

CA-ClimateNews: A Human-Annotated Multilingual Benchmark for Climate-Relevance Classification of Central Asian News

Adilet Metinov¹ Zarina Uvalieva¹ Bektemir Kumarbai uulu¹
Nurtilek Alibekov¹ Tynchtykbek Tashbaltaev¹

¹Airun Intelligence Lab, Bdigital LLC
metinovadilet@gmail.com

Abstract

Climate change is acute in Central Asia (glacier loss, the Aral Sea, and drought exposure), yet there are no NLP resources for climate discourse in the region’s low-resource Turkic languages. We present **CA-ClimateNews**, the first human-annotated multilingual dataset for climate-relevance classification of Central Asian news, covering **Kyrgyz, Kazakh, Uzbek, and (regionally grounded) Russian**. The corpus comprises **3,140** news items labeled at two levels (binary climate relevance and an 8-way topic category) by two annotators per language, with a 100-item blind double-annotation overlap adjudicated to gold. Inter-annotator agreement is almost perfect (Cohen’s κ 0.86–0.98 on relevance). Because we used transparent LLM-assisted pre-labeling, we additionally run a **between-subjects** experiment showing that **displaying model pre-labels induces anchoring**: an independent group shown the pre-label overrode it 3–7 $\times$ less than a blind group. Our released labels are the blind version. We report baselines: fine-tuned multilingual encoders reach 88–96 relevance macro-F1 on the balanced benchmark; a Russian-only model loses up to ~ 7 points transferring zero-shot to the Turkic languages *even at matched training size* (least, ~ 2 , for Kazakh), and a related Turkic source is no better, isolating a cross-lingual gap that in-language supervision closes. The dataset, guidelines, and code are publicly released.¹

1 Introduction

Central Asia is on the front line of climate change: shrinking Tian Shan and Pamir glaciers (Farinotti et al., 2015) feed the region’s rivers, the Aral Sea has largely disappeared (Micklin, 2007), and heat and drought threaten agriculture and water security (Shaw et al., 2022). Understanding how this is discussed in regional media requires being able to *find*

and *categorize* climate content, a text-classification problem. Yet no annotated resources exist for climate discourse in the region’s low-resource Turkic languages (Kyrgyz, Kazakh, Uzbek), and multilingual climate NLP has focused almost entirely on English (Diggelmann et al., 2020; Webersinke et al., 2021; Coan et al., 2021; Laud et al., 2023).

We close this gap with **CA-ClimateNews**, a human-annotated benchmark for two-level climate-relevance classification of Central Asian news in four languages. Building it surfaced a striking property of the domain: climate is a *needle in a haystack*: only $\sim 1\%$ of a random news sample is climate relevant (1.00%; 58/5,780 LLM-labeled random items, 95% Wilson CI [0.8, 1.3]%), ranging from 0.34% (Kyrgyz) to 1.7% (Uzbek). This scarcity is precisely why the resource did not previously exist and why targeted collection is required.

Our contributions are:

1. **The first multilingual, human-annotated climate-relevance dataset for Central Asia** (3,140 items; Kyrgyz/Kazakh/Uzbek/Russian), with two-level labels, near-perfect inter-annotator agreement, and adjudicated gold.
2. **A between-subjects anchoring result**: an independent annotator group shown the LLM pre-label overrides it 3–7 $\times$ less than a blind group, a concrete, quantified caution for LLM-assisted labeling pipelines.
3. **Baselines** revealing the low-resource gap: multilingual and monolingual fine-tuning, an LLM pre-labeling baseline, and an RU $\rightarrow$ Turkic zero-shot transfer experiment with matched-data and Turkic $\rightarrow$ Turkic controls, showing a cross-lingual drop of up to ~ 7 points (~ 2 for Kazakh) that a related same-script source does not avoid and that subword overlap does not explain.

¹<https://github.com/ma-bdigital/CA-ClimateNews>

2 Related Work

Climate NLP is overwhelmingly English: claim verification (Diggelmann et al., 2020), topic/relevance detection (Varini et al., 2020), environmental-claim detection (Stammach et al., 2023), domain language models (Webersinke et al., 2021), contrarian-claim detection (Coan et al., 2021), and multi-task benchmarks (Laud et al., 2023). We found *no* climate or environmental text-classification resource for Kyrgyz, Kazakh, or Uzbek, a gap this paper addresses. The closest methodological cousin is MASAKHANEWS (Adelani et al., 2023), a multilingual human-annotated news-topic dataset for African languages whose pipeline, like many, had annotators correct classifier pre-labels, the practice we scrutinize in §6.

Central Asian NLP has grown but not toward climate. Kazakh has NER (Yeshpanov et al., 2022); Uzbek has news topic classification (Kuriyozov et al., 2023) and monolingual encoders (Mansurov and Mansurov, 2021; Kuriyozov et al., 2024); Kyrgyz has a multilabel topic benchmark (Alekseev et al., 2023) and, recently, a first monolingual encoder (Metinov et al., 2025); and Turkic understanding benchmarks exist (TUMLU (Isbarov et al., 2025), Kardeş-NLU (Senel et al., 2024), SIB-200 (Adelani et al., 2024)), with recent surveys mapping the landscape (Veitsman and Hartmann, 2025; Alekseev and Turatali, 2024), but none targets climate or news relevance. Cross-lingual coverage comes from multilingual encoders: XLM-R (Conneau et al., 2020), Glot500 (ImaniGooghari et al., 2023), mmBERT (Marone et al., 2025), though these over-fragment Turkic text (Rust et al., 2021), and zero-shot transfer (Conneau et al., 2018; Hu et al., 2020; Ruder et al., 2021) falters for low-resource, distant targets (Joshi et al., 2020; Senel et al., 2024), the regime our RU→Turkic experiment probes.

LLM-assisted annotation. LLMs increasingly serve as data annotators (Gilardi et al., 2023; Wang et al., 2021; Ding et al., 2023), so showing their pre-labels to human verifiers is now common practice. But pre-annotation is long known to bias annotators toward the suggestion (Fort and Sagot, 2010), an instance of automation bias and anchoring (Skitka et al., 2000; Tversky and Kahneman, 1974). We quantify this effect for LLM pre-labels (§6).

Lang	Items	%REL	Date range
RU	800	52.0	2009–2026
KK	740	49.6	2003–2026
KY	800	47.8	2003–2026
UZ	800	52.4	2015–2026
All	3,140	50.4	2003–2026

Table 1: Corpus overview. %REL = share labeled climate-relevant (L1). KK totals 740 rather than 800 because, after deduplication, its outlet archives yielded fewer distinct climate-relevant items to balance against.

3 Dataset

Sources and collection. We collected news items (headline, lead, URL, date) from nine public Central Asian outlets via RSS and archive sitemaps: Azattyk (KY), 24.kg (KY/RU), Egemen and Azattyq (KK), Kun.uz, Daryo.uz and Gazeta.uz (UZ), and Tengrinews, Kaktus and Sputnik (RU). Because climate content is rare (~1% of random samples), the climate portion was collected by keyword/slug search over full outlet archives, and the non-relevant portion by random sampling. Crucially, the non-relevant half deliberately mixes *hard negatives*, climate-adjacent items (generic weather, pollution, or agriculture with no climate framing) drawn from the same keyword harvest, with easy random negatives, so a classifier cannot succeed by keyword presence alone. All items are deduplicated by normalized URL and normalized title. Both the keyword-seeded and random components are reported for transparency (§7). Language codes label the *article* language; the Russian portion is drawn from Central Asian outlets across all three countries, making it regionally grounded rather than Moscow-centric.

Size and balance. Table 1 summarizes the corpus: 3,140 items, balanced at 50.4% climate-relevant per the annotation design.

Label schema. Each item is labeled at two levels. **L1 (relevance, binary):** RELEVANT if substantively about climate change, its impacts, or responses; else NOT_RELEVANT (including generic weather or pollution with no climate framing). **L2 (topic, 8-way, only when L1=RELEVANT):** WATER_CRYO, DISASTER, AGRI_LAND, ENERGY, POLICY, SCIENCE_GEN, HEALTH_SOCIETY, OTHER_CLIMATE. Hard-case rules (weather/ disaster events relevant only if linked to climate; energy relevant only with renewable-transition/emissions/policy framing) are given verbatim in the guidelines; L2 takes the sin-

Lang	L1 κ	L1 %agr	L2 κ	L2 %agr
RU	0.859	93	0.826	86
KK	0.858	93	0.932	95
KY	0.860	93	1.000	100
UZ	0.980	99	0.872	90

Table 2: Inter-annotator agreement on the blind 100-item overlap. All Cohen’s $\kappa \geq 0.81$ (almost perfect).

gle *dominant* frame, ties breaking to the more specific category. The final L2 distribution is dominated by WATER_CRYO (464) and POLICY (426), reflecting the region’s glacier/water and policy-oriented coverage.

4 Annotation Protocol

Two native-speaker annotators labeled each language, following a shared guidelines document (in Russian, the common language) designed per standard annotation practice (Pustejovsky and Stubbs, 2012). The protocol was: (1) a 30-item Russian *calibration* round to align annotators; (2) a 100-item *blind double-annotation overlap* per language (LLM pre-labels hidden) for agreement; (3) *adjudication* of disagreements into gold; (4) blind single-annotation of the remaining items, split disjointly and source-balanced between the two annotators. Every item received a human decision; no item ships on an LLM label alone. Separately, for the anchoring experiment (§6), an *independent* group of two native speakers per language annotated the same items with the LLM pre-label visible; those labels are used only for that comparison and are not released.

5 Inter-Annotator Agreement

On the 100-item blind overlap per language, Cohen’s κ (Cohen, 1960) is in the “almost perfect” band (Landis and Koch, 1977) for both levels (Table 2); see Artstein and Poesio (2008) on agreement measures. L2 κ is computed only on the items both annotators called relevant (42/40/47/52 for RU/KK/KY/UZ), so KY’s $\kappa = 1.000$ is perfect agreement on 47 items. The 35 disagreements were adjudicated into gold, each decision released per item. We further **triple-annotate the held-out test set**: two more annotators re-labeled all 630 items *blind* (three-way L1 κ 0.75–0.99, 90.3% unanimous; App. B), and majority adjudication yields the released *test gold* used for all baseline metrics.

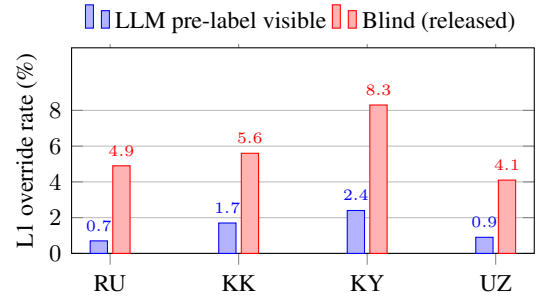


Figure 1: L1 override rate for an *LLM-visible* annotator group vs. an independent *blind* group (two native speakers per language each). The blind group overrides 3–7× more often, exposing anchoring; its rate approaches the LLM-error estimate (2–14%) and the human–human floor (1–7%).

6 The LLM Pre-label Anchoring Effect

We used transparent LLM-assisted pre-labeling: a large language model (Anthropic’s Claude Sonnet 5), prompted with the annotation guidelines at default decoding, proposed L1/L2 labels for humans to verify or correct; the exact prompt and model version are released with the code. To test whether the pre-label biases annotators, we ran a **between-subjects** experiment: for each language, one group of two native speakers annotated with the pre-label *visible* (and were told it was model-generated), and a *separate, independent* group of two native speakers annotated the same items *blind* (pre-label hidden). Because the groups are disjoint, the comparison is free of the within-annotator ordering, learning, and memory confounds that a single-group before/after design would incur.

Each item was single-annotated; an override is any human label that differs from the model’s. Over the full single-annotated bulk (N=2,740; 640–700/language, *not* the overlap), the blind group overrode the model on 157 items vs. the visible group’s 39 (5.7% vs. 1.4%; 118 more corrections), and the blind rate exceeds the visible rate in *all four* languages by a large margin (3–7×; mean +4.3 points; Figure 1). Because each condition has only two annotators, an item-level test would overstate significance. But the conditions use *disjoint* annotator sets (sixteen people, none in both), so the 4/4 replication is across distinct individuals, not just languages (App. C). The comparison is against the model’s *fixed* pre-labels, so it does not depend on gold, and the blind rate approaches the LLM-error and human–human floors (Figure 1). **Recommendation:** hide model pre-labels, or elicit

Lang	L1 Acc	L1 κ	L2 Acc	L2 κ
RU	94.8	0.895	85.5	0.822
KK	93.2	0.865	89.8	0.852
KY	91.5	0.831	85.1	0.795
UZ	96.1	0.922	78.0	0.727
All	93.9	0.878	84.4	0.802

Table 3: LLM pre-label baseline vs. human labels (% and κ). L2 on items both call relevant.

the human label first; our released labels are the blind version.

7 Baselines

LLM pre-labeling. Against the human labels, the LLM pre-labeler reaches 93.9% L1 accuracy (κ 0.878) with high recall (98.5%) but lower precision (90.3%): it over-cludes, functioning as a strong recall-oriented pre-filter that still needs human review (Table 3). L2 accuracy is 84.4% (κ 0.802); the top confusions (OTHER_CLIMATE $\leftrightarrow$ SCIENCE_GEN, WATER_CRYO $\leftrightarrow$ DISASTER) coincide with the categories the two humans disagreed on: genuine frame ambiguity, not random error.

Supervised and transfer baselines. These *characterize* the benchmark; they are untuned reference points, not tuned systems. We fine-tune encoders on a 70/10/20 split stratified by language $\times$ (L1,L2), and report per-language macro-F1 on the held-out test set (Table 4). We compare: (i) a majority-class floor; (ii) massively multilingual encoders: XLM-R (Conneau et al., 2020), Glot500 (Imani-Googhari et al., 2023), mmBERT (Marone et al., 2025), mBERT (Devlin et al., 2019), fine-tuned jointly on all four languages; (iii) per-language monolingual encoders (Kaz-RoBERTa (Sagyndyk et al., 2025), TahrirchiBERT, KyrgyzBERT2 (the latest release of Metinov et al., 2025), RuBERT); and (iv) **cross-lingual transfer**: XLM-R fine-tuned on Russian only and evaluated *zero-shot* on the three Turkic languages, with a matched-data *in-language control* (XLM-R on the same $\sim$ 560 target items) and a *Turkic $\rightarrow$ Turkic control* (KK $\rightarrow$ KY) that together separate cross-lingual loss from data quantity and from source distance (Table 5). Main baselines and the KK $\rightarrow$ KY control report the mean over three seeds (13/42/100); the RU $\rightarrow$ Turkic and in-language rows use eight.

Results. On the balanced (50/50) test set, all multilingual encoders reach 88–96 L1 macro-F1, far above the $\sim$ 33 majority floor; no single en-

Model	KY	KK	UZ	RU
Majority	30.7	32.6	34.8	34.7
<i>Multilingual (joint fine-tune)</i>				
XLM-R	90.9	87.9	89.6	96.2
Glott500	93.2	87.7	91.7	95.2
mBERT	91.1	87.5	90.6	95.4
mmBERT	91.9	84.1	94.6	94.8
Best mono.	87.7	88.3	81.1	92.3

Table 4: Level-1 (climate relevance) macro-F1 ($\times$ 100), mean over 3 seeds (13/42/100); per-cell std 0.4–3.3 (largest for mBERT on KK). Best per language in **bold**. Best mono. = KyrgyzBERT2 / Kaz-RoBERTa / TahrirchiBERT / RuBERT, each fine-tuned on its own language. Monolingual Kaz-RoBERTa edges the best multilingual on KK by 0.4 (within seed variance); elsewhere the multilingual encoders lead by $\sim$ 4 on RU and 5–14 on KY/UZ.

XLM-R training setup	KY	KK	UZ
In-language (mono, L-only)	91.3	88.3	83.7
RU-only $\rightarrow$ zero-shot	86.8	86.3	77.1
KK-only $\rightarrow$ zero-shot [†]	86.2	–	–
Δ (in-lang – RU-only)	+4.5	+2.0	+6.6

Table 5: RU $\rightarrow$ Turkic transfer at a matched $\sim$ 560-item budget (Level-1 macro-F1 $\times$ 100, mean over 8 seeds; std $\leq$ 4.0). These matched-budget in-language numbers use only $\sim$ 560 items, so fall below the full-data joint fine-tune of Table 4 (notably UZ). Zero-shot from Russian costs 2.0–6.6 points; the Kyrgyz/Uzbek gaps are positive in all 8 seeds, the Kazakh gap (+2.0) within seed variance. [†]Turkic $\rightarrow$ Turkic control (KK $\rightarrow$ KY), mean over 3 seeds.

coder dominates (best multilingual: Glot500 on KY, XLM-R on KK/RU, mmBERT on UZ). No monolingual clearly beats them (Kaz-RoBERTa is level on Kazakh within seed variance; the Kyrgyz/Uzbek encoders trail $\sim$ 6/ $\sim$ 14; Table 4). Our main characterization result is the **RU $\rightarrow$ Turkic transfer gap** (Table 5): at a matched budget ($\sim$ 560 items, 8 seeds), zero-shot transfer from Russian costs 2.0–6.6 macro-F1 points versus in-language training (smallest for Kazakh). This is *not* straightforwardly explained by script or lexical distance: a same-family Turkic source transfers no better (KK $\rightarrow$ KY 86.2; Table 5), and subword overlap does not track the gap, reflecting the value of *in-language* supervision that no related source substitutes. L2 (8-way topic) is harder and noisier (best mmBERT 0.58–0.75; Appendix A). We release CA-ClimateNews (near-perfect agreement, adjudicated gold, transparent LLM assistance, an anchor-

ing caution) with baselines showing a RU→Turkic gap that in-language supervision closes.

Limitations

The held-out *test set* is triple-annotated and the 400 overlap items are gold (§5); the remaining training items are single-annotated blind. Human-human and LLM-error rates are estimated on 100 items/language (sampling noise; the anchoring *direction* is robust across all four languages). The corpus is news register only, uses a headline+lead unit (the climate frame is set in the lead), and a single-label 8-way scheme; we make no dialectal-coverage claims. Collection mixes keyword-seeded (climate) and random (non-climate) sampling, and includes state-affiliated outlets; both are disclosed. Because the climate-relevant items are keyword/slug-retrieved, climate coverage that avoids the seed vocabulary is under-represented in the positive class (the hard-negative design guards the negative class but not this); the benchmark therefore measures relevance discrimination rather than open-world recall. Some items lack a parsed date (mostly one outlet); date ranges differ by language (KY and KK reach back to 2003, RU to 2009, UZ only to 2015) because of archive depth, so outlet and period are confounded with language. The test set is balanced (~50% relevant) for stable macro-F1; at the ~1% deployment prevalence the models are best read as a high-recall *retrieval* filter, not final classifiers. The low resulting precision is the base-rate effect, not a model flaw: at ~90% recall, prevalence-adjusted precision ranges from ~4% (Kyrgyz) to 31% (Russian) (Table 9), yet the triage value is large: a reviewer then inspects only 4–15% of the stream (7–26× fewer items), and raising the threshold to 80% recall lifts precision to 52–100% on Russian/Uzbek. The prevalence estimates (~1% overall; ~1,400–1,500 random items/language) rest on few positives (as few as five for Kyrgyz), so per-language values carry wide confidence intervals and Table 9’s deployment numbers are illustrative.

Ethics and Data Statement

We release headlines, leads, URLs, access dates, and labels, not full article text. The **annotations/labels are released under CC BY 4.0** and the **code under the MIT License**; the headline and lead excerpts remain the property of their respective outlets and are provided as short excerpts under

fair-use/quotation norms, with a takedown process for any outlet that requests it. All labels are keyed to source URLs and access dates, so the text can be re-fetched from source if preferred. The study involved four native-speaker annotators per language (two who produced the released blind labels and two independent annotators for the LLM-visible anchoring condition), plus two further annotators who blind-re-labeled the held-out test set, all university-educated, aged 20–30, residing in the respective countries, and compensated at local professional rates. Known biases: news register only, a keyword-seeded climate portion, and state-affiliated outlets in the mix.

Acknowledgments

We thank the native-speaker annotators for Kyrgyz, Kazakh, Uzbek, and Russian for their careful work, and the anonymous reviewers for their helpful feedback.

References

- David Ifeoluwa Adelani, Marek Masiak, Israel Abebe Azime, et al. 2023. MasakhaNEWS: News topic classification for African languages. In *Proceedings of IJCNLP-AACL (Volume 1: Long Papers)*.
- David Ifeoluwa Adelani et al. 2024. SIB-200: A simple, inclusive, and big benchmark for topic classification in 200+ languages and dialects. In *Proceedings of EACL*. ArXiv:2309.07445.
- Anton Alekseev, Sergey I. Nikolenko, and Gulnara Kabaeva. 2023. Benchmarking multilabel topic classification in the Kyrgyz language. *arXiv preprint arXiv:2308.15952*.
- Anton Alekseev and Timur Turatali. 2024. KyrgyzNLP: Challenges, progress, and future. *arXiv preprint arXiv:2411.05503*.
- Ron Artstein and Massimo Poesio. 2008. Survey article: Inter-coder agreement for computational linguistics. *Computational Linguistics*, 34(4):555–596.
- Travis G Coan, Constantine Boussalis, John Cook, and Mirjam O Nanko. 2021. Computer-assisted classification of contrarian claims about climate change. *Scientific Reports*, 11.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and Psychological Measurement*, 20(1).
- 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 ACL*.
- Alexis Conneau, Ruty Rinott, Guillaume Lample, Adina Williams, Samuel R. Bowman, Holger Schwenk, and Veselin Stoyanov. 2018. XNLI: Evaluating cross-lingual sentence representations. In *Proceedings of EMNLP*, pages 2475–2485.
- Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of NAACL-HLT*.
- Thomas Diggelmann, Jordan Boyd-Graber, Jannis Bulian, Massimiliano Ciaramita, and Markus Leippold. 2020. CLIMATE-FEVER: A dataset for verification of real-world climate claims. *arXiv preprint arXiv:2012.00614*.
- Bosheng Ding, Chengwei Qin, Linlin Liu, Yew Ken Chia, Boyang Li, Shafiq Joty, and Lidong Bing. 2023. Is GPT-3 a good data annotator? In *Proceedings of ACL (Volume 1: Long Papers)*, pages 11173–11195.
- Daniel Farinotti, Laurent Longuevergne, Geir Moholdt, Doris Duethmann, Thomas Mölg, Tobias Bolch, Sergiy Vorogushyn, and Andreas Güntner. 2015. Substantial glacier mass loss in the Tien Shan over the past 50 years. *Nature Geoscience*, 8(9):716–722.
- Karën Fort and Benoît Sagot. 2010. Influence of pre-annotation on POS-tagged corpus development. In *Proceedings of the Fourth Linguistic Annotation Workshop (LAW IV)*, pages 56–63.
- Fabrizio Gilardi, Meysam Alizadeh, and Maël Kubli. 2023. ChatGPT outperforms crowd workers for text-annotation tasks. *Proceedings of the National Academy of Sciences*, 120(30):e2305016120.
- Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation. In *Proceedings of ICML*, volume 119, pages 4411–4421.
- Ayyoob ImaniGooghari et al. 2023. Glot500: Scaling multilingual corpora and language models to 500 languages. In *Proceedings of ACL*. ArXiv:2305.12182.
- Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, et al. 2025. TUMLU: A unified and native language understanding benchmark for Turkic languages. *arXiv preprint arXiv:2502.11020*.
- Pratik Joshi, Sebastin Santy, Amar Budhiraja, Kalika Bali, and Monojit Choudhury. 2020. The state and fate of linguistic diversity and inclusion in the NLP world. In *Proceedings of ACL*, pages 6282–6293.
- Elmurod Kuriyozov, David Vilares, and Carlos Gómez-Rodríguez. 2024. BERTbek: A pretrained language model for Uzbek. In *Proceedings of SIGUL*.
- Elmurod Kuriyozov et al. 2023. Text classification dataset and analysis for Uzbek language. *arXiv preprint arXiv:2302.14494*.
- J Richard Landis and Gary G Koch. 1977. The measurement of observer agreement for categorical data. *Biometrics*, 33(1).
- Tanmay Laud, Daniel Spokoyny, Tom Corringham, and Taylor Berg-Kirkpatrick. 2023. ClimaBench: A benchmark dataset for climate change text understanding in english. In *arXiv preprint arXiv:2301.04253*.
- B. Mansurov and A. Mansurov. 2021. UzBERT: pre-training a BERT model for Uzbek. *arXiv preprint arXiv:2108.09814*.
- Marc Marone et al. 2025. mmBERT: A modern multilingual encoder. *arXiv preprint arXiv:2509.06888*.
- Adilet Metinov, A. Kudakeeva, and A. Kabaeva. 2025. KyrgyzBERT: A compact, efficient language model for Kyrgyz. *arXiv preprint arXiv:2511.20182*.
- Philip Micklin. 2007. The Aral Sea disaster. *Annual Review of Earth and Planetary Sciences*, 35:47–72.
- James Pustejovsky and Amber Stubbs. 2012. *Natural Language Annotation for Machine Learning*. O’Reilly Media.
- Sebastian Ruder, Noah Constant, Jan Botha, Aditya Siddhant, Orhan Firat, Jinlan Fu, Pengfei Liu, Junjie Hu, Dan Garrette, Graham Neubig, and Melvin Johnson. 2021. XTREME-R: Towards more challenging and nuanced multilingual evaluation. In *Proceedings of EMNLP*, pages 10215–10245.
- Phillip Rust, Jonas Pfeiffer, Ivan Vulić, Sebastian Ruder, and Iryna Gurevych. 2021. How good is your tokenizer? on the monolingual performance of multilingual language models. In *Proceedings of ACL-IJCNLP (Volume 1: Long Papers)*, pages 3118–3135.
- Beksultan Sagyndyk, Sanzhar Murzakhmetov, and Kirill Yakunin. 2025. Kaz-RoBERTa conversational: Technical report. *TechRxiv*. DOI 10.36227/techrxiv.175942902.25827042/v1.
- Lütfi Kerem Senel, Benedikt Ebing, Konul Baghirova, Hinrich Schütze, and Goran Glavaš. 2024. Kardeş-NLU: Transfer to low-resource languages with the help of a high-resource cousin – a benchmark and evaluation for Turkic languages. In *Proceedings of EACL (Volume 1: Long Papers)*, pages 1672–1688.
- R. Shaw, Y. Luo, T. S. Cheong, S. Abdul Halim, S. Chaturvedi, M. Hashizume, G. E. Insarov, Y. Ishikawa, M. Jafari, A. Kitoh, J. Pulhin, C. Singh, K. Vasant, and Z. Zhang. 2022. Asia. In H.-O. Pörtner et al., editors, *Climate Change 2022: Impacts, Adaptation and Vulnerability. Contribution of Working Group II to the Sixth Assessment Report of the IPCC*, pages 1457–1579. Cambridge University Press.

- Linda J Skitka, Kathleen Mosier, and Mark D Burdick. 2000. Automation bias and errors: Are crews better than individuals? *The International Journal of Aviation Psychology*, 10(1).
- Dominik Stammbach, Nicolas Webersinke, Julia Bingler, Mathias Kraus, and Markus Leippold. 2023. Environmental claim detection. In *Proceedings of ACL (Volume 2: Short Papers)*, pages 1051–1066.
- Amos Tversky and Daniel Kahneman. 1974. Judgment under uncertainty: Heuristics and biases. *Science*, 185(4157).
- Francesco S. Varini, Jordan Boyd-Graber, Massimiliano Ciaramita, and Markus Leippold. 2020. ClimateText: A dataset for climate change topic detection. In *NeurIPS 2020 Workshop on Tackling Climate Change with Machine Learning*. ArXiv:2012.00483.
- Yana Veitsman and Mareike Hartmann. 2025. Recent advancements and challenges of Turkic Central Asian language processing. In *Proceedings of the First Workshop on Language Models for Low-Resource Languages (LoResLM)*, pages 309–324.
- Shuohang Wang, Yang Liu, Yichong Xu, Chenguang Zhu, and Michael Zeng. 2021. Want to reduce labeling cost? GPT-3 can help. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 4195–4205.
- Nicolas Webersinke, Mathias Kraus, Julia Anna Bingler, and Markus Leippold. 2021. ClimateBERT: A pre-trained language model for climate-related text. In *arXiv preprint arXiv:2110.12010*.
- Rustem Yeshpanov, Yerbolat Khassanov, and Huseyin Atakan Varol. 2022. KazNERD: Kazakh named entity recognition dataset. In *Proceedings of LREC*. ArXiv:2111.13419.

Model	KY	KK	UZ	RU
Majority	9.4	8.9	6.4	5.5
XLM-R	58.0	62.5	49.1	48.3
Glott500	29.2	28.8	35.3	32.2
mBERT	55.6	63.8	48.3	52.9
mmBERT	70.2	75.3	65.4	58.2
Best mono.	41.1	37.8	22.0	36.3

Table 6: Level-2 (8-way topic) macro-F1 ($\times 100$) on relevant test items, mean over 3 seeds (13/42/100). mmBERT dominates; Glot500 collapses and is high-variance (std up to 13) on this fine-grained, small-sample task.

Lang	A-B	A-M	B-M	unan.	flips
RU	0.900	0.875	0.875	91.2%	6
KK	0.840	0.852	0.799	87.2%	7
KY	0.869	0.749	0.799	85.4%	13
UZ	0.988	0.938	0.951	96.9%	4
All	0.901	0.854	0.857	90.3%	30

Table 7: Test-set Level-1 agreement across three independent annotators. A, B = the two new blind annotators; M = existing master label. unan. = three-way unanimity; flips = L1 labels changed by majority adjudication.

A Level-2 Topic Results

Table 6 reports per-language macro-F1 for the 8-way topic task (L2), evaluated on the relevant items of each language’s test set (~ 75 –90 items). Because several categories are rare, macro-F1 is noisy; treat these as indicative. mmBERT is strongest overall. Glot500’s collapse (L2 ~ 30 vs. L1 ~ 91) is a training-instability artifact, not a representational limit: its L2 macro-F1 varied widely across seeds (std up to 13, vs. ≤ 5 for the other multilingual encoders); we report it unchanged for transparency. The small monolingual models also struggle on this fine-grained task.

B Test-Set Triple Annotation

Two annotators independently re-labeled the entire held-out test set *blind*; together with the existing label this gives three independent labels per item. Table 7 reports pairwise Level-1 Cohen’s κ (A, B = the two new annotators; M = the existing master label), three-way unanimity, and the number of labels changed by majority adjudication. All pairwise κ are in the “almost perfect” band; majority voting corrects 30 of 630 L1 labels (4.8%), which become the released test gold used for all Section 7 metrics.

Lang	N	vis.	blind	vis.%	blind%	$\times$
RU	700	5	34	0.7	4.9	6.8
KK	640	11	36	1.7	5.6	3.3
KY	700	17	58	2.4	8.3	3.4
UZ	700	6	29	0.9	4.1	4.8

Table 8: Anchoring override counts by condition over the single-annotated bulk. vis./blind = items (and %) where the human label differed from the model in the LLM-visible / blind condition; $\times$ = blind/visible ratio. We report effect sizes and four-way replication rather than an item-level p -value (see text).

C Anchoring Significance

The anchoring override rates (§6, Figure 1) are measured over the full single-annotated bulk, not the 100-item overlap. Table 8 gives the per-language item counts N , the visible- and blind-condition override counts and rates, and the blind/visible ratio. Because each condition has only two annotators who split the items disjointly, override decisions are not independent within a condition, so an item-level test (e.g. a two-proportion z treating 2,740 items as independent) would badly overstate significance. We therefore analyze at the language level: the blind rate exceeds the visible rate in all four languages by a large, consistent margin (mean $+4.3$ points, 3 – $7\times$; one-sided sign test $p=0.06$). Because the two conditions use disjoint annotator sets (sixteen people, none in both), this four-way replication is across distinct individuals; we rest the anchoring claim on the effect size and replication rather than on a single small-sample p -value.

D Deployment Operating Points

Balanced macro-F1 measures discrimination; deployment on the natural $\sim 1\%$ climate stream is a retrieval task. Table 9 reports, for each language’s best encoder, the precision–recall tradeoff *at the natural prevalence*: at each recall target, the fraction of the stream flagged (a reviewer’s reading load), the prevalence-adjusted precision, and the resulting load reduction. Precision is tunable: raising the threshold to 80% recall lifts precision substantially (to 100% on Russian, 52% on Uzbek) while flagging under 3% of the stream. Kazakh and Kyrgyz stay low-precision even then, reflecting higher false-positive rates.

Lang	model	recall	flagged	prec	load
RU	XLM-R	90%	3.8%	31%	26×
RU	XLM-R	80%	1.0%	100%	96×
KK	XLM-R	90%	14.5%	5%	7×
KK	XLM-R	80%	7.5%	8%	13×
KY	Glott500	90%	7.4%	4%	14×
KY	Glott500	80%	1.7%	16%	59×
UZ	mmBERT	90%	5.4%	28%	19×
UZ	mmBERT	80%	2.7%	52%	37×

Table 9: Deployment operating points at the natural per-language prevalence (KY 0.34%, KK 0.74%, UZ 1.69%, RU 1.30%). flagged = share of the stream predicted relevant (reading load); prec = prevalence-adjusted precision; load = reading-load reduction ($1/\text{flagged}$) vs. reading everything.