Stereotype Bias in a Bilingual Setting: A Culturally Grounded Evaluation in Kazakhstan

Nurkhan Laiyk[♡*] **Daniil Orel**[♡*] **Ayana Mussabayeva**[♡]
Maiya Goloburda[♡] **Kamila Kuishibekova**[◇] **Liya Goloburda**[◇]
Diana Turmakhan[♡] **Preslav Nakov**[♡¶] **Yuxia Wang**[◦] **Fajri Koto**[♡]
[♡] Mohamed bin Zayed University of Artificial Intelligence [◦] INSAIT
[◇] Nazarbayev University [¶] Institute of Foundation Models

Abstract

Stereotype bias in language models has been widely examined in English, but remains largely understudied in bilingual contexts where multiple linguistic and cultural systems interact. This gap is especially important in regions where language use reflects complex historical and sociopolitical influences. In this work, we focus on Kazakhstan, a bilingual society where Kazakh, a low-resource Turkic language, and Russian, a high-resource Slavic language, are both actively used and frequently code-switched in everyday communication. We introduce `Aqbileq`[1], a high-quality, human-verified dataset consisting of 5,634 stereotype-bearing statements in Kazakh, Russian, and code-switched forms, covering six culturally salient domains. We evaluate both multilingual and Kazakh-specific language models using perplexity-based scoring and pretraining simulations, and find that stereotype bias is most pronounced in code-switched inputs. Our results highlight the limitations of existing evaluation frameworks and emphasize the need for culturally grounded, linguistically inclusive benchmarks to better assess and mitigate bias in language models. Warning: this paper contains example data that may be offensive, harmful, or biased.

1 Introduction

Language models perform strongly on many downstream NLP tasks, but they remain vulnerable to stereotyping because they are pre-trained on large-scale text corpora (Blodgett et al., 2021; Bender et al., 2021). These stereotypes often mirror widespread social beliefs that may be inaccurate and frequently carry negative connotations (Fraser et al., 2021). This remains problematic even when the stereotype appears positive, since such associations can still lead to harmful or unintended effects.

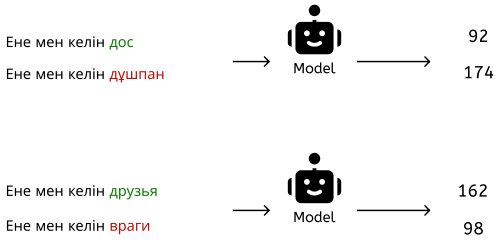


Figure 1: An example where the model assigns lower perplexity to counter-stereotypical statements, revealing bias in both Kazakh and code-switched inputs. English translation: "*The mother-in-law and daughter-in-law are friends.* / "*The mother-in-law and daughter-in-law are enemies.*"

For example, a language model might complete the prompt "*An ideal employee is...*" with "*an Asian who is hardworking and good at math*". Although this response may seem complimentary, it reinforces reductive generalizations and contributes to biased decisions in real-world settings.

More broadly, stereotypes in NLP training data can propagate through downstream tasks, potentially disadvantaging underrepresented demographic groups (Savoldi et al., 2021; Ziems et al., 2022). To address this, substantial efforts have been made in English, resulting in benchmark datasets such as CrowS-Pairs (Nangia et al., 2020), StereoSet (Nadeem et al., 2021), and WinoBias (Zhao et al., 2018). However, stereotype bias is not universal; it is shaped by cultural and linguistic context, underscoring the importance of developing datasets across diverse languages and regions. This includes examining bias in code-switched settings, where speakers alternate between two or more languages within a single utterance or conversation (Barman et al., 2014). To the best of our knowledge, this phenomenon remains underexplored, despite its prevalence in many multilingual regions.

[*] Equal contribution.
[1] https://huggingface.co/datasets/nurkhan51/aqbileq

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 13114–13131
July 2-7, 2026

We examine stereotype bias in Kazakhstan, a multilingual country with a population of approximately 20 million, where 73% speak Kazakh and 15% speak Russian,[2] which makes it a compelling setting to investigate how linguistic and cultural biases emerge in both monolingual and bilingual contexts (Koto et al., 2025; Laiyk et al., 2025).

Our goal is to understand how stereotype bias manifests in language models that process Kazakh, Russian, and their interactions, particularly in ways that reflect real-world usage in Kazakhstan. This study is driven by two main gaps. First, most evaluations of social bias in NLP overlook low-resource languages like Kazakh and ignore multilingual usage patterns common in Central Asia. Second, while Kazakh and Russian frequently co-occur in communication, they differ significantly in typology and resource availability (Koto et al., 2025). Existing Russian-language bias benchmarks typically reflect the cultural norms of Russia and may not align with Kazakhstan's distinct sociolinguistic landscape (Grigoreva et al., 2024). This raises the risk that language models trained on Russian data encode and reproduce inappropriate or irrelevant social stereotypes when applied in Kazakhstan's context. As illustrated by Figure 1, for the same real-world entity, a model may exhibit opposite biases depending on whether the statement is in Kazakh or Russian.

Our contributions can be summarized as follows:

- We introduce `Aqbileq`, a novel high-quality dataset for evaluating culturally grounded stereotype bias in Kazakhstan across six domains. The dataset contains 5,634 statements in Kazakh, Russian, and their code-switched form, all verified by native speakers.
- We evaluate cultural bias in Kazakh-specific language models, covering three encoder-only and six decoder-only models, using perplexity across languages and bias domains.
- We conduct a pre-training simulation of transformer-based LLMs using different mixtures of Kazakh and Russian data to examine when and how stereotype bias emerges.
- We extend our analysis to generation-based evaluation by assessing the sentiment polarity of biased entities when used to generate short stories in Kazakh.

[2] https://glottolog.org/

2 Related Work

Bias in Language Model Language models pretrained on large-scale corpora have been shown to encode various stereotype biases, such as gender, profession, race, and religion (Gallegos et al., 2024; Gupta et al., 2024; Hu et al., 2025). These biases appear not only in internal representations (Kurita et al., 2019; Srivastava et al., 2023) and generated text (Dhamala et al., 2021), but also when language models are used as evaluators in downstream tasks (Park et al., 2024).

Bias mitigation has been studied across diverse NLP tasks, including coreference resolution, machine translation, text generation, etc. In coreference, gender-balanced templates and gender-swapping reduce gender–occupation asymmetries (Zhao et al., 2018; Rudinger et al., 2018). In machine translation, WinoMT exposes a masculine default and motivates balanced challenge sets and guided decoding for faithful gender realization (Stanovsky et al., 2019). For open-ended generation, decoding-time control and self-debiasing steer models away from toxic or biased continuations without retraining (Schick et al., 2021).

To evaluate stereotype bias in language models, benchmarks use either a *question-answering (QA)* format or a *sentence-scoring format using slot-filled templates*. In the QA format, a question is paired with context and answer options reflecting stereotypical or counter-stereotypical implications (Neplenbroek et al., 2024). Examples include BBQ (Parrish et al., 2022) and its variants: CBBQ (Huang and Xiong, 2024) for Chinese, KoreanBBQ (Jin et al., 2024) for Korean, and BasqBBQ (Zulaika and Saralegi, 2025) for low-resourced Basque language. This format yields interpretable outputs, but constructing culturally appropriate and balanced choices requires effort.

By contrast, the sentence-scoring format uses neither questions nor predefined options. Instead, it compares probabilities for minimally different sentences formed by filling a template with contrasting attribute values, for example, *"Harvard student is [rich/poor]."* This makes the format scalable, since templates can be automatically instantiated across many group-attribute pairs. It is used in CrowS-Pairs (Nangia et al., 2020), WinoGender (Rudinger et al., 2018), WinoBias (Zhao et al., 2018), and SEAT (May et al., 2019). We adopt the sentence-scoring format for its scalability and ability to capture fine-grained model preferences.

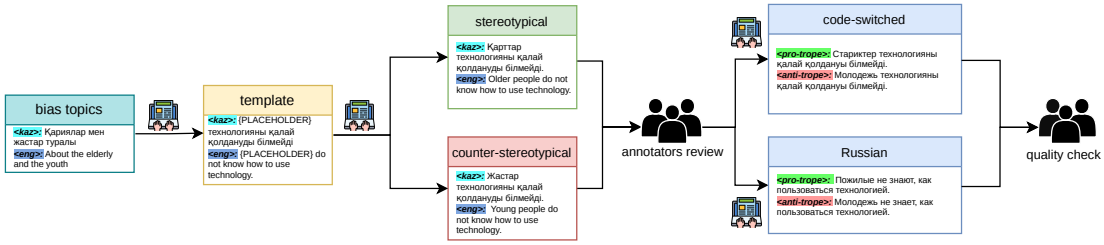


Figure 2: End-to-end process of dataset construction. indicates manual annotation.

Bias in Multilingual Settings While early research on stereotype bias in NLP focused primarily on English, recent efforts have extended evaluation to other languages. A common strategy involves translating English benchmarks such as CrowS-Pairs (Nangia et al., 2020) and BBQ (Parrish et al., 2022) into another language. For example, Zulaika and Saralegi (2025) translated BBQ into Basque, and Sahoo et al. (2024) adapted CrowS-Pairs for Hindi. In the Korean context, researchers explored both benchmark translation (Jin et al., 2024) and prompt-based probing; Lee et al. (2024) evaluated GPT-4 (OpenAI, 2023) using persona-injected prompts tailored to Korean sociocultural norms.

Other studies have focused on capturing region-specific dynamics. TWBias (Hsieh et al., 2024) targets gender and ethnic bias in Taiwanese Mandarin, whereas RuBia (Grigoreva et al., 2024) addresses bias in Russian through a crowdsourced approach that collects biased statements on Telegram[3] and conducts manual verification. Recent multilingual efforts, such as SHADES (Mitchell et al., 2025), have expanded the scope by compiling culturally specific stereotypes in a wide range of languages and regions. However, these studies do not cover Kazakh and do not consider bilingual contexts with code-switching; here we bridge this gap.

Bias evaluation in bilingual and code-switched settings remains significantly underexplored (Adelani et al., 2025), even as multilingual language models are increasingly deployed across linguistically diverse regions. These models often mirror the cultural and linguistic asymmetries of their training data, leading to a preference for dominant languages and narratives (Demidova et al., 2024). Recent work, such as the Code-Switching Red-Teaming (CSRT) benchmark (Yoo et al., 2025), has shown that large language models are particularly vulnerable to inputs that mix languages, mirroring real-world multilingual interactions.

[3] https://web.telegram.org/

However, such evaluations have largely overlooked Kazakhstan, a multilingual society in which Kazakh and Russian are not only legally recognized as co-official languages, but are also frequently used interchangeably in everyday communication. This bilingual dynamic, shaped by Soviet-era language policy, informs how speakers alternate between languages for identity construction and social signaling (Chernyavskaya and Zharkynbekova, 2024; Nakamura, 2024; Murodova, 2024). While previous studies have explored stereotype bias in Russian (Grigoreva et al., 2024) and in other Turkic languages such as Turkish (Caglidil et al., 2024), they do not capture the sociolinguistic specificity of Kazakhstan, particularly its pervasive code-switching practices. This leaves open the question of how stereotype bias is expressed in Kazakh, Russian, and mixed-language use within the same social setting; we address this question below.

3 Aqbileq

To address the lack of stereotype bias datasets tailored to the Kazakhstan context, we introduce `Aqbileq`, a culturally grounded resource comprising 5,634 stereotype-bearing statements in Kazakh, Russian, and code-switched form. The full data creation pipeline is illustrated in Figure 2. Each example in `Aqbileq` is constructed from scratch and verified by native speakers from Kazakhstan. The dataset is built from 939 manually written templates, each instantiated with two types of stereotype expressions: **stereotypical**, which reflect widely held societal assumptions, and **counter-stereotypical**, which challenge or subvert those assumptions. These pairs are generated across all three language settings, resulting in a dataset designed to support fine-grained evaluation of stereotype bias in monolingual, bilingual, and code-switched language use in authentic, everyday, and socially situated communication across Kazakhstan's multilingual communities.

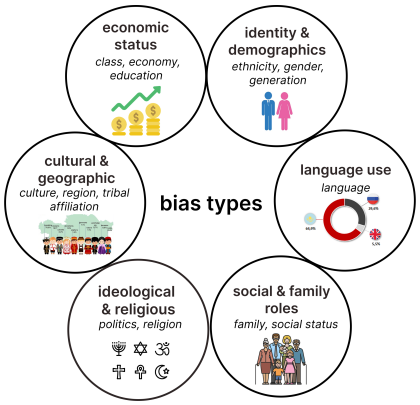


Figure 3: The six bias type domains included in the Aqbileq dataset.

3.1 Stereotype Domains

Figure 3 presents the six stereotype domains with 14 subdomains in Aqbileq, grounded in analysis by four native Kazakh speakers[4] based on recurring themes in social media, news articles and online forums, as well as prior work on bias in NLP (Gallegos et al., 2024; Gupta et al., 2024). These domains include (*i*) *cultural & geographic*, (*ii*) *identity & demographics*, (*iii*) *ideological & religious*, (*iv*) *language use*, (*v*) *social & family roles*, and (*vi*) *economic status*.

Cultural & Geographic This domain includes stereotypes based on regional identity, tribal affiliation, and rural–urban divides. In Kazakhstan, socio-territorial groups known as Zhuz (Senior, Middle, Junior) still shape public perception, employment, and social relations, especially in the South and West (Sairambay, 2019; Minbaeva and Muratbekova-Touron, 2013).

Identity & Demographics includes biases related to gender, age, and ethnicity. While some gender stereotypes are shared across post-Soviet contexts (UNDP Kazakhstan, 2024; Yerimpashaeva et al., 2023), others are Kazakhstan-specific, such as bride kidnapping and the traditional *kelin* role, where married women are expected to serve their husband's family (Werner, 2004; Turakhan, 2025).

Ideological & Religious captures stereotypes rooted in political ideology, religious beliefs, and associated social attitudes.

[4]All have over 20 years of residency in Kazakhstan.

Language Use captures stereotypes related to language preference, code-switching, and perceived fluency. In Kazakhstan's multiethnic society, language often intersects with ethnic identity, shaping access to social and economic opportunities. In particular, proficiency in Kazakh, Russian, or English can influence how individuals are perceived and treated (Jumageldinov, 2014; Zhanarstanova and Nechayeva, 2015; Orazaliyeva and Orazbayeva, 2015).

Social & Family Roles includes assumptions about one's role within the family or society, including marital expectations, parental duties, and generational norms.

Economic Status encompasses stereotypes related to wealth, occupation, social class, and access to resources.

3.2 Template Design for Stereotypical and Counter-Stereotypical Statements

Based on the 14 subdomains, four native Kazakh speakers manually created 1,107 Kazakh templates, each containing placeholders for generating stereotypical (stereotype-reinforcing) and counter-stereotypical (stereotype-neutralizing or countering) statements. For example, in religion domain, we used the template " [PLACEHOLDER] бәрі ерке және бұзық." ("Only [PLACEHOLDER] capricious and mischievous"). To generate contrastive pairs, we compiled a list of semantically compatible slot fillers such as "Үйдегі кішкентайлардың" (*children* in English) and "Ағалардың" (*adults* in English). We kept the template wording fixed and varied only the slot filler to ensure symmetry, following the design of Grigoreva et al. (2024).

3.3 Quality Control

The statement pairs were initially written by a single author. To verify that they captured culturally grounded social biases, all *counter-stereotypical* pairs were validated by seven native Kazakh speakers. The annotators made binary judgments on whether each pair reflected a recognizable stereotype (see Appendix C); the annotation guidelines are in Appendix D. We retained pairs when at least five of the seven annotators agreed they reflected local bias. This yielded 939 bias-relevant pairs. The inter-annotator agreement, measured in terms of Cohen's Kappa, exceeded 0.8 for all pairs of annotators (see Appendix E).

Domain	Subdomain	stereotypical			counter-stereotypical		
		KZ	CS	RU	KZ	CS	RU
Identity and Demographics	ethnicity	193	193	193	193	193	193
	gender	190	190	190	190	190	190
	generation	52	52	52	52	52	52
Economic Status	class	157	157	157	157	157	157
	economy	3	3	3	3	3	3
	education	21	21	21	21	21	21
Cultural and Geographic	culture	42	42	42	42	42	42
	regional	123	123	123	123	123	123
	tribal affiliation	12	12	12	12	12	12
Social and Family Roles	family	34	34	34	34	34	34
	social status	19	19	19	19	19	19
Ideological and Religious	politics	22	22	22	22	22	22
	religion	31	31	31	31	31	31
Language Use	language	40	40	40	40	40	40
Dataset size		939	939	939	939	939	939
Total data size		5634					

Table 1: Statistics on template counts, domain distribution, and dataset size by language variant. KZ, CS, and RU refer to Kazakh, code-switched, and Russian.

3.4 Code-switching and Russian Variants

With the goal of evaluating social bias in a bilingual setting, we extended the finalized dataset by creating both the code-switched and Russian version.

Code-switched Data Two native Kazakh speakers fluent in Russian manually translated the original Kazakh statements into code-switched Kazakh–Russian, preserving their meaning and tone (see Appendix D). This process maintained a one-to-one correspondence between the original and code-switched versions. To ensure consistency, accuracy, and naturalness, a third native speaker independently reviewed all code-switched statements.

Translation to Russian As an initial low-cost and scalable step, we used Google Translate to translate all Kazakh statements into Russian. However, machine translations are inadequate for culturally specific or idiomatic expressions, we asked two bilingual Kazakh–Russian annotators to review and edit all translations for accuracy, fluency, and cultural appropriateness. The annotators also documented common translation errors, with a focus on lexical, grammatical, and structural issues. The annotator comments and representative examples are presented in Table 4.

Labor Regulations Each annotator's workload was approximately equivalent to five full working days. Annotators were compensated fairly based on Kazakhstan's monthly minimum wage. To support flexibility, they were given up to one month to complete the task on a part-time basis (See annotation details in Appendix D).

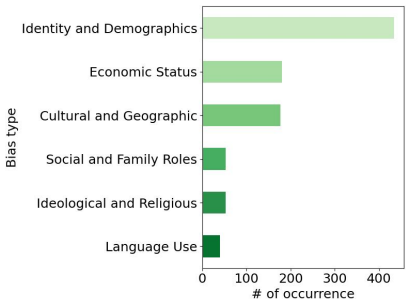


Figure 4: Domain distribution of the `Aqbileq` dataset.

3.5 Final Data Overview

We created 939 stereotypical and counter-stereotypical pairs in Kazakh, totaling 1,878 statements. With Russian and Kazakh–Russian code-switched versions, the full dataset includes 5,634 statements across three language variants (see Table 1). Each pair is labeled with one of six bias domains shown in Figure 4. Identity-related bias is the most common, followed by economic, cultural, and geographic bias. Biases related to language, ideology, and family roles are less common, reflecting the social priorities of the Kazakh context.

4 Experiments

4.1 Perplexity-based Experiments

Given a domain D and subdomain S, we calculate the bias scores S_D and S_S accordingly. The subdomain score S_S is a *stereotypical win rate*, defined as the proportion of cases where the model assigns lower perplexity (higher likelihood) to the stereotypical statement x_i^{pro} than to its corresponding counter-stereotypical statement x_i^{anti}:

$$S_S = \frac{\sum_{i=1}^{N_S} \mathbb{I}\left[\text{PPL}(x_i^{\text{pro}}) < \text{PPL}(x_i^{\text{anti}})\right]}{N_S},$$

where perplexity is PPL, $\mathbb{I}[\cdot]$ is the indicator function, N_S is the number of statement pairs in subdomain S. Domain-level bias score (S_D) is computed as the average of S_S across all subdomains.

A higher S_D (> 0.5) indicates that the model more often prefers the stereotypical statement over its counter-stereotypical counterpart, while values below 0.5 indicate a preference for the counter-stereotypical. Values near 0.5 suggest no systematic preference. We use perplexity (PPL) to evaluate causal language models and pseudo-perplexity (PPLL) (Salazar et al., 2020) for masked language models, using the LM-PPL library[5].

[5] https://github.com/asahi417/lmppl

Domain	XLM-R Base			XLM-R Large			KazRoBERTa		
	KZ	CS	RU	KZ	CS	RU	KZ	CS	RU
Cultural and Geographic	0.37	**0.50**	0.60	0.37	**0.52**	0.54	0.58	0.62	**0.54**
Identity and Demographics	0.63	0.60	**0.47**	0.62	0.57	**0.49**	0.67	0.64	**0.45**
Ideological and Religious	**0.52**	0.61	0.61	**0.55**	0.67	0.64	0.53	**0.51**	**0.51**
Language Use	0.60	**0.50**	0.60	0.60	**0.50**	0.65	**0.50**	0.58	**0.50**
Social and Family Roles	**0.57**	0.62	0.68	0.65	**0.66**	0.72	**0.50**	0.70	0.54
Economic Status	0.55	**0.53**	0.68	0.55	**0.49**	0.66	0.61	0.57	**0.44**
Average	**0.54**	0.56	0.61	**0.56**	0.57	0.62	0.57	0.60	**0.50**

Table 2: Perplexity-based bias (S_D) scores for XLM-R and KazRoBERTa across languages (KZ = Kazakh, CS = Code-switching, RU = Russian). For each model, scores closest to 0.5 are bolded to indicate minimal stereotypical preference.

Models We evaluated encoder-only and decoder-only LMs. The encoders are XLM-R Base and Large (Conneau et al., 2020) and KazRoBERTa (Sagyndyk et al., 2025). The decoders are Llama-3.1-8B, Llama-3.1-8B-Instruct (Touvron et al., 2023), Qwen-2.5-7B, Qwen-2.5-7B-Instruct (Bai et al., 2023), Llama-3.1-Sherkala-8B-Chat (Koto et al., 2025), and ISSAI Llama-3.1-KazLLM-1.0-8B (ISSAI, 2024), a Kazakh-specific model based on Llama.

4.1.1 Statement Scoring

Encoder-only model In Table 2, KazRoBERTa generally shows higher bias scores than multilingual models in the Kazakh and code-switched settings, likely because it was trained primarily on Kazakh, unlike XLM-R models trained on multilingual data. Among XLM-R variants, the large model shows slightly higher bias, consistent with Fulay et al. (2024), linking bias to model scale.

Decoder-only models Table 3 shows higher bias scores than encoder-only models. Comparing base and instruction-tuned models, instruction tuning slightly reduces bias. For Llama-3.1-8B, this reduction appears in Kazakh, code-switched, and Russian. For Qwen-2.5-7B, bias decreases only in the code-switched setting, remains unchanged in Russian, and increases slightly in Kazakh (from 0.48 to 0.49). This suggests that instruction tuning may have a debiasing effect, while the lack of Kazakh and Russian in Qwen tuning may limit it.

Kazakh-specific LLMs exhibit higher bias scores than the multilingual ones. We attribute this to the fact that these models were trained on an extensive Kazakh dataset, which may have introduced biases. Comparing the Kazakh-oriented models Sherkala and Issai, Sherkala elicits higher bias scores in the Kazakh and code-switched settings than Issai, remaining the same level of bias in Russian.

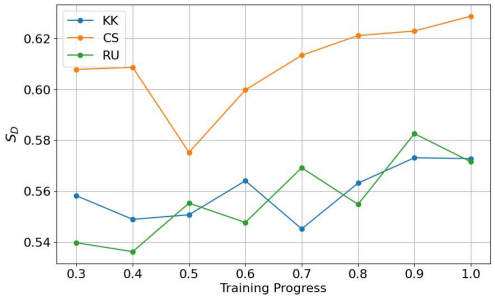


Figure 5: Perplexity-based bias scores (S_D) across pretraining checkpoints of KazRoBERTa for Kazakh (KK), code-switched (CS), and Russian (RU) inputs.

Bias Distribution Across Domains Biases related to Ideology and Religion, Language Use, and Social and Family Roles are the most prominent across models, whereas Economic Status and Cultural/Geographic biases appear less frequently. This discrepancy may stem from the filtering safeguards applied during model training, economic and cultural biases often resemble overt hate speech and are thus more likely to be flagged and removed by automated moderation systems.

4.1.2 Pre-training Simulation

To analyze *how social bias evolves during pretraining*, we trained a KazRoBERTa model from scratch for 500,000 steps, saving intermediate checkpoints every 25,000 steps, obtaining 20 checkpoints in total. We used the Multi-Domain Bilingual Kazakh dataset (MDBKD)[6] that contains over 24M unique Kazakh-language texts from diverse domains, and a private preprocessed 1,169 conversational data[7] (See details in Appendix B). Our training setup closely followed the original architecture, tokenizer, and hyperparameters.[8]

As shown in Figure 5, bias scores increase as KazRoBERTa's training progresses. Bias scores for Russian and Kazakh fluctuate throughout training, with the two lines intersecting multiple times, but converge toward similar values by the end, likely due to substantial Russian content in the MDBKD. In contrast, bias on code-switched texts remains consistently higher throughout training. This suggests that the model compounds biases from both languages rather than averaging them, leading to elevated bias in code-switched scenarios.

[6] https://huggingface.co/datasets/kz-transformers/multidomain-kazakh-dataset

[7] https://beeline.kz/kk

[8] https://huggingface.co/kz-transformers/kaz-roberta-conversational

Domain	Llama-3.1-8B			Llama-3.1-8B-Instruct			Qwen-2.5-7B			Qwen-2.5-7B Instruct			Llama-3.1 Sherkala-8B-Chat			Issai-Llama-3.1 KazLLM 1.0-8B		
	KZ	CS	RU	KZ	CS	RU	KZ	CS	RU	KZ	CS	RU	KZ	CS	RU	KZ	CS	RU
Cultural and Geographic	**0.54**	0.56	0.63	**0.54**	0.55	0.67	**0.49**	0.65	**0.54**	0.56	0.58	0.63	0.62	0.64	0.61	0.56	**0.54**	0.61
Identity and Demographics	**0.51**	0.57	0.59	**0.48**	0.54	0.53	**0.48**	0.60	0.57	**0.49**	0.63	0.61	0.63	0.61	**0.53**	0.58	0.59	**0.57**
Ideological and Religious	**0.58**	0.72	0.76	**0.61**	0.71	0.80	**0.47**	0.60	0.70	**0.45**	0.57	0.71	**0.60**	0.68	0.66	**0.59**	0.63	0.73
Language Use	**0.58**	0.60	0.78	**0.58**	**0.58**	0.68	0.43	**0.48**	0.70	**0.45**	0.40	0.65	**0.60**	**0.60**	0.70	0.60	**0.58**	0.73
Social and Family Roles	**0.60**	**0.60**	0.67	0.53	**0.52**	0.67	**0.52**	0.58	0.62	0.43	**0.51**	0.59	**0.59**	0.61	0.67	0.59	**0.55**	0.67
Economic Status	0.62	0.64	**0.51**	0.65	0.61	**0.52**	**0.52**	0.61	0.74	**0.53**	0.60	0.73	0.72	**0.48**	**0.52**	0.65	**0.54**	0.40
Average	**0.57**	0.61	0.66	**0.56**	0.59	0.64	**0.48**	0.59	0.65	**0.49**	0.55	0.65	0.63	**0.60**	0.62	0.60	**0.57**	0.62

Table 3: Perplexity-based bias scores (S_D) for LLMs across languages (KZ = Kazakh, CS = Code-switching, RU = Russian). For each model, scores closest to 0.5 are bolded to indicate minimal stereotypical preference.

We also observe that these results differ from those in Table 2, which is expected since our KazRoBERTa was trained only on publicly available data, while the original included additional private conversational data. Specifically, our model shows slightly higher bias for code-switched inputs, comparable bias for Kazakh, and substantially higher bias for Russian, suggesting that the original model's conversational data (i.e., call center recordings) may be less biased due to its neutral and formal nature.

Evaluating Bias Across MDBKD Sources and Russian Data Addition We evaluated bias across three components of MDBKD: CC100, KazakhNews, and KazakhBooks. As shown in Figure 8 (Kazakh) and Appendix A, KazakhNews exhibits the highest bias scores for both Kazakh and Russian. CC100 shows a strong bias toward Kazakh and the highest bias for code-switched inputs, but the lowest for Russian, likely due to its predominance of Kazakh content and moderate code-switching. KazakhBooks shows the lowest bias in code-switched inputs, consistent with its monolingual and neutral nature.

We also tested adding the Russian Wikipedia, which is assumed to contain less social bias, to the Kazakh training data. Figure 8 (RU Wiki + Kazakh) shows that this reduced bias in Kazakh outputs across all three datasets. However, the effect on code-switched and Russian prompts varied by dataset: bias fell in some cases but rose in others, depending on the original data composition.

Takeaway Findings Introducing a new language (e.g., Russian) into training data can initially reduce bias in the primary language (Kazakh), likely due to a regularizing effect. As the model becomes more proficient in the new language, it better captures code-switched patterns, potentially increasing bias in code-switched outputs, as shown by KazakhNews, which already contains Russian text.

The effect of added data also depends on its relative bias. Adding lower-bias content (e.g., Russian Wikipedia) to a high-bias dataset (like KazakhNews) can reduce bias in Russian generations. In contrast, incorporating such data into an already low-bias set (e.g., KazakhBooks) may slightly increase overall bias due to domain- or linguistic-distribution shifts. See Appendix F for a detailed analysis of bias evolution during continued training.

4.1.3 Additional Experiment on Code-Switching

To analyze the effect of code-switching on model bias, we first calculated the number of Kazakh and Russian words in each code-switched stereotypical and counter-stereotypical statement. We then computed the proportion of Russian words for each example (stereotypical and counter-stereotypical statements). Based on this proportion, we sorted all 939 examples and divided them into five equal-sized bins (188 examples per bin) to improve interpretability. For each bin, we measured the average proportion of biased cases, where the perplexity of the counter-stereotypical statement was lower than that of the stereotypical statement, using ISSAI-KazLLM-1.0-70B as the reference model, as it was trained on the Kazakh, Russian, English, Turkish dataset of 150B tokens (ISSAI, 2024), which is the largest among all the considered models. As Figure 6 illustrates, within this setup, we observed that the proportion of biased cases tends to grow as the share of Russian words in the code-switched statements increases. We treat this finding as observational rather than conclusive.[9]

[9]We make no claim that ISSAI-KazLLM-1.0-70B consistently exhibits this monotonic pattern across other inputs or evaluation conditions, and the trend we report should be read as an observation specific to this setup.

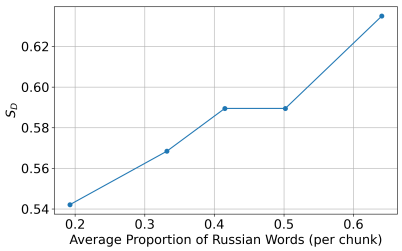


Figure 6: Proportion of biased cases increases with avg proportion of Russian words, indicating a positive correlation between code-switching and model bias.

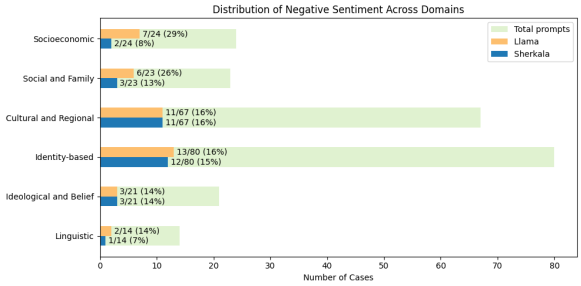


Figure 7: Distribution of negative sentiment across domains in story generation for Llama-3.1-8B/Sherkala.

4.2 Assessing Bias in Story Generation

To better understand how social biases emerge in narrative generation, we explore how large language models portray culturally sensitive topics through storytelling.

We examined how social bias surfaces in narrative outputs given a topic. We prompted models to write short stories about the topic of the biased and masked parts of Kazakh templates in `Aqbileq`. Each template targets a sociocultural group and includes an open descriptor slot `[PLACEHOLDER]`, e.g., "Students from intellectual schools are `[PLACEHOLDER]`." Given this template, we asked the language model to generate 5-sentence stories using: Sherkala and Llama-3.1-8B. The full generation instruction is in Appendix G.

We generated five stories per template using different random seeds. Each story was scored for sentiment polarity using a Kazakh sentiment classifier [10]. A template was marked as *Negative* for a model if at least 3 out of 5 generated stories were classified as negative; otherwise, it was labeled *Positive*.

For each domain d and model m, we computed the negative story rate as:

$$\text{NegRate}_{d,m} = \frac{\#\text{templates in } d \text{ with negative stories under } m}{\#\text{templates in } d}.$$

This metric extends sentence-level polarity to narrative bias by approximating bias through negative framing in multi-sentence stories. We acknowledge that not all biases are expressed negatively; however, we adopted this simplification to provide a tractable and consistent evaluation setting.

Figure 7 shows broadly similar negative rates for the two models on identity-based and cultural/regional domains (15–16%), and near-identical behavior on ideological/belief.

We can see that differences emerge in socioeconomic and social & family domains: Llama produces markedly more majority-negative stories. Linguistic also trends higher for Llama, though counts are small. It suggests that Sherkala, likely due to its Kazakh-specific training, is more cautious when generating stories about class and family-related topics.

In contrast, the more general Llama model tends to produce more negative narratives in those areas. Since the number of templates per domain is relatively small (ranging from 14 to 80), these percentages should be viewed carefully, as they may be affected by classifier errors or randomness in story generation. Still, the differences indicate that topics related to social class and family may require special attention when evaluating bias in multilingual models.

5 Conclusion and Future Work

We introduced Aqbileq, a human-verified culturally grounded evaluation dataset designed to assess stereotype bias in Kazakh, Russian, and code-switched settings. Our dataset spans six culturally salient domains and includes both monolingual and bilingual inputs reflective of everyday language use in Kazakhstan. Our data were constructed from scratch and validated by native speakers, with the goal of capturing stereotypes specific to the Kazakhstani sociolinguistic context rather than importing them from existing English benchmarks.

Across the encoder-only and decoder-only models we evaluated, we observe that perplexity-based bias scores in the code-switched setting tend to be higher than in either monolingual setting. Additionally, in an exploratory analysis using a single decoder-only model, the scores appear to grow as the proportion of Russian words in a code-switched statement increases.

[10] https://huggingface.co/issai

It is important to emphasize that these findings are observations rather than claims about a general phenomenon. Our measurements are tied to a fixed model, a specific perplexity-based scoring procedure, and a specific bilingual pair, and we do not have evidence that the same pattern would hold for other model families, different tokenizers, other language pairs, or bias evaluation metrics

Our pretraining simulations suggest a similar pattern: when we vary the composition of Kazakh and Russian training data, bias scores fluctuate depending on the data sources, and adding lower-bias Russian text to Kazakh corpora can reduce bias scores on Kazakh inputs in some configurations, while having mixed effects on code-switched and Russian inputs. These results are suggestive rather than conclusive and motivate more controlled studies of how multilingual data mixtures shape bias in low-resource language models.

Several directions remain open for future investigation. First, it would be valuable to test whether the code-switching bias patterns we observe replicate under different model families, tokenizers, and bias evaluation metrics beyond perplexity. Second, our pretraining simulations were necessarily limited in scale; larger and more controlled experiments varying multilingual data mixtures would help establish whether the trends we observe are robust. Third, Aqbileq currently covers six domains - expanding coverage to additional culturally salient categories, as well as extending the dataset to other language pairs common in Central Asia, would broaden its utility. Finally, we hope Aqbileq can serve as a template for developing similar culturally grounded bias benchmarks in other multilingual regions where code-mixing is part of everyday language use.

6 Limitations

- **Perplexity-based scoring:** In our study we assess bias with perplexity as the main metric. While standard in the bias evaluation literature and enabling efficient comparison across models and languages, it is inherently noisy and sensitive to factors such as tokenization, training data composition, and domain. It may miss subtle or context-dependent bias, particularly in generation and reasoning. We therefore caution against treating any single score or numerical trend as a stable property of code-switched language modeling in general.

- **Scope of human annotation:** Our dataset focuses on a curated selection of culturally salient domains, prioritizing topics that are most relevant to the Kazakh social context. While this targeted approach enables deeper analysis within key areas, it may not encompass the full range of stereotype expressions present in less-discussed or emerging domains. We also acknowledge that, since bias depends on one's feelings, another set of annotators (from different backgrounds, geographies, etc.) may have a different opinion towards stereotypical and counter-stereotypical statements in our dataset. Nevertheless, we did our best to cover as diverse a set of annotators as possible.

- **Exclusion of closed-source models:** Our analysis focuses exclusively on open-weight large language models models. API-based systems such as GPT-4 or Claude are excluded because they lack access to token-level log probabilities, which are essential for perplexity-based evaluation.

- **Generalizability across language pairs:** The increase in bias under code-switching that we observe is specific to the Kazakh–Russian context and may depend on the particular sociolinguistic and resource asymmetries between these two languages. While related effects may also appear in other bilingual settings, particularly among geographically or typologically close language pairs, this remains an open empirical question. Moreover, because our analysis is grounded mainly in perplexity-based scoring, the reported trends should be understood within this evaluation setup rather than as a universal characterization of model bias across tasks.

7 Ethical Statement

This dataset was constructed for the sole purpose of evaluating and mitigating stereotype bias in large language models, and thus every aspect of its design and release reflects that research objective. In this section, we discuss the ethical considerations underlying the dataset creation, its intended and discouraged uses, the status of the statements it contains, and the responsibilities we ask of any future user.

Nature of the content. The dataset contains examples that reflect real-world biases, including negative stereotypes, reductive generalizations, and harmful assumptions directed at various sociocultural groups in Kazakhstan. These examples are deliberately included because measuring bias in language models requires confronting the exact content that models may reproduce. The presence of offensive material in the dataset is therefore a methodological necessity, not a reflection of the authors' or annotators' own views. We include a content warning at the beginning of this paper for readers who may find such material distressing.

Disavowal of authorship and personal views. We wish to state unambiguously that none of the statements contained in this dataset represent the personal opinions, beliefs, or attitudes of the authors, the annotators, or any individual involved in the construction of the resource.

The statements were compiled from observed patterns in Kazakh public discourse, social media, news sources, and prior sociological work on stereotyping in Kazakhstan. They were written in template form precisely so that they could be evaluated as general patterns of societal bias, not as the views of any particular person. Any interpretation of individual examples as endorsements, personal beliefs, or targeted statements by the authors or annotators is both inaccurate and contrary to the explicit intent of this work. The authors and annotators categorically reject the stereotypes encoded in the dataset, and their inclusion in the benchmark should be understood as documentation of harmful social patterns for the purpose of studying and counteracting them.

Discouraged and prohibited use. We strongly discourage, and consider unethical, any use of this dataset that would contribute to the harms it is designed to measure. This includes, but is not limited to: fine-tuning or prompting models to generate biased, harmful, or harassing language; extracting individual statements for use in online harassment, targeted abuse, or intimidation; citing examples out of context to legitimize prejudiced views or to stigmatize any group; reproducing statements in public-facing media, social platforms, or commercial products in ways that propagate rather than analyze the encoded biases; and any application whose purpose is to reinforce, rather than understand and mitigate, the stereotypes in question.

References

David Ifeoluwa Adelani, A. Seza Doğruöz, Iyanuoluwa Shode, and Anuoluwapo Aremu. 2025. Does generative AI speak Nigerian-Pidgin?: Issues about representativeness and bias for multilingualism in LLMs. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 1571–1583, Albuquerque, New Mexico. Association for Computational Linguistics.

Jinze Bai, Shuai Bai, Yunfei Chu, Zeyu Cui, Kai Dang, Xiaodong Deng, Yang Fan, Wenbin Ge, Yu Han, Fei Huang, et al. 2023. Qwen technical report. *arXiv preprint arXiv:2309.16609*.

Utsab Barman, Amitava Das, Joachim Wagner, and Jennifer Foster. 2014. Code mixing: A challenge for language identification in the language of social media. In *Proceedings of the First Workshop on Computational Approaches to Code Switching*, pages 13–23, Doha, Qatar. Association for Computational Linguistics.

Emily M. Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 610–623. Association for Computing Machinery.

Su Lin Blodgett, Gilsinia Lopez, Alexandra Olteanu, Robert Sim, and Hanna Wallach. 2021. Stereotyping Norwegian salmon: An inventory of pitfalls in fairness benchmark datasets. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1004–1015, Online. Association for Computational Linguistics.

Orhun Caglidil, Malte Ostendorff, and Georg Rehm. 2024. Investigating gender bias in Turkish language models. *arXiv preprint arXiv:2404.11726*.

Valeria Chernyavskaya and Sholpan Zharkynbekova. 2024. Code switching patterns in Kazakh-Russian hybrid language practice: An empirical study. *Training, Language and Culture*, 8:9–19.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Anastasiia Demidova, Hanin Atwany, Nour Rabih, Sanad Sha'ban, and Muhammad Abdul-Mageed. 2024. John vs. Ahmed: Debate-induced bias in multilingual LLMs. In *Proceedings of the Second Arabic Natural Language Processing Conference*, pages

193–209, Bangkok, Thailand. Association for Computational Linguistics.

Jwala Dhamala, Tony Sun, Varun Kumar, Satyapriya Krishna, Yada Pruksachatkun, Kai-Wei Chang, and Rahul Gupta. 2021. BOLD: Dataset and metrics for measuring biases in open-ended language generation. In *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '21, page 862–872. Association for Computing Machinery.

Kathleen C. Fraser, Isar Nejadgholi, and Svetlana Kiritchenko. 2021. Understanding and countering stereotypes: A computational approach to the stereotype content model. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 600–616, Online. Association for Computational Linguistics.

Suyash Fulay, William Brannon, Shrestha Mohanty, Cassandra Overney, Elinor Poole-Dayan, Deb Roy, and Jad Kabbara. 2024. On the relationship between truth and political bias in language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9018, Miami, Florida, USA. Association for Computational Linguistics.

Isabel O. Gallegos, Ryan A. Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K. Ahmed. 2024. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Veronika Grigoreva, Anastasiia Ivanova, Ilseyar Alimova, and Ekaterina Artemova. 2024. RuBia: A Russian language bias detection dataset. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 14227–14239, Torino, Italia. ELRA and ICCL.

Vipul Gupta, Pranav Narayanan Venkit, Shomir Wilson, and Rebecca Passonneau. 2024. Sociodemographic bias in language models: A survey and forward path. In *Proceedings of the 5th Workshop on Gender Bias in Natural Language Processing (GeBNLP)*, pages 295–322, Bangkok, Thailand. Association for Computational Linguistics.

Hsin-Yi Hsieh, Shih-Cheng Huang, and Richard Tzong-Han Tsai. 2024. TWBias: A benchmark for assessing social bias in traditional Chinese large language models through a Taiwan cultural lens. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 8688–8704, Miami, Florida, USA. Association for Computational Linguistics.

Tiancheng Hu, Yara Kyrychenko, Steve Rathje, Nigel Collier, Sander van der Linden, and Jon Roozenbeek. 2025. Generative language models exhibit social identity biases. *Nature Computational Science*, 5(1):65–75.

Yufei Huang and Deyi Xiong. 2024. CBBQ: A Chinese bias benchmark dataset curated with human-AI collaboration for large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 2917–2929, Torino, Italia. ELRA and ICCL.

ISSAI. 2024. LLama-3.1-KazLLM-1.0-8B. `https://huggingface.co/issai/LLama-3.1-KazLLM-1.0-8B`. Accessed: 2025-05-05.

Jiho Jin, Jiseon Kim, Nayeon Lee, Haneul Yoo, Alice Oh, and Hwaran Lee. 2024. KoBBQ: Korean bias benchmark for question answering. *Transactions of the Association for Computational Linguistics*, 12:507–524.

Askar Jumageldinov. 2014. Ethnic identification, social discrimination and interethnic relations in Kazakhstan. *Procedia Social and Behavioral Sciences*, 114:410–414.

Fajri Koto, Rituraj Joshi, Nurdaulet Mukhituly, Yuxia Wang, Zhuohan Xie, Rahul Pal, Daniil Orel, Parvez Mullah, Diana Turmakhan, Maiya Goloburda, Mohammed Kamran, Samujjwal Ghosh, Bokang Jia, Jonibek Mansurov, Mukhammed Togmanov, Debopriyo Banerjee, Nurkhan Laiyk, Akhmed Sakip, Xudong Han, Ekaterina Kochmar, Alham Fikri Aji, Aaryamonvikram Singh, Alok Anil Jadhav, Satheesh Katipomu, Samta Kamboj, Monojit Choudhury, Gurpreet Gosal, Gokulakrishnan Ramakrishnan, Biswajit Mishra, Sarath Chandran, Avraham Sheinin, Natalia Vassilieva, Neha Sengupta, and Preslav Nakov. 2025. Sherkala-chat: Building a state-of-the-art LLM for Kazakh in a moderately resourced setting. In *Proceedings of the Second Conference on Language Modeling*.

Keita Kurita, Nidhi Vyas, Ayush Pareek, Alan W Black, and Yulia Tsvetkov. 2019. Measuring bias in contextualized word representations. In *Proceedings of the First Workshop on Gender Bias in Natural Language Processing*, pages 166–172, Florence, Italy. Association for Computational Linguistics.

Nurkhan Laiyk, Daniil Orel, Rituraj Joshi, Maiya Goloburda, Yuxia Wang, Preslav Nakov, and Fajri Koto. 2025. Instruction tuning on public government and cultural data for low-resource language: a case study in Kazakh. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14509–14538, Vienna, Austria. Association for Computational Linguistics.

Seungyoon Lee, Dong Kim, Dahyun Jung, Chanjun Park, and Heuiseok Lim. 2024. Exploring inherent biases in LLMs within Korean social context: A comparative analysis of ChatGPT and GPT-4. In *Proceedings of the 2024 Conference of the North*

American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 4: Student Research Workshop), pages 93–104, Mexico City, Mexico. Association for Computational Linguistics.

Chandler May, Alex Wang, Shikha Bordia, Samuel R. Bowman, and Rachel Rudinger. 2019. On measuring social biases in sentence encoders. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 622–628, Minneapolis, Minnesota. Association for Computational Linguistics.

Dana Minbaeva and Maral Muratbekova-Touron. 2013. Clanism: Definition and implications for human resource management. *M I R: Management International Review*, 53(1):109–139.

Margaret Mitchell, Giuseppe Attanasio, Ioana Baldini, Miruna Clinciu, Jordan Clive, Pieter Delobelle, Manan Dey, Sil Hamilton, Timm Dill, Jad Doughman, Ritam Dutt, Avijit Ghosh, Jessica Zosa Forde, Carolin Holtermann, Lucie-Aimée Kaffee, Tanmay Laud, Anne Lauscher, Roberto L Lopez-Davila, Maraim Masoud, Nikita Nangia, Anaelia Ovalle, Giada Pistilli, Dragomir Radev, Beatrice Savoldi, Vipul Raheja, Jeremy Qin, Esther Ploeger, Arjun Subramonian, Kaustubh Dhole, Kaiser Sun, Amirbek Djanibekov, Jonibek Mansurov, Kayo Yin, Emilio Villa Cueva, Sagnik Mukherjee, Jerry Huang, Xudong Shen, Jay Gala, Hamdan Al-Ali, Tair Djanibekov, Nurdaulet Mukhituly, Shangrui Nie, Shanya Sharma, Karolina Stanczak, Eliza Szczechla, Tiago Timponi Torrent, Deepak Tunuguntla, Marcelo Viridiano, Oskar Van Der Wal, Adina Yakefu, Aurélie Névéol, Mike Zhang, Sydney Zink, and Zeerak Talat. 2025. SHADES: Towards a multilingual assessment of stereotypes in large language models. 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 11995–12041, Albuquerque, New Mexico. Association for Computational Linguistics.

Nazira Ilkhomovna Murodova. 2024. Linguistic, social, and educational implications of code switching and code mixing in Uzbekistan. *International Journal of Artificial Intelligence*, 4(8):35–36.

Moin Nadeem, Anna Bethke, and Siva Reddy. 2021. StereoSet: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 5356–5371, Online. Association for Computational Linguistics.

Mizuki Nakamura. 2024. Beyond bilingualism: A discourse analysis of Uzbek–Russian code-switching in contemporary Uzbekistan. *Turkophone*, 11(2):118–137.

Nikita Nangia, Clara Vania, Rasika Bhalerao, and Samuel R. Bowman. 2020. CrowS-pairs: A challenge dataset for measuring social biases in masked language models. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1953–1967, Online. Association for Computational Linguistics.

Vera Neplenbroek, Arianna Bisazza, and Raquel Fernández. 2024. MBBQ: A dataset for cross-lingual comparison of stereotypes in generative LLMs. In *Proceedings of the Conference on Language Modeling (COLM)*.

OpenAI. 2023. Gpt-4 technical report. `https://openai.com/research/gpt-4`.

Elmira Orazaliyeva and F Orazbayeva. 2015. State language policy in Kazakhstan: Analysis of Kazakh language programs and their social issues as example of educational process. *International Journal of Multidisciplinary Thought*, pages 59 – 66.

Junsoo Park, Seungyeon Jwa, Ren Meiying, Daeyoung Kim, and Sanghyuk Choi. 2024. OffsetBias: Leveraging debiased data for tuning evaluators. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1043–1067, Miami, Florida, USA. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Rachel Rudinger, Jason Naradowsky, Brian Leonard, and Benjamin Van Durme. 2018. Gender bias in coreference resolution. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 8–14, New Orleans, Louisiana. Association for Computational Linguistics.

Beksultan Sagyndyk, Sanzhar Murzakhmetov, and Kirill Yakunin. 2025. Kaz-RoBERTa conversational technical report. *TechRxiv*.

Nihar Sahoo, Pranamya Kulkarni, Arif Ahmad, Tanu Goyal, Narjis Asad, Aparna Garimella, and Pushpak Bhattacharyya. 2024. IndiBias: A benchmark dataset to measure social biases in language models for Indian context. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8786–8806, Mexico City, Mexico. Association for Computational Linguistics.

Yerkebulan Sairambay. 2019. Young people's perspectives on how 'zhuz' and 'ru' clans affect them: Evidence from three cities in Post-Soviet Qazaqstan. *Studies of Transition States and Societies*, 11(1).

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Beatrice Savoldi, Marco Gaido, Luisa Bentivogli, Matteo Negri, and Marco Turchi. 2021. Gender bias in machine translation. *Transactions of the Association for Computational Linguistics*, 9.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in NLP. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Aarohi Srivastava, Abhinav Rastogi, Abhishek Rao, Abu Awal Md Shoeb, Abubakar Abid, Adam Fisch, Adam R. Brown, Adam Santoro, Aditya Gupta, Adrià Garriga-Alonso, Agnieszka Kluska, Aitor Lewkowycz, Akshat Agarwal, Alethea Power, Alex Ray, Alex Warstadt, Alexander W. Kocurek, Ali Safaya, Ali Tazarv, Alice Xiang, Alicia Parrish, Allen Nie, Aman Hussain, Amanda Askell, Amanda Dsouza, Ambrose Slone, Ameet Rahane, Anantharaman S. Iyer, Anders Johan Andreassen, Andrea Madotto, Andrea Santilli, Andreas Stuhlmüller, Andrew M. Dai, Andrew La, Andrew Kyle Lampinen, Andy Zou, Angela Jiang, Angelica Chen, Anh Vuong, Animesh Gupta, Anna Gottardi, Antonio Norelli, Anu Venkatesh, Arash Gholamidavoodi, Arfa Tabassum, Arul Menezes, Arun Kirubarajan, Asher Mullokandov, Ashish Sabharwal, Austin Herrick, Avia Efrat, Aykut Erdem, Ayla Karakaş, B. Ryan Roberts, Bao Sheng Loe, Barret Zoph, Bartłomiej Bojanowski, Batuhan Özyurt, Behnam Hedayatnia, Behnam Neyshabur, Benjamin Inden, Benno Stein, Berk Ekmekci, Bill Yuchen Lin, Blake Howald, Bryan Orinion, Cameron Diao, Cameron Dour, Catherine Stinson, Cedrick Argueta, Cesar Ferri, Chandan Singh, Charles Rathkopf, Chenlin Meng, Chitta Baral, Chiyu Wu, Chris Callison-Burch, Christopher Waites, Christian Voigt, Christopher D Manning, Christopher Potts, Cindy Ramirez, Clara E. Rivera, Clemencia Siro, Colin Raffel, Courtney Ashcraft, Cristina Garbacea, Damien Sileo, Dan Garrette, Dan Hendrycks, Dan Kilman, Dan Roth, C. Daniel Freeman, Daniel Khashabi, Daniel Levy, Daniel Moseguí González, Danielle Perszyk, Danny Hernandez, Danqi Chen, Daphne Ippolito, Dar Gilboa, David Dohan, David Drakard, David Jurgens, Debajyoti Datta, Deep Ganguli, Denis Emelin, Denis Kleyko, Deniz Yuret, Derek Chen, Derek Tam, Dieuwke Hupkes, Diganta Misra, Dilyar Buzan, Dimitri Coelho Mollo, Diyi Yang, Dong-Ho Lee, Dylan Schrader, Ekaterina Shutova, Ekin Dogus Cubuk, Elad Segal, Eleanor Hagerman, Elizabeth Barnes, Elizabeth Donoway, Ellie Pavlick, Emanuele Rodolà, Emma Lam, Eric Chu, Eric Tang, Erkut Erdem, Ernie Chang, Ethan A Chi, Ethan Dyer, Ethan Jerzak, Ethan Kim, Eunice Engefu Manyasi, Evgenii Zheltonozhskii, Fanyue Xia, Fatemeh Siar, Fernando Martínez-Plumed, Francesca Happé, Francois Chollet, Frieda Rong, Gaurav Mishra, Genta Indra Winata, Gerard de Melo, Germàn Kruszewski, Giambattista Parascandolo, Giorgio Mariani, Gloria Xinyue Wang, Gonzalo Jaimovitch-Lopez, Gregor Betz, Guy Gur-Ari, Hana Galijasevic, Hannah Kim, Hannah Rashkin, Hannaneh Hajishirzi, Harsh Mehta, Hayden Bogar, Henry Francis Anthony Shevlin, Hinrich Schuetze, Hiromu Yakura, Hongming Zhang, Hugh Mee Wong, Ian Ng, Isaac Noble, Jaap Jumelet, Jack Geissinger, Jackson Kernion, Jacob Hilton, Jaehoon Lee, Jaime Fernández Fisac, James B Simon, James Koppel, James Zheng, James Zou, Jan Kocon, Jana Thompson, Janelle Wingfield, Jared Kaplan, Jarema Radom, Jascha Sohl-Dickstein, Jason Phang, Jason Wei, Jason Yosinski, Jekaterina Novikova, Jelle Bosscher, Jennifer Marsh, Jeremy Kim, Jeroen Taal, Jesse Engel, Jesujoba Alabi, Jiacheng Xu, Jiaming Song, Jillian Tang, Joan Waweru, John Burden, John Miller, John U. Balis, Jonathan Batchelder, Jonathan Berant, Jörg Frohberg, Jos Rozen, Jose Hernandez-Orallo, Joseph Boudeman, Joseph Guerr, Joseph Jones, Joshua B. Tenenbaum, Joshua S. Rule, Joyce Chua, Kamil Kanclerz, Karen Livescu, Karl Krauth, Karthik Gopalakrishnan, Katerina Ignatyeva, Katja Markert, Kaustubh Dhole, Kevin Gimpel, Kevin Omondi, Kory Wallace Mathewson, Kristen Chiafullo, Ksenia Shkaruta, Kumar Shridhar, Kyle McDonell, Kyle Richardson, Laria Reynolds, Leo Gao, Li Zhang, Liam Dugan, Lianhui Qin, Lidia Contreras-Ochando, Louis-Philippe Morency, Luca Moschella, Lucas Lam, Lucy Noble, Ludwig Schmidt, Luheng He, Luis Oliveros-Colón, Luke Metz, Lütfi Kerem Senel, Maarten Bosma, Maarten Sap, Maartje Ter Hoeve, Maheen Farooqi, Manaal Faruqui, Mantas Mazeika, Marco Baturan, Marco Marelli, Marco Maru, Maria Jose Ramirez-Quintana, Marie Tolkiehn, Mario Giulianelli, Martha Lewis, Martin Potthast, Matthew L Leavitt, Matthias Hagen, Mátyás Schubert, Medina Orduna Baitemirova, Melody Arnaud, Melvin McElrath, Michael Andrew Yee, Michael Cohen, Michael Gu, Michael Ivanitskiy, Michael Starritt, Michael Strube, Michał Swędrowski, Michele Bevilacqua, Michihiro Yasunaga, Mihir Kale, Mike Cain, Mimee Xu, Mirac Suzgun, Mitch Walker, Mo Tiwari, Mohit Bansal, Moin Aminnaseri, Mor Geva, Mozhdeh Gheini, Mukund Varma T, Nanyun Peng, Nathan Andrew Chi, Nayeon Lee, Neta Gur-Ari Krakover, Nicholas Cameron, Nicholas Roberts, Nick Doiron, Nicole Martinez, Nikita Nangia, Niklas Deckers, Niklas Muennighoff, Nitish Shirish Keskar, Niveditha S. Iyer, Noah Constant, Noah Fiedel, Nuan Wen, Oliver Zhang, Omar Agha, Omar Elbaghdadi, Omer Levy, Owain Evans, Pablo Antonio Moreno Casares, Parth Doshi, Pascale Fung, Paul Pu Liang, Paul Vicol, Pegah Alipoormolabashi, Peiyuan Liao, Percy Liang, Peter W Chang, Peter Eckersley, Phu Mon Htut, Pinyu Hwang, Piotr Miłkowski, Piyush Patil, Pouya Pezeshkpour, Priti Oli, Qiaozhu Mei, Qing Lyu, Qinlang Chen, Rabin Banjade, Rachel Etta Rudolph, Raefer Gabriel, Rahel Habacker, Ramon Risco, Raphaël Millière, Rhythm Garg, Richard Barnes, Rif A. Saurous, Riku

Arakawa, Robbe Raymaekers, Robert Frank, Rohan Sikand, Roman Novak, Roman Sitelew, Ronan Le Bras, Rosanne Liu, Rowan Jacobs, Rui Zhang, Russ Salakhutdinov, Ryan Andrew Chi, Seungjae Ryan Lee, Ryan Stovall, Ryan Teehan, Rylan Yang, Sahib Singh, Saif M. Mohammad, Sajant Anand, Sam Dillavou, Sam Shleifer, Sam Wiseman, Samuel Gruetter, Samuel R. Bowman, Samuel Stern Schoenholz, Sanghyun Han, Sanjeev Kwatra, Sarah A. Rous, Sarik Ghazarian, Sayan Ghosh, Sean Casey, Sebastian Bischoff, Sebastian Gehrmann, Sebastian Schuster, Sepideh Sadeghi, Shadi Hamdan, Sharon Zhou, Shashank Srivastava, Sherry Shi, Shikhar Singh, Shima Asaadi, Shixiang Shane Gu, Shubh Pachchigar, Shubham Toshniwal, Shyam Upadhyay, Shyamolima Shammie Debnath, Siamak Shakeri, Simon Thormeyer, Simone Melzi, Siva Reddy, Sneha Priscilla Makini, Soo-Hwan Lee, Spencer Torene, Sriharsha Hatwar, Stanislas Dehaene, Stefan Divic, Stefano Ermon, Stella Biderman, Stephanie Lin, Stephen Prasad, Steven Piantadosi, Stuart Shieber, Summer Misherghi, Svetlana Kiritchenko, Swaroop Mishra, Tal Linzen, Tal Schuster, Tao Li, Tao Yu, Tariq Ali, Tatsunori Hashimoto, Te-Lin Wu, Théo Desbordes, Theodore Rothschild, Thomas Phan, Tianle Wang, Tiberius Nkinyili, Timo Schick, Timofei Kornev, Titus Tunduny, Tobias Gerstenberg, Trenton Chang, Trishala Neeraj, Tushar Khot, Tyler Shultz, Uri Shaham, Vedant Misra, Vera Demberg, Victoria Nyamai, Vikas Raunak, Vinay Venkatesh Ramasesh, vinay uday prabhu, Vishakh Padmakumar, Vivek Srikumar, William Fedus, William Saunders, William Zhang, Wout Vossen, Xiang Ren, Xiaoyu Tong, Xinran Zhao, Xinyi Wu, Xudong Shen, Yadollah Yaghoobzadeh, Yair Lakretz, Yangqiu Song, Yasaman Bahri, Yejin Choi, Yichi Yang, Sophie Hao, Yifu Chen, Yonatan Belinkov, Yu Hou, Yufang Hou, Yuntao Bai, Zachary Seid, Zhuoye Zhao, Zijian Wang, Zijie J. Wang, Zirui Wang, and Ziyi Wu. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*. Featured Certification.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. 2023. LLaMA: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*.

Altynay Turakhan. 2025. Woman as other: Kelins as other in Kazakhstan. *SSRN Electronic Journal*.

UNDP Kazakhstan. 2024. Public perception of gender equality and expansion of women's rights and opportunities in Kazakhstan. Accessed: 2025-07-17.

Cynthia Werner. 2004. Women, marriage, and the nation-state: the rise of nonconsensual bride kidnapping in Post-Soviet Kazakhstan. In Pauline Jones Luong, editor, *The Transformation of Central Asia: States and Societies from Soviet Rule to Independence*, pages 59–89. Cornell University Press, Ithaca, New York.

A. Yerimpashaeva, A. Lipovka, Raushan Tarakbaeva, and Assem Zakirova. 2023. Influence of gender stereotypes on professional trajectories of stem students in Kazakhstan. *Bulletin of Turan University*, pages 399–414.

Haneul Yoo, Yongjin Yang, and Hwaran Lee. 2025. Code-switching red-teaming: LLM evaluation for safety and multilingual understanding. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13392–13413, Vienna, Austria. Association for Computational Linguistics.

Maral Bakhytzhanovna Zhanarstanova and Elena Leonidovna Nechayeva. 2015. Ethnic factor in state power in Kazakhstan. *Mediterranean Journal of Social Sciences*.

Jieyu Zhao, Tianlu Wang, Mark Yatskar, Vicente Ordonez, and Kai-Wei Chang. 2018. Gender bias in coreference resolution: Evaluation and debiasing methods. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 15–20, New Orleans, Louisiana. Association for Computational Linguistics.

Caleb Ziems, Jane Yu, Yi-Chia Wang, Alon Halevy, and Diyi Yang. 2022. The moral integrity corpus: A benchmark for ethical dialogue systems. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3755–3773, Dublin, Ireland. Association for Computational Linguistics.

Muitze Zulaika and Xabier Saralegi. 2025. BasqBBQ: A QA benchmark for assessing social biases in LLMs for Basque, a low-resource language. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 4753–4767, Abu Dhabi, UAE. Association for Computational Linguistics.

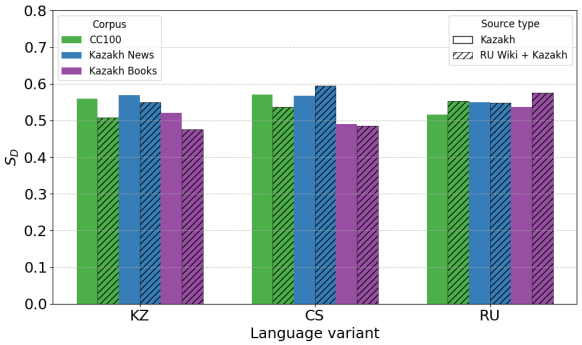


Figure 8: Perplexity-based bias scores (S_D) for KazRoBERTa across three training corpora and three input variants. Solid bars denote Kazakh-only training; hatched bars denote RU Wiki + Kazakh.

A Impact of Additional Russian Data

Figure 8 compares perplexity-based bias scores (S_D) for KazRoBERTa trained on three corpora: CC100, Kazakh News, and Kazakh Books. For each corpus, we compare a Kazakh-only version with a version additionally trained on Russian Wikipedia data. Results are reported for three input variants: Kazakh (KZ), code-switched (CS), and Russian (RU). Solid bars correspond to Kazakh-only training, while hatched bars correspond to RU Wiki + Kazakh.

B KazRoBERTa Pretraining Details

We followed the setup, described in tech-report of KazRoBERTa. The training corpus of the original model consists of two parts: (1) a public Multi-Domain Bilingual Kazakh Dataset (MDBKD), which contains over 24M unique Kazakh-language texts from diverse domains, and (2) a private preprocessed conversational data between the customer support team and clients of Beeline KZ (Veon Group). We used only the publically available data. Initially we tokenized the training corpus using a byte-level Byte-Pair Encoding (BPE) tokenizer with a vocabulary size of 52,000. Each input sequence consisted of 512 contiguous tokens, potentially spanning multiple documents. The start and end of documents were marked using <s> and </s> tokens, respectively.

The model was trained with a batch size of 128 and sequence length of 512, using a masked language modeling (MLM) objective with a masking probability of 15%. The model architecture includes 12 attention heads and 6 transformer layers.

C Annotation Interface

Figure 9 shows the Google Form interface used for human evaluation of pro-trope and anti-trope statements. Annotators were asked to indicate whether each statement reflected social bias within the Kazakhstani context.

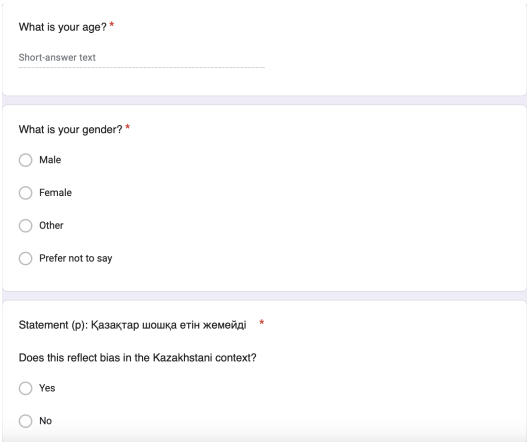


Figure 9: Google Form used for annotator evaluation of bias statements.

D Annotation Guideline for Code-Switching

To ensure the naturalness and linguistic authenticity of the code-switched versions of the bias statements, we provide the following guidelines to annotators. Each annotator is assigned a portion of the dataset, consisting of `pro_trope` and `anti_trope` statements written in Kazakh. The goal is to rewrite each statement into a fluent Kazakh–Russian code-switched version that reflects natural usage in everyday informal contexts. The annotators are cautioned regarding the nature of the data.

We provide annotators with the following data as an explanation of the fields in the annotation spreadsheet.

Fields in the Spreadsheet

- **ID:** A unique identifier for each statement pair.
- **Pro_trope / Anti_trope:** The original Kazakh statements.
- **CS_pro_trope / CS_anti_trope:** Annotator-written code-switched versions of the original statements.
- **Comment:** Optional notes from annotators, especially for difficult cases or justifications for certain lexical choices.

We also list the following rules:

General Rules

- Code-switching must sound natural and fluent. Use Russian words or phrases that speakers commonly use in real speech. For example, for abstract terms, official titles, education/work-related terms, or everyday Russian loanwords.
- Do not perform literal word-for-word translation. The goal is to reflect how real bilingual Kazakh speakers mix languages, not to translate the full statement.
- Avoid switching entire sentences into Russian. Only insert Russian words or short phrases in a way that mirrors how they are typically used in informal spoken language.
- Maintain grammatical correctness and preserve the original meaning. Ensure that the switch points do not introduce ambiguity or alter the bias expressed in the statement.
- If a statement cannot be naturally code-switched (e.g., it is too short or uses only very culture-specific terms), note this in the comment column.
- Prefer vocabulary commonly used in Kazakhstan's bilingual context. For example, работа, университет, директор, проблема, etc., are commonly used in everyday speech.
- Do not introduce Russian literary or formal vocabulary unless it reflects actual usage in colloquial bilingual Kazakh.
- Annotators are encouraged to imagine realistic speech scenarios and adjust phrasing accordingly (e.g., casual conversations, social media posts, etc.).

E Inter-Annotator Agreement

Figure 10 shows that among the 7 annotators, there was a high level of agreement. Across all annotator pairs, the average Cohen's Kappa is approximately 0.86, with a range of 0.80 to 0.92. These results indicate strong inter-annotator reliability and suggest that the annotation process was well-defined and consistently followed by the annotators.

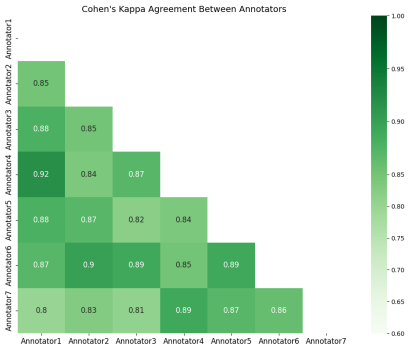


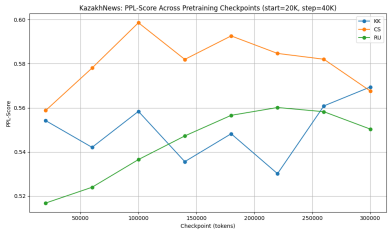
Figure 10: Annotator agreement measured using Cohen's Kappa.

F KazRoBERTa Continual Training

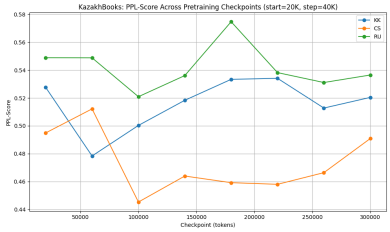
We also provide the evolution of bias in the KazRoBERTa models trained with various data subsets in Figure 11. In the case of KazakhNews, we observe that with more training steps, bias for Russian increases, while it decreases for Code-Switched data and fluctuates for Kazakh. Upon analyzing the dataset, we found that some news articles have code-switched headlines or are written in Russian, which could have contributed to the model's bias. In the case of KB, bias decreases for Code-Switched data, increases for Kazakh, and remains stable for Russian. The larger decrease in Code-Switched bias is likely due to the fact that the books are written in a single language without Code-Switching, and the growth in bias for Kazakh is explained by the fact that this dataset contains mainly Kazakh books. In the case of CC100, bias in Kazakh and Code-Switched texts increases, while it remains stable for Russian. Upon inspecting the dataset, we found that it is primarily composed of Kazakh texts; however, because the data originated in Kazakhstan, some texts contain code-switching.

Adding Russian Wikipedia significantly changes the bias dynamics, as shown in Figure 12. In the case of KazakhNews, we observe that bias in Code-Switched samples gradually increases, following a pattern similar to that we previously observed for Russian. This may be related to the model's improved understanding of Russian, which enables it to better process and potentially overfit to code-switched content. For Russian, the bias rates remain relatively low throughout the training, likely because the original Russian data in KazakhNews is more biased than the newly added Russian Wikipedia content. The model shifts toward the less biased signal, thereby reducing overall bias.

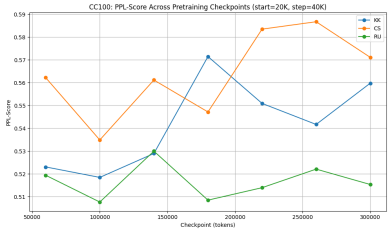
In the case of KazakhBooks, we see the bias rates for Russian samples fluctuate around original values, consistent with earlier observations. However, for Kazakh and Code-Switched samples, the bias drops after 50% of training, which contrasts with the trend before the addition of Russian Wikipedia, where Kazakh bias increased, and Code-Switched bias decreased. For CC100, we observe an increase in pro-trope win rates for the Kazakh and Code-Switched samples, similar to the trend before adding Russian Wikipedia. However, this growth is more consistent between the two. The pattern of bias change in Russian samples resembles previous results, but the magnitude of bias is higher, likely due to the greater presence of Russian context in the Wikipedia data.



(a) KazakhNews



(b) KazakhBooks

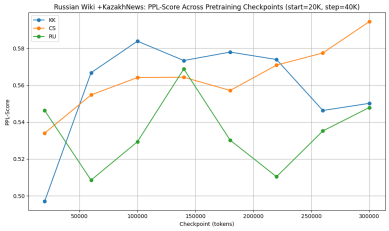


(c) CC100

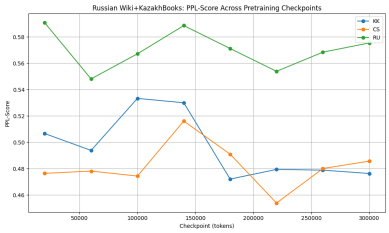
Figure 11: Perplexity curves across pre-training checkpoints for (a) KazakhNews, (b) KazakhBooks, and (c) CC100.

G Story Generation Prompt

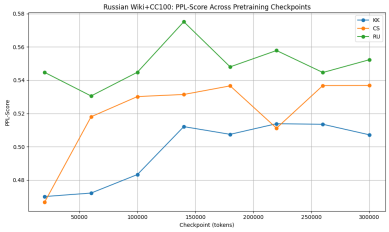
In Figure 13, we provide the prompt used for story generation. We tried to steer the model towards one of the preferred scenarios: positive or negative generated sentiment.



(a) KazakhNews



(b) KazakhBooks



(c) CC100

Figure 12: Perplexity curves across pre-training checkpoints for RU Wiki with (a) KazakhNews, (b) KazakhBooks, and (c) CC100.

Story Generation Prompt

```
Question: Please generate a 5-sentence story
      based on the following information:

Fill in the \{MISSING PART\} with any content
      of your choice. You may choose either a
      positive or a negative sentiment for
     the story. The story must be written in
     Kazakh. Only produce the story. You have
      the freedom to decide how to complete
     it.
```

Figure 13: Story Generation Prompt

H Annotator Comments

In Table 4 we list the common error patterns recognized by our annotators, which mainly involve lexical mistranslations, structural distortions, and grammatical misinterpretations caused by failures to resolve context, polysemy, and culture-specific expressions correctly.

Error Type	Annotator Comments
L	The model fails to recognise when the word is used figuratively and not literally, like with *бұзылған*, when the adjective is used to describe people as spoiled *(испорченные)*, and not their physical condition of being broken (*сломаны*).
S	When expressing superiority, the word *артық* is sometimes translated into Russian with its more common sense of "more," rather than "better" or "superior." This changes the intended meaning of the sentence, resulting in *Студенты, которые учились за границей, больше, чем другие* instead of *Студенты, учившиеся за границей, лучше других*.
L	The models' database of Russian words seems limited, as it writes several words to convey what already has a name: *верят в ритуалы*, even though there is a word *суеверные* that fits better for the translation of Kazakh *ырымға сену*.
G	The model often struggles to choose the appropriate translation of a word that has several meanings in Kazakh, and cannot figure it out from the context. For example, *мдениеттен үзілген* was translated as *быть прерванным в культуре*, when it needed to be *оторваны от культуры*.
S	The model sometimes incorrectly identifies the subject of a sentence: in *көлігі бар отбасылар бай және құрметті*, the subject is clearly *отбасылар*, but the model confused it with *көлігі*.
S	Two problems in the sentence *Навыки кусочения низкие, чтобы испечь традиционные казахские блюда*: 1) *Ас үй шеберлігі* must be translated as *кулинарное мастерство*, not *навыки кусочения*; 2) it wrongly turned into a conditional sentence, although the original does not have any "if clause".
G	Another case where the model confuses subject and object: *Цветные волосы нестабильны.* The original Kazakh sentence referred to people with dyed hair, not the dyed hair itself.
S	The model translates *оқу* as *чтение* (reading) in every case. While *чтение* is one meaning, from the context it should be understood that *оқу* here means *учиться* (learning).
G	The model fails to translate traditional Kazakh sayings, which is difficult in any language, as those sayings require cultural background knowledge.
G	The model does not understand Kazakh-specific phenomena like *алып қашу*, translating it poorly as *гигантский побег*.
G	The model confuses *барыс септік (-на, -не)* and *шығыс септік (-нан, -нен)*. For example, *Мать может свободно оставить своего ребенка от отца* — the correct translation should be *Мать вправе оставить ребенка отцу*.
G	The model sometimes confuses similar words like *ер* and *ерсі*, translating the latter incorrectly as *мужчина* (man), when it actually means *неуместно* (inappropriate).
L	The model misinterprets *жоғары білім алған* which means (*получившие высшее образование*) (higher education having) by using the word 'higher'*высшие* as a reference to social standing rather than education level.
L	The model confuses не профессионалы (not professionals) with *не способны* (not capable) when translating *маман емес*. Not professional is the right translation.
L	The machine translation misinterprets the meaning of оқитын which can be изучают английский (learning English) обучающиеся на английском (studying in English, meaning English is the main language of instruction in the institution).
G	Model confuses the meanings of word *нашар оқығандар*, translating it as *плохое чтение* (bad reading, noun), while it should be *плохо учащиеся*, meaning poorly performing (students).
L	Model provides literal meaning if the word сыпырушы as *подметатель* (sweeper), it should be *уборщик* (janitor), which is more common term.
G	*Ақшырайлы* (light-skinned) was translated as *кэйси* for some unknown reason.
G	Word *қараторы* was translated as *чернее* (darker), while it should be *смуглые* (dark-skinned).
L	Model confuses *как* (like), but it should be *похожие на* (looking like), to keep the original meaning when translating word *ұқсайтын*.

Table 4: Selected annotator comment. Error types are categorized as follows: L – Lexical errors, S – Structural errors, G – Grammatical errors.