methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Mengjie Zhao and Hinrich Schütze. 2021. Discrete and soft prompting for multilingual models. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8547–8555, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. 2021. Calibrate before use: Improving few-shot performance of language models. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 12697–12706. PMLR.

A Appendix

A.1 Tasks and Datasets

In our experiments, we consider 16 tasks spanning the following task types - classification, sequence to sequence labeling and generation. Below we review the experimental setups and datasets used for benchmarking for these two tasks. A list of all the datasets with the languages covered by them can be found in Table 3.

A.1.1 Classification

These tasks involve classifying a single sentence or a group of sentences into a finite number of discrete labels. For each dataset, we measure the performance of different models in terms of classification accuracy. For prompt-based models in particular, since we add no constraint on the output space of the LLM we compute the exact match between the generated output and a verbalized label to determine if the example was classified correctly. We run experiments for all the prompting strategies that we discussed in the previous sections for each dataset. The details of each dataset that we use for benchmarking are given below:

Dataset	Task	Languages
XNLI	Natural Language Inference	15
Indic-XNLI	Natural Language Inference	11
GLUECoS	Natural Language Inference	2
PAWS-X	Paraphrase Identification	7
XCOPA	Commonsense Reasoning	10
XStoryCloze	Commonsense Reasoning	11
TyDiQA-GoldP	Question Answering	9
MLQA	Question Answering	6
XQuAD	Question Answering	11
IndicQA	Question Answering	10
UDPOS	Part of Speech Tagging	38
PANX	NER	48
WinoMT	Gender Bias	8
GLUECoS	Sentiment Analysis	2
Jigsaw	Toxicity Classification	6
XLSum	Summarization	44

Table 3: Datasets and Language coverage of the datasets that MEGA presents evaluation for.

1. Natural Language Inference: XNLI (Conneau et al., 2018) is a dataset for cross-lingual Natural Language Inference, which consists of professional translations of the MNLI (Wang et al., 2018) corpus into 14 languages. We also consider IndicXNLI (Aggarwal et al., 2022) that translates the XNLI dataset into 11 Indic languages by using Machine Translation, followed by validation by native speakers.

2. Paraphrase Identification: PAWS-X (Yang et al., 2019b) is a paraphrase identification dataset professionally translated from the PAWS (Zhang et al., 2019) dataset into six typologically diverse languages.

3. Commonsense Reasoning: XCOPA (Ponti et al., 2020) is a commonsense reasoning dataset, which is a translation of the COPA (Roemmele et al., 2011) dataset into 11 typologically diverse languages, including very low-resource languages such as Eastern Apurímac Quechua and Haitian Creole.

XStoryCloze (Lin et al., 2022b) is created by translating the English StoryCloze (Mostafazadeh et al., 2017) dataset using professional translators into 10 typologically diverse languages.

A.1.2 Question Answering

We focus on Span Prediction type of Question Answering (QA) tasks in our experiments, where given a context and a question the task is to predict the answer within the context. One major challenge that we come across for multilingual evaluation of QA tasks is that for many languages we often cannot fit the context and question pairs for the few-shot and text examples in the maximum context size of 4096 for the DV003 model. This is mainly attributed to the poor performance of GPT's tokenizer on many non-latin script languages which results in over-tokenizing the words in these languages.

To overcome this issue we follow two steps. First, for the few-shot examples we only provide the line within the paragraph containing the answer as the context. Second, for the test example, we index the chunks of the context using the embeddings from the `text-embedding-ada-002` model. Given the question, the closest chunk in the full context is retrieved and used in the prompt for the test example. We use a maximum chunk size of 100 in our experiments and use the implementation for retrieval provided in the **LangChain**[4] library. By doing this,we minimize the space taken by the context tokens in our prompt.

Note that, for newer GPT models i.e. GPT-3.5-Turbo and GPT-4 which support longer context lengths, we do not use this retrieval strategy for QA tasks and prompt the models to obtain the answers directly. For each task, we calculate the Exact Match and F1 score as defined in Rajpurkar

[4] https://github.com/hwchase17/langchain

et al. (2016a). For our experiments we consider the following four tasks:

1. TyDiQA (Clark et al., 2020) is a QA dataset covering 11 typologically diverse languages. The task consists of two sub-tasks - passage selection and minimum answer span (Gold-P). For our experiments, we consider the Gold-P task and evaluate Monolingual and Zero-Shot Cross-Lingual prompting strategies. Since the labels do not directly transfer one-to-one across translation for QA tasks as they do for classification and require the use of alignment algorithms, we skip translate-test prompting for this task.

2. MLQA (Lewis et al., 2020) is an extractive QA dataset translated into 7 languages by professional translators. The task has two variants, the first where the question, context, and answer are all in the same language; and the second, where the question is in a different language than the context and answer. We consider the former variant of the task in our experiments. For MLQA, translate-test splits are also available, where each language's test data has been translated into English with answers aligned using the attention scores. There is no training data available for MLQA, and we use SQuAD'sRajpurkar et al. (2016a) training data for selecting few-shot examples in English and validation data for MLQA in other languages to get their few-shot examples. This way, we are able to evaluate for all three prompting setups.

3. XQuAD (Artetxe et al., 2020) consists of professional translations of a subset of the SQuaD dataset (Rajpurkar et al., 2016b) into 10 languages. XQuAD only has validation datasets available publicly, hence we evaluate the models on them. Like MLQA we use English SQuAD data for few-shot examples and since we cannot use validation data in other languages for few-shot, we only evaluate for zero-shot cross-lingual setup for this task.

4. IndicQA (Doddapaneni et al., 2022) is a manually curated cloze-style reading comprehension dataset that can be used for evaluating question-answering models in 11 Indic languages. The context paragraphs are chosen from Wikipedia articles whose topics are closely related to Indic culture, history,etc. The publicly available test set has about 2000 sentences that we carry out our evaluation on.

A.2 Sequences Labeling

In the sequence labeling task, a sequence of tokens (such as words) to be labeled are provided to the system.

A.2.1 Part of Speech Tagging

UDPOS (Zeman et al., 2020) is a dataset for Part of Speech Tagging taken from the Universal Dependencies 2.5 from the XTREME (Hu et al., 2020) benchmark. We benchmark a subset of the languages available in UDPOS.

A.2.2 Named Entity Recognition

PANX (Pan et al., 2017) or WikiANN is a Named Entity Recognition dataset consisting of Wikipedia sentences tagged with Person, Organization and Location.

For both tasks we use the linguistic structure prompting approach of Blevins et al. (2022) to define the prompts. The exact prompts used can be found in §A.4. Given the nature of both tasks, which would involve token alignment across the translation, we do not evaluate the translate-test prompting strategies for these setups. Also, since both tasks involve > 30 languages, to make the best use of the compute resources we only evaluate GPT-3.5-Turbo in a monolingual setup for these two tasks. Finally, we evaluate the first 1000 examples for each language for these datasets given the large number of languages. We have recomputed all baselines with this specification as well.

A.3 Generation

A.3.1 Summarization

The XLSum (Hasan et al., 2021a) dataset contains article-summary pairs across 44 typologically diverse languages, ranging from high to very low-resource.

For a similar reason as the tagging datasets, we only evaluate on first 1000 examples of the test sets in different languages and recompute the baselines on the same testset using the weights of the XL-SUM pretrained model, opensourced by the authors (Hasan et al., 2021b).

A.3.2 Code-switching datasets

All the datasets we consider so far are monolingual, however, a majority of the world's population speaks more than one language, leading to language contact phenomena such as code-switching (Doğruöz et al., 2021; Sitaram et al., 2019). We include two code-switching datasets in MEGA to benchmark the performance of generative models.

GLUECoS-NLI (Khanuja et al., 2020a) is a code-mixed NLI dataset in Hindi-English, consist-