XSEMPLR: Cross-Lingual Semantic Parsing in Multiple Natural Languages and Meaning Representations

Yusen Zhang[1] **Jun Wang**[2] **Zhiguo Wang**[2] **Rui Zhang**[1]
[1]Penn State University [2]AWS AI Labs
{yfz5488,rmz5227}@psu.edu, {juwanga,zhiguow}@amazon.com

Abstract

Cross-Lingual Semantic Parsing (CLSP) aims to translate queries in multiple natural languages (NLs) into meaning representations (MRs) such as SQL, lambda calculus, and logic forms. However, existing CLSP models are separately proposed and evaluated on datasets of limited tasks and applications, impeding a comprehensive and unified evaluation of CLSP on a diverse range of NLs and MRs. To this end, we present XSEMPLR, a unified benchmark for cross-lingual semantic parsing featured with 22 natural languages and 8 meaning representations by examining and selecting 9 existing datasets to cover 5 tasks and 164 domains. We use XSEMPLR to conduct a comprehensive benchmark study on a wide range of multilingual language models including encoder-based models (mBERT, XLM-R), encoder-decoder models (mBART, mT5), and decoder-based models (Codex, BLOOM). We design 6 experiment settings covering various lingual combinations (monolingual, multilingual, cross-lingual) and numbers of learning samples (full dataset, few-shot, and zero-shot). Our experiments show that encoder-decoder models (mT5) achieve the highest performance compared with other popular models, and multilingual training can further improve the average performance. Notably, multilingual large language models (e.g., BLOOM) are still inadequate to perform CLSP tasks. We also find that the performance gap between monolingual training and cross-lingual transfer learning is still significant for multilingual models, though it can be mitigated by cross-lingual few-shot training. Our dataset and code are available at https://github.com/psunlpgroup/XSemPLR.

1 Introduction

Cross-Lingual Semantic Parsing (CLSP) aims to translate queries in multiple natural languages (NLs) into meaning representations (MRs) (Li et al., 2020; Xu et al., 2020a; Dou et al., 2022; Sherborne and Lapata, 2021, 2022). As demonstrated in Figure 1, Cross-Lingual Semantic Parsing covers natural languages for geographically diverse users and various meaning representations, empowering applications such as natural language interfaces to databases, question answering over knowledge graphs, virtual assistants, smart home device control, human-robot interaction, and code generation.

However, current research on CLSP has three drawbacks. First, most existing research focuses on semantic parsing in English (Zelle and Mooney, 1996; Wang et al., 2015; Yu et al., 2018), limiting the development of multilingual information access systems for users in other languages. Second, current datasets have a poor coverage of NLs and MRs. Although there are encouraging efforts in developing CLSP models (Li et al., 2020; Dou et al., 2022; Sherborne and Lapata, 2022), their experiments only cover a few NLs and MRs, impeding comprehensive and unified evaluation on a diverse range of tasks. Third, due to the lack of a comprehensive CLSP benchmark, the performance of multilingual language models on CLSP is understudied. Some pretrained language models are proposed to solve cross-lingual tasks such as XLM-R (Conneau et al., 2019) and mT5 (Xue et al., 2020), while other large language models are designed for code generation such as Codex (Chen et al., 2021a) and BLOOM (Scao et al., 2022). However, little research has focused on evaluating models on CLSP.

In this paper, we propose XSEMPLR, a unified benchmark for cross-lingual semantic parsing featured with 22 natural languages and 8 meaning representations as summarized in Table 1. In order to cover a large variety of languages and meaning representations, we first select 9 high-quality CLSP datasets and then clean and format them in a unified manner. Then, we conduct a comprehensive benchmarking study on three categories of multilingual language models including pretrained encoder-

Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics
Volume 1: Long Papers, pages 15918–15947

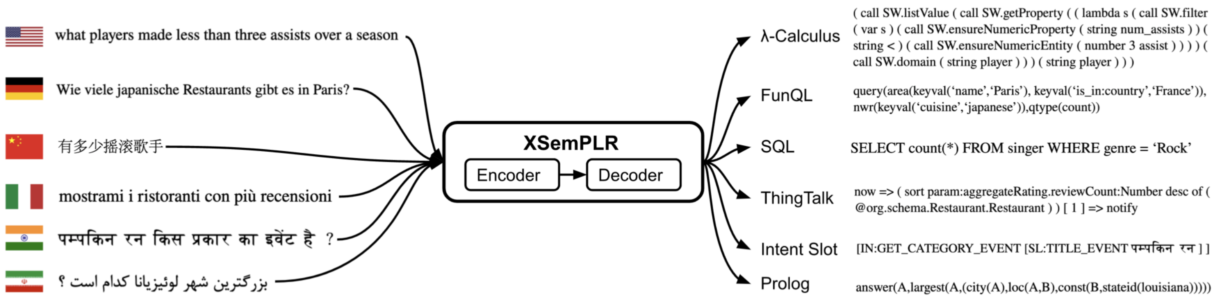

Figure 1: Overview of Cross-Lingual Semantic Parsing over various natural languages and meaning representations.

based models augmented with pointer generator (mBERT, XLM-R), pretrained encoder-decoder models (mBART, mT5), and decoder-based large language models (Codex, BLOOM). To evaluate these models, we design 6 experiment settings covering various lingual combinations and learning sample scales, including Monolingual (and Monolingual Few-shot), Multilingual, and Cross-lingual Zero-Shot/Few-Shot Transfer.

Our results show that the encoder-decoder model (mT5) yields the best performance on monolingual evaluation compared with other models. Then, we pick two models with the best monolingual performance (i.e., mT5 and XLM-R) to conduct few-shot and zero-shot cross-lingual transfer learning from English to other low-resource languages. Results show a significant performance gap between monolingual training (Taget NL -> Target NL[1]) and cross-lingual transfer learning (En -> Target NL). Furthermore, we find that this gap can be significantly reduced by few-shot learning on target NL. We further train these two models in a multilingual setting and find such training can boost the performance in some of the languages, while, however, it usually hurts the performance in English. Finally, we test two large language models Codex (Chen et al., 2021a) and BLOOM (Scao et al., 2022). We find the performance gap of cross-lingual transfer learning is significant for these two models as well.

Our contributions are summarized as follows: (1) We propose XSEMPLR to unify and benchmark 9 datasets covering 5 tasks, 22 natural languages, and 8 meaning representations for cross-lingual semantic parsing; (2) We perform a holistic evaluation of 3 groups of state-of-the-art multilingual language models on XSEMPLR, demonstrating noticeable performance gaps of cross-lingual transfer models comparing English and other languages; (3)

[1]We use A -> B to denote the model finetuned on NL A and tested on NL B.

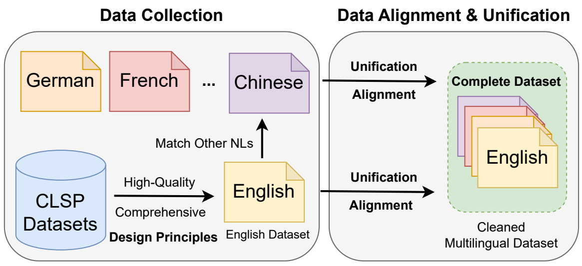

Figure 2: Construction pipeline of XSEMPLR.

We show two effective strategies for boosting performance in low-resource languages: multilingual training and cross-lingual transfer learning.

2 XSEMPLR Benchmark

Figure 2 shows the construction pipeline of XSEMPLR. We first select 9 CLSP datasets according to our design principles. Then, we collect other NLs of the selected datasets. Finally, we clean the datasets by removing outliers and performing alignment between different languages.

2.1 Design Principles

We carefully pick 9 datasets from all available semantic parsing datasets to construct XSEMPLR according to two principles. First, the picked datasets need to have **high quality**, which means they are either annotated by humans or augmented with careful crafting (Moradshahi et al., 2020), and the translation of user inputs are provided by humans instead of machine translation models. Second, XSEMPLR needs to be **comprehensive** (Hu et al., 2020), which means including diverse NLs and MRs for a broad range of tasks and applications.

2.2 Data Collection

Table 1 summarizes the characteristics and statistics of different datasets in XSEMPLR.

Multilingual ATIS (MATIS) contains user questions for a flight-booking task. We collect the origi-

Task	Dataset	Meaning Representation	Language	Executable	Domain	Train	Dev	Test
NLI for Databases	MATIS	SQL	7	✓	1	4303	481	444
NLI for Databases	MGeoQuery	SQL,Lambda,FunQL,Prolog	8	✓	1	548	49	277
NLI for Databases	MSpider	SQL	3	✓	138	8095	1034	–
NLI for Databases	MNLmaps	Functional Query Language	2	✓	1	1500	–	880
QA on Knowledge Graph	MOvernight	Lambda Calculus	3	✓	8	8754	2188	2740
QA on Knowledge Graph	MCWQ	SPARQL	4	✓	1	4006	733	648
QA on Web	MSchema2QA	ThingTalk Query Language	11	✓	2	8932	–	971
Task-Oriented DST	MTOP	Hierarchical Intent and Slot	6	✗	11	5446	863	1245
Code Generation	MCoNaLa	Python	4	✓	1	1903	476	896

Table 1: Datasets in XSEMPLR. We assemble 9 datasets in various domains for 5 semantic parsing tasks. It covers 8 meaning representations. The questions cover 22 languages in 15 language families. Train/Dev/Test columns indicate the number of MRs each paired with multiple NLs.

nal English questions from ATIS (Price, 1990; Dahl et al., 1994) and add the translations from Xu et al. (2020b). For MRs, we focus on the task of Natural Language Interface (NLI) to databases and thus collect SQL from Iyer et al. (2017) and Finegan-Dollak et al. (2018).

Multilingual GeoQuery (MGeoQuery) contains user questions about US geography. We collect original English questions from GeoQuery (Zelle and Mooney, 1996) and add other translations (Lu and Ng, 2011; Jones et al., 2012; Susanto and Lu, 2017b). GeoQuery has several MRs available. We collect Prolog and Lambda Calculus from Guo et al. (2020), FunQL from Susanto and Lu (2017b), and SQL from Finegan-Dollak et al. (2018) [2].

Multilingual Spider (MSpider) is a human-annotated complex and cross-domain text-to-SQL datasets. We collect Spider (Yu et al., 2018) with English questions and add other NLs from Min et al. (2019) and Nguyen et al. (2020).

Multilingual NLmaps (MNLmaps) is a Natural Language Interface to query the OpenStreetMap database. We collect NLMaps (Lawrence and Riezler, 2016) in English, and add translations in German (Haas and Riezler, 2016).

Multilingual Overnight (MOvernight) is a multi-domain semantic parsing dataset in lambda DCS. We include English Overnight (Wang et al., 2015) and add translations from Sherborne et al. (2020).

Multilingual Schema2QA (MSchema2QA) is a question answering dataset over schema.org web data in ThingTalk Query Language. We include training examples with all 11 available languages and pair them with the MR in the corresponding language following Moradshahi et al. (2020) and Xu et al. (2020a). To make the dataset size comparable to others, we include 5% of the training set.

MCWQ is a multilingual knowledge-based question answering dataset grounded in Wikidata (Cui et al., 2021). We collect all questions in MCWQ in 4 languages. The split follows maximum compound divergence (MCD) (Keysers et al., 2020) so that the test set contains novel compounds to evaluate compositionality generalization ability.

MTOP is a multilingual semantic parsing dataset for task-oriented dialogs with meaning representations of hierarchical intent and slot annotations (Gupta et al., 2018; Li et al., 2020). We include examples with all 6 languages and pair the translations with the compositional decoupled representation in the corresponding language.

MCoNaLa is a multilingual code generation benchmark for Python by extending English CoNaLa (Yin et al., 2018; Wang et al., 2022). We include all 4 languages.

2.3 Data Alignment and Unification

We perform data alignment and unification over 9 datasets to construct a unified high-quality benchmark. To be specific, for the first 6 datasets introduced in Section 2.2, because each of them has multiple parts proposed in different work, we merge these parts by aligning the same user question in different languages into the same meaning representation. For the other 3 datasets, we directly use the entire samples since no other parts need to be merged. We also try to unify the language of MRs (e.g., adopting a single form of SQL queries; keeping only one English MR when there is more than one in MTOP). We also remove a few samples in MATIS and MGeoQuery with no MRs. We provide more details in Appendix including the examples of each dataset (Table 5), data construction (Ap-

[2] We report averaged scores of 4 MRs in the tables, unless otherwise specified.

pendix A), natural languages (Appendix A), and meaning representations (Appendix A).

2.4 Evaluation Metrics

We evaluate the predicted results using various automatic metrics. For the Spider dataset, we follow Yu et al. (2018) and use their proposed tool for evaluation [3]. For the other datasets, we simply use exact matching, i.e., token-by-token string comparison, to see if the prediction is the same as the ground truth label. For a fair comparison with state-of-the-art models, we also use the metrics proposed in their models, including Execution Score, Denotation Accuracy, and Code BLEU (Section 4.2).

2.5 Data Analysis

Natural Languages XSEMPLR contains diverse and abundant natural languages in both high-resource and low-resource groups, including 22 languages belonging to 15 language families (Appendix A). Most state-of-the-art performances are achieved in English and a few other high-resource languages. However, the lack of information in the low-resource languages brings unanswered questions to model generalization. Therefore, both these 2 types of languages are included in XSEMPLR, to form a unified cross-lingual dataset for semantic parsing. Among these 22 languages, English is the most resourced language with many popular datasets in semantic parsing. Some languages spoken in Western Europe are also relatively high-resource languages, such as German and Spanish. We also involve many low-resource languages as well, such as Vietnamese and Thai.

Meaning Representations XSEMPLR includes 8 meaning representations for different applications: Prolog, Lambda Calculus, Functional Query Language (FunQL), SQL, ThingTalk Query Language, SPARQL, Python, and Hierarchical intent and slot. All of them can be executed against underlying databases or knowledge graphs, except for the last one which is designed for complex compositional requests in task-oriented dialogues. The first four are domain-specific because they contain specific predicates defined for a given domain, while the last four are considered open-domain and open-ontology (Guo et al., 2020). It is also worth noting that these MRs are not equivalent to their general expressiveness. For example, the ThingTalk query language is a subset of SQL in expressiveness (Moradshahi et al., 2020), and FunQL is less expressive than Lambda Calculus partially due to the lack of variables and quantifiers.

[3]All numbers reported in the paper is "Exact Set Match without Values" in https://yale-lily.github.io/spider.

3 Experiment Setup

We describe our evaluation settings and models for a comprehensive benchmark study on XSEMPLR.

3.1 Evaluation Settings

We consider the following 6 settings for training and testing.

Translate-Test. We train a model on the English training data and translate target NL test data to English using the public Google NMT system (Wu et al., 2016). This setting uses one semantic parsing model trained on English but also relies on available machine translation models for other languages. This serves as a strong yet practical baseline for other settings.

Monolingual. We train a monolingual model on each target NL training data. This setting creates one model per target NL. In addition to benchmarking them, we design this setting for two reasons: (1) It helps the comparison between monolingual and cross-lingual performance; (2) We pick the best models from this setting to further conduct cross-lingual and few-shot/zero-shot experiments. Additionally, since some target NL training data can be expensive to obtain, we also test a **Monolingual Few-shot** setting by training monolingual models with only 10% training data.

Multilingual. Thanks to the progress in multilingual embeddings and pretrained multilingual language models, we can train one multilingual model on all NL training data. This setting uses only one model to serve all NLs.

Cross-lingual Zero-shot Transfer. Models are trained only on English NL data and then tested on a target-NL test set. This setting uses one model for all target NLs and evaluates the cross-lingual transfer ability without any target-NL training data. Besides, to test the value of additional target NL training data, we finetune the model on 10% target-NL training data. This **Cross-lingual Few-shot Transfer** setting creates one model per target NL. We use these two settings to evaluate the capability

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa‡	Average
Translate-Test										
mT5	44.50	53.88	45.26	66.36	59.69	19.85	3.18★	29.78★	8.13	36.74
Monolingual										
mBERT+PTR	30.63	72.18	40.40	83.82	57.47	23.46	52.53	75.41	5.87	49.09
XLM-R+PTR	31.31	71.41	47.30	85.17	59.10	23.53	62.37	80.36	7.69	52.03
mBART	41.93	62.29	33.31	83.19	59.60	30.02	50.35	75.76	6.78	49.25
mT5	**53.15**	**74.26**	**50.73**	**91.65**	**66.29**	**30.15**	**65.16**	**81.83**	**10.29**	**58.16**
Monolingual Few-Shot										
XLM-R+PTR	23.44	17.91	36.04	19.77	40.74	5.64	**49.00**	60.42	0.38	28.15
mT5	**24.85**	25.48	**38.10**	26.93	**53.59**	**7.68**	33.27	**61.90**	1.05	**30.32**
Codex†	18.02	**31.93**	30.66	**34.26**	3.43	2.93	21.62	10.08	**13.87**	18.53
BLOOM†	0.00	17.84	2.13	12.16	0.62	0.00	5.21	5.16	8.40	5.72
Multilingual										
XLM-R+PTR	39.72	71.35	**40.20**	85.91	61.03	**30.79**	**61.82**	81.68	–	59.06
mT5	**54.45**	**76.57**	32.30	**91.31**	**67.55**	28.51	60.92	**82.95**	–	**61.82**
Cross-lingual Zero-Shot Transfer										
XLM-R+PTR	6.05	**39.85**	18.53	**60.23**	36.77	**4.27**	20.22	51.46	0.12	26.39
mT5	**31.85**	27.35	**41.93**	34.89	**52.68**	4.06	**44.04**	**50.18**	0.77	**31.97**
Codex†	16.31	28.53	27.56	32.05	2.99	2.16	19.57	14.08	**8.35**	16.84
BLOOM†	0.00	11.29	1.70	7.05	0.38	0.00	3.93	1.67	6.16	3.58
Cross-lingual Few-Shot Transfer										
XLM-R+PTR	15.71	51.08	43.68	64.89	52.03	20.16	53.51	72.79	–	46.73
mT5	**49.57**	**57.31**	**49.42**	**71.70**	**62.53**	**24.85**	**59.24**	**74.83**	–	**56.18**

Table 2: Results on XSEMPLR. We consider 6 settings including 2 Monolingual, 1 Multilingual, and 2 Cross-lingual settings, and one Translate-Test setting. Each number is averaged across different languages in that dataset. † Codex/BLOOM are evaluated on only two settings as we apply 8-shot in-context learning without finetuning the model parameters. ‡ Two settings are not applicable to MCoNaLa because it has no training set on NLs other than English. ★ Translate-Test performances on MSchem2QA and MTOP are especially low because the MR of these data also contains tokens in target languages.

of the model to transfer from a fine-tuned model of high-resource NL to a low-resource test set.

3.2 Models

We evaluate three different groups of multilingual language models on XSEMPLR.

Multilingual Pretrained Encoders with Pointer-based Decoders (Enc-PTR). The first group is multilingual pretrained encoders with decoders augmented with pointers. Both encoders and decoders use Transformers (Vaswani et al., 2017). The decoder uses pointers to copy entities from natural language inputs to generate meaning representations (Rongali et al., 2020; Prakash et al., 2020). We use two types of multilingual pretrained encoders, mBERT (Devlin et al., 2018) and XLM-R (Conneau et al., 2019), and both are trained on web data covering over 100 languages.

Multilingual Pretrained Encoder-Decoder Models (Enc-Dec). The second group uses pretrained encoder-decoder models, including mBART (Chipman et al., 2022) and mT5 (Xue et al., 2020) which uses text-to-text denoising objective for pretraining over multilingual corpora.

Multilingual Large Language Models (LLMs). The third group is multilingual large language models based on GPT (Brown et al., 2020) including Codex (Chen et al., 2021a) and BLOOM (Scao et al., 2022). Codex is fine-tuned on publicly available code from GitHub. While it is not trained on a multilingual corpus, it has shown cross-lingual semantic parsing capabilities (Shi et al., 2022b). BLOOM is a 176B-parameter multilingual language model pretrained on 46 natural and 13 programming languages from the ROOTS corpus (Laurençon et al., 2022). We mainly use these models to evaluate the ability of few-shot learning using in-context learning without any further finetuning. Specifically, we append 8 samples and the test query to predict the MR. For Monolingual Few-shot, samples and the query are in the same NL, while for Cross-lingual Zero-shot Transfer, samples are in English and the query is in the target NL.

4 Results and Analysis

Table 2 shows the performance of all 6 models on 6 settings. Our results and analysis aim to answer the following research questions:

- RQ 1: What is the best model and training strategy for performance, and how does it compare with previous state-of-the-art? (Section 4.1, 4.2)
- RQ 2: How capable are the current multilingual LLMs on the task of CLSP? (Section 4.3)
- RQ 3: What is the effect of few-shot learning? (Section 4.4)
- RQ 4: What is the effect of multilingual learning? (Section 4.5)
- RQ 5: What is the effect of cross-lingual transfer learning? (Section 4.6)
- RQ 6: How performance varies across different natural languages and meaning representations? (Section 4.7, 4.8)

4.1 Analysis of Monolingual

We obtain the following main findings on Monolingual setting:

Enc-Dec (mT5) obtains the best performance. Among the two transformer-based pointer generators, XLM-R+Transformer (XLM-R+PTR) (52.03[4]) performs slightly better than mBERT+Transformer (mBERT+PTR) (49.09). Among mBART and mT5, mT5 (58.16) outperforms mBART (49.25) by a large margin. Besides, although mT5 outperforms XLM-R by 6.13, XLM-R is still able to outperform mBART by 2.78. Thus, we pick mT5 among mT5/mBART, and XLM-R among XLM-R/mBERT to conduct the experiments on the other settings.

Next, we evaluate mT5 model on Translation-Test setting. As shown in the table, mT5 in Monolingual setting outperforms Translation-Test by a large margin (58.16 vs. 36.74). This shows that multilingual language models are more effective than Translation-Test methods. In other words, it is necessary to train a multilingual model even though we have a high-quality translation system.

4.2 Comparison with SOTA

Table 3 lists the performance of mT5 in Monolingual setting with the previous state-of-the-art. Some of the previous work use denotation accuracy and execution accuracy which are different from the exact match we use. To make our results comparable with previous work, we apply the evaluation tools of previous work to XSEMPLR. As shown in the table, Enc-Dec (mT5) outperforms previous work on all NLs of MSchema2QA, MCWQ,

[4]If not specified, the numbers in this section are the averaged exact matching scores across all NLs.

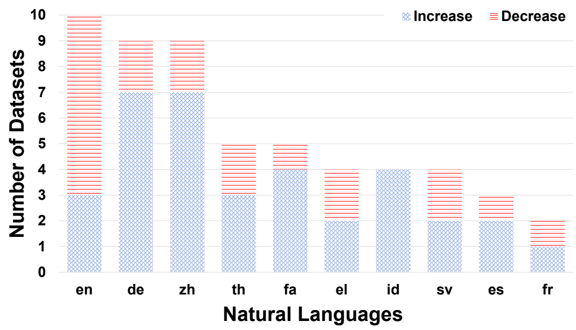

Figure 3: Effect of multilingual training with mT5 on different NLs. X-axis is the NL that was included in at least two datasets. Y-axis is the number of datasets that the performance increases/decreases of this NL after multilingual training. Performance of English (high resource NLs) are easier to drop in multilingual training.

MNLMaps, MATIS datasets and obtains comparable results on the others.

4.3 Analysis of Codex and BLOOM

We evaluate Codex and BLOOM to test the performance of in-context learning of large language models. As shown in Table 2, LLMs (Codex and BLOOM) are outperformed by mT5 model by a large margin for both Few-shot (11.79/24.60) and Zero-shot (15.13/28.39) settings. This suggests that multilingual LLMs are still inadequate for cross-lingual semantic parsing tasks.

4.4 Comparison between Few-shot Settings

We also test the Enc-Dec (mT5) and Enc-PTR (XLM-R) models on two types of few-shot experiments, including Monolingual and Cross-lingual Few-Shot.

As can be seen, mT5 of cross-lingual few-shot outperforms monolingual few-shot by a large 22.21 exact match score (excluding MCoNaLa), while XLM-R has a smaller gain of 15.12. We can summarize two observations: 1) pretraining on the English NL can significantly boost the performance of few-shot on target NLs (En + Target Few-shot -> Target NL), and 2) the model with higher cross-lingual capability gains more improvement, such as mT5 gains more than XLM-R. Both observations demonstrate the capability of cross-lingual models to transfer knowledge from the source to the target NLs.

4.5 Analysis of Multilingual Training

We compare the performance of Monolingual and Multilingual settings. As can be seen in Table 2,

Dataset	Language	SOTA (Source)	XSEMPLR	Metric
MSpider	English	77.10 (Li et al., 2023)	67.60	Exact Match
	English	81.00 (Li et al., 2023)	69.10	Execution
	Vietnamese	69.00 (Shi et al., 2022a)	43.00	Exact Match
	Vietnamese	64.50 (Shi et al., 2022a)	42.00	Execution
	Chinese	66.1★ (Shi et al., 2022a)	39.90	Exact Match
MSchema2QA	Arabic	29.17 (Moradshahi et al., 2020)	53.55	Exact Match
	German	51.84 (Moradshahi et al., 2020)	72.19	Exact Match
	Spanish	56.01 (Moradshahi et al., 2020)	68.69	Exact Match
	Farsi	54.88 (Moradshahi et al., 2020)	60.25	Exact Match
	Finnish	52.43 (Moradshahi et al., 2020)	68.28	Exact Match
	Italian	54.87 (Moradshahi et al., 2020)	67.97	Exact Match
	Japanese	46.27 (Moradshahi et al., 2020)	62.41	Exact Match
	Polish	49.69 (Moradshahi et al., 2020)	60.87	Exact Match
	Turkish	56.84 (Moradshahi et al., 2020)	70.03	Exact Match
	Chinese	36.60 (Moradshahi et al., 2020)	56.54	Exact Match
MCWQ	English	27.70 (Cui et al., 2022)	39.29	Exact Match
	Hebrew	16.60 (Cui et al., 2022)	33.02	Exact Match
	Kannada	16.60 (Cui et al., 2022)	23.74	Exact Match
	Chinese	23.00 (Cui et al., 2022)	24.56	Exact Match
MNLMaps	English	85.70 (Duong et al., 2017)	92.73	Exact Match
	German	83.00 (Duong et al., 2017)	90.57	Exact Match
MATIS	English	77.20 (Sherborne and Lapata, 2023)	83.78	Denotation accuracy
	Farsi	67.80 (Sherborne and Lapata, 2023)	80.59	Denotation accuracy
	Portuguese	66.10 (Sherborne and Lapata, 2023)	78.60	Denotation accuracy
	Spanish	64.10 (Sherborne and Lapata, 2023)	76.58	Denotation accuracy
	German	66.60 (Sherborne and Lapata, 2023)	80.63	Denotation accuracy
	Chinese	64.90 (Sherborne and Lapata, 2023)	78.38	Denotation accuracy
MGeoQuery†	English	90.00 (Zou and Lu, 2018)	79.06	Denotation accuracy
	Thai	86.10 (Zou and Lu, 2018)	72.56	Denotation accuracy
	German	76.80 (Zou and Lu, 2018)	73.29	Denotation accuracy
	Greek	83.20 (Zou and Lu, 2018)	76.90	Denotation accuracy
	Chinese	82.10 (Zou and Lu, 2018)	75.81	Denotation accuracy
	Indonesian	83.90 (Zou and Lu, 2018)	80.14	Denotation accuracy
	Swedish	83.90 (Zou and Lu, 2018)	79.78	Denotation accuracy
	Farsi	76.80 (Zou and Lu, 2018)	69.68	Denotation accuracy
MOvernight	English	81.90 (Sherborne and Lapata, 2021)	69.38‡	Denotation accuracy
	German	66.20 (Sherborne and Lapata, 2021)	66.90‡	Denotation accuracy
	Chinese	66.00 (Sherborne and Lapata, 2021)	62.59‡	Denotation accuracy
MCoNaLa	Russian	9.56 (Wang et al., 2022)	6.38	Code BLEU-4
	Spanish	2.64 (Wang et al., 2022)	2.55	Code BLEU-4
	Japanese	9.90 (Wang et al., 2022)	7.66	Code BLEU-4

Table 3: Comparison between mT5 monolingual and state-of-the-art models, except that MCoNaLa dataset uses cross-lingual zero-shot settings because the dataset only contains English training samples. mT5 obtains better or comparable performance on all datasets. ★ Previous SOTA model only contains exact match scores for Chinese. † The SOTA model of MGeoQuery uses Lambda as MR while XSEMPLR uses SQL. ‡ The SOTA model of MOvernight uses denotation accuracy and XSEMPLR uses exact match.

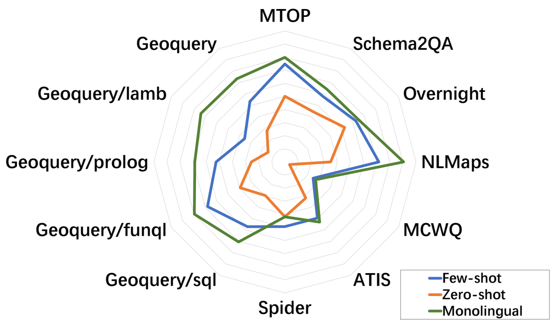

Figure 4: The performance of cross-lingual Few/Zero-shot (mT5) on different datasets and languages. MGeoQuery/* indicates a single MR; MGeoQuery is the averaged score across 4 MRs. Each neighbor grey circle has a 10 score difference, and the center of the circle indicates a 0 score. The cross-lingual transfer performance gap is significant for the zero-shot setting. However, few-shot training shrinks this gap greatly.

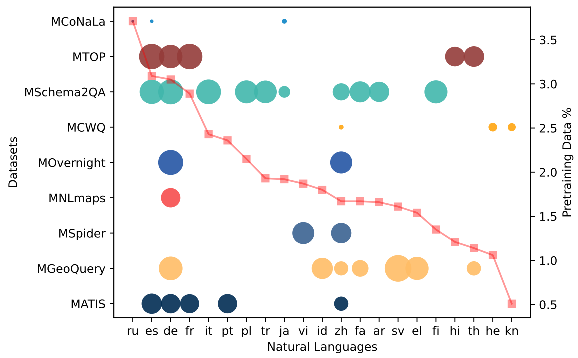

Figure 5: Left vertical axis: The performance of cross-lingual zero-shot mT5 models on different datasets over different languages. Larger dots indicate higher accuracy. Right vertical axis: Red line indicates the percentage of different languages in the mT5 training data. Chinese/German has the largest/smallest performance loss for transfer learning. Additionally, performance and pretraining data size have no evident correlation.

mT5 improves by 2.31 on MGeoQuery, and XLM-R improves by 8.41 on MATIS dataset. This demonstrates that Enc-Dec/Enc-PTR (mT5/XLM-R) can be improved by training in a mixture of various languages. However, not all datasets can boost performance via such training. The average change of mT5/XLM-R is around -2/+2 points.

We further explore the reason for the performance drop in multilingual training. As shown in Figure 3, most of the major NLs can obtain performance gain, except that English performance drops in 7 datasets and gains in 3 datasets. This is known as "Curse of Multilinguality" (Pfeiffer et al., 2022). Similarly in CLSP, performance of English (high resource NLs) is easier to drop in multilingual training.

4.6 Cross-lingual Performance Gap

To examine the transfer ability of the cross-lingual models, we investigate the performance difference between the Monolingual and Cross-lingual Few/Zero-shot for each dataset using mT5. As shown in Figure 4, by examining the distance between green and orange lines, we find that for the zero-shot setting, the cross-lingual transfer performance gap is significant, which is even larger than 50% on the NLmaps dataset, demonstrating the limitation of current cross-lingual models. However, by examining the difference between orange and blue lines, we also find that using even 10% of samples in target data, the transfer gap will be shortened rapidly. The few-shot gap usually shrinks to around half of the zero-shot gap, e.g., the Schema2QA dataset. For MATIS, the gap even shrinks to around 5 which is very close to the performance of the monolingual setting.

4.7 Analysis over Natural Languages

We pick the best model mT5 and analyze its performance in the zero-shot setting in Figure 5. Results show that the performance of Chinese transfer learning (En -> Zh) and English monolingual training (En -> En) usually is the largest compared with transfer learning of other NLs. On the other hand, German usually has the smallest transfer performance loss. This is probably because of two reasons. First, the Chinese data source is less than German when pretraining on mT5. Second, the language family of English is closer to German (IE: Germanic) compared with Chinese (Sino-Tibetan). This phenomenon is discussed in Hu et al. (2020), and we find this conclusion also holds for cross-lingual semantic parsing tasks.

4.8 Analysis over Meaning Representations

Table 4 shows the performance of mT5 on various MRs in MGeoQuery. In almost all languages, FunQL outperforms the other three meaning representations, and SQL obtains the worst performance. This is consistent with the observation of Guo et al. (2020). We speculate that there are two possible reasons: (1) the grammar of SQL is more complex than the others, and FunQL enjoys much easier grammar (Li et al., 2022), and (2) FunQL contains a number of brackets that provide information of

	SQL	Prolog	Lambda	FunQL
English	76.50	81.59	76.50	**89.89**
German	68.23	64.26	**72.20**	71.83
Thai	68.59	63.90	70.04	**76.17**
Chinese	70.04	63.18	74.37	**77.62**
Farsi	64.98	61.73	64.62	**75.45**
Greek	71.84	75.81	78.70	**85.92**
Indonesian	75.09	75.09	78.34	**87.00**
Swedish	75.45	77.26	79.78	**84.48**
Average	71.34	70.35	74.32	**81.04**

Table 4: Monolingual performance of mT5 on MGeoQuery. FunQL/SQL obtains the best/worst performance.

structure to the models (Shu et al., 2021).

5 Related Work

Cross-lingual Semantic Parsing Most semantic parsing datasets are originally in English such as GeoQuery (Zelle and Mooney, 1996), ATIS (Finegan-Dollak et al., 2018), Overnight (Wang et al., 2015), and Spider (Yu et al., 2018). Cross-lingual Semantic Parsing datasets are usually constructed by translating the English user questions into other languages (Dou et al., 2022; Athiwaratkun et al., 2022). For example, Lu and Ng (2011) translate GeoQuery English queries to create a Chinese version. Min et al. (2019) and Nguyen et al. (2020) create Chinese and the Vietnamese translation of Spider. However, existing CLSP datasets follow different formats and are independently studied as separate efforts. We aim to provide a unified benchmark and modeling framework to facilitate systematic evaluation and generalizable methodology.

Multilingual Language Models There has been significant progress in multilingual language models. MUSE (Conneau et al., 2017) aligns monolingual word embeddings in an unsupervised way without using any parallel corpora. XLM (Lample and Conneau, 2019) is a pretrained language model based on RoBERTa (Liu et al., 2019) which offers cross-lingual contextualized word representations. Similarly, mBERT is developed as the multilingual version of BERT Devlin et al. (2018). XLM-R (Conneau et al., 2019) outperforms mBERT and XLM in sequence labeling, classification, and question answering. Focusing on sequence-to-sequence tasks such as machine translation, mBART (Liu et al., 2020) extends BART by introducing multilingual denoising pretraining. mT5 (Xue et al., 2020) extends T5 by pretraining on the multilingual dataset mC4. Multilingual large language models have been proposed such as BLOOM (Scao et al., 2022) and XGLM (Lin et al., 2022). From multilingual embeddings to multilingual large language models, there have been more effective representations as well as more languages covered (Srivastava et al., 2022). We aim to systematically evaluate these models on CLSP, which is understudied by existing work.

Cross-lingual NLP Benchmarks Cross-lingual benchmarks have been established in many NLP tasks. XNLI is a large-scale corpus aimed to provide a standardized evaluation set (Conneau et al., 2018). Hu et al. (2020) developed XTREME to evaluate how well multilingual representations in 40 languages can generalize. XGLUE is another dataset used to implement evaluation in various cross-lingual tasks (Liang et al., 2020). MLQA (Lewis et al., 2019), XQuAD (Artetxe et al., 2019), and XOR QA (Asai et al., 2020) are three evaluation frameworks for cross-lingual question answering. Sun and Duh (2020) introduce CLIRMatrix by collecting multilingual datasets from Wikipedia for cross-lingual information retrieval (Zbib et al., 2019; Oard et al., 2019; Zhang et al., 2019; Shi et al., 2021; Chen et al., 2021b). For cross-lingual summarization, NLCS was built by Zhu et al. (2019) to tackle the problem of the divided summarization and translation. Nonetheless, there is no unified benchmark for CLSP, and thus we are unable to calibrate the performance of multilingual language models on CLSP.

6 Conclusion

We build XSemPLR, a unified benchmark for cross-lingual semantic parsing with multiple natural languages and meaning representations. We conduct a comprehensive benchmark study on three representative types of multilingual language models. Our results show that mT5 with monolingual training yields the best performance, while notably multilingual LLMs are still inadequate to perform cross-lingual semantic parsing tasks. Moreover, the performance gap between monolingual training and cross-lingual transfer learning is still significant. These findings call for both improved semantic parsing capabilities of multilingual LLMs and stronger cross-lingual transfer learning techniques for semantic parsing.

Limitations

While we cover a wide range of different factors of cross-lingual semantic parsing (e.g., tasks, datasets, natural languages, meaning representations, domains), we cannot include all possible dimensions along with these aspects. Furthermore, we focus on the linguistic generalization ability for semantic parsing because the questions are translated from the English datasets. In the future, we will explore questions raised by native speakers in each language to study the model ability under variations in cultural backgrounds and information-seeking needs.

Acknowledgment

We thank Victoria Lin, Bailin Wang, Robin Jia, Ice Pasupat, Tianze Shi, Bing Xiang, Luke Zettlemoyer for their early feedback and discussions. We thank Peng Shi, Yucheng Nie, Junru Liu, Tom Sherborne, Harsh Maniar, Xiangyu Dong, Chen Wang, Songlin Hou, Haoran Zhang, Nan Zhang, and Sarkar Das for their valuable help and comments.

References

Mikel Artetxe, Sebastian Ruder, and Dani Yogatama. 2019. On the cross-lingual transferability of monolingual representations. *arXiv preprint arXiv:1910.11856*.

Akari Asai, Jungo Kasai, Jonathan H Clark, Kenton Lee, Eunsol Choi, and Hannaneh Hajishirzi. 2020. Xor qa: Cross-lingual open-retrieval question answering. *arXiv preprint arXiv:2010.11856*.

Ben Athiwaratkun, Sanjay Krishna Gouda, Zijian Wang, Xiaopeng Li, Yuchen Tian, Ming Tan, Wasi Uddin Ahmad, Shiqi Wang, Qing Sun, Mingyue Shang, et al. 2022. Multi-lingual evaluation of code generation models. *arXiv preprint arXiv:2210.14868*.

Jonathan Berant, Andrew Chou, Roy Frostig, and Percy Liang. 2013. Semantic parsing on freebase from question-answer pairs. In *Proceedings of the 2013 conference on empirical methods in natural language processing*, pages 1533–1544.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. 2020. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901.

Mark Chen, Jerry Tworek, Heewoo Jun, Qiming Yuan, Henrique Ponde de Oliveira Pinto, Jared Kaplan, Harri Edwards, Yuri Burda, Nicholas Joseph, Greg Brockman, et al. 2021a. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Yanda Chen, Chris Kedzie, Suraj Nair, Petra Galuščáková, Rui Zhang, Douglas W Oard, and Kathleen McKeown. 2021b. Cross-language sentence selection via data augmentation and rationale training. *arXiv preprint arXiv:2106.02293*.

Hugh A Chipman, Edward I George, Robert E McCulloch, and Thomas S Shively. 2022. mbart: Multidimensional monotone bart. *Bayesian Analysis*, 17(2):515–544.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *arXiv preprint arXiv:1911.02116*.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *arXiv preprint arXiv:1710.04087*.

Alexis Conneau, Guillaume Lample, Ruty Rinott, Adina Williams, Samuel R Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. Xnli: Evaluating cross-lingual sentence representations. *arXiv preprint arXiv:1809.05053*.

Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2021. Multilingual compositional wikidata questions. *arXiv preprint arXiv:2108.03509*.

Ruixiang Cui, Rahul Aralikatte, Heather Lent, and Daniel Hershcovich. 2022. Compositional generalization in multilingual semantic parsing over wikidata. *Transactions of the Association for Computational Linguistics*, 10:937–955.

Deborah A Dahl, Madeleine Bates, Michael K Brown, William M Fisher, Kate Hunicke-Smith, David S Pallett, Christine Pao, Alexander Rudnicky, and Elizabeth Shriberg. 1994. Expanding the scope of the atis task: The atis-3 corpus. In *HUMAN LANGUAGE TECHNOLOGY: Proceedings of a Workshop held at Plainsboro, New Jersey, March 8-11, 1994*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Longxu Dou, Yan Gao, Mingyang Pan, Dingzirui Wang, Wanxiang Che, Dechen Zhan, and Jian-Guang Lou. 2022. Multispider: Towards benchmarking multilingual text-to-sql semantic parsing. *arXiv preprint arXiv:2212.13492*.

Long Duong, Hadi Afshar, Dominique Estival, Glen Pink, Philip R Cohen, and Mark Johnson. 2017. Multilingual semantic parsing and code-switching. In

Proceedings of the 21st Conference on Computational Natural Language Learning (CoNLL 2017), pages 379–389.

Catherine Finegan-Dollak, Jonathan K Kummerfeld, Li Zhang, Karthik Ramanathan, Sesh Sadasivam, Rui Zhang, and Dragomir Radev. 2018. Improving text-to-sql evaluation methodology. *arXiv preprint arXiv:1806.09029*.

Jiaqi Guo, Qian Liu, Jian-Guang Lou, Zhenwen Li, Xueqing Liu, Tao Xie, and Ting Liu. 2020. Benchmarking meaning representations in neural semantic parsing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1520–1540.

Sonal Gupta, Rushin Shah, Mrinal Mohit, Anuj Kumar, and Mike Lewis. 2018. Semantic parsing for task oriented dialog using hierarchical representations. *arXiv preprint arXiv:1810.07942*.

Carolin Haas and Stefan Riezler. 2016. A corpus and semantic parser for multilingual natural language querying of openstreetmap. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 740–750.

Simeng Han, Hailey Schoelkopf, Yilun Zhao, Zhenting Qi, Martin Riddell, Luke Benson, Lucy Sun, Ekaterina Zubova, Yujie Qiao, Matthew Burtell, et al. 2022. Folio: Natural language reasoning with first-order logic. *arXiv preprint arXiv:2209.00840*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *International Conference on Machine Learning*, pages 4411–4421. PMLR.

Srinivasan Iyer, Ioannis Konstas, Alvin Cheung, Jayant Krishnamurthy, and Luke Zettlemoyer. 2017. Learning a neural semantic parser from user feedback. *arXiv preprint arXiv:1704.08760*.

Bevan Jones, Mark Johnson, and Sharon Goldwater. 2012. Semantic parsing with bayesian tree transducers. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 488–496.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, Dmitry Tsarkov, Xiao Wang, Marc van Zee, and Olivier Bousquet. 2020. Measuring compositional generalization: A comprehensive method on realistic data. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

Daniel Keysers, Nathanael Schärli, Nathan Scales, Hylke Buisman, Daniel Furrer, Sergii Kashubin, Nikola Momchev, Danila Sinopalnikov, Lukasz Stafiniak, Tibor Tihon, et al. 2019. Measuring compositional generalization: A comprehensive method on realistic data. *arXiv preprint arXiv:1912.09713*.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *Advances in Neural Information Processing Systems (NeurIPS)*.

Hugo Laurençon, Lucile Saulnier, Thomas Wang, Christopher Akiki, Albert Villanova del Moral, Teven Le Scao, Leandro Von Werra, Chenghao Mou, Eduardo González Ponferrada, Huu Nguyen, et al. 2022. The bigscience roots corpus: A 1.6 tb composite multilingual dataset. In *Thirty-sixth Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Carolin Lawrence and Stefan Riezler. 2016. Nlmaps: A natural language interface to query openstreetmap. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: System Demonstrations*, pages 6–10.

Patrick Lewis, Barlas Oğuz, Ruty Rinott, Sebastian Riedel, and Holger Schwenk. 2019. Mlqa: Evaluating cross-lingual extractive question answering. *arXiv preprint arXiv:1910.07475*.

Haoran Li, Abhinav Arora, Shuohui Chen, Anchit Gupta, Sonal Gupta, and Yashar Mehdad. 2020. Mtop: A comprehensive multilingual task-oriented semantic parsing benchmark. *arXiv preprint arXiv:2008.09335*.

Jinyang Li, Binyuan Hui, Reynold Cheng, Bowen Qin, Chenhao Ma, Nan Huo, Fei Huang, Wenyu Du, Luo Si, and Yongbin Li. 2023. Graphix-t5: Mixing pre-trained transformers with graph-aware layers for text-to-sql parsing. *arXiv preprint arXiv:2301.07507*.

Zhenwen Li, Jiaqi Guo, Qian Liu, Jian-Guang Lou, and Tao Xie. 2022. Exploring the secrets behind the learning difficulty of meaning representations for semantic parsing. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3616–3625, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Yaobo Liang, Nan Duan, Yeyun Gong, Ning Wu, Fenfei Guo, Weizhen Qi, Ming Gong, Linjun Shou, Daxin Jiang, Guihong Cao, et al. 2020. Xglue: A new benchmark dataset for cross-lingual pre-training, understanding and generation. *arXiv preprint arXiv:2004.01401*.

Xi Victoria Lin, Todor Mihaylov, Mikel Artetxe, Tianlu Wang, Shuohui Chen, Daniel Simig, Myle Ott, Naman Goyal, Shruti Bhosale, Jingfei Du, et al. 2022. Few-shot learning with multilingual generative language models. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 9019–9052.

Yinhan Liu, Jiatao Gu, Naman Goyal, Xian Li, Sergey Edunov, Marjan Ghazvininejad, Mike Lewis, and

Luke Zettlemoyer. 2020. Multilingual denoising pre-training for neural machine translation. *Transactions of the Association for Computational Linguistics*, 8:726–742.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Wei Lu and Hwee Tou Ng. 2011. A probabilistic forest-to-string model for language generation from typed lambda calculus expressions. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 1611–1622.

Qingkai Min, Yuefeng Shi, and Yue Zhang. 2019. A pilot study for chinese sql semantic parsing. *arXiv preprint arXiv:1909.13293*.

Mehrad Moradshahi, Giovanni Campagna, Sina J Semnani, Silei Xu, and Monica S Lam. 2020. Localizing open-ontology qa semantic parsers in a day using machine translation. *arXiv preprint arXiv:2010.05106*.

Anh Tuan Nguyen, Mai Hoang Dao, and Dat Quoc Nguyen. 2020. A pilot study of text-to-sql semantic parsing for vietnamese. *arXiv preprint arXiv:2010.01891*.

Douglas W Oard, Marine Carpuat, Petra Galuščáková, Joseph Barrow, Suraj Nair, Xing Niu, Han-Chin Shing, Weijia Xu, Elena Zotkina, Kathleen McKeown, et al. 2019. Surprise languages: rapid-response cross-language ir. In *ACM NTCIR-14 Conference*, volume 10.

Jonas Pfeiffer, Naman Goyal, Xi Lin, Xian Li, James Cross, Sebastian Riedel, and Mikel Artetxe. 2022. Lifting the curse of multilinguality by pre-training modular transformers. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 3479–3495, Seattle, United States. Association for Computational Linguistics.

Prafull Prakash, Saurabh Kumar Shashidhar, Wenlong Zhao, Subendhu Rongali, Haidar Khan, and Michael Kayser. 2020. Compressing transformer-based semantic parsing models using compositional code embeddings. *arXiv preprint arXiv:2010.05002*.

Patti Price. 1990. Evaluation of spoken language systems: The atis domain. In *Speech and Natural Language: Proceedings of a Workshop Held at Hidden Valley, Pennsylvania, June 24-27, 1990*.

Subendhu Rongali, Luca Soldaini, Emilio Monti, and Wael Hamza. 2020. Don't parse, generate! a sequence to sequence architecture for task-oriented semantic parsing. In *Proceedings of The Web Conference 2020*, pages 2962–2968.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, Alexandra Sasha Luccioni, François Yvon, Matthias Gallé, et al. 2022. Bloom: A 176b-parameter open-access multilingual language model. *arXiv preprint arXiv:2211.05100*.

Tom Sherborne and Mirella Lapata. 2021. Zero-shot cross-lingual semantic parsing. *arXiv preprint arXiv:2104.07554*.

Tom Sherborne and Mirella Lapata. 2022. Zero-shot cross-lingual semantic parsing. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 4134–4153, Dublin, Ireland. Association for Computational Linguistics.

Tom Sherborne and Mirella Lapata. 2023. Meta-learning a cross-lingual manifold for semantic parsing. *Transactions of the Association for Computational Linguistics*, 11:49–67.

Tom Sherborne, Yumo Xu, and Mirella Lapata. 2020. Bootstrapping a crosslingual semantic parser. *arXiv preprint arXiv:2004.02585*.

Peng Shi, Linfeng Song, Lifeng Jin, Haitao Mi, He Bai, Jimmy Lin, and Dong Yu. 2022a. Cross-lingual text-to-SQL semantic parsing with representation mixup. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 5296–5306, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2021. Cross-lingual training of dense retrievers for document retrieval. In *Proceedings of the 1st Workshop on Multilingual Representation Learning*, pages 251–253.

Peng Shi, Rui Zhang, He Bai, and Jimmy Lin. 2022b. Xricl: Cross-lingual retrieval-augmented in-context learning for cross-lingual text-to-sql semantic parsing. *arXiv preprint arXiv:2210.13693*.

Chang Shu, Yusen Zhang, Xiangyu Dong, Peng Shi, Tao Yu, and Rui Zhang. 2021. Logic-consistency text generation from semantic parses. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4414–4426, Online. Association for Computational Linguistics.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, et al. 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*.

Shuo Sun and Kevin Duh. 2020. Clirmatrix: A massively large collection of bilingual and multilingual datasets for cross-lingual information retrieval. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4160–4170.

Raymond Hendy Susanto and Wei Lu. 2017a. Neural architectures for multilingual semantic parsing. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 38–44.

Raymond Hendy Susanto and Wei Lu. 2017b. Semantic parsing with neural hybrid trees. In *Proceedings of the AAAI Conference on Artificial Intelligence*.

Shyam Upadhyay, Manaal Faruqui, Gokhan Tür, Hakkani-Tür Dilek, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *2018 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 6034–6038. IEEE.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *arXiv preprint arXiv:1706.03762*.

Yushi Wang, Jonathan Berant, and Percy Liang. 2015. Building a semantic parser overnight. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1332–1342.

Zhiruo Wang, Grace Cuenca, Shuyan Zhou, Frank F Xu, and Graham Neubig. 2022. Mconala: A benchmark for code generation from multiple natural languages. *arXiv preprint arXiv:2203.08388*.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Silei Xu, Giovanni Campagna, Jian Li, and Monica S Lam. 2020a. Schema2qa: High-quality and low-cost q&a agents for the structured web. In *Proceedings of the 29th ACM International Conference on Information & Knowledge Management*, pages 1685–1694.

Weijia Xu, Batool Haider, and Saab Mansour. 2020b. End-to-end slot alignment and recognition for cross-lingual nlu. *arXiv preprint arXiv:2004.14353*.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2020. mT5: A massively multilingual pre-trained text-to-text transformer.

Pengcheng Yin, Bowen Deng, Edgar Chen, Bogdan Vasilescu, and Graham Neubig. 2018. Learning to mine aligned code and natural language pairs from stack overflow. In *2018 IEEE/ACM 15th international conference on mining software repositories (MSR)*, pages 476–486. IEEE.

Tao Yu, Rui Zhang, Kai Yang, Michihiro Yasunaga, Dongxu Wang, Zifan Li, James Ma, Irene Li, Qingning Yao, Shanelle Roman, et al. 2018. Spider: A large-scale human-labeled dataset for complex and cross-domain semantic parsing and text-to-sql task. *arXiv preprint arXiv:1809.08887*.

Rabih Zbib, Lingjun Zhao, Damianos Karakos, William Hartmann, Jay DeYoung, Zhongqiang Huang, Zhuolin Jiang, Noah Rivkin, Le Zhang, Richard Schwartz, et al. 2019. Neural-network lexical translation for cross-lingual ir from text and speech. In *Proceedings of the 42nd International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 645–654.

John M Zelle and Raymond J Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the national conference on artificial intelligence*, pages 1050–1055.

Luke S Zettlemoyer and Michael Collins. 2012. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*.

Rui Zhang, Caitlin Westerfield, Sungrok Shim, Garrett Bingham, Alexander Fabbri, Neha Verma, William Hu, and Dragomir Radev. 2019. Improving low-resource cross-lingual document retrieval by reranking with deep bilingual representations. *arXiv preprint arXiv:1906.03492*.

Junnan Zhu, Qian Wang, Yining Wang, Yu Zhou, Jiajun Zhang, Shaonan Wang, and Chengqing Zong. 2019. Ncls: Neural cross-lingual summarization. *arXiv preprint arXiv:1909.00156*.

Yanyan Zou and Wei Lu. 2018. Learning cross-lingual distributed logical representations for semantic parsing. *arXiv preprint arXiv:1806.05461*.

A Data Construction Details

In this section, we introduce the details of data collection, natural languages, meaning representations, and dataset statistics.

A.1 Data Collection

Multilingual ATIS ATIS (Price, 1990; Dahl et al., 1994) contains user questions for a flight-booking task. The original user questions are in English. We add the translations in Spanish, German, French, Portuguese, Japanese, Chinese from Xu et al. (2020b). Furthermore, Upadhyay et al. (2018) provide translations in Hindi and Turkish but only for a subset of utterances. Susanto and Lu (2017a) provide translations in Indonesian and Chinese, and Sherborne et al. (2020) provide translations in Chinese and German, but neither is available through LDC. Therefore, we don't include these. For meaning representations, we focus on the task of NLI to databases and thus collect

SQL from Iyer et al. (2017); Finegan-Dollak et al. (2018), while there are other formats available such as logical forms (Zettlemoyer and Collins, 2012) and BIO tags for slot and intent (Upadhyay et al., 2018). To unify SQL formats across datasets, we rewrite the SQL queries following the format of Spider (Yu et al., 2018). We follow the question splits from Finegan-Dollak et al. (2018). Through manual inspection, we discard 52 examples which do not have aligned translations from Xu et al. (2020b). This gives 5228 examples with 4303 training, 481 dev, and 444 test.

Multilingual GeoQuery GeoQuery (Zelle and Mooney, 1996) contains user questions about US geography. The original user questions are in English. One of the earliest work on cross-lingual semantic parsing is on the Chinese version of GeoQuery created by Lu and Ng (2011). Later, Jones et al. (2012) create German, Greek, and Thai translations, and Susanto and Lu (2017b) create Indonesian, Swedish, and Farsi translations. We include all these 8 languages. Furthermore, GeoQuery has several meaning representations available. To include multiple meaning representations, we collect Prolog and Lambda Calculus from Guo et al. (2020), FunQL from Susanto and Lu (2017b), and SQL from Finegan-Dollak et al. (2018). To unify SQL formats across datasets, we rewrite the SQL queries following the format of Spider (Yu et al., 2018). We follow the question splits from Finegan-Dollak et al. (2018). Through manual inspection, we discard 3 examples that do not have corresponding FunQL representations. This gives 874 examples with 548 training, 49 dev, and 277 test.

Multilingual Spider Spider (Yu et al., 2018) is a human-annotated complex and cross-domain text-to-SQL datasets. The original Spider uses English utterances and database schemas. To include utterances in other languages, we include the Chinese version (Min et al., 2019) and the syllable-level Vietnamese version (Nguyen et al., 2020). In this way, each SQL query is paired with a database schema in English and an utterance in three languages. Because the test set is not public, we include only the training and dev set. We also exclude GeoQuery examples from its training set because we use the full version of GeoQuery separately. This creates 8095 training examples and 1034 dev examples following the original splits (Yu et al., 2018).

Multilingual NLmaps NLMaps (Lawrence and Riezler, 2016) is a Natural Language Interface to query the OpenStreetMap database about geographical facts. The original questions are in English, and later Haas and Riezler (2016) provide translations in German. The meaning representation is Functional Query Language designed for OpenStreetMap, which is similar to FunQL of GeoQuery. We follow the original split with 1500 training and 880 test examples.

Multilingual Overnight Overnight (Wang et al., 2015) is a multi-domain semantic parsing dataset with lambda DCS logical forms executable in SEMPRE (Berant et al., 2013). The questions cover 8 domains in Calendar, Blocks, Housing, Restaurants, Recipes, Publications, Social, Basketball. The original dataset is in English, and Sherborne et al. (2020) provide translation in German and Chinese. They use machine translation for the training set and human translation on the dev and test sets. We include the Baidu Translation for Chinese and Google Translate for German. We merge all the domains together as a single dataset and follow the original split with 8754 training, 2188 dev, and 2740 test examples.

MCWQ MCWQ (Cui et al., 2021) is a multilingual knowledge-based question answering dataset grounded in Wikidata. This is created by adapting the CFQ (Compositional Freebase Questions) dataset (Keysers et al., 2019) by translating the queries into SQARQL for Wikidata. The questions are in four languages including Hebrew, Kannada, Chinese, and English. The split follows maximum compound divergence (MCD) so that the test set contains novel compounds to test compositionality generalization ability. We follow the MCD3 splits with 4006 training, 733 dev, and 648 test examples.

Multilingual Schema2QA Schema2QA (Xu et al., 2020a) is an open-ontology question answering dataset over scraped Schema.org web data with meaning representations in ThingTalk Query Language. Moradshahi et al. (2020) extend the original dataset with utterances in English, Arabic, German, Spanish, Farsi, Finnish, Italian, Japanese, Polish, Turkish, Chinese. The questions cover 2 domains in hotels and restaurants. The training examples are automatically generated based on template-based synthesis, crowdsourced paraphrasing, and machine translation. The test examples are crowdsourced and manually annotated by an expert with human translations. We include training examples with all 11 languages available and pair the translations with the query in corresponding language. To make the dataset size comparable to others, we include 5% of the training set. This gives 8932 training examples and 971 test examples. We also include a no-value version of the query, because the entities in the translated utterances are localized in the new languages and thus do not align well with the values in English queries.

MTOP MTOP (Li et al., 2020) is a multilingual task-oriented semantic parsing dataset with meaning representations based on hierarchical intent and slot annotations (Gupta et al., 2018). It covers 11 domains in Alarm, Calling, Event, Messaging, Music, News, People, Recipes, Reminder, Timer, Weather. It includes 6 languages in English, German, French, Spanish, Hindi, Thai. We include examples with all 6 languages available and pair the translations with the compositional decoupled representation in corresponding language. This gives 5446 training, 863 dev, 1245 test examples.

MCoNaLa MCoNaLa (Wang et al., 2022) is a code generation benchmark which requires to generate Python code. It collects English examples from the English Code/Natural Language Challenge (CoNaLa (Yin et al., 2018)) dataset and further annotates a total of 896 NL-code pairs in three languages, including Spanish, Japanese, and Russian. The training and dev set contains 1903 and 476 English examples, separately.

A.2 Language Details

We assemble 9 datasets in various domains for 5 semantic parsing tasks. It covers 8 meaning representations: SQL, Lambda Calculus, Functional Query Language, Prolog, SPARQL, ThingTalk Query Language, Python, Hierarchical Intent and Slot. The questions covers 22 languages in 15 language families: Arabic(Afro-Asiatic), Chinese(Sino-Tibetan), English(IE: Germanic), Farsi(IE: Iranian), Finnish(Uralic), French(IE: Romance), German(IE: Germanic), Greek(IE: Greek), Hebrew(Afro-Asiatic), Hindi(IE: Indo-Aryan), Indonesian(Austronesian), Italian(IE: Romance), Japanese(Japonic), Kannada(Dravidian), Polish(IE: Slavic), Portuguese(IE: Romance), Russian(IE: Slavic), Spanish(IE: Romance), Swedish(IE: Germanic), Thai(Kra-Dai), Turkish(Turkic), Vietnamese(Austro-Asiatic). Each dataset has English for cross-lingual transfer over other languages.

A.3 Meaning Representation Details

Prolog uses first-order logic augmented with higher-order predicates for quantification and aggregation. Lambda Calculus is a formal system for computation, and it represents all first-order logic and naturally supports higher-order functions with constants, quantifiers, logical connectors, and lambda abstract. FunQL is a variable-free language, and it encodes compositionality using nested function-argument structures. SQL is the query language based upon relational algebra to handle relations among entities and variables in databases. The last two, ThingTalk Query Language (Xu et al., 2020a) and Hierarchical intent and slot (Gupta et al., 2018) are recently proposed for Question Answering on Web and Task-Oriented Dialogue State Tracking, respectively. Python is a high-level, general-purpose programming language. Its design philosophy emphasizes code readability with the use of significant indentation.

A.4 Dataset Statistics

Figure 6 shows the statistics of the dataset. As can be seen, the top 3 NLs with the most samples in XSEMPLR are English, Chinese, and German, while the top 3 MRs are Lambda, SQL, and ThingTalk.

B Experiment Details

We introduce the training settings and input/output format for all experiments and settings in this section.

B.1 Training Settings

For experiments on LSTM model (Table 7), we use OpenNMT[5] as the implementation. For Transformer-PTR models, we use Pytorch[6] as the implementation. For Codex and BLOOM models, we use OpenAI API[7] and Huggingface API[8], respectively, and for mT5 and mBART models, we leverage Huggingface[9] as implementation. For each model, we train 300 epochs on MGeoquery due to the less number of training instances in this

[5] https://opennmt.net/
[6] https://pytorch.org/
[7] https://platform.openai.com/docs/api-reference
[8] https://huggingface.co/inference-api
[9] https://huggingface.co/

Dataset	Utterance	Meaning Representation (MR)
ATIS	Liste todos os voos que chegam ao General Mitchell International de várias cidades	SELECT DISTINCT T3.FLIGHT_ID FROM CITY AS T1 JOIN AIRPORT_SERVICE AS T2 ON T1.CITY_CODE = T2.CITY_CODE JOIN FLIGHT AS T3 ON T3.FROM_AIRPORT = T2.AIRPORT_CODE JOIN AIRPORT AS T4 ON T3.TO_AIRPORT = T4.AIRPORT_CODE WHERE T4.AIRPORT_CODE = "MKE"
GeoQuery	بزرگترین شهر لوئیزیانا کدام است ؟	answer(A,largest(A,(city(A),loc(A,B),const(B,stateid(louisiana)))))
Spider	有多少摇滚歌手	SELECT count(*) FROM singer WHERE genre = 'Rock'
NLmaps	Wie viele japanische Restaurants gibt es in Paris?	query(area(keyval('name','Paris'), keyval('is_in:country','France')), nwr(keyval('cuisine','japanese')),qtype(count))
Overnight	what players made less than three assists over a season	(call SW.listValue (call SW.getProperty ((lambda s (call SW.filter (var s) (call SW.ensureNumericProperty (string num_assists)) (string <) (call SW.ensureNumericEntity (number 3 assist)))) (call SW.domain (string player))) (string player)))
MCWQ	האם M0 התחתן הילד של עם M1	ASK WHERE ?x0 wdt:P749 M0 . ?x0 wdt:P26 M1 . FILTER (?x0 != M1)
Schema2QA	mostrami i ristoranti con più recensioni	now => (sort param:aggregateRating.reviewCount:Number desc of (@org.schema.Restaurant.Restaurant)) [1] => notify
MTOP	पम्पकिन रन किस प्रकार का इवेंट है ?	[IN:GET_CATEGORY_EVENT [SL:TITLE_EVENT पम्पकिन रन]]
MCoNaLa	タプルdataを空白区切りで表示する	for i in data: print(' '.join(str(j) for j in i))

Table 5: Examples of each dataset in XSEMPLR including diverse languages and meaning representations. ATIS: Portuguese-SQL, Geoquery: Farsi-Prolog, Spider: Vietnamese-SQL, NLmaps: German-FunQL, Overnight: English-Lambda Calculus, MCWQ: Hebrew-SPARQL, Schema2QA: Arabic-ThingTalk Query Language, MTOP: Hindi-Hierarchical Intent and Slot, MCoNaLa: Japanese-Python.

dataset and 100 epochs on the rest of the datasets. The learning rate is chosen from {1e-5, 3e-5, 5e-5, 1e-4} according to the parameter search on the dev set.

For Codex and BLOOM, the maximum length of the generated sequence is set to 256 tokens. For Codex, the temperature is set to 0. For BLOOM, if the generated result does not contain complete MR, we append the generated results to the input and redo the generation and repeat this process over again until the generated result is complete. However, the maximum API call of one sample is set to 5 times. After 5 calls, we use the generated result as the final result. We use default settings for the rest of the parameters.

We run the model on 8 RTX A6000 GPUs, and it takes from hours to several days according to the data size. The model architecture from Huggingface is mT5-large, mBART-large, and mBERT-base. For Codex and BLOOM, we use code-davinci-002[10], and bigscience/bloom. The batch size is set to 16 for training mT5/mBART and 32 for training Transformer-PTR models.

B.2 Input/Output Format

For input of the Transformer-PTR models, we directly feed the query into the model. For MSpider, we append the table to the end of the sequence with the format "[CLS] Query [SEP] Schema name [SEP] Table 1 [SEP] Table 2 ...", each table is represented by "table name.column name". We add "table name.*" to each table to represent all columns. For instance[11]:

```
[CLS] how many singers do we have? [SEP] * [SEP] stadium.* stadium.stadium_-id stadium.location stadium.name stadium.capacity stadium.highest stadium.lowest stadium.average [SEP] singer.* singer.singer_id singer.name singer.country singer.song_name singer.song_release_year singer.age singer.is_male [SEP] concert.* concert.concert_id concert.concert_-
```

[10] code-davinci-002 has been deprecated

[11] In these examples, we use "-" to connect the words crossing lines.

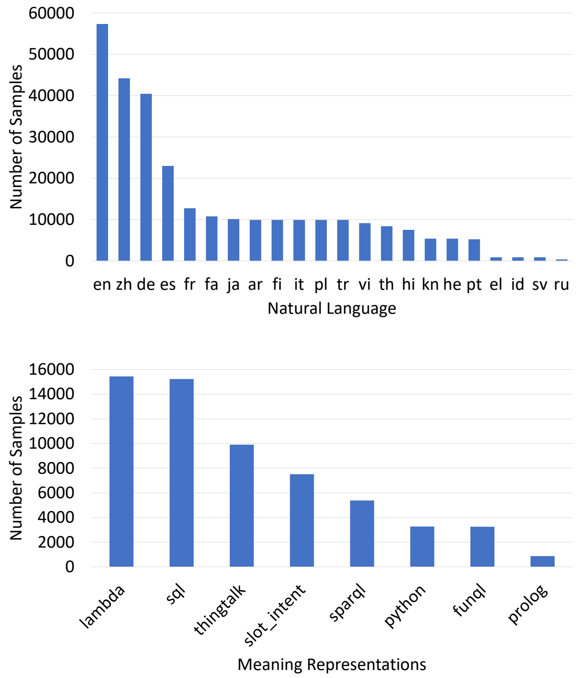

Figure 6: Distribution of 22 natural languages and 8 meaning representations. Each number of bar represents the sum of samples across all datasets.

```
name concert.theme concert.stadium_id concert.year [SEP] singer_in_concert.* singer_in_concert.concert_id singer_in_-concert.singer_id [SEP]
```

As for the output, we scan the tokens in the label and replace the ones that appear in the source text with "@ptrN" where "N" is a natural number showing the index of the token in the source text. We remove the "FROM" clause in SQL. In this way, the pointer network can easily identify which tokens are copied from the source. For instance:

```
[CLS] select count ( @ptr19 ) [SEP] concert_singer
```

For mT5 and mBART, we use the tokenizers provided by Huggingface to tokenize the queries. And for MSipder dataset, we append the table columns one by one to the tail, separated by "||". For instance:

```
how many singers do we have? || stadium.stadium_id || stadium.location || stadium.name || stadium.capacity || stadium.highest || stadium.lowest || stadium.average || singer.singer_-id || singer.name || singer.country singer.song_name || singer.song_-release_year || singer.age singer.is_-male || concert.concert_id || concert.concert_name || concert.theme || concert.stadium_id || concert.year || singer_in_concert.concert_id || singer_-in_concert.singer_id
```

The output is simply the MR itself.

```
select count ( singer_id ) from singer
```

For Codex and BLOOM, we use 8-shot in-context learning (Han et al., 2022). Specifically, we concatenate 8 pairs of examples and a query as the input. For MSpider, we additionally list the information of the schema including table names and column names of each example. It is worth noting that the number of examples of BLOOM for in-context learning decrease to 4 on MATIS dataset and decreases to 1 on MSpider dataset because the number of tokens exceeds the input limit. The example of MSpider input is listed as follows:

```
# Translate the following sentences into sql:

# Question:
# Who performed the song named "Le Pop"?

# The information of tables:
# 0. Table name is: Songs. The table columns are as follows: SongId, Title
# 1. Table name is: Albums. The table columns are as follows: AId, Title, Year, Label, Type
# 2. Table name is: Band. The table columns are

-- 3 Tables Ignored --

# 6. Table name is: Vocals. The table columns are as follows: SongId, Bandmate, Type

# Translation results are as follows:
# SELECT T2.firstname , T2.lastname FROM Performance AS T1 JOIN Band AS T2 ON T1.bandmate = T2.id JOIN Songs AS T3 ON
```

```
T3.SongId = T1.SongId WHERE T3.Title =
"Le Pop"

-- More Examples Ignored --

# Translate the following sentences
into sql:

# Question:
# Tell me the types of the policy used
by the customer named "Dayana Robel".

# The information of tables:

-- 6 Tables Ignored --

# Translation results are as follows:
```

The expected output is the MR with a starting symbol "#".

```
# SELECT DISTINCT t3.policy_type_code
FROM customers AS t1 JOIN customers_-
policies AS t2 ON t1.customer_id =
t2.customer_id JOIN available_policies
AS t3 ON t2.policy_id = t3.policy_id
WHERE t1.customer_name = "Dayana Robel"
```

B.3 Experiment Path

The experiments are done in the following order: we first evaluate 2 Enc-PTR and 2 Enc-Dec baseline models in the Monolingual setting. Then, we pick two of them with the best performance to evaluate on all the other settings. Finally, we evaluate LLMs using in-context learning in two finetuning-free settings.

C Results and Discussions

This section lists the results for each NL and MR and introduces the comparison with SOTA, training data size and few-shot learning, and error analysis.

C.1 Results for Each NL and MR

We list some of the results of our models on various datasets and languages. Table 7, 8, 9, 11, 10 show the Monolingual performance of LSTM, mBERT+PTR, XLM-R+PTR, mBART, and mT5. Table 12, 13, 14, 15 shows the Monolingual Few-Shot performance of XLM-R+PTR, mT5, Codex, and BLOOM. Table 16, and 17 show the Multilingual performance of XLM-R+PTR, and mT5. Table 18, 19, 20, 21 show the Cross-lingual Zero-Shot Transfer performance of XLM-R+PTR, mT5, Codex, and BLOOM. Table 22, 23 show the Cross-lingual Few-Shot Transfer performance of XLM-R+PTR, and mT5.

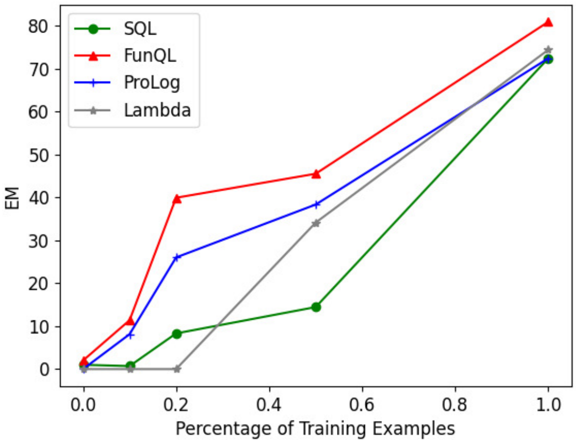

Figure 7: Exact Matching (EM) scores on the MGeoQuery dataset using mT5 as a monolingual learner.

C.2 Training Data Size and Few-shot Learning

Figure 7 displays the averaged Exact Matching scores (EM) across all languages on MGeoQuery dataset, where each line represents a meaning representation, and each dot on the line represents a few-shot experiment using such meaning representation. The X-axis is the percentage of data we use to train the model. Results show that the performance was largely influenced by the number of samples in the training set. The performance can be as high as 70% if given sufficient data, while training on 10% of training data may lead to 0 scores. Besides, among all four MRs, the performance of FunQL increases most steadily, showing its robustness.

C.3 Error Analysis

We conduct error analysis on MGeoQuery dataset. First, we select the English split with SQL MR, and compare the golden MR and the predictions generated by mT5. We classify the errors into 4 types:

- Syntax error: The prediction contains a syntax error. In other words, SQL engine can parse the predictions because of the grammar issues.

Error Type	Description	Proportion(%)
Syntax error	Incorrect program syntax (invalid grammar)	**17.14**
Semantic error		**64.27**
Token	Incorrect or missing column/value/operator	22.85
Structure	Incorrect program structure (valid grammar)	41.42
Incorrect Exact Match	Incorrect exact match with the correct execution answer	**18.47**

Table 6: Error analysis on MGeoQuery English test set. The MR is SQL.

- Token error: one of the two types of semantic errors. Predictions contain wrong column names, values (such as strings and numbers), and operators (not including keywords).
- Structure error: one of the two types of semantic errors. Predictions contain wrong structures. It means some keywords of SQL are incorrect or missing.
- Incorrect Exact Match: although the exact match shows the prediction is different from the golden one, the execution results are the same.

As shown in Table 6, most of the errors are semantic errors (64.27%) in which the structure error is around two times of token error (41.42% vs. 22.85%). Syntax error and incorrect exact match occupy around 18% of errors respectively.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ★	MSchema2QA	MTOP
English	48.9	76.8	15.8	72.2	22.4	92	48.1	78.6
Arabic	–	–	–	–	–	–	33.1	–
Chinese	44.6	61.2	10.2	–	20.8	75.1	35.9	–
Farsi	–	52.0	–	–	–	–	24.4	–
Finnish	–	–	–	–	–	–	24.7	–
French	47.5	–	–	–	–	–	–	60.8
German	47.7	59.5	–	64.9	2.1	–	38.3	58.5
Greek	–	51.4	–	–	–	–	–	–
Hebrew	–	–	–	–	–	74.0	–	–
Hindi	–	–	–	–	–	–	–	58.6
Indonesian	–	69.3	–	–	–	–	–	–
Italian	–	–	–	–	–	–	33.8	–
Japanese	2.7	–	–	–	–	–	49.6	–
Kannada	–	–	–	–	–	77.7	–	–
Polish	–	–	–	–	–	–	31.4	–
Portuguese	46.4	–	–	–	–	–	–	–
Spanish	7.2	–	–	–	–	–	44.5	63.8
Swedish	–	63.3	–	–	–	–	–	–
Thai	–	48.6	–	–	–	–	–	60.0
Turkish	–	–	–	–	–	–	41.4	–
Vietnamese	–	–	8.6	–	–	–	–	–

Table 7: The performance of LSTM with Monolingual setting on different datasets and different languages.★ We use random split in LSTM experiments rather than MCD3 split for MCWQ dataset.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
English	37.33	88.08	55.4	85.8	61.82	35.49	64.98	86.58	5.87
Arabic	–	–	–	–	–	–	48.09	–	–
Chinese	32.26	63.9	42.6	–	53.36	22.38	43.87	–	–
Farsi	–	80.86	–	–	–	–	46.65	–	–
Finnish	–	–	–	–	–	–	50.26	–	–
French	34.21	–	–	–	–	–	–	75.18	–
German	37.56	85.92	–	81.84	57.22	–	60.56	73.16	–
Greek	–	86.64	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	24.38	–	–	–
Hindi	–	–	–	–	–	–	–	70.97	–
Indonesian	–	84.84	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	50.26	–	–
Japanese	–	–	–	–	–	–	48.97	–	–
Kannada	–	–	–	–	–	11.57	–	–	–
Polish	–	–	–	–	–	–	45.31	–	–
Portuguese	36.64	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	–
Spanish	5.76	–	–	–	–	–	62.51	77.2	–
Swedish	–	87.36	–	–	–	–	–	–	–
Thai	–	81.58	–	–	–	–	–	69.36	–
Turkish	–	–	–	–	–	–	56.33	–	–
Vietnamese	–	–	23.2	–	–	–	–	–	–

Table 8: The performance of mBERT+PTR with Monolingual setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
English	36.71	88.45	53.1	86.02	62.99	37.19	73.53	90.54	7.69
Arabic	–	–	–	–	–	–	58.08	–	–
Chinese	34.91	77.98	44.1	–	56.93	19.29	48.92	–	–
Farsi	–	81.23	–	–	–	–	60.56	–	–
Finnish	–	–	–	–	–	–	64.26	–	–
French	38.31	–	–	–	–	–	–	78.58	–
German	38.28	89.17	–	84.32	59.27	–	68.59	79.22	–
Greek	–	85.92	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	14.66	–	–	–
Hindi	–	–	–	–	–	–	–	77.93	–
Indonesian	–	88.81	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	63.44	–	–
Japanese	–	–	–	–	–	–	55.26	–	–
Kannada	–	–	–	–	–	22.99	–	–	–
Polish	–	–	–	–	–	–	55.82	–	–
Portuguese	34.01	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	–
Spanish	5.63	–	–	–	–	–	68.59	81.16	–
Swedish	–	89.17	–	–	–	–	–	–	–
Thai	–	85.56	–	–	–	–	–	74.7	–
Turkish	–	–	–	–	–	–	69	–	–
Vietnamese	–	–	44.7	–	–	–	–	–	–

Table 9: The performance of XLM-R+PTR with Monolingual setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
English	45.72	74.01	52.32	86.82	65.18	38.12	57.16	87.87	6.78
Arabic	–	–	–	–	–	–	44.59	–	–
Chinese	45.72	65.34	16.48	–	56.93	25.35	42.95	–	–
Farsi	–	59.57	–	–	–	–	38.72	–	–
Finnish	–	–	–	–	–	–	54.48	–	–
French	47.97	–	–	–	–	–	–	74.05	–
German	50.23	57.76	–	79.55	56.68	–	59.11	75.18	–
Greek	–	49.46	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	33.95	–	–	–
Hindi	–	–	–	–	–	–	–	72.59	–
Indonesian	–	74.01	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	46.65	–	–
Japanese	–	–	–	–	–	–	53.76	–	–
Kannada	–	–	–	–	–	22.69	–	–	–
Polish	–	–	–	–	–	–	43.56	–	–
Portuguese	43.47	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	–
Spanish	18.47	–	–	–	–	–	57.67	78.66	–
Swedish	–	68.95	–	–	–	–	–	–	–
Thai	–	58.12	–	–	–	–	–	66.21	–
Turkish	–	–	–	–	–	–	46.65	–	–
Vietnamese	–	–	14.31	–	–	–	–	–	–

Table 10: The performance of mBART with Monolingual setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
English	53.60	89.89	68.30	92.73	69.38	39.29	76.00	91.67	10.29
Arabic	–	–	–	–	–	–	53.55	–	–
Chinese	52.48	77.62	54.90	–	62.59	24.56	56.54	–	–
Farsi	–	75.45	–	–	–	–	60.25	–	–
Finnish	–	–	–	–	–	–	68.28	–	–
French	53.60	–	–	–	–	–	–	82.30	–
German	52.93	71.83	–	90.57	66.90	–	72.19	82.38	–
Greek	–	85.92	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	33.02	–	–	–
Hindi	–	–	–	–	–	–	–	78.98	–
Indonesian	–	87.00	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	67.97	–	–
Japanese	–	–	–	–	–	–	62.41	–	–
Kannada	–	–	–	–	–	23.74	–	–	–
Polish	–	–	–	–	–	–	60.87	–	–
Portuguese	53.15	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	–
Spanish	53.13	–	–	–	–	–	68.69	83.91	–
Swedish	–	84.48	–	–	–	–	–	–	–
Thai	–	76.17	–	–	–	–	–	71.71	–
Turkish	–	–	–	–	–	–	70.03	–	–
Vietnamese	–	–	57.15	–	–	–	–	–	–

Table 11: The performance of mT5 with Monolingual setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
English	29.50	27.01	43.44	20.68	47.88	9.41	58.91	69.36	0.38
Arabic	–	–	–	–	–	–	48.71	–	–
Chinese	28.11	6.51	33.76	–	34.85	6.02	34.91	–	–
Farsi	–	6.04	–	–	–	–	37.69	–	–
Finnish	–	–	–	–	–	–	56.13	–	–
French	37.80	–	–	–	–	–	–	58.21	–
German	5.85	21.50	–	18.86	39.49	–	57.57	60.55	–
Greek	–	26.20	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	1.08	–	–	–
Hindi	–	–	–	–	–	–	–	59.66	–
Indonesian	–	25.47	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	48.09	–	–
Japanese	–	–	–	–	–	–	41.55	–	–
Kannada	–	–	–	–	–	6.02	–	–	–
Polish	–	–	–	–	–	–	40.99	–	–
Portuguese	37.33	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	–
Spanish	2.02	–	–	–	–	–	54.48	62.09	–
Swedish	–	23.40	–	–	–	–	–	–	–
Thai	–	7.14	–	–	–	–	–	52.63	–
Turkish	–	–	–	–	–	–	59.94	–	–
Vietnamese	–	–	30.92	–	–	–	–	–	–

Table 12: The performance of XLM-R+PTR with Monolingual Few-shot setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
English	31.98	33.26	40.81	36.25	59.48	10.80	39.24	72.43	1.05
Arabic	–	–	–	–	–	–	24.20	–	–
Chinese	32.88	16.25	33.46	–	48.47	4.63	19.26	–	–
Farsi	–	17.69	–	–	–	–	23.27	–	–
Finnish	–	–	–	–	–	–	35.84	–	–
French	28.60	–	–	–	–	–	–	62.81	–
German	20.27	23.82	–	26.93	52.81	–	36.05	60.91	–
Greek	–	29.88	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	6.20	–	–	–
Hindi	–	–	–	–	–	–	–	61.20	–
Indonesian	–	30.42	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	45.73	–	–
Japanese	–	–	–	–	–	–	29.66	–	–
Kannada	–	–	–	–	–	9.10	–	–	–
Polish	–	–	–	–	–	–	28.94	–	–
Portuguese	27.93	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	–
Spanish	7.43	–	–	–	–	–	47.89	59.90	–
Swedish	–	32.40	–	–	–	–	–	–	–
Thai	–	21.21	–	–	–	–	–	54.16	–
Turkish	–	–	–	–	–	–	35.84	–	–
Vietnamese	–	–	37.04	–	–	–	–	–	–

Table 13: The performance of mT5 with Monolingual Few-shot setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
English	17.79	34.39	34.43	36.82	4.34	4.48	22.97	20.21	13.87
Arabic	–	–	–	–	–	–	16.79	–	–
Chinese	16.89	31.77	27.85	–	2.74	3.86	18.85	–	–
Farsi	–	27.71	–	–	–	–	17.61	–	–
Finnish	–	–	–	–	–	–	21.52	–	–
French	18.47	–	–	–	–	–	–	17.46	–
German	18.24	31.59	–	31.70	3.21	–	20.60	18.51	–
Greek	–	33.03	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	2.47	–	–	–
Hindi	–	–	–	–	–	–	–	0.49	–
Indonesian	–	32.49	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	24.30	–	–
Japanese	–	–	–	–	–	–	19.36	–	–
Kannada	–	–	–	–	–	0.93	–	–	–
Polish	–	–	–	–	–	–	20.70	–	–
Portuguese	18.24	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	–
Spanish	18.47	–	–	–	–	–	24.30	1.13	–
Swedish	–	33.85	–	–	–	–	–	–	–
Thai	–	30.60	–	–	–	–	–	2.67	–
Turkish	–	–	–	–	–	–	30.79	–	–
Vietnamese	–	–	29.69	–	–	–	–	–	–

Table 14: The performance of Codex with Monolingual Few-shot setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
English	0.00	21.66	2.22	15.23	0.91	0.00	9.68	7.03	8.40
Arabic	–	–	–	–	–	–	5.87	–	–
Chinese	0.00	20.76	2.71	–	0.62	0.00	4.43	–	–
Farsi	–	11.64	–	–	–	–	1.96	–	–
Finnish	–	–	–	–	–	–	3.71	–	–
French	0.00	–	–	–	–	–	–	5.25	–
German	0.00	19.86	–	9.09	0.33	–	8.24	5.66	–
Greek	–	18.05	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	0.00	–	–	–
Hindi	–	–	–	–	–	–	–	5.50	–
Indonesian	–	22.48	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	5.77	–	–
Japanese	–	–	–	–	–	–	4.02	–	–
Kannada	–	–	–	–	–	0.00	–	–	–
Polish	–	–	–	–	–	–	2.99	–	–
Portuguese	0.00	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	–
Spanish	0.00	–	–	–	–	–	8.75	4.77	–
Swedish	–	19.59	–	–	–	–	–	–	–
Thai	–	8.66	–	–	–	–	–	2.75	–
Turkish	–	–	–	–	–	–	1.96	–	–
Vietnamese	–	–	1.45	–	–	–	–	–	–

Table 15: The performance of BLOOM with Monolingual Few-shot setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP
English	40.05	76.42	36.63	85.91	63.69	32.72	61.32	89.57
Arabic	–	–	–	–	–	–	65.19	–
Chinese	40.84	69.37	45.70	–	58.07	31.94	68.25	–
Farsi	–	66.85	–	–	–	–	62.62	–
Finnish	–	–	–	–	–	–	62.00	–
French	41.30	–	–	–	–	–	–	82.54
German	39.68	68.38	–	85.91	61.35	–	70.44	81.00
Greek	–	73.80	–	–	–	–	–	–
Hebrew	–	–	–	–	–	28.86	–	–
Hindi	–	–	–	–	–	–	–	78.74
Indonesian	–	75.24	–	–	–	–	–	–
Italian	–	–	–	–	–	–	57.88	–
Japanese	–	–	–	–	–	–	59.32	–
Kannada	–	–	–	–	–	29.63	–	–
Polish	–	–	–	–	–	–	64.12	–
Portuguese	42.46	–	–	–	–	–	–	–
Spanish	34.03	–	–	–	–	–	54.58	81.73
Swedish	–	74.52	–	–	–	–	–	–
Thai	–	66.22	–	–	–	–	–	76.48
Turkish	–	–	–	–	–	–	54.27	–
Vietnamese	–	–	38.28	–	–	–	–	–

Table 16: The performance of XLM-R+PTR with Multilingual setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP
English	58.45	82.04	36.07	92.27	70.33	29.94	69.52	90.61
Arabic	–	–	–	–	–	–	56.09	–
Chinese	49.83	75.99	30.66	–	63.98	28.24	58.15	–
Farsi	–	71.48	–	–	–	–	55.17	–
Finnish	–	–	–	–	–	–	62.96	–
French	55.00	–	–	–	–	–	–	78.47
German	60.12	74.19	–	90.34	68.36	–	65.27	83.46
Greek	–	79.61	–	–	–	–	–	–
Hebrew	–	–	–	–	–	28.55	–	–
Hindi	–	–	–	–	–	–	–	85.58
Indonesian	–	80.42	–	–	–	–	–	–
Italian	–	–	–	–	–	–	58.10	–
Japanese	–	–	–	–	–	–	62.55	–
Kannada	–	–	–	–	–	27.31	–	–
Polish	–	–	–	–	–	–	56.23	–
Portuguese	48.47	–	–	–	–	–	–	–
Spanish	54.85	–	–	–	–	–	63.31	84.12
Swedish	–	79.33	–	–	–	–	–	–
Thai	–	69.50	–	–	–	–	–	75.48
Turkish	–	–	–	–	–	–	62.78	–
Vietnamese	–	–	30.17	–	–	–	–	–

Table 17: The performance of mT5 with Multilingual setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ-MCD3	MSchema2QA	MTOP	MCoNaLa
Arabic	–	–	–	–	–	–	3.91	–	–
Chinese	0.92	12.83	20.30	–	23.80	2.16	0.51	–	–
Farsi	–	17.80	–	–	–	–	18.33	–	–
Finnish	–	–	–	–	–	–	26.98	–	–
French	2.15	–	–	–	–	–	–	59.90	–
German	1.61	51.13	–	60.23	49.74	–	40.37	56.27	–
Greek	–	58.44	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	5.56	–	–	–
Hindi	–	–	–	–	–	–	–	44.14	–
Indonesian	–	56.19	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	32.96	–	–
Japanese	–	–	–	–	–	–	0.31	–	0.20
Kannada	–	–	–	–	–	5.09	–	–	–
Polish	–	–	–	–	–	–	29.97	–	–
Portuguese	0.23	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	0.07
Spanish	25.35	–	–	–	–	–	39.24	62.65	0.10
Swedish	–	65.22	–	–	–	–	–	–	–
Thai	–	17.35	–	–	–	–	–	34.36	–
Turkish	–	–	–	–	–	–	9.58	–	–
Vietnamese	–	–	16.76	–	–	–	–	–	–

Table 18: The performance of XLM-R+PTR with Cross-lingual Zero-Shot Transfer setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
Arabic	–	–	–	–	–	–	38.31	–	–
Chinese	18.02	17.69	38.59	–	45.91	1.39	26.67	–	–
Farsi	–	25.27	–	–	–	–	41.40	–	–
Finnish	–	–	–	–	–	–	50.26	–	–
French	33.56	–	–	–	–	–	–	61.92	–
German	34.68	53.43	–	34.89	59.45	–	59.32	52.22	–
Greek	–	50.90	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	5.86	–	–	–
Hindi	–	–	–	–	–	–	–	35.89	–
Indonesian	–	42.24	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	58.50	–	–
Japanese	–	–	–	–	–	–	11.64	–	1.43
Kannada	–	–	–	–	–	4.94	–	–	–
Polish	–	–	–	–	–	–	49.95	–	–
Portuguese	34.46	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	0.29
Spanish	38.51	–	–	–	–	–	55.82	61.36	0.59
Swedish	–	68.23	–	–	–	–	–	–	–
Thai	–	18.05	–	–	–	–	–	39.53	–
Turkish	–	–	–	–	–	–	48.51	–	–
Vietnamese	–	–	45.26	–	–	–	–	–	–

Table 19: The performance of mT5 with Cross-lingual Zero-Shot Transfer setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
Arabic	–	–	–	–	–	–	17.82	–	–
Chinese	12.61	26.62	27.18	–	2.70	3.55	17.40	–	–
Farsi	–	25.36	–	–	–	–	16.79	–	–
Finnish	–	–	–	–	–	–	22.35	–	–
French	17.57	–	–	–	–	–	–	15.76	–
German	18.24	30.23	–	32.05	3.28	–	20.19	17.87	–
Greek	–	30.96	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	1.54	–	–	–
Hindi	–	–	–	–	–	–	–	7.92	–
Indonesian	–	31.04	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	23.48	–	–
Japanese	–	–	–	–	–	–	16.48	–	12.86
Kannada	–	–	–	–	–	1.39	–	–	–
Polish	–	–	–	–	–	–	19.26	–	–
Portuguese	17.57	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	9.57
Spanish	15.54	–	–	–	–	–	21.11	16.73	2.64
Swedish	–	31.77	–	–	–	–	–	–	–
Thai	–	23.74	–	–	–	–	–	12.13	–
Turkish	–	–	–	–	–	–	20.80	–	–
Vietnamese	–	–	27.95	–	–	–	–	–	–

Table 20: The performance of Codex with Cross-lingual Zero-Shot Transfer setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP	MCoNaLa
Arabic	–	–	–	–	–	–	5.66	–	–
Chinese	0.00	16.07	2.61	–	0.47	0.00	4.63	–	–
Farsi	–	3.34	–	–	–	–	1.54	–	–
Finnish	–	–	–	–	–	–	1.13	–	–
French	0.00	–	–	–	–	–	–	1.54	–
German	0.00	16.43	–	7.05	0.29	–	6.49	1.94	–
Greek	–	9.84	–	–	–	–	–	–	–
Hebrew	–	–	–	–	–	0.00	–	–	–
Hindi	–	–	–	–	–	–	–	1.78	–
Indonesian	–	18.50	–	–	–	–	–	–	–
Italian	–	–	–	–	–	–	5.66	–	–
Japanese	–	–	–	–	–	–	2.37	–	0.08
Kannada	–	–	–	–	–	0.00	–	–	–
Polish	–	–	–	–	–	–	3.71	–	–
Portuguese	0.00	–	–	–	–	–	–	–	–
Russian	–	–	–	–	–	–	–	–	0.09
Spanish	0.00	–	–	–	–	–	7.83	2.26	0.04
Swedish	–	14.62	–	–	–	–	–	–	–
Thai	–	0.27	–	–	–	–	–	0.81	–
Turkish	–	–	–	–	–	–	0.31	–	–
Vietnamese	–	–	0.79	–	–	–	–	–	–

Table 21: The performance of BLOOM with Cross-lingual Zero-Shot Transfer setting on different datasets and different languages.

	ATIS	GeoQuery	Spider	NLmaps	Overnight	MCWQ-MCD3	Schema2QA	MTOP
Arabic	–	–	–	–	–	–	53.66	–
Chinese	4.16	23.22	44.12	–	46.61	14.35	37.49	–
Farsi	–	29.00	–	–	–	–	46.55	–
Finnish	–	–	–	–	–	–	57.16	–
French	24.40	–	–	–	–	–	–	75.10
German	23.27	65.31	–	64.89	57.44	–	61.77	73.81
Greek	–	70.91	–	–	–	–	–	–
Hebrew	–	–	–	–	–	22.53	–	–
Hindi	–	–	–	–	–	–	–	72.35
Indonesian	–	71.90	–	–	–	–	–	–
Italian	–	–	–	–	–	–	58.29	–
Japanese	–	–	–	–	–	–	39.79	–
Kannada	–	–	–	–	–	23.61	–	–
Polish	–	–	–	–	–	–	53.45	–
Portuguese	23.27	–	–	–	–	–	–	–
Spanish	3.46	–	–	–	–	–	63.72	78.33
Swedish	–	68.38	–	–	–	–	–	–
Thai	–	28.82	–	–	–	–	–	64.35
Turkish	–	–	–	–	–	–	63.23	–
Vietnamese	–	–	43.24	–	–	–	–	–

Table 22: The performance of XLM-R+PTR with Cross-lingual Few-Shot Transfer setting on different datasets and different languages.

	MATIS	MGeoQuery	MSpider	MNLmaps	MOvernight	MCWQ	MSchema2QA	MTOP
Arabic	–	–	–	–	–	–	47.89	–
Chinese	48.65	44.32	44.39	–	60.40	29.48	53.35	–
Farsi	–	44.23	–	–	–	–	42.22	–
Finnish	–	–	–	–	–	–	61.48	–
French	50.45	–	–	–	–	–	–	62.81
German	50.32	56.95	–	71.70	64.67	–	68.80	80.68
Greek	–	60.11	–	–	–	–	–	–
Hebrew	–	–	–	–	–	26.85	–	–
Indonesian	–	58.40	–	–	–	–	–	–
Italian	–	–	–	–	–	–	66.63	–
Japanese	–	–	–	–	–	–	45.73	–
Kannada	–	–	–	–	–	18.21	–	–
Polish	–	–	–	–	–	–	57.98	–
Portuguese	49.32	–	–	–	–	–	–	–
Spanish	49.10	–	–	–	–	–	65.81	83.51
Swedish	–	64.71	–	–	–	–	–	–
Thai	–	44.49	–	–	–	–	–	71.71
Turkish	–	–	–	–	–	–	69.00	–
Vietnamese	–	–	54.45	–	–	–	–	–

Table 23: The performance of mT5 with Cross-lingual Few-Shot Transfer setting on different datasets and different languages.

ACL 2023 Responsible NLP Checklist

A For every submission:

☑ A1. Did you describe the limitations of your work?
Last section before Appendix, no number

☒ A2. Did you discuss any potential risks of your work?
In this paper, we mainly propose a benchmark and evaluate current SOTA models. Since every component is from previous work it is safe to use.

☑ A3. Do the abstract and introduction summarize the paper's main claims?
Abstract is located at the beginning of the paper.

☒ A4. Have you used AI writing assistants when working on this paper?
Left blank.

B ☑ Did you use or create scientific artifacts?

Section 2

☑ B1. Did you cite the creators of artifacts you used?
Section 2

☑ B2. Did you discuss the license or terms for use and / or distribution of any artifacts?
We maintain a list of licenses and ensure each of them can be used. We will publish the list upon acceptance.

☑ B3. Did you discuss if your use of existing artifact(s) was consistent with their intended use, provided that it was specified? For the artifacts you create, do you specify intended use and whether that is compatible with the original access conditions (in particular, derivatives of data accessed for research purposes should not be used outside of research contexts)?
Section 2

☐ B4. Did you discuss the steps taken to check whether the data that was collected / used contains any information that names or uniquely identifies individual people or offensive content, and the steps taken to protect / anonymize it?
Not applicable. Left blank.

☑ B5. Did you provide documentation of the artifacts, e.g., coverage of domains, languages, and linguistic phenomena, demographic groups represented, etc.?
The document will be published upon acceptance

☑ B6. Did you report relevant statistics like the number of examples, details of train / test / dev splits, etc. for the data that you used / created? Even for commonly-used benchmark datasets, include the number of examples in train / validation / test splits, as these provide necessary context for a reader to understand experimental results. For example, small differences in accuracy on large test sets may be significant, while on small test sets they may not be.
In Table 1, we discuss the data splits and data statistics.

C ☑ Did you run computational experiments?

Section 3

☑ C1. Did you report the number of parameters in the models used, the total computational budget (e.g., GPU hours), and computing infrastructure used?
Appendix D

The Responsible NLP Checklist used at ACL 2023 is adopted from NAACL 2022, with the addition of a question on AI writing assistance.

☑ C2. Did you discuss the experimental setup, including hyperparameter search and best-found hyperparameter values?
Appendix D

☑ C3. Did you report descriptive statistics about your results (e.g., error bars around results, summary statistics from sets of experiments), and is it transparent whether you are reporting the max, mean, etc. or just a single run?
We run all experiments once with unified settings.

☑ C4. If you used existing packages (e.g., for preprocessing, for normalization, or for evaluation), did you report the implementation, model, and parameter settings used (e.g., NLTK, Spacy, ROUGE, etc.)?
Appendix D

D ☒ Did you use human annotators (e.g., crowdworkers) or research with human participants?

Left blank.

☐ D1. Did you report the full text of instructions given to participants, including e.g., screenshots, disclaimers of any risks to participants or annotators, etc.?
No response.

☐ D2. Did you report information about how you recruited (e.g., crowdsourcing platform, students) and paid participants, and discuss if such payment is adequate given the participants' demographic (e.g., country of residence)?
No response.

☐ D3. Did you discuss whether and how consent was obtained from people whose data you're using/curating? For example, if you collected data via crowdsourcing, did your instructions to crowdworkers explain how the data would be used?
No response.

☐ D4. Was the data collection protocol approved (or determined exempt) by an ethics review board?
No response.

☐ D5. Did you report the basic demographic and geographic characteristics of the annotator population that is the source of the data?
No response.