MASSIVE-Agents: A Benchmark for Multilingual Function-Calling in 52 Languages

Mayank Kulkarni **Vittorio Mazzia** **Judith Gaspers**

Christopher Hench **Jack FitzGerald**

Amazon AGI*

Correspondence: {maykul,vmazzia,gaspers,henchc,jgmf}@amazon.com

Abstract

We present MASSIVE-Agents, a new benchmark for assessing multilingual function calling across 52 languages. We created MASSIVE-Agents by cleaning the original MASSIVE dataset and then reformatting it for evaluation within the Berkeley Function-Calling Leaderboard (BFCL) framework. The full benchmark comprises 47,020 samples with an average of 904 samples per language, covering 55 different functions and 286 arguments. We benchmarked 21 models using Amazon Bedrock and present the results along with associated analyses. MASSIVE-Agents is challenging, with the top model, Nova Premier, achieving an average Abstract Syntax Tree (AST) Accuracy of 34.05% across all languages, with performance varying significantly from 57.37% for English to as low as 6.81% for Amharic. Some models, particularly smaller ones, yielded a score of zero for the more difficult languages. Additionally, we provide results from ablations using a custom 1-shot prompt, ablations with prompts translated into different languages, and comparisons based on model latency.

1 Introduction

Large Language Models (LLMs) and Foundation Models (FMs) are now commonly used as agents, where the term *agent* is used to describe a software module that has access to real-world tools and can determine if using a tool is necessary to achieve the user's goal. The agent must choose from a list of available tools, generate the correct syntax to successfully invoke the tool, and handle the tool's response (Xi et al., 2023; Yao et al., 2023; Schick et al., 2023; Patil et al., 2023). The agent also often then formulates human-friendly responses.

Existing benchmarks for agentic performance are often limited to evaluating agents with queries only in English, such as the Berkeley Function Calling Leaderboard (BFCL) (Yan et al., 2024), API-Bank (Li et al., 2023), Nexus Function Calling (team, 2023), and τ-bench (Yao et al., 2024). To help address this gap, we adapted a portion of the MASSIVE dataset (FitzGerald et al., 2023) into a new benchmark we call MASSIVE-AGENTS. Our contributions include the following:

1. We reformatted the MASSIVE dataset for tool use using the BFCL framework,
2. We performed substantial automatic and manual data filtering on the original data to ensure the highest quality of the final data, reducing the original test set from 152k to 47k utterances,
3. We conducted benchmarking and analysis of 21 models across major performance tiers, and
4. We release the new benchmark data and code publicly[1].

2 Method

Building upon the MASSIVE (FitzGerald et al., 2023; Bastianelli et al., 2020) benchmark published in 2022, we created the MASSIVE-AGENTS multilingual natural language understanding dataset designed for evaluating FMs agentic capabilities. In the following, we first provide more details on our dataset specification and the original MASSIVE dataset. Subsequently, we describe the filtering and transformations we conducted on the MASSIVE dataset to create MASSIVE-AGENTS .

2.1 The MASSIVE source dataset

The original MASSIVE dataset (FitzGerald et al., 2023) is a massively multilingual benchmark that comprises parallel data across 52 languages and was designed for evaluating Spoken Language Understanding (SLU) models. Consequently, the dataset comprises user utterances directed toward

*All authors were associated with Amazon at the time of publication.

[1] https://github.com/alexa/massive

voice-controlled devices, which are human annotated with intent and slot labels, spanning 60 intents, 55 slot types, and 18 domains. Moreover, meta information, including quality scores, e.g., regarding grammar and spelling, is available. We selected MASSIVE as our source dataset as it i) covers many languages with parallel data, as this enables easy analysis across languages, and ii) the format/SLU labels are well suited for conversion into the desired agent format, including function calling tests.

2.2 Dataset specification

We aim to create a multilingual dataset designed for evaluating FMs agentic capabilities, as prior research has demonstrated there are unique challenges when dealing with low-resource languages in addition to typologically diverse languages (Simpson et al., 2008; Strassel and Tracey, 2016; Lakew et al., 2020; Marivate et al., 2020; Maguerresse et al., 2020; Goyal et al., 2021). To align with a widely used benchmark for measuring agentic performance, i.e., the Berkeley Function Calling Leaderboard (BFCL), we represent the data accordingly via function calling tests, i.e., pythonic function calls. We manually specify the set of supported functions based on the slot and intent matrix used for the original MASSIVE dataset, defining which arguments are required, optional, and if any of the arguments have default values. In the latter case, we also specify default behaviors for a set of characteristics.

We assume that user utterances are directed toward a voice interface; thus, the structure of the language used may be more terse and task-directed. We furthermore assume that the corresponding device has multimodal capabilities, including apps that can be launched, such as email, social media, music, and media (Spotify, YouTube, . . .) in addition to native capabilities, such as asking about the weather (forecast). Based on this, we define default experiences for each function, such as the email app opening up with pre-filled content about the sender and recipient or the device reading out calendar events for the specified day.

Note that function and user behavior specifications are needed to determine which utterances from the original MASSIVE dataset should be transformed for MASSIVE-AGENTS .

2.3 Data Filtering

While the MASSIVE data format generally allows for conversion into the desired format that can be used by agents, creating a clean and high-quality dataset comes with many challenges and requires extensive filtering efforts.

2.3.1 Manual Filtering

We choose English as our familiar anchor language for data filtering, as the MASSIVE dataset is a parallel corpus, and we propagate filters across lesser-known languages. We manually filtered out 933 utterances from the original MASSIVE dataset that did not meet our quality standards, resulting in 1,978 carefully curated utterances as our representative dataset. The resulting data points were reviewed by another annotator and further validated by independent annotations on 20% of the data and demonstrating high inter-annotator agreement. We filter out utterances containing Missing or incomplete annotations, Incorrect annotations, Ambiguous utterances, Unsupported behaviors, Multi-Turn utterances, Multi-Function utterances, Low frequency mismatched arguments, Duplicated arguments and Open ended/catch all functions. We provide a comprehensive overview of these filters in App A.1.

2.3.2 Automated Filtering

We conduct programmatic-based filtering [2] for every language based on certain thresholds. The criteria on which the threshold is defined are a combination of the inter-annotator agreement across intent slots and an automated mechanism to judge the grammar and spelling. All of which are present in the metadata of the annotations of the original MASSIVE dataset. As a fail-safe, we further drop any utterances that contain duplicated arguments that were not flagged in the manual filtering round.

In order to maintain the parallel nature of the dataset to the degree that it is possible, we take an intersection of this high-quality automated, filtered data with the identifiers of the manually filtered high-quality annotation curated for English. This results in a 47,020 sample dataset across all languages with an average of 900 samples per language.

2.4 Transformation to functions

We convert intents to function names by replacing the "_" with a "." to effectively define a

[2] https://github.com/alexa/massive

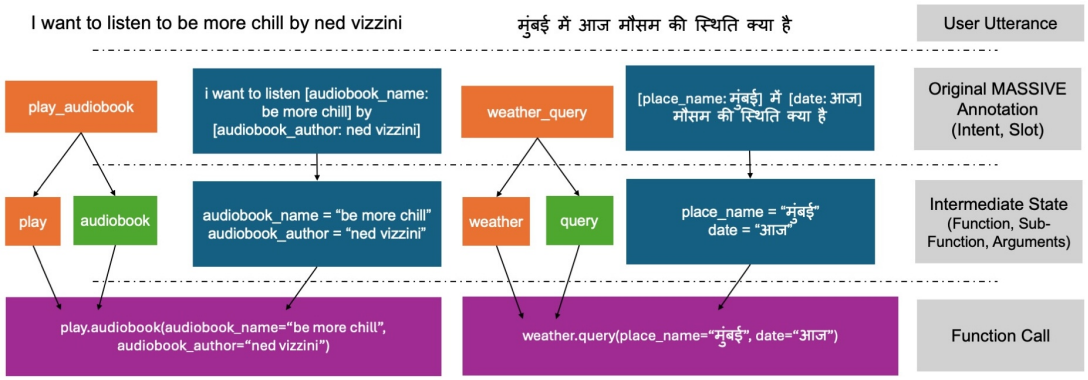

Figure 1: Transformation of two prompts in English (left) and Hindi (right), from the original NLU-style Intent-Slot annotation in the original MASSIVE dataset, to building intermediate states of a parent function and sub-function and capturing the slot name and values as arguments. These are finally combined as the function call in each language. While we run experiments with descriptions of the functions in target languages, we also maintain English function and argument names for simplicity.

Namespace and a corresponding function. With this approach, for instance, $email\_addcontact$, and $email\_sendemail$ all are grouped under the $email$ Namespace and are different function calls as $email.addcontact()$ and $email.sendemail()$. Further, we parse out slot names and slot values, and these constitute the argument names and argument values for the functions. We provide an illustration of this transformation in Fig. 1.

Given that the utterances are all transcriptions from a voice-based interface, we do not post-process any date timestamps and expect the model to be able to copy appropriate argument values from the utterance verbatim as directed in the prompt instructions for evaluation.

We also make the following considerations while transforming the data to functions:

2.4.1 QA Style Questions

Through analysis of the annotated utterances, we find that there exists a lot of ambiguity in the slot annotations for question-answering based intents. This is primarily due to the limitations of expressiveness with the slot tagging approach; this is exemplified in the $qa\_factoid$ intent, "how large is [$place\_name$: alaska]". To mitigate issues, we use the entire utterance as a required $query$ argument to the function and discard the actual slot annotations. Thus resulting in the transformation to $qa.factoid(query$="how large is alaska"). We transform $qa.factoid()$, $qa.maths()$, $qa.currency()$ and $datetime.convert()$ functions using this methodology. These functions also test the model's ability to follow instructions to leverage the grounded 'QA' functions in the prompt instead of attempting to answer parametrically.

Function call with no arguments We define the behavior of the function when no arguments are provided or required. For example, $recommendation.events()$ would return a list of events within a 20 mile radius of the user's location or $email.query()$ would return all unread emails over the past two days. This tests the model's ability to understand what the user is trying to achieve when the utterance is underspecified, such as "check emails".

2.5 Function and Argument Annotations

MASSIVE-AGENTS incorporates 55 different functions and 200 arguments with 53 distinct argument names. We human annotate descriptions for each of the functions with the description of the function itself, the default behavior without arguments, and the default user experience. The complete function description is generated by combining the three descriptions in order, respectively, with prefixes in between.

For example, the $alarm.query()$ description reads as "Query date, time, or event name for any or all alarms currently set. If no arguments are provided, then it returns the date(s), time, and event name for all alarms currently set. The default user experience is that it plays and shows the date(s), times, and event names of all alarms currently set".

Further, for each of the 200 arguments corresponding to their function, we have unique descriptions and also human annotations for the default argument values.

2.6 Sampling

We also create a condensed version of the dataset, i.e., 10k samples, to facilitate function testing and development. We do so by stratifying across functions and within functions across arguments; specifically, we create a 3-level graph of locales, functions, and arguments. After determining the number of samples needed per locale, we iterate through the locales. For each locale, we cycle through a randomized list of functions and randomly choose an argument tuple before cycling to the next function. Once the utterances from an argument set are exhausted, that entry is removed, and this holds true for functions as well. We aim to get a relatively even distribution across locales, functions, and sets of arguments, and we empirically demonstrate that the 10k version results track those of the full dataset and are thus representative.

2.7 Dataset Statistics

The Full and 10k datasets contain 47,020 and 9,741 samples, respectively, with 286 and 279 unique function and argument name combinations along with an average of 904 samples per language and 187 samples per language. We also see that due to filtering and quality control, some languages, such as Khmer (km-KH) have a much smaller coverage of only 53 samples; however, most other languages are fairly well represented in the dataset. App. A.2 contains a complete analysis of the statistics.

3 Benchmarking

In this section, we present a comprehensive evaluation of various language models performance on the MASSIVE-AGENTS benchmark. Specifically, we assess FMs ability to generate accurate function calls across multiple languages, test splits, and different prompting strategies. Limiting our analysis to the family of non-thinking models available in Amazon Bedrock at the time of writing, our evaluation encompasses open-source models ranging from 1B to 405B parameters (Dubey et al., 2024), and proprietary models, including releases such as Claude 3.5 v2 Sonnet (Anthropic, 2024), and Nova models (Intelligence, 2024a,b). We examine both zero-shot and one-shot scenarios across all languages, using either an initial English-only prompt or prompts translated into different target languages.

[3]Output speed obtained from https://artificialanalysis.ai/ on 19 December 2024.

In Sec. 3.1, we detail our experimental setup, including our prompting strategies and evaluation metrics. Sec. 3.2 presents our main results, analyzing performance across different languages and models. Finally, in Sec. 3.3, we conduct ablation studies to understand the impact of different prompting strategies and their interaction with various languages.

3.1 Experimental Setup

Tab. 1 provides a comprehensive overview of all evaluated models, including their parameter counts, maximum context lengths, release dates, and output speed (tokens/sec) as an example of performance reference.

For our main experiments, we employ two distinct prompting strategies. The first uses the BFCL (Yan et al., 2024) zero-shot template, while the second is a custom one-shot template with MASSIVE-AGENTS -specific instructions to enhance performance. Both templates are challenging and computationally demanding due to the context length, typically exceeding 12,000 tokens. However, while BFCL positions output syntax instructions before the set of actions, our custom template places instructions after the actions, which empirically should lower the number of syntax errors. Additionally, our custom template includes specific instructions encouraging models to preserve argument values exactly as they appear in the query, as well as to only make the function call without comments or clarifying questions. Instructions and function descriptions are in English for both templates. The complete prompt templates are provided in App. A.4.

To ensure reproducibility across all experiments, we utilize greedy decoding for all models. We conduct our evaluation on two dataset splits: the 10k split is tested with both prompting strategies, while the full MASSIVE-AGENTS dataset is evaluated on a subset of models using only the BFCL zero-shot template due to computational requirements. The full split evaluation serves as a way to validate the correlation between the two splits, helping to confirm that the 10k split is a reliable proxy for model performance assessment.

We employ two primary metrics for evaluation:

- **Function Selection Accuracy (FSA):** This metric evaluates the proportion of queries for which the model correctly identifies the intended function. It focuses solely on select-

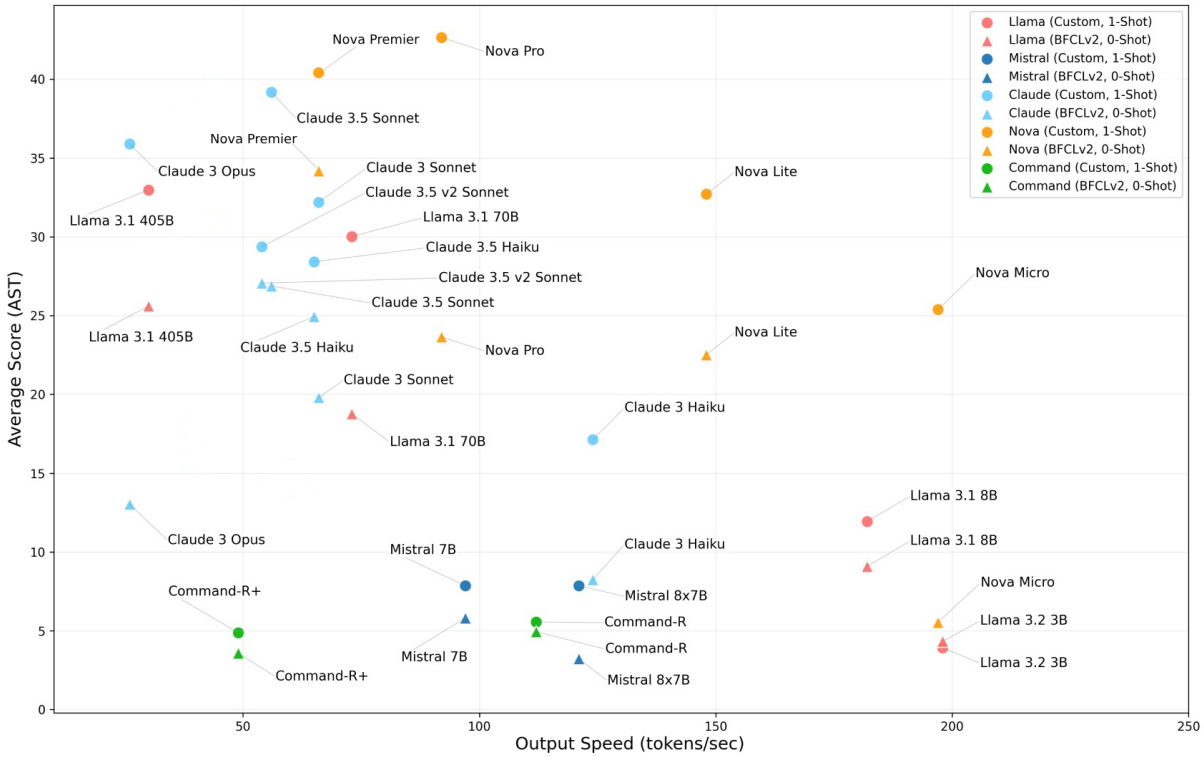

Figure 2: Comparison of different benchmarked models is plotted by output speed (tokens/sec)[3] against the average AST accuracy. Triangular markers represent results for the BFCL zero-shot prompt, while circular markers represent those for the custom one-shot prompt. All results were obtained using the 10k split.

ing the correct function, without considering the accuracy of arguments. The aggregated results report the average across all tested languages. A higher FSA score indicates better alignment between the user's query and the model's function selection.

- **Abstract Syntax Tree Accuracy (AST):** AST measures the overall correctness of the generated function call, including both the function selection and the parsing of arguments into the correct structure. It evaluates whether the entire abstract syntax tree of the generated response matches the ground truth perfectly. A perfect match requires that both the function name and all its arguments are correct, reflecting the model's ability to fully comprehend and generate precise function calls. Even if a more permissive evaluation of arguments could result in significantly higher outcomes, we prefer to adhere to an evaluation consistent with the public BFCL. The aggregated results report the average across all tested languages.

3.2 Results and Analysis

Fig. 2, Tab. 2-3, and appendix Tab. 5-10 present a comprehensive overview of our evaluation results across different models, prompting strategies, and dataset splits. Unless otherwise stated, all reported average scores are macro averages: we compute the accuracy for each language individually and then average across languages.

Looking at zero-shot performance on the 10k split (Tab. 2, 5, and 6), we observe several key trends. First, larger models generally perform better, with Nova Premier achieving the highest average AST score (34.05%) and Claude 3.5 v2 Sonnet the highest FSA score (86.61%). This trend is consistent with model scale, as evidenced by the progression in the Llama family, where performance steadily improves from 1B (AST: 0.23%, FSA: 2.50%) to 405B parameters (AST: 25.61%, FSA: 81.07%).

As expected, the one-shot setup with our custom prompt (Tab. 2, 7, and 8) yields substantial improvements across all models. The performance gains are particularly pronounced in FSA, with Claude 3.5 v2 Sonnet reaching 90.83% average accuracy (Tab. 8), suggesting that the model bet-

Model	Params	Max CL	DoR	Output Speed
Llama 3.1 8B	8B	128k	07/24	182
Llama 3.1 70B	70B	128k	07/24	73
Llama 3.1 405B	405B	128k	07/24	30
Llama 3.2 1B	1B	131k	09/24	560
Llama 3.2 3B	3B	131k	09/24	198
Llama 3.2 11B	11B	128k	09/24	131
Mistral 7B	7B	32k	09/23	97
Mixtral 8x7B	56B	32k	12/23	121
Command R	35B	128k	08/24	112
Command R+	104B	128k	08/24	49
Claude 2.1	NP	200k	07/23	NP
Claude 3 Haiku	NP	200k	03/24	124
Claude 3 Sonnet	NP	200k	03/24	66
Claude 3 Opus	NP	200k	03/24	26
Claude 3.5 Sonnet	NP	200k	06/24	56
Claude 3.5 Haiku	NP	200k	10/24	65
Claude 3.5 v2 Sonnet	NP	200k	10/24	54
Nova Micro	NP	128k	12/24	197
Nova Lite	NP	300k	12/24	148
Nova Pro	NP	300k	12/24	92
Nova Premier	NP	1M	04/25	66

Table 1: Summary table of the models evaluated in this study, detailing each model's parameter count, maximum context length (CL), release date (DoY), and Output Speed (tokens/sec)[3] performance.

ter understands function selection when provided with an example and more tailored instructions. Similarly, AST scores show notable improvements across Claude 3 Opus, Nova Micro, and Nova Pro. For Claude 3 Opus, the AST zero-shot BFCL score improved from 13.03% to 35.90% with a one-shot custom prompt. Nova Micro saw an increase from 5.53% (AST zero-shot BFCL) to 25.38% (AST one-shot custom prompt), while Nova Pro improved from 23.65% to 42.65% under the same conditions. Despite these advancements, AST results remain below 50%, highlighting that while models are becoming more proficient at selecting the correct function, they still face challenges in precise argument parsing across all languages, as particularly shown by the Sonic family of models.

The comparison between 10k and full splits (Tab. 3, 9, and 10) demonstrates strong correlation in performance patterns, validating our use of the 10k split as a reliable proxy for model evaluation. This correlation is evident in both high-performing models, such as Nova Premier, Claude 3.5 Sonnet, and smaller models, such as Llama 3.1 8B, where relative performance rankings remain consistent across both splits.

When examining language-specific performance, we observe significant variations across different language families and scripts. Germanic languages (e.g., English, German, Dutch) and Romance languages (e.g., Italian, French, Spanish) consistently show comparatively high performance across all models. For instance, in the zero-shot setting, English achieves the maximum AST score (57.37%) with Nova Premier, while Thai (th-TH) and Italian (it-IT) also demonstrate strong performance with scores above 40% for several models. On the other hand, we observe substantial performance degradation for certain language groups. Low-resource languages like Amharic (am-ET) and Javanese (jv-ID) consistently appear among the lowest-performing languages across all models and setups. This performance gap is particularly evident in zero-shot scenarios, where these languages often score below 5% in AST.

Finally, the error analysis shown in Fig. 3 categorizes models into three size groups across the spectrum from small models like Llama 3.2 1B to large ones like Llama 3.1 405B and Nova Premier, showing their error distributions on a logarithmic scale with dotted lines indicating maximum values. The analysis reveals a clear evolution of error patterns as model size increases. Small models like Llama 3.2 1B primarily fail due to basic syntax errors, reflecting their fundamental struggle with function call structure and with instruction following due to the size of the context. This pattern shifts significantly in medium and large models, where errors stem mainly from wrong function selection and missing/invalid parameters rather than syntax issues. Notably, across the entire size spectrum, we do not observe significant hallucination issues, suggesting that models remain constrained to the provided function specifications when generating function calls.

For comprehensive analysis, App. A.3 presents complete results for all languages across both BFCL zero-shot and one-shot setups, as well as the full split evaluation, providing detailed insights into model performance across the entire language spectrum.

3.3 Ablation Study

We conduct an ablation study on prompt translation strategies using seven representative models: three large models (Claude 3.5 v2 Sonnet, Nova Pro, and Claude 3.5 Sonnet) and four smaller ones (Llama 3.1 8B, Claude 3.5 Haiku, Nova Micro, and Claude

Model	AST, BFCL Prompt 0-Shot, 10k					AST, Custom Prompt, 1-Shot, 10k				
	Avg	Max Lang	Max	Min Lang	Min	Avg	Max Lang	Max	Min Lang	Min
Nova Premier	34.05	57.37	en-US	6.81	am-ET	40.40	63.68	en-US	12.57	am-ET
Claude 3.5 v2 Sonnet	27.06	53.68	en-US	17.37	kn-IN	29.37	63.68	en-US	17.37	kn-IN
Llama 3.1 405B	25.61	46.32	en-US	12.04	am-ET	32.97	58.64	th-TH	17.28	cy-GB
Claude 3.5 Haiku	24.94	43.16	en-US	12.04	cy-GB	28.42	57.37	en-US	14.29	jv-ID
Nova Pro	23.65	47.89	en-US	10.53	pl-PL	42.65	59.47	en-US	25.26	kn-IN
Nova Lite	22.52	43.16	en-US	12.11	pl-PL	32.70	55.79	en-US	17.46	jv-ID
Llama 3.1 70B	18.77	40.84	th-TH	6.28	mn-MN	30.02	54.97	th-TH	13.09	cy-GB
Claude 3 Opus	13.03	27.22	th-TH	4.19	hu-HU	35.90	59.47	en-US	22.11	kn-IN
Llama 3.2 11B	9.78	21.05	en-US	1.57	am-ET	18.41	32.63	en-US	2.62	am-ET
Llama 3.1 8B	9.09	20.00	en-US	0.52	am-ET	11.94	32.63	zh-TW	1.04	ka-GE
Mistral 7B	5.81	20.53	en-US	0.00	multi	7.86	21.58	en-US	0.00	multi
Nova Micro	5.53	17.89	en-US	0.00	km-KH	25.38	48.95	en-US	12.70	jv-ID
Llama 3.2 3B	4.33	18.42	en-US	0.52	multi	3.91	14.21	en-US	0.00	multi
Command R+	3.58	16.32	en-US	0.00	multi	4.88	24.74	en-US	0.00	multi
Mixtral 8x7B	3.23	12.11	en-US	0.00	multi	7.86	25.26	en-US	0.00	am-ET
Claude 2.1	0.67	4.74	sv-SE	0.00	multi	0.46	2.63	es-ES	0.00	multi
Llama 3.2 1B	0.23	1.59	de-DE	0.00	multi	0.11	2.10	en-US	0.00	multi

Table 2: Summary table of AST results for the BFCL zero-shot and custom one-shot prompts on the 10k split. For each evaluated model, macro average, maximum, and minimum scores are reported.

Figure 3: Error distribution analysis across model sizes on a logarithmic scale. Dotted lines indicate maximum values. The progression shows a shift from syntax errors in small models to function selection and parameter errors in larger models.

3 Haiku). Using Amazon Translate[4], we create variants of the BFCL zero-shot template where we translate separately the initial instructions and function descriptions for six languages: Italian (it-IT), German (de-DE), Spanish (es-ES), Japanese (ja-JP), Hindi (hi-IN), and Chinese (zh-CN). We then evaluate models using either partial translation (only descriptions) or full translation (both instructions and descriptions), comparing these ap-

[4] https://aws.amazon.com/it/translate/

	AST, BFCL Prompt, 0-Shot, Full				
Model	Avg	Max Lang	Max	Min Lang	Min
Nova Premier	37.93	58.34	de-DE	6.30	am-ET
Claude 3.5 Sonnet	28.55	45.93	th-TH	19.36	ur-PK
Claude 3.5 Haiku	27.18	40.27	en-US	17.05	jv-ID
Nova Lite	25.94	41.55	en-US	13.21	km-KH
Nova Pro	25.16	45.65	en-US	15.40	mn-MN
Llama 3.1 70B	19.29	39.34	th-TH	8.56	mn-MN
Claude 3 Opus	13.02	27.91	th-TH	4.13	jv-ID
Llama 3.1 8B	9.40	19.62	en-US	1.97	am-ET
Mistral 7B	7.01	18.80	en-US	0.24	my-MM
Nova Micro	5.58	15.42	en-US	0.00	km-KH
Mixtral 8x7B	3.83	9.63	en-US	0.00	multi
Claude 2.1	0.58	3.11	sv-SE	0.00	multi

Table 3: Summary table of AST results for the BFCL zero-shot prompt on the full split. For each evaluated model, the macro average across languages, maximum, and minimum scores are reported. The results closely mirror those from the 10k split, demonstrating that the 10k split serves as a reliable proxy for the full test set.

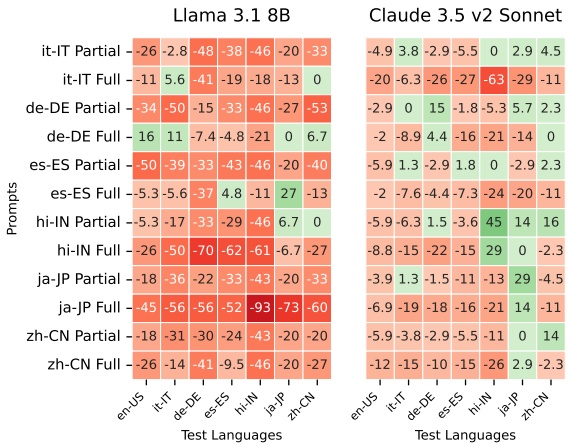

Figure 4: Relative AST performance changes in percentage points when using translated prompts compared to the English-only baseline. Results for Llama 3.1 8B and Claude 3.5 v2 Sonnet across six test languages, showing both partial (function descriptions only) and full (instructions and descriptions) translation prompts.

proaches against an English-only baseline.

The results, visualized in Fig. 4 and App.A.5, reveal distinct patterns across model sizes. In general, smaller models like Llama 3.1 8B show significant performance degradation when moving away from English instructions, particularly for non-Latin scripts. For instance, Japanese translations cause performance drops of up to 73% in AST scores when evaluated on the same language of the prompt. Conversely, larger models demonstrate overall improvements with prompt translations. Nova Pro, and Claude 3.5 v2 Sonnet show consistent improvements in both AST and FSA metrics when provided with only translated descriptions. The gains are particularly notable in AST Claude 3.5 v2 Sonnet (improvements of 1-44%), and Nova Pro (1-26%). However, interestingly, these improvements diminish or even reverse when using fully translated prompts either for AST or FSA, suggesting a preference for English instruction. On the other hand, Claude 3 and 3.5 Haiku models show a slightly more varied pattern, with FSA scores actually improving in some cases when only descriptions are translated, suggesting better function identification with language-specific descriptions. This is not the case for AST, where noticeable and significant degradations are observed in most of the evaluations.

Additionally, an interesting observation emerged from this ablation study: Claude 3.5 Sonnet and Nova Micro show substantial improvements across nearly all translation combinations (Fig. 7 - 10). The gains are explained by both models' tendency to generate explanatory comments or additional questions before function calls when prompted in English, leading to syntax errors despite explicit instructions against this behavior. This tendency diminishes with translated prompts, resulting in cleaner function calls. Interestingly, this same improvement is achieved with the custom one-shot template, which more explicitly prohibits additional commentary (Claude 3.5 Sonnet from 26.88% to 39.18%, Nova Micro from 5.53% to 25.38%). While BFCLv2 zero-shot includes similar instructions before the action list (as shown in App. A.4), the positioning appears less effective in preventing this behavior. The complete results of this ablation study, including detailed AST and FSA scores for all models and language combinations, are provided in App. A.5.

4 Conclusion

When discussing artificial general intelligence, emphasis is often given to a model's performance across a breadth of tasks and domains. We contend that full multilingualism across all of the Earth's languages should also be a requirement. MASSIVE and subsequently MASSIVE-AGENTS are unique in that the languages are sampled across a range of typologies and language resource levels. We hope that this new benchmark will prove valuable to language model researchers worldwide toward the goal of greater multilinguality of future foundation models.

5 Limitations

This work covers a broad set of experiments across 52 languages and 21 model types; however, there exist several limitations that we call out in this section. As the dataset is adapted from the original MASSIVE dataset, which was based on voice assistant requests, this also restricts the paradigm in which the dataset exists and may not be entirely representative of the function calling abilities or types of requests that may be currently asked of LLMs. Specifically, the dataset is restricted to single-turn utterances and also static workflows that were achievable through the last generation of voice assistants. Further, the evaluation expects the verbatim use of an argument value from the voice transcription and does not allow for normalization of date/time or for aliases and thus is a stricter evaluation. We also find that the MASSIVE dataset is not well suited for transformation to multiple function calling due to a lack of good annotations with multiple function calls and thus do not evaluate multiple function calling and leave this as future work. However, we also call out that despite these limitations, the dataset is still challenging, and contemporary LLMs still struggle to achieve remarkable performance. Further, due to the contemporary nature of reasoning style models, we have been unable to benchmark these as a part of this work and aim to also benchmark the same in the near future. Finally, our language ablation that considers prompts in other languages only covers high-resource languages for which we could also find speakers to verify our machine-translated prompts, and this in the future could be extended to more languages and to look for latent improvements that may exist in similar language roots or geographical closeness.

References

Anthropic. 2024. Model card addendum: Claude 3.5 haiku and upgraded claude 3.5 sonnet. Accessed: 2024-11-17.

Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. SLURP: A spoken language understanding resource package. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7252–7262, Online. Association for Computational Linguistics.

Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Amy Yang, Angela Fan, et al. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jack FitzGerald, Christopher Hench, Charith Peris, Scott Mackie, Kay Rottmann, Ana Sanchez, Aaron Nash, Liam Urbach, Vishesh Kakarala, Richa Singh, Swetha Ranganath, Laurie Crist, Misha Britan, Wouter Leeuwis, Gokhan Tur, and Prem Natarajan. 2023. MASSIVE: A 1M-example multilingual natural language understanding dataset with 51 typologically-diverse languages. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4277–4302, Toronto, Canada. Association for Computational Linguistics.

Naman Goyal, Cynthia Gao, Vishrav Chaudhary, Peng-Jen Chen, Guillaume Wenzek, Da Ju, Sanjana Krishnan, Marc'Aurelio Ranzato, Francisco Guzman, and Angela Fan. 2021. The flores-101 evaluation benchmark for low-resource and multilingual machine translation.

Amazon Artificial General Intelligence. 2024a. The amazon nova family of models: Technical report and model card. *Amazon Technical Reports*.

Amazon Artificial General Intelligence. 2024b. Amazon nova premier: Technical report and model card. *Amazon Technical Reports*.

Surafel M. Lakew, Matteo Negri, and Marco Turchi. 2020. Low resource neural machine translation: A benchmark for five african languages.

Minghao Li, Yingxiu Zhao, Bowen Yu, Feifan Song, Hangyu Li, Haiyang Yu, Zhoujun Li, Fei Huang, and Yongbin Li. 2023. Api-bank: A comprehensive benchmark for tool-augmented llms.

Alexandre Magueresse, Vincent Carles, and Evan Heetderks. 2020. Low-resource languages: A review of past work and future challenges.

Vukosi Marivate, Tshephisho Sefara, Vongani Chabalala, Keamogetswe Makhaya, Tumisho Mokgonyane, Rethabile Mokoena, and Abiodun Modupe. 2020. Investigating an approach for low resource language dataset creation, curation and classification: Setswana and sepedi. In *Proceedings of the first workshop on Resources for African Indigenous Languages*, pages 15–20, Marseille, France. European Language Resources Association (ELRA).

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. 2023. Gorilla: Large language model connected with massive apis.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessi, Roberta Raileanu, Maria Lomeli, Eric Hambro, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. 2023. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, pages 68539–68551. Curran Associates, Inc.

Heather Simpson, Christopher Cieri, Kazuaki Maeda, Kathryn Baker, and Boyan Onyshkevych. 2008. Human language technology resources for less commonly taught languages: Lessons learned toward creation of basic language resources. In *SALTMIL Workshop on interoperability between people in the creation of language resources for less-resourced languages*, pages 7–11.

Stephanie Strassel and Jennifer Tracey. 2016. LORELEI language packs: Data, tools, and resources for technology development in low resource languages. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3273–3280, Portorož, Slovenia. European Language Resources Association (ELRA).

Nexusflow.ai team. 2023. Nexusraven-v2: Surpassing gpt-4 for zero-shot function calling.

Zhiheng Xi, Wenxiang Chen, Xin Guo, Wei He, Yiwen Ding, Boyang Hong, Ming Zhang, Junzhe Wang, Senjie Jin, Enyu Zhou, Rui Zheng, Xiaoran Fan, Xiao Wang, Limao Xiong, Yuhao Zhou, Weiran Wang, Changhao Jiang, Yicheng Zou, Xiangyang Liu, Zhangyue Yin, Shihan Dou, Rongxiang Weng, Wensen Cheng, Qi Zhang, Wenjuan Qin, Yongyan Zheng, Xipeng Qiu, Xuanjing Huang, and Tao Gui. 2023. The rise and potential of large language model based agents: A survey.

Fanjia Yan, Huanzhi Mao, Charlie Cheng-Jie Ji, Tianjun Zhang, Shishir G. Patil, Ion Stoica, and Joseph E. Gonzalez. 2024. Berkeley function calling leaderboard.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2024. τ-bench: A benchmark for tool-agent-user interaction in real-world domains.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. 2023. ReAct: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*.

Appendix

A.1 Manual Filtering Details

As discussed in 2.3.1 we perform significant manual filtering of the data to ensure only high-quality data points make it through the pipeline. We use the following guidelines or buckets in order to decide whether or not an utterance is deemed to be high quality.

- **Missing or Incomplete Annotations**: We find some utterances are missing annotations for crucial information either due to a lack of appropriate slot types or poor quality annotation. One such example is "add lowes hardware to my contact emails located in [$place\_name$: cleveland ohio]", where "lowes hardware" is missing an annotation for $business\_name$.
- **Incorrect Annotations**: We find some utterances to have incorrect annotations due to poor annotation quality. An example of such is '[song_name : comfort my ears] with [artist_name : arijit singh]' where 'comfort my ears' is not a song name.
- **Ambiguous Utterances**: We find some utterances missing context on the expected behavior or item being referenced, examples of such include "remove that item", "respond to email" or "set a reminder for this".
- **Unsupported Behaviors**: We observe that some utterances expect behaviors of future function calls or request functionality not supported by the function. Examples are 'notifications for news' and 'explanation of the song' respectively.
- **Multi-Turn Utterances**: We drop a handful of utterances that are leading utterances, expecting the model to produce a response, before the user responds with the action the user wants the model to take, thus requiring the model to ingest more than one user turn before taking an action.
- **Multi-Function Utterances**: We drop the handful of utterances that require multiple function calls to be made, as the annotations for MASSIVE did not support this (although they would still constitute valid requests). Often the functions themselves are repeated with different arguments and not necessarily chaining multiple functions to achieve a workflow. For example, 'i love this music can you please save this to my dance playlist and remember that i like it' requires two different function calls, and 'where can i get a shot of [food_type : tequila] and some to go [food_type : mexican] food' requires the same function call twice. Thus we decided to focus on high quality on single function calling and leave multiple function calls as future work.
- **Low-frequency mismatched arguments**: We find some utterances with very rarely requested argument types where the argument type itself did not align with the rest of the function. Such an example is, "set a repeating reminder alarm for the [$media\_type$: facebook] event i have to attend on [$date$: seventh april]", which $media\_type$ is an odd placed functionality for an alarm function.
- **Duplicated Arguments**: We also ignore the handful of utterances where there are multiple annotations for the same argument name, as these are often poor-quality utterances or typically require multiple intents. 'where can i get a shot of [food_type : tequila] and some to go [food_type : mexican] food' is again an example of this.
- **Open-ended and catch all functions**: We explicitly drop the functions of $general\_greet$, $general\_quirky$, and $cooking\_query$ as these are all open-ended utterances, making it hard to define arguments and also rationalize having these as functions. Further, we also drop functions of $audio\_volume\_other$ and $music\_settings$ as these serve as catch all functions for the audio and music setups and often have poor-quality utterances.

A.2 Dataset Statistics

We compute statistics on both the full and 10k datasets in Table 4 to present the number of functions covered, the namespaces covered, unique arguments (ARG), unique combinations of function and argument keys (function-ARG), and the total number of samples. We see that across both the full and 10k datasets there is similar coverage for the function-ARG combinations, which is further evidence that these datasets correlate well with each other. We also see that due to annotation quality issues, some languages, such as Khmer (km-KH) have a much smaller coverage of only 53 samples; however, most other languages are fairly well represented in the dataset.

	10k					Full				
Lang	Function	Namespace	ARG	Function-ARG	Total	Function	Namespace	ARG	Function-ARG	Total
All	55	18	53	279	9741	55	18	53	286	47020
tr-TR	55	18	39	108	189	55	18	49	209	1163
sq-AL	54	18	42	111	188	54	18	47	206	1301
fr-FR	55	18	46	113	189	55	18	47	195	972
de-DE	53	18	39	110	189	53	18	44	150	737
pl-PL	55	18	45	123	190	55	18	51	231	1359
sl-SL	53	18	41	115	192	53	18	47	204	1122
vi-VN	48	17	35	108	188	48	17	37	134	435
da-DK	52	18	37	109	188	52	18	45	184	992
el-GR	51	18	44	111	191	51	18	46	187	950
cy-GB	54	18	43	112	191	54	18	47	203	1227
es-ES	55	18	43	112	190	55	18	49	187	913
mn-MN	46	17	38	100	191	46	17	41	136	409
ru-RU	55	18	41	113	189	55	18	49	206	1185
is-IS	53	18	41	108	188	53	18	43	176	879
ca-ES	55	18	42	122	189	55	18	47	202	1026
lv-LV	55	18	44	115	192	55	18	49	186	896
ar-SA	52	18	40	105	191	52	18	42	159	602
kn-IN	55	18	48	124	190	55	18	53	253	1768
hu-HU	52	17	39	107	191	52	17	44	186	988
nl-NL	55	18	43	113	190	55	18	50	189	1005
hy-AM	49	18	43	110	192	49	18	44	158	670
km-KH	16	9	13	30	53	16	9	13	30	53
ms-MY	55	18	43	120	191	55	18	50	222	1251
ja-JP	54	18	44	118	190	54	18	47	200	1079
zh-CN	52	17	41	113	188	52	17	47	218	1380
ro-RO	54	18	41	110	185	54	18	49	198	904
hi-IN	39	16	33	64	185	39	16	33	64	185
he-IL	55	18	46	118	192	55	18	49	210	1218
ml-IN	55	18	45	117	189	55	18	53	228	1277
pt-PT	53	18	43	107	190	53	18	46	153	536
tl-PH	50	17	38	103	192	50	17	39	174	929
zh-TW	51	18	37	96	190	51	18	41	130	525
nb-NO	52	18	41	100	191	52	18	43	162	863
ko-KR	55	18	39	115	192	55	18	49	178	840
it-IT	51	18	43	94	189	51	18	45	129	602
id-ID	55	18	46	121	188	55	18	53	238	1335
ta-IN	51	18	39	104	192	51	18	43	157	605
th-TH	50	17	36	96	191	50	17	41	119	455
az-AZ	52	18	40	109	192	52	18	43	166	657
sv-SE	52	18	43	107	190	52	18	48	198	1126
sw-KE	55	18	40	114	191	55	18	48	208	1162
te-IN	52	18	41	100	191	52	18	43	126	478
en-US	55	18	42	104	190	55	18	53	270	1952
bn-BD	49	18	38	99	191	49	18	41	133	605
am-ET	40	17	41	89	191	40	17	41	92	254
my-MM	54	18	44	112	191	54	18	48	185	819
jv-ID	55	18	50	119	189	55	18	53	221	1161
ur-PK	34	14	28	68	190	34	14	28	68	248
ka-GE	54	18	40	112	192	54	18	43	152	736
fi-FI	55	18	42	108	191	55	18	47	198	1088
af-ZA	55	18	45	113	188	55	18	49	199	1054
fa-IR	55	18	45	116	188	55	18	52	202	1044

Table 4: Dataset Statistics for the 10k and Full dataset

A.3 Results for all languages on 10k and full splits

In this appendix, we present results for all languages and models tested, both for the 10k and full splits. Specifically, Tables 5 and 6 provide AST and FSA results for the 10k split, obtained using the BFCL zero-shot template. In contrast, Tables 7 and 8 present data for both metrics on the same split but using the custom one-shot prompt. Lastly, Tables 9 and 10 report scores for AST and FSA on the full split with the BFCL zero-shot template.

A.4 BFCL and Custom one-shot prompts

To help reproducibility, this section outlines the complete prompts used across all experiments. Figure 5 depicts the original BFCL prompt, which does not feature any examples and provides instructions before listing the available functions. As a result, due to the extensive number of functions and the corresponding large context, models may produce a higher percentage of syntax errors. Conversely, Figure 6 presents our custom prompt, which incorporates tailored instructions for the MASSIVE-Agents dataset and includes a fixed example to help the model follow the correct syntax. Except for the ablation study on prompt translation, both prompts are static and always provided in English.

```
You are an expert in composing functions. You are
     given a question and a set of possible
     functions.
Based on the question, you will need to make one or
     more function/tool calls to achieve the purpose
     .
If none of the function can be used, point it out.
     If the given question lacks the parameters
     required by the function,
also point it out. You should only return the
     function call in tools call sections.

If you decide to invoke any of the function(s), you
     MUST put it in the format of [func_name1(
     params_name1=params_value1, params_name2=
     params_value2...), func_name2(params)]\n
You SHOULD NOT include any other text in the
     response.

Here is a list of functions in JSON format that you
     can invoke.

{functions} User: {question} Bot:
```

Figure 5: BFCL zero-shot prompt, showcasing placeholders for `functions` and the input `question`.

A.5 Ablation Study - Additional Material

This appendix section provides a comprehensive report on our ablation study of prompt translation, presenting AST and FSA results for all large models (Claude 3.5 v2 Sonnet, Nova Pro, and Claude 3.5 Sonnet) and small models (Llama 3.1 8B, Claude 3.5 Haiku, Nova Micro, and Claude 3 Haiku) tested. All variations are based on the BFCL zero-shot template, utilizing Amazon Translate. Specifically, Figures 7 and 8 display results for all models and languages using the AST metric, while Tables 9 and 10 present results for the FSA metric.

```
A chat between a curious User and an artificial
     intelligence Bot. The Bot gives helpful,
     detailed, and polite answers to the User's
     questions. In this session, the model has
     access to external functionalities.

The following actions are available:
{functions}

Model Instructions:
- Function calls must adhere to the programming
     language's syntax.
- Select the most suitable action from the given
     list.
- Extract relevant arguments for function calls from
      the user's utterance.
- Do not ask any clarifying questions.
- Preserve the original argument values as provided
     in the user's utterance.
- Always respond with a function call (e.g., Action:
      ActionOne()). Do not add further explanations
     or comments.

Example Interaction:
User: set an alarm for nine am this week
Bot: Action: alarm.set(time="nine am", date="this
     week")

User: {question} Bot:
```

Figure 6: Custom one-shot custom prompt, showcasing placeholders for `functions` and the input `question`.

Language	Llama 3.2 1B	Claude 2.1	Mixtral 8x7B	Command R+	Llama 3.2 3B	Command R	Nova Micro	Mistral 7B	Claude 3 Haiku	Llama 3.1 8B	Llama 3.2 11B	Claude 3 Opus	Llama 3.1 70B	Claude 3 Sonnet	Nova Lite	Nova Pro	Claude 3.5 Haiku	Llama 3.1 405B	Claude 3.5 Sonnet	Claude 3.5 v2 Sonnet	Nova Premier
Average	0.23	0.67	3.23	3.58	4.33	4.94	5.53	5.81	8.24	9.09	9.78	13.03	18.77	19.80	22.52	23.65	24.94	25.61	26.88	27.06	34.05
af-ZA	0.00	0.53	4.79	12.23	3.72	9.04	5.85	6.38	5.85	13.30	11.70	9.57	15.96	20.21	23.40	27.66	24.47	27.66	30.32	27.66	41.49
am-ET	0.00	0.00	0.00	0.00	0.52	0.00	1.05	0.52	3.14	0.52	1.57	8.38	6.81	14.14	16.23	17.28	19.89	12.04	25.13	22.51	6.81
ar-SA	0.52	0.52	2.62	0.00	4.19	0.00	5.76	7.85	7.85	6.28	4.71	10.47	15.71	18.32	18.32	20.94	22.51	19.89	26.70	25.66	36.13
az-AZ	0.00	1.56	3.12	1.04	3.12	2.08	5.21	2.08	8.33	4.17	5.21	13.02	17.71	20.31	27.08	21.88	22.92	18.75	25.00	27.60	36.98
bn-BD	0.00	0.00	2.09	0.00	2.09	0.00	3.14	4.19	6.28	3.14	3.67	12.04	13.09	15.71	14.66	17.28	17.80	21.99	17.80	21.47	31.94
ca-ES	0.53	0.53	6.35	5.82	6.35	9.00	6.35	6.35	9.00	7.94	8.47	10.58	17.99	24.34	19.05	24.87	26.98	30.69	25.40	27.51	34.92
cy-GB	0.00	0.00	1.05	1.05	1.05	2.09	1.05	1.57	3.14	4.19	3.67	5.24	6.81	12.04	16.75	19.89	12.04	15.71	19.89	20.42	19.89
da-DK	0.53	1.06	3.19	9.04	2.13	10.11	6.38	10.11	7.98	12.23	14.89	10.11	25.00	26.60	28.19	32.98	25.00	39.36	32.45	35.64	48.40
de-DE	1.59	1.06	4.23	10.58	8.47	13.76	11.64	12.70	9.52	14.29	12.70	23.28	29.63	28.04	25.93	23.81	32.80	31.75	34.92	35.98	49.73
el-GR	0.00	0.00	5.76	0.00	2.62	0.00	1.57	4.71	4.19	7.33	7.33	14.66	18.32	20.42	20.94	22.51	25.66	24.61	30.89	25.13	44.50
en-US	0.00	0.00	12.11	16.32	18.42	22.11	17.89	20.53	15.79	20.00	21.05	17.89	34.21	32.10	43.16	47.89	43.16	46.32	38.42	53.68	57.37
es-ES	0.53	0.00	3.68	10.00	9.47	14.74	6.32	7.37	9.47	11.05	12.63	13.68	23.68	23.16	22.11	22.63	27.37	30.00	32.63	28.95	45.79
fa-IR	0.53	1.06	3.19	0.00	0.53	0.00	3.72	9.57	10.64	10.11	12.77	17.55	17.55	22.34	28.19	22.87	25.00	25.53	24.47	22.34	45.21
fi-FI	0.00	0.00	3.67	4.19	4.71	5.76	4.71	2.62	4.19	3.14	3.67	7.85	16.23	15.18	17.28	18.85	18.85	21.47	16.23	19.89	30.37
fr-FR	0.53	1.06	2.65	9.00	7.94	17.46	7.41	9.52	7.94	14.82	14.29	13.23	29.10	29.63	29.63	31.22	34.39	35.98	37.57	38.62	47.09
he-IL	0.00	0.00	2.08	0.00	0.52	0.00	0.52	7.29	3.12	5.21	5.73	8.85	11.98	17.19	20.83	22.40	19.79	19.27	18.75	19.27	26.04
hi-IN	0.54	2.16	1.08	0.00	4.87	0.54	9.19	4.87	8.65	15.14	14.05	18.38	23.24	19.46	34.05	25.41	27.57	35.13	27.03	20.54	57.34
hu-HU	0.00	2.62	4.71	3.67	2.62	3.14	7.33	6.28	11.00	6.81	8.38	4.19	14.14	19.89	19.37	23.04	23.04	20.94	23.56	25.66	30.89
hy-AM	0.00	0.00	0.00	0.00	1.04	0.00	3.12	1.04	8.85	3.65	4.17	8.33	13.02	16.15	15.62	18.23	17.19	13.54	18.23	18.23	11.98
id-ID	0.00	0.00	4.25	11.70	2.66	12.77	5.85	5.85	10.64	11.17	12.77	13.30	28.19	20.75	22.34	27.13	32.45	32.98	30.85	32.98	43.62
is-IS	0.00	0.00	1.60	1.06	1.06	1.06	3.19	3.72	5.32	4.79	5.85	6.92	11.17	15.43	18.09	19.15	15.96	17.55	19.15	21.28	29.79
it-IT	1.06	2.12	6.88	12.70	12.17	25.93	12.70	12.70	17.46	19.05	19.58	22.22	32.27	33.33	33.86	35.45	41.27	43.91	41.27	41.80	52.38
ja-JP	0.53	0.53	5.79	3.16	6.32	2.63	6.84	7.90	7.37	7.90	10.00	14.74	17.37	15.79	22.63	24.74	23.68	20.53	23.16	18.42	43.16
jv-ID	0.00	0.00	0.53	2.12	0.53	3.70	1.59	2.12	4.76	6.35	6.35	4.23	12.17	7.94	12.17	17.46	12.70	18.52	14.29	18.52	20.11
ka-GE	0.00	0.00	1.04	0.00	1.04	0.00	1.56	2.08	7.29	6.25	5.73	13.02	14.06	19.27	16.67	21.88	23.44	19.79	30.73	27.08	11.46
km-KH	0.00	0.00	0.00	0.00	5.66	0.00	0.00	1.89	11.32	16.98	16.98	20.76	16.98	24.53	13.21	26.41	22.64	16.98	30.19	28.30	13.21
kn-IN	0.00	1.58	1.58	0.00	3.16	0.00	2.63	0.53	10.00	8.42	8.95	14.21	11.05	12.63	15.79	15.26	14.21	13.16	16.84	17.37	14.21
ko-KR	1.04	0.52	9.90	0.52	4.69	0.00	8.85	9.90	10.42	10.42	12.50	18.23	19.79	20.31	23.44	21.35	27.60	29.69	29.69	26.04	25.52
lv-LV	0.00	0.52	1.04	0.00	3.12	2.08	6.77	3.12	3.65	8.85	7.29	7.29	11.46	16.67	17.71	20.83	23.96	21.35	24.48	26.56	33.33
ml-IN	0.00	0.53	0.53	0.00	0.53	0.00	2.12	0.00	6.88	7.94	9.00	14.82	12.17	16.93	19.58	20.11	20.64	15.34	19.05	23.28	19.58
mn-MN	0.00	0.00	1.57	0.00	1.57	0.00	2.62	1.05	8.38	4.71	5.76	11.52	6.28	16.23	17.28	18.32	22.51	16.23	22.51	22.51	22.51
ms-MY	0.00	0.52	4.19	9.95	1.57	10.47	5.76	3.14	6.28	11.52	11.00	9.95	17.28	19.37	24.08	26.18	27.75	28.80	27.22	30.37	40.31
my-MM	0.00	0.00	1.57	0.00	0.52	0.00	2.09	0.00	8.90	7.85	8.90	16.23	13.09	19.89	23.56	16.75	28.27	21.47	30.89	26.70	8.38
nb-NO	0.00	1.57	3.67	7.85	3.14	9.95	3.67	6.81	4.71	13.09	11.00	9.95	15.18	18.85	21.99	28.27	21.99	26.18	28.80	32.98	43.46
nl-NL	1.05	1.58	4.21	7.90	8.95	15.26	8.42	8.42	14.21	10.00	14.74	10.53	24.21	26.32	25.26	30.00	27.37	34.21	33.68	31.05	45.79
pl-PL	0.53	0.00	0.53	2.63	2.10	4.21	4.74	4.74	5.26	4.74	4.74	6.84	13.68	13.16	12.11	10.53	17.37	15.26	16.84	18.95	29.47
pt-PT	0.00	0.53	3.68	11.58	7.37	17.37	13.16	10.00	12.11	15.26	16.84	17.37	30.00	31.05	27.89	30.00	39.47	40.53	39.47	37.90	48.42
ro-RO	0.54	0.00	5.41	6.49	4.32	10.27	7.57	6.49	10.27	10.27	12.43	10.27	22.70	17.84	25.41	25.95	30.27	35.68	31.89	34.05	43.24
ru-RU	0.53	0.00	3.70	0.00	7.94	0.00	6.88	9.52	10.58	6.88	7.94	15.87	23.28	21.69	22.75	24.34	24.34	25.93	30.16	28.57	35.98
sl-SL	0.00	0.00	2.08	4.17	2.08	1.56	3.12	4.69	4.17	4.17	5.21	9.38	13.02	15.10	18.23	21.88	20.31	22.40	20.31	23.44	33.33
sq-AL	0.00	1.06	2.13	0.53	1.60	1.60	4.25	2.13	3.19	5.32	4.79	5.32	13.30	14.36	17.55	17.02	15.96	19.15	17.02	22.34	28.19
sv-SE	0.00	4.74	4.74	3.16	4.74	7.37	4.74	7.90	4.74	9.47	11.58	12.63	23.68	22.63	27.37	28.42	30.53	33.16	34.21	35.26	43.16
sw-KE	0.00	0.00	2.62	3.67	2.09	2.62	4.71	2.62	5.24	6.81	8.38	5.76	9.95	18.85	18.32	18.85	18.85	19.89	16.75	21.47	31.94
ta-IN	0.00	0.52	0.52	0.00	6.25	0.00	5.21	1.56	9.38	7.29	6.77	12.50	13.54	15.62	22.92	17.19	20.83	15.10	24.48	23.44	16.67
te-IN	0.00	0.00	0.52	0.00	6.28	0.00	4.19	1.05	12.04	8.90	9.95	25.66	17.80	17.28	22.51	24.08	21.99	19.37	28.27	27.75	25.13
th-TH	0.00	2.62	1.57	0.00	12.57	0.00	8.38	10.47	14.66	17.80	17.80	27.22	40.84	27.22	33.51	33.51	32.46	44.50	45.03	31.94	37.17
tl-PH	0.00	0.00	1.04	4.69	0.52	7.81	3.65	4.17	3.12	6.77	9.90	8.33	20.31	19.79	18.75	18.75	26.56	23.44	23.44	27.60	36.46
tr-TR	0.00	1.06	3.17	3.70	3.70	4.76	7.41	4.76	7.41	7.41	9.00	6.88	19.05	12.17	23.81	19.05	19.58	22.22	19.58	17.99	31.22
ur-PK	0.53	2.10	2.63	0.00	6.32	0.00	8.42	7.37	13.68	9.47	10.53	21.05	21.05	18.95	32.10	24.21	28.42	21.05	22.11	27.89	38.42
vi-VN	0.53	0.53	4.25	0.00	3.72	0.00	2.13	7.98	5.32	9.57	12.23	13.30	28.19	18.62	23.94	26.06	29.25	39.36	34.04	31.38	38.83
zh-CN	0.00	0.00	5.85	0.53	4.79	2.13	6.92	7.45	7.98	7.98	10.11	18.62	18.62	19.15	20.75	27.66	30.32	28.72	29.79	23.40	37.77
zh-TW	0.53	0.00	8.95	5.26	7.90	3.68	9.47	14.21	16.84	15.79	13.16	25.26	34.21	26.84	34.74	31.58	37.37	38.42	36.32	31.58	45.26

Table 5: BFCL zero-shot prompt on 10k split. AST for all tested languages.

Language	Llama 3.2 1B	Claude 2.1	Nova Micro	Command R+	Mixtral 8x7B	Command R	Llama 3.2 3B	Mistral 7B	Claude 3 Opus	Claude 3 Haiku	Llama 3.1 8B	Llama 3.2 11B	Nova Pro	Nova Lite	Claude 3 Sonnet	Llama 3.1 70B	Claude 3.5 Sonnet	Claude 3.5 Haiku	Nova Premier	Llama 3.1 405B	Claude 3.5 v2 Sonnet
Average	2.50	2.61	12.20	13.73	16.55	21.03	21.42	22.39	36.95	40.37	53.67	55.63	65.29	68.61	69.91	70.34	72.83	77.23	78.77	81.07	86.61
af-ZA	1.60	2.66	13.83	31.91	14.89	34.04	18.09	17.02	23.94	28.19	52.13	53.72	64.89	67.55	60.64	70.21	77.66	73.40	78.19	80.32	85.11
am-ET	0.52	0.00	2.09	0.00	1.57	0.52	0.52	0.52	29.84	36.65	15.71	17.28	53.40	53.93	65.45	49.74	68.59	73.82	53.40	69.63	83.25
ar-SA	3.66	2.62	12.04	1.05	14.14	2.09	20.94	27.75	40.84	36.65	51.83	52.88	70.68	67.02	72.25	72.77	73.82	79.06	81.15	81.15	87.96
az-AZ	0.52	6.77	12.50	14.06	12.50	14.58	16.15	15.62	37.50	38.02	56.25	61.98	59.90	71.88	67.71	73.44	72.40	71.88	81.77	80.73	87.50
bn-BD	1.57	1.57	7.85	0.00	19.90	0.00	21.99	21.47	39.79	42.41	51.83	53.93	62.83	70.68	78.01	81.15	64.92	74.87	79.58	84.82	86.39
ca-ES	2.65	4.23	11.64	20.63	19.58	32.80	22.22	28.04	23.81	32.28	55.03	57.67	66.14	63.49	70.90	70.90	67.72	76.72	79.37	81.48	84.66
cy-GB	0.52	0.00	4.71	8.38	7.33	14.14	5.76	6.81	13.61	19.90	29.84	28.27	54.97	57.07	52.36	48.69	56.54	49.21	48.17	61.78	67.54
da-DK	4.79	2.66	15.96	28.19	18.62	38.83	16.49	25.00	21.28	26.06	53.72	56.91	72.87	73.94	65.96	62.77	71.28	72.87	85.11	85.11	89.36
de-DE	5.29	2.12	22.22	40.21	21.16	49.74	32.80	37.57	38.10	38.10	60.85	59.26	59.26	68.78	66.14	73.54	76.19	79.37	85.19	80.95	86.77
el-GR	1.57	3.66	2.62	0.00	23.04	2.62	17.80	25.13	43.98	47.12	59.69	61.78	73.30	69.11	75.92	73.30	83.77	79.58	87.43	84.82	89.53
en-US	3.68	0.00	32.11	33.68	28.42	60.00	54.21	47.89	30.00	36.32	63.68	67.37	80.53	82.11	58.42	75.26	66.84	84.74	91.58	87.89	89.47
es-ES	3.16	3.16	16.32	32.11	28.95	57.89	40.53	38.95	30.53	39.47	62.11	62.63	72.11	66.84	71.58	70.53	82.11	81.05	90.53	86.84	91.05
fa-IR	4.26	3.72	6.91	2.66	20.21	2.66	21.81	29.26	48.40	46.28	62.23	64.36	62.77	71.81	77.13	69.15	73.94	76.60	80.85	83.51	86.17
fi-FI	1.05	0.00	17.80	18.32	19.37	29.84	16.23	18.32	28.80	35.08	47.64	52.36	64.40	74.87	78.53	70.16	74.35	80.10	85.86	83.25	88.48
fr-FR	4.76	2.12	14.81	29.63	22.22	59.26	36.51	33.86	29.10	34.92	58.20	59.26	77.78	75.66	69.31	82.54	84.13	84.13	87.83	88.36	93.65
he-IL	1.56	0.00	2.60	0.00	16.15	2.08	6.77	27.08	40.62	40.10	56.77	60.94	78.12	78.12	80.73	79.17	80.73	82.81	82.81	83.85	90.10
hi-IN	3.24	8.11	22.16	1.62	15.14	1.08	30.81	19.46	57.84	47.03	61.62	62.16	64.86	67.03	82.16	84.86	68.65	86.49	87.03	85.95	93.51
hu-HU	3.14	9.95	17.80	14.66	19.37	19.90	19.90	24.08	20.42	37.70	52.88	56.02	61.26	69.63	72.77	62.83	73.82	76.96	81.15	79.58	84.29
hy-AM	1.04	1.04	7.29	0.00	7.81	2.60	9.38	3.12	38.54	45.83	54.69	59.90	71.88	69.79	69.27	75.00	72.92	77.60	61.46	84.90	83.33
id-ID	2.66	0.53	14.89	36.70	22.87	45.74	26.06	28.72	32.98	40.96	55.32	61.17	74.47	71.81	66.49	65.96	75.53	81.91	81.38	82.98	85.64
is-IS	0.00	1.06	10.11	7.98	9.57	5.32	5.85	8.51	21.28	26.06	36.17	36.17	55.85	62.23	61.17	57.98	66.49	70.21	78.72	70.74	84.04
it-IT	3.17	6.88	23.81	33.33	17.46	62.96	38.62	40.74	41.80	41.80	73.02	75.13	75.13	73.54	75.13	72.49	80.42	83.60	92.06	91.01	95.24
ja-JP	6.84	1.58	22.11	12.11	25.79	18.42	30.53	34.74	48.42	58.42	62.63	67.37	72.63	68.95	78.42	81.05	70.00	84.74	84.21	82.11	87.89
jv-ID	0.53	0.00	3.70	8.99	5.29	18.52	7.94	7.94	12.17	18.52	29.63	30.16	47.62	45.50	35.45	43.39	43.92	42.86	53.44	54.50	61.90
ka-GE	0.52	0.00	7.29	0.00	10.42	2.08	7.81	5.73	43.75	41.67	48.96	48.96	65.62	62.50	75.52	68.23	75.52	77.60	58.33	77.60	88.54
km-KH	3.77	0.00	0.00	0.00	3.77	11.32	24.53	7.55	60.38	58.49	67.92	62.26	66.04	67.92	77.36	64.15	79.25	75.47	64.15	71.70	84.91
kn-IN	0.53	2.11	5.26	1.05	8.42	1.58	18.42	5.79	61.05	56.32	51.05	53.16	56.84	69.47	72.63	73.68	66.32	73.68	75.79	81.58	82.63
ko-KR	3.12	1.04	22.40	4.17	42.71	2.08	20.83	39.58	54.17	62.50	67.19	69.79	75.00	77.60	83.85	78.65	86.98	89.58	72.92	88.54	96.88
lv-LV	0.52	1.04	12.50	3.65	5.73	12.50	14.06	9.38	24.48	28.65	46.35	46.88	67.71	64.06	71.35	67.71	71.88	76.04	80.73	79.17	85.42
ml-IN	1.06	2.65	5.82	1.06	4.76	1.59	14.81	3.70	57.67	48.15	59.26	60.85	58.73	66.14	71.96	76.72	66.67	80.42	77.25	86.24	91.01
mn-MN	2.62	0.00	6.28	0.00	5.76	2.09	6.28	3.66	49.21	46.07	38.74	39.27	57.59	62.30	70.68	63.87	74.87	72.25	71.20	74.87	85.86
ms-MY	1.57	0.52	12.57	30.37	20.94	39.27	25.65	18.85	24.61	29.84	51.83	53.93	71.20	70.68	67.02	60.73	72.25	74.87	82.20	79.06	82.72
my-MM	0.52	0.00	7.33	0.00	4.19	1.57	3.66	1.05	44.50	43.46	41.88	43.98	47.12	55.50	68.06	60.73	71.73	74.35	46.60	74.87	84.82
nb-NO	3.14	5.24	11.52	29.84	19.37	41.88	16.75	25.65	25.65	25.65	60.21	60.21	70.68	68.59	65.45	58.64	72.25	74.35	82.20	83.25	90.58
nl-NL	5.26	4.74	15.26	28.95	20.53	54.74	30.53	31.05	27.37	37.37	60.00	62.63	70.00	71.58	61.05	70.00	76.32	80.00	88.95	87.89	90.00
pl-PL	4.21	0.00	17.37	30.00	21.58	38.95	30.53	34.21	36.84	44.21	61.05	66.84	66.32	74.74	74.74	78.95	78.42	84.74	89.47	88.42	89.47
pt-PT	5.79	2.11	23.68	31.05	22.63	61.58	30.00	42.11	34.74	42.63	62.63	64.21	67.37	66.84	68.95	65.26	81.05	83.16	86.32	85.26	90.00
ro-RO	4.32	0.54	13.51	23.78	16.22	32.97	23.78	21.62	25.41	34.59	51.35	54.59	67.57	70.81	65.95	72.97	73.51	78.92	84.86	86.49	87.57
ru-RU	2.12	1.06	13.76	0.53	26.98	2.12	29.10	36.51	41.80	50.26	65.08	67.20	69.31	69.84	70.90	84.13	79.37	83.60	87.83	87.30	90.48
sl-SL	1.56	0.00	15.10	23.96	17.19	21.35	9.90	27.60	23.44	33.85	50.00	48.96	68.75	71.35	64.58	68.23	71.35	77.60	81.77	77.60	85.42
sq-AL	3.19	9.04	7.45	9.04	13.30	13.83	10.11	12.23	19.15	26.06	37.77	38.83	54.79	61.17	59.04	53.19	58.51	56.38	71.28	68.09	75.00
sv-SE	2.63	13.16	9.47	22.63	19.47	38.42	26.84	32.11	27.37	31.58	56.32	57.89	66.84	76.32	67.89	63.68	78.42	82.63	90.00	83.68	90.00
sw-KE	2.09	1.05	6.81	16.23	4.19	13.09	12.57	12.04	18.85	20.94	31.94	31.94	54.45	58.12	60.73	62.30	52.36	62.30	73.30	70.68	72.77
ta-IN	1.04	2.60	11.46	2.08	10.42	1.56	19.27	4.69	58.85	52.08	54.17	55.73	63.54	68.75	72.40	80.73	79.69	82.81	71.88	82.81	89.06
te-IN	2.09	1.05	5.24	1.05	3.14	1.57	21.99	5.24	59.16	58.12	57.59	60.21	61.78	73.30	78.53	82.72	72.77	78.01	83.77	84.29	90.58
th-TH	0.52	4.71	13.61	0.00	13.61	1.05	44.50	32.98	57.07	57.59	69.63	73.82	68.06	77.49	82.72	87.96	86.91	86.39	80.63	85.34	89.01
tl-PH	2.60	1.04	6.77	24.48	15.10	27.08	16.67	21.35	20.83	23.96	46.88	47.92	56.77	59.90	58.85	61.98	59.90	70.83	73.44	71.35	79.17
tr-TR	2.12	2.65	17.99	23.81	18.52	30.69	16.93	19.05	34.92	49.74	58.73	61.38	67.72	75.13	68.25	70.90	62.43	83.60	86.77	86.77	89.95
ur-PK	3.16	9.47	16.32	2.63	15.79	4.21	28.42	27.89	57.89	48.42	52.63	57.37	54.21	75.26	76.32	74.21	71.58	81.05	84.74	81.05	91.58
vi-VN	2.13	2.66	2.13	0.53	21.28	3.72	28.19	27.13	36.17	36.70	53.19	55.85	59.04	67.02	70.21	77.66	76.60	78.19	81.91	80.85	87.77
zh-CN	3.72	2.13	12.77	13.30	33.51	28.72	31.38	42.02	47.34	52.66	61.70	65.96	71.81	70.21	76.60	78.19	79.26	85.11	84.57	85.11	88.30
zh-TW	2.11	0.53	16.84	13.68	29.47	26.32	32.63	45.79	55.26	63.68	63.16	63.68	72.63	73.68	81.58	84.74	84.21	87.37	84.74	89.47	91.58

Table 6: BFCL zero-shot prompt on 10k split. FSA for all tested languages.

Language	Llama 3.2 1B	Claude 2.1	Llama 3.2 3B	Command R+	Command R	Mistral 7B	Mixtral 8x7B	Llama 3.1 8B	Claude 3 Haiku	Llama 3.2 11B	Nova Micro	Claude 3.5 Haiku	Claude 3.5 v2 Sonnet	Llama 3.1 70B	Claude 3 Sonnet	Nova Lite	Llama 3.1 405B	Claude 3 Opus	Claude 3.5 Sonnet	Nova Premier	Nova Pro
Average	0.11	0.46	3.91	4.88	5.56	7.86	7.86	11.94	17.13	18.41	25.38	28.42	29.37	30.02	32.20	32.70	32.97	35.90	39.18	40.40	42.65
af-ZA	0.00	0.00	4.79	14.36	9.57	9.57	7.45	10.64	15.96	17.55	29.25	30.85	33.51	30.32	34.58	31.91	31.91	38.83	39.89	46.28	45.21
am-ET	0.00	0.00	0.00	0.00	0.52	0.00	0.00	1.05	12.04	2.62	16.75	19.89	20.94	15.71	20.94	21.47	17.80	29.84	31.94	12.57	33.51
ar-SA	0.00	1.05	1.57	0.00	0.00	8.38	6.28	13.61	13.09	16.75	21.47	27.22	25.66	21.47	28.27	29.84	27.22	34.03	30.37	45.03	34.55
az-AZ	0.00	0.00	0.52	0.52	1.04	5.73	7.29	8.33	15.10	14.06	26.56	26.04	26.04	25.52	30.21	32.29	25.52	31.77	27.60	34.90	41.67
bn-BD	0.00	0.00	3.67	0.00	0.00	5.76	7.33	7.85	14.14	14.66	17.28	19.37	20.42	18.85	23.04	26.70	33.51	29.32	31.94	37.70	35.60
ca-ES	0.00	0.53	3.17	8.47	12.17	10.05	12.70	8.47	12.70	16.40	24.87	30.16	29.10	26.46	37.04	27.51	35.45	38.09	42.86	43.39	44.97
cy-GB	0.00	0.00	0.00	0.52	2.62	2.09	1.57	1.57	9.42	6.28	17.28	18.32	23.04	13.09	26.70	22.51	17.28	32.98	31.41	19.89	37.70
da-DK	0.00	1.06	3.72	11.70	13.30	14.89	10.64	17.02	21.81	21.81	29.79	34.58	36.70	38.30	42.02	41.49	45.75	39.89	46.81	53.72	53.72
de-DE	0.00	0.53	12.17	14.82	14.82	14.29	11.11	20.64	23.81	28.04	33.86	40.21	44.44	42.33	45.50	44.97	48.68	52.91	49.73	61.38	52.91
el-GR	0.00	0.00	1.57	0.00	0.00	9.42	6.81	12.04	14.14	18.85	24.61	28.27	26.70	33.51	33.51	35.60	32.46	43.98	41.36	53.93	45.03
en-US	2.10	1.58	14.21	24.74	34.74	21.58	25.26	13.16	44.74	32.63	48.95	57.37	63.68	52.11	50.00	55.79	45.26	59.47	62.63	63.68	59.47
es-ES	0.53	2.63	8.42	11.58	14.74	12.11	11.58	21.58	17.37	26.32	28.42	35.26	31.58	35.26	40.00	34.74	36.84	40.00	45.79	53.68	44.74
fa-IR	0.00	0.53	2.66	0.00	0.00	10.11	7.45	13.30	20.75	21.81	27.66	28.19	27.13	36.70	35.11	34.58	35.11	31.91	39.89	50.53	51.06
fi-FI	0.00	0.00	1.57	3.67	3.67	3.67	6.28	6.81	9.95	9.95	21.47	22.51	21.47	24.61	25.66	30.37	26.18	23.04	28.27	39.79	37.17
fr-FR	0.00	1.06	8.47	8.47	16.93	10.58	17.99	22.75	20.64	24.34	33.33	37.57	40.21	40.74	41.27	41.27	39.15	38.09	46.56	53.97	49.73
he-IL	0.00	0.00	0.52	0.00	0.52	5.73	4.17	8.33	13.54	14.58	19.79	22.40	20.83	18.23	23.44	25.00	21.88	26.04	28.12	28.65	32.81
hi-IN	0.00	0.00	5.95	0.00	0.00	7.57	11.35	15.68	15.68	24.32	32.43	23.24	21.62	40.00	36.76	40.54	45.41	48.65	47.57	54.60	56.22
hu-HU	0.00	0.52	1.57	2.62	2.09	8.38	6.28	8.38	13.09	16.23	22.51	27.75	29.32	20.94	29.32	28.80	26.70	24.08	33.51	39.27	38.74
hy-AM	0.00	0.00	0.00	0.00	0.52	2.08	5.21	3.65	14.06	10.42	17.71	19.79	19.27	19.79	23.44	23.44	17.71	23.96	24.48	13.54	33.33
id-ID	0.00	0.53	5.32	15.96	14.36	9.57	10.11	16.49	18.62	25.53	30.32	32.45	36.70	34.58	38.30	32.98	38.83	40.96	48.40	47.87	46.28
is-IS	0.00	0.53	1.06	0.53	0.53	5.32	4.79	7.45	12.23	9.04	19.15	24.47	21.28	23.40	29.25	27.13	26.06	31.38	31.38	38.83	41.49
it-IT	0.00	0.00	8.47	18.52	22.22	13.76	16.40	25.93	30.16	31.22	38.62	42.86	42.86	46.56	52.38	46.56	50.27	51.32	58.20	60.85	52.91
ja-JP	0.00	0.00	8.42	4.21	1.05	12.11	8.42	15.79	20.53	23.16	28.42	31.58	17.89	27.89	27.89	39.47	34.21	40.53	32.10	41.58	45.26
jv-ID	0.00	0.00	0.00	2.12	4.23	2.65	3.17	4.23	4.76	10.58	12.70	14.29	23.28	14.82	24.87	17.46	20.11	24.34	30.16	24.87	25.93
ka-GE	0.00	0.52	0.00	0.00	1.04	2.60	3.65	1.04	17.71	15.62	19.27	22.40	27.08	26.56	26.04	31.77	27.60	32.81	38.54	19.79	41.15
km-KH	0.00	0.00	3.77	0.00	5.66	0.00	1.89	9.43	20.76	18.87	18.87	22.64	24.53	39.62	37.74	30.19	35.85	35.85	41.51	35.85	35.85
kn-IN	0.00	0.00	2.10	0.00	0.53	1.05	2.63	5.26	13.16	13.16	16.84	14.74	17.37	14.74	19.47	25.26	17.37	22.11	27.37	17.37	25.26
ko-KR	0.52	0.00	6.25	0.00	1.04	10.94	13.02	21.88	20.31	22.92	24.48	30.21	27.08	35.94	29.17	33.33	36.46	37.50	33.33	33.33	40.10
lv-LV	0.00	0.00	2.08	1.56	1.56	5.73	7.81	6.25	12.50	12.50	25.00	28.12	28.12	19.79	28.65	28.12	27.08	31.77	33.85	43.75	40.62
ml-IN	0.00	0.00	1.06	0.00	0.00	0.00	1.59	9.00	14.82	15.87	20.11	23.28	23.81	20.64	25.40	26.98	17.46	29.10	33.33	20.11	34.39
mn-MN	0.00	0.00	0.00	0.00	1.05	0.00	2.09	3.67	13.09	12.57	15.18	22.51	24.61	16.75	21.47	25.13	20.94	27.75	27.75	27.22	32.98
ms-MY	0.52	0.00	4.19	16.75	14.14	7.85	8.90	15.18	14.14	20.42	27.75	28.80	32.98	31.41	39.27	34.55	32.98	36.13	45.55	47.64	41.36
my-MM	0.00	0.00	0.00	0.00	0.52	0.52	2.62	3.67	18.32	15.18	20.94	26.70	29.32	20.94	31.94	30.89	35.08	41.36	43.98	17.80	37.70
nb-NO	0.00	1.57	4.19	9.42	12.04	7.85	7.33	12.04	16.23	21.99	30.37	30.37	40.84	30.89	35.60	34.55	30.89	32.98	43.98	49.74	46.07
nl-NL	0.00	2.10	7.37	14.21	14.21	10.53	7.90	12.63	18.42	20.00	30.53	31.58	36.84	34.74	37.90	40.00	40.00	43.68	43.16	49.47	48.42
pl-PL	0.53	0.53	4.21	6.84	3.16	9.47	4.74	10.00	10.00	13.68	17.37	20.00	20.00	25.26	21.58	21.58	23.68	24.21	28.42	40.00	28.95
pt-PT	0.00	1.58	6.32	15.26	18.95	15.79	14.74	26.32	21.58	29.47	37.90	43.68	41.58	40.00	45.26	43.16	47.89	44.74	51.58	55.79	53.68
ro-RO	0.54	0.00	3.78	8.65	9.73	11.35	12.43	19.46	22.16	21.62	27.57	38.38	38.92	36.22	41.08	35.13	40.54	32.43	49.73	50.81	50.81
ru-RU	0.00	1.59	3.70	0.00	0.00	14.82	9.52	13.23	16.93	19.58	24.34	26.46	29.63	30.69	31.75	31.75	31.75	32.80	35.98	43.91	40.21
sl-SL	0.00	1.56	2.60	4.69	1.56	9.38	9.90	5.21	13.54	9.90	22.40	23.44	23.44	24.48	27.08	29.17	26.04	26.04	28.65	44.27	40.62
sq-AL	0.00	0.53	2.13	1.60	3.72	5.32	4.79	4.25	12.77	12.23	19.15	21.28	24.47	25.00	26.06	21.81	23.94	23.40	33.51	33.51	34.58
sv-SE	0.00	1.05	9.47	8.95	11.05	15.26	11.05	16.84	21.58	22.63	27.89	37.90	38.42	36.32	38.95	38.42	39.47	42.63	46.84	46.84	50.00
sw-KE	0.00	0.00	1.05	5.76	2.09	2.09	6.81	7.85	10.47	9.95	17.80	20.42	24.08	25.13	27.22	23.04	34.03	27.22	36.65	35.60	34.03
ta-IN	0.00	0.00	2.08	0.00	1.04	0.52	1.56	5.73	18.23	18.75	21.35	20.83	23.96	25.52	22.92	27.08	22.40	28.65	34.90	21.88	34.38
te-IN	0.00	0.00	6.28	0.00	1.57	0.52	2.62	4.19	17.80	23.56	23.56	24.61	28.27	24.61	26.70	34.55	29.84	38.74	38.22	30.37	42.41
th-TH	0.52	0.00	11.00	0.00	0.00	12.57	10.47	26.18	25.13	31.94	32.46	40.31	39.27	54.97	47.64	43.46	58.64	58.64	62.30	43.46	57.59
tl-PH	0.00	1.04	2.60	10.42	5.21	5.21	8.33	9.90	13.54	13.54	30.73	31.77	34.90	33.33	33.33	33.33	32.81	43.75	46.35	47.40	46.35
tr-TR	0.00	0.00	2.12	2.12	3.70	6.88	7.41	10.05	10.05	11.64	25.93	22.22	23.81	26.46	23.28	28.57	28.57	26.46	32.80	32.80	32.27
ur-PK	0.00	0.00	2.63	0.00	1.58	6.32	12.11	12.63	22.63	18.95	30.00	26.32	28.42	37.90	26.32	41.05	43.16	35.79	32.63	48.95	53.16
vi-VN	0.00	1.06	2.13	0.00	1.06	9.57	7.45	13.83	20.21	22.34	25.00	30.85	34.04	43.62	33.51	33.51	48.40	35.64	46.81	47.87	54.25
zh-CN	0.00	0.00	4.79	0.53	1.06	11.70	6.92	17.55	19.68	24.47	28.19	34.58	23.94	30.32	32.45	36.17	37.77	42.55	40.43	42.02	45.75
zh-TW	0.53	0.00	7.37	4.21	2.10	19.47	8.95	32.63	26.84	26.84	35.79	38.95	33.68	47.89	37.37	45.26	44.74	56.84	48.42	52.63	54.74

Table 7: Custom one-shot prompt on 10k split. AST for all tested languages.

Language	Claude 2.1	Llama 3.2 1B	Command R+	Command R	Mistral 7B	Llama 3.2 3B	Mixtral 8x7B	Llama 3.1 8B	Claude 3 Haiku	Llama 3.2 11B	Nova Micro	Claude 3 Opus	Claude 3.5 Haiku	Nova Premier	Llama 3.1 405B	Llama 3.1 70B	Nova Lite	Claude 3 Sonnet	Nova Pro	Claude 3.5 Sonnet	Claude 3.5 v2 Sonnet
Average	1.14	3.30	11.95	14.88	23.75	23.92	27.93	41.33	63.05	65.58	72.87	75.63	78.26	80.84	81.24	81.88	83.72	83.72	87.34	90.66	90.83
af-ZA	0.00	2.66	35.64	28.19	27.66	23.94	27.13	42.02	47.34	61.17	71.28	74.47	80.32	81.91	84.04	80.85	87.23	84.57	89.36	91.49	91.49
am-ET	0.00	0.00	0.00	0.52	0.00	1.57	5.76	1.05	63.35	20.42	51.31	76.44	70.68	52.36	69.11	57.59	67.54	70.68	76.96	87.96	89.01
ar-SA	2.09	2.62	0.52	2.09	29.32	21.99	22.51	59.16	66.49	66.49	74.87	78.01	82.20	86.39	87.43	86.39	83.77	86.39	87.96	91.62	92.15
az-AZ	0.00	2.08	5.21	4.69	23.96	14.58	26.56	34.38	60.94	60.42	71.35	70.83	75.52	79.69	82.29	82.29	85.42	81.25	89.58	91.67	91.15
bn-BD	0.00	2.09	0.00	0.52	22.51	40.31	28.27	43.46	73.30	74.35	74.35	81.15	78.01	79.58	86.39	85.86	85.86	84.29	86.91	91.10	91.10
ca-ES	2.12	4.23	15.34	22.75	25.40	26.46	38.10	47.62	49.21	59.26	77.25	80.42	79.37	82.54	84.13	79.89	82.54	83.60	87.30	89.95	89.95
cy-GB	0.00	2.09	7.33	13.09	7.85	3.14	15.71	5.24	41.36	30.37	47.12	65.97	54.45	50.79	62.30	58.64	66.49	65.45	74.35	79.58	79.58
da-DK	2.66	4.79	27.13	32.98	34.57	24.47	32.98	51.60	60.11	65.43	77.66	69.68	79.79	85.64	88.83	86.17	91.49	87.77	90.43	92.55	95.74
de-DE	2.65	5.82	33.86	31.75	34.39	33.86	40.74	53.44	62.96	77.78	78.31	79.37	83.07	89.42	87.30	85.71	88.89	85.71	88.89	92.06	92.59
el-GR	0.00	1.05	0.00	1.57	30.37	20.42	36.65	60.73	70.68	75.92	76.96	77.49	79.58	93.72	87.96	89.53	86.39	87.43	89.53	93.72	92.67
en-US	3.68	9.47	37.89	71.58	42.11	48.95	49.47	30.53	76.84	81.05	85.26	92.63	88.95	93.16	78.42	87.37	89.47	85.79	92.11	93.68	93.68
es-ES	6.84	5.79	25.79	37.37	28.95	44.21	40.00	59.47	58.95	76.84	77.37	71.58	85.26	92.11	90.00	85.26	87.37	87.37	90.00	92.63	93.68
fa-IR	1.06	3.72	0.53	3.19	33.51	25.53	30.85	46.28	72.87	72.87	78.19	75.53	79.79	84.57	88.30	86.17	82.45	87.23	87.77	90.96	91.49
fi-FI	0.52	2.62	9.95	13.09	24.61	20.42	33.51	32.98	53.93	62.83	71.73	66.49	78.53	88.48	86.39	84.29	86.91	89.01	89.01	93.19	93.19
fr-FR	4.23	3.70	18.52	44.97	30.16	40.21	47.62	65.08	70.90	77.78	79.37	64.55	87.30	90.48	88.36	91.01	91.01	89.95	92.06	92.59	94.71
he-IL	0.52	0.00	0.00	2.08	21.35	8.85	22.92	54.17	69.27	68.75	81.77	78.65	84.38	83.85	88.02	86.98	86.46	89.58	92.19	94.27	94.79
hi-IN	0.00	1.62	0.00	1.08	32.97	30.27	29.73	38.38	72.43	69.73	81.62	85.95	87.03	79.46	88.65	92.43	91.35	85.41	89.19	94.59	94.59
hu-HU	1.57	2.62	7.85	6.28	26.18	20.42	30.89	40.31	56.54	65.45	71.20	67.54	78.53	84.82	83.77	76.96	83.25	82.72	89.01	89.01	90.05
hy-AM	0.00	0.00	0.00	2.08	10.42	2.08	18.23	18.23	68.23	63.54	70.31	74.48	74.48	65.10	79.69	81.77	82.81	79.17	86.98	90.10	90.62
id-ID	1.06	4.26	42.55	40.43	30.32	37.77	29.79	53.19	59.04	73.40	78.19	80.32	79.79	85.11	85.64	84.04	85.64	86.17	92.55	93.09	91.49
is-IS	0.53	1.60	3.19	3.72	12.77	11.17	22.87	28.72	49.47	42.02	63.83	70.21	73.40	79.26	79.79	68.09	78.72	78.19	82.45	89.36	90.43
it-IT	0.00	10.05	35.45	46.03	32.80	38.62	39.68	69.84	65.61	83.60	85.71	82.01	88.89	96.30	94.71	92.59	93.65	94.18	93.12	98.41	97.35
ja-JP	0.00	5.79	8.95	7.89	28.95	33.16	30.00	56.32	77.37	77.89	81.58	83.68	87.89	79.47	85.26	83.68	85.79	86.32	91.05	91.05	91.05
jv-ID	0.00	2.12	8.99	15.87	10.05	7.94	14.81	11.64	18.52	33.86	41.80	48.15	31.75	51.85	56.08	50.26	55.56	57.67	58.20	70.37	68.78
ka-GE	0.52	0.00	0.00	1.56	9.38	2.60	18.75	10.94	69.79	60.42	68.75	76.56	68.75	68.75	81.25	81.25	81.77	82.29	87.50	93.75	94.27
km-KH	0.00	5.66	0.00	9.43	5.66	13.21	22.64	22.64	69.81	71.70	66.04	84.91	54.72	71.70	83.02	83.02	77.36	81.13	83.02	86.79	86.79
kn-IN	0.00	0.53	0.53	2.63	6.84	25.79	13.68	24.74	66.32	70.53	68.95	78.42	71.05	78.42	82.11	80.00	84.21	76.32	86.32	86.84	85.79
ko-KR	0.00	4.69	0.52	3.65	38.54	28.12	35.42	61.98	82.29	76.56	84.38	89.58	91.67	77.08	93.75	88.54	89.58	89.06	94.27	94.27	96.88
lv-LV	0.00	2.08	2.60	2.60	17.71	10.94	29.69	16.15	60.94	57.29	69.79	78.12	78.65	82.81	79.69	73.96	84.38	84.38	86.46	91.15	92.71
ml-IN	0.53	0.00	0.00	1.59	1.06	11.64	9.52	29.10	75.13	69.84	76.19	84.66	77.25	76.72	86.77	78.84	84.13	81.48	88.89	89.95	92.06
mn-MN	0.00	0.00	0.00	2.62	5.24	2.62	15.71	25.65	57.07	39.79	57.59	81.15	72.25	72.25	75.39	67.02	75.39	77.49	86.39	89.53	87.96
ms-MY	0.52	3.14	39.27	39.27	26.70	29.84	27.75	56.54	58.64	71.20	71.20	71.20	73.82	80.63	86.91	82.20	83.77	88.48	85.86	89.01	86.91
my-MM	0.00	0.00	0.52	1.57	1.57	1.57	13.61	20.94	67.54	51.31	61.26	80.10	69.63	58.64	79.58	67.54	71.20	75.92	82.72	89.53	90.05
nb-NO	2.09	2.62	24.08	30.89	29.32	25.65	30.89	50.26	57.59	67.54	75.92	72.25	81.68	85.86	84.82	84.29	82.72	84.29	86.39	92.15	93.19
nl-NL	3.68	4.74	34.74	40.00	30.00	33.16	27.37	43.68	59.47	72.63	78.95	80.53	83.68	87.37	88.95	82.63	88.95	90.00	88.42	93.16	93.68
pl-PL	2.11	4.74	21.58	21.58	40.53	36.32	33.68	48.95	63.68	74.74	80.00	65.79	87.37	92.11	88.95	90.53	90.53	87.89	93.16	93.68	93.16
pt-PT	3.68	6.32	28.42	38.42	34.74	36.84	42.11	68.42	63.16	78.42	81.05	79.47	86.32	90.00	91.58	87.89	88.95	91.58	92.63	93.68	93.68
ro-RO	1.62	4.86	22.70	27.57	27.57	28.65	32.43	51.89	60.54	67.03	70.81	65.95	83.24	89.73	88.65	81.62	83.24	89.73	90.81	93.51	92.97
ru-RU	1.59	4.76	0.00	1.06	41.80	35.45	40.21	69.84	66.67	80.42	76.19	67.72	83.60	82.01	91.01	90.48	87.30	84.66	89.95	94.18	94.71
sl-SL	4.69	2.08	18.23	14.06	29.17	17.19	38.02	38.54	53.65	54.17	70.83	59.90	78.65	85.42	83.33	78.65	80.21	81.77	88.54	87.50	89.06
sq-AL	0.53	2.66	2.66	9.04	13.30	11.70	22.87	12.23	51.60	46.28	61.17	57.98	67.02	73.40	72.87	71.28	71.28	72.34	76.60	80.32	81.38
sv-SE	1.05	3.16	22.63	23.68	35.79	31.58	34.21	55.26	66.32	75.79	77.89	78.95	88.95	90.53	90.00	90.53	90.53	88.95	91.58	94.74	94.21
sw-KE	0.00	2.09	17.28	7.85	9.42	12.04	19.37	34.03	36.65	43.46	52.36	57.07	62.83	73.30	82.20	73.30	69.63	73.30	76.44	81.15	81.15
ta-IN	0.00	0.52	0.00	2.60	4.69	27.60	9.38	16.67	71.88	67.19	76.04	81.77	77.08	77.08	83.33	84.38	81.77	79.17	86.98	92.71	91.67
te-IN	0.00	2.09	1.57	1.57	4.19	27.75	8.90	9.42	75.39	71.20	72.25	86.39	76.96	78.01	85.86	85.86	88.48	84.82	85.86	91.62	90.58
th-TH	0.00	4.19	0.52	2.09	28.80	32.98	31.41	51.83	77.49	76.44	78.53	89.53	88.48	80.10	89.01	91.10	89.01	92.15	90.58	94.24	91.62
tl-PH	3.65	3.65	25.52	25.00	17.19	16.67	29.69	41.15	38.02	55.21	72.92	73.44	72.40	83.33	74.48	78.65	80.73	80.73	84.38	85.94	89.06
tr-TR	1.59	4.23	15.34	11.64	32.80	25.93	27.51	53.44	62.96	69.84	79.37	67.72	83.60	86.77	87.83	86.77	88.89	86.24	88.89	92.06	92.06
ur-PK	0.00	3.16	1.58	3.16	26.84	42.63	32.11	51.05	68.95	78.42	76.84	82.11	80.00	88.42	92.11	88.42	90.53	87.37	93.68	91.58	92.11
vi-VN	1.06	5.32	3.72	2.13	32.45	29.26	29.26	40.96	71.28	64.89	72.34	68.09	78.19	83.51	82.98	85.11	84.04	85.64	86.70	90.96	89.89
zh-CN	0.53	6.91	3.72	5.85	37.23	32.45	25.53	63.30	80.85	77.66	80.32	86.17	90.43	85.11	87.23	88.30	87.23	88.30	90.43	90.96	89.89
zh-TW	0.53	6.84	8.95	6.84	45.26	33.16	35.79	75.79	79.47	78.95	83.16	91.58	92.11	88.42	93.16	91.58	91.58	92.11	91.05	94.21	94.21

Table 8: Custom one-shot prompt on 10k split. FSA for all tested languages.

Language	Claude 2.1	Mixtral 8x7B	Nova Micro	Mistral 7B	Claude 3 Haiku	Llama 3.1 8B	Claude 3 Opus	Llama 3.1 70B	Claude 3 Sonnet	Nova Pro	Nova Lite	Claude 3.5 Haiku	Claude 3.5 Sonnet	Nova Premier
Average	0.58	3.83	5.58	7.01	7.73	9.40	13.02	19.29	21.29	25.16	25.94	27.18	28.55	37.93
af-ZA	1.04	5.03	7.50	6.64	5.98	10.15	10.25	19.92	24.29	30.27	29.22	28.94	34.06	46.77
am-ET	0.00	0.00	1.57	0.39	7.87	1.97	8.66	10.24	14.17	17.72	20.08	20.47	24.02	6.30
ar-SA	0.17	3.82	3.49	6.98	5.65	6.15	12.13	15.95	20.76	24.25	22.43	25.41	28.24	38.21
az-AZ	0.76	1.83	3.96	1.37	5.48	3.50	8.83	16.89	15.37	16.44	21.77	18.11	19.63	30.59
bn-BD	0.50	2.65	3.80	6.45	6.28	6.45	12.89	14.71	17.69	18.18	21.32	20.83	22.48	32.40
ca-ES	0.78	5.65	6.33	10.04	7.90	8.09	9.36	18.52	26.80	27.58	24.46	29.73	30.51	43.96
cy-GB	0.00	1.71	2.61	1.22	4.24	5.30	6.19	10.19	18.01	23.80	24.20	18.58	22.00	27.79
da-DK	0.91	5.54	6.45	10.89	5.75	10.28	7.76	21.37	26.71	31.45	30.75	29.13	29.64	46.77
de-DE	0.68	6.92	10.72	15.33	9.90	14.65	19.54	29.85	28.90	29.17	32.02	37.45	35.55	58.34
el-GR	0.63	5.37	2.42	7.47	7.79	10.00	16.42	18.11	20.00	24.53	25.58	26.11	32.53	45.26
en-US	0.10	9.63	15.42	18.80	12.40	19.62	15.21	34.73	28.38	45.65	41.55	40.27	36.94	56.87
es-ES	0.44	5.80	9.20	13.58	8.87	10.73	12.27	23.77	28.37	29.79	29.13	34.50	36.80	48.63
fa-IR	0.96	4.31	3.35	9.20	9.77	9.48	18.68	22.03	22.61	25.10	32.47	28.54	28.54	49.81
fi-FI	0.00	3.68	6.80	4.23	4.96	4.96	7.17	14.71	18.29	20.50	21.60	21.41	22.34	34.84
fr-FR	0.72	6.79	7.82	12.24	10.49	12.45	12.65	25.41	27.26	32.10	30.76	36.21	37.35	51.95
he-IL	0.16	3.20	2.05	7.14	5.50	6.32	10.35	14.12	19.05	26.44	25.53	23.64	23.89	34.73
hi-IN	2.16	1.08	9.19	5.41	10.81	14.59	17.84	24.32	22.70	24.87	34.05	27.03	24.87	58.12
hu-HU	1.32	4.55	5.06	6.17	7.08	4.96	5.67	14.78	19.03	20.95	21.76	23.68	23.58	35.53
hy-AM	0.00	0.45	2.24	0.60	4.92	4.48	11.34	11.79	15.52	20.00	18.06	19.11	20.00	12.39
id-ID	0.15	6.52	6.97	9.51	8.02	12.96	11.24	22.40	24.20	32.66	31.24	34.68	33.93	48.76
is-IS	0.11	1.82	4.89	4.66	3.75	6.37	7.62	14.34	19.79	23.78	23.44	22.07	25.14	35.15
it-IT	1.00	6.48	9.97	11.13	11.79	15.45	16.28	31.23	31.73	34.22	32.23	39.04	38.04	57.14
ja-JP	0.65	6.49	9.36	11.03	10.56	11.77	20.02	20.95	22.61	30.31	30.12	29.94	28.45	44.76
jv-ID	0.00	2.07	3.10	2.67	3.45	7.15	4.13	15.07	13.69	21.79	17.31	17.05	20.41	32.30
ka-GE	0.00	0.68	3.12	2.31	5.44	4.75	11.68	12.09	18.75	21.20	20.24	21.74	28.53	18.48
km-KH	0.00	0.00	0.00	1.89	11.32	16.98	18.87	13.21	18.87	26.41	13.21	24.53	32.08	15.09
kn-IN	0.68	2.09	2.77	2.04	6.79	6.11	17.65	10.75	15.78	18.44	21.72	19.06	22.79	19.23
ko-KR	0.36	8.45	9.76	15.00	11.31	11.43	24.05	22.74	26.31	29.88	33.93	33.69	34.52	33.33
lv-LV	0.22	1.56	6.03	3.12	4.69	5.80	6.14	11.94	18.64	22.66	21.76	24.78	25.22	36.61
ml-IN	0.47	0.55	2.66	0.63	6.73	5.33	13.39	10.26	16.84	15.50	22.40	19.58	20.05	17.15
mn-MN	0.00	0.73	1.96	0.49	5.87	2.69	11.00	8.56	14.18	15.40	16.63	19.80	20.54	26.65
ms-MY	0.24	3.44	4.72	7.43	6.08	14.15	10.79	21.02	21.10	29.90	27.82	30.22	31.18	46.76
my-MM	0.00	0.61	2.20	0.24	8.43	7.20	19.54	16.12	19.66	18.19	23.93	24.54	29.79	10.01
nb-NO	1.39	4.52	4.17	9.27	7.76	12.63	10.89	19.35	24.68	29.55	28.51	30.13	30.71	49.25
nl-NL	1.29	6.77	6.27	11.84	8.56	11.14	12.64	23.98	26.27	31.24	29.65	33.83	34.53	51.94
pl-PL	0.00	3.83	7.65	7.80	7.29	8.24	8.39	17.59	18.98	21.19	23.33	24.95	25.09	39.59
pt-PT	0.93	5.41	12.31	13.25	10.26	15.30	17.35	27.43	26.68	30.78	28.92	36.38	36.19	53.73
ro-RO	0.22	5.09	7.96	7.85	8.85	10.95	9.74	24.23	22.46	28.65	27.66	33.41	31.30	46.68
ru-RU	0.00	4.64	4.56	11.14	6.25	5.74	10.29	18.40	18.65	22.62	20.25	23.04	24.22	40.67
sl-SL	0.09	4.90	6.59	6.51	4.01	4.99	7.49	13.01	16.67	22.01	21.48	22.82	21.57	36.45
sq-AL	1.77	2.77	4.69	2.84	4.61	6.23	7.15	12.99	16.99	20.45	19.75	21.52	23.52	39.05
sv-SE	3.11	7.19	4.71	11.90	7.81	12.88	10.48	23.98	28.06	31.53	32.15	34.64	35.08	52.40
sw-KE	0.17	2.32	3.70	2.07	3.27	7.14	7.49	14.89	19.36	21.34	22.46	20.74	22.98	38.90
ta-IN	0.83	0.50	3.64	1.65	10.08	7.11	17.52	16.20	19.17	18.18	24.79	23.80	26.12	17.69
te-IN	0.00	1.46	4.39	1.67	9.21	10.04	19.25	17.16	16.53	21.76	24.69	23.85	28.87	23.85
th-TH	1.98	1.98	6.37	8.35	14.72	17.80	27.91	39.34	27.03	30.11	33.63	33.85	45.93	37.36
tl-PH	0.00	3.23	3.88	3.77	3.66	10.98	8.07	17.76	21.64	23.25	22.71	27.34	25.40	40.90
tr-TR	0.60	3.87	8.60	5.67	9.72	8.60	10.83	21.15	20.55	24.59	29.92	27.86	24.42	40.59
ur-PK	1.61	2.82	8.06	8.87	10.08	8.87	19.36	19.76	18.55	18.95	32.66	27.82	19.36	37.10
vi-VN	0.23	4.14	2.07	8.28	8.28	11.49	15.86	26.67	17.70	24.83	23.91	28.05	31.04	38.85
zh-CN	0.43	6.67	6.02	11.81	8.91	13.12	17.97	25.87	24.49	30.80	28.77	34.35	35.73	44.35
zh-TW	0.19	7.43	7.24	13.71	12.95	13.33	22.86	31.43	26.48	27.24	30.86	35.05	38.67	42.29

Table 9: BFCL zero-shot prompt on full split. AST for all tested languages.

Language	Claude 2.1	Nova Micro	Mixtral 8x7B	Mistral 7B	Claude 3 Opus	Claude 3 Haiku	Llama 3.1 8B	Nova Pro	Nova Lite	Llama 3.1 70B	Claude 3 Sonnet	Claude 3.5 Sonnet	Claude 3.5 Haiku	Nova Premier
Average	2.10	11.24	16.91	23.45	35.80	39.41	52.00	64.60	68.13	69.56	69.72	71.29	77.38	80.03
af-ZA	1.99	13.66	13.57	17.65	20.02	22.39	44.50	64.71	66.60	66.60	58.35	72.39	71.73	78.65
am-ET	0.00	2.76	1.57	0.79	33.86	44.49	23.23	55.12	55.12	51.97	68.90	64.17	75.20	56.30
ar-SA	2.16	9.14	16.78	25.75	34.39	35.22	45.35	70.43	64.45	70.10	71.43	72.59	75.91	80.56
az-AZ	5.02	9.74	12.48	10.81	33.03	41.70	52.05	57.08	67.73	67.28	64.84	68.19	72.15	80.37
bn-BD	1.32	7.44	21.16	20.17	45.29	40.99	59.34	63.80	72.07	79.83	74.38	67.11	79.17	83.64
ca-ES	4.29	10.14	22.32	33.24	23.00	30.02	49.51	68.91	61.79	67.64	69.59	71.54	75.34	81.87
cy-GB	0.24	5.30	6.28	6.44	18.01	25.92	27.14	60.72	63.90	53.87	60.88	57.70	57.95	58.76
da-DK	2.62	13.81	21.47	29.13	21.47	27.12	49.19	66.53	69.25	60.28	64.72	66.13	71.98	81.75
de-DE	1.49	21.30	24.29	37.99	36.64	35.14	62.96	62.82	70.28	78.02	69.06	77.20	83.45	90.09
el-GR	3.68	3.89	23.05	25.68	46.84	46.00	59.37	71.68	72.32	70.63	72.84	81.89	83.16	88.42
en-US	0.10	30.84	29.97	49.18	27.46	32.79	66.39	81.56	79.30	77.15	58.50	63.93	83.40	89.55
es-ES	1.86	17.42	29.13	45.56	27.16	39.87	56.96	71.63	69.33	63.42	72.84	79.96	81.93	88.39
fa-IR	2.97	5.17	23.37	31.70	47.80	46.36	55.17	63.22	74.71	70.88	77.11	72.22	81.32	83.72
fi-FI	0.09	19.12	21.32	20.86	27.39	37.50	47.15	65.07	71.60	68.11	73.53	71.78	79.60	85.11
fr-FR	2.26	15.84	23.46	38.68	29.94	38.48	59.47	75.10	73.77	78.40	67.08	79.73	82.00	88.07
he-IL	0.57	3.37	17.16	25.45	36.12	37.93	54.84	73.56	74.88	75.21	75.45	73.32	78.98	83.74
hi-IN	8.11	22.16	15.14	19.46	57.30	47.57	61.62	63.24	67.03	86.49	83.78	70.27	85.41	87.03
hu-HU	5.77	12.55	17.61	21.05	19.84	34.82	48.68	60.83	67.00	60.22	70.34	71.66	75.91	82.79
hy-AM	0.45	5.97	6.12	4.48	40.45	42.24	51.04	71.64	70.15	71.49	76.27	71.04	78.06	65.37
id-ID	0.52	14.98	24.42	28.91	26.37	33.26	53.41	70.79	70.41	62.55	62.17	73.48	78.35	81.27
is-IS	0.46	9.33	10.81	9.67	22.18	30.15	35.27	58.13	63.14	60.86	65.98	67.69	72.70	80.55
it-IT	3.65	18.27	16.11	35.55	35.55	37.04	68.60	70.76	71.10	73.59	71.10	76.08	81.23	91.20
ja-JP	1.30	20.11	28.17	32.72	48.19	56.53	58.20	69.51	67.47	78.59	78.04	69.42	84.99	83.32
jv-ID	0.00	6.55	7.67	8.79	12.83	18.43	33.76	50.65	44.88	48.49	44.88	49.53	46.43	59.43
ka-GE	0.27	6.39	8.56	6.52	43.21	36.82	50.14	64.40	67.93	79.21	72.96	73.10	75.68	68.89
km-KH	0.00	0.00	3.77	7.55	64.15	60.38	64.15	69.81	67.92	67.92	77.36	77.36	77.36	66.04
kn-IN	1.19	5.26	7.18	8.20	59.73	52.43	52.38	56.11	67.19	72.85	70.98	68.55	75.00	77.09
ko-KR	0.83	18.57	38.57	39.52	53.33	62.86	58.93	69.64	79.29	78.21	81.43	84.29	87.38	76.19
lv-LV	1.00	11.94	9.38	12.17	23.10	30.13	39.62	68.30	66.07	69.87	71.65	73.55	76.00	81.25
ml-IN	2.04	6.50	4.39	5.17	61.55	54.74	52.86	57.48	70.79	77.37	77.06	63.12	79.56	78.94
mn-MN	0.00	5.38	3.42	2.44	48.66	44.50	34.23	55.01	61.86	65.04	71.15	69.93	72.13	73.35
ms-MY	0.56	9.75	17.51	25.82	25.50	28.78	50.36	68.67	65.47	61.31	57.95	68.67	73.38	80.74
my-MM	0.00	7.20	2.81	1.34	46.89	42.86	38.58	46.64	62.27	62.03	71.31	71.67	71.79	53.85
nb-NO	3.94	8.34	17.96	24.33	24.22	28.74	57.24	67.09	67.21	59.21	65.47	65.70	75.67	85.40
nl-NL	3.28	13.03	23.38	35.52	28.56	33.83	58.91	71.94	73.63	70.95	61.59	77.91	83.28	88.36
pl-PL	0.22	17.14	23.40	35.32	31.27	42.46	59.31	69.02	74.32	75.64	73.29	77.85	84.55	88.52
pt-PT	2.24	21.27	24.44	41.98	33.02	36.75	59.70	67.16	66.79	62.13	67.35	74.81	80.41	84.70
ro-RO	0.44	12.50	17.70	27.99	21.35	33.08	52.77	68.81	71.35	70.58	66.04	74.00	79.42	86.06
ru-RU	1.35	8.78	24.30	39.66	38.48	44.39	62.87	69.79	69.87	79.32	70.38	75.95	81.77	87.68
sl-SL	0.09	15.33	21.57	27.90	22.01	30.48	42.34	64.71	67.74	65.24	64.08	68.18	75.58	82.17
sq-AL	8.69	7.61	13.53	9.30	18.68	25.83	36.20	54.50	58.19	54.65	59.80	58.80	63.26	74.56
sv-SE	9.06	9.41	23.27	35.70	25.04	31.17	58.70	67.67	73.53	63.94	68.12	73.80	83.04	88.72
sw-KE	0.43	6.54	7.14	8.86	18.16	18.42	33.48	53.53	56.11	59.38	56.02	54.30	58.78	72.03
ta-IN	1.98	6.94	7.60	6.94	53.72	51.74	52.73	59.01	66.45	78.35	72.56	71.07	79.01	74.55
te-IN	2.30	5.44	3.77	6.28	50.42	53.97	54.81	58.16	67.99	79.92	74.90	70.29	78.87	80.13
th-TH	3.30	9.89	14.07	34.51	55.16	53.85	68.57	64.84	73.85	84.62	82.42	86.15	87.69	82.42
tl-PH	0.75	8.93	14.32	17.44	19.38	20.56	47.15	59.20	58.56	56.73	63.40	61.46	70.94	75.24
tr-TR	1.89	18.23	17.28	23.22	31.30	47.12	55.98	66.90	70.59	69.65	69.30	64.75	82.46	85.38
ur-PK	7.26	14.11	15.73	33.87	52.82	46.37	55.65	48.79	76.21	71.77	74.60	70.56	80.65	87.10
vi-VN	2.76	2.53	20.23	32.18	44.37	41.84	60.46	66.21	72.18	76.78	71.72	78.62	79.54	83.91
zh-CN	1.38	13.12	30.22	43.33	42.61	54.57	61.59	69.86	67.61	76.52	77.54	78.77	85.58	83.62
zh-TW	0.95	15.62	30.48	46.86	53.90	58.48	61.33	68.57	73.71	86.48	83.05	84.76	88.76	84.76

Table 10: BFCL zero-shot prompt on full split. FSA for all tested languages.

Figure 7: AST performance changes (percentage points) for prompt translation experiments with Nova Micro, Claude 3.5 Haiku, Nova Pro, and Claude 3.5 v2 Sonnet models. Heatmaps show relative changes from the English-only baseline for four models across different prompt translation strategies and evaluation languages. Green/red indicates improvements/degradations, respectively. All variants are derived from the BFCL zero-shot template.

Figure 8: AST performance changes (percentage points) for prompt translation experiments with Llama 3.1 8B, Claude 3 Haiku, and Claude 3.5 Sonnet models. Heatmaps show relative changes from the English-only baseline for four models across different prompt translation strategies and evaluation languages. Green/red indicates improvements/degradations, respectively. All variants are derived from the BFCL zero-shot template.

Figure 9: FSA performance changes (percentage points) for prompt translation experiments with Nova Micro, Claude 3.5 Haiku, Nova Pro, and Claude 3.5 v2 Sonnet models. Heatmaps show relative changes from the English-only baseline for four models across different prompt translation strategies and evaluation languages. Green/red indicates improvements/degradations, respectively. All variants are derived from the BFCL zero-shot template.

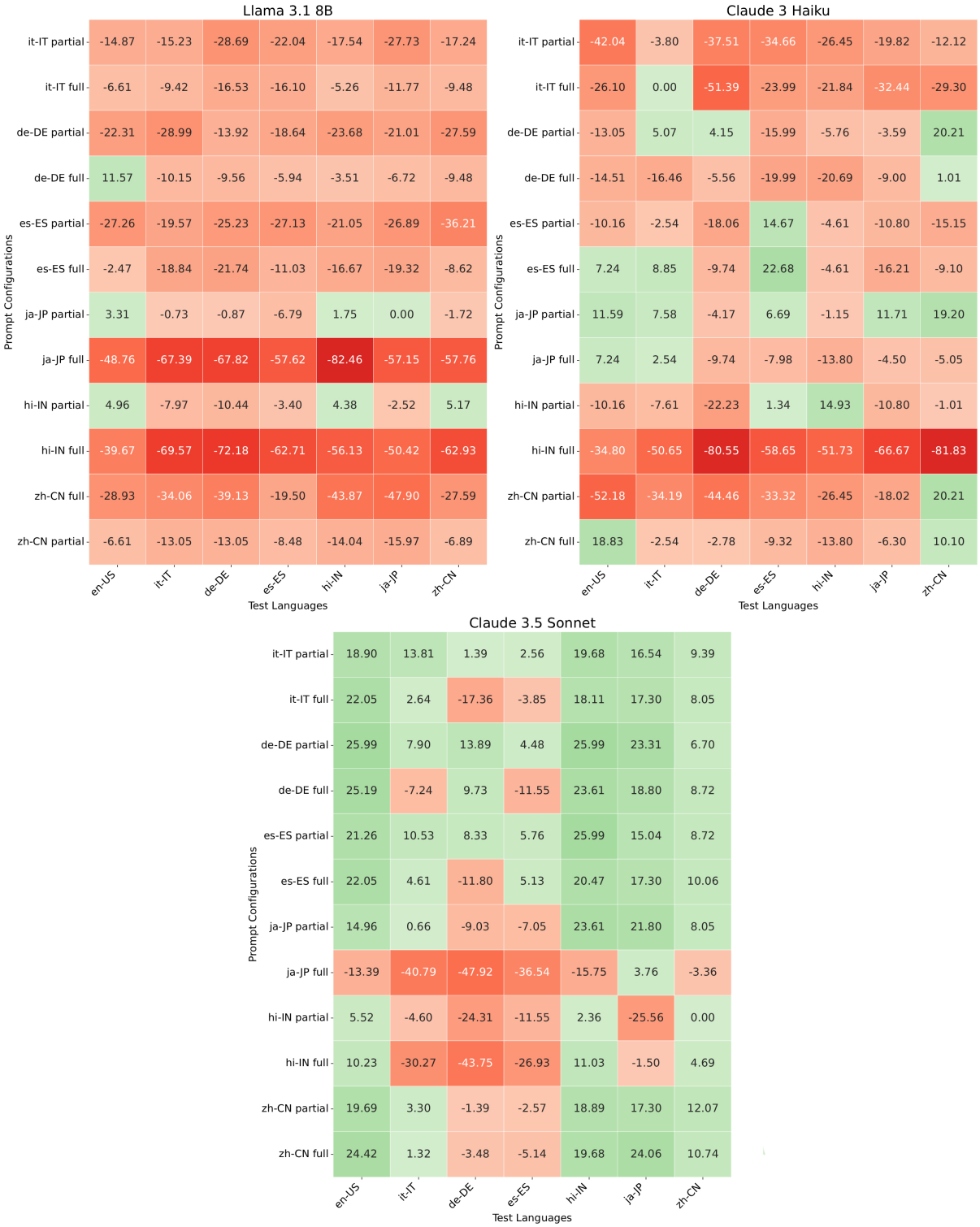

Figure 10: FSA performance changes (percentage points) for prompt translation experiments with Llama 3.1 8B, Claude 3 Haiku, and Claude 3.5 Sonnet models. Heatmaps show relative changes from the English-only baseline for four models across different prompt translation strategies and evaluation languages. Green/red indicates improvements/degradations, respectively. All variants are derived from the BFCL zero-shot template.