Instruction Tuning on Public Government and Cultural Data for Low-Resource Language: a Case Study in Kazakh

Nurkhan Laiyk[1*] **Daniil Orel**[1*] **Rituraj Joshi**[3] **Maiya Goloburda**[1]
Yuxia Wang[1] **Preslav Nakov**[1,2] **Fajri Koto**[1]
[1]Mohamed bin Zayed University of Artificial Intelligence
[2]Institute of Foundation Models
[3]Cerebras Systems
{nurkhan.laiyk,daniil.orel}@mbzuai.ac.ae

Abstract

Instruction tuning in low-resource languages remains underexplored due to limited text data, particularly in government and cultural domains. To address this, we introduce and open-source a large-scale (10,600 samples) instruction-following (IFT) dataset,[1] covering key institutional and cultural knowledge relevant to Kazakhstan. Our dataset enhances LLMs' understanding of procedural, legal, and structural governance topics. We employ LLM-assisted data generation, comparing open-weight and closed-weight models for dataset construction, and select GPT-4o as the backbone. Each entity of our dataset undergoes full manual verification to ensure high quality. We also show that fine-tuning Qwen, Falcon, and Gemma on our dataset leads to consistent performance improvements in both multiple-choice and generative tasks, demonstrating the potential of LLM-assisted instruction tuning for low-resource languages.

1 Introduction

Instruction tuning enhances large language models (LLMs) by fine-tuning them on structured prompts, improving their ability to follow human instructions across various tasks such as question answering and summarization (Ouyang et al., 2022). While extensive instruction-tuning datasets exist for English, such as FLAN (Longpre et al., 2023), P3 (Sanh et al., 2021), and Dolly (Conover et al., 2023), efforts in low-resource languages remain limited. This gap is particularly evident in domain-specific applications where multilingual LLMs often provide generic or inaccurate responses due to a lack of localized training data (Li et al., 2023).

A key challenge in adapting LLMs to underrepresented languages is the scarcity of high-quality instruction data (Li et al., 2023). Multilingual models may process low-resource languages at a technical level (OpenAI, 2024), but their practical effectiveness is often constrained by an incomplete understanding of region-specific socio-political structures and cultural contexts. For example, when asked about administrative procedures like obtaining a passport in a particular country, models tend to default to well-documented cases rather than providing precise, localized information. Similarly, cultural narratives—such as folklore, literature, and traditions—are often missing from instruction datasets (Conover et al., 2023), limiting the models' ability to generate contextually appropriate responses. While prior work relied on translation (Sengupta et al., 2023) or template-based techniques (Cahyawijaya et al., 2024) to build instruction-tuning datasets, it does not fully reflect the actual local context, as direct translations often fail to capture the nuances of regional governance, customs, and linguistic variations.

Building instruction datasets from scratch is costly, making large-scale manual data collection impractical for many low-resource languages. To address this, we adopt an LLM-assisted dataset generation approach (Liu et al., 2022; Cahyawijaya et al., 2023; Zhang et al., 2024), followed by full human validation. Specifically, we use a single-prompt method where LLMs process high-quality unlabeled text from public government and cultural sources to extract both factual information and corresponding instructions. These domains are highly relevant for real-world applications, but remain underexplored for instruction-tuning, particularly in the context of government data.

To demonstrate the effectiveness of this approach, we introduce an instruction-tuning dataset for Kazakh that integrates both institutional[2] (`GovSet`) and cultural (`CultSet`) domains. We choose Kazakh as our case study because it re-

[*] These authors contributed equally.

[1] https://huggingface.co/datasets/nurkhan5l/kazakh-ift

[2] Our study incorporates administrative, procedural, legal, structural, and other government-related types of information.

Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 14509–14538

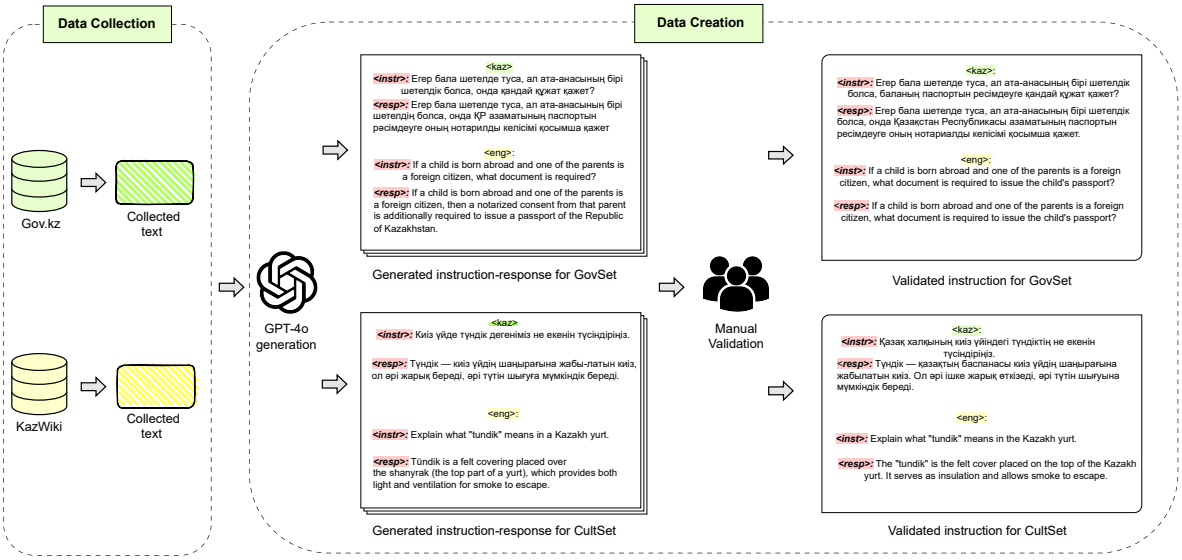


Figure 1: Overview of the end-to-end process for constructing GovSet and CultSet datasets. English translations are for illustration purposes.

mains underrepresented in NLP (Joshi et al., 2020), despite having approximately 20 million speakers. Prior research on Kazakh NLP has primarily focused on classic tasks such as named entity recognition (Yeshpanov et al., 2022) and sentiment analysis (Yeshpanov and Varol, 2024), leaving more advanced applications like instruction tuning, largely unexplored.

Our contributions are as follows:

- We create an open-source, high-quality, manually verified large scale (10K samples) IFT dataset, which covers both cultural, and institutional knowledge, relevant to Kazakhstan.
- We contribute new domain knowledge on essential institutional topics, including procedural, legal, structural, and other key aspects of public governance, enhancing LLMs' understanding of these critical areas.
- We compare the efficacy of open-weight and closed-weight LLMs in LLM-assisted dataset construction for low-resource languages and underrepresented cultures.
- We demonstrate that fine-tuning on our dataset results in consistent improvements in both multiple-choice and generative tasks. These results highlight the impact of incorporating localized knowledge into instruction tuning and demonstrate the potential of LLM-assisted approaches for expanding instruction datasets in other low-resource languages.

2 Related Work

2.1 Instruction-Tuning Datasets in English

There are three main strategies for creating English instruction-tuning datasets: human-curated datasets, templatized NLP tasks, and synthetic data generation using LLMs.

Human-curated datasets, such as Open Assistant (Köpf et al., 2023) and Dolly (Conover et al., 2023), rely heavily on human annotation. While this approach ensures high-quality data, it is expensive and difficult to scale across multiple languages. To reduce costs, datasets like the Public Pool of Prompts (P3) (Sanh et al., 2021), SuperNatural Instructions (Wang et al., 2022), and FLAN (Longpre et al., 2023) reformat existing NLP tasks into instruction-based formats. However, these datasets primarily focus on specific NLU tasks rather than general-purpose instruction following, limiting their applicability.

Prior work on instruction dataset creation has largely relied on generating data from existing language models without incorporating external real-world knowledge. Self-Instruct (Wang et al., 2023) expands an initial set of human-written instructions by iteratively generating new tasks using the model's own outputs, while Honovich et al. (2023) create instruction-tuning datasets by conditioning on a few example instructions. These methods rely solely on sampling data from language models or predefined topics, rather than grounding them in external knowledge, making them less suitable for

capturing domain-specific or culturally relevant understanding. Unlike these approaches, our work focuses on Kazakh, a low-resource language, and constructs an instruction dataset by leveraging external, factual sources such as governmental and cultural texts, ensuring alignment with real-world contexts.

2.2 Instruction Tuning Datasets for Medium- to Low-Resource Languages

While human-curated datasets are often expensive and require native speakers, prior work has explored automatic dataset generation using machine translation for low- to medium-resource languages. Sengupta et al. (2023) applied this approach to develop JAIS, an Arabic-centric language model, by translating various English instruction datasets into Arabic. Similarly, Li et al. (2023) translated Alpaca (Taori et al., 2023) into 52 languages. More recently, Alyafeai et al. (2024) introduced an Arabic instruction dataset by translating Alpaca into Arabic and then performing manual edits and localization to ensure relevance to the Arabic context. Although this method is scalable, the quality of translation remains inconsistent for low-resource languages, as noted by Li et al. (2023). Additionally, machine-translated datasets often introduce Anglocentric biases, limiting their ability to capture culturally diverse perspectives.

Several frameworks have been proposed to improve instruction tuning for low-resource languages. MURI (Köksal et al., 2024) generates multilingual instruction datasets using reverse instruction generation and translation, but none of its data have been validated by native speakers. Li et al. (2024) improve upon translation-based methods by prompting LLMs in English while requiring responses in low-resource languages (e.g., Urdu), allowing models to leverage their internal knowledge of the target language's local context. However, our approach differs by grounding responses in factual information from reliable sources, such as government data, rather than relying solely on the model's internal knowledge. Additionally, our dataset undergoes full human validation to ensure accuracy and relevance. Meanwhile, Cahyawijaya et al. (2023) focused on the linguistic aspects of instruction tuning by denoising low-resource language text and prompting models to reconstruct complete sentences. While this enhances fluency, it differs from our approach, which prioritizes grounding instructions in externally verified, domain-specific knowledge rather than refining linguistic quality alone.

2.3 Existing datasets in Kazakh

While there has been significant progress in developing Kazakh datasets, the majority of high-quality Kazakh datasets are related to speech (Mussakhojayeva et al., 2024, 2022, 2021). In terms of textual data, existing resources primarily focus on question answering and reading comprehension rather than instruction tuning. For instance, KazQAD (Yeshpanov et al., 2024) is a Kazakh open-domain question answering (ODQA) dataset that can be used in both reading comprehension and full ODQA settings, as well as for information retrieval experiments. Similarly, Belebele (Bandarkar et al., 2024) is another dataset that, while useful for multilingual machine reading comprehension, is not explicitly designed for instruction tuning. Belebele covers 122 languages, including Kazakh, and comprises 900 multiple-choice questions associated with 488 distinct passages from the Flores-200 dataset.

Despite progress in Kazakh NLP resources, no existing instruction-tuning dataset incorporates cultural or domain-specific knowledge, focuses on real-world applications, and undergoes full human validation. This limits the ability of LLMs to process Kazakh-language instructions effectively in practical and locally relevant contexts, which we aim to address in this work.

3 Background

Kazakh Cultural Heritage. Kazakhstan has a rich cultural heritage that reflects a blend of nomadic traditions, Soviet influences, and modern developments. The country's nomadic heritage is evident in many aspects of daily life, from its architecture, with *yurts* (traditional felt tents) still used in rural areas, to its customs of communal gatherings and feasts, such as the celebration of the Turkic New Year, *Nauryz*.

Kazakhstan's Soviet past has also left a lasting imprint on its culture. Many cities still bear the architectural marks of Soviet planning, while the era also shaped the country's education and scientific institutions, fostering a strong tradition in mathematics and engineering—most notably reflected in the *Baikonur Cosmodrome*, the world's first and largest space launch facility.

Alongside all of this, modern developments have

transformed Kazakhstan. The capital, Astana, is a prime example of this shift, with its futuristic skyline and ambitious urban projects. Investments in technology, renewable energy, and digital infrastructure have propelled Kazakhstan onto the global stage, while cultural revitalization efforts have fostered a renewed interest in the Kazakh language, music, and art.

At its core, Kazakhstan's culture is shaped by beliefs and values, social practices, language, artistic expression, and material culture. Understanding these components is crucial for ensuring accurate and meaningful representation.

Kazakhstan's Institutional Structure and Public Governance. Kazakhstan is a presidential republic that has prioritized modernization since its independence in 1991, particularly in governance and legal systems. The 1995 Constitution established the legal foundation, defining citizens' rights and the structure of government. A major step in this modernization has been the digitalization of public services. Kazakhstan ranks among the top 25 countries in the UN E-Government Development Index (EGDI) (Nations, n.d.), with the *eGov* platform serving as a centralized portal for services like business registration, tax payments, and social benefits.

These efforts reflect a broader national context shaped by Kazakhstan's cultural heritage, nomadic traditions, and growing digital infrastructure. Platforms like *eGov* highlight the integration of technology into daily governance. As the country continues to modernize, it is essential that language models accurately represent these unique characteristics to support cultural understanding and global relevance.

4 Data

4.1 Document Source

GovSet We manually collected 1,376 texts from the official Kazakhstan e-Government portal (gov.kz[3]), the primary and most comprehensive platform for all public services, governmental processes, and administrative resources in the country. As the central hub for Kazakhstan's digital governance, gov.kz consolidates a wide range of essential information into a single system, covering diverse aspects of public administration, legal frameworks, citizen services, and governmental initiatives. By incorporating these texts, we ensure that the dataset captures essential institutional aspects of life in Kazakhstan, including its governmental structure and public services. This enrichment enhances instruction-tuning applications, making them more linguistically appropriate and contextually informed.

CultSet We automatically collected 4,400 texts from Kazakh Wikipedia,[4] specifically focusing on pages related to Kazakh culture. These pages were identified based on metadata that explicitly indicated their relevance to Kazakh cultural topics. The parsed texts include various aspects of Kazakh traditions, heritage, arts, and historical practices, providing a rich source of culturally relevant content. This ensures that the dataset reflects the depth and diversity of Kazakh culture, making it suitable for instruction-tuning tasks that require a culturally grounded perspective.

4.2 LLM-assisted Data Generation

We benchmark one open-weight LLM: LLaMA 3.1-70B (Touvron et al., 2023), and three closed-weight LLMs: GPT-4o (OpenAI, 2024), Gemini-1.5 (DeepMind, 2024), and Claude-3.5-Sonnet (Anthropic, 2024), to assess their effectiveness in assisting dataset creation. These models were selected based on their strong performance in multilingual benchmarks. However, their capability in generating instruction datasets specific to Kazakh government and cultural data remains uncertain.

We design a prompt (see Appendix A.3) that instructs LLMs to first extract factual information from a given Kazakh document and then generate an instruction dataset based on the extracted content. Table 2 provides detailed statistics on the source documents and the resulting instruction fine-tuning (IFT) dataset using GPT-4o. Specifically, we use 4,400 Kazakh cultural Wikipedia documents and 1,376 Kazakh government data sources, generating a total of 10,600 IFT instances. Of these, 58% belong to the government public data category (GovSet), while the remaining samples are derived from Wikipedia (CultSet). Examples of generated IFT data can be found in Table 15 and Table 17.

Human Evaluation Across LLMs For each LLM, we sampled 100 generated IFT instances, drawn from 25 randomly selected GovSet and 25 CultSet documents. Additionally, we randomly

[3] https://www.gov.kz

[4] kk.wikipedia.org

sampled 100 instances from MURI (Köksal et al., 2024), which also includes Kazakh IFT data, to provide a comparative quality assessment. Two native Kazakh speakers were recruited to manually evaluate the generated data based on the following criteria:

- **Correctness**: The factual accuracy and alignment with the original text. A high score indicates that the generated pair adheres closely to the source material without introducing errors or inaccuracies.
- **Fluency**: The grammatical and stylistic quality of the generated text. A higher score reflects well-structured, natural, and polished language.
- **Completeness**: The degree to which the instruction-response pair is clear, contextually grounded, and free from ambiguity. High scores indicate that the pair is fully self-contained, with enough context to make it understandable.

All criteria were rated on a Likert scale from 1 to 5, with 5 representing the highest quality. A detailed evaluation rubric is provided in Table 8.

Table 1 presents the quality assessment of various LLMs in generating IFT data for Kazakh. The inter-annotator agreement, measured using Pearson correlation, is high (ranging from 0.68 to 0.70) across correctness, completeness, and fluency, indicating strong reliability in the evaluation process (see Appendix K.1 for further details).

Among the evaluated models, GPT-4o achieved the highest performance across all three criteria. In contrast, LLaMA-3.1 (70B) lagged significantly, scoring nearly 0.8–1 point lower in all aspects. Notably, MURI's quality was lower than GPT-4o despite both relying on OpenAI models. This discrepancy is likely due to MURI's reliance on machine translation, where Kazakh text is first translated into English before generating instructions, followed by a final back-translation into Kazakh. This multi-step translation process can introduce errors due to cumulative translation inaccuracies. Additionally, MURI is entirely LLM-generated without human validation, further affecting its quality.

4.3 Manual Post-Editing

Given GPT-4o's strong performance, we use it for large-scale IFT data generation while ensuring quality through full human verification. We employ 12 expert annotators, all native Kazakh speakers with advanced degrees in World Languages, Literature, or Political Science from top Kazakhstani universities. Their extensive experience—having lived in Kazakhstan for over 25 years—equips them with the necessary linguistic and cultural expertise.

Model	Correctness	Completeness	Fluency
Llama 3.1 (70B)	3.54	3.45	3.07
Claude	3.74	3.48	3.09
Gemini 1.5	3.85	3.64	3.32
GPT-4o	**4.38**	**4.29**	**4.04**
MURI	3.87	3.52	3.41

Table 1: Human evaluation on LLM-generated instruction datasets.

	CultSet	GovSet
Collected text	4,400	1,376
Avg. lengths (#char) of collected text	245	179
Generated IFT pairs	4,400	6,200
Avg. lengths (#char) of instruction	85	76
Avg. length (#) of output	453	215
# of unique tokens	62,449	24,304

Table 2: Statistics of GPT-4o generated IFT dataset.

To maintain consistency, annotators received detailed guidelines outlining task objectives, evaluation criteria, and examples of high-quality IFT pairs (see Appendix F). They were responsible for manually reviewing and correcting errors in the generated data. Before starting the main annotation process, all candidates completed a pilot task to assess their understanding of project requirements and their ability to refine IFT pairs accurately. Only those who met the evaluation criteria were selected. Each annotator's workload was equivalent to five full working days, and they were compensated fairly based on Kazakhstan's monthly minimum wage. To accommodate flexibility, annotators were given up to one month to complete the task while working part-time.

Table 3 summarizes the error types identified during manual post-editing of GPT-4o-generated data across the two document sources. Annotators found that `CultSet` had a higher proportion of "No error" cases (28.32%) compared to `GovSet` (19.47%), suggesting variations in data quality.

Structural errors were the most common in both datasets, accounting for over 28% in `CultSet` and 33% in `GovSet`. These errors involve grammatically correct but poorly structured responses, including issues with logical flow, organization, and unnatural phrasing for a Kazakh speaker. Addition-

Error Type	% of Questions	
	CultSet	GovSet
No error	28.32%	19.47%
Wrong language	0.07%	0.14%
Structural	**28.45%**	**33.58%**
Grammatical	25.24%	28.73%
Lexical	17.92%	18.08%

Table 3: Distribution of error types in GPT-4o-generated IFT data from CultSet and GovSet, identified during manual post-editing.

ally, grammatical and lexical errors were frequently observed, with annotators noting that GPT-4o occasionally replaces Kazakh words with Russian equivalents, even when the correct Kazakh term is explicitly provided in the original text. For a detailed breakdown of annotator observations, see Appendix H.

4.4 Final Data Overview

As shown in Table 2, the final dataset consists of 4,400 CultSet and 6,200 GovSet IFT instances, totaling 10,600 high-quality samples. We split the dataset into 90% training and 10% test, where the training data is used for full fine-tuning of LLMs, and the test set is used for generation evaluation in our experiments.

Since both CultSet and GovSet are topic-based, we include their respective topics as metadata in the final IFT dataset (see Table 13 and Table 14 for topic definitions). Figure 2 illustrates the topic distribution of the dataset. The most common topics in CultSet include Kazakh literature, traditions, and media, while GovSet primarily covers legal assistance, the healthcare system, real estate laws, and education in Kazakhstan. Examples of GPT-4o-generated IFT data can be found in Table 15 and Table 17.

Table 2 further highlights a notable difference between the two subsets: the average output length in CultSet is significantly longer and includes more unique tokens than GovSet. This difference stems from the nature of GovSet responses, which are strictly factual and concise, whereas CultSet responses tend to be more diverse and expressive.

5 Experiments

We conducted two experiments: multiple-choice questions (MCQ) and text generation evaluation. We will detail each evaluation in the following sections.

Model Selection For both MCQ and generation evaluations, we use three instruction-tuned models: Gemma-2-9b-instruct (Gemma) (Team et al., 2024), Qwen-2.5-7b-instruct (Qwen) (Qwen et al., 2025), and Falcon-3-10b-instruct (Falcon) (Team, 2024). While these LLMs offer multilingual capabilities, none were specifically trained for Kazakh, allowing us to assume that our IFT data is novel to them.

Fine-tuning We performed full fine-tuning on Gemma-2-9b-instruct (Gemma), Qwen-2.5-7b-instruct (Qwen), and Falcon-3-10b-instruct (Falcon) using the AdamW optimizer with hyperparameters $\beta_1 = 0.9$, $\beta_2 = 0.95$, $\epsilon = 1e{-}5$, and a weight decay of 0.1. We scaled the gradient norms using a maximum norm clipping value of 1.0. The learning rate was kept constant throughout the fine-tuning without any warm-up or decay with a value of $1e{-}6$ for Gemma and Falcon, and $1e{-}5$ for Qwen. The batch size used was 16, and we packed multiple documents until the maximum sequence length was 8,192 tokens. Cross-document attention is disabled by modifying attention masks so the tokens of a document only attend to the tokens from the same document in a causal way. No adjustments were made to the original tokenizer for each model.

Baseline As a baseline, we include the Kazakh Alpaca dataset,[5] which has been translated and localized into Kazakh. For each model, we conduct full fine-tuning with (1) our training dataset, (2) Alpaca, and (3) a combination of Alpaca and our training dataset.

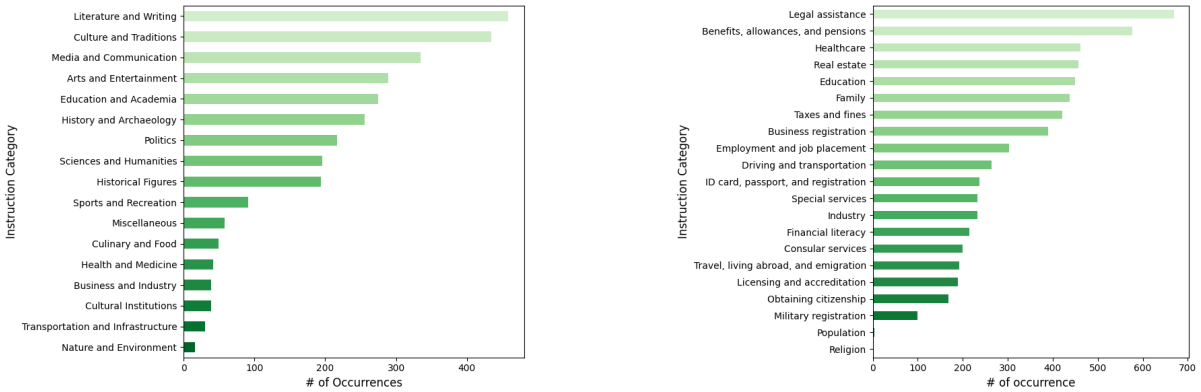
5.1 Multiple-choice Question Evaluation

Dataset A dedicated open-source Kazakh NLP community[6] has collaboratively developed and crowd-sourced multiple hand-crafted benchmarks to assess the factual knowledge of LLMs in Kazakh. We use three multiple-choice question (MCQ) datasets: (1) Dastur-MC (Sagyndyk et al., 2024b), which evaluates knowledge of Kazakh traditions, (2) Kazakh Constitution-MC (Sagyndyk et al., 2024a), which focuses on Kazakhstan's legal system, and (3) Kazakh Unified National (Sagyndyk et al., 2024c), which assesses citizen's rights, legal protections, and societal knowledge (referred to as the "Human Rights and Society" dataset).[7]

[5] https://huggingface.co/datasets/AmanMussa/kazakh-instruction-v2

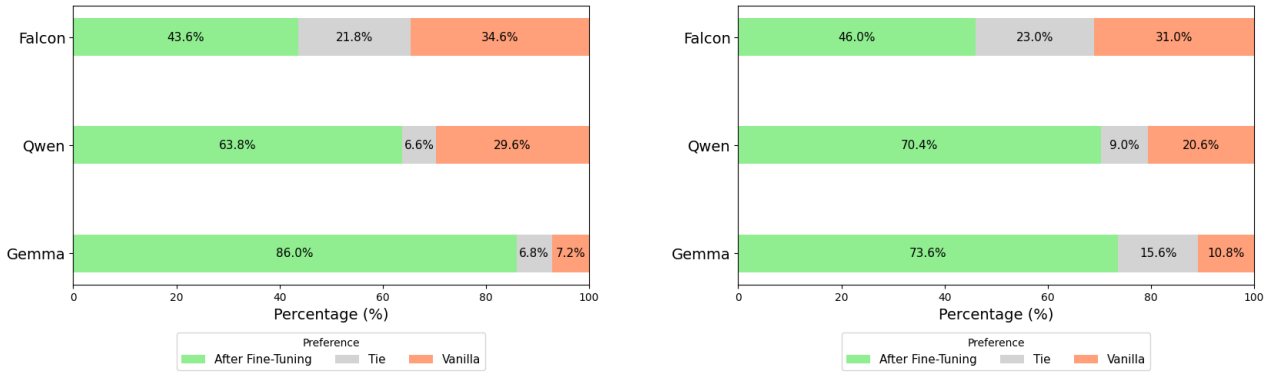
[6] https://huggingface.co/kz-transformers

[7] Examples of test questions are provided in Appendix L.



(a) CultSet (b) GovSet

Figure 2: Topic distribution of GPT-4 generated IFT dataset in `CultSet` and `GovSet`.



(a) CultSet (b) GovSet

Figure 3: Distribution of preferences for (a) `CultSet` and (b) `GovSet` datasets across models. The charts illustrate the percentage of 'Tie', 'Vanilla', and 'After Fine-Tuning' preferences in each dataset.

Model	Vanilla	RAG	Alpaca	Ours	Alpaca + Ours
Dastur					
Gemma	0.498	0.533	0.513	0.543	**0.566**
Qwen	0.403	0.410	0.421	0.443	**0.465**
Falcon	0.450	0.460	0.458	0.464	**0.471**
Constitution					
Gemma	0.600	**0.655**	0.627	0.640	0.650
Qwen	0.520	0.523	0.609	0.670	**0.680**
Falcon	0.430	0.386	0.450	0.490	**0.520**
Human Rights and Society					
Gemma	0.405	0.450	0.430	0.465	**0.480**
Qwen	0.300	0.325	0.330	0.365	**0.375**
Falcon	0.215	0.220	0.234	0.250	**0.275**

Table 4: Zero-shot accuracies of language models in different datasets: (1) Dastur, (2) Constitution, and (3) Human Rights and Society

Each dataset consists of multiple-choice questions with four answer options, only one of which is correct. We selected these evaluation benchmarks because they align with the focus of our instruction fine-tuning dataset and are not derived from our document sources (`CultSet` and `GovSet`). These datasets cover culturally significant topics, legal frameworks, and citizen-government interactions, reflecting real-world applications that our fine-tuned models aim to support.

Since no documented quality assurance process was available for the three datasets, we conducted a manual verification to ensure the accuracy of the questions. To maintain a fair and valid comparison, only the manually verified samples were used in our evaluation. For the Dastur-MC dataset, we randomly sampled 300 questions and manually verified their correctness. The same process was applied to the Kazakh Constitution-MC and Human Rights and Society datasets, with 200 randomly selected questions from each.

Setup In addition to the fine-tuned models, we include retrieval-augmented generation (RAG) without fine-tuning to estimate the upper bound of the

	CultSet		GovSet	
	Vanilla	**After FT**	**Vanilla**	**After FT**
Gemma	15.76	**24.87**	16.12	**25.10**
Falcon	25.96	**27.98**	26.17	**28.70**
Qwen	**27.64**	26.63	**30.27**	28.42

Table 5: ROUGE-L comparison on `CultSet` and `GovSet` before and after fine-tuning.

original models' performance. For RAG, we use BM25 encoding, as no specialized Kazakh retrieval encoder is available. For each question, we retrieve the top two matching text chunks (each 256 symbols long) from the training texts of our IFT corpus and provide them as additional context.

To assess the capabilities of the model, we use the `LM Eval Harness` (Gao et al., 2024) framework in a zero-shot setting. During evaluation, the answer is selected based on the alphabetical option with the highest likelihood.

Result Table 4 presents the zero-shot evaluation results across different models and techniques. Overall, our fine-tuned dataset consistently outperforms other approaches across datasets and models. The only exception is the Constitution dataset, where RAG performs better with Gemma. Models fine-tuned on Kazakh Alpaca show some improvement, though it remains lower than that achieved with our instruction fine-tuning (IFT) dataset.

Combining parts of our IFT dataset with the translated Alpaca dataset yields the highest performance gains. This aligns with prior studies (Brief et al., 2024; Wang et al., 2024), which suggest that incorporating general chat instructions alongside domain-specific ones enhances model performance.

For RAG-enhanced models, performance generally exceeds that of the vanilla models, except for Falcon on the Constitution dataset. However, fine-tuned models consistently achieve higher scores than their RAG-enhanced counterparts. We hypothesize that this is due to the models' limited proficiency in Kazakh, which may hinder their ability to fully understand the retrieved context. As a result, despite the additional information provided by RAG, the models may struggle to extract the necessary details to select the correct answer in MCQs.

5.2 Generation Evaluation

We evaluate generation performance using our test set, which consists of 500 questions from both `CultSet` and `GovSet` (excluded from fine-tuning). We compare the best-performing models from Section 5.1 against their vanilla counterparts. In this section, "After Fine-Tuning" refers to models fine-tuned on Alpaca + Our Data, while "Vanilla" refers to the original instruct models.

Automatic Evaluation with ROUGE and BERTScore As shown in Table 5, fine-tuned models generally outperform their vanilla counterparts, except for Qwen, where fine-tuning results in a lower ROUGE-L score (Lin, 2004). However, a lower ROUGE-L does not necessarily indicate worse performance—it may be due to Qwen generating different phrasings compared to the gold answers.

To further validate the quality of generated responses, we also evaluate BERTScore (Zhang et al.). We use Kaz-RoBERTa[8] as the encoder model, as it is one of the few open-source Kazakh-language transformers. The BERTScore results in Table 6 align well with the ROUGE-L scores. However, since Kazakh is a low-resource language, BERTScore should be considered a reference point rather than a definitive metric, as Kaz-RoBERTa embeddings may not perfectly capture synonym relationships.

Preference Evaluation with GPT-4o We conducted a 1-to-1 preference evaluation using the LLM-as-a-judge approach. Specifically, we prompted GPT-4o to compare responses from different models and determine whether each response wins, loses, or ties. The prompt includes the instruction and the gold response as context for GPT-4o.[9] As shown in Figure 3, the results align with ROUGE-L and BERTScore, confirming that fine-tuned models generally produce improved outputs. Compared to Falcon, Qwen and Gemma exhibit more significant improvements (63%–80% winning rate), likely because their pre-trained versions were less optimized for the task, making fine-tuning more impactful.

Additionally, we analyze the win rate across topics in `CultSet` and `GovSet`, as shown in Appendix B. The results indicate that the impact of fine-tuning varies by topic and is not always consistent. In `CultSet`, fine-tuning Qwen with our IFT data yields the most improvement in Cultural Institutions and Culture & Traditions, while the

[8] Huggingface model: kaz-roberta-conversational

[9] The prompt used for comparison is provided in Appendix A.2.

		CultSet			GovSet		
		Precision	**Recall**	**F1**	**Precision**	**Recall**	**F1**
Vanilla	Gemma	29.26	33.47	30.92	27.36	34.81	30.39
	Falcon	23.29	28.17	25.20	20.38	24.68	22.11
	Qwen	40.58	47.46	43.40	36.57	44.14	39.50
After Fine-Tuning	Gemma	41.94	46.36	43.62	40.27	44.90	42.00
	Falcon	24.59	29.68	26.64	23.78	27.73	25.36
	Qwen	39.64	45.40	41.82	36.28	40.20	37.59

Table 6: BERTScore Precision, Recall, and F1 for `CultSet` and `GovSet`.

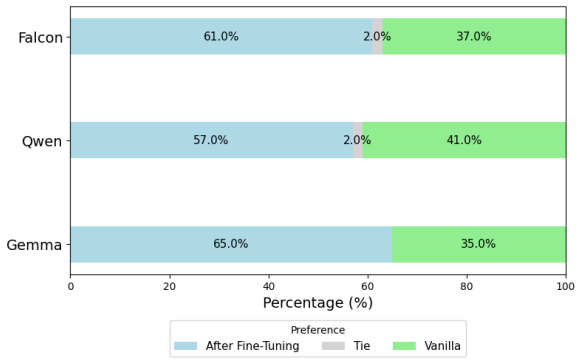


Figure 4: Conversational data preference evaluation.

gains are smaller in Science & Humanities and even lead to a decline in performance for Education & Academia. In `GovSet`, fine-tuning Qwen with our dataset significantly enhances performance in Legal Assistance, though the improvement is less noticeable in Employment-related topics.

While LLM-based evaluations provide scalable comparisons, they may not fully capture human judgment nuances, making human evaluation essential for validating model preferences. Therefore, three human annotators conducted a preference evaluation on a randomly sampled 100 examples for each model (Gemma, Qwen, and Falcon) across both `CultSet` and `GovSet`. Their judgments were compared against the GPT-based preference evaluation **to assess alignment**. We computed Cohen's Kappa between GPT-4o and the annotators, obtaining 0.63 for `CultSet` and 0.68 for `GovSet`, indicating substantial agreement. We have also calculated the agreement rate between annotators (detailed in Appendix K.2). The results show that GPT's alignment with human preferences is moderate, with better agreement on `GovSet` than `CultSet`.

Conversational Evaluation. As an extension of these experiments, we generated a set of 100 conversations for both `CultSet` and `GovSet` combined, covering topics presented in Figure 2. These conversations were intentionally left unfinished using a special prompt, as detailed in Appendix A.2. Both the original and fine-tuned models were tasked with generating the most appropriate continuation for each conversation. Examples of the resulting texts are shown in Appendix M. To evaluate the quality of the responses, we employed an LLM-as-a-judge framework. The results, presented in Figure 4, indicate that models fine-tuned on domain-specific data produced significantly more coherent and contextually appropriate responses compared to their pre-fine-tuning counterparts. We also see that in the conversational settings there are less ties, compared to simple question answering.

6 Conclusion and Future Work

We introduced a culturally and institutionally aligned instruction-tuning dataset for Kazakh, aiming to enhance practical knowledge representation and address the specific needs of public governmental data processing in Kazakh. Through a carefully designed data collection pipeline, we generated instruction-tuning examples using GPT-4o and ensured their quality via a manual correction and localization to capture Kazakh linguistic and cultural nuances accurately.

The evaluation results show that this approach substantially improved the model's factual knowledge and the understanding of low-resource languages. It also shows that after such fine-tuning, the model's responses are much better in terms of correctness and soundness, as assessed by native speakers and LLM as a judge.

In future work, we plan to apply this methodology to other languages and dialects. We further aim to work towards streamlining and automating the process as much as possible. We will also focus more on the modeling part of the experiments, and open-source culturally and institutionally relevant models for low-resource languages, including Kazakh.

7 Limitations

We aim to establish a robust instruction-tuning dataset for Kazakh, authentically reflecting the cultural and linguistic richness of the language. Unlike many existing datasets, which rely on translated resources or machine-generated responses, our dataset is entirely crafted from Kazakh-specific content, ensuring greater alignment with the cultural values and linguistic nuances of the region. However, we recognize several limitations in our work:

- **Cultural Representation:** The dataset emphasizes topics deeply rooted in Kazakh culture, traditions, and societal norms, ensuring relevance and cultural authenticity. However, certain culturally sensitive topics, such as those involving religious matters, were intentionally omitted to avoid controversy and maintain neutrality.
- **Language Variations:** Kazakh is a rich language with significant regional variations in vocabulary and usage. While our dataset primarily focuses on standard Kazakh, it does not explicitly account for regional dialects or variations, potentially limiting its applicability to speakers outside the standard dialect's scope.
- **Modeling Limitations:** Our works is a proof of concept, and it was not aimed at creation of SOTA models for Kazakh. That is why we experiment with smaller models and do not apply any training tricks such as tokenizer adaptaion for Kazakh.
- **Possible Data Drift:** We also acknowledge that despite of being very conservative by nature, some institutional procedures can change over time, that is why it is possible that the data provided in our IFT dataset will get less actual. To handle this issue we are planning updating the datasets annually.

8 Ethics

We adhered to the internal policies of web resources while scraping data and included only publicly available information verified by authorities.

While our method enhances LLMs' understanding of Kazakhstan's institutional nuances, users should not blindly trust generated responses. LLM outputs serve as a starting point, and users remain responsible for fact-checking due to potential hallucinations.

All human subjects in our study provided informed consent, were fully aware of the study's objectives, and had the right to withdraw at any time. They were also appropriately compensated as part of their job.

References

Zaid Alyafeai, Khalid Almubarak, Ahmed Ashraf, Deema Alnuhait, Saied Alshahrani, Gubran Abdulrahman, Gamil Ahmed, Qais Gawah, Zead Saleh, Mustafa Ghaleb, Yousef Ali, and Maged Al-shaibani. 2024. CIDAR: Culturally relevant instruction dataset for arabic. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 12878–12901, Bangkok, Thailand. Association for Computational Linguistics.

Anthropic. 2024. Claude ai: An overview. https://www.anthropic.com.

Lucas Bandarkar, Davis Liang, Benjamin Muller, Mikel Artetxe, Satya Narayan Shukla, Donald Husa, Naman Goyal, Abhinandan Krishnan, Luke Zettlemoyer, and Madian Khabsa. 2024. The belebele benchmark: a parallel reading comprehension dataset in 122 language variants. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 749–775, Bangkok, Thailand. Association for Computational Linguistics.

Meni Brief, Oded Ovadia, Gil Shenderovitz, Noga Ben Yoash, Rachel Lemberg, and Eitam Sheetrit. 2024. Mixing it up: The cocktail effect of multi-task fine-tuning on llm performance – a case study in finance.

Samuel Cahyawijaya, Holy Lovenia, Fajri Koto, Rifki Putri, Wawan Cenggoro, Jhonson Lee, Salsabil Akbar, Emmanuel Dave, Nuurshadieq Nuurshadieq, Muhammad Mahendra, Rr Putri, Bryan Wilie, Genta Winata, Alham Aji, Ayu Purwarianti, and Pascale Fung. 2024. Cendol: Open instruction-tuned generative large language models for Indonesian languages. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14899–14914, Bangkok, Thailand. Association for Computational Linguistics.

Samuel Cahyawijaya, Holy Lovenia, Tiezheng Yu, Willy Chung, and Pascale Fung. 2023. InstructAlign: High-and-low resource language alignment via continual crosslingual instruction tuning. In *Proceedings of the First Workshop in South East Asian Language Processing*, pages 55–78, Nusa Dua, Bali, Indonesia. Association for Computational Linguistics.

Mike Conover, Matt Hayes, Ankit Mathur, Jianwei Xie, Jun Wan, Sam Shah, Ali Ghodsi, Patrick Wendell, Matei Zaharia, and Reynold Xin. 2023. Free dolly: Introducing the world's first truly open instruction-tuned llm. Last accessed 2024-01-15.