# Graphical Abstract

**Enhancing Zero-Shot Multilingual Semantic Parsing: A Framework Leveraging Large Language Models for Data Augmentation and Advanced Prompting Techniques**

Dinh-Truong Do, Minh-Phuong Nguyen, Le-Minh Nguyen

# Highlights

**Enhancing Zero-Shot Multilingual Semantic Parsing: A Framework Leveraging Large Language Models for Data Augmentation and Advanced Prompting Techniques**

Dinh-Truong Do, Minh-Phuong Nguyen, Le-Minh Nguyen

- In this paper, we present a comprehensive framework that leverages the power of large language models to augment multilingual data. Our framework introduces three novel multilingual semantic parsing chain-of-thought (CoT) prompting techniques that incrementally guide LLMs to step-by-step parse complex logical structures in new languages.
- We also propose a method for selecting a small and generalized subset of annotated data. This enables us to achieve impressive results using significantly less data, reducing annotation costs and effort.
- To validate the effectiveness of our framework, we conducted extensive experiments on two widely recognized multilingual semantic parsing datasets, MTOP and MASSIVE. Our results showcase new state-of-the-art performance on zero-shot-semantic parsing tasks for these datasets.

# Enhancing Zero-Shot Multilingual Semantic Parsing: A Framework Leveraging Large Language Models for Data Augmentation and Advanced Prompting Techniques

Dinh-Truong Do[a], Minh-Phuong Nguyen[a], Le-Minh Nguyen[a]

[a]*Japan Advanced Institute of Science and Technology, Japan*

**Abstract**

In recent years, significant progress has been made in semantic parsing tasks due to the introduction of pre-trained language models. However, there remains a notable gap between English and other languages because of the limited availability of annotated data. One promising approach to bridge this gap is to augment multilingual datasets using labeled English data and then train semantic parsers with this enhanced dataset (known as zero-shot multilingual semantic parsing). In this study, we propose a novel framework for the zero-shot multilingual semantic parsing task through LLM-driven data augmentation. Our approach leverages annotated English data — consisting of sentences and their corresponding semantic representations — to generate augmented data in target languages. This is achieved through the implementation of multilingual chain-of-thought (CoT) prompting techniques that incrementally construct semantic forms. By deconstructing complex semantic structures into sub-fragments, our framework guides LLMs to progressively construct accurate semantic forms in target languages without the need for target language demonstration examples, enabling zero-shot learning. We demonstrate the effectiveness of our framework on two benchmark datasets, MTOP and MASSIVE, achieving state-of-the-art (SOTA) performance in zero-shot multilingual semantic parsing. Additionally, we analyze the effectiveness of the proposed framework under stricter constraints, where only limited annotated data in English is available. Our experiments indicate that our framework, trained using 600 samples of English data ( 5% of the full dataset), can achieve 80% of the performance of a system trained on the full dataset. This demonstrates the effectiveness of our proposed framework under conditions of limited annotated data.

*Keywords:* Zero-shot Multilingual Semantic Parsing, Large Language Model, Chain-of-thought Prompting

## 1. Introduction

Semantic parsing has been a significant research area in NLP for decades. It involves creating semantic parsers that understand user intentions, *e.g., create an alarm*, and identify entities within those intentions, *e.g., date time* (Do et al., 2023a; Mansimov and Zhang, 2022). These parsers are used for various purposes, such as creating virtual assistants that can understand and respond to user commands (Campagna et al., 2019) or generating computer code from natural language instructions (Li et al., 2022). One exciting development in this area is the creation of multilingual semantic parsers. These parsers aim to understand multiple languages, eliminating the need for separate parsers for each language. The emergence of multilingual pre-trained language models has significantly boosted the effectiveness of multilingual semantic parsers (Conneau et al., 2020; Xue et al., 2021; Muennighoff et al., 2023). These pre-trained language models provide a strong foundation for understanding meaning across diverse languages. However, fine-tuning these models to create multilingual semantic parsers often requires substantial amounts of annotated data, which can be labor-intensive and challenging to acquire, particularly for a wide range of languages. Existing semantic parsing datasets are predominantly English-focused, posing a significant challenge for researchers aiming to develop techniques that function across diverse languages with limited resources. Addressing this issue is essential for advancing multilingual NLP capabilities.

Overcoming the lack of annotated data for multilingual semantic parsing has been a major challenge and attracted significant research interest (Gritta et al., 2022; Nicosia et al., 2021). Based on the characteristics of the proposed methods, prior studies can be divided into two main strategies: (1) exploiting English data alone with a cross-lingual objective when training to boost multilingual capabilities (Yang et al., 2021; Sherborne and Lapata, 2022), and (2) augmenting multilingual datasets of target languages and then training a multilingual parser on augmented data (Nicosia and Piccinno, 2022; Awasthi et al., 2023). This paper follows the seconds strategy, using English-augmented datasets to enhance a multilingual semantic parser, addressing data scarcity and leveraging augmented datasets' strengths.

Recent advancements in Large Language Models (LLMs) have demonstrated their capability in following instructions and enhancing data through augmentation techniques (Touvron et al., 2023; Wu et al., 2023b). These models have also shown promise in multilingual semantic parsing. (Awasthi et al., 2023) illustrated that with just a few carefully chosen examples in the target language, it's possible to augment other input-semantic representation pairs. This approach taps into the strength of in-context learning (Raventós et al., 2024). Despite these advancements, the potential of LLMs to enhance zero-shot multilingual semantic parsing—where parsing is done without any specific demonstrations in the target language—remains largely untapped. This is because LLMs often struggle to generate novel semantic structures in a zero-shot manner, as they haven't encountered such representations during pre-training. Further research into refining prompts with LLMs and drawing on their natural language abilities could unlock their potential for zero-shot transfer in multilingual semantic parsing.

In this paper, we introduce the Zero-MParser framework, designed to improve zero-shot multilingual semantic parsing. This framework leverages the cross-lingual capabilities of LLMs to expand data from existing English data into new languages and integrates the state-of-the-art grammar-enhanced recursive insertion-based semantic parsing method (Do et al., 2023b) to build the multilingual semantic parser. Specifically, our framework consists of three key phases: LLM-based augmentation, noisy data filtering, and multilingual semantic parsing. First, in the LLM-based augmentation, we translate English utterances into the target languages using off-the-shelf translation tools. We then propose three novel multilingual semantic parsing CoT prompting techniques that incrementally guide LLMs to step-by-step parse complex logical structures in new languages. This creates a multilingual dataset for training. For example, given an English sentence and its semantic form, the sentence is translated into German (Figure 1). The LLM then generates a "silver" German semantic representation, e.g., `"[IN:CREATE_CALL [SL:ORDINAL dritten] [SL:CONTACT Richard]]"`. Notably, unlike past approaches (Awasthi et al., 2023) that rely on a few examples in new languages, we use only the translated target utterance to steer LLMs to predict the semantic form in the target language. In the second phase, the noisy data being generated in the first phase can hurt the performance of parsers; therefore, we proposed a filtering method to filter noisy data in the generated instances to keep only high-quality augmented data. In the final phase, multilingual semantic parsing, we employ the grammar-enhanced

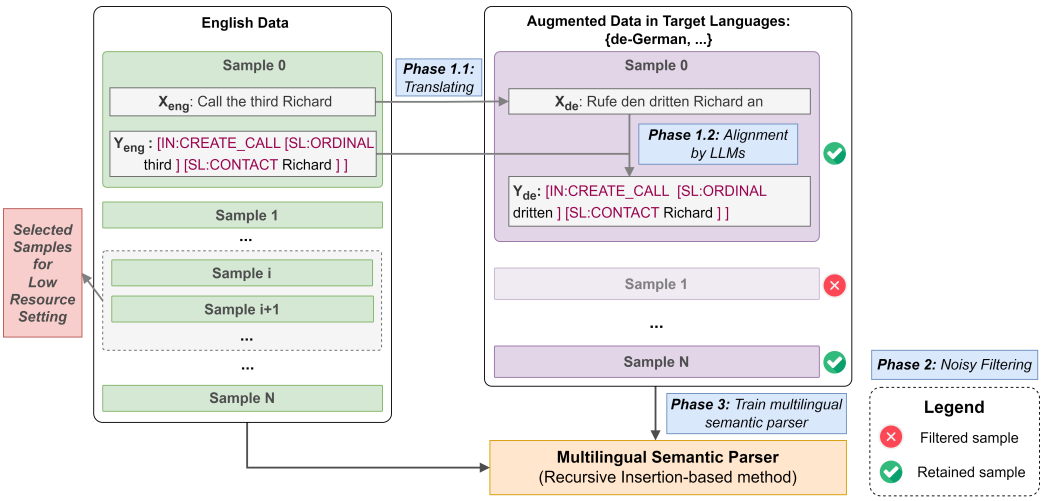


Figure 1: Our framework utilizes LLMs to augment data in the target languages for zero-shot multilingual semantic parsing.

recursive insertion-based model learned from English data and augmented data. Experiments on established multilingual semantic parsing datasets show Zero-MParser's efficacy. In zero-shot multilingual semantic parsing, it achieves SOTA exact match scores on the MTOP dataset, exceeding leading methods by 2.2 points. Especially no human effort requirements as compared to previous approaches (Awasthi et al., 2023). It also shows promise on MASSIVE, demonstrating its potential to advance multilingual semantic parsing. The details of the proposed framework will be discussed in Section 3.1.

To further analyze the proposed framework, we evaluate its performance under stricter constraints, where not only is there no data in the target languages, but also only limited annotated data in English is available (low-resource English data). To achieve this, we introduce three selective compression methods to create a low-resource subset from the full annotated English dataset: random selection, label-covered-based sampling, and a mix of random and label-covered-based sampling (Section 3.2). Our experiments show that using just 5% of the annotated data, we can achieve 80% of the performance obtained with the entire dataset. This demonstrates the effectiveness of our proposed framework under the constraints of limited English data. Analyzing the results of each selective method provides valuable guidance for annotators on how to build a small but effective annotated dataset,

ensuring the creation of an efficient system.

In summary, this paper makes three key contributions to the field of zero-shot multilingual semantic parsing:

- **First**, we present a comprehensive framework that leverages the power of large language models to augment multilingual data, leading to state-of-the-art zero-shot multilingual semantic parsing performance.
- **Second**, we introduce a method for selecting a small and generalized subset of annotated data. This enables us to achieve impressive results using significantly less data, reducing annotation costs and effort.
- **Finally**, we demonstrate the effectiveness of our entire framework through extensive experiments on two widely recognized multilingual semantic parsing datasets, MTOP and MASSIVE.

*Outline.* This paper consists of seven sections, with Section 1 serving as the introduction. The remaining sections are outlined as follows:

- In Section 2, a literature review of prior works relevant to the fundamental concepts and methods explored in this paper is provided. It delves into topics such as zero-shot multilingual semantic parsing and the advancement of semantic parsing with LLM.
- In Section 3, a detail of the proposed framework is presented, comprising three distinct phases: LLM-based augmentation, noisy data filtering, and multilingual semantic parsing (Section 3.1). Additionally, the second part of this section introduces selection strategies aimed at creating subsets of training data to evaluate the effectiveness of the proposed framework under strict constraints of limited English annotated data (Section 3.2).
- Section 4 outlines the experimental settings, evaluation methods, and primary results derived from the conducted experiments.
- Section 5 involves an in-depth analysis focused on gaining profound insights into the performance and behavior of the proposed method.
- Section 6 provides suggestions and directions for potential future improvements.
- Sections 7 serves as the concluding section, summarizing the main findings and conclusions drawn from the research.

## 2. Related Work

Semantic parsing, the task of converting natural language utterances into formal meaning representations such as logical forms, has seen significant advancements, particularly in multilingual and zero-shot settings (Sherborne et al., 2023; Wu et al., 2023a). In this section, we review prior work on zero-shot multilingual semantic parsing and discuss how LLMs have further advanced the field of semantic parsing.

### *2.1. Zero-shot Multilingual Semantic Parsing*

Recent years have seen remarkable progress in semantic parsing tasks (Chen et al., 2020). However, a significant disparity persists between English and other languages due to the scarcity of human-annotated data in non-English languages. Zero-shot multilingual semantic parsing has emerged as a crucial task aimed at addressing this challenge. It involves developing models that can understand and parse languages without human-annotated data in the target languages. This task is essential for extending natural language understanding to low-resource languages, where acquiring human-annotated datasets is often prohibitively expensive and time-consuming (Nicosia et al., 2021).

Various approaches have been proposed to tackle the problems of zero-shot multilingual semantic parsing. Nicosia et al. (2021) developed a technique called "Translate-and-Fill" (TaF) to augment training data for multilingual semantic parsing. TaF trains a filler on English examples and then leverages the cross-lingual capabilities of language models to generate similar examples automatically in other languages without needing any training data in those languages. Xia and Monti (2021) demonstrated a machine translation-based method to bootstrap training data for multilingual parsing, mitigating data scarcity issues through transfer learning with pre-trained multilingual encoders. Sherborne and Lapata (2022) proposed a multi-task encoder-decoder model. This model learned from English-logical form paired data and unlabeled text in each target language, transferring parsing knowledge without needing individual annotations for target languages. Additionally, a first-order meta-learning algorithm was developed by Sherborne and Lapata (2023) to train a multilingual semantic parser with maximal sample efficiency during cross-lingual transfer. This algorithm leverages data from high-resource languages to train the parser while simultaneously optimizing

its ability to perform well on low-resource languages, maximizing knowledge transfer.

### *2.2. Advancing Semantic Parsing with LLMs*

Large language models (LLMs) are language models with a large number of parameters, ranging from billions, as in the case of T5 (Raffel et al., 2020), to hundreds of billions, as seen with models like LLaMA-2 (Touvron et al., 2023). These models are typically based on the Transformer architecture and start their pre-training process with a large amount of unlabelled text data (Vaswani et al., 2017). After this pre-training, they are fine-tuned to refine their capabilities to interpret human commands or to perform specific tasks in natural language processing (NLP). Key advancements, such as in-context learning (ICL) (Brown et al., 2020) and chain-of-thought (CoT) prompting (Wei et al., 2024), enable LLMs to generate intermediate reasoning steps, which enhance their problem-solving capacity for complex tasks (Fei et al., 2023; Xu et al., 2024; Fei et al., 2024). These approaches have shown notable improvements in tasks demanding multi-step reasoning and logical inference.

In the realm of semantic parsing task, researchers have developed various strategies to utilize the in-context learning feature of LLMs to overcome the hurdles in semantic parsing. Shin and Van Durme (2022) indicate that LLMs pre-trained with programming languages, like Codex (Chen et al., 2021), are more adept at semantic parsing tasks than models like GPT-3 (Brown et al., 2020), which are mainly trained with natural language text. An et al. (2023) showed that the efficacy of ICL in semantic parsing significantly depends on the selection of the demonstration set, which is influenced by diversity, similarity, and complexity. Diversity refers to the range of patterns seen across contexts, similarity considers the common structures in specific expressions, and complexity pertains to the depth of information in each instance. Mekala et al. (2023) present ZEROTOP, a method for zero-shot task-oriented semantic parsing that transforms the challenge of semantic parsing into a combination of abstract and extractive question-answering tasks. These questions are then fed into an LLM, which produces answers to construct the desired semantic representation. In addition, Levy et al. (2023) proposed an alternative method that involves designing comprehensive demonstrations that include essential sub-logical structures for predicting new inputs in semantic parsing.

Conceptually, our study aligns with the research (Awasthi et al., 2023) that showcased the use of LLMs for converting English datasets into various

languages through ICL. In that research, carefully chosen demonstrations in target languages were used to generate input for LLMs, and the LLMs' output served as augmented data for training a multilingual semantic parser. However, our method is uniquely geared towards a zero-shot multilingual scenario, which does not require any demonstration samples in the target language for effective training. Additionally, instead of using LLMs like mT5 (Xue et al., 2021) to train the final semantic parser, we employ a smaller language model, XLM-Roberta (Conneau et al., 2020). This choice significantly enhances the practicality of our method for resource-constrained environments, such as mobile devices, where smaller, fine-tuned models can operate efficiently. This ensures broader applicability and practicality in real-world scenarios.

## 3. Methodology

Our method revolves around zero-shot multilingual semantic parsing, where we aim to develop a multilingual semantic parser for target languages despite having human-annotated data only for English. This section is divided into two parts. Firstly, we present the Zero-MParser framework designed to address the challenges of zero-shot multilingual semantic parsing.This framework consists of three distinct phases: LLM-based augmentation, noisy data filtering, and multilingual semantic parsing (Section 3.1). Secondly, we describe the process of constructing low-resource English subsets to evaluate the effectiveness of the proposed framework under the constraint where not only is there no human-annotated data in target languages, but also the available English data is limited (Section 3.2).

### *3.1. Zero-MParser Framework For Zero-shot Multilingual Semantic Parsing*

Our study aims to predict the semantic representation of a user utterance in a target language ($y_{tgt}$), given the training samples in English only ($\mathcal{D}_{en}$). The training annotated dataset comprises paired instances presented in the format of an utterance paired with its corresponding logical structure, denoted as $\mathcal{D}_{en} = \{(x_{en}^{(i)}, y_{en}^{(i)})\}$.

Our Zero-MParser framework, detailed in Figure 2, tackles zero-shot multilingual semantic parsing in three key phases: *LLM-based data augmentation*, *noisy data filtering*, and *multilingual semantic parsing*. First, we augment the English dataset by translating utterances and generating

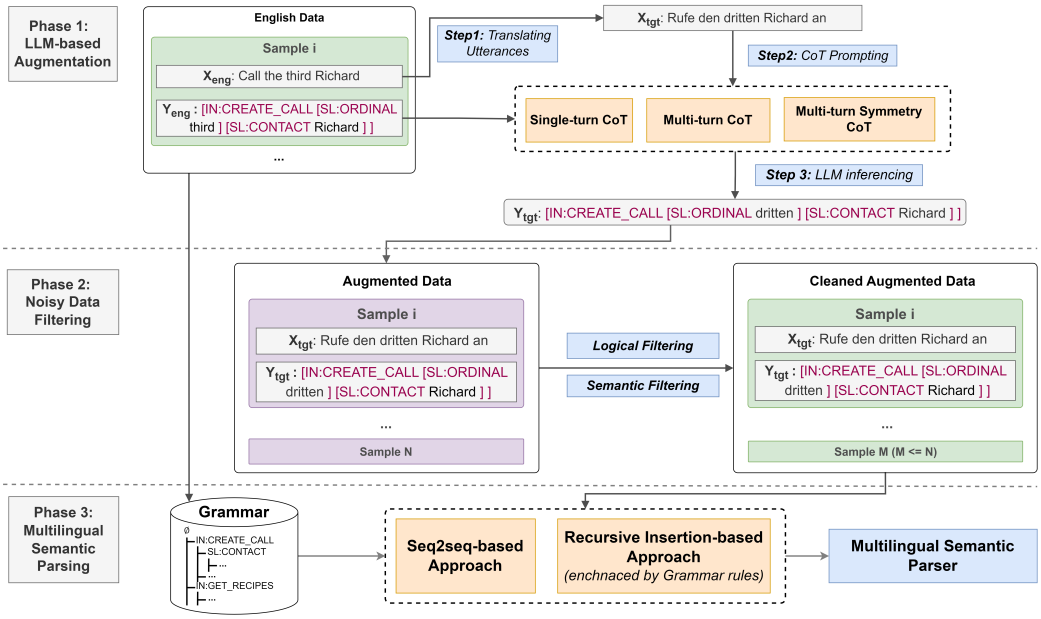


Figure 2: An overview of our Zero-MParser framework. The dashed box contains multiple approaches, and we use one approach at a time.

their logical forms in target languages like German. Crucially, these logical forms, despite being expressed in different languages, share a common semantic structure, a key feature of semantic parsing for tasks like virtual assistants where equivalent user intent should yield the same action regardless of language. As an illustration, Figure 2 demonstrates that the logical forms for English and French utterances maintain the same semantic structure `[IN:CREATE_CALL [SL:ORDINAL ] [SL:CONTACT ] ]`. After this phase, we have an augmented dataset, $\mathcal{D}_{tgt} = (x_{tgt}^{(i)}, y_{tgt}^{(i)})$, where $x_{tgt}^{(i)}$ is the translated utterance from $x_{en}^{(i)}$, and $y_{tgt}^{(i)}$ is the logical form in the target language. Next, we filter this augmented dataset to remove invalid or noisy samples, ensuring high-quality training data. Finally, we combine the original English data with the augmented data $D_{en} \cup \mathcal{D}_{tgt|tgt\in\{fr,de,\cdots\}}$ and train a multilingual semantic parser on this combined data, enabling it to understand and interpret user utterances in multiple languages.

### *3.1.1. LLM-based Augmentation*

Our preliminary experiments indicate that while LLMs show adeptness in understanding sentences in natural language, they struggle with generalizing

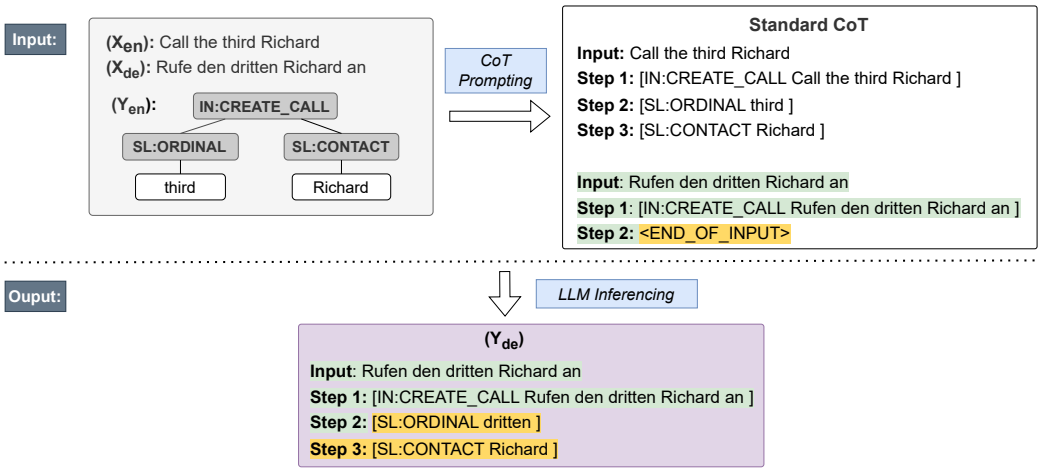


Figure 3: An example of singe-turn CoT prompting. LLM's generated text is highlighted in yellow.

complex semantic patterns. This issue arises because LLMs are predominantly trained on vast corpora of natural language text, which prepares them well for tasks like question answering or machine translation but is insufficient for semantic parsing tasks. Addressing this challenge, we implement a method that breaks down the full semantic structure into smaller, more natural segments, known as sub-semantic fragments, thereby untangling the complexity of the semantic structure step by step (as shown in Figure 3). We describe a semantic fragment as a non-terminal node within a semantic parsed tree, which includes both the semantic labels and the associated span of text (for example, `[SL:CONTACT` *Richard* `]`). This methodology allows each fragment to appear more naturally language-like in comparison to the entire, complex semantic structure. By presenting these fragments along with their translations, we empower LLMs to generalize similar fragments in new languages using CoT prompting techniques. Specifically, we first translate the $x_{en}^{(i)}$ to $x_{tar}^{i}$ using a off-the-shelf translation tool. Then, we propose three distinct methodologies that leverage the CoT prompting technique to assist LLMs in incrementally generating the semantic structure in the target language ($y_{tgt}^{(i)}$), utilizing the given triplet ($x_{en}^{(i)}$, $y_{en}^{(i)}$, $x_{tgt}^{(i)}$) as a basis. These methodologies are: Single-turn CoT, Multi-turn CoT, and Multi-turn symmetry CoT.

*Single-turn CoT.* In this approach, LLMs are tasked with producing all the semantic fragments for the target language in a single step, as illustrated in Figure 3. Given that utterances with the same meaning have the same semantic structure across different languages, we prompt the LLMs with the root node (`[IN:CREATE_CALL` ***Refen den*** `...]`) to serve as a trigger, motivating the LLMs to sequentially generate subsequent semantic fragments. Subsequently, we use these produced fragments to piece together the complete semantic structure in the target language.

*Multi-turn CoT.* As mentioned earlier, the semantic structure exhibits consistency across languages. However, the single-turn CoT approach lacks a guarantee that LLMs will maintain this consistency when generating the logical form. Therefore, in the multi-turn CoT approach, we utilize LLMs to progressively generate text spans associated with each semantic frame. Each span is generated one at a time. For example, in Figure 4, during the first turn, LLMs predict the span `"dritten"` as the text span for `SL:ORDINAL`, and in the second turns, LLMs predict the span `"Richard"` corresponding to `SL:CONTACT` (last turn).

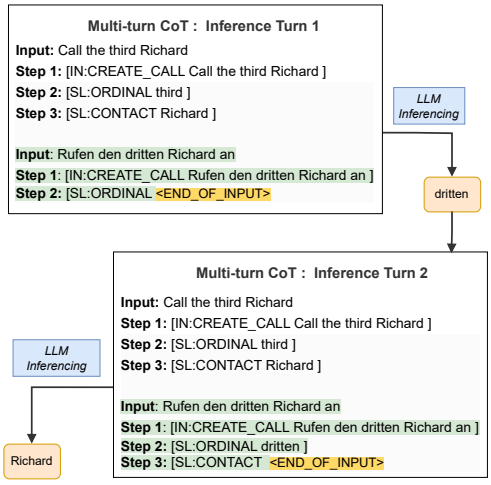


Figure 4: Multi-turn CoT prompting.

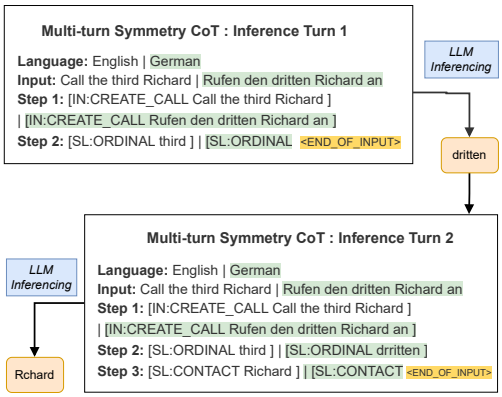


Figure 5: Multi-turn symmetry CoT.

*Multi-turn symmetry CoT.* Unlike the multi-turn CoT prompting approach, our multi-turn symmetry CoT strategy improves the alignment of semantic fragments between English and the target language. This strategy entails

pairing each semantic fragment from both languages, as depicted in Figure 5. Semantic fragments in two languages are separated by a special token, denoted as "|". This alignment facilitates the transfer of information, fostering better coordination and coherence between steps in different languages and consequently boosting the overall effectiveness of the parsing process.

Each prompting technique offers unique advantages that address different aspects of the task, allowing us to determine the most effective strategy through comparative analysis and empirical evaluation. For instance, the "Single-turn CoT" technique is efficient as it requires only a single inference call to the LLM, making it resource-friendly. The "Multi-turn CoT" technique helps mitigate the issue of incorrect logical tokens by iteratively refining the output. Lastly, the "Multi-turn CoT Symmetry" technique enhances the alignment between the source and target languages, improving translation accuracy and consistency.

*3.1.2. Noisy Data Filtering*

To mitigate the impact of potentially noisy generated text, we introduce a rigorous filtering procedure with two distinct steps. The initial step involves logical filtering, which eliminates all augmentations that do not yield an accurate semantic parsed tree. Subsequently, the second step involves semantic filtering, aiming to exclude samples that produce incorrect semantic representation, such as inaccurate span predictions. The overarching objective of this entire process is to curate our augmented data, upholding a high standard of quality.

*Noisy Logical Data Filtering.* The noisy logical data filtering process involves the following sub-steps:

- **Step 1: Alignment Verification**. Initially, we discard samples where spans are not contained within the translated utterance, ensuring coherence and relevance to the intended content.
- **Step 2: Label and Intent Validation**. Subsequently, we scrutinize samples for unrecognized label slots or intents by comparing them to a reference label set from English data. Samples with unacknowledged labels or intents are excluded from consideration.
- **Step 3: Semantic Parsing Tree Compatibility**. Additionally, we evaluate whether samples can be accurately converted into semantic parsed trees. Any samples failing this conversion are also excluded.

*Noisy Semantic Data Filtering.* Our preliminary experiments, along with the findings from previous studies (Do et al., 2023b; Awasthi et al., 2023), indicate that a significant source of error in semantic parsing lies in incorrectly predicted slot spans. This implies that while the label is predicted accurately, the corresponding span for these labels is often incorrect. Therefore, by reducing the occurrence of wrongly assigned spans in the augmented dataset, we aim to enhance the overall performance of the model. Based on this point, we introduce three semantic filtering methods.

- **Method 1: Sentence Embedding-based.** Figure 6 shows an overview of this filtering approach. The process begins by considering the augmented logical form in the target language alongside its counterpart in the English language. Subsequently, we extract slot spans from both English and the target languages. For each span pair, we calculate the similarity score using multilingual Sentence-BERT (Reimers and Gurevych, 2019), which is pre-trained on multilingual texts. Next, the score of an augmented sample is determined by finding the minimum value from the list of similarity scores. This approach implies that the weakest point serves as the representative of the slot spans. Alternatively, another combining score method, like meaning, could be employed. Experimental results suggest that the choice between mean and min as the combining score method yields approximately the same outcomes. Based on the obtained scores, we drop the top $\lambda$ samples with the lowest scores, where the value of $\lambda$ represents a hyperparameter.

- **Method 2: Cross-alignment Sentence Embedding-based.** The primary distinction between this approach and the one mentioned earlier lies in the swapping of spans between the source and target languages (Figure 6). This adjustment is made because the previous method neglects to consider the context of the span. It is crucial to take into account whether the spans of the two texts are similar or not. Specifically, we swap the span from the source language to the target language and vice versa to address this contextual consideration.

- **Method 3: Back Translation-based Semantic Filtering.** If two spans in two languages are similar, it is more likely that back translation can be performed from the target language to the source language without introducing any mislabeling. Building on this idea, we

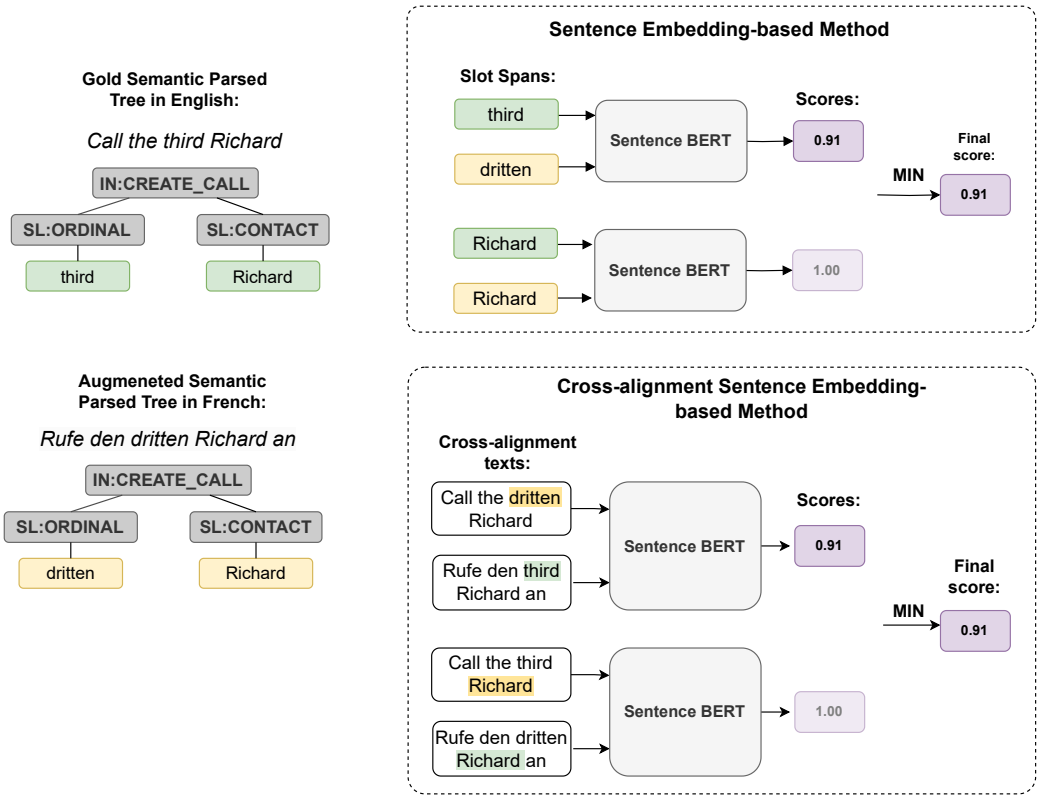


Figure 6: Sentence Embedding-based Semantic Scoring Methods.

replicate the data augmentation process by translating from the target language, changing the order of languages in Figure 5 to have the target language first and English second. We then compute the similarity for each back-translated slot span in English compared to the original span. This results in a list of similarity scores corresponding to each slot. We use the ROUGE score (Lin, 2004) as a metric to quantify this similarity, measuring lexical similarity to ensure that content is preserved during back translation. Finally, the score of an augmented sample is determined by finding the minimum value from the list of similarity scores.

*3.1.3. Multilingual Semantic Parsing.*

Following the filtering step, we now have high-quality training data ready to be used for training semantic parsers. In this phase, we utilize two multi-

lingual semantic parsing methods: the seq2seq-based and recursive insertion-based methods.

*Seq2seq-based Method.* This method employs a traditional seq2seq technique for multilingual semantic parsing, building upon previous studies (Awasthi et al., 2023; Nicosia et al., 2021). It utilizes a pre-trained encoder-decoder model designed for multilingual tasks as its foundation (Xue et al., 2021), and training on augmented data. The loss function is formulated as the negative log-likelihood of the actual tokens in the output sequence as follows:

$$Loss = -\sum_{t=1}^{T} \log(p\ (\ y_t \mid y_{<t}, x)) \tag{1}$$

Here, $T$ represents the length of the output sequence, $y_t$ denotes the true token at time step $t$, $y_{<t}$ covers the sequence of tokens leading up to time step $t-1$, $x$ represents the input sequence from the augmented training data, and $p\ (\ y_t \mid y_{<t}, x)$ indicates the model's predicted probability of token $y_t$ at time step $t$. This prediction takes into account the sequence of tokens up to time step $t-1$ and the input sequence $x$.

*Recursive Insertion-based Method.* In this method, we implement a recursive-insertion-based approach (Mansimov and Zhang, 2022), incorporating grammar constraints for enhancement. As the backbone, we use an encoder-only multilingual pre-trained language model (Conneau et al., 2020). The parsing procedure is conceptualized as the step-by-step generation of sub-parsed trees, where the result of each prior step becomes the input for the subsequent one (refer to Table 1). The output of each step includes the label, the start position of the label, and the end position of the label.

| **Step** | **Linearized representation of full logical form** | **Start** | **End** | **Label** |
|---|---|---|---|---|
| $\mathcal{P}_0$ | Call the third Richard | - | - | - |
| $\mathcal{P}_1$ | [IN:CREATE_CALL Call the third Richard ] | 0 | 4 | IN:CREATE_CALL |
| $\mathcal{P}_2$ | [IN:CREATE_CALL Call the [SL:ORDINAL third ] Richard ] | 2 | 3 | SL:ORDINAL |
| $\mathcal{P}_3$ | [IN:CREATE_CALL Call the [SL:ORDINAL third ] [SL:CONTACT Richard ] ] | 3 | 4 | SL:CONTACT |

Table 1: An illustrative sequence of incremental trees in the parsing process using the recursive insertion-based method approach.

We first extract grammar from the English semantic parsing tree, representing the logical form, to disregard less promising label predictions. Specifically, we derive parent-child grammar rules from the semantic parsed tree of English data $\mathcal{G} = \{A \rightarrow B \mid A, B \text{ are non-terminal nodes}\}$, for instance, `IN:CREATE_CALL` $\rightarrow$ `SL:CONTACT`. Notably, the schema is identical in both English and the target language, differing only in the arrangement of slots within each intent (Awasthi et al., 2023). Consequently, the grammar extracted from English data has broad applicability across target languages. With the obtained grammar, we proceed to train a multilingual semantic parsing model for all languages, utilizing the grammar-based RINE (Recursive Insertion-based Encoder) model (Do et al., 2023b).

To train the semantic parser using the recursive insertion-based method, we consider the sequence representation of a sub-parsed tree $\mathcal{P}_i = [x_1, ..., x_k, ..., x_m]$, where $m$ denotes the total number of tokens in the sequence. The loss function includes node label prediction, start position prediction, and end position prediction, with a grammar-based penalty to eliminate less promising node label predictions as follows:

$$s^{penalty} = \begin{cases} 0 & \text{if} \quad p^{label} \in \mathcal{C} \\ +\infty & \text{otherwise} \end{cases} \tag{2}$$

$$Loss_{label} = \text{CE}(g^{label}, p^{label}) + s^{penalty} \tag{3}$$

$$Loss_{start} = \sum_k \text{CE}(g_k^{start}, p_k^{start}) \tag{4}$$

$$Loss_{end} = \sum_k \text{CE}(g_k^{end}, p_k^{end}) \tag{5}$$

$$Loss_{final} = Loss_{label} + Loss_{start} + Loss_{end} \tag{6}$$

Here, $\mathcal{C}$ denotes the candidates obtained by the grammar $\mathcal{G}$ and parent node $U$ ($\mathcal{C} = \mathcal{G}(U)$). "CE" stands for cross-entropy loss. The probabilities of node label, start position, and end position are represented by $p^{label}$, $p^{start}$, and $p^{end}$, respectively, using the hidden state obtained from the encoder. Meanwhile, $g^{label}$ represents the ground-truth node label, and $g_k^{start}$ and $g_k^{end}$ are the ground-truth start and end positions. The loss function $Loss_{label}$ combines the cross-entropy loss between the predicted node label $p^{label}$ and the ground truth with the grammar-based penalty $s^{penalty}$. Similarly, $Loss_{start}$ and $Loss_{end}$ compute the cross-entropy losses between the

predicted *start* and *end* positions $p_k^{start}$ and $p_k^{end}$, respectively, and their corresponding ground truth.

### *3.2. Zero-shot Multilingual Semantic Parsing with Limited English Data*

To answer the question: How effective is the proposed framework under the stricter scenarios of having a limited amount of annotated English data available? In this section, we introduce methods for constructing these limited data subsets. Specifically, we present three approaches for creating limited data subsets $\mathcal{S}_{en}$ from fully annotated data $\mathcal{D}_{en}$ ($\mathcal{S}_{en} \subset \mathcal{D}_{en}$): *Random Sampling Method*, *Label-Covered-based Method*, and *Mixed Random and Label-Covered-based Method*. After obtaining the subset $\mathcal{S}_{en}$, we run our framework on this subset and analyze the obtained results (Section 4.3.2).

#### *3.2.1. Random Sampling Method*

We introduce a hyperparameter, denoted as $k$, to determine the desired number of training samples to retain. With this parameter, we randomly sample $k$ samples of the dataset $\mathcal{D}_{en}$. The purpose of this method is to maintain the label distribution by randomly selecting a portion of the dataset. This ensures that the model focuses on the crucial aspects of label distribution, making it easier to learn.

#### *3.2.2. Label Covered-based Method*

In this approach, our goal is to create a set of samples, denoted as$\mathcal{S}_{en}$, that ensures each intent/slot label in the training data $\mathcal{D}_{en}$ appears at least once in $\mathcal{S}_{en}$. To achieve this, we introduce Algorithm 1, designed to guarantee that each intent/slot label is covered at least once in the selected samples.

We randomly select a sample $e$ from the training set $\mathcal{D}_{en}$ (line 3) and remove $e$ from the training pool (line 4). We then extract the set of covered labels in $e$ (line 5). If the intersection between the set of label pools and the set of covered labels $t_{convered}$ is not empty (line 6), we add $e$ to $\mathcal{S}_{en}$. The set of uncovered labels $\mathcal{T}$ is updated to exclude those already covered in $t_{convered}$ (line 8). This selection process continues until the desired number of demonstrations is reached, potentially resulting in more than one example chosen for each label (lines 2).

#### *3.2.3. Mix Random and Label Covered-based Method*

The label-covered method is effective when ensuring that each label has at least one sample in the training data. However, it has the potential to

**Algorithm 1:** Label Covered-based Sampling Method

**Input :** Pool of English training samples $\mathcal{D}_{en}$; List of uncovered intent/slot labels $\mathcal{T}$; Desired number of output samples $k$

**Output:** Set of selected English training samples $\mathcal{S}_{en}$

$\mathcal{S}_{en} = \emptyset$
**while** $|\mathcal{S}_{en}| < k$ **do**
    Random select an example $e \in \mathcal{D}_{en}$
    Remove $e$ from $\mathcal{D}_{en}$
    $t_{covered}$ = set of covered labels in $e$
    **if** *intersection*$(t_{covered}, T)$ *is not* $\emptyset$ **then**
        Add $e$ to $\mathcal{D}_{en}$
        Remove from $T$ labels that appear in $t_{covered}$
    **end**
    **if** $|\mathcal{S}_{en}| == k$ **then**
        break
    **end**
**end**

harm performance by breaking the distribution of the labels, making it more challenging for the model to learn. Given these considerations, we aim to combine the advantages of both methods mentioned above by using a mix of them. This entails selecting one sample using the first approach, the second sample using the second approach, the third sample using the first approach again, and so forth until the desired number of training samples $k$ is reached.

## 4. Experiments

In this section, we outline our experimental settings and results for two configurations of multilingual semantic parsing tasks: zero-shot multilingual semantic parsing with full English data and zero-shot multilingual semantic parsing with limited English source data.

### *4.1. Datasets*

We conducted evaluations on two well-known multilingual semantic parsing datasets: MTOP (Li et al., 2021) and MASSIVE (FitzGerald et al., 2023). Examples of each dataset are provided in Table 2.

| Dataset | Example |
|---|---|
| MTOP | **Utterance**: What's next week weather in Vegas<br>**Logical Form**: [IN:GET_WEATHER [SL:DATE_TIME next week ] [SL:LOCATION Vegas ] ] |
| MASSIVE | **Utterance**: please raise the lights to max<br>**Intent**: iot_hue_lightup<br>**Raw Logical Form**: please raise the lights [change_amount : to max]<br>**Converted Logical Form**: [IN:IOT_HUE_LIGHTUP [SL:CHANGE_AMOUNT to max ] ] |

Table 2: An example utterance-logical form pair for each of the datasets.

*MTOP.* This dataset comprises multilingual semantic parsing samples in six languages: English, German, French, Hindi, Thai, and Spanish. Each sample consists of a user utterance along with its corresponding semantic logical form in hierarchical representation (Table 2). Specifically, the logical forms in the dataset contain labels for user *intents* prefixed with "IN:", and relevant entities called *slots* prefixed with "SL:". The dataset spans 11 different domains, encompassing 117 intents and 78 slots. On average, each language has around 12.3K samples for training, 1.5K for development, and 2.7K for testing.

*MASSIVE.* This dataset covers a diverse set of languages, totaling 51, for semantic parsing. With 18 domains, 60 intents, and 50 slots, the dataset averages approximately 11.5K phrases for training, 2K for development, and 3K for testing per language. For our research, we focused on six of these languages: English, German, Spanish, French, Hindi, and Thai. To work with this dataset, we converted the logical form of MASSIVE data into hierarchical representation (Table 2).

In this experiment, our emphasis is on zero-shot multilingual semantic parsing, which means that in this setting, the annotated training data comprises only samples from the English language, with no training samples available for the other five languages. Additionally, in the zero-shot multilingual semantic parsing with limited English data, we utilize a small set of English training data extracted using the proposed methods presented in Section 3.2.

### *4.2. Experimental Setup*

Our method comprises three main phases: *LLMs-based Augmentation*, *Noisy data Filtering*, and *Multilingual Semantic Parsing*.

*LLMs-based Augmenting.* In this phase, we utilized three versions of the Llama-2 models, with 7B, 13B, and 70B parameters (Touvron et al., 2023). To ensure consistency in LLMs outcomes, we fixed the temperature value at 0, as done in prior studies (Levy et al., 2023; Zhuo et al., 2023).

*Noisy data Filtering.* For the sentence embedding-based methods, we employed the multilingual bi-encoder pre-trained model: `distiluse-base-multilingual-cased` (Reimers and Gurevych, 2019) to calculate the similarity of two texts. The selection for the retention percentage $\lambda$ was made from the options {5%, 10%, 20%, **30%**, 40%, 50%}[1].

*Seq2seq-based Multilingual Semantic Parsing.* We initialized our seq2seq semantic parser with the mT5-Large pre-trained checkpoint (1.2B parameters) (Xue et al., 2021). Fine-tuning involved a mix of English gold data and augmented data from various languages. We used the Adam optimizer (Kingma and Ba, 2014) with a learning rate of 1e-5, a warm-up phase of 1000 steps, and a batch size of 32. The training process lasted for 10,000 steps, taking approximately 25 hours on a single A100 80GB GPU. The best checkpoint, determined by development set performance, was selected for test set predictions.

*Recursive Insertion-based Multilingual Semantic Parsing.* We employed the XLM-Roberta model (355M parameters) (Conneau et al., 2020) as an encoder. The training process used the Adam optimizer (Kingma and Ba, 2014) with a warm-up step count of 1000. The learning rate was selected from the options {**1e-05**, 5e-05, 1e-06}. The training duration encompassed 50 epochs (1000 steps), and the grammar rules were derived from the training set of the English gold data.

*Evaluation Metric.* Following the previous research (Awasthi et al., 2023), we employ agnostic exact match (EM) as our primary evaluation metric. This metric disregards the sequence of slots within an intent when comparing logical forms, facilitating accurate assessment. We evaluate our approach and compare it to other methods using this exact match score.

*Baselines.* We replicated the TAF method (Nicosia et al., 2021; Awasthi et al., 2023) as a baseline, maintaining the hyperparameters consistent with the values presented in the original papers.

[1]The **bold** value is the best performance on the development set.

### *4.3. Main Results*

#### *4.3.1. Zero-shot Multilingual Semantic Parsing with full English data*

We evaluated our proposed method on the MTOP dataset, comparing them with approaches from previous studies, including zero-shot, few-shot, and the Translate, Augment, and Fill (TAF) strategy (Awasthi et al., 2023; Nicosia et al., 2021). Specifically, we assessed various approaches:

- **(1):** A seq2seq model trained exclusively on English data, representing the zero-shot setting (Awasthi et al., 2023).
- **(2):** A seq2seq model that incorporates a few manually selected samples of the target languages (≈250 samples for each language) into the English training data (Awasthi et al., 2023).
- **(3):** A seq2seq model incorporating data augmentation using the TAF technique (Awasthi et al., 2023).
- **(4):** Our reproduced version of (1).
- (**5**)**:** Our reproduced version of (3).
- **(6):** Our zero-shot model is trained exclusively on English data using the recursive insertion-based method (RINE) and grammatical structure constraints.
- **(7):** Our proposed seq2seq-based Zero-MParser method.
- **(8):** Our proposed RINE-based Zero-MParser method utilizes grammatical structure constraints.

In Table 3, our proposed method showcased superior performance, surpassing previous state-of-the-art results, especially the TAF method (Awasthi et al., 2023; Nicosia et al., 2021), by a margin of 2.2 in exact match (EM) scores, as indicated by the results in rows (3) and (8). This progress was particularly notable for languages with limited resources, such as Hindi and Thai, as highlighted in rows (7) and (8). When evaluating the performance of seq2seq semantic parsing methods in rows (1), (2), (3), and (7), our Zero-MParser seq2seq-based model achieved the most impressive outcomes. This emphasizes the effectiveness of our data augmentation technique, further bolstered by our multilingual CoT prompting strategies. Additionally,

| Method | de | es | fr | hi | th | Average |
|---|---|---|---|---|---|---|
| (1) Seq2seq Zero-shot (Awasthi et al., 2023) | 54.4 | 57.8 | 62.8 | 42.3 | 42.1 | 51.9 |
| (2) Seq2seq Few-shot (Awasthi et al., 2023) | 62.8 | 69.5 | 65.9 | 55.3 | 53.9 | 61.5 |
| (3) Seq2seq TAF (Awasthi et al., 2023) | 75.0 | 74.9 | 78.0 | 63.0 | 60.8 | 70.3 |
| *(Our methods)* | | | | | | |
| (4) Seq2seq Zero-shot | 54.0 | 58.9 | 58.9 | 44.1 | 38.3 | 50.8 |
| (5) Seq2seq TAF | 73.2 | **75.2** | **78.5** | 61.9 | 62.6 | 70.3 |
| (6) RINE-based Zero-shot | 63.5 | 68.1 | 70.3 | 54.4 | 43.9 | 60.0 |
| (7) Seq2seq Zero-MParser | 73.9 | 71.9 | 76.2 | 71.0 | 62.4 | 71.1 |
| (8) RINE-based Zero-MParser | **75.2** | 73.7 | 78.0 | **71.0** | **64.8** | **72.5** |

Table 3: Comparing Performance through EM Scores on the MTOP Test Dataset

implementing the RINE method with grammatical rules in rows (6) and (8) resulted in even greater performance improvements over the seq2seq methods, underscoring the significance of incorporating grammatical insights and breaking down the parsing task into smaller, manageable steps.

*MASSIVE.* Figure 7 depicts the comparison between the seq2seq TAF method baseline and our Zero-MParser across five languages in the MASSIVE datasets. Our approach demonstrates superiority over the TAF method in four out of five languages. When averaging across all languages, our method outperforms the TAF method by a 0.9 EM score. The exception is observed in the Spanish language, where our Zero-MParser performs less effectively than the TAF method. This outcome aligns with the results obtained from the MTOP dataset. One possible explanation is that the utilized LLMs may not possess sufficient capabilities in Spanish. Investigating the primary reason behind this discrepancy presents an intriguing research avenue.

*4.3.2. Zero-shot Multilingual Semantic Parsing with low-resource source data*

To assess the effectiveness of our framework in a low-resource setting with an English dataset, we generate low-resource data using three discussed sampling methods outlined in Section 3.2.1. Each sampling method involves creating subsets with a corresponding percentage value relative to fully annotated data; specifically, we generate subsets for {1%, 3%, 5%, 7%, 10%, 25%, 50%, 80%}. Subsequently, we run our framework on each subset and obtain final results on the MTOP test set.

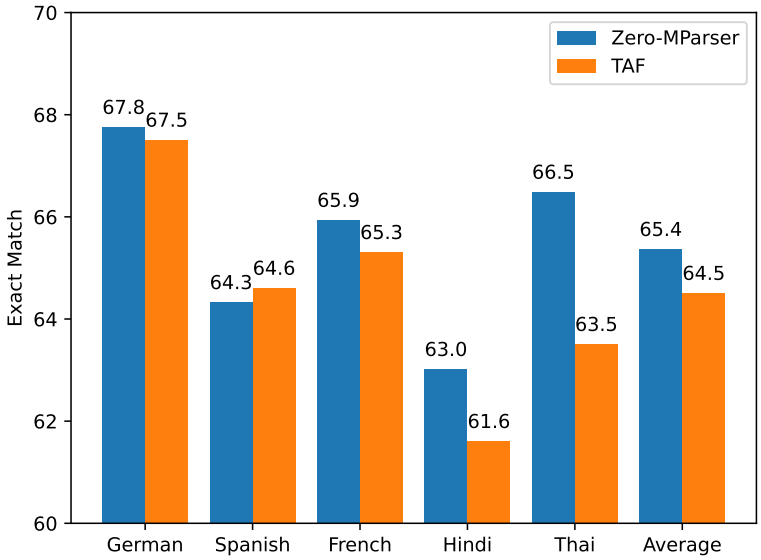


Figure 7: Results of Zero-MParser and the TAF baseline on the MASSIVE test set.

The results from the three sampling methods are depicted in Figure 8. Notably, in both the random sampling and the mixed random and label-covered methods, our framework achieves over 80% performance compared to fully annotated data with just 5% of annotated data. Specifically, it attains 81% (58.9/72.5) with random sampling and 83% (60.4/72.5) with mixed sampling. This demonstrates the efficacy of our framework even in the presence of limited resources for English-annotated data.

Furthermore, by comparing the results of the three sampling methods, it is evident that random sampling consistently produces better results than label-covered sampling in each annotated percentage setting. We posit that random sampling preserves the label distribution in the annotated data, making it easier for models to learn. Conversely, label-covered sampling ensures each label exists at least once in the selected subsets but can disrupt the label distribution, thereby impacting performance. Finally, the mixed random and label-covered method combines the strengths of both approaches and achieves the highest results. This suggests a guideline for annotators: not only should they aim to cover as many labels as possible, but also strive to maintain a balanced distribution of labels. For instance, if two annotators are involved, the first can focus on capturing the maximum number of labels,

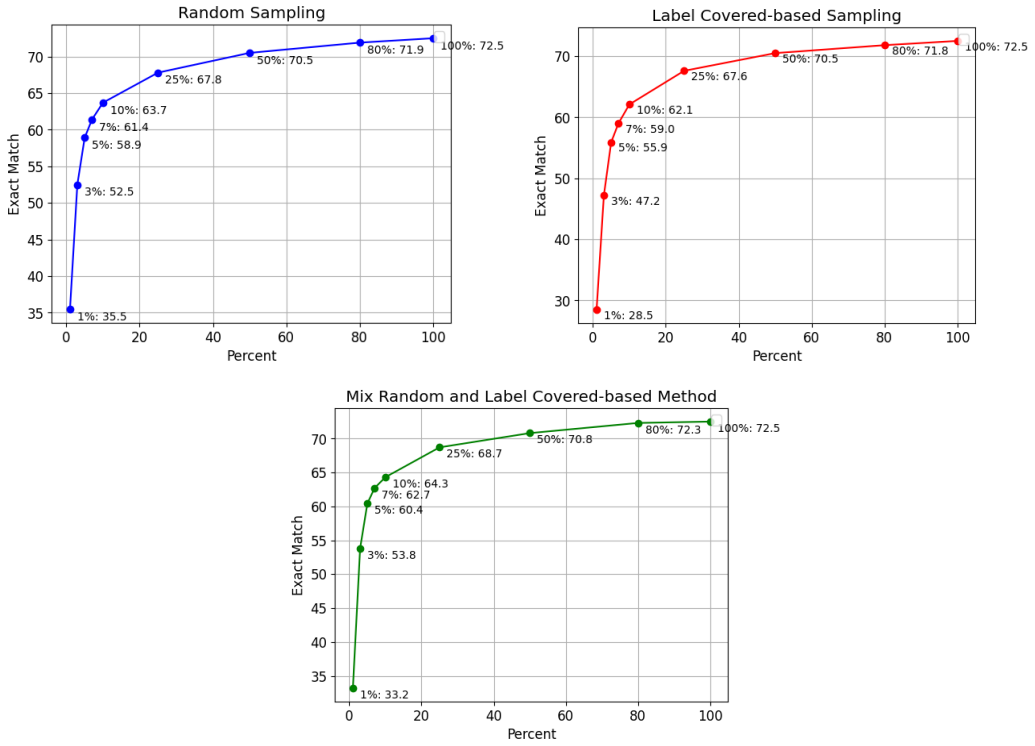


Figure 8: Comparing the accuracy of Zero-MParser with different percent of annotated English data using different data sampling strategies.

while the second adopts a more standard annotation approach.

## 5. Analysis

In this section, we conduct an in-depth analysis to enhance our understanding of the framework. All experiments are conducted on the MTOP development set.

### *5.1. Ablation Study*

To assess the impact of each component in our framework, we conducted an ablation study on the validation set of the MTOP dataset (Table 4). Our findings indicate a significant decrease in results when the multilingual augmenting phase is excluded, with the EM score dropping by 11.5 points compared to the full-setting model. Additionally, omitting the noisy data filtering phase resulted in a lower EM score compared to the baseline, showing

| Method | de | es | fr | hi | th | Average | Δ |
|---|---|---|---|---|---|---|---|
| RINE-based Zero-MParser | **72.4** | **72.4** | **77.9** | 65.4 | **62.2** | **70.1** | - |
| - without Multilingual Augmenting | 62.3 | 66.1 | 69.0 | 49.2 | 43.3 | 58.0 | -11.5 |
| - without Noisy Data Filtering | 72.1 | 71.7 | 75.9 | 65.9 | 60.4 | 69.2 | -0.9 |
| - without Recursive Insertion-based method | 71.2 | 70.4 | 74.8 | **66.0** | 61.0 | 68.7 | -1.4 |

Table 4: Ablation study results on the MTOP development set. The symbol Δ represents the variance in average EM scores between the full-setting model and other methods.

| Method | de | es | fr | hi | th | Average |
|---|---|---|---|---|---|---|
| Llama2-7B | 68.7 | 69.5 | 73.6 | 59.9 | 53.6 | 65.0 |
| Llama2-13B | 71.5 | 71.3 | 75.8 | 59.3 | 54.4 | 66.5 |
| Llama2-70B | **72.4** | **72.4** | **77.9** | **65.4** | **62.2** | **70.1** |

Table 5: LLM Size Impact: Outcomes with different sizes of LLM.

a decrease of 0.9 points. Furthermore, not utilizing the recursive insertion-based method led to a performance drop of 1.4 EM score. These outcomes underscore the importance of each component in our framework.

### *5.2. The impact of LLM sizes*

Table 5 displays the outcomes obtained with various parameter sizes of LLama-2 (Touvron et al., 2023). It is evident that larger LLM models consistently yield superior performance, with model 70B producing the best results, followed by model 13B, and ultimately model 7B. However, it is crucial to emphasize the effectiveness of smaller models as well. For instance, despite being only 1/10th the size of LLama 2 70B, LLama 7B still achieves 93% of the larger model's performance. This observation suggests that one can select the LLM size based on specific computational constraints without experiencing a significant decline in performance.

### *5.3. The impact of CoT strategies*

Our analysis, as outlined in Table 6, indicates that among the three examined multilingual CoT approaches (standard, multi-turn, and multi-turn symmetry), the multi-turn symmetry strategy stands out as the most effective. This notable success can be credited to its inherent capability to improve

| Method | de | es | fr | hi | th | Average |
|---|---|---|---|---|---|---|
| Standard CoT | 71.3 | 70.2 | 75.5 | 59.0 | 56.6 | 66.5 |
| Multi-turn CoT | 72.1 | 70.4 | 75.6 | 61.4 | 56.7 | 67.3 |
| Multi-turn symmetry CoT | **72.4** | **72.4** | **77.9** | **65.4** | **62.2** | **70.1** |

Table 6: Impact of CoT prompting technique: The use of multi-turn CoT approaches leads to more effective datasets, which in turn produce higher EM scores

| Method | Num. Filtered Sample $(\lambda)^2$ | de | es | fr | hi | th | Average |
|---|---|---|---|---|---|---|---|
| Without semantic filtering | _ | 72.1 | 71.7 | 75.9 | 65.9 | 60.4 | 69.2 |
| - With Sentence Embedding-based Semantic Filtering | 2223 (5%) | 72.7 | 71.8 | 77.1 | 64.6 | 60.2 | 69.3 |
| - With Cross Alignment Sentence Embedding-based Semantic Filtering | 2223 (5%) | **73.0** | **72.7** | 76.4 | **65.9** | 61.3 | 69.8 |
| - With Back Translation-based Semantic Filtering | 13342 (30%) | 72.4 | 72.4 | **77.9** | 65.4 | **62.2** | **70.1** |

Table 7: Impact of semantic filtering method.

alignment and coherence between corresponding steps in both English and the target languages.

### *5.4. The impact of filtering methods*

In Table 7, we present the performance of our models using three semantic filtering strategies and without employing semantic filtering and number of discarded samples. The best results are achieved with the back translation method. Omitting the semantic filtering step results in a 0.9 EM score decrease compared to the filtering phase. These results demonstrate the effectiveness of the semantic filtering step. Additionally, the back translation-based semantic filtering method proves to be the most effective, filtering 30% of the augmented data while still achieving the highest performance among all methods.

| Method | de | es | fr | hi | th | Average |
|---|---|---|---|---|---|---|
| MBART English-to-Many (Tang et al., 2020) | 71.3 | 70.2 | 75.5 | 59.0 | 56.6 | 66.5 |
| Google | 72.4 | 72.4 | 77.9 | 65.4 | **62.2** | 70.1 |
| Oracle | **72.6** | **72.7** | **78.3** | **66.1** | 61.0 | **70.2** |

Table 8: Translation Impact: Outcomes with Google's translation tool are close to the results with human translators (Oracle).

### *5.5. The Impact of Translated Text*

Our analysis, as presented in Table 8, assesses the impact of various machine translation tools on the ultimate performance of our method. We contrast expressions translated by the freely accessible mBART English-to-Many model (Tang et al., 2020), Google Translate, and an optimal scenario utilizing human-annotated translations from the dataset. Google Translate demonstrates a substantial improvement over mBART, elevating the EM score by 2.4 points. Remarkably, its outcomes closely align with those attained through human expertise, underscoring its effectiveness in multilingual semantic parsing tasks, a finding consistent with previous research methods, aligning with previous works (Shi et al., 2022; Li et al., 2014).

### *5.6. Can LLMs Be Directly Applied for Zero-Shot Multilingual Semantic Parsing?*

To further evaluate the effectiveness of our proposed framework, we address the question: "Is it necessary to involve dataset augmentation followed by parser training, or can we directly apply LLMs for this task?". To investigate this, we follow a specific process. Given an utterance in the target language, we first identify the top-k similar utterances in the English training data using the multilingual Sentence-BER model, distiluse-base-multilingual-cased2 (Reimers and Gurevych, 2019). Next, we prompt the LLMs using the CoT method, similar to Single-turn CoT (Section 3.1.1), employing the top-k similar English utterances and their corresponding logical forms. The LLM

[2] $\lambda$ indicates the percentage of filtered samples per augmented samples. We tested $\lambda$ values of {5%, 10%, 20%, 30%, 40%, 50%} and reported the best-performing value for each method.

(Llama-2-70b) then predicts the CoT steps for the utterance in the target language. An example of the input is illustrated in Figure 9.

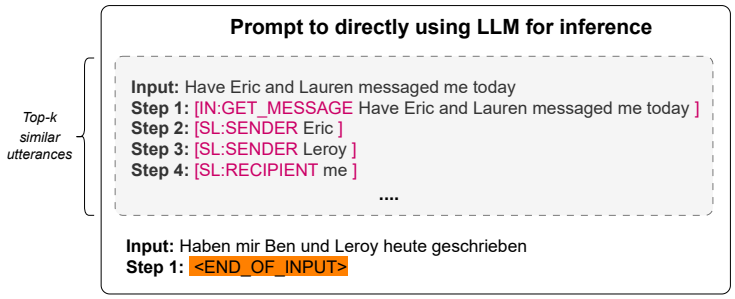


Figure 9: Example input for using the LLM for inferencing.

Table 9 presents the results of our proposed framework, Zero-MParser, compared with the method of directly using LLMs for inferencing with different top-k values {1, 3, 5}. The results indicate a significant performance gap of 20.7 EM scores. These findings underscore the necessity of the proposed method for performing zero-shot multilingual semantic parsing. Additionally, as the value of k increases, accuracy improves. However, this comes with increased GPU memory requirements and longer inference times. This trade-off between resource requirements, inference time, and accuracy is a critical consideration. Nevertheless, with the rapid advancements in the capabilities of LLMs, directly applying LLMs to the task of multilingual semantic parsing shows great promise for future research endeavors.

| **Method** | **de** | **es** | **fr** | **hi** | **th** | **Average** |
|---|---|---|---|---|---|---|
| Zero-MParser | **72.4** | **72.4** | **77.9** | **65.4** | **62.2** | **70.1** |
| Llama-2-70b for inferencing | | | | | | |
| - top-k similar utterances = 1 | 31.3 | 27.3 | 33.2 | 31.6 | 23.7 | 29.4 |
| - top-k similar utterances = 3 | 47.8 | 43.3 | 51.3 | 37.0 | 36.7 | 43.2 |
| - top-k similar utterances = 5 | 54.8 | 50.3 | 57.0 | 42.2 | 42.6 | 49.4 |

Table 9: Comparison of method using LLM for inferencing and Zero-MParser.

### *5.7. Efficiency Analysis of the Data Augmentation Process*

In this section, we evaluate the efficiency of our proposed data augmentation framework in comparison to manual annotation and the baseline method, TAF. The key metrics, summarized in Table 10, clearly demonstrate the substantial advantages of our approach. Remarkably, our method generates the dataset in just 72 hours—a significant reduction from the 75–125 days (Li et al., 2021) required for manual annotation—while achieving an accuracy close to the 96% benchmark of human-labeled data. Although our framework requires more time to produce augmented data than TAF, it delivers significantly higher accuracy. This trade-off underscores a favorable balance between efficiency and performance, making our method an attractive alternative. Furthermore, our approach only needs to be executed once to generate high-quality training data, thereby enhancing its practicality for real-world applications.

| **Metric** | **Human Annotation** | **Zero-MParser** | **TAF** |
|---|---|---|---|
| Time Consumption | 75 $\sim$ 125 days[3] | 72 hours | 10 hours |
| Computational Resources | _ | 1 GPU A100 80GB | 1 GPU A100 80GB |
| Throughput | _ | 1087 samples/hour | 7833 samples/hour |
| Final Accuracy | 75.2 | 72.5 | 70.3 |

Table 10: Comparison of Efficiency Metrics Across Human Annotation, TAF, and Zero-MParser, including time consumption (total time required for data generation), computational resources (hardware used), throughput (samples generated per hour), and final accuracy (performance of the semantic parser on MTOP test set).

### *5.8. Case Study*

Table 11 presents several example utterances from the MTOP dataset to illustrate the performance differences between our model and the baseline. In the first example, our model correctly predicted the span `"Haustier - Adoptions"` for the slot `SL:GROUP`, whereas the baseline model predicted `"Haustier - Adoptionen"`. This discrepancy highlights the effectiveness of our semantic parser in accurately predicting spans based on their positions in the input utterance. Compared to the baseline, which relies on sequence generation that does not guarantee correct alignment with the input. The

[3]The estimated duration of 75–125 days is derived from the original paper, which approximates 15–25 days per language.

second example involves a ground-truth tree with a complex structure containing multiple intents and slots, requiring the model to correctly predict each semantic fragment. Our model was able to return the correct answer, whereas the baseline model failed to do so. This suggests that our chain-of-thought prompting augmentation method generates more accurate and complex training samples, enabling the semantic parser to learn and predict complex structures more effectively.

| **Type** | **Ouput** |
|---|---|
| Input | Rufe meine Haustier - Adoptions - Gruppe auf Whatsapp an (Call my pet adoption group on Whatsapp) |
| Ground-Truth | [IN: CREATE_CALL [SL:GROUP Haustier - Adoptions ] [SL:NAME_APP Whatsapp ] ] |
| Baseline | [IN:CREATE_CALL [SL:GROUP Haustier - Adoptionen ] [SL:NAME_APP Whatsapp ] ] ✗ |
| Zero-MParser | [IN:CREATE_CALL [SL:GROUP Haustier - Adoptions ] [SL:NAME_APP Whatsapp ] ] ✓ |
| Input | Rufe meine Haustier - Adoptions - Gruppe auf Whatsapp an (Call my pet adoption group on Whatsapp) |
| Ground-Truth | [IN:CREATE_REMINDER [SL:PERSON_REMINDED mich ] [SL:DATE_TIME 30 Minuten vor ] [SL:TODO [IN:GET_TODO [SL:TODO meinem geplanten Anruf ] [SL:DATE_TIME am Nachmittag ] ] ] ] |
| Baseline | [IN:CREATE_REMINDER [SL:PERSON_REMINDED mich ] [SL:DATE_TIME 30 Minuten vor ] [SL:TODO [IN:CREATE_CALL ] ] [SL:DATE_TIME am Nachmittag ] ] ✗ |
| Zero-MParser | [IN:CREATE_REMINDER [SL:PERSON_REMINDED mich ] [SL:DATE_TIME 30 Minuten vor ] [SL:TODO [IN:GET_TODO [SL:TODO meinem geplanten Anruf ] [SL:DATE_TIME am Nachmittag ] ] ] ] ✓ |

Table 11: Comparison of outputs of baseline (TAF) and our Zero-MParser model on the validation set of MTOP dataset.

Despite the advancements over the baseline, our method still makes some mistakes. To better understand the limitations of our approach, we undertake an error evaluation to scrutinize the predictions generated by our Zero-MParser framework in comparison to the gold data. We categorize errors into five primary types: Wrong Intent, Wrong Slot Label, Wrong Slot Span, Extra Slot, and Missing Slot. Examples of these error categories, along with their corresponding percentages in the overall errors, are outlined in Table 12. The analysis highlights that the predominant error category is "Wrong Span Prediction," indicating accurate schema prediction but incorrect spans within each slot. This highlights an area for future research to improve the model's ability to precisely locate and extract slot spans,

potentially through enhanced span detection mechanisms or more effective alignment strategies.

| Error | Example |
|---|---|
| Wrong Slot Span (43.3%) | Utterance: Wer arbeitet bei Long John Silver 's ? (Who works at Long John Silvers ?)<br>Prediction: [IN:GET_CONTACT [SL:EMPLOYER Long John Silver ] ]<br>Gold: [IN:GET_CONTACT [SL:EMPLOYER Long John Silver 's ] ] |
| Extra Slot (16.9%) | Utterance: Schick eine Videonachricht an den Smoothie - Chat (send a vide message to the smoothie chat)<br>Prediction: [IN:SEND_MESSAGE [SL:TYPE_CONTENT Videonachricht ] [SL:GROUP Smoothie ] ]<br>Gold: [IN:SEND_MESSAGE [SL:GROUP Smoothie ] ] |
| Wrong Intent (15.0%) | Utterance: Starte meinen Timer neu (start my timer over timer)<br>Prediction: [IN:RESTART_TIMER [SL:METHOD_TIMER Timer ] ]<br>Gold: [IN:RESUME_TIMER [SL:METHOD_TIMER Timer ] ] |
| Missing Slot (14.7%) | Utterance: Spiele einen bestimmten Rap - Künstler (Play a certain rap artist)<br>Prediction: [IN:PLAY_MUSIC [SL:MUSIC_GENRE Rap ] ]<br>Gold: [IN:PLAY_MUSIC [SL:MUSIC_GENRE Rap ] [SL:MUSIC_TYPE Künstler ] ] |
| Wrong Slot Label (10.1%) | Utterance: Welche Neuigkeiten gibt es in der Musikbranche ? (what's the news in the music industry)<br>Prediction: [IN:GET_STORIES_NEWS [SL:NEWS_TYPE Neuigkeiten ] [SL:NEWS_TOPIC Musikbranche ] ]<br>Gold: [IN:GET_STORIES_NEWS [SL:NEWS_TYPE Neuigkeiten ] [SL:NEWS_CATEGORY Musikbranche ] ] |

Table 12: Examples of error categories on the dev set of MTOP dataset.

## 6. Limitations

There are three main limitations to our work as follows:

- **(I)** We observed that larger LLMs tend to achieve better performance. However, utilizing larger LLMs demands substantial computing resources, which may not be accessible or feasible for all researchers or organizations. Future work should focus on making LLMs more efficient and capable of running on smaller devices.

- **(II)** Our experiments primarily evaluated the framework on five non-English languages, including German, French, Spanish, Hindi, and Thai. While these languages provide valuable insights, it is essential

to extend the evaluation to include a broader range of languages. This remains an important avenue for future research.

- **(III)** Our framework relies on LLMs to generate augmented data, but not all generated text is usable. There is some noisy text that needs to be filtered out. Further research is necessary to understand why LLMs produce incorrect samples and how to enable LLMs to return better results, thereby potentially improving the overall effectiveness of our method.

## 7. Conclusion

In conclusion, this paper addresses the challenges of zero-shot multilingual semantic parsing by introducing the Zero-MParser framework, which leverages LLMs to enhance performance. Through three key phases—LLM-based augmentation, noisy data filtering, and multilingual semantic parsing—our framework achieves state-of-the-art results in zero-shot settings, notably surpassing leading methods by 1.43 points in exact match scores on the MTOP dataset. Additionally, we explore the framework's performance under the constraint of limited English annotations, revealing the significant impact of sample selection on model effectiveness. The development of a selective compression method for semantic parsing datasets allows us to achieve 80% of full dataset performance using just 5% of annotated data. This not only demonstrates the framework's robustness but also provides valuable guidance for annotators dealing with minimal resources. In summary, our contributions lie in presenting a holistic framework, introducing an efficient data selection method, and validating the effectiveness of the entire approach through extensive experiments on MTOP and MASSIVE datasets.

## Acknowledgments

This work is supported partly by AOARD grant FA23862214039. The views and conclusions contained herein are those of the authors only and should not be interpreted as representing those of the U.S. Government.

## References

An, S., Lin, Z., Fu, Q., Chen, B., Zheng, N., Lou, J.G., Zhang, D., 2023. How do in-context examples affect compositional generalization?,

in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 11027–11052. URL: https://aclanthology.org/2023.acl-long.618.

Awasthi, A., Gupta, N., Samanta, B., Dave, S., Sarawagi, S., Talukdar, P., 2023. Bootstrapping multilingual semantic parsers using large language models, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia. pp. 2455–2467. URL: https://aclanthology.org/2023.eacl-main.180.

Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.

Campagna, G., Xu, S., Moradshahi, M., Socher, R., Lam, M.S., 2019. Genie: A generator of natural language semantic parsers for virtual assistant commands, in: Proceedings of the 40th ACM SIGPLAN Conference on Programming Language Design and Implementation, pp. 394–410.

Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al., 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 URL: https://arxiv.org/abs/2107.03374.

Chen, X., Ghoshal, A., Mehdad, Y., Zettlemoyer, L., Gupta, S., 2020. Low-resource domain adaptation for compositional task-oriented semantic parsing, in: Webber, B., Cohn, T., He, Y., Liu, Y. (Eds.), Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP), Association for Computational Linguistics, Online. pp. 5090–5100. doi:10.18653/v1/2020.emnlp-main.413.

Conneau, A., Khandelwal, K., Goyal, N., Chaudhary, V., Wenzek, G., Guzmán, F., Grave, E., Ott, M., Zettlemoyer, L., Stoyanov, V., 2020. Unsupervised cross-lingual representation learning at scale, in: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, Association for Computational Linguistics, Online.

pp. 8440–8451. URL: https://aclanthology.org/2020.acl-main.747, doi:10.18653/v1/2020.acl-main.747.

Do, D.T., Nguyen, M.P., Nguyen, L.M., 2023a. Gram: Grammar-based refined-label representing mechanism in the hierarchical semantic parsing task, in: International Conference on Applications of Natural Language to Information Systems, Springer. pp. 339–351.

Do, T., Nguyen, P., Nguyen, M., 2023b. StructSP: Efficient fine-tuning of task-oriented dialog system by using structure-aware boosting and grammar constraints, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Findings of the Association for Computational Linguistics: ACL 2023, Association for Computational Linguistics, Toronto, Canada. pp. 10206–10220. URL: https://aclanthology.org/2023.findings-acl.648, doi:10.18653/v1/2023.findings-acl.648.

Fei, H., Li, B., Liu, Q., Bing, L., Li, F., Chua, T.S., 2023. Reasoning implicit sentiment with chain-of-thought prompting, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), Association for Computational Linguistics, Toronto, Canada. pp. 1171–1182. doi:10.18653/v1/2023.acl-short.101.

Fei, H., Wu, S., Ji, W., Zhang, H., Zhang, M., Lee, M.L., Hsu, W., 2024. Video-of-thought: Step-by-step video reasoning from perception to cognition, in: International Conference on Machine Learning. URL: https://api.semanticscholar.org/CorpusID:270556477.

FitzGerald, J., Hench, C., Peris, C., Mackie, S., Rottmann, K., Sanchez, A., Nash, A., Urbach, L., Kakarala, V., Singh, R., Ranganath, S., Crist, L., Britan, M., Leeuwis, W., Tur, G., Natarajan, P., 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), Association for Computational Linguistics, Toronto, Canada. pp. 4277–4302. URL: https://aclanthology.org/2023.acl-long.235, doi:10.18653/v1/2023.acl-long.235.

Gritta, M., Hu, R., Iacobacci, I., 2022. CrossAligner & co: Zero-shot transfer methods for task-oriented cross-lingual natural language under-

standing, in: Findings of the Association for Computational Linguistics: ACL 2022, Association for Computational Linguistics, Dublin, Ireland. pp. 4048–4061. URL: https://aclanthology.org/2022.findings-acl.319, doi:10.18653/v1/2022.findings-acl.319.

Kingma, D.P., Ba, J., 2014. Adam: A method for stochastic optimization. arXiv preprint arXiv:1412.6980 .

Levy, I., Bogin, B., Berant, J., 2023. Diverse demonstrations improve in-context compositional generalization, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 1401–1422. URL: https://aclanthology.org/2023.acl-long.78.

Li, H., Arora, A., Chen, S., Gupta, A., Gupta, S., Mehdad, Y., 2021. MTOP: A comprehensive multilingual task-oriented semantic parsing benchmark, in: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, Association for Computational Linguistics, Online. pp. 2950–2962. URL: https://aclanthology.org/2021.eacl-main.257, doi:10.18653/v1/2021.eacl-main.257.

Li, H., Graesser, A.C., Cai, Z., 2014. Comparison of google translation with human translation, in: the twenty-seventh international flairs conference. URL: https://cdn.aaai.org/ocs/7864/7864-36722-1-PB.pdf.

Li, Z., Wu, Y., Peng, B., Chen, X., Sun, Z., Liu, Y., Paul, D., 2022. Setransformer: A transformer-based code semantic parser for code comment generation. IEEE Transactions on Reliability 72, 258–273.

Lin, C.Y., 2004. Rouge: A package for automatic evaluation of summaries, in: Text summarization branches out, pp. 74–81.

Mansimov, E., Zhang, Y., 2022. Semantic parsing in task-oriented dialog with recursive insertion-based encoder, in: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 11067–11075. doi:https://doi.org/10.1609/aaai.v36i10.21355.

Mekala, D., Wolfe, J., Roy, S., 2023. ZEROTOP: Zero-shot task-oriented semantic parsing using large language models, in:

Bouamor, H., Pino, J., Bali, K. (Eds.), Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing, Association for Computational Linguistics, Singapore. pp. 5792–5799. URL: https://aclanthology.org/2023.emnlp-main.354, doi:10.18653/v1/2023.emnlp-main.354.

Muennighoff, N., Wang, T., Sutawika, L., Roberts, A., Biderman, S., Le Scao, T., Bari, M.S., Shen, S., Yong, Z.X., Schoelkopf, H., Tang, X., Radev, D., Aji, A.F., Almubarak, K., Albanie, S., Alyafeai, Z., Webson, A., Raff, E., Raffel, C., 2023. Crosslingual generalization through multitask finetuning, in: Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 15991–16111. URL: https://aclanthology.org/2023.acl-long.891, doi:10.18653/v1/2023.acl-long.891.

Nicosia, M., Piccinno, F., 2022. Byte-level massively multilingual semantic parsing, in: FitzGerald, J., Rottmann, K., Hirschberg, J., Bansal, M., Rumshisky, A., Peris, C., Hench, C. (Eds.), Proceedings of the Massively Multilingual Natural Language Understanding Workshop (MMNLU-22), Association for Computational Linguistics, Abu Dhabi, United Arab Emirates (Hybrid). pp. 25–34. URL: https://aclanthology.org/2022.mmnlu-1.3, doi:10.18653/v1/2022.mmnlu-1.3.

Nicosia, M., Qu, Z., Altun, Y., 2021. Translate & Fill: Improving zero-shot multilingual semantic parsing with synthetic data, in: Findings of the Association for Computational Linguistics: EMNLP 2021, Association for Computational Linguistics, Punta Cana, Dominican Republic. pp. 3272–3284. URL: https://aclanthology.org/2021.findings-emnlp.279, doi:10.18653/v1/2021.findings-emnlp.279.

Raffel, C., Shazeer, N., Roberts, A., Lee, K., Narang, S., Matena, M., Zhou, Y., Li, W., Liu, P.J., 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. The Journal of Machine Learning Research 21, 5485–5551. URL: https://dl.acm.org/doi/abs/10.5555/3455716.3455856.

Raventós, A., Paul, M., Chen, F., Ganguli, S., 2024. Pretraining task diver-

sity and the emergence of non-bayesian in-context learning for regression. Advances in Neural Information Processing Systems 36.

Reimers, N., Gurevych, I., 2019. Sentence-BERT: Sentence embeddings using Siamese BERT-networks, in: Inui, K., Jiang, J., Ng, V., Wan, X. (Eds.), Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP), Association for Computational Linguistics, Hong Kong, China. pp. 3982–3992. URL: https://aclanthology.org/D19-1410, doi:10.18653/v1/D19-1410.

Sherborne, T., Hosking, T., Lapata, M., 2023. Optimal transport posterior alignment for cross-lingual semantic parsing. Transactions of the Association for Computational Linguistics 11, 1432–1450. URL: https://aclanthology.org/2023.tacl-1.81.

Sherborne, T., Lapata, M., 2022. Zero-shot cross-lingual semantic parsing, in: Muresan, S., Nakov, P., Villavicencio, A. (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Dublin, Ireland. pp. 4134–4153. doi:10.18653/v1/2022.acl-long.285.

Sherborne, T., Lapata, M., 2023. Meta-Learning a Cross-lingual Manifold for Semantic Parsing. Transactions of the Association for Computational Linguistics 11, 49–67. URL: https://doi.org/10.1162/tacl_a_00533.

Shi, P., Song, L., Jin, L., Mi, H., Bai, H., Lin, J., Yu, D., 2022. Cross-lingual text-to-SQL semantic parsing with representation mixup, in: Findings of the Association for Computational Linguistics: EMNLP 2022, Association for Computational Linguistics, Abu Dhabi, United Arab Emirates. pp. 5296–5306. URL: https://aclanthology.org/2022.findings-emnlp.388, doi:10.18653/v1/2022.findings-emnlp.388.

Shin, R., Van Durme, B., 2022. Few-shot semantic parsing with language models trained on code, in: Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Seattle, United States. pp. 5417–5425. URL: https://aclanthology.org/2022.naacl-main.396.

Tang, Y., Tran, C., Li, X., Chen, P.J., Goyal, N., Chaudhary, V., Gu, J., Fan, A., 2020. Multilingual translation with extensible multilingual pretraining and finetuning. arXiv preprint arXiv:2008.00401 URL: https://arxiv.org/abs/2008.00401.

Touvron, H., Martin, L., Stone, K., Albert, P., Almahairi, A., Babaei, Y., Bashlykov, N., Batra, S., Bhargava, P., Bhosale, S., et al., 2023. Llama 2: Open foundation and fine-tuned chat models. arXiv preprint arXiv:2307.09288 .

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need. Advances in neural information processing systems 30.

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E.H., Le, Q.V., Zhou, D., 2024. Chain-of-thought prompting elicits reasoning in large language models, in: Proceedings of the 36th International Conference on Neural Information Processing Systems, Curran Associates Inc., Red Hook, NY, USA. URL: https://dl.acm.org/doi/10.5555/3600270.3602070.

Wu, S., Xin, C., Lin, H., Han, X., Liu, C., Chen, J., Yang, F., Wan, G., Sun, L., 2023a. Ambiguous learning from retrieval: Towards zero-shot semantic parsing, in: Rogers, A., Boyd-Graber, J., Okazaki, N. (Eds.), Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Toronto, Canada. pp. 14081–14094. doi:10.18653/v1/2023.acl-long.787.

Wu, T., He, S., Liu, J., Sun, S., Liu, K., Han, Q.L., Tang, Y., 2023b. A brief overview of chatgpt: The history, status quo and potential future development. IEEE/CAA Journal of Automatica Sinica 10, 1122–1136.

Xia, M., Monti, E., 2021. Multilingual neural semantic parsing for low-resourced languages, in: Ku, L.W., Nastase, V., Vulić, I. (Eds.), Proceedings of *SEM 2021: The Tenth Joint Conference on Lexical and Computational Semantics, Association for Computational Linguistics, Online. pp. 185–194. doi:10.18653/v1/2021.starsem-1.17.

Xu, J., Fei, H., Pan, L., Liu, Q., Lee, M.L., Hsu, W., 2024. Faithful logical reasoning via symbolic chain-of-thought, in: Ku, L.W., Martins, A.,

Srikumar, V. (Eds.), Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), Association for Computational Linguistics, Bangkok, Thailand. pp. 13326–13365. doi:10.18653/v1/2024.acl-long.720.

Xue, L., Constant, N., Roberts, A., Kale, M., Al-Rfou, R., Siddhant, A., Barua, A., Raffel, C., 2021. mT5: A massively multilingual pre-trained text-to-text transformer, in: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Association for Computational Linguistics, Online. pp. 483–498. URL: https://aclanthology.org/2021.naacl-main.41, doi:10.18653/v1/2021.naacl-main.41.

Yang, J., Fancellu, F., Webber, B., Yang, D., 2021. Frustratingly simple but surprisingly strong: Using language-independent features for zero-shot cross-lingual semantic parsing, in: Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pp. 5848–5856.

Zhuo, T.Y., Li, Z., Huang, Y., Shiri, F., Wang, W., Haffari, G., Li, Y.F., 2023. On robustness of prompt-based semantic parsing with large pre-trained language model: An empirical study on codex, in: Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics, Association for Computational Linguistics, Dubrovnik, Croatia. pp. 1090–1102. URL: https://aclanthology.org/2023.eacl-main.77, doi:10.18653/v1/2023.eacl-main.77.