Processing Inconsistency Predicts Language Competence: LLM Evaluation Without Answer Labels on Turkic Languages

Ilya Galyukshev
Central University / Moscow
galyukshev.ilya@gmail.com

Ilseyar Alimova
Applied AI Institute / Moscow
alimovailseyar@gmail.com

Abstract

Most languages lack labeled evaluation benchmarks for large language models (LLMs). Creating such benchmarks requires native speakers, domain expertise, and answer annotation—resources unavailable for the vast majority of languages. We investigate whether a model's internal processing signals—such as generation entropy and tokenizer statistics—correlate with its actual accuracy on a language, with the long-term goal of estimating language competence without labeled data. Our key observation is that for languages a model does not know, both tokenizer segmentation and generation entropy become highly variable across questions, whereas for known languages they remain consistent. We call this the *inconsistency hypothesis* and test it on 11 instruction-tuned LLMs (1B–70B parameters) across 14 language–script varieties (12 Turkic plus English and Russian controls). We extract over 25 processing features per model–language pair; individually, even the strongest correlate only moderately with accuracy (Pearson $|r|$ up to 0.55). Yet combining just three complementary features—a tokenizer coverage ratio, entropy variability, and the model's English/Russian benchmark score—explains 75% of accuracy variance in leave-one-language-out evaluation, nearly doubling the 44% explained by a model-mean baseline. The variability of processing signals (standard deviation) consistently outperforms mean values as a predictor across all five model families, but only for greedy-pass measures; sampling-based measures show no such pattern.[1]

1 Introduction

Most of the world's approximately 7,000 languages lack any natural language understanding (NLU) evaluation benchmark (Joshi et al., 2020). Creating labeled benchmarks requires native speakers, domain expertise, and funding. This gap is particularly acute for Turkic languages: a large language family spanning multiple scripts and regions, with rich agglutinative morphology that challenges subword tokenizers. While TUMLU (Isbarov et al., 2025) recently introduced natively authored benchmarks for Turkic languages, creating such resources remains costly and language-by-language.

We hypothesize that a model's internal processing signals—extracted from its responses to questions, without using answer keys—correlate with its actual competence in a given language. We call this the *inconsistency hypothesis*: a model that knows a language well should process different questions with similar confidence, whereas a model unfamiliar with a language will handle some questions confidently and struggle with others. This cross-question *variability* in processing signals, rather than their average level, may serve as a diagnostic of language competence.

The contribution of this paper is as follows: (i) Empirical test of the inconsistency hypothesis across 11 instruction-tuned LLMs (1B–70B) from five model families, covering 8 Turkic languages in 12 script varieties, with English and Russian controls; (ii) Large-scale evaluation on low-resource Turkic languages, including multi-script variants, extending multilingual model analysis beyond high-resource benchmarks; (iii) Label-free competence proxy based on internal model signals, combining tokenizer- and generation-level features without using ground-truth annotations.

2 Related Work

Our work builds on three lines of research: tokenizer-based multilingual analysis, uncertainty estimation for LLMs, and Turkic language evaluation.

[1]Code and data: https://github.com/IlyaGalyukshev/Multilingual-Uncertainty

Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop), pages 1074–1086

Tokenizer features and multilingual quality. Ali et al. (2024) find that common tokenizer metrics such as fertility are not always predictive of downstream performance. Ahia et al. (2023) show that tokenizer design introduces systematic cost disparities across languages. Tsvetkov and Kipnis (2024) propose information parity—the ratio of negative log-likelihood (NLL) on parallel corpora—which correlates with task performance but requires parallel text. Our method does not require parallel text, though our experimental design includes parallel script variants and matched EN/RU questions that serve as natural controls (§3.2). We find that fertility alone is a weak predictor of accuracy in our setting (§4.2). These approaches all require either parallel text or downstream task labels; we ask whether unlabeled processing signals suffice.

Uncertainty estimation for LLMs. Kuhn et al. (2023) introduce semantic entropy over sampled outputs. Manakul et al. (2023) propose SelfCheckGPT for hallucination detection via sample consistency. Lin et al. (2023) derive several uncertainty measures, including spectral graph methods, from response similarity matrices. Fadeeva et al. (2023) consolidate methods in LM-Polygraph; Ulmer et al. (2024) show that LLM confidence can be calibrated from generations alone, without requiring access to internal probabilities. These works focus on per-question or per-response uncertainty. Our approach differs in the aggregation level: we use the *cross-question standard deviation* of per-question uncertainty at the model–language level as the diagnostic signal. Tang et al. (2024) propose LAPE for language-specific neuron identification; Mondal et al. (2025) show these neurons do not reliably improve cross-lingual transfer. We compute LAPE but find it adds negligible predictive power (Appendix B).

Turkic language evaluation. TUMLU (Isbarov et al., 2025) provides natively authored multiple-choice questions for Turkic languages across 11 domains, with several languages available in multiple scripts. We use TUMLU as our primary benchmark and supplement it with English and Russian subsets from Global MMLU (Singh et al., 2025).

3 Experimental Setup

Our experimental pipeline has three stages: (1) we extract processing features from model responses to questions using a simple prompt, (2) we measure each model's accuracy on each language using a separate structured evaluation prompt, and (3) we measure how well these features—individually and in combination—predict accuracy. This section describes the models, data, and features.

3.1 Models

We evaluate 11 instruction-tuned models from five families: Llama (Grattafiori et al., 2024), Qwen (Team, 2024; Yang et al., 2025), Gemma (Team et al., 2024; Gemma Team et al., 2025), and GigaChat3 (Mamedov et al., 2025). The models range from 1B to 70B parameters (Table 1), providing diversity in both architecture and scale to test the robustness of our findings.

Model	Family	Params
Llama-3.2-1B	Llama	1B
Llama-3.1-8B	Llama	8B
Llama-3.3-70B	Llama	70B
Qwen2.5-3B/7B/14B	Qwen	3–14B
Qwen3.5-35B-A3B	Qwen	35B
Gemma-2-9B	Gemma-2	9B
Gemma-3-4B/12B	Gemma-3	4–12B
GigaChat3-10B	GigaChat3	10B

Table 1: Models evaluated. All are instruction-tuned. Qwen3.5 (3B active) and GigaChat3 (1.8B active) use MoE; the rest are dense.

Two models use MoE architectures: GigaChat3-10B (MoE with a dense first feed-forward network (FFN) layer) and Qwen3.5-35B (pure MoE in the FFN layers, with routed experts and a shared expert in all 40 layers). Gemma-2-9B is primarily English-centric and not positioned as multilingual. The remaining nine models are dense transformers.

3.2 Languages and Benchmarks

We use two benchmarks: TUMLU for Turkic languages and Global MMLU for English and Russian. Both consist of multiple-choice questions with four answer options. Table 2 summarizes the dataset statistics.

TUMLU questions are natively authored across domains including biology, chemistry, geography, history, mathematics, physics, and language/literature; the number of categories varies by language (5–10). TUMLU covers 9 Turkic languages including Kyrgyz; we exclude Kyrgyz because its questions were not available at the time of our experiments. EN/RU subsets of Global MMLU contain 2,157 identical questions in both languages, drawn from categories matching TUMLU domains.

Language	Abbr.	Script(s)	Source	#Q	#Cat
Azerbaijani	az	Lat	T	735	7
Crimean Tatar	ct-l, ct-c	Lat, Cyr	T	380	7
Karakalpak	kar	Lat	T	240	5
Kazakh	kk-c, kk-l	Cyr, Lat	T	944	10
Tatar	tat	Cyr	T	839	9
Turkish	tr	Lat	T	945	9
Uyghur	uy-a, uy-l	Arab, Lat	T	519	5
Uzbek	uz-l, uz-c	Lat, Cyr	T	735	7
English	en	Lat	G	2157	7
Russian	ru	Cyr	G	2157	7

Table 2: Languages and benchmarks. T = TUMLU; G = Global MMLU. #Q = questions; #Cat = subject categories. Languages with two scripts share identical (transliterated) questions.

Four Turkic languages appear in two scripts; within each pair, questions are identical (transliterated), so accuracy differences reflect purely model–script interactions. This gives us 14 language–script varieties in total.

3.3 Feature Extraction

We extract features in four groups: *tokenizer-based* (no model inference), *logit-based* and *attention-based* (one greedy pass per question), and *sampling-based* (10 stochastic samples per question). All features are computed from the same questions used for evaluation but with a simpler prompt: a one-sentence-answer instruction in the target language, without JSON formatting or system instructions (Appendix H). Token and question budgets are equalized across languages (Appendix D).

Tokenizer-based features are corpus-level statistics computed by running each model's tokenizer on formatted MCQ prompts (question, four answer choices, and instruction) *without* the chat template, truncated to a common token budget (the minimum total tokens across all 14 languages for a given model). The features are: fertility f (tokens per word), subword length mean and standard deviation (μ_c, σ_c), unique token fraction (distinct IDs divided by vocabulary size), and shared vocabulary fractions with English, Russian, and Turkish (fraction of target language's unique token IDs also appearing in the high-resource language's token set).

Uncertainty-based features are computed per question and aggregated as mean (μ) and standard deviation (σ) across questions. The same prompts are wrapped in the model's chat template and decoded greedily (max 64 new tokens). A forward pass yields per-token log-probabilities and entropy: mean token NLL ($\overline{\mathrm{NLL}} = -\frac{1}{n}\sum_i \log p(y_i \mid y_{<i}, x)$), sequence NLL, and mean token entropy $\bar{H} = \frac{1}{n}\sum_i H(p(\cdot \mid y_{<i}, x))$, where H is Shannon entropy. The cross-question σ_H captures how variable the model's uncertainty is across questions for a given language. We additionally compute two attention-based metrics (requiring eager attention implementation): RAUQ (Vazhentsev et al., 2025), which recurrently aggregates token probabilities and attention to preceding tokens across the middle-third layers ($\alpha = 0.2$), and Focus (Zhang et al., 2023), which scores keyword tokens (selected by IDF, adapted from the original spaCy-based approach) using probability corrections and attention-based penalty propagation ($\gamma = 0.9$).

Sampling-based features (Lin et al., 2023): 10 samples ($T_{\text{sample}} = 0.9$, $\text{top}_p = 0.95$) yield four uncertainty metrics—lexical similarity (negative mean pairwise $1 - \text{BLEU}$, with n-gram weights following LM-Polygraph; Fadeeva et al., 2023) and three graph-theoretic measures (DegMat, EigValLaplacian, Eccentricity) computed on a Jaccard word-overlap similarity graph (adapted from the original semantic similarity formulation). These do not improve the proxy at 10× the cost (§5).

3.4 Accuracy Evaluation

Accuracy is measured using a separate prompt format: a shared English system instruction asks the model to return a JSON object with a short chain-of-thought and a single-letter answer (A/B/C/D), while the user message uses a target-language JSON request (Appendix H). Answer choices are deterministically shuffled per question to control for position bias. The model generates greedily with `max_new_tokens=512`. Responses are parsed as JSON, with language-specific regex fallback; unparseable responses are scored as incorrect (Appendix G). To prevent overlap between features and the evaluation target, proxy regression is evaluated against accuracy on questions *not* used for feature extraction (§4).

4 Results

We present results in two parts: individual feature analysis and the combined proxy. Additional experiments (feature ablation, bridge-language effects, script effects) are in §5. All proxy regression metrics are reported against accuracy on *unseen* questions—those not used for feature extraction—

to ensure no overlap between features and the evaluation target. Descriptive statistics and feature correlations use total accuracy.

Accuracy across 11 models and 14 varieties ranges from 9.9% (Gemma-2-9B on Tatar, below the 25% random baseline) to 93.8% (Qwen3.5-35B on English), with mean Turkic accuracy of 41.0% compared to 75.7% for English and 66.2% for Russian. Qwen3.5-35B dominates: its lowest Turkic accuracy (Karakalpak, 60.0%) exceeds the highest Turkic accuracy of seven other models. Model size is a significant but incomplete predictor: Gemma-2-9B (mean Turkic 25.2%) underperforms the smaller Qwen2.5-3B (34.5%), likely reflecting English-centric training data rather than capacity. Script pairs reveal consistent patterns: Kazakh Cyrillic outperforms Latin in 9/11 models (mean gap +7.6 pp), while Uzbek and Crimean Tatar show smaller Latin advantages (§5). Across all languages, the accuracy gap between English and Turkic targets ranges from 8 pp (Qwen3.5-35B on Turkish) to 55 pp (Gemma-2-9B on Tatar), illustrating the wide variation our proxy aims to capture. The full accuracy table is in Appendix A.

4.1 Individual Feature Analysis

Table 3 shows the top 15 features ranked by the absolute value of their Pearson correlation with accuracy across all 154 model–language pairs. No single feature is a strong predictor: correlations range from $|r| = 0.38$ to 0.55.

Feature	Pearson r	Kendall τ
σ_c (subword length std)	+0.55	+0.34
Eccentricity mean	−0.55	−0.41
σ_H (entropy std)	−0.54	−0.39
μ_c (subword length mean)	+0.52	+0.29
Sequence NLL std	−0.49	−0.36
μ_H (entropy mean)	−0.49	−0.34
Eigenvalue Laplacian mean	+0.47	+0.33
Fertility (f)	−0.46	−0.26
Eigenvalue Laplacian std	−0.46	−0.33
Focus std	−0.42	−0.31
NLL std	−0.42	−0.31
NLL mean	−0.42	−0.30
Degree matrix mean	−0.41	−0.31
Lexical similarity mean	+0.39	+0.30
Unique token fraction	+0.38	+0.19

Table 3: Top 15 features by |Pearson r| with accuracy ($n = 154$ model–language pairs, total accuracy).

A consistent pattern emerges: for both tokenizer and uncertainty features, the *standard deviation* outperforms the corresponding *mean* as an accuracy predictor. At the tokenizer level, σ_c ($r = +0.55$) outranks μ_c ($r = +0.52$). At the generation level, σ_H ($r = -0.54$) outranks μ_H ($r = -0.49$).

Testing the inconsistency pattern. We compare dependent correlations using Williams' test (Williams, 1959). This test is designed for the situation where two predictors (here, σ_H and μ_H) are both correlated with the same target variable (accuracy) and with each other; it asks whether one predictor has a significantly stronger correlation with the target than the other. Pearson r measures linear association between two variables ($r = \pm 1$: perfect linear relationship; $r = 0$: no linear relationship). For each model family, we compute $|r(\sigma_H, \text{accuracy})|$ vs. $|r(\mu_H, \text{accuracy})|$ (Table 4):

Family	n	$\lvert r(\sigma_H)\rvert$	$\lvert r(\mu_H)\rvert$	p
Llama	42	0.82	0.77	.11
Qwen	56	0.63	0.49	.06
Gemma-2	14	0.52	0.44	.58
Gemma-3	**28**	**0.66**	**0.32**	**.01**
GigaChat3	14	0.75	0.73	.85

Table 4: Entropy standard deviation (σ_H) vs. mean (μ_H) as accuracy predictors, by model family (Pearson $|r|$, n = model–language pairs in family). Williams' test p-value for the difference. Bold: $p < 0.05$.

σ_H outperforms μ_H in all five families (Table 4). To test whether this pattern generalizes beyond entropy, we extend the comparison to all five greedy-pass measures—three logit-based (entropy, token NLL, sequence NLL) and two attention-based (RAUQ, Focus)—across all five families, giving 25 family–measure combinations. The standard deviation outperforms the mean (i.e., achieves higher $|r|$ with accuracy) in 21 of these 25 combinations. A binomial test gives $p < 0.001$, though this overstates significance because several measures are correlated (e.g., entropy and NLL share logit information); a conservative grouping into two independent clusters (logit-based and attention-based) still finds the pattern in 4 of 5 families (§8).

In contrast, for the four sampling-based graph metrics (lexical similarity, DegMat, EigValLaplacian, eccentricity; 10 samples per question), the standard deviation outperforms the mean in only 4 of 20 family–measure combinations—consistent with chance. This is expected: sampling already averages over multiple stochastic outputs per question, diluting the per-question inconsistency signal. Sampling features also add no predictive power to the proxy at 10× the computational cost (Ap-

pendix C).

Robustness to architecture. Qwen3.5-35B (pure MoE) exhibits inflated μ_H relative to Qwen2.5-14B on most languages, likely because MoE routing distributes probability across varied generation paths. Yet σ_H, averaged across all 14 languages, remains nearly identical between the two models, suggesting that σ_H is more robust to architectural effects that contaminate the mean. Since σ_H is computed from generated text, models with severe output truncation (high unparseable rates) might show artificially low σ_H; however, partial correlation controlling for unparseable rate confirms that σ_H retains a substantial association with accuracy (Appendix G).

4.2 Combined Proxy

Individual features correlate only moderately with accuracy. We now test whether combining features into a single proxy yields stronger prediction. The proxy uses three features selected by nested cross-validation from the eight strongest candidates:

- σ_c/f (tokenizer coverage ratio): the ratio of subword length variability to fertility. High values indicate well-represented languages with long, meaningful tokens and few tokens per word.
- σ_H (entropy variability): processing inconsistency, as described in §4.1.
- Mean EN/RU accuracy from Global MMLU: a model capability anchor, obtainable from a single evaluation run on a publicly available benchmark.

We combine these using Ridge regression, a linear model with regularization that prevents overfitting to small datasets. The regularization strength is selected separately for each feature configuration via generalized cross-validation (GCV; `sklearn.linear_model.RidgeCV` with $\alpha \in [10^{-2}, 10^{3}]$, 100 log-spaced candidates), which estimates leave-one-out error analytically without refitting.

Results. We evaluate the proxy in leave-one-language-out (LOO) mode: for each language, the proxy is trained on the remaining 13 languages and predicts accuracy on the held-out language. We report three metrics: R^2 (proportion of variance explained; 1.0 = perfect), Kendall τ (rank correlation; 1.0 = perfect ranking), and mean absolute error (MAE) in percentage points (pp). Table 5 shows results for all configurations.

Method	R^2	τ	MAE
Model-mean baseline	0.44	0.43	11.1 pp
Fertility only	0.09	0.23	15.0 pp
σ_H only	0.23	0.35	13.2 pp
Without EN/RU (3 features)	0.47	0.48	10.9 pp
Full proxy (3 features)	**0.75**	**0.71**	**7.7 pp**

Table 5: Proxy performance (LOO by language, accuracy on unseen questions, $n = 154$). Model-mean baseline predicts each pair using that model's mean accuracy on the remaining 13 languages. Without EN/RU uses σ_c, σ_H, and unique token fraction.

The full proxy explains 75% of accuracy variance ($R^2 = 0.75$, $\tau = 0.71$, MAE $= 7.7$ pp), nearly doubling the 44% achieved by the model-mean baseline. Bootstrap 95% CI: $R^2 \in [0.67, 0.80]$. Standardized coefficients show that the EN/RU anchor and the tokenizer coverage ratio contribute most ($+11.3$ and $+9.8$ pp per standard deviation, respectively), while σ_H contributes -2.3 pp. Removing σ_H and keeping only EN/RU accuracy and σ_c/f yields $R^2 = 0.73$, confirming that the tokenizer ratio carries most of the language-level signal. Figure 1 shows prediction errors across all 154 model–language pairs: the proxy systematically underpredicts Qwen3.5-35B (the strongest model) and tends to overpredict on the lowest-resource languages.

The anchor-free proxy (without EN/RU accuracy) captures *language coverage*—how well the tokenizer and representations handle a given language—exceeding the model-mean baseline ($R^2 = 0.47$ vs. 0.44), but it cannot distinguish *model capability*: it ranks languages within a model but not models against each other. Adding the EN/RU anchor resolves this: if we were to use each model's true mean accuracy across all languages (an idealized upper bound), the proxy would reach $R^2 = 0.81$; the EN/RU anchor nearly matches this because EN/RU accuracy correlates at $r = 0.94$ with model-mean accuracy. Table 6 shows per-model proxy diagnostics: the proxy works well for mid-range models (MAE < 8 pp) but systematically underpredicts Qwen3.5-35B (bias -9.8 pp) and overpredicts GigaChat3 (bias $+6.3$ pp).

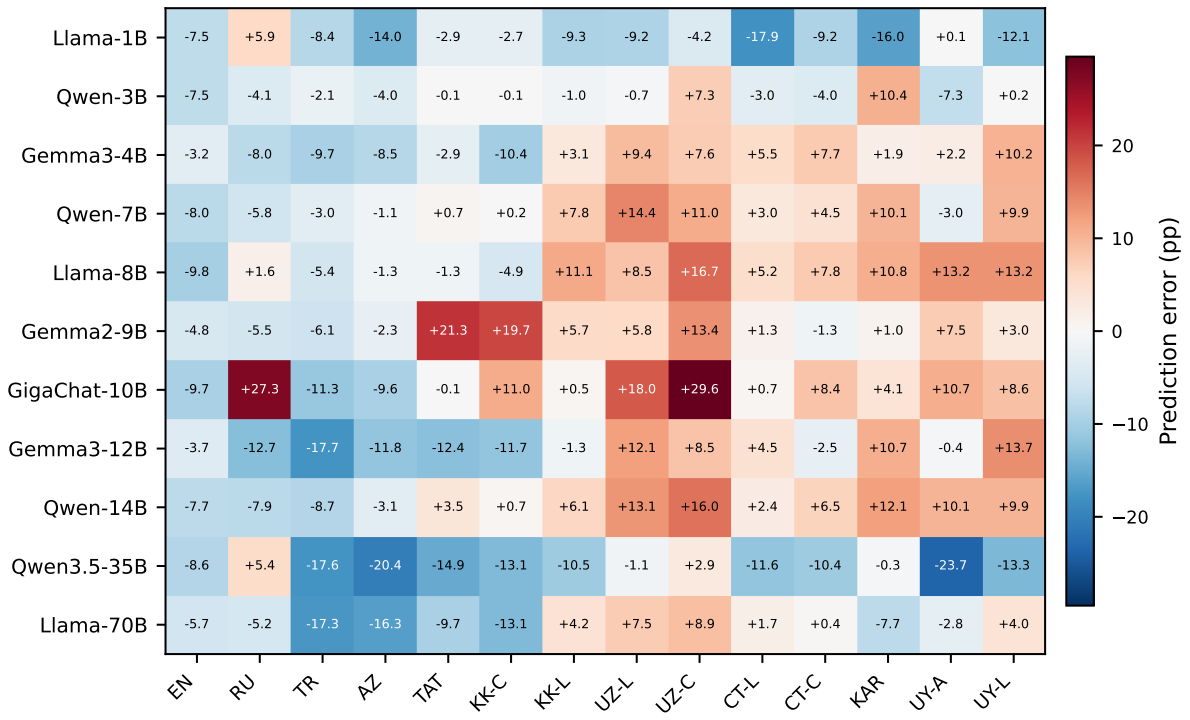

Figure 1: Proxy prediction error (predicted − actual accuracy, pp) across 11 models and 14 language–script varieties. Blue: underprediction; red: overprediction. The proxy systematically underpredicts Qwen3.5-35B and overpredicts on low-resource languages.

Model	MAE	τ	Bias
Qwen2.5-3B	3.7	0.47	−1.1
Qwen2.5-7B	5.9	0.43	+2.9
Gemma-3-4B	6.5	0.43	+0.3
Gemma-2-9B	7.1	0.56	+4.2
Llama-3.3-70B	7.5	0.36	−3.6
Qwen2.5-14B	7.7	0.52	+3.8
Llama-3.1-8B	7.9	0.71	+4.7
Llama-3.2-1B	8.5	0.21	−7.7
Gemma-3-12B	8.8	0.32	−1.8
GigaChat3-10B	10.7	0.36	+6.3
Qwen3.5-35B	**11.0**	**0.36**	**−9.8**

Table 6: Per-model proxy performance (MAE and bias in pp, τ = within-model language ranking). Full proxy, LOO by language, unseen accuracy.

4.3 Generalization

As Figure 2 shows, the proxy features extracted from ∼150 questions predict accuracy on held-out unseen questions (20–2,028 per language–model pair), using a different prompt format. The proxy–accuracy format mismatch (one-sentence prompt vs. JSON evaluation) provides additional evidence of robustness. The subset accuracy (measured on the ∼150 proxy questions) correlates at $r = 0.99$ with full-dataset accuracy, suggesting that ∼150 questions are sufficient for stable feature estimation. For the practical task of deciding which of two languages a model handles better, the proxy answers correctly in 72% of within-model pairwise comparisons (716/1,001 pairs). Cross-model generalization is analyzed in §5.

5 Ablation Studies

We analyze the contribution of individual proxy components and investigate additional factors that affect prediction.

Feature ablation. Table 5 shows how performance degrades as features are removed. Removing the EN/RU anchor (R^2: 0.75 → 0.47) eliminates the model capability signal: the anchor-free proxy ranks languages within a model but cannot compare across models, since EN/RU accuracy correlates at $r = 0.94$ with model-mean accuracy. Removing σ_H while keeping the anchor and σ_c/f yields $R^2 = 0.73$ ($\Delta R^2 = 0.02$), confirming that the tokenizer coverage ratio carries most of the language-level signal. Using fertility alone gives $R^2 = 0.09$—worse than the model-mean baseline ($R^2 = 0.44$)—consistent with prior findings that fertility is not reliably predictive of downstream performance (Ali et al., 2024). Extended baselines

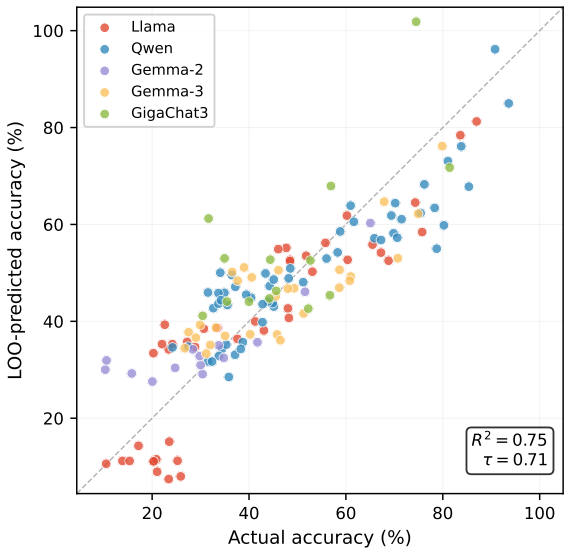

Figure 2: Predicted vs. actual accuracy (leave-one-language-out, 154 model–language pairs). Points colored by model family. $R^2 = 0.75$, $\tau = 0.71$.

in Appendix E.

Sampling features. Adding sampling-based graph metrics (10 samples per question at 10× the cost) does not improve LOO prediction: anchor-free R^2 remains at 0.47 with or without sampling features (Appendix C). MoE models show inflated sampling diversity independent of competence, likely because stochastic routing introduces variability unrelated to language knowledge.

Cross-model generalization. Leave-one-model-out evaluation ($R^2 = 0.55$, $\tau = 0.68$) is substantially weaker than leave-one-language-out ($R^2 = 0.75$). Notably, removing σ_H *improves* LOO-by-model to $R^2 = 0.67$, suggesting that σ_H captures model-family-specific patterns that do not transfer across architectures. The tokenizer feature σ_c/f combined with the EN/RU anchor is the most robust combination across both evaluation protocols.

Bridge language effect. Turkish vocabulary overlap (the fraction of shared tokens between Turkish and a target language) correlates positively with accuracy within the five Latin-script Turkic languages for all 11 models ($r = +0.69$ to $+0.99$; 11/11 positive, binomial $p < 0.001$). This is not a script-identity effect: English overlap (also Latin) correlates negatively ($r = -0.05$). The ranking (Azerbaijani > Crimean Tatar > Kazakh-Latin > Uyghur-Latin > Uzbek-Latin) matches known lexical proximity to Turkish. The effect is strongest for mid-range models (Qwen2.5-7B: $r = +0.99$) and weakest for the most capable model (Qwen3.5-35B: $r = +0.85$), suggesting that the pattern is most pronounced when a model relies on a dominant related language rather than having direct training data for each target. Full per-model correlations in Appendix F.

Script effects. For four Turkic languages available in two scripts, paired t-tests across 11 models reveal that Kazakh Cyrillic significantly outperforms Latin (+7.6 pp, $p = 0.038$), consistent with a Russian bridge: models with strong Russian benefit from shared Cyrillic subword tokens. Uzbek and Crimean Tatar show smaller, non-significant Latin advantages. Uyghur shows no script preference ($p = 0.90$), as Arabic script shares few tokens with any high-resource language. Figure 3 shows individual model differences; script pair statistics in Appendix F.

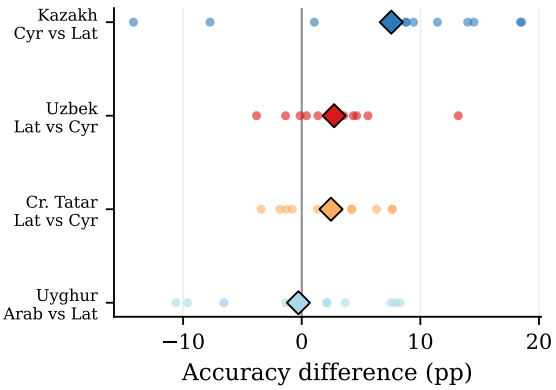

Figure 3: Script pair accuracy differences across 11 models. Each dot = one model; diamond = mean. Kazakh Cyrillic significantly outperforms Latin ($p = 0.038$).

6 Discussion

We discuss the implications of our findings, outline a practical protocol for applying the proxy, and identify directions for future work.

The inconsistency hypothesis. Our results support the hypothesis that cross-question *variability* of processing signals is a better diagnostic of language competence than their average level. The pattern is consistent across all five model families for greedy-pass measures (Table 4), and extends from generation entropy to tokenizer features (σ_c outranks μ_c; Table 3). For a well-covered language, the tokenizer segments most words into similarly-sized subwords; for a poorly covered language,

some words are fragmented into single characters while others are captured by borrowed tokens, producing high σ_c. This tokenizer-level inconsistency is why σ_c enters the final proxy rather than μ_c. However, individual significance tests are underpowered due to small per-family samples, and the 25 greedy-pass comparisons are not independent (§8). We view this as evidence of a consistent pattern rather than a definitive statistical proof; notably, nested cross-validation selects σ_H in all 14 LOO folds (Appendix E), and the signal requires only a single greedy pass—sampling adds nothing.

Practical protocol. (1) Compute σ_c/f on target-language MCQ prompts (no model inference needed). (2) Run greedy decoding on $\sim$130 of those questions (answer keys not used); compute σ_H. (3) Add the model's EN/RU accuracy from any existing benchmark. Without the EN/RU anchor, the proxy is useful for within-model language ranking but not cross-model comparison.

When is the proxy useful? The proxy is most valuable as a triage tool: before investing in expensive human evaluation or creating a new benchmark, a researcher can estimate a model's competence from existing MCQ questions without using answer keys. For language comparison within a single model, the anchor-free proxy suffices. For model comparison on a fixed language, the EN/RU anchor is needed but requires only a single additional evaluation run.

Connection to information parity. Tsvetkov and Kipnis (2024) show that NLL ratios on parallel corpora predict multilingual performance. Our quasi-parallel design—identical questions across script pairs, matched EN/RU categories—offers an indirect connection: σ_H computed on these quasi-parallel prompts captures similar information without requiring aligned sentence-level parallel text. A direct comparison on languages where both methods are applicable is a natural next step.

Future directions. Three extensions seem most promising: (a) testing the inconsistency hypothesis on typologically distant language families (Bantu, Dravidian, Austronesian) where bridge-language dynamics may differ; (b) exploring raw-text features (e.g., perplexity on Wikipedia snippets) to remove the MCQ dependency entirely; and (c) developing a capability proxy that requires no labeled data at all—our attempts using English entropy as a model-quality signal did not improve prediction, but architectural features (parameter count, vocabulary size) or cross-lingual transfer metrics may prove more informative.

7 Conclusion

We investigated whether internal processing signals of LLMs correlate with their accuracy on Turkic languages. Our experiments on 11 models across 14 language–script varieties show that the cross-question standard deviation of both tokenizer and uncertainty features is consistently more predictive than the mean, supporting the inconsistency hypothesis. A three-feature proxy achieves leave-one-language-out $R^2 = 0.75$, with the tokenizer coverage ratio contributing most of the language-level signal. These correlations are encouraging as a step toward estimating language competence without labeled benchmarks, though the current approach still relies on existing MCQ resources and a high-resource language anchor. We hope this work motivates further exploration of model-internal signals for low-resource language evaluation.

8 Limitations

We identify several limitations of the current study that should be considered when interpreting the results.

Two failure modes. The proxy has complementary weaknesses at both extremes. For the most capable model (Qwen3.5-35B), the EN/RU anchor still substantially underpredicts Turkic accuracy (Table 6). For models with severe output truncation (e.g., Gemma-2-9B on Tatar and Kazakh-Cyrillic, where over 60% of responses are unparseable; Table 11), σ_H is anomalously low because the truncated generation is too short for entropy variability to manifest. A capability proxy without any labeled data remains elusive: we tested several entropy-based model-quality signals, but none improved prediction.

Input requirements. Our features are extracted from MCQ prompts (without using answer keys), not from raw text. The current study relies on existing benchmarks (TUMLU, Global MMLU); for languages without any MCQ resource, automatic translation or question generation may introduce artifacts—a risk we have not tested.

Generalization. All evaluation uses TUMLU and Global MMLU. Whether the proxy transfers to other benchmarks, task formats, or typologically

distant language families is unknown and should not be assumed.

Statistical power. Although the aggregate inconsistency pattern is strong across greedy-pass measures, individual Williams tests reach significance only for one family (Gemma-3). The bridge-language analysis uses only five languages per model. The 25 greedy-pass comparisons are not fully independent, as several measures (entropy, NLL, sequence NLL) are strongly correlated. A conservative analysis grouping measures into two independent clusters finds the pattern in most but not all families, without reaching statistical significance.

Feature selection. The candidate features were pre-selected from a larger pool via exploratory correlation analysis. Pre-registration would strengthen claims.

Ethical Considerations

This work uses only publicly available benchmarks (TUMLU, Global MMLU) and open-weight models. No personal data is collected or generated. The proxy is intended as a diagnostic tool for researchers, not as a deployment-readiness criterion: a favorable score does not guarantee safe or reliable performance and should be supplemented with human evaluation.

Acknowledgments

The work was supported by the grant for research centers in the field of AI provided by the Ministry of Economic Development of the Russian Federation in accordance with the agreement 000000C313925P4F0002 and the agreement with Skoltech №139-10-2025-033.

References

Orevaoghene Ahia, Sachin Kumar, Hila Gonen, Jungo Kasai, David R Mortensen, Noah A Smith, and Yulia Tsvetkov. 2023. Do all languages cost the same? tokenization in the era of commercial language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9904–9923.

Mehdi Ali, Michael Fromm, Klaudia Thellmann, Richard Rutmann, Max Lübbering, Johannes Leveling, Katrin Klug, Jan Ebert, Niclas Doll, Jasper Buschhoff, and 1 others. 2024. Tokenizer choice for llm training: Negligible or crucial? In *Findings of the Association for Computational Linguistics: NAACL 2024*, pages 3907–3924.

Ekaterina Fadeeva, Roman Vashurin, Akim Tsvigun, Artem Vazhentsev, Sergey Petrakov, Kirill Fedyanin, Daniil Vasilev, Elizaveta Goncharova, Alexander Panchenko, Maxim Panov, and 1 others. 2023. Lm-polygraph: Uncertainty estimation for language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 446–461.

Gemma Team and 1 others. 2025. Gemma 3 technical report. *arXiv preprint arXiv:2503.19786*.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, and 1 others. 2024. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Jafar Isbarov, Arofat Akhundjanova, Mammad Hajili, Kavsar Huseynova, Dmitry Gaynullin, Anar Rzayev, Osman Tursun, Aizirek Turdubaeva, Ilshat Saetov, Rinat Kharisov, and 1 others. 2025. Tumlu: A unified and native language understanding benchmark for turkic languages. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 22816–22838.

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 the 58th annual meeting of the association for computational linguistics*, pages 6282–6293.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. 2023. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. *arXiv preprint arXiv:2302.09664*.

Zhen Lin, Shubhendu Trivedi, and Jimeng Sun. 2023. Generating with confidence: Uncertainty quantification for black-box large language models. *arXiv preprint arXiv:2305.19187*.

Valentin Mamedov, Evgenii Kosarev, Gregory Leleytner, Ilya Shchuckin, Valeriy Berezovskiy, Daniil Smirnov, Dmitry Kozlov, Sergei Averkiev, Lukyanenko Ivan, Aleksandr Proshunin, and 1 others. 2025. Gigachat family: Efficient russian language modeling through mixture of experts architecture. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 3: System Demonstrations)*, pages 93–106.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Soumen Kumar Mondal, Sayambhu Sen, Abhishek Singhania, and Preethi Jyothi. 2025. Language-specific neurons do not facilitate cross-lingual transfer. In *The Sixth Workshop on Insights from Negative Results in NLP*, pages 46–62.

Shivalika Singh, Angelika Romanou, Clémentine Fourrier, David Ifeoluwa Adelani, Jian Gang Ngui, Daniel Vila-Suero, Peerat Limkonchotiwat, Kelly Marchisio, Wei Qi Leong, Yosephine Susanto, and 1 others. 2025. Global mmlu: Understanding and addressing cultural and linguistic biases in multilingual evaluation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 18761–18799.

Tianyi Tang, Wenyang Luo, Haoyang Huang, Dongdong Zhang, Xiaolei Wang, Wayne Xin Zhao, Furu Wei, and Ji-Rong Wen. 2024. Language-specific neurons: The key to multilingual capabilities in large language models. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5701–5715.

Gemma Team, Morgane Riviere, Shreya Pathak, Pier Giuseppe Sessa, Cassidy Hardin, Surya Bhupatiraju, Léonard Hussenot, Thomas Mesnard, Bobak Shahriari, Alexandre Ramé, and 1 others. 2024. Gemma 2: Improving open language models at a practical size. *arXiv preprint arXiv:2408.00118*.

Qwen Team. 2024. Qwen2.5: A party of foundation models.

Alexander Tsvetkov and Alon Kipnis. 2024. Information parity: Measuring and predicting the multilingual capabilities of language models. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7971–7989.

Dennis Ulmer, Martin Gubri, Hwaran Lee, Sangdoo Yun, and Seong Oh. 2024. Calibrating large language models using their generations only. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15440–15459.

Artem Vazhentsev, Lyudmila Rvanova, Gleb Kuzmin, Ekaterina Fadeeva, Ivan Lazichny, Alexander Panchenko, Maxim Panov, Timothy Baldwin, Mrinmaya Sachan, Preslav Nakov, and 1 others. 2025. Uncertainty-aware attention heads: Efficient unsupervised uncertainty quantification for llms. *arXiv preprint arXiv:2505.20045*.

Evan J Williams. 1959. The comparison of regression variables. *Journal of the Royal Statistical Society: Series B (Methodological)*, 21(2):396–399.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, and 1 others. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Tianhang Zhang, Lin Qiu, Qipeng Guo, Cheng Deng, Yue Zhang, Zheng Zhang, Chenghu Zhou, Xinbing Wang, and Luoyi Fu. 2023. Enhancing uncertainty-based hallucination detection with stronger focus. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 915–932.

A Full Accuracy Table

Table 7 provides the complete accuracy matrix referenced in §4.

Model	az	ct-l	ct-c	en	kk-c	kk-l	kar	ru	tat	tr	uy-a	uy-l	uz-l	uz-c
Llama-3.2-1B	25.3	23.2	18.9	48.6	13.8	21.5	23.8	28.8	16.7	23.5	10.8	21.4	19.6	15.2
Qwen2.5-3B	37.3	37.4	38.7	69.3	34.3	33.3	26.7	58.1	35.8	45.5	33.9	31.8	32.4	26.8
Gemma-3-4B	46.0	32.9	28.7	68.3	46.6	32.1	33.8	58.8	40.8	52.5	29.3	27.2	30.2	29.8
Qwen2.5-7B	44.1	42.1	40.8	81.0	44.1	35.3	32.5	69.9	42.5	56.4	40.7	32.8	30.6	34.4
Llama-3.1-8B	38.0	34.0	30.8	74.5	43.3	24.8	24.6	60.3	41.5	48.2	20.0	21.4	27.1	23.5
Gemma-2-9B	34.6	33.7	27.4	65.0	10.7	24.9	25.8	51.5	9.9	42.5	19.9	29.5	28.6	15.4
GigaChat3-10B	52.5	45.3	37.6	81.6	58.2	44.2	40.8	74.9	54.2	56.4	28.7	35.3	36.2	31.6
Gemma-3-12B	62.7	44.5	46.3	80.0	59.5	48.1	41.2	74.8	62.1	70.6	43.7	35.5	40.0	41.4
Qwen2.5-14B	51.6	48.4	40.8	83.7	49.6	40.1	37.1	76.2	46.0	66.0	30.6	37.2	37.3	35.0
Qwen3.5-35B	**79.2**	**71.8**	**72.6**	**93.8**	**76.4**	**67.6**	**60.0**	**90.9**	**78.3**	**85.7**	**78.4**	**70.9**	**60.7**	**60.8**
Llama-3.3-70B	69.1	53.2	56.6	87.2	67.5	49.0	56.2	83.9	66.4	75.9	50.3	46.6	47.8	46.4

Table 7: Total accuracy (%) across all 154 model–language pairs. Models sorted by parameter count.

B LAPE Analysis

Following Tang et al. (2024), we identify language-specific neurons via LAPE (Language-specific Activation Probability Entropy). For each FFN layer, we hook the gate projection and record which neurons activate (> 0) across all tokens in the truncated corpus (using the same prompts as tokenizer features, without chat template). For each neuron, the activation probability is computed per language and normalized; the entropy of this distribution measures language specificity (low entropy = neuron activates preferentially for specific languages). We select the top 1% lowest-entropy neurons that pass an activation frequency filter. The number of language-specific neurons (as a fraction of total neurons) serves as the LAPE feature.

Globally, LAPE has near-zero correlation with accuracy ($r = -0.11$, $n = 140$ excluding Qwen3.5-35B). Per-model correlations are heterogeneous: Llama-3.2-1B ($r = -0.60$, $p = 0.02$), Qwen2.5-14B ($r = -0.60$, $p = 0.02$) are negative; GigaChat3-10B trends positive ($r = +0.53$, $p = 0.05$). LAPE adds $\Delta R^2 < 0.001$ to the proxy.

Qwen3.5-35B reports identical LAPE (1.001%) for all 14 languages—a methodological artifact: all 40 FFN layers use MoE, and only the shared expert `gate_proj` is hooked, yielding language-uniform activations. GigaChat3-10B avoids this because its first layer uses a dense FFN.

C Sampling Features

For each question, we generate 10 samples ($T_{\text{sample}} = 0.9$, $\text{top}_p = 0.95$) and compute four uncertainty metrics (Lin et al., 2023). Lexical similarity is the negative mean pairwise $(1 - \text{BLEU})$ across samples, using n-gram weights following LM-Polygraph (Fadeeva et al., 2023). The remaining three metrics (DegMat, EigValLaplacian, Eccentricity) are computed on a pairwise Jaccard word-overlap similarity graph (see §3.3 for definitions). In-sample R^2 increases from 0.64 to 0.67 when adding sampling features, but LOO-by-language R^2 does not improve (best anchor-free $R^2 = 0.47$ with or without sampling features). MoE models show inflated sampling diversity independent of competence.

D Budget Equalization

Token budget (tokenizer features): formatted MCQ prompts without chat template, truncated to $\min(\text{total\_tokens})$ across 14 languages per model. **Question budget** (uncertainty features): same prompts wrapped in chat template, equalized to the minimum number of questions fitting within the token budget (129–220 per model). **Evaluation**: full datasets, JSON-structured prompt, `max_new_tokens` $= 512$.

E Per-Model Results and Baselines

Nested CV ($n = 154$). 8 candidates (pre-selected by $|r| > 0.2$ with accuracy, pairwise $|r| < 0.8$): fertility, σ_c, σ_H, eccentricity mean, eccentricity infs count, unique token fraction, sequence NLL std,

RAUQ std. 10/14 outer folds select σ_c, σ_H, eccentricity mean; 4/14 select σ_c, σ_H, unique token fraction. Both triplets give identical LOO-by-language $R^2 = 0.47$. We report the latter because eccentricity is a sampling-based feature, while unique token fraction is a tokenizer statistic (no GPU cost)—making the anchor-free proxy fully inference-free for its tokenizer component.

Extended baselines (Table 8; extends Table 5 from §4.2). The model-mean baseline predicts each model–language pair using that model's mean accuracy on the remaining 13 languages (leave-one-language-out). The row "EN/RU + σ_c/f" shows that σ_H adds only $\Delta R^2 = 0.02$ over the two-feature anchor proxy, indicating that the tokenizer coverage ratio carries most of the language-level signal.

Method	R^2	τ	**MAE**
Model-mean (by language)	0.44	0.43	11.1 pp
Fertility only	0.09	0.23	15.0 pp
Eccentricity only	0.24	0.36	13.6 pp
σ_H only	0.23	0.35	13.2 pp
σ_H + unique frac	0.36	0.45	12.0 pp
Without EN/RU (3 feat.)	0.47	0.48	10.9 pp
EN/RU + σ_c/f	0.73	0.67	8.1 pp
Full proxy	**0.75**	**0.71**	**7.7 pp**

Table 8: Extended baselines (LOO by language, unseen accuracy, $n = 154$).

Leave-one-model-out. As a stricter generalization test, we hold out all 14 language–script varieties for one model and train on the remaining 10. This tests whether the proxy transfers to an entirely unseen model family. Results: full proxy $R^2 = 0.55$, $\tau = 0.68$, MAE $= 9.6$ pp. Notably, removing σ_H improves LOO-by-model to $R^2 = 0.67$, suggesting that σ_H captures model-family-specific patterns that do not transfer. The tokenizer feature σ_c/f combined with the EN/RU anchor is the most robust combination across both LOO protocols. LOO-by-language ($R^2 = 0.75$) remains the appropriate evaluation for the typical use case: estimating a *known* model's competence on a *new* language.

F Bridge Language Details

Table 9 reports per-model Turkish vocabulary overlap correlations (referenced in §5). Table 10 reports script pair differences.

Model	r	**min (4/5)**	**max (4/5)**
Qwen2.5-7B	+0.99	+0.97	+1.00
Qwen2.5-14B	+0.97	+0.95	+1.00
Gemma-2-9B	+0.97	+0.96	+1.00
Qwen2.5-3B	+0.95	+0.92	+0.99
Qwen3.5-35B	+0.85	+0.73	+0.93
Llama-3.2-1B	+0.83	+0.74	+0.93
Llama-3.1-8B	+0.81	+0.69	+0.97
Gemma-3-4B	+0.77	+0.42	+0.91
GigaChat3-10B	+0.76	+0.73	+0.89
Llama-3.3-70B	+0.71	+0.67	+0.87
Gemma-3-12B	+0.69	+0.16	+0.83

Table 9: Turkish vocabulary overlap vs. accuracy within 5 Latin-script Turkic languages (Pearson r, total accuracy). min/max (4/5): range of r when each language is excluded in turn.

G Unparseable Response Rates

Table 11 shows unparseable rates by model (referenced in §4.1 and §8). Responses that cannot be parsed into A/B/C/D are counted as incorrect (§3.4).

Pair	Winner	Wins	Δ	p
Kazakh Cyr/Lat	Cyr	9/11	+7.6 pp	**.038**
Uzbek Lat/Cyr	Lat	8/11	+2.7 pp	.070
CT Lat/Cyr	Lat	7/11	+2.5 pp	.064
Uyghur Arab/Lat	Arab	6/11	−0.3 pp	.898

Table 10: Script pair accuracy differences (paired t-test, two-tailed, across 11 models, total accuracy).

Model	Mean %	Range
Qwen2.5-7B	1.3	0.0–3.4
Qwen2.5-3B	1.8	0.1–5.4
Llama-3.3-70B	4.0	0.8–8.5
Qwen3.5-35B	5.0	1.0–12.5
Gemma-3-12B	7.0	2.4–13.2
Gemma-3-4B	7.3	1.8–15.5
Qwen2.5-14B	7.6	0.6–17.9
GigaChat3-10B	13.1	2.3–32.1
Gemma-2-9B	16.3	0.4–62.9
Llama-3.1-8B	16.9	1.3–38.7
Llama-3.2-1B	30.0	5.9–63.6

Table 11: Unparseable response rates (%) by model, sorted by mean. Range = min–max across 14 languages.

H Prompt Templates

Feature extraction prompts (§3.3) use a one-sentence-answer instruction in the target language (e.g., English: "Answer in one sentence.", Turkish: "Cevabı tek bir cümleyle yazın."). Each prompt follows the format: "`[Question label]: {question} / {choices} / [instruction]`", where labels and instructions are in the target language. Tokenizer features use these prompts without the chat template; logit-based, attention-based, and sampling-based features wrap them in the model's chat template.

Evaluation prompts (§3.4) use a different format. A shared English system instruction precedes each prompt:

```
Answer the following multiple choice question. Provide your response as a JSON object with
two fields:
- "reasoning": your step-by-step reasoning process (from 1 to 3 sentences)
- "answer": a single letter (A, B, C, or D) representing your final answer

RETURN ONLY JSON (NO PROSE, NO MARKDOWN, NO COMMENTS, NO EXTRA WORDS)
Example format:
{"reasoning": "Let me analyze each option...", "answer": "B"}
```

The per-language user message replaces the one-sentence instruction with a target-language JSON request (e.g., "JSON formatında cavab verin:" for Azerbaijani). Full templates for all 14 varieties are in the released code.