Estonian WinoGrande Dataset: Comparative Analysis of LLM Performance on Human and Machine Translation

Marii Ojastu[1,2,*], Hele-Andra Kuulmets[1], Aleksei Dorkin[1], Marika Borovikova[2], Dage Särg[3], Kairit Sirts[1,*]

[1]TartuNLP, Institute of Computer Science
[2]Department of Translation Studies, Institute of Foreign Language and Cultures
[3]Institute of Genomics
University of Tartu, Estonia
{firstname.lastname}@ut.ee

Abstract

In this paper, we present a localized and culturally adapted Estonian translation of the test set from the widely used commonsense reasoning benchmark, WinoGrande. We detail the translation and adaptation process carried out by translation specialists and evaluate the performance of both proprietary and open source models on the human translated benchmark. Additionally, we explore the feasibility of achieving high-quality machine translation by incorporating insights from the manual translation process into the design of a detailed prompt. This prompt is specifically tailored to address both the linguistic characteristics of Estonian and the unique translation challenges posed by the WinoGrande dataset. Our findings show that model performance on the human translated Estonian dataset is slightly lower than on the original English test set, while performance on machine-translated data is notably worse. Additionally, our experiments indicate that prompt engineering offers limited improvement in translation quality or model accuracy, and highlight the importance of involving language specialists in dataset translation and adaptation to ensure reliable and interpretable evaluations of language competency and reasoning in large language models.

Keywords: WinoGrande, benchmark translation, Estonian

1. Introduction

The WinoGrande dataset (Sakaguchi et al., 2019) is a widely used benchmark for evaluating the commonsense reasoning abilities of large language models (LLMs) (Gemma Team, 2024; Jiang et al., 2024; Touvron et al., 2023; Grattafiori et al., 2024). However, this dataset is exclusively in English, limiting evaluation of commonsense reasoning only to that language. Meanwhile, large language models have recently become increasingly multilingual, creating the need for comparable benchmarks that assess these capabilities beyond English.

A common approach to creating benchmarks for other languages is to translate an existing benchmark from English to the target language. To reduce translation costs, machine translation (MT) systems are often used for this purpose (Lai et al., 2023; Foroutan et al., 2023; Thellmann et al., 2024; Singh et al., 2024; Dang et al., 2024; Raihan et al., 2025). However, evaluation results on machine-translated data tend to differ from those on human-annotated data (Kreutzer et al., 2025). This difference is not easily predictable and may arise from multiple factors such as models exploiting translation artifacts (Artetxe et al., 2020) or translation inaccuracies in test data that obscure the benchmark task (Plaza et al., 2024). Moreover, machine translation lacks mechanisms for localization or cultural adaptation, resulting in datasets that are semantically distant from natural language use by native speakers. Finally, the WinoGrande benchmark is particularly difficult to translate across languages due to the inherent nature of the task and the strict constraints each example must satisfy.

Schema	The door opened louder than the window because the _ had less grease on its hinges.
Option 1	window
Option 2	door

Table 1: An example from the WinoGrande dataset.

Considering the aforementioned aspects, we manually translated, localized and culturally adapted the WinoGrande test set, which consists of 1,767 instances, into Estonian, a mid-resource Finno-Ugric language, to support development and evaluation of language models for this language. We discuss in detail the linguistic and translation challenges encountered in the process, including the correction of ambiguous and incorrectly labelled examples that were identified during this work.

We compared this newly created manually translated dataset to two machine translated versions produced by OpenAI models. Our experiments with open and proprietary models show higher accuracy on the manually translated, localized, adapted

Proceedings of the Fifteenth Language Resources and Evaluation Conference (LREC 2026), pages 6904–6918
11–16 May 2026.

and corrected WinoGrande than on the machine-translated versions, with gains in both the localized-adapted and corrected subsets.

To further examine the differences, we manually identified machine translated examples where the meaning has shifted from the original meaning. We observe that, on average, models score lower on these examples than on human-translated equivalents, indicating potential label–sentence inconsistencies introduced by machine translation.

Among the 1,767 examples in the translated WinoGrande test set, we manually corrected 89 (5% of the entire test set), while localization and adaptation was applied to 53 examples (3%). Among the remaining semantically comparable samples, up to 15.2% lost their original semantics during machine translation. Taken together, these results demonstrate the ongoing infeasibility of using current state-of-the-art LLMs to reliably translate WinoGrande sentences into Estonian.[1]

2. WinoGrande Dataset

The Winograd Schema Challenge (WSC) (Levesque et al., 2012) is a pronoun reference disambiguation task proposed to evaluate commonsense knowledge in AI systems and has inspired a range of subsequent benchmarks (Sakaguchi et al., 2019; Rudinger et al., 2018; Zhang et al., 2020) for testing AI capabilities. The original WSC is a set of few hundred expert-crafted examples intended only to assess AI capabilities. However, it was later demonstrated to be solvable with fine-tuning deep neural networks on auxiliary datasets without particular advancements in commonsense reasoning being made (Kocijan et al., 2019; Sakaguchi et al., 2019).

To address this issue, Sakaguchi et al. (2019) introduced WinoGrande, a large-scale ambiguity resolution dataset, in total consisting of 44k crowd-sourced problems inspired by the original Winograd Schema Challenge. The dataset consists of sentences where in the first part of a sentence two nouns are mentioned and the second part contains a blank which corresponds to the mention of one of the nouns (see Table 1). Some sentences are paired (hereafter twin sentences), have a lexical overlap on 70% and share the same set of answer options (see Table 5). The objective is to decide which of the two nouns correctly fills the blank.

According to Sakaguchi et al. (2019), the sentences are designed to meet two additional criteria. First, the answer options must be unambiguous, meaning that humans would easily be able to select the correct option without considerable effort.

[1]The human translated Estonian WinoGrande dataset is available at `https://huggingface.co/datasets/tartuNLP/winogrande_et`

WSC translations	**Type**
French (Amsili and Seminck, 2017)	HT
Portuguese (Melo et al., 2019)	HT
Japanese (Shibata et al., 2015)	HT
Chinese (Bernard and Han, 2020)	HT
Hungarian (Vadász and Ligeti-Nagy, 2022)	MT+PE
Thai (Artkaew, 2025)	HT
Hebrew (Shwartz, 2024)	N/A
WG translations	**Type**
Icelandic (Snæbjarnarson et al., 2022)	HT
8 African languages (Alhanai et al., 2025)	HT
Arabic + dialects (Mousi et al., 2025)	MT+PE
Italian (Moroni et al., 2024)	MT
Kyrgys (Turatali et al., 2025)	MT+PE
Nepali (Nyachhyon et al., 2025)	MT+PE
Egyptian (Shang et al., 2025)	MT
Romanian (Masala et al., 2024)	MT
Lithuanian (Nakvosas et al., 2025)	MT
Korean (Kim et al., 2025)	HT
Basque* (Corral et al., 2024)	HT
Persian (Farsi et al., 2025)	MT+PE

Table 2: Overview of Winograd Schema Challenge (WSC) and WinoGrande (WG) dataset translations across languages. HT indicates human translation, MT indicates machine translation, and PE indicates post-editing. Asterisk (*) marks a translation of a subset of samples.

The authors report that human performance on the test set is 94%. Secondly, the correct answer option should not be derivable solely from the local context or simple word associations. To enforce the second constraint, the authors applied a bias-reduction algorithm that extends the idea of filtering out human-detectable lexical biases to the embedding space. The algorithm identifies and removes examples that a model could solve by exploiting statistical regularities in word embeddings rather than true reasoning. Despite extensive manual and automatic filtering, the dataset has been reported to have several issues—such as instances solvable through simple word correlations, poorly written examples, sentences revealing the correct answer, or examples that are genuinely difficult to understand (Kocijan et al., 2023)—and is not completely free of artifacts (Elazar et al., 2021).

Although the WinoGrande dataset has turned out to be less challenging for large language models than anticipated (Lourie et al., 2021), it measures the progress only for English, leaving it unclear how well the success of solving ambiguity resolution tasks transfers to other languages. To answer this question, both the Winograd Schema Challenge and WinoGrande have been translated into several other languages (Table 2). Additional translations exist (Shavrina et al., 2020; Aparovich et al., 2025; Žagar and Robnik-Šikonja, 2022), for a re-

Ex.	English Schema	Estonian Schema	Comment
(A)	I much prefer the necklace that I have over the bracelet of my friend because the _ is cheap. **Option 1**: necklace **Option 2**: bracelet	Ma eelistan oma sõbra käevõru asemel pigem oma kaelakeed, sest see _ on maitsetu. **Option 1**: kaelakee **Option 2**: käevõru	This example shows the need for deliberate choices to avoid ambiguity. In English, *cheap* implies both low price and low quality, while its Estonian equivalent *odav* refers mainly to low price. In translation, *cheap* is rendered as *maitsetu* ("tacky") rather than literally, preserving the schema's intended clarity and disambiguation.
(B)	The girl ate less beans than meat on the date because the _ made her full. **Option 1**: beans **Option 2**: meat	**Human translated:** Tüdruk sõi kohtingul ube vähem kui liha, sest tal sai _ kõht täis. **Option 1**: ubadest **Option 2**: lihast **Machine translated:** Tüdruk sõi kohtingul vähem ube kui liha, sest _ tegi ta kõhu täis. **Option 1**: oad **Option 2**: liha	The machine translation introduces a number mismatch: in English, *beans* (plural) and *meat* (singular) pose no issue, but in Estonian, the verb *tegi* ("made") agrees only with the singular. This allows the model to rely on grammar rather than reasoning, since the correct answer is also the only grammatically fitting one. The human translation fixes this by rephrasing the schema as *became full from the _* instead of the literal _ *made her full*, removing the grammatical cue.

Table 3: Examples of translated WinoGrande schemas.

cast version of the Winograd Schema Challenge dataset that is part of the SuperGLUE benchmark (Wang et al., 2020). However, as Amsili and Seminck (2017) noted, translating these tasks is not a straightforward process and requires solving several linguistic challenges such as the disagreement of translated nouns in number or gender or general ambiguity of the translated schema.

Another issue with translating these datasets to other languages is the culture- and region-specific knowledge that some of these examples assume and which might be in the role of commonsense knowledge that is needed to resolve ambiguity. However, for the speakers of non-English languages, this type of knowledge can not be expected to be part of commonsense knowledge meaning that such examples must be specifically adapted for the target language. For instance, Snæbjarnarson et al. (2022) reports having to do cultural adaptation of some of the examples. Alhanai et al. (2025) assessed cultural appropriateness of WinoGrande translations into 11 African languages and found that 20.6% examples can be considered culturally inappropriate.

3. Human Translated Dataset

This section presents the steps involved in creating a human-translated Estonian version of the WinoGrande dataset. It first describes the translation process, followed by the description of localization, cultural adaptation, and error correction. Finally, to ensure translation quality, we report inter-annotator agreement on this newly created dataset.

3.1. Translation Process

The translation of the WinoGrande test set into Estonian was carried out by a master's student in translation studies as part of their thesis work. The translations were revised in collaboration with a professional translator, who is a master's level expert in translation studies. Both translators are co-authors of this paper.

The translation aimed to preserve as many features of the original dataset as possible. In case of twin sentences, a 70% lexical overlap was maintained between the sentences, following Sakaguchi et al. (2019). Furthermore, since both sentences in a pair were required to share the same set of answer options, which were identical not only in lexical form but also in grammatical case, deliberate manual adjustments were necessary in the translation. Notably, achieving this is more difficult in Estonian, as its agglutinative structure complicates the preservation of morphological uniformity.

When a direct translation resulted in ambiguous schemas—often because the Estonian equivalent of a word had a broader or narrower meaning than its English counterpart—problematic words were substituted with alternatives that preserved the intended objective of the schema, similarly to Amsili and Seminck (2017). An example of this type of adjustment is presented in Table 3 (A).

Similarly to Amsili and Seminck (2017), adjustments were made in cases where a direct translation of the nouns would have resulted in disagreement in number and would have made the instance trivial to solve based on verb morphology. An exam-

English Schema	**Culturally Adapted Estonian Schema**
(A) Samantha lived in the city while Patricia lived in the desert, so _ thought it common to see a cactus in their yard. **Option 1**: Samantha **Option 2**: Patricia	Sandra elas mandril, kuid Piret elas saarel. Seega oli _ jaoks oma aias kadaka nägemine tavapärane. **Option 1**: Sandra **Option 2**: Piret
(B) Laura beat Erin in a game of Mortal Kombat, then _ said congratulations on the win. **Option 1**: Laura **Option 2**: Erin	Laura tegi Erele lauamängus „Mees, kes teadis ussisõnu" pähe, mispeale õnnitles _ teda võidu puhul. **Option 1**: Laura **Option 2**: Ere

Table 4: Examples of culturally adapted WinoGrande schemas in English and Estonian. In Sample A, correct resolution requires cultural knowledge: the original contrasts city and desert (e.g., cacti in deserts), while the Estonian version replaces these with mainland (*mandril*) and island (*saarel*), and cactus with juniper (*kadaka*), a plant associated with Estonian islands. In Sample B, the game name is culturally adapted, but this knowledge is not necessary for reasoning.

ple of such adjustment is presented in Table 3 (B).

Finally, in order to allow models' evaluation in the few-shot setting, we also translated six schemas from the development set. We selected instances that represent the variability of the tasks, for instance, sentences from the social and physical domain and single and twin sentences.

3.2. Localization, Adaptation and Error Correction

The dataset was localized by replacing geographical locations, brands, foods, activities, and animal species with culturally and regionally appropriate Estonian equivalents. A total of 53 samples were localized. As a result of the localization process, two types of samples can be identified: those in which the culturally adapted information is necessary for reasoning (Table 4, A) , and those in which it is not (Table 4, B). Names were adapted in all samples, with English names replaced by typical Estonian names.

During the manual translation process, 89 samples in the original dataset were found to be either inherently ambiguous or had the wrong answer marked as correct. Such issues were addressed in the human translated dataset by adjusting the schema or editing its answer options to eliminate ambiguity and ensure correctness. The labels marking those problematic items in the original WinoGrande dataset will be released together with the Estonian WinoGrande dataset.

3.3. Inter-Annotator Agreement

The human translation and adaptation was carried out so that the original answer options were always retained. For example, if the correct answer in the original English schema was option 1, the translated and adapted version preserved the same correct option. To assess human-level agreement on the resulting dataset, two additional annotators (both co-authors) independently labeled all items in the translated dataset. Annotators were instructed to select the correct answer option if it was clearly inferable from the context, or to mark it "undecidable" if the item appeared ambiguous and the correct answer could not be determined.

The Cohen's kappa between the two annotators is 0.816, and the Fleiss' kappa between all three annotations (including the original) was 0.855, both values falling into the very high agreement range according to standard interpretation conventions (Landis and Koch, 1977). The accuracy was 95% for Annotator 1 and 92% for Annotator 2. Annotator 1 found 22 items undecidable, while Annotator 2 labeled 106 tasks as such. In the majority of the cases, the annotators disagreed on which items were undecidable, with the overlap being only in eight cases.[2]

4. Machine Translated Dataset

We compare the human-translated dataset with two machine-translated versions produced using GPT-4o and GPT-4.1. The first version was generated with a short, generic translation prompt, while the second was produced using longer, detailed prompts designed to address the issues observed in the first machine translated version.

4.1. Simple Prompt

After producing the translations with the simple zero-shot translation prompt (A.1), we manually analyzed them for any systematic issues that could hinder the objective of the tasks or result in deviations from the required format, thereby potentially affecting the benchmark results.

The manual analysis of the machine translated data revealed the existence of translated schemas that can be solved based on grammatical cues rather than commonsense reasoning (see an example in Table 3 (B)). This can lead to predictions that are either incorrect or correct for the wrong reasons (Elazar et al., 2021; McCoy et al., 2019),

[2]These samples are marked in our dataset.

Language	Twin Sentence 1	Twin Sentence 2
English	I spilled water and jello on the buttons of the remote. It being sticky is probably not because of the _. **Option 1**: jello **Option 2**: water	I spilled water and jello on the buttons of the remote. It's being sticky probably because of the _. **Option 2**: jello **Option 2**: water
Estonian Machine translation Simple prompt	Ma ajasin vett ja želeed pultnuppudele. Selle kleepuv olemine ei ole tõenäoliselt tingitud _. **Option 1**: želee **Option 2**: vesi	Ma ajasin vett ja tarretist puldi nuppudele. See on tõenäoliselt kleepuv _ tõttu. **Option 1**: tarretis **Option 2**: vesi
Estonian Machine translation Detailed prompt	Ma valasin vett ja želeed puldinuppudele. Selle kleepuvus ei ole tõenäoliselt _ tõttu. **Option 1**: želee **Option 2**: vee	Ma valasin vett ja želeed puldinuppudele. See on tõenäoliselt kleepuv _ tõttu. **Option 1**: želee **Option 2**: vee
Estonian Human translation	Ma ajasin puldi nuppudele vett ja tarretist. Selle kleepumine ei tulene ilmselt _. **Option 1**: tarretisest **Option 2**: veest	Ma ajasin puldi nuppudele vett ja tarretist. Selle kleepumine tuleneb ilmselt _. **Option 1**: tarretisest **Option 2**: veest

Table 5: The machine-translated (Simple prompt) twin sentences show substantial lexical divergence (marked with underline). Additionally, the answer options differ across the two sentences. Finally, the options are rendered in the nominative case, which does not fit the grammatical context of the sentence. The machine translation using the detailed prompt results in less lexical divergence between twin sentences and ensures that their corresponding options remain consistent with the intended format, with each pair of twin sentences being assigned identical options. The human translation is also given for the reference.

making the schema unreliable for its intended purpose. Additionally, machine translation sometimes introduced ambiguity into the schema (see an example in Table 3 (A)).

In WinoGrande dataset, the answer options always appear in the sentence. What distinguishes the English dataset from the Estonian dataset is the way the answers appear in the schema. In the English dataset, the answer and its antecedent typically appear in the same form, except for changes in the possessive case for antecedents. In contrast, in the Estonian dataset both the answer and its antecedent can appear in any of the 14 grammatical cases, with the case of the antecedent usually being different from the case of the answer. This nuance becomes apparent in the machine-translated dataset, where the answer options are translated independently of the sentence context and are usually rendered in the nominative case. Consequently, they often fail to fit grammatically within the sentence, which may require a different case for proper syntactic agreement (see Table 5).

For twin sentences, we observed that machine translation often fails to maintain structural and lexical consistency. The machine translated sentence pairs frequently diverge in wording, and the answer choices may be translated inconsistently (see Table 5). While this misalignment may not affect overall performance metrics, it undermines the integrity of the dataset's original format.

4.2. Detailed Prompt

Next, we developed two detailed prompts—one for single sentences (A.2) and the other for the twin sentences (A.3)—and used them to generate a revised translation of the dataset using OpenAI GPT-4.1. The prompts were supplemented with five single and ten twin few-shot examples, respectively. The additional samples were manually selected and translated to reflect different linguistic challenges in the task translation.

These prompts specifically targeted systematic shortcomings identified in the initial machine translation with the simple prompt, such as incorrect inflection of answer options, insufficient lexical overlap between twin sentences (less than 70%), numerical mismatches, and overly literal translations that could compromise the interpretability of the schema. The prompt was supplemented with a list of Estonian names, each labeled with gender (male or female) and inflected across all 14 grammatical cases, to enable automatic localization of names in the translation. The instruction for substituting the names with Estonian ones from the list was included in the prompt.

5. Benchmarking Models

Our comparison of the newly created datasets is based on their applicability as reliable benchmarks for evaluating commonsense reasoning in Estonian. For that purpose, we first used them to obtain and compare overall results across a range of open and closed LLMs (this section). We then conducted a subset-level analysis (Section 6) of these results to better understand the differences in results across the datasets.

The set of models evaluated is following:

- Five proprietary models: Gemini 2.5 Pro, Gemini 2.5 Flash (Gemini Team, Google, 2025), Claude Sonnet 4.5 (Anthropic, 2025), GPT-4.1 (OpenAI et al., 2024), GPT-5 (OpenAI, 2025);
- Four open models in the moderate to large range: Llama 3.3-70B and Llama 3.1-405B (Llama Team, 2024), Gemma 3-27B (Gemma Team, 2025), and Qwen 2.5-72B (Yang et al., 2025);
- Three open models in the smaller range: Llama 3.1-8B (Llama Team, 2024), Apertus 8B (Hernández-Cano et al., 2025), EuroLLM 9B (Martins et al., 2024).

All models were evaluated on all four versions of the WinoGrande dataset: the original English dataset, the human translated Estonian WinoGrande, and the two machine-translated Estonian versions of the dataset. The Gemini 2.5 Pro model was evaluated with minimum thinking (128 token budget), Gemini 2.5 Flash and GPT-5 were evaluated in no thinking mode, and Claude Sonnet 4.5 was evaluated in default thinking mode. All models were prompted in few-shot setting using three examples from the translated development set items. The accuracy results are presented in Table 6.

Compared to the original English WinoGrande test set, the performance of proprietary models on the Estonian human-translated dataset was generally similar, with the average difference being only 0.4%.

Performance differences were more pronounced among open models in the moderate to large model range, where on average, the models performed 5.5% better on English compared to human translated Estonian dataset. While most models performed worse on the Estonian dataset, interestingly Gemma 3-27B performed slightly better.

Among smaller models, the average performance was the same on both English and Estonian datasets, with Llama 3.1-8B and Apertus-8B performing only slightly above chance level on both datasets. EuroLLM-9B was the strongest of the small models according to accuracy, and performed 1.7% better on the Estonian dataset than on the English one.

Comparing the machine translated Estonian WinoGrande versions to the human translated version reveals that all models obtain higher performance on the human translated dataset. A detailed investigation of this observation is presented in Section 6.

	EN	HT	MT	
			Simple	Detailed
Proprietary models				
Gemini 2.5 Pro	89.6	91.1	83.9	82.7
Gemini 2.5 Flash	87.8	88.6	82.2	81.8
Claude Sonnet 4.5	94.7	93.7	88.3	87.4
GPT-4.1	85.4	82.6	76.1	77.2
GPT-5	83.5	82.9	76.6	76.4
Average	88.2	87.8	81.4	81.1
Open models: moderate to large range				
Gemma 3-27B	73.6	74.6	68.8	69.6
Qwen 2.5-72B	83.4	71.4	65.7	66.2
Llama 3.3-70B	79.9	74.8	69.4	69.7
Llama 3.1-405B	84.3	78.4	72.0	72.7
Average	80.3	74.8	68.9	69.6
Open models: smaller range				
Llama 3.1-8B	55.5	54.2	53.8	53.1
Apertus-8B	51.9	51.5	51.2	51.2
EuroLLM-9B	59.6	61.3	59.4	57.7
Average	55.7	55.7	54.8	54.0

Table 6: Accuracy results reported across models for the WinoGrande English test set (EN), Estonian human-translated version (HT), the Estonian machine-translated version using a simple prompt (MT Simple), and the detailed prompt (MT Detailed)

Contrasting the results on the two machine translation versions reveals only minor differences, suggesting that the translations obtained with the detailed prompt did not yield an substantial effect on the models' performance.

6. Subset-Level Analysis

We analyzed the impact of cultural adaptation, and error correction in human translation on the overall results by calculating the models' accuracies for subsets of samples that were culturally adapted (n = 53), corrected (n = 89), and the rest that were deemed semantically comparable (n = 1,617). The 8 samples that both annotators found ambiguous were excluded from the subset analysis. For these analyses, we compared the results of open models of moderate to large size, as smaller models often performed near random even on the English dataset. We compared the results across three versions of the dataset—the original English, the human translated Estonian, and the machine trans-

	EN	HT	MT
Culturally adapted $n = 53$			
Gemma 3-27B	84.9	81.1	77.8
Qwen 2.5-72B	86.8	77.4	73.6
Llama 3.3-70B	88.7	81.1	81.5
Llama 3.1-405B	92.5	83.0	81.1
Average	88.2	80.7	78.5
Corrected $n = 89$			
Gemma 3-27B	49.4	71.9	41.6
Qwen 2.5-72B	58.4	68.5	42.7
Llama 3.3-70B	61.8	73.0	47.2
Llama 3.1-405B	56.2	78.7	48.3
Average	56.5	73.0	44.9
Semantically comparable $n = 1,617$			
Gemma 3-27B	74.5	74.6	70.9
Qwen 2.5-72B	84.6	71.2	67.3
Llama 3.3-70B	80.5	74.6	70.4
Llama 3.1-405B	85.6	78.2	74.0
Average	81.3	74.7	70.7

Table 7: Accuracy on different subsets for the English (EN), Estonian human-translated (HT), and the Estonian machine-translated version using the detailed prompt (MT).

lated Estonian created with the detailed prompt.[3]

6.1. The Impact of Cultural Adaptation

We compared how models perform on culturally adapted examples versus those human translated examples that did not undergo such adaptation. During the human translation process, 53 schemas were thoroughly adapted by editing the content of the schema to reflect cultural relevance. We observe that, across all models, accuracies are on average 6% higher on the human translated subset that was culturally adapted (Table 7, *Culturally adapted* samples) compared to the human translated subset that did not undergo such adaptation (Table 7, *Semantically comparable* samples). Regardless of the slightly higher accuracies, we do not interpret this as evidence of the models' stronger ability to reason based on cultural knowledge. This is because the nature of the adaptations varies: some introduce contextually relevant information—such as geographic references—that require cultural or situational understanding, while others involve changes unrelated to reasoning (Table 4).

[3]The machine translated dataset created with the simple prompt shows very similar results.

Model	MT Simple		MT Detailed	
	MT	HT	MT	HT
Altered meaning	$n = 246$		$n = 191$	
Gemma 3-27B	63.4	77.2	62.3	74.9
Qwen 2.5-72B	67.1	72.0	65.5	69.6
Llama 3.3-70B	67.5	80.1	65.5	75.4
Llama 3.1-405B	66.3	82.1	63.9	73.8
Average	66.1	77.9	64.3	73.4
Retained meaning	$n = 1371$		$n = 1426$	
Gemma 3-27B	71.2	74.1	72.1	74.5
Qwen 2.5-72B	66.5	71.1	67.5	71.5
Llama 3.3-70B	70.8	73.6	71.1	74.5
Llama 3.1-405B	74.9	77.5	75.3	78.8
Average	70.8	74.1	71.5	74.8

Table 8: Comparison of the accuracy scores for schemas where machine translation (MT) either altered or retained the original meaning.

6.2. The Impact of Incorrect Schemas

Problematic English schemas were identified manually during the translation process. The results presented in Table 7 (*Corrected* samples) confirm the lower quality of these English schemas, as the performance on this English subset is close to the chance level.

Because these schemas were corrected during the human translation process, performance on the human translated Estonian samples is considerably higher compared to the same subset in English. This improvement in accuracies highlights the importance of quality assurance in schema translation. Importantly, the results indicate that the adjustments preserved the integrity of the tasks and schemas were not rendered trivial by the edits, as the outcomes from the subset of corrected schemas do not differ considerably from those of the subset of 1,617 human translated schemas that did not undergo such corrections (Table 7, *Semantically comparable* samples).

The performance of the machine translated version of this subset is very low, even below the chance level. This demonstrates that the errors in the original English dataset propagate during the machine translation process in unexpected ways, potentially amplifying the original errors.

6.3. The Impact of Machine Translation

To study the effect of machine translation on the results, we had a qualified translator (the same person who did the human translation) manually analyze the subset of 1,617 semantically comparable samples to identify instances where the machine translation had altered the meaning of the original

Version	Sentence	Option 1	Option 2
EN	So _ was sorry because Christine's cat was bitten by Jessica's dog when they got into a fight.	Christine	Jessica
MT-detailed	Nii et _ oli kahju, sest Kristiina kass hammustas Jessika koera, kui nad kaklesid.	Kristiina	Jessika
Back translation	So _ was upset, because Kristiina's cat bit Jessica's dog when they were fighting.	Kristiina	Jessica

Table 9: Example showing original English sentence, its machine-translated version in Estonian, and the back translation with Open AI GPT-5. The schema has become ambiguous in translation and the meaning has also shifted (e.g., "the cat bit the dog" versus "the dog bit the cat").

sentence. In this labeling, meaning was the sole evaluation criterion; grammatical issues, unnatural phrasing, or other surface-level errors were not considered problematic unless they affected the sentence's meaning. The guiding question was: *Does the machine-translated schema convey the same meaning as the original English schema?*.[4]

The analysis of the machine translations generated with the simple prompt identified 246 (15.2%) schemas in which the meaning was lost or altered. The translations produced with the detailed prompt resulted in 191 (11.8%) such schemas. An example of a schema with a shifted meaning is shown in Table 9.

We then compared results on the subset of schemas where machine translation had altered the meaning to those where machine translation had retained the meaning (**MT** columns in Table 8). The shifts in meaning have a clear impact on performance—with one exception (Qwen 2.5-72B), all models perform worse on machine translated schemas with altered meaning. With the simple prompt, the difference in accuracies between the averages of the two subsets is 4.7% and with the detailed prompt it is 7.2%. However, it must be noted that, besides human-perceived shift in meaning, machine translated instances contain other artifacts that may affect the accuracy even when the meaning is retained. Thus, we also compared the subset of schemas with retained meaning with the human translated version of the same subset (rows in the lower part of Table 8). This comparison shows that the accuracies on the machine translated datasets are lower (3.8% with the simple prompt and 3.3% with the detailed prompt) indicating that there are issues beyond altered meaning in the machine translated data that also potentially affect task integrity.

7. Discussion

In this paper, we presented the Estonian translation of the WinoGrande test set, which has been translated and culturally adapted by translation specialists and annotated to ensure that translated Estonian schemas satisfy the requirements of the WinoGrande benchmark. We evaluated model performance on the Estonian dataset to assess their reasoning abilities in Estonian. We also explored whether the insights gained from the translation process can be used to engineer a detailed machine translation prompt capable of producing a machine translated dataset of comparable quality.

We observed that the performance for proprietary models was notably good for all models on the English and Estonian dataset, and the results were similar on both datasets. The difference was more pronounced for the open models. The open models performed slightly worse on the Estonian dataset. Smaller range models demonstrated generally poor performance on the task itself, with the exception of EuroLLM-9B, which is specifically trained for European languages.

During the translation process, we identified a number of flawed schemas in the original dataset. Unlike the original Winograd Schema Challenge dataset, which was developed by experts, WinoGrande was crowd-sourced. This difference in methodology likely accounts for the variability in schema quality and also speaks for the importance of involving language specialists in the development of benchmark tasks requiring advanced linguistic competence.

We noted a substantial difference in the number of items the annotators labeled as undecidable. Although the discrepancy is numerically large, we do not consider it significant, as the annotators' accuracy remained close to the reported human performance level (94%) on the English test set (Sakaguchi et al., 2019).

Our experiments with machine translation showed that, despite fine-tuning the translation prompt to account for the dataset's intricacies and the specific target language, the detailed prompt failed to produce a dataset of human-comparable quality. It is also important to highlight that constructing an effective prompt tailored to a specific language and dataset demands detailed knowl-

[4]These annotations will be released with the dataset.

edge of the dataset's structure and purpose as well as the potential linguistic challenges it presents in a given language. As such, the process still requires significant input from language experts to ensure the prompt is both linguistically and contextually appropriate. Our analysis of the machine-translated dataset focused solely on shifts in meaning during translation. We did not evaluate the translations for grammatical accuracy or the naturalness of the language. Further research could explore how these factors influence model performance.

8. Conclusion

The Estonian translation of the WinoGrande test set provides a valuable linguistic benchmark for Estonian and supports the development and evaluation of multilingual language models. The experimental results demonstrate strong performance from proprietary language models in Estonian, while outcomes from open-source models show greater variability. The comparison with machine translation shows that prompt engineering offers limited benefit to translation quality and the results obtained with machine translation may not accurately reflect the models' language comprehension or reasoning. When a model is presented with incoherent or semantically distorted input alongside answer options, it will return a prediction. However, such predictions can be uninterpretable because they do not provide meaningful insight into the model's linguistic understanding or reasoning capabilities, since the input lacks a coherent structure to support such inference.

Acknowledgements

This work was supported by the National Program for Estonian Language Technology Program (project EKTB104) funded by the Estonian Ministry of Education and Research, and partially supported by the Estonian Research Council Grant PSG721.

9. References

Tuka Alhanai, Adam Kasumovic, Mohammad M. Ghassemi, Aven Zitzelberger, Jessica M. Lundin, and Guillaume Chabot-Couture. 2025. Bridging the Gap: Enhancing LLM Performance for Low-Resource African Languages with New Benchmarks, Fine-Tuning, and Cultural Adjustments. *Proceedings of the AAAI Conference on Artificial Intelligence*, 39(27):27802–27812.

Pascal Amsili and Olga Seminck. 2017. A Google-Proof Collection of French Winograd Schemas. In *Proceedings of the 2nd Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2017)*, pages 24–29, Valencia, Spain. Association for Computational Linguistics.

Anthropic. 2025. System Card: Claude Sonnet 4.5. `https://assets.anthropic.com/m/12f214efcc2f457a/original/Claude-Sonnet-4-5-System-Card.pdf`.

Maksim Aparovich, Volha Harytskaya, Vladislav Poritski, Oksana Volchek, and Pavel Smrz. 2025. BelarusianGLUE: Towards a Natural Language Understanding Benchmark for Belarusian. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 511–527, Vienna, Austria. Association for Computational Linguistics.

Mikel Artetxe, Gorka Labaka, and Eneko Agirre. 2020. Translation Artifacts in Cross-lingual Transfer Learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 7674–7684, Online. Association for Computational Linguistics.

Phakphum Artkaew. 2025. Thai Winograd Schemas: A Benchmark for Thai Commonsense Reasoning. In *Proceedings of the Second Workshop in South East Asian Language Processing*, pages 42–51, Online. Association for Computational Linguistics.

Timothée Bernard and Ting Han. 2020. Mandarinograd: A Chinese Collection of Winograd Schemas. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 21–26, Marseille, France. European Language Resources Association.

Ander Corral, Ixak Sarasua, and Xabier Saralegi. 2024. Pipeline Analysis for Developing Instruct LLMs in Low-Resource Languages: A Case Study on Basque.

John Dang, Shivalika Singh, Daniel D'souza, Arash Ahmadian, Alejandro Salamanca, Madeline Smith, Aidan Peppin, Sungjin Hong, Manoj Govindassamy, Terrence Zhao, et al. 2024. Aya expanse: Combining research breakthroughs for a new multilingual frontier. *arXiv preprint arXiv:2412.04261*.

Yanai Elazar, Hongming Zhang, Yoav Goldberg, and Dan Roth. 2021. Back to Square One: Artifact Detection, Training and Commonsense Disentanglement in the Winograd Schema. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages

10486–10500, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Farhan Farsi, Farnaz Aghababaloo, Shahriar Shariati Motlagh, Parsa Ghofrani, MohammadAli SadraeiJavaheri, Shayan Bali, Amirhossein Shabani, Farbod Bijary, Ghazal Zamaninejad, AmirMohammad Salehoof, and Saeedeh Momtazi. 2025. MELAC: Massive Evaluation of Large Language Models with Alignment of Culture in Persian Language.

Negar Foroutan, Mohammadreza Banaei, Karl Aberer, and Antoine Bosselut. 2023. Breaking the Language Barrier: Improving Cross-Lingual Reasoning with Structured Self-Attention. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 9422–9442, Singapore. Association for Computational Linguistics.

Gemini Team, Google. 2025. Gemini 2.5: Pushing the Frontier with Advanced Reasoning, Multimodality, Long Context, and Next Generation Agentic Capabilities. *arXiv preprint arXiv:2507.06261*.

Google Gemma Team. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Google DeepMind Gemma Team. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, et al. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Alejandro Hernández-Cano, Alexander Hägele, Allen Hao Huang, Angelika Romanou, Antoni-Joan Solergibert, Barna Pasztor, Bettina Messmer, Dhia Garbaya, Eduard Frank Ďurech, Ido Hakimi, Juan García Giraldo, Mete Ismayilzada, Negar Foroutan, Skander Moalla, Tiancheng Chen, Vinko Sabolčec, Yixuan Xu, Michael Aerni, Badr AlKhamissi, Ines Altemir Marinas, Mohammad Hossein Amani, Matin Ansaripour, Ilia Badanin, Harold Benoit, Emanuela Boros, Nicholas Browning, Fabian Bösch, Maximilian Böther, Niklas Canova, Camille Challier, Clement Charmillot, Jonathan Coles, Jan Deriu, Arnout Devos, Lukas Drescher, Daniil Dzenhaliou, Maud Ehrmann, Dongyang Fan, Simin Fan, Silin Gao, Miguel Gila, María Grandury, Diba Hashemi, Alexander Hoyle, Jiaming Jiang, Mark Klein, Andrei Kucharavy, Anastasiia Kucherenko, Frederike Lübeck, Roman Machacek, Theofilos Manitaras, Andreas Marfurt, Kyle Matoba, Simon Matrenok, Henrique Mendoncça, Fawzi Roberto Mohamed, Syrielle Montariol, Luca Mouchel, Sven Najem-Meyer, Jingwei Ni, Gennaro Oliva, Matteo Pagliardini, Elia Palme, Andrei Panferov, Léo Paoletti, Marco Passerini, Ivan Pavlov, Auguste Poiroux, Kaustubh Ponkshe, Nathan Ranchin, Javi Rando, Mathieu Sauser, Jakhongir Saydaliev, Muhammad Ali Sayfiddinov, Marian Schneider, Stefano Schuppli, Marco Scialanga, Andrei Semenov, Kumar Shridhar, Raghav Singhal, Anna Sotnikova, Alexander Sternfeld, Ayush Kumar Tarun, Paul Teiletche, Jannis Vamvas, Xiaozhe Yao, Hao Zhao Alexander Ilic, Ana Klimovic, Andreas Krause, Caglar Gulcehre, David Rosenthal, Elliott Ash, Florian Tramèr, Joost VandeVondele, Livio Veraldi, Martin Rajman, Thomas Schulthess, Torsten Hoefler, Antoine Bosselut, Martin Jaggi, and Imanol Schlag. 2025. Apertus: Democratizing Open and Compliant LLMs for Global Language Environments.

Albert Q Jiang, Alexandre Sablayrolles, Antoine Roux, Arthur Mensch, Blanche Savary, Chris Bamford, Devendra Singh Chaplot, Diego de las Casas, Emma Bou Hanna, Florian Bressand, et al. 2024. Mixtral of Experts. *arXiv preprint arXiv:2401.04088*.

Hyeonwoo Kim, Dahyun Kim, Jihoo Kim, Sukyung Lee, Yungi Kim, and Chanjun Park. 2025. Open Ko-LLM Leaderboard2: Bridging Foundational and Practical Evaluation for Korean LLMs.

Vid Kocijan, Ana-Maria Cretu, Oana-Maria Camburu, Yordan Yordanov, and Thomas Lukasiewicz. 2019. A Surprisingly Robust Trick for the Winograd Schema Challenge. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 4837–4842, Florence, Italy. Association for Computational Linguistics.

Vid Kocijan, Ernest Davis, Thomas Lukasiewicz, Gary Marcus, and Leora Morgenstern. 2023. The Defeat of the Winograd Schema Challenge.

Julia Kreutzer, Eleftheria Briakou, Sweta Agrawal, Marzieh Fadaee, and Kocmi Tom. 2025. D\'ej\a vu: Multilingual llm evaluation through the lens of machine translation evaluation. *arXiv preprint arXiv:2504.11829*.

Viet Lai, Chien Nguyen, Nghia Ngo, Thuat Nguyen, Franck Dernoncourt, Ryan Rossi, and Thien Nguyen. 2023. Okapi: Instruction-tuned Large Language Models in Multiple Languages with Reinforcement Learning from Human Feedback.

In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 318–327, Singapore. Association for Computational Linguistics.

J Richard Landis and Gary G Koch. 1977. The Measurement of Observer Agreement for Categorical Data. *Biometrics*, pages 159–174.

Hector J. Levesque, Ernest Davis, and Leora Morgenstern. 2012. The Winograd Schema Challenge. In *Proceedings of the Thirteenth International Conference on Principles of Knowledge Representation and Reasoning*, KR'12, page 552–561. AAAI Press.

AI @ Meta Llama Team. 2024. The Llama 3 Herd of Models. *arXiv preprint arXiv:2407.21783*.

Nicholas Lourie, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2021. UNICORN on RAINBOW: A Universal Commonsense Reasoning Model on a New Multitask Benchmark. *Proceedings of the AAAI Conference on Artificial Intelligence*, 35(15):13480–13488.

Pedro Henrique Martins, Patrick Fernandes, João Alves, Nuno M. Guerreiro, Ricardo Rei, Duarte M. Alves, José Pombal, Amin Farajian, Manuel Faysse, Mateusz Klimaszewski, Pierre Colombo, Barry Haddow, José G. C. de Souza, Alexandra Birch, and André F. T. Martins. 2024. EuroLLM: Multilingual Language Models for Europe.

Mihai Masala, Denis Ilie-Ablachim, Alexandru Dima, Dragos Georgian Corlatescu, Miruna-Andreea Zavelca, Ovio Olaru, Simina-Maria Terian, Andrei Terian, Marius Leordeanu, Horia Velicu, Marius Popescu, Mihai Dascalu, and Traian Rebedea. 2024. "Vorbești Românește?" A Recipe to Train Powerful Romanian LLMs with English Instructions. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11632–11647, Miami, Florida, USA. Association for Computational Linguistics.

R Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448.

Gabriela Melo, Vinicius Imaizumi, and Fábio Cozman. 2019. Winograd Schemas in Portuguese. In *Anais do XVI Encontro Nacional de Inteligência Artificial e Computacional*, pages 787–798, Porto Alegre, RS, Brasil. SBC.

Luca Moroni, Simone Conia, Federico Martelli, and Roberto Navigli. 2024. Towards a More Comprehensive Evaluation for Italian LLMs. In *Proceedings of the Tenth Italian Conference on Computational Linguistics (CLiC-it 2024)*, pages 584–599, Pisa, Italy. CEUR Workshop Proceedings.

Basel Mousi, Nadir Durrani, Fatema Ahmad, Md. Arid Hasan, Maram Hasanain, Tameem Kabbani, Fahim Dalvi, Shammur Absar Chowdhury, and Firoj Alam. 2025. AraDiCE: Benchmarks for dialectal and cultural capabilities in LLMs. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4186–4218, Abu Dhabi, UAE. Association for Computational Linguistics.

Artūras Nakvosas, Povilas Daniušis, and Vytas Mulevičius. 2025. Open Llama2 Models for the Lithuanian Language. *Informatica*, page 385–406.

Jinu Nyachhyon, Mridul Sharma, Prajwal Thapa, and Bal Krishna Bal. 2025. Consolidating and Developing Benchmarking Datasets for the Nepali Natural Language Understanding Tasks.

OpenAI. 2025. GPT-5 System Card. `https://cdn.openai.com/gpt-5-system-card.pdf`.

OpenAI, Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, Red Avila, Igor Babuschkin, Suchir Balaji, Valerie Balcom, Paul Baltescu, Haiming Bao, Mohammad Bavarian, Jeff Belgum, Irwan Bello, Jake Berdine, Gabriel Bernadett-Shapiro, Christopher Berner, Lenny Bogdonoff, Oleg Boiko, Madelaine Boyd, Anna-Luisa Brakman, Greg Brockman, Tim Brooks, Miles Brundage, Kevin Button, Trevor Cai, Rosie Campbell, Andrew Cann, Brittany Carey, Chelsea Carlson, Rory Carmichael, Brooke Chan, Che Chang, Fotis Chantzis, Derek Chen, Sully Chen, Ruby Chen, Jason Chen, Mark Chen, Ben Chess, Chester Cho, Casey Chu, Hyung Won Chung, Dave Cummings, Jeremiah Currier, Yunxing Dai, Cory Decareaux, Thomas Degry, Noah Deutsch, Damien Deville, Arka Dhar, David Dohan, Steve Dowling, Sheila Dunning, Adrien Ecoffet, Atty Eleti, Tyna Eloundou, David Farhi, Liam Fedus, Niko Felix, Simón Posada Fishman, Juston Forte, Isabella Fulford, Leo Gao, Elie Georges, Christian Gibson, Vik Goel, Tarun Gogineni, Gabriel Goh, Rapha Gontijo-Lopes, Jonathan Gordon, Morgan Grafstein, Scott Gray, Ryan Greene, Joshua Gross, Shixiang Shane Gu, Yufei Guo, Chris Hallacy, Jesse Han, Jeff Harris, Yuchen He, Mike Heaton, Johannes Heidecke, Chris Hesse, Alan Hickey, Wade Hickey, Peter Hoeschele, Brandon Houghton, Kenny Hsu, Shengli Hu, Xin Hu, Joost Huizinga, Shantanu Jain, Shawn Jain, Joanne Jang, Angela Jiang, Roger Jiang, Haozhun Jin, Denny Jin, Shino Jomoto, Billie Jonn, Heewoo Jun, Tomer Kaftan, Łukasz Kaiser, Ali Kamali, Ingmar Kanitscheider, Nitish Shirish Keskar, Tabarak Khan, Logan Kilpatrick, Jong Wook Kim, Christina Kim, Yongjik Kim, Jan Hendrik Kirchner, Jamie Kiros, Matt Knight, Daniel Kokotajlo, Łukasz Kondraciuk, Andrew Kondrich, Aris Konstantinidis, Kyle Kosic, Gretchen Krueger, Vishal Kuo, Michael Lampe, Ikai Lan, Teddy Lee, Jan Leike, Jade Leung, Daniel Levy, Chak Ming Li, Rachel Lim, Molly Lin, Stephanie Lin, Mateusz Litwin, Theresa Lopez, Ryan Lowe, Patricia Lue, Anna Makanju, Kim Malfacini, Sam Manning, Todor Markov, Yaniv Markovski, Bianca Martin, Katie Mayer, Andrew Mayne, Bob McGrew, Scott Mayer McKinney, Christine McLeavey, Paul McMillan, Jake McNeil, David Medina, Aalok Mehta, Jacob Menick, Luke Metz, Andrey Mishchenko, Pamela Mishkin, Vinnie Monaco, Evan Morikawa, Daniel Mossing, Tong Mu, Mira Murati, Oleg Murk, David Mély, Ashvin Nair, Reiichiro Nakano, Rajeev Nayak, Arvind Neelakantan, Richard Ngo, Hyeonwoo Noh, Long Ouyang, Cullen O'Keefe, Jakub Pachocki, Alex Paino, Joe Palermo, Ashley Pantuliano, Giambattista Parascandolo, Joel Parish, Emy Parparita, Alex Passos, Mikhail Pavlov, Andrew Peng, Adam Perelman, Filipe de Avila Belbute Peres, Michael Petrov, Henrique Ponde de Oliveira Pinto, Michael, Pokorny, Michelle Pokrass, Vitchyr H. Pong, Tolly Powell, Alethea Power, Boris Power, Elizabeth Proehl, Raul Puri, Alec Radford, Jack Rae, Aditya Ramesh, Cameron Raymond, Francis Real, Kendra Rimbach, Carl Ross, Bob Rotsted, Henri Roussez, Nick Ryder, Mario Saltarelli, Ted Sanders, Shibani Santurkar, Girish Sastry, Heather Schmidt, David Schnurr, John Schulman, Daniel Selsam, Kyla Sheppard, Toki Sherbakov, Jessica Shieh, Sarah Shoker, Pranav Shyam, Szymon Sidor, Eric Sigler, Maddie Simens, Jordan Sitkin, Katarina Slama, Ian Sohl, Benjamin Sokolowsky, Yang Song, Natalie Staudacher, Felipe Petroski Such, Natalie Summers, Ilya Sutskever, Jie Tang, Nikolas Tezak, Madeleine B. Thompson, Phil Tillet, Amin Tootoonchian, Elizabeth Tseng, Preston Tuggle, Nick Turley, Jerry Tworek, Juan Felipe Cerón Uribe, Andrea Vallone, Arun Vijayvergiya, Chelsea Voss, Carroll Wainwright, Justin Jay Wang, Alvin Wang, Ben Wang, Jonathan Ward, Jason Wei, CJ Weinmann, Akila Welihinda, Peter Welinder, Jiayi Weng, Lilian Weng, Matt Wiethoff, Dave Willner, Clemens Winter, Samuel Wolrich, Hannah Wong, Lauren Workman, Sherwin Wu, Jeff Wu, Michael Wu, Kai Xiao, Tao Xu, Sarah Yoo, Kevin Yu, Qiming Yuan, Wojciech Zaremba, Rowan Zellers, Chong Zhang, Marvin Zhang, Shengjia Zhao, Tianhao Zheng, Juntang Zhuang, William Zhuk, and Barret Zoph. 2024. GPT-4 Technical Report.

Irene Plaza, Nina Melero, Cristina del Pozo, Javier Conde, Pedro Reviriego, Marina Mayor-Rocher, and María Grandury. 2024. Spanish and LLM Benchmarks: is MMLU Lost in Translation?

Nishat Raihan, Antonios Anastasopoulos, and Marcos Zampieri. 2025. mHumanEval - A Multilingual Benchmark to Evaluate Large Language Models for Code Generation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 11432–11461, Albuquerque, New Mexico. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender Bias in Coreference Resolution.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. 2019. WinoGrande: An Adversarial Winograd Schema Challenge at Scale.

Guokan Shang, Hadi Abdine, Ahmad Chamma, Amr Mohamed, Mohamed Anwar, Abdelaziz Bounhar, Omar El Herraoui, Preslav Nakov, Michalis Vazirgiannis, and Eric Xing. 2025. Nile-Chat: Egyptian Language Models for Arabic and Latin Scripts.

Tatiana Shavrina, Alena Fenogenova, Emelyanov Anton, Denis Shevelev, Ekaterina Artemova, Valentin Malykh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. RussianSuperGLUE: A Russian Language Understanding Evaluation Benchmark. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4717–4726, Online. Association for Computational Linguistics.

Tomohide Shibata, Shotaro Kohama, and Sadao Kurohashi. 2015. winograd schema challenge. In *Proceedings of the Annual Meeting of the Association for Natural Language Processing (NLP2015)*, pages 493–496, Kyoto, Japan. Association for Natural Language Processing.

Vered Shwartz. 2024. Winograd schema challenge datasets. `https://www.cs.ubc.ca/~vshwartz/resources/winograd_he.html`. Accessed: 2026-02-27.

Shivalika Singh, Freddie Vargus, Daniel D'souza, Börje F. Karlsson, Abinaya Mahendiran, Wei-Yin Ko, Herumb Shandilya, Jay Patel, Deividas Mataciunas, Laura O'Mahony, Mike Zhang, Ramith Hettiarachchi, Joseph Wilson, Marina Machado, Luisa Moura, Dominik Krzemiński, Hakimeh Fadaei, Irem Ergun, Ifeoma Okoh, Aisha Alaagib, Oshan Mudannayake, Zaid Alyafeai, Vu Chien, Sebastian Ruder, Surya Guthikonda, Emad Alghamdi, Sebastian Gehrmann, Niklas Muennighoff, Max Bartolo, Julia Kreutzer, Ahmet Üstün, Marzieh Fadaee, and Sara Hooker. 2024. Aya Dataset: An Open-Access Collection for Multilingual Instruction Tuning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11521–11567, Bangkok, Thailand. Association for Computational Linguistics.

Vésteinn Snæbjarnarson, Haukur Barri Símonarson, Pétur Orri Ragnarsson, Svanhvít Lilja Ingólfsdóttir, Haukur Jónsson, Vilhjalmur Thorsteinsson, and Hafsteinn Einarsson. 2022. A Warm Start and a Clean Crawled Corpus - A Recipe for Good Language Models. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 4356–4366, Marseille, France. European Language Resources Association.

Klaudia Thellmann, Bernhard Stadler, Michael Fromm, Jasper Schulze Buschhoff, Alex Jude, Fabio Barth, Johannes Leveling, Nicolas Flores-Herr, Joachim Köhler, René Jäkel, and Mehdi Ali. 2024. Towards Multilingual LLM Evaluation for European Languages.

Hugo Touvron, Louis Martin, Kevin Stone, Peter Albert, Amjad Almahairi, Yasmine Babaei, Nikolay Bashlykov, Soumya Batra, Prajjwal Bhargava, Shruti Bhosale, et al. 2023. Llama 2: Open Foundation and Fine-Tuned Chat Models. *arXiv preprint arXiv:2307.09288*.

Timur Turatali, Aida Turdubaeva, Islam Zhenishbekov, Zhoomart Suranbaev, Anton Alekseev, and Rustem Izmailov. 2025. Bridging the Gap in Less-Resourced Languages: Building a Benchmark for Kyrgyz Language Models. In *2025 10th International Conference on Computer Science and Engineering (UBMK)*, pages 1673–1677.

Noémi Vadász and Noémi Ligeti-Nagy. 2022. Winograd Schemata and Other Datasets for Anaphora Resolution in Hungarian. *Acta Linguistica Academica*.

Alex Wang, Yada Pruksachatkun, Nikita Nangia, Amanpreet Singh, Julian Michael, Felix Hill, Omer Levy, and Samuel R. Bowman. 2020. SuperGLUE: A Stickier Benchmark for General-Purpose Language Understanding Systems.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiaxi Yang, Jing Zhou, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Lianghao Deng, Mei Li, Mingfeng Xue, Mingze Li, Pei Zhang, Peng Wang, Qin Zhu, Rui Men, Ruize Gao, Shixuan Liu, Shuang Luo, Tianhao Li, Tianyi Tang, Wenbiao Yin, Xingzhang Ren, Xinyu Wang, Xinyu Zhang, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yinger Zhang, Yu Wan, Yuqiong Liu, Zekun Wang, Zeyu Cui, Zhenru Zhang, Zhipeng Zhou, and Zihan Qiu. 2025. Qwen3 Technical Report.

Aleš Žagar and Marko Robnik-Šikonja. 2022. Slovene SuperGLUE benchmark: Translation and evaluation. In *Proceedings of the Thirteenth Language Resources and Evaluation Conference*, pages 2058–2065, Marseille, France. European Language Resources Association.

Hongming Zhang, Xinran Zhao, and Yangqiu Song. 2020. WinoWhy: A Deep Diagnosis of Essential Commonsense Knowledge for Answering Winograd Schema Challenge. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5736–5745, Online. Association for Computational Linguistics.

A. Translation Prompts

A.1. Simple Translation

> Please translate the given example in JSON format into Estonian. Retain the JSON structure. Do not translate the keys. Translate the values only. Please output only the JSON object, nothing else.

A.2. Detailed Prompt for Single Items

> Please translate the given example in JSON format into Estonian so that it reads as if originally written in Estonian: natural, fluent, and culturally appropriate. It consists of a sentence and two answer options.
>
> Do not translate word for word. Always choose the contextually correct meaning of each word, especially adjectives, idioms, and figurative expressions - never use a literal dictionary equivalent if it does not fit the situation. Ensure that both answer options remain grammatically and semantically valid in the sentence, without revealing or implying which is correct. Pay particular attention to number and case: both answer options must agree in grammatical form so they can fit the same syntactic slot. If the two answer options translations differ in grammatical number (singular vs plural), replace one option with a contextually appropriate alternative so that both options agree in form (e.g. both plural nouns, both singular nouns). The replacement must still preserve the fairness and neutrality of the question.
>
> Replace all names with appropriate ones from the provided table, matching gender and applying the correct Estonian case endings. Select appropriate names at random. A name may be reused across different sentences, but not within the same sentence. Maintain neutrality and difficulty: the question must stay fair, challenging, and unbiased.
>
> Retain the JSON structure. Do not translate the keys. Translate the values only. Please output only the JSON object, nothing else. Do not escape with backticks or add line breaks.

A.3. Translation Prompt for Twin Sentences

The following are sentence pairs with a gap expressed with the underscore and two answer options. Your task is to translate each sentence pair into Estonian so that it reads as if originally written in Estonian: natural, fluent, and culturally appropriate.

The sentence pairs have a lexical overlap of at least 70%, and their translations must also maintain at least 70% overlap. This overlap should be preserved consistently, without introducing unnecessary variation. Do not vary style, synonyms, or sentence structure between them. Answer options must also remain exactly the same in these sentence pairs.

Do not translate word for word. Always choose the contextually correct meaning of each word, especially adjectives, idioms, and figurative expressions — never use a literal dictionary equivalent if it does not fit the situation. Ensure that both answer options remain grammatically and semantically valid in the sentences, without revealing or implying which is correct. Pay particular attention to number and case: both answer options must agree in grammatical form so they can fit the same syntactic slot.

If the two answer options translations differ in grammatical number (singular vs plural), replace one option with a contextually appropriate alternative so that both options agree in form (e.g. both plural nouns, both singular nouns). The replacement must still preserve the fairness and neutrality of the question.

Do not fill the gap. Replace all names with randomly chosen ones from the provided table, matching gender and applying correct Estonian case endings. Both names in a sentence must be the same gender, and names may be reused across sentence pairs but not within the same sentence. For sentence pairs, use the same names consistently in both sentences.

Do not add names where none exist. Maintain neutrality and difficulty: the question must stay fair, challenging, and unbiased. You will receive a JSON list containing two examples as input. Your output should also be JSON list containing the translated sentences and answer options. Make sure you translate only the values and not the keys. Please output only the JSON list, nothing else. Do not escape with backticks or add addtional line breaks.