*.
- Bei Chen, Fengji Zhang, Anh Nguyen, Daoguang Zan, Zeqi Lin, Jian-Guang Lou, and Weizhu Chen. 2022. Codet: Code generation with generated tests. *arXiv preprint arXiv:2207.10397*.
- Madeline Endres, Sarah Fakhoury, Saikat Chakraborty, and Shuvendu K Lahiri. 2023. Formalizing natural language intent into program specifications via large language models. *arXiv preprint arXiv:2310.01831*.
- Stuart I Feldman. 1990. A fortran to c converter. In *ACM SIGPLAN Fortran Forum*, volume 9, pages 21–22. ACM New York, NY, USA.
- Galois. 2018. C2Rust.
- Shuzheng Gao, Cuiyun Gao, Yulan He, Jichuan Zeng, Lunyu Nie, Xin Xia, and Michael Lyu. 2023. Code structure-guided transformer for source code summarization. *ACM Transactions on Software Engineering and Methodology*, 32(1):1–32.
- Yuxiu Gao and Chen Lyu. 2022. M2ts: Multi-scale multi-modal approach based on transformer for source code summarization. In *Proceedings of the 30th IEEE/ACM International Conference on Program Comprehension*, pages 24–35.
- Zi Gong, Cuiyun Gao, Yasheng Wang, Wenchao Gu, Yun Peng, and Zenglin Xu. 2022. Source code summarization with structural relative position guided transformer. In *2022 IEEE International Conference on Software Analysis, Evolution and Reengineering (SANER)*, pages 13–24. IEEE.
- Google. 2021a. [Rust in the Android platform](#).
- Google. 2021b. [Rust in the Linux kernel](#).
- Yujia Li, David Choi, Junyoung Chung, Nate Kushman, Julian Schrittwieser, Rémi Leblond, Tom Eccles, James Keeling, Felix Gimeno, Agustin Dal Lago, et al. 2022. Competition-level code generation with alphacode. *Science*, 378(6624):1092–1097.
- Jiawei Liu, Chunqiu Steven Xia, Yuyao Wang, and Lingming Zhang. 2023. [Is your code generated by chatgpt really correct? rigorous evaluation of large language models for code generation](#). In *Advances in Neural Information Processing Systems*, volume 36, pages 21558–21572. Curran Associates, Inc.
- Christopher Mims. 2024. [The invisible \\$1.52 trillion problem: Clunky old software](#). *The Wall Street Journal*.
- Marcus J Min, Yangruibo Ding, Luca Buratti, Saurabh Pujar, Gail Kaiser, Suman Jana, and Baishakhi Ray. 2023. Beyond accuracy: Evaluating self-consistency of code large language models with identitychain. *arXiv preprint arXiv:2310.14053*.
- Rangeet Pan, Ali Reza Ibrahimzada, Rahul Krishna, Divya Sankar, Lambert Pouguem Wassi, Michele Merler, Boris Sobolev, Raju Pavuluri, Saurabh Sinha, and Reyhaneh Jabbarvand. 2024. Lost in translation: A study of bugs introduced by large language models while translating code. In *Proceedings of the IEEE/ACM 46th International Conference on Software Engineering*, pages 1–13.
- Jeff H Perkins and Michael D Ernst. 2004. Efficient incremental algorithms for dynamic detection of likely invariants. In *proceedings of the 12th ACM SIGSOFT twelfth International Symposium on Foundations of Software Engineering*, pages 23–32.
- Ruchir Puri, David S Kung, Geert Janssen, Wei Zhang, Giacomo Domeniconi, Vladimir Zolotov, Julian Dolby, Jie Chen, Mihir Choudhury, Lindsey Decker, et al. 2021. Codenet: A large-scale ai for code dataset for learning a diversity of coding tasks. *arXiv preprint arXiv:2105.12655*.

- Stephen E Robertson and Steve Walker. 1994. Some simple effective approximations to the 2-poisson model for probabilistic weighted retrieval. In *SIGIR'94: Proceedings of the Seventeenth Annual International ACM-SIGIR Conference on Research and Development in Information Retrieval, organised by Dublin City University*, pages 232–241. Springer.
- Baptiste Roziere, Marie-Anne Lachaux, Lowik Chanusot, and Guillaume Lample. 2020. Unsupervised translation of programming languages. *Advances in neural information processing systems*, 33:20601–20611.
- Baptiste Roziere, Jie M Zhang, Francois Charton, Mark Harman, Gabriel Synnaeve, and Guillaume Lample. 2021. Leveraging automated unit tests for unsupervised code translation. *arXiv preprint arXiv:2110.06773*.
- Ze Tang, Chuanyi Li, Jidong Ge, Xiaoyu Shen, Zheling Zhu, and Bin Luo. 2021. Ast-transformer: Encoding abstract syntax trees efficiently for code summarization. In *2021 36th IEEE/ACM International Conference on Automated Software Engineering (ASE)*, pages 1193–1195. IEEE.
- Hongqiu Wu, Hai Zhao, and Min Zhang. 2021. **Code summarization with structure-induced transformer**. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1078–1090, Online. Association for Computational Linguistics.
- Zhen Yang, Fang Liu, Zhongxing Yu, Jacky Wai Keung, Jia Li, Shuo Liu, Yifan Hong, Xiaoxue Ma, Zhi Jin, and Ge Li. 2024. Exploring and unleashing the power of large language models in automated code translation. *arXiv preprint arXiv:2404.14646*.
- Eric Zelikman, Qian Huang, Gabriel Poesia, Noah Goodman, and Nick Haber. 2023. Parsel: Algorithmic reasoning with language models by composing decompositions. *Advances in Neural Information Processing Systems*, 36:31466–31523.

A Example Prompts

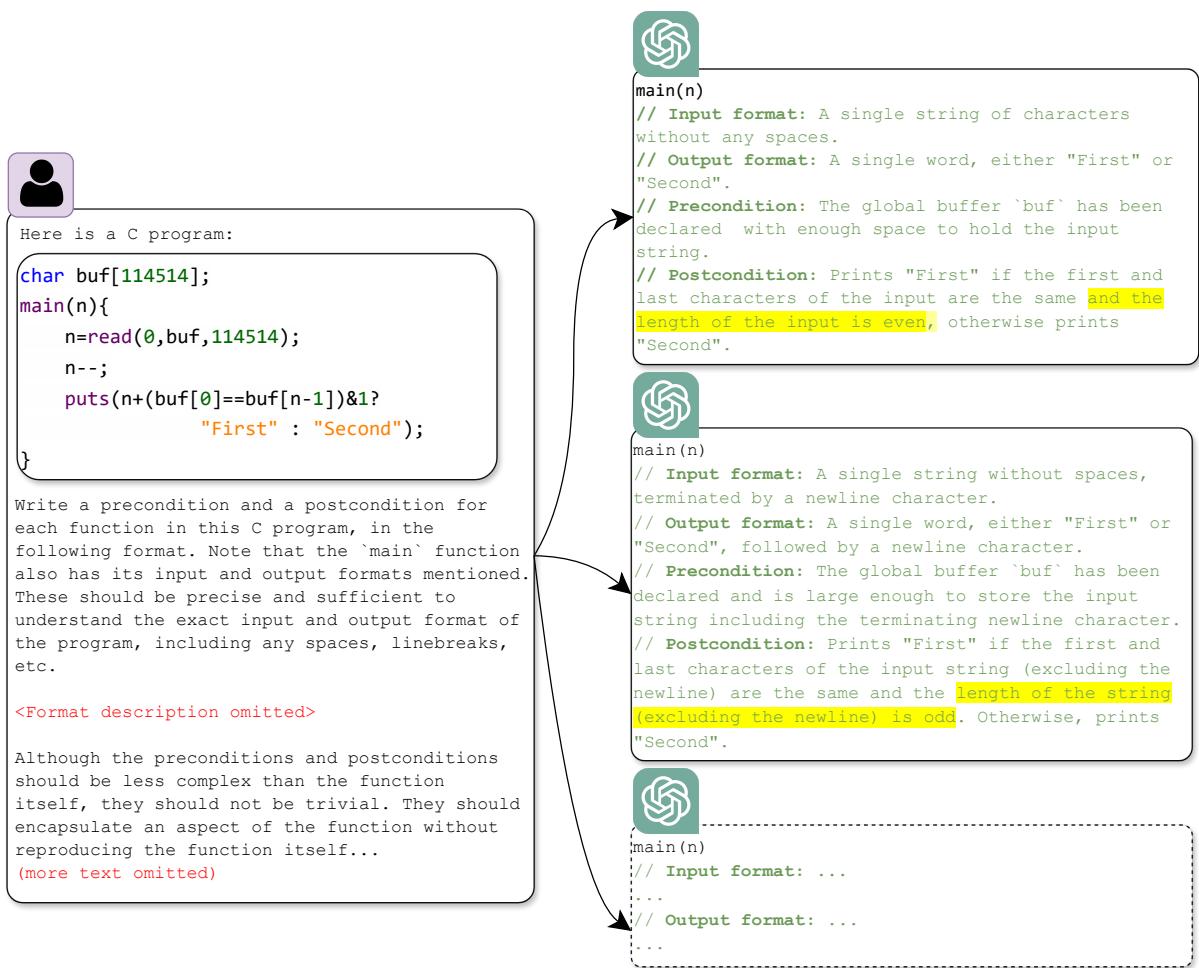


Figure 5: The first step of our approach involves generating multiple candidate specifications from a given program. This is the actual output of gpt-4o on our motivating example from Figure 1.

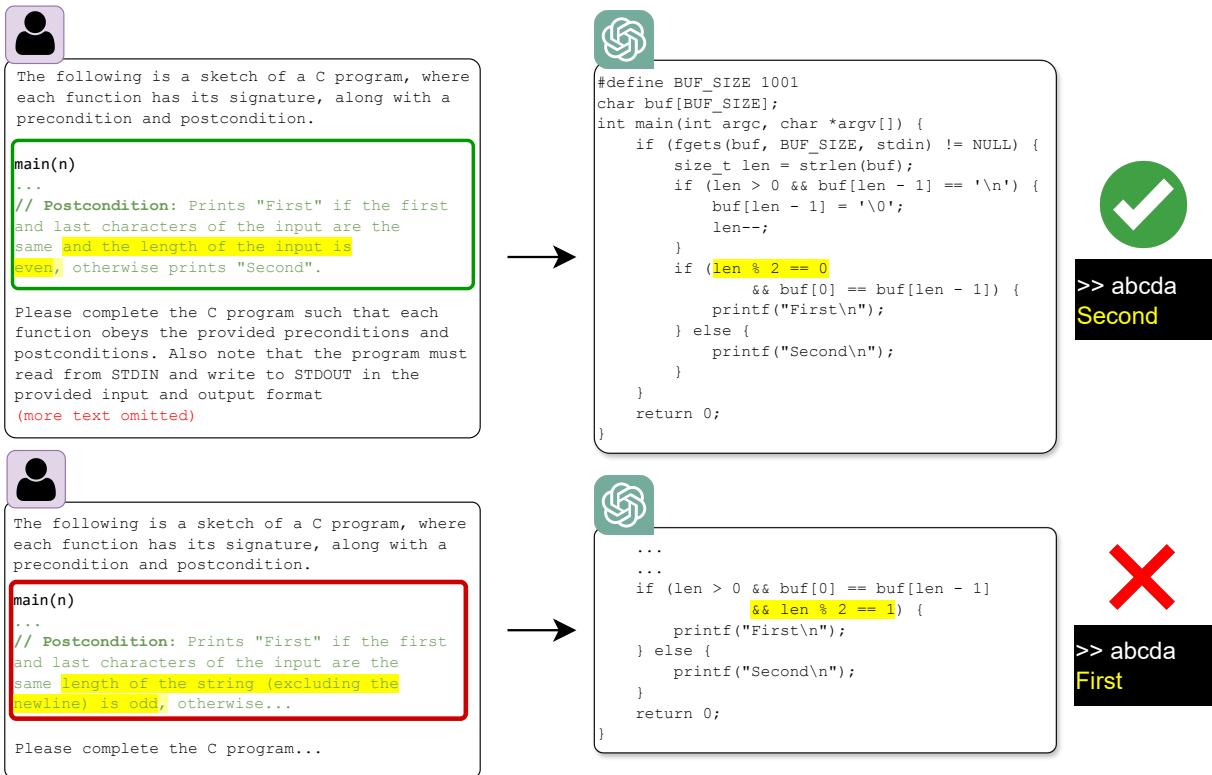


Figure 6: Filtering out incorrect specifications. We re-generate C code using each specification, and pick the specifications corresponding to re-generated C programs that pass the test case.

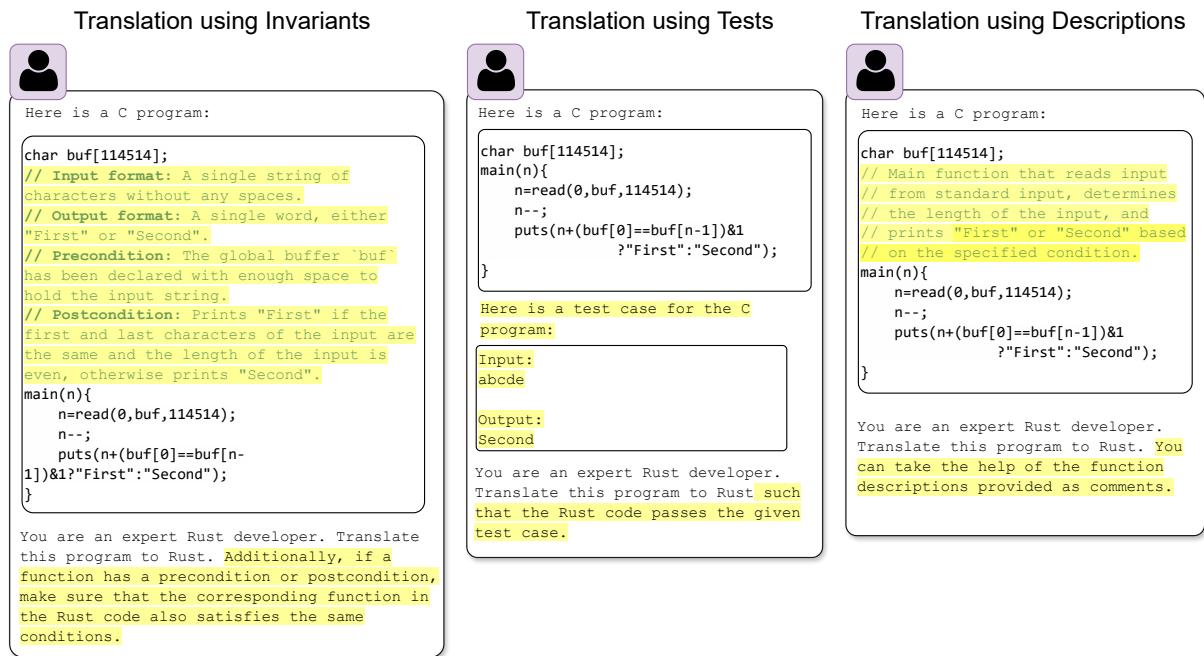


Figure 7: Using different modalities of specifications to perform specification-augmented translation. These are the actual prompts and specifications generated by SPECTRA.