Resources and Few-shot Learners
for In-context Learning in Slavic Languages

Michal Štefánik◇ and **Marek Kadlčík**◇ and **Piotr Gramacki**♣ and **Petr Sojka**◇

◇Faculty of Informatics,
Masaryk University, Czechia

♣Department of Artificial Intelligence,
Wrocław University of Science and Technology, Poland

Abstract

Despite the rapid recent progress in creating accurate and compact in-context learners, most recent work focuses on in-context learning (ICL) for tasks in English. However, the ability to interact with users of languages outside English presents a great potential for broadening the applicability of language technologies to non-English speakers.

In this work, we collect the infrastructure necessary for training and evaluation of ICL in a selection of Slavic languages[1]: Czech, Polish, and Russian. We link a diverse set of datasets and cast these into a unified instructional format through a set of transformations and newly-crafted templates written purely in target languages. Using the newly-curated dataset, we evaluate a set of the most recent in-context learners and compare their results to the supervised baselines. Finally, we train, evaluate and publish a set of in-context learning models that we train on the collected resources and compare their performance to previous work.

We find that ICL models tuned on English are also able to learn some tasks from non-English contexts, but multilingual instruction fine-tuning consistently improves the ICL ability. We also find that the massive multitask training can be outperformed by single-task training in the target language, uncovering the potential for specializing in-context learners to the language(s) of their application.

1 Introduction

The emergent ability of very large language models to understand unseen tasks from natural input text (Brown et al., 2020a), referred to as In-context Learning (ICL), recently motivated a large body of work focused specifically on creating more efficient models able to understand a new task from human

---

[1]All our templates and models are available on https://github.com/fewshot-goes-multilingual/slavic-incontext-learning
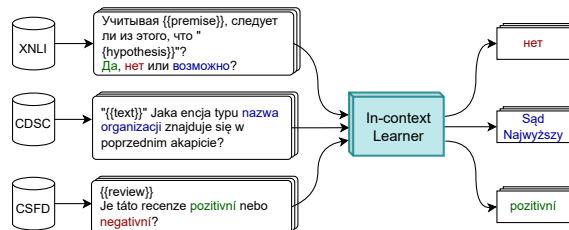

Figure 1: In this work, we transform Czech, Polish, and Russian datasets for diverse task types into a unified instructional format through a set of templates curated by the native speakers of target languages. The resulting collection enables an evaluation of existing in-context learners as well as the creation of new in-context learners interacting fully in the target language.

instructions (Min et al., 2022; Sanh et al., 2022; Wei et al., 2022; Chung et al., 2022). The ICL models presented in these works reduce the number of parameters compared to the first in-context learners by orders of magnitude. In exchange, they assume that the generalization to new tasks emerges from a vast mixture of diverse training tasks seen in the training process.

The data volume and diversity requirements might also be the factor that substantially limits the application of current ICL models mainly to English. Acquiring a large and diverse set of tasks is relatively easy for English, which is in the spotlight of the NLP community. Unfortunately, there are fewer datasets in other languages, and the collection of new ones is costly. Previous work addresses this problem by automatic translation of some English datasets (Chandra et al., 2021), or by a cross-lingual training (Mishra et al., 2022) and evaluation (Conneau et al., 2018). However, such approaches do not resemble the use of instruction models by non-English speakers, expecting the models to interact *solely* in their native language.

This work evaluates the quality of in-context learning achievable in non-English languages to this date, specifically focusing on applicability in

few-shot in-context learning for interaction in selected Slavic languages (Figure 1). Further, we assess the possibilities of further improvement under the assumption of limited data availability in the target language. We formulate these goals in two research questions:

RQ1: *How well can recent in-context few-shot learners* **perform** *in the interaction purely within our chosen, non-English languages?*

RQ2: *Can the improvements of in-context learning in a large-resource language* **transfer** *to lower resource, target languages?*

Given very limited previous work in in-context learning in our target languages, within our work, we first (i) survey and transfer a diverse set of datasets to instructional format through a set of transformations and newly-collected database of prompting templates with both the instructions and labels written in our target language(s). Our collected tasks include datasets for Named Entity Recognition, Sentiment Classification, Natural language Inference, and Question Answering in our target languages. After collecting the datasets of diverse tasks in the ICL-compatible format, we (ii) survey and evaluate in-context few-shot learners that can be applied to our target languages. Finally, we (iii) explore the possibility of further improving the in-context learners specific for our target languages along two axes: (a) by increasing models' exposure to target-language data and (b) by improving ICL ability in high-resource language, evaluating the cross-lingual transfer of such improvements.

This paper is structured as follows. Section 2 overviews the standard settings of in-context few-shot learning and surveys the previous work in this direction. Section 3 describes the evaluation datasets that we use and covers datasets' selection and unification process and templates database collection. Section 4 presents the settings used for training our in-context learners for Czech, Polish, and Russian. Finally, Section 5 presents the evaluation results, including existing and newly-trained in-context learners in the supervised baseline.

2 Background

In-context learners In-context learning from both human prompt and a set of input-output examples is initially observed as an emergent ability of GPT-3 (Brown et al., 2020b) trained on a vast collection of unlabelled texts for Causal language modeling (CLM) objective (Radford and Narasimhan, 2018). Subsequent work reproduces ICL ability and open-sources the resulting models, such as BLOOM (Scao et al., 2022) or OPT (Zhang et al., 2022). However, in-context learners trained in a solely unsupervised fashion are impractically large and hence, expensive for conventional use; In unsupervised settings, the ICL ability seems to emerge only when using far over 10 billion parameters (Brown et al., 2020b), thus requiring an extensive infrastructure to perform a single inference.

Computational overhead is addressed by a series of smaller models trained *specifically* for in-context learning. The smaller in-context learners are trained with a large mixture of tasks converted to a consistent sequence-to-sequence format via human-written *templates* (Bach et al., 2022) that define the input prompts for each task in the collection. A popular use of this framework includes prefixing the input sequence with natural-language *instructions*, such as the ones given to human annotators (Mishra et al., 2022). Large-scale instruction-based prompting in training over 1,600 tasks is also adopted in training TK-INSTRUCT (Wang et al., 2022) that we assess in our evaluations.

Recently, more attention has been dedicated to a selection of in-context training *tasks* under the assumption that some training tasks might be more beneficial for the emergence of in-context learning than others. In this direction, FLAN-T5 of Chung et al. (2022) further extends a database of tasks with the ones requiring multi-step reasoning in a *Chain-of-Thought* manner, where additionally to the correct prediction, the model is trained to predict a *sequence of steps* mapping the input to an output.

In-context Few-shot learning In-context learners are easily applicable in few-shot evaluation settings, where a small set of demonstrations for a given task exists. Given a dataset $\mathcal{D} : \{(x_1 \rightarrow Y_1), \ldots, (x_i \rightarrow Y_i)\} \in \mathcal{D}$ containing pairs of *input* x_j with associated *label* Y_j, an *in-context few-shot learner* $\Theta(x) \rightarrow y$ aims to predict a correct $y_{k+1} \equiv Y_{k+1}$ given *input text* containing a sequence of k input-output *demonstrations*, and the predicted input x_{k+1} (Štefánik and Kadlčík, 2022; Gao et al., 2022):

$$\Theta([x_1 \rightarrow Y_1, \ldots, x_k \rightarrow Y_k], x_{k+1}) \rightarrow y_{k+1} \quad (1)$$

	Name	Task	Size	Templates
	CNEC (Ševčíková et al., 2007)	NER	19k	3
	CSFD (this work)	Clf.	30k	3
cs	FBCom (Brychcín and Habernal, 2013)	Clf.	7k	3
	MALL (Brychcín and Habernal, 2013)	Clf.	30k	3
	SQAD (Medveď, 2022)	QA	8k	4
	CTKFacts (Ullrich et al., 2022)	NLI	5k	7
	PoliticAds (Augustyniak et al., 2020)	NER	1k	4
pl	KPWR (Broda et al., 2012)	NER	9k	4
	Polemo (Kocoń et al., 2019)	Clf.	8k	4
	CDSC (Wróblewska et al., 2017)	NLI	10k	4
	Polyglot (Al-Rfou et al., 2015)	NER	136k	3
ru	CEDR (Sboev et al., 2021)	Clf.	9k	3
	SberQuAD (Efimov et al., 2019)	QA	74k	4
	XNLI (Conneau et al., 2018)	NLI	399k	7

Table 1: Overview of datasets that we transform to a sequence-to-sequence format through manually-crafted templates in target languages.

Contrary to the standard supervised learning, in in-context learning, model Θ is *not* updated. Thus, it can rely solely on its ability to understand the task from input text.

Similarly to humans, the specific wording of input, i.e., *prompt* x_j, might play a large difference in the evaluation performance of the model. A prompt formulation optimal for one model type is likely not optimal for another (Lu et al., 2022). Therefore, in order to fairly compare different in-context learners, one should evaluate in-context learners on a larger set of diverse prompts (Bach et al., 2022). With this motivation, we also collect multiple prompts for each task, with a focus on their mutual diversity.

3 Datasets

The evaluation and training of new in-context learners for our target languages require (i) a collection of datasets for a representative range of tasks, and (ii) the transformation of these datasets into a unified, self-containing sequence-to-sequence form of inputs and outputs. Thus, one of our main contributions is the adaptation of the datasets for Czech, Polish, and Russian in a range of tasks: Named entity recognition, Sentiment classification, Natural language inference, and Question answering. The overview of the datasets for our target languages is shown in Table 1.

This section overviews the datasets in the target languages that we transformed, followed by a description of the process of constructing the templates for these datasets.

3.1 Data Collections in Target Languages

Contrary to English, labelled resources in our target languages for some tasks are relatively sparse, which conditions us to undertake some compromises in the diversity of the resources that we proceed with. The following text also covers the transformation that we had to perform with these datasets to cast them into a unified sequence-to-sequence format.

3.1.1 Czech Datasets

Contrary to Polish with a larger base of speakers, Czech datasets include all tasks that we aim to collect, including NER, Classification, QA, and NLI.

CNEC (Ševčíková et al., 2007) dataset for **NER** presents entities in the context of radio transcripts and news articles, featuring a relatively large collection of more than 10,000 original texts. We transform this dataset into sequence-to-sequence form by querying a specific type of entity, where we only use samples containing at most one occurrence of the requested entity to avoid ambiguity.

We note that all **classification** datasets that we find for evaluation are focused on a specific case of sentiment classification. Nevertheless, the volume, quality, and variance of sentiment classification datasets are relatively high; (i) **CSFD** presents a set of 30,000 public reviews from the movie critiques with diverse vocabulary and the challenging end task of predicting the corresponding star rating (0–5). The dataset is balanced, with each rating having a similar number of occurrences. To evaluate the models in a natural language, instead of predicting a specific numeric rating for each review, we transform the dataset labels to *positive/negative* classification, omitting samples with rating=3. (ii) **MALL** (Brychcín and Habernal, 2013) dataset is a semantically less complex collection of product reviews of online store products, and (iii) **FBCom** (Brychcín and Habernal, 2013) features a collection of scraped but verified Facebook comments presenting a sample of informal language. The latter two datasets come with three-class targets (positive/neutral/negative).

The only available Czech **QA** dataset, **SQAD** (Medveď, 2022), also builds a dataset on Wikipedia, containing the original articles in a full length, associated with manually-crafted questions and associated answer texts. To avoid the overhead of models' inference with full Wikipedia articles

in a few-shot format, we synthesize the contexts containing answers by sequencing paragraphs containing the first answer occurrence. Thus, our curated context paragraphs resemble the format of the commonly-known English SQuAD dataset (Rajpurkar et al., 2016). We note that the original version of the dataset contains a strong statistical bias, with around half of the questions having the answer at the beginning of the article. To avoid exploiting this bias in evaluation, we randomly removed 90% of the questions whose answer starts in the first 50 characters.

Finally, **CTKFacts** (Ullrich et al., 2022) introduces a collection of **NLI** examples containing premises extracted from Wikipedia, with manually-crafted hypotheses to assess given the premises, in standard NLI settings.

3.1.2 Polish

The Polish datasets for our desired tasks are smaller than Czech, and contrary to Czech, to the date of writing, we find no publicly-available Polish QA dataset. However, we find two Polish **NER** datasets. **PoliticAds** (Augustyniak et al., 2020) presents input texts in a relatively unconventional domain of political advertising. A lot of entities are largely context-dependent, thus presenting adaptation challenges for general-domain models. Therefore, we complement this quite small and specific dataset with the **KPWR** (Broda et al., 2012) dataset. However, original KPWR has a very fine granularity of entities; thus, we transform the target entities to a second-level type (i.e. mapping entity *name-location-city* simply to *location*). After disambiguation analogical to CNEC, we obtain a sequence-to-sequence dataset with 9,000 inputs.

Consistently to Czech, we enrich the set with **Polemo** dataset (Kocoń et al., 2019) for sentiment **classification**, which contains a human-annotated set of consumer reviews from the domains of *medicine*, *hotels*, *products*, and *university*. Finally, we find **CDSC** dataset for **NLI** (Wróblewska et al., 2017), featuring a collection of premise-hypothesis pairs from a wide range of 46 thematic groups.

3.1.3 Russian

Being the language with a much larger speaker base, Russian is also the richest in resources. Thus, we pick the datasets for our tasks of interest that we assess as having the highest quality. **Polyglot** (Al-Rfou et al., 2015) is a large **NER** dataset curated from references to Wikipedia sites. We transform the datasets to per-entity-type prompt format, creating multiple prompts from each sample, resulting in more than 100 k input-output entity pairs. Consistently with other languages, we further include in the collection a **CEDR** dataset for sentiment **classification** originating in social media (Sboev et al., 2021). While its domain is not representative of many use cases, we assess the quality of annotations as superior to its alternatives and the number of labels (5) as practical for few-shot evaluation with reasonably long contexts.

SberQuAD (Efimov et al., 2019) is an extractive **QA** dataset comparable with English SQuAD in both the size and domain; Its 74,000 question-context-answer tuples are manually collected with the contexts originating in Wikipedia. Contrary to SQuAD, a small portion of questions has several different answers in the context, making the correct prediction ambiguous in some cases; We omit these cases in evaluations. Finally, we choose an **XNLI** dataset (Conneau et al., 2018) for evaluating **NLI** in Russian for its heterogeneity and size. However, other quality alternatives exist (see, e.g. Shavrina et al. (2020)), and our templates can be used with any other Russian NLI dataset as well.

3.2 Templates

For each of the referenced datasets, we write a new template mapping the samples of the dataset into a sequence-to-sequence format. To reinforce templates' heterogeneity, we start by reviewing existing templates of the analogical tasks in English, collected within BigScience's P3 project (Sanh et al., 2022). From existing templates, we pick a set of mutually most-distinct templates for each task and proceed to the writing phase. The resulting number of templates for each dataset was chosen subjectively to maintain a high level of heterogeneity among the templates of each dataset.

Inspired by the existing templates, we ask our target-language volunteer native speakers to write the templates in a form that they find "the most natural to ask for the solution for a given task from a human with a native understanding of their target language". We make sure that all the templates contain the exact-matching form of the expected response (i.e., *label*) so that the domain of possible answers is clearly enclosed by the prompt. The examples of some curated templates can be found in Table 2. A full list of the collected templates can be found in Appendix A.

Lang	Task	Template
cs	NER	{{text}} Jaká entita typu {{label_type}} se nachází v předchozím odstavci?
cs	Clf.	{{comment}} Je tato recenze {{"pozitivní, neutrální nebo negativní"}}?
cs	QA	{{context}} Q: {{question}} S odkazem na sekci výše je správná odpověď na danou otázku
cs	NLI	Za předpokladu, že {{evidence}} vyplývá, že {{claim}}? Ano, ne, nebo možná?
pl	clf.	"{{text}}" Ten tekst jest pozytywny, negatywny, neutralny czy dwuznaczny?
pl	NLI	Oceń czy poniższe zdania są zgodne ze sobą - tak, nie czy nie wiadomo? Zdanie A: {{premise}} Zdanie B: {{hypothesis}} Zgodność:
pl	NER	Jaka encja typu {{label_type_selected}} znajduje się w następującym tekście? "{{text}}"
ru	NER	{{text}} Какой объект типа {{label_type}} находится в предыдущем абзаце?
ru	NLI	Примите за истину следующее: {{premise}} Тогда следующее утверждение: "{{hypothesis}}" есть "правда", "ложь" или "неубедительно"?
ru	QA	Посмотрите на абзац ниже и ответьте на следующий вопрос: Абзац: {{context}} Вопрос: {{question}}
ru	Clf.	{{text}} Каково настроение этого обзора? радость, печаль, удивление, страх или гнев?

Table 2: Examples of instruction templates for each of the language + task pair that we collect in this work. A full list of templates collected in this work by our native speakers can be found in Appendix A Table 6.

We do not identify any instructional templates for the Named Entity Recognition task in the previous work. This is likely due to the complexity of fair evaluation of prediction containing a *sequence* of prediction, necessary for collecting *all* predictions for the prompted entity type; an evaluation of sequences is difficult by using the commonly-used generative measures. After consideration, we decided to reformulate the NER tasks in the form of information extraction, where we filter out samples where prompted entity type occurs more than once. This makes the task easier, but on the other hand, the evaluation is not biased by the models' ability to order predictions correctly. Based on that, we assume that such evaluation corresponds better to in-context learners' ability to identify entities.

4 Experiments

Making in-context learning in our target languages finally possible through the transformations described in the previous section, our first objective is to assess the current state-of-the-art of the recent in-context few-shot learners when used in the interaction *exclusively* in the target language (**RQ1**). We follow by outlining the perspectives in further enhancing the quality of target-language in-context few-shot learners by assessing the potential of cross-lingual transfer (**RQ2**).

4.1 In-context Few-shot Learning Evaluation

The overview of previous work on in-context learning covered in Section 2 shows a shifting interest from the over-parametrization to the scaling of diverse training tasks (Wang et al., 2022) and more explicit reasoning schemes, such as a Chain-of-Thought (Chung et al., 2022), where in addition

to the final result, the model learns to predict the reasoning path that has led to the prediction. Our evaluation aims to assess how these aspects impact the quality of in-context few-shot learning in our target languages.

Multilingual fine-tuning To this date, we identify only one in-context learners' family that claims to support all our target languages: MTK-INSTRUCT (Wang et al., 2022). While its English counterpart (TK-INSTRUCT) fine-tunes T5 models (Raffel et al., 2020) on 1,616 tasks with English prompts, inputs, and targets, MTK-INSTRUCT is additionally fine-tuned on 576 tasks with inputs in 55 diverse languages, including Czech, Polish and Russian. Still, the instructional templates for these languages were written in English due to easier quality assurance. Thus, it remains an open question whether such-acquired in-context learning skills transfer to an interaction *solely* in the target language.

Hence, we assess the benefit of multilingual training by measuring and comparing the performance of English-only TK-INSTRUCT and multilingual TK-INSTRUCT of the same size (3 B parameters).

Fine-tuning strategy We evaluate the impact of a set of objectives of FLAN-T5 (Chung et al., 2022) complementary to a sole scaling of tasks of TK-INSTRUCT. Notably, these include (i) additional fine-tuning for a zero-shot setting, i.e. without presenting the model with demonstrations, (ii) fine-tuning for generating Chain-of-Thought, i.e. a sequence of steps leading the model to the answer, that is purposed to enhance the model's reasoning ability.

The evaluations of the impact of a fine-tuning strategy are also complemented by the assessment of our newly-trained in-context learners, trained on a single task type (QA), including the data in a target language; We detail our approach to train these models in Section 4.2.

Model size Finally, we evaluate both TK-INSTRUCT and FLAN-T5 in two different sizes: in a 700-million and in a four-times bigger, 3-billion-parameters variant. While it is perhaps not a surprising finding that the larger model would also perform better in the unseen language, the experiments in this axis assess the scale of improvement that can be expected by increasing computational costs for larger models, as compared to other adjustments.

4.2 Cross-lingual Transfer

In addition to the evaluation of existing in-context learners, we are interested in assessing how much the ICL in lower-resource languages can benefit from the improvements in a large-resource language (**RQ2**). This is particularly relevant given the fast pace of progress in general in-context learning focused primarily on English, naturally raising a question on how applicable these results are in languages for which data resources are sparser.

However, having no control over the specific data and training configuration of the existing models, we assess the scale of cross-lingual transfer by fine-tuning our own in-context learners that differ in the configuration in a large-resource language (English) while fixing the configuration in the target language. By also considering the choices of the previous work (Sanh et al., 2022), we pick the Question Answering as the one that we assume is crucial for obtaining in-context learning ability while also being available in our target languages.

Therefore, in our experiments, we *permute* only the English QA dataset and mix it in training with the QA dataset of the target language. We train in-context learners with three different configurations; (i) using *no* English QA dataset, (ii) using the standard SQuAD (Rajpurkar et al., 2016) containing more than 90,000 question-context-answer tuples, and (iii) using a lesser-known AdversarialQA (AQA) dataset (Bartolo et al., 2021) containing 30,000 more complex questions that *exploit* the flaws of QA models trained on SQuAD, making its samples complementary to SQuAD. Finally, we measure the impact of this change in Czech and

Russian, for which the target-language QA datasets are available.

All our newly-trained in-context learners (further referred to as mTK-QA$_{SQuAD}$ and mTK-QA$_{AQA}$) are based on mT5 model (Xue et al., 2021) of 1.3-billion parameter size. We make our newly-trained in-context learners for both Czech[2] and Russian[3] publicly available for any use.

5 Results

Consistently with the previous work (Sanh et al., 2022; Wang et al., 2022), we jointly report the ROUGE-L score (Lin, 2004) over all the evaluation datasets (which we transform and create templates for (§3)) and all the evaluated in-context learners (§4.1), including the newly-trained ones introduced in this work (§4.2). To ease the readability, we split the reports by language, to the results on Czech datasets in Table 3, Russian datasets in Table 5, and Polish datasets in Table 4.

As a reference of the resulting ICL performance, for each dataset, we also train a **baseline model** that is also based on mT5 model (Xue et al., 2021), fine-tuned on the training split of the dataset transformed to a sequence-to-sequence format through a mixture of *all* the templates that we curated. Details on the training and evaluation configuration that we use can be found in Appendix B.

Multilingual training helps in most cases A comparison of mTK-INSTRUCT to TK-INSTRUCT of the same size through all languages (Tables 3, 5, 4) evaluates the significance of including the training data from the target language(s). Note that mT5, a base model for mTk-instruct, was pretrained on mC4 balanced over languages, but mTk-instruct was finetuned on only 15 Polish, 5 Russian, and 2 Czech datasets making it about 1% of all data. Additionally, the training prompts for these datasets were English.

Still, we see that mTK-INSTRUCT is better than its English-finetuned counterpart in all evaluation datasets, except two Czech sentiment classification tasks. However, in some cases, the differences are relatively small; For instance, in the case of Polish CDSC, where English Tk-Instruct ends only 2.8 points behind the multilingual counterpart. The

---

[2] https://huggingface.co/ fewshot-goes-multilingual/mTk-SQuAD_ en-SQAD_cs-1B
[3] https://huggingface.co/ fewshot-goes-multilingual/ mTk-AdversarialQA_en-SberQuAD_ru-1B

Dataset + task Model	CNEC NER	CSFD Clf.	FBCom Clf.	MALL Clf.	SQAD QA	CTKFacts NLI
Supervised (mT5-1B)	67.9± 9.1	82.4±4.5	49.3±10.3	42.8±10.8	88.3±5.3	56.1±10.9
Tκ-Instruct (700M)	15.3± 6.7	14.1±7.1	25.2± 7.2	25.5± 8.4	5.6±4.8	**54.7**±8.2
Tκ-Instruct (3B)	32.8± 9.1	20.9±8.1	23.0± 7.4	25.1± 6.9	34.0±9.0	47.8±9.8
T5-FLAN (700M)	41.1±10.0	0.0± 0.0	0.0± 0.0	0.0± 0.0	46.5±8.4	30.3±9.3
T5-FLAN (3B)	49.6±10.4	0.0± 0.0	0.0± 0.0	0.1± 0.1	51.6±9.1	34.7±10.7
mTκ-Instruct (3B)	62.5± 8.9	**90.2**±4.2	10.8±6.2	9.9±7.0	67.9±8.6	44.0±10.1
mTκ-QA$_{none}$(1B)	72.0± 9.0	45.9±9.1	29.2±8.2	32.1±8.9	85.0±7.0*	35.4±10.5
mTκ-QA$_{SQUAD}$(1B)	72.3± 9.1	72.9±6.6	**32.1**±9.0	**34.7**±9.2	**87.8**±5.3*	46.9±10.1
mTκ-QA$_{AQA}$(1B)	**77.0**± 7.8	59.8±8.8	27.6±8.6	29.8±9.9	87.1±6.6*	42.7±10.7

Table 3: **In-context learners' performance in Czech:** ROUGE-L scores of selected in-context learners in Czech interaction using the listed datasets, for the best-performing template of each model. In-context learners were shown **three** demonstrations of each task. Included confidence intervals ($\alpha = 0.05$) are computed using bootstrapped evaluation (sample groups $n = 100$, repeats $r = 200$). Results marked with * denote cases where the held-out set of the listed dataset was used in training.

Dataset + task Model	PoliticAds NER	KPWR NER	Polemo Clf.	CDSC NLI
Supervised (mT5-1B)	5.9±5.1	63.3±10.3	51.9±9.9	75.5±8.5
Tκ-Instruct (700M)	5.6±4.3	8.6±5.4	28.3±8.6	52.3±8.2
Tκ-Instruct (3B)	17.6±8.1	54.6±11.2	19.5±8.4	67.8±8.8
T5-FLAN (700M)	6.8±5.5	33.8±9.8	24.3±8.6	10.0±6.4
T5-FLAN (3B)	18.4±7.3	60.5±7.8	**43.0**±9.0	**71.5**±9.0
mTκ-Instruct (3B)	**32.1**±9.6	**67.6**±8.4	25.4±8.6	70.6±8.2

Table 4: **In-context learners' performance in Polish:** ROUGE-L scores of selected in-context learners in Polish interaction using the listed datasets. Configuration of evaluation is identical to Table 3.

Dataset + task Model	Polyglot NER	CEDR Clf.	SberQAD QA	XNLI NLI
Supervised (mT5-1B)	54.3±10.8	48.6±9.6	86.4±6.5	51.5±11.5
Tκ-Instruct (700M)	0.1±0.5	12.2±6.8	0.6±1.1	12.9± 6.9
Tκ-Instruct (3B)	3.6±3.9	17.7±8.3	8.1±4.1	22.2± 8.2
T5-FLAN (700M)	1.0±1.6	15.1±6.1	11.4±4.8	13.8± 6.2
T5-FLAN (3B)	2.0±2.5	24.4±7.4	19.6±5.6	26.0± 9.0
mTκ-Instruct (3B)	57.6±11.2	**33.0**±9.9	73.7±6.7	**35.3**±10.3
mTκ-QA$_{none}$(1B)	53.3±8.4	17.9±8.1	**89.1**±5.2*	19.6± 7.5
mTκ-QA$_{SQUAD}$(1B)	50.3±9.3	7.5±4.5	84.6±6.0*	23.8± 8.8
mTκ-QA$_{AQA}$(1B)	**66.3**±10.9	27.0±9.9	86.0±5.6*	32.3± 8.3

Table 5: **In-context learners' performance in Russian:** ROUGE-L scores of selected in-context learners in Russian interaction using the listed datasets. Configuration of evaluation is identical to Table 3.

error analysis of mTk-Instruct on two flawing classification tasks (FBCom and MALL) has shown that despite purely Czech prompts, the model generates English responses. This could be explained by a semantic similarity of our tasks to some of the model's fine-tuning datasets, but in our evaluation, we consider the divergence from the prompted language of interaction a valid failure.

Inconsistent benefits of CoT training Comparing the performance of T5-FLAN models with Tk-instruct models of the corresponding size, we find that T5-FLAN is superior in 17 out of 28 cases. However, the differences are often relatively small, and the performance of both in-context learners in these cases remains below the usable level nevertheless. Therefore, while it seems that fine-tuning to a Chain-of-Thought reasoning allows the modeling of features that are applicable also in some multilingual settings, these do not generalize over all in-context learning scenarios. Notably, T5-FLAN perhaps surprisingly fails on classification in Czech,

where it shows an inability to understand the task even from the given demonstrations. On the other hand, we note that in two of four evaluation cases in Polish, the larger T5-FLAN performs superiorly to even multilingual mTk-Instruct of the same size.

Model size matters The comparisons of T5-FLAN and Tk-Instruct in their two size variants show the superiority of the larger model with the exceptions in 3 out of 28 cases, suggesting that model size can be an even more important condition of accurate in-context learning ability than utilization of target-language data in training.

It is also worth noticing that the difference in performance between two *sizes* of T5-FLAN are often very large; For instance, note the difference between Polish CDSC or Russian NLI. This suggests that the different sizes of T5-FLAN might, in fact, be very distinct in their representations.

Cross-lingual transfer A comparison of mTĸ-QA models that we train with and without the high-resource QA dataset (§4.2) outlines the potential for improvement of ICL in lower-resource languages with adjustments in the high-resource language. We see that including a complementary QA dataset in other-than-evaluated language can help in in-context learning of *all* new tasks, with improvements over 60% in Czech CSFD, or Russian XNLI.

Additionally, using a higher-quality AdversarialQA can also significantly, though not consistently, improve ICL ability for some tasks. For instance, note the difference of 12.9 points in sentiment classification of the Czech CSFD dataset or of 16 in Russian NER. This relatively large sensitivity to the data configuration in a high-resource language, from which we aim to transfer the ICL ability, suggests that recent and future improvements in models' ICL measured in English might also be directly applicable to other languages.

In-context learners trained on a single task are comparable to multi-task learners While outperforming the in-context learners trained on a much larger scale of tasks was not our initial objective, we note that at least one of our in-context learners trained using a single (QA) task *out-performs* mTk-Instruct in 6 out of 10 Czech and Russian evaluations. In *all* other cases, a QA model performs within the confidence interval of mTk-Instruct. Additionally, in 4 out of 10 cases, at least one of our QA models performs comparably or better than the supervised baseline. Hence, rather than a weak performance of mTk-Instruct, this result underlines the efficiency of Question answering as a proxy task for generalizing to the unseen tasks. We also find this result encouraging for creating in-context learners specialized to other target languages, with a perspective to outperform generic state-of-the-art learners in a similar methodology.

6 Conclusion

This paper documents our work in creating the evaluation benchmark for in-context learning for Czech, Polish, and Russian. We transform selected datasets into a compatible format, and with the aid of volunteer native speakers, we create templates for these datasets exclusively in the evaluated language. However, our templates can be applied to any other dataset of the supported types (NER, Classification, QA, and NLI).

In the interaction that is purely in the language(s) of our interest, we evaluate a set of recent in-context learners that we consider state-of-the-art in this area. We find that even in-context learners trained dominantly on English data might perform considerably well and even outperform a fully supervised baseline in some cases. However, on average, massive multilingual pre-training and instruction-based fine-tuning still largely improve the ICL ability.

Finally, we train a set of in-context learners specifically for our target languages by mixing the large QA datasets in English with smaller QA datasets in our target languages; In both Czech and Russian, such-created learners perform better or comparably to mTk-Instruct trained on a vastly larger collection of over 2,000 tasks from 55 languages. We believe that this finding will motivate future work in creating specialized but more accurate in-context learners also for other languages outside English.

We publicly release all data transformations, templates, and the newly-created in-context learners for any use.

Limitations

Templates While the templates that we curate with the help of native speakers were picked to maximize their mutual diversity, we acknowledge that the volumes of templates that we create for some datasets do not cover the full variance of possible prompts of our tasks. Therefore, our templates might not be optimal for our evaluated in-context learners.

Models In-context learners fine-tuned specifically for in-context instruction learning, including our introduced ones, are orders of magnitude smaller than the original language models acquired from sole pre-training like 175-billion-parameter GPT-3 (Brown et al., 2020b), but still remain compute-demanding for widespread deployment; We notice the inference time of a single sample for our 1B models to range between 3 and 10 seconds on a four-core CPU typical for middle-level personal computers to this date.

Analogically, also the application of our methodology (§4.2) to other languages with similar size of the base model (1.3 B) constrains the users to use dedicated GPU hardware with a minimum of

30 GB memory. We train our assessed in-context learners using Nvidia A100 GPUs with 80 GB VRAM, where the convergence of a single mT5-based model takes approximately 40 hours of computing.

References

Rami Al-Rfou, Vivek Kulkarni, Bryan Perozzi, and Steven Skiena. 2015. Polyglot-NER: Massive multilingual named entity recognition. *Proc. of the 2015 SIAM International Conference on Data Mining, Vancouver, British Columbia, Canada.*

Lukasz Augustyniak, Krzysztof Rajda, Tomasz Kajdanowicz, and Michał Bernaczyk. 2020. Political Advertising Dataset: the use case of the Polish 2020 Presidential Elections. In *Proceedings of the The Fourth Widening Natural Language Processing Workshop*, pages 110–114, Seattle, USA. ACL.

Stephen H. Bach, Victor Sanh, Zheng-Xin Yong, Albert Webson, Colin Raffel, Nihal V. Nayak, Abheesht Sharma, Taewoon Kim, M Saiful Bari, Thibault Fevry, Zaid Alyafeai, Manan Dey, Andrea Santilli, Zhiqing Sun, Srulik Ben-David, Canwen Xu, Gunjan Chhablani, Han Wang, Jason Alan Fries, Maged S. Al-shaibani, Shanya Sharma, Urmish Thakker, Khalid Almubarak, Xiangru Tang, Xiangru Tang, Mike Tian-Jian Jiang, and Alexander M. Rush. 2022. Promptsource: An integrated development environment and repository for natural language prompts.

Max Bartolo, Tristan Thrush, Robin Jia, Sebastian Riedel, Pontus Stenetorp, and Douwe Kiela. 2021. Improving question answering model robustness with synthetic adversarial data generation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 8830–8848, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Bartosz Broda, Michał Marcińczuk, Marek Maziarz, Adam Radziszewski, and Adam Wardyński. 2012. KPWr: Towards a free corpus of Polish. In *Proc. of the Eighth International Conference on Language Resources and Evaluation (LREC'12)*, pages 3218–3222, Istanbul, Turkey. European Language Resources Association (ELRA).

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020a. Language Models are Few-Shot Learners. In *Advances in NIPS*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, Sandhini Agarwal, Ariel Herbert-Voss, Gretchen Krueger, Tom Henighan, Rewon Child, Aditya Ramesh, Daniel Ziegler, Jeffrey Wu, Clemens Winter, Chris Hesse, Mark Chen, Eric Sigler, Mateusz Litwin, Scott Gray, Benjamin Chess, Jack Clark, Christopher Berner, Sam McCandlish, Alec Radford, Ilya Sutskever, and Dario Amodei. 2020b. Language models are few-shot learners. In *Advances in Neural Information Processing Systems*, volume 33, pages 1877–1901. Curran Associates, Inc.

Tomáš Brychcín and Ivan Habernal. 2013. Unsupervised Improving of Sentiment Analysis Using Global Target Context. In *Proceedings of the International Conference Recent Advances in Natural Language Processing RANLP 2013*, pages 122–128, Hissar, Bulgaria. INCOMA Ltd. Shoumen, BULGARIA.

Andreas Chandra, Affandy Fahrizain, Ibrahim, and Simon Willyanto Laufried. 2021. A Survey on non-English Question Answering Dataset.

Hyung Won Chung, Le Hou, Shayne Longpre, Barret Zoph, Yi Tay, William Fedus, Yunxuan Li, Xuezhi Wang, Mostafa Dehghani, Siddhartha Brahma, Albert Webson, Shixiang Shane Gu, Zhuyun Dai, Mirac Suzgun, Xinyun Chen, Aakanksha Chowdhery, Alex Castro-Ros, Marie Pellat, Kevin Robinson, Dasha Valter, Sharan Narang, Gaurav Mishra, Adams Yu, Vincent Zhao, Yanping Huang, Andrew Dai, Hongkun Yu, Slav Petrov, Ed H. Chi, Jeff Dean, Jacob Devlin, Adam Roberts, Denny Zhou, Quoc V. Le, and Jason Wei. 2022. Scaling Instruction-Finetuned Language Models. *arXiv e-prints*, page arXiv:2210.11416.

Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating Cross-lingual Sentence Representations. In *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2475–2485, Brussels, Belgium. ACL.

Pavel Efimov, Leonid Boytsov, and Pavel Braslavski. 2019. SberQuAD – Russian Reading Comprehension Dataset: Description and Analysis. *CoRR*, abs/1912.09723.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. 2022. PAL: Program-aided Language Models.

Jan Kocoń, Piotr Miłkowski, and Monika Zaśko-Zielińska. 2019. Multi-level sentiment analysis of PolEmo 2.0: Extended corpus of multi-domain consumer reviews. In *Proc. of the 23rd Conference on Computational Natural Language Learning (CoNLL)*, pages 980–991, Hong Kong, China. ACL.

Chin-Yew Lin. 2004. ROUGE: A Package for Automatic Evaluation of Summaries. In *Text Summarization Branches Out*, pages 74–81, Barcelona, Spain. ACL.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8086–8098, Dublin, Ireland. Association for Computational Linguistics.

Marek Medveď. 2022. SQAD 3.2. LINDAT/CLARIAH-CZ digital library at the Institute of Formal and Applied Linguistics (ÚFAL), Faculty of Mathematics and Physics, Charles University.

Sewon Min, Mike Lewis, Luke Zettlemoyer, and Hannaneh Hajishirzi. 2022. MetaICL: Learning to learn in context. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2791–2809, Seattle, United States. Association for Computational Linguistics.

Swaroop Mishra, Daniel Khashabi, Chitta Baral, and Hannaneh Hajishirzi. 2022. Cross-task generalization via natural language crowdsourcing instructions. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3470–3487, Dublin, Ireland. Association for Computational Linguistics.

Alec Radford and Karthik Narasimhan. 2018. Improving Language Understanding by Generative Pre-Training.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(146):1–67.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2383–2392, Austin, USA. ACL.

Victor Sanh, Albert Webson, Colin Raffel, Stephen Bach, Lintang Sutawika, Zaid Alyafeai, Antoine Chaffin, Arnaud Stiegler, Arun Raja, Manan Dey, M Saiful Bari, Canwen Xu, Urmish Thakker, Shanya Sharma Sharma, Eliza Szczechla, Taewoon Kim, Gunjan Chhablani, Nihal Nayak, Debajyoti Datta, Jonathan Chang, Mike Tian-Jian Jiang, Han Wang, Matteo Manica, Sheng Shen, Zheng Xin Yong, Harshit Pandey, Rachel Bawden, Thomas Wang, Trishala Neeraj, Jos Rozen, Abheesht Sharma, Andrea Santilli, Thibault Fevry, Jason Alan Fries, Ryan Teehan, Teven Le Scao, Stella Biderman, Leo Gao, Thomas Wolf, and Alexander M Rush. 2022. Multitask prompted training enables zero-shot task generalization. In *International Conference on Learning Representations*.

Alexander Sboev, Aleksandr Naumov, and Roman Rybka. 2021. Data-Driven Model for Emotion Detection in Russian Texts. *Procedia Computer Science*, 190:637–642. 2020 Annual International Conference on Brain-Inspired Cognitive Architectures for Artificial Intelligence: Eleventh Annual Meeting of the BICA Society.

Teven Le Scao, Angela Fan, Christopher Akiki, Ellie Pavlick, Suzana Ilić, Daniel Hesslow, Roman Castagné, and BigScience. Workshop:. 2022. BLOOM: A 176B-Parameter Open-Access Multilingual Language Model.

Magda Ševčíková, Zdeněk Žabokrtský, and Oldřich Krůza. 2007. Named Entities in Czech: Annotating Data and Developing NE Tagger. In *Proc. of the 10th International Conference on Text, Speech and Dialogue*, volume 4629 of *LNCS*, pages 188–195, Berlin / Heidelberg. Springer.

Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. In *Proc. of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726. ACL.

Michal Štefánik and Marek Kadlčík. 2022. What is Not in the Context? Evaluation of Few-shot Learners with Informative Demonstrations.

Michal Štefánik, Vít Novotný, Nikola Groverová, and Petr Sojka. 2022. Adaptor: Objective-Centric Adaptation Framework for Language Models. In *Proceedings of the 60th Annual Meeting of the ACL: System Demonstrations*, pages 261–269, Dublin, Ireland. ACL.

Herbert Ullrich, Jan Drchal, Martin Rýpar, Hana Vincourová, and Václav Moravec. 2022. CsFEVER and

CTKFacts: Acquiring Czech data for fact verification.

Yizhong Wang, Swaroop Mishra, Pegah Alipoormolabashi, Yeganeh Kordi, Amirreza Mirzaei, Atharva Naik, Arjun Ashok, Arut Selvan Dhanasekaran, Anjana Arunkumar, David Stap, Eshaan Pathak, Giannis Karamanolakis, Haizhi Lai, Ishan Purohit, Ishani Mondal, Jacob Anderson, Kirby Kuznia, Krima Doshi, Kuntal Kumar Pal, Maitreya Patel, Mehrad Moradshahi, Mihir Parmar, Mirali Purohit, Neeraj Varshney, Phani Rohitha Kaza, Pulkit Verma, Ravsehaj Singh Puri, Rushang Karia, Savan Doshi, Shailaja Keyur Sampat, Siddhartha Mishra, Sujan Reddy A, Sumanta Patro, Tanay Dixit, and Xudong Shen. 2022. Super-NaturalInstructions: Generalization via declarative instructions on 1600+ NLP tasks. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 5085–5109, Abu Dhabi, United Arab Emirates. Association for Computational Linguistics.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M. Dai, and Quoc V Le. 2022. Finetuned Language Models are Zero-Shot Learners. In *Proc. of International Conference on Learning Representations*.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Remi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proc. of the 2020 Conf. EMNLP: System Demonstrations*, pages 38–45. ACL.

Alina Wróblewska, Krasnowska-Kieraś, and Katarzyna. 2017. Polish evaluation dataset for compositional distributional semantics models. In *Proc. of the 55th Annual Meeting of the ACL (Volume 1: Long Papers)*, pages 784–792, Vancouver, Canada. ACL.

Linting Xue, Noah Constant, Adam Roberts, Mihir Kale, Rami Al-Rfou, Aditya Siddhant, Aditya Barua, and Colin Raffel. 2021. mT5: A massively multilingual pre-trained text-to-text transformer. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 483–498, Online. Association for Computational Linguistics.

Susan Zhang, Stephen Roller, Naman Goyal, Mikel Artetxe, Moya Chen, Shuohui Chen, Christopher Dewan, Mona Diab, Xian Li, Xi Victoria Lin, Todor Mihaylov, Myle Ott, Sam Shleifer, Kurt Shuster, Daniel Simig, Punit Singh Koura, Anjali Sridhar, Tianlu Wang, and Luke Zettlemoyer. 2022. Opt: Open pretrained transformer language models.

A   Table of templates

Table 6 contains a full list of templates collected within this work, including the segments filled from the transformed datasets.

B   Details of training and evaluation configuration

All models trained within this work, including the baselines are based on the mT5-Large model trained on the referenced dataset(s) using Batch size=30, learning rate = $2 \cdot 10^{-5}$ and early stopping with the patience of 10 evaluations (i.e. 2,000 updates) based on the evaluation loss on a held-out set of data of all training datasets. Where the validation split was provided, we use it as the held-out evaluation set, otherwise, we slice out the last 200 samples of the training data for this purpose. For a simple tracking of multi-dataset training, as well as for convenient bulk training of all supervised baselines, we used Adaptor library (Štefánik et al., 2022) in version 0.2.0, with Hugging Face Transformers library (Wolf et al., 2020), version 4.19.1 as backend. For each training, we used a single Nvidia A100 with 80 GB of GPU memory.

In all evaluations, we used greedy search generation with a default configuration of *generate* method in version 4.19.1.

Lang	Task	Template
cs	NER	{{text}} {{label_type}} v tomto textu je
cs	NER	Jaká entita typu {{label_type}} se nachází v následujícím textu? {{text}}
cs	NER	{{text}} Jaká entita typu {{label_type}} se nachází v předchozím odstavci?
cs	Clf.	Jaký sentiment vyjadřuje následující filmová recenze? {{comment}}
cs	Clf.	{{comment}} Shledal recenzent tento film {{"dobrým nebo zlým"}}?
cs	Clf.	{{comment}} Je tato recenze {{"pozitivní nebo negativní"}}?
cs	Clf.	{{comment}} Je tento komentář {{"pozitivní, neutrální nebo negativní"}}?
cs	Clf.	{{comment}} Jaký je sentiment tohoto komentáře? {{"pozitivní, neutrální nebo negativní"}}?
cs	Clf.	Jaký sentiment má následující komentář? {{comment}}
cs	Clf.	{{comment}} Je tato recenze {{"pozitivní, neutrální nebo negativní"}}?
cs	Clf.	Jaký sentiment má následující recenze? {{comment}}
cs	Clf.	{{comment}} Jaký je sentiment této recenze? {{"pozitivní, neutrální nebo negativní"}}?
cs	QA	{{context}} Q: {{question}} S odkazem na sekci výše je správná odpověď na danou otázku
cs	QA	Podívejte se na odstavec níže a odpovězte na následující otázku: Odstavec: {{context}} Otázka: {{question}}
cs	QA	{{context}} S odkazem na výše uvedený odstavec, {{question}}
cs	QA	{{context}} Otázka: {{question}} Odpověď:
cs	NLI	{{evidence}} Otázka: {{claim}} Pravda, nepravda, nebo ani jedno?
cs	NLI	{{evidence}} Za uvedeného předpokladu a na základě znalostí o světě, "{{claim}}" je určitě pravda, nepravda, nebo není jasné?
cs	NLI	{{evidence}} Na základě předchozího odstavce, je to pravda, že "{{claim}}"? Ne, možná, nebo ano?
cs	NLI	Za předpokladu, že {{evidence}} vyplývá, že {{claim}}? Ano, ne, nebo možná?
cs	NLI	Předpokládejme následovné: {{evidence}} Pak musí být pravda, že "{{claim}}"? Ano, ne, nebo možná?
cs	NLI	Předpokládáme, že {{evidence}} Je možné předpokládat, že "{{claim}}" je pravda? Ano, ne, nebo možná?
cs	NLI	Předpokládejme následovné: {{evidence}} Pak následující tvrzení: "{{claim}}" je pravda, nepravda, nebo nejasné?
pl	clf.	"{{text}}" Ten tekst jest pozytywny, negatywny, neutralny czy dwuznaczny?
pl	clf.	Oceń ten tekst jako pozytywny, negatywny, neutralny lub dwuznaczny. Tekst: {{text}}
pl	clf.	Oceń wydźwięk tego tekstu jako pozytywny, negatywny, neutralny lub dwuznaczny. Tekst: {{text}} Wydźwięk:
pl	clf.	"{{text}}" Jaka jest ta recenzja? Jest pozytywna, negatywna, neutralna czy dwuznaczna?:
pl	NLI	"{{sentence_A}}" Na podstawie tego, można powiedzieć, że zdanie "{{sentence_B}}" jest potwierdzeniem, zaprzeczeniem czy niezwiązane?
pl	NLI	Oceń czy poniższe zdania są zgodne ze sobą - tak, nie czy nie wiadomo? Zdanie A: {{sentence_A}} Zdanie B: {{sentence_B}} Zgodność:
pl	NLI	Hipotezę i przesłankę można powiązać jako potwierdzenie, zaprzeczenie lub niezwiązane. Hipoteza: {{sentence_A}} Przesłanka: {{sentence_B}} Powiązanie:
pl	NLI	Hipoteza: {{sentence_A}} Przesłanka: {{sentence_B}} Czy przesłanka jest dla hipotezy potwierdzeniem, zaprzeczeniem czy jest niezwiązana?
pl	NER	"{{text}}" {{label_type_selected}} w tym tekście to
pl	NER	Znajdź encje typu {{label_type_selected}} w następującym tekście: {{text}}
pl	NER	Jaka encja typu {{label_type_selected}} znajduje się w następującym tekście? "{{text}}"
pl	NER	"{{text}}" Jaka encja typu {{label_type_selected}} znajduje się w poprzednim akapicie?
pl	NER	"{{text}}" {{label_type_selected}} w tym tekście to
pl	NER	Znajdź encje typu {{label_type_selected}} w następującym tekście: {{text}}
pl	NER	Jaka encja typu {{label_type_selected}} znajduje się w następującym tekście? "{{text}}"
pl	NER	"{{text}}" Jaka encja typu {{label_type_selected}} znajduje się w poprzednim akapicie?
ru	NER	{{text}} {{label_type}} в этом тексте:
ru	NER	Какой объект типа {{label_type}} встречается в следующем тексте? {{text}}
ru	NER	{{text}} Какой объект типа {{label_type}} находится в предыдущем абзаце?
ru	NLI	{{premise}} Используя только приведенное выше описание и то, что вы знаете о мир, "{{hypothesis}}" определенно верна, неверна или неубедительна?
ru	NLI	{{premise}} Верно ли, исходя из предыдущего отрывка, что "{{hypothesis}}"? Да, нет, а может быть?
ru	NLI	Учитывая {{premise}}, следует ли из этого, что "{{hypothesis}}"? Да, нет или возможно?
ru	NLI	{{premise}} Имеем ли мы право говорить, что "{{hypothesis}}"? Да, нет, или может быть?
ru	NLI	Учитывая, что {{premise}} Следовательно, должно быть верно, что "{{hypothesis}}"? Да, нет, а Возможно?
ru	NLI	Учитывая {{premise}} Должны ли мы предположить, что "{{hypothesis}}" верна? Да, нет или возможно?
ru	NLI	Примите за истину следующее: {{premise}} Тогда следующее утверждение: "{{hypothesis}}" есть "правда", "ложь" или "неубедительно"?
ru	QA	{{context}} Ответ на вопрос: {{question}}
ru	QA	Посмотрите на абзац ниже и ответьте на следующий вопрос: Абзац: {{context}} Вопрос: {{question}}
ru	QA	{{context}}\n\nСо ссылкой на абзац выше, {{question}}
ru	QA	{{context}} Вопрос: {{question}} Отвечать:
ru	Clf.	{{text}} Это обзор радят, печал, удивление, страх или гнев?
ru	Clf.	Каково настроение следующего обзора? {{text}} Варианты: радость, печаль, удивление, страх, гнев
ru	Clf.	{{text}} Каково настроение этого обзора? радость, печаль, удивление, страх или гнев?

Table 6: Templates for all languages and all task types that we collect in this work. Templates were written by native speakers of the template's language.