

A Multimodal Corpus for Natural Language Understanding in Russian and Tatar

Alsu Khabirzyanova^{*1}, Zaryana Damashova^{*1}, and Ilseyar Alimova²

¹HSE University ²Skoltech

afkhabirzyanova1@gmail.com, z10damashova@gmail.com,
alimovailseyar@gmail.com

Abstract

Task-oriented dialogue systems rely on natural language understanding (NLU) to map utterances to intents and slot values. Although voice assistants are widely deployed, Russian and — especially — Tatar remain low-resource for NLU: annotated corpora are scarce and manual cultural adaptation is costly.

To address this gap, we construct parallel Russian and Tatar NLU resources based on the xSID benchmark: test and validation splits are manually translated and culturally adapted; the training set is machine-translated with large language models (LLMs) and manually corrected. We additionally record native-speaker audio for both languages and evaluate spoken NLU through automatic speech recognition (ASR)→NLU.

In-language training substantially outperforms English-only fine-tuning; typologically close pivot languages transfer better than distant ones on slot filling; and instruction-tuned generative models achieve strong intent scores but trail fine-tuned encoders on slot extraction.¹

1 Introduction

Natural language understanding (NLU) is a core component of task-oriented dialogue (ToD) systems: given a user utterance, the system must detect the intent and fill slots with argument values. The intent denotes the overall goal of the utterance; slots denote the argument spans that specify what is being requested. Figure 1 illustrates this annotation scheme with Russian and Tatar examples from our corpora. Voice assistants are now commonplace, yet most NLU benchmarks and pretrained models remain English-centric (Razumovskaia et al., 2022; Hu et al., 2023b). Russian is under-resourced relative to English; Tatar – a Turkic language spoken by

Lang.	Language Family	Annotation
ru	Indo-European	Сегодня будет дождь?
tt	Turkic	Бүген яңгыр явармы?
en	Indo-European	Is it going to rain today?

intent: weather/find
slots: datetime, weather/attribute

Figure 1: Examples of annotation for Russian and Tatar in our corpora with intent `weather/find` and slots `datetime` and `weather/attribute`.

more than five million people in Russia – lacks publicly available ToD-NLU benchmarks altogether. Building annotated corpora is expensive: machine-translated data introduces *translationese* and slot-boundary errors, while native collection or professional translation scales poorly (Razumovskaia et al., 2022; Majewska et al., 2023).

We present two new corpora for Russian and Tatar NLU, consisting of text and speech data and derived from xSID-0.7 (Krückl et al., 2025; van der Goot et al., 2021a) through a common processing pipeline. Both languages follow the same annotation schema (18 intents, 44 slots, BIO tagging), the same train/validation/test design (translated vs. culturally adapted variants), and the same evaluation protocol. The English training pool (37,173 utterances after deduplication) is machine-translated and automatically adapted; test and validation are produced manually. Seven Russian and nine Tatar non-professional speakers record the adapted benchmark on consumer microphones in everyday acoustic conditions; transcriptions feed the same NLU models as gold text.

Our contributions:

1. Two new text NLU corpora for Russian and Tatar with manually curated benchmarks and scalable training data.

^{*}Equal contribution.

¹Corpora and code are publicly available at:
<https://github.com/zaryana223/Text-and-Audio-Corpora-for-Natural-Language-Understanding-in-Russian-and-Tatar>

2. Two new audio NLU corpora for speech-based information extraction.
3. A systematic study of cross-lingual transfer with pivot languages, including symmetric Russian—Tatar experiments.
4. An evaluation of instruction-tuned generative LLMs alongside fine-tuned encoders in zero-shot and few-shot regimes.

2 Related Work

Intent detection and slot filling are core ToD-NLU subtasks (Ni et al., 2023). Despite the breadth of language diversity, only a small fraction of the world’s languages are represented in NLP research (Joshi et al., 2020); Russian and Tatar are no exception to this pattern. Commercial voice-command datasets such as SNIPS (Coucke et al., 2018) cover single languages; multilingual benchmarks followed to study cross-lingual transfer (Schuster et al., 2019), and surveys document the persistent gap between English and other languages (Razumovskaia et al., 2022).

A common recipe translates an English corpus and fine-tunes a multilingual encoder (Schuster et al., 2019), but MT pipelines are brittle for slot filling due to *translationese* and boundary errors (Winkler et al., 2024). Native-speaker collection avoids these artefacts but scales poorly: outline-based approaches (Majewska et al., 2023) and culturally adapted multi-parallel datasets (Hu et al., 2023a) address this at the cost of annotation effort. Translated and naturally phrased test sets yield systematically different scores (Winkler et al., 2024), and dialectal in-the-wild corpora reveal further gaps (Mæhlum and Scherrer, 2024). Performance disparities between languages also reflect tokenization and model architecture, not data volume alone (Ahia et al., 2023; Hu et al., 2023b).

xSID (van der Goot et al., 2021a) introduced a cross-lingual slot and intent detection benchmark across 13 languages built via expert translation from English, exposing a recurring challenge: translated evaluation data may not reflect natural target-language usage. Auxiliary non-English tasks partially close the cross-lingual gap, with masked LM proving most effective for slot filling; MaChAmp supports such multi-task training (van der Goot et al., 2021b). xSID-0.7 (Krückl et al., 2025) revisits the English training pool and label schema; related work further extends the benchmark with naturally collected Bavarian and Lithua-

nian queries (Winkler et al., 2024).

Generative LLMs perform intent detection via in-context learning and chain-of-thought prompting, competing with fine-tuned encoders on some benchmarks (Arora et al., 2024). Even on established ToD benchmarks, instruction-tuned LLMs underperform fine-tuned models on explicit slot tracking while benefiting from few-shot in-domain examples (Hudeček and Dušek, 2023). ILLU-MINER frames intent classification and slot filling as one generation task for instruction-tuned models and improves slot F1 over encoder baselines (Mirza et al., 2024). Generative models also target multi-intent joint detection and slot filling (Li and Zhu, 2026).

Broader Russian NLU evaluation is anchored by RussianSuperGLUE (Shavrina et al., 2020) and contextual benchmarks that expose shallow modelling (Chirkin et al., 2025); colloquial perturbations hurt slot F1 more than intent accuracy even in related high-resource settings (Artemova et al., 2024). SLURP packages complex English spoken NLU across 18 domains with ASR cascade evaluation (Bastianelli et al., 2020). The Russian audio subset from our corpora was later integrated into a broader multimodal Russian evaluation suite (Chervyakov et al., 2026).

3 Corpora

3.1 Shared pipeline

Figure 2 summarizes the construction pipeline of the **test and validation** splits, which is **identical** for Russian and Tatar. Starting from the English xSID-0.7 test and validation releases (Krückl et al., 2025), we (1) manually translate utterances; (2) produce culturally adapted variants; (3) record the adapted benchmark with native speakers on consumer microphones; and (4) transcribe the audio with language-specific ASR. Gold manual text and ASR transcriptions are evaluated with the same NLU models and metrics.

The **translated** split is a manual calque of the English benchmark: wording and slot boundaries follow the source, and many named entities remain foreign (e.g., US cities or Western media titles).

The **adapted** split localises the same intents for Russian or Tatar users: we replace slot values with culturally familiar analogues and rephrase surrounding text where a literal translation would sound unnatural (Majewska et al., 2023; Razumovskaia et al., 2022). For each

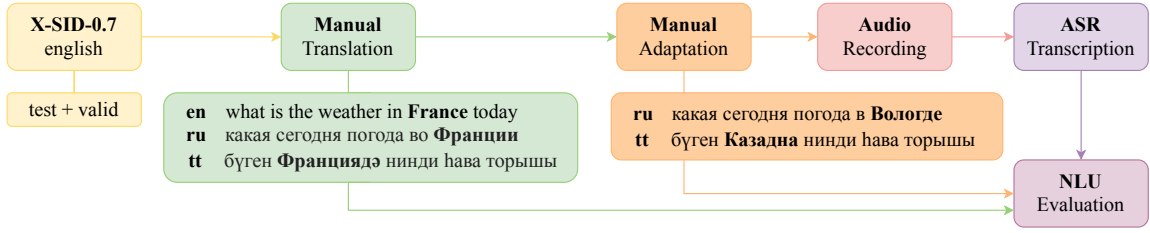


Figure 2: Unified corpus construction pipeline (identical for Russian and Tatar).

slot category we compiled candidate replacement lists: `location` (cities, metro stations, districts), `restaurant_name`, `cuisine`, and `served_dish`; `artist`, `album`, `playlist`, and `service for music`; `movie_name` and `cinema location for screenings`; and `object_name` (books, games, series, paintings).

Candidates were collected from Wikipedia, Yandex Music, and Google search and chosen manually by utterance semantics and approximate span length, so the original task and slot size stay comparable (Russian: nationwide entities; Tatar: Tatarstan toponyms, local restaurants/cinemas, and regional media). Both variants share the same intent labels and BIO schema; only surface forms and entity choices differ. Test and validation adaptation is fully manual; the training set uses the same replacement policy via an automatically built lexicon.

The **training** set (37,173 utterances after deduplication) is built separately at scale. We piloted GPT-5, DeepSeek, and Gemini-2.5-Flash on held-out examples and selected DeepSeek as the primary translator based on output quality and accessibility. Separate Russian and Tatar prompts enforce CoNLL-only output, entity-handling conventions, and the annotation rules in §3.3: output format; transliteration and title adaptation; Russian tense and *my* normalisation in alarm/reminder contexts; and Tatar possessive suffixes, temporal case marking, and slot-contiguity constraints under agglutinative word order. Translated outputs underwent pattern-based BIO correction, manual correction of remaining errors, and automatic entity adaptation via the replacement lexicon. Full prompt specifications are given in Appendix B (§B.1–B.2).

3.2 Base schema and statistics

Both corpora inherit the xSID-0.7 inventory: 18 intents and 44 slot types with BIO annotation. In

Split	RU	TT
Test (trans.)	532	500
Test (adapt.)	500	500
Val. (trans.)	300	300
Val. (adapt.)	300	300
Training (dedup., trans. + adapt.): 37,173		
Audio (test/val.)	500/300	500/300
Speakers	7	9

Table 1: Corpus statistics for Russian (RU) and Tatar (TT) NLU resources.

the English training pool, `weather/find` and `datetime` dominate, which skews per-class intent scores.

Each language provides translated and adapted test and validation sets; Table 1 summarises corpus sizes and audio coverage.

Table 2 compares mean BIO entity length (tokens per span) on the manually annotated benchmarks. Translated Russian and Tatar spans are almost the same length (1.69 vs. 1.70 tokens on pooled test and validation). On the adapted **test** sets (500 utterances each), Tatar spans are shorter than Russian (1.51 vs. 1.59; $\Delta = -0.08$): cultural adaptation replaces long foreign names with compact local forms in Russian, whereas Tatar often realises the same slots through shorter adapted lemmas without extra function words. Agglutinative morphology (possessive suffixes on alarms/reminders, multi-token `datetime` expressions) would be expected to widen Tatar spans, yet cultural adaptation to shorter local forms dominates, yielding shorter adapted spans overall.

3.3 Language-specific annotation features

Russian:

- English *my* in alarm/reminder contexts is rendered as *all*, *current*, or left as *my*, depending on number and naturalness.

Manual benchmark	Russian	Tatar	Δ
Translated (test + val)	1.69	1.70	+0.02
Adapted (test + val)	1.59	1.51	-0.08

Table 2: Mean BIO slot span length (tokens per entity) on manually annotated Russian and Tatar benchmarks. Δ is Tatar minus Russian. Translated and adapted rows pool test and validation.

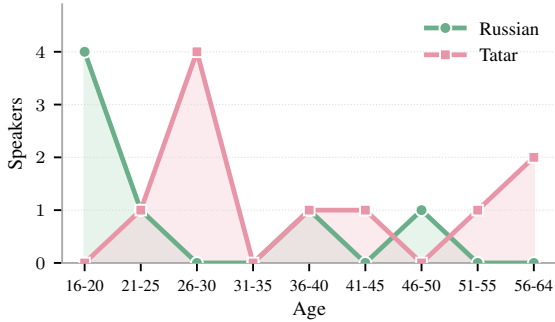


Figure 3: Speaker age distribution in the Russian and Tatar audio corpora.

- *Present continuous* queries are rendered as a progressive for ongoing events, or as a future form when a temporal offset is explicit (e.g., *in 15 seconds*).
- Early-morning clock times (*I am-3 am*) are normalised as “*N hour(s) at night*” (Russian: *N chas(a) nochi*) inside `datetime` spans.

Tatar:

- Stacked possessive constructions such as *miña minem* (lit. ‘to-me my’) are avoided, since this redundant doubling of the possessive — a literal artefact of word-for-word translation from English — does not occur in natural Tatar speech.
- When English has one contiguous slot but Tatar word order would break it, annotators rephrase the utterance to keep a single span where possible.
- Loanwords are written following Tatar orthographic conventions.

3.4 Audio recording

Non-professional speakers (seven Russian, nine Tatar) record adapted audio on smartphone microphones in everyday conditions; Figure 3 shows the age distribution. One Russian validation subset (150 utterances, one speaker) was used to select the ASR model for the full test set.

4 Experiments

4.1 Models

We evaluate three model families under a common experimental setup (Section 4.2): encoder fine-tuning on English, translated, and culturally adapted data; cross-lingual pivot transfer from eleven xSID languages; ASR→NLU cascades on the adapted audio test; and zero- and few-shot generative prompting.

4.1.1 Text encoders

We fine-tune six multilingual encoders with the MaChAmp framework (van der Goot et al., 2021b) on joint intent classification and BIO slot labelling. **Russian:** mBERT (Devlin et al., 2019) (104 languages; MLM+NSP), RemBERT (Chung et al., 2020) (110 languages; MLM), mDeBERTa-v3-base (He et al., 2021) (CC100, 16 languages including Russian), EuroBERT-210m/610m (Boizard et al., 2025) (15 European languages; balanced pretraining), and mmBERT (Marone et al., 2025) (1,800+ languages); pretraining is English-dominated except EuroBERT. **Tatar:** the same family with Glot500-base (Imani et al., 2023) instead of EuroBERT — XLM-R continued-pretrained on Glot500-c (~500 predominantly low-resource languages, including Turkic); EuroBERT is optionally reported as a no-Tatar-pretraining baseline.

4.1.2 Generative LLMs

Six instruction-tuned models (2B–9B) are evaluated without fine-tuning: Qwen2.5-3B/7B-Instruct, Gemma-2-2B/9B-it, Phi-4-mini-instruct, and Mistral-7B-Instruct-v0.3.

4.1.3 ASR models

Russian: Vosk (vosk-model-ru-0.42) and GigaAM v3 RNNT (Kutsakov et al., 2025); GigaAM is used on the test set (validation WER/CER 0.1225/0.0866 vs. 0.1569/0.0982 for Vosk).

Tatar: Söyle, a noise-robust multilingual ASR for Turkic languages (TatSC) (Mussakhojayeva et al., 2024).

4.2 Experimental Setup

4.2.1 Metrics

Following xSID (van der Goot et al., 2021a), we report **intent accuracy** (I_{acc}) and span-level slot F1 (seqeval), as in the original benchmark tables.

For compact comparison we report an averaged score (**Avg.**), by analogy with the xSID summary

columns (van der Goot et al., 2021a):

$$\text{Avg.} = \frac{I_{\text{acc}} + \text{Slot F1}}{2}. \quad (1)$$

For the spoken Russian and Tatar pipeline, where ASR output length may differ from gold, Span F1 in Eq. (1) is replaced by the mean per-utterance BIO F1 over 500 adapted test sentences:

$$\text{Slot F1}_{500} = \frac{1}{500} \sum_{i=1}^{500} \text{BIO F1}(u_i), \quad (2)$$

where u_i is the i -th test utterance and $\text{BIO F1}(u_i)$ is the weighted token-level F1 over BIO tags when gold and hypothesised sequences have equal length, and 0 otherwise.

4.3 Training configurations

We fine-tune six multilingual encoders for Russian and Tatar under six configurations: training on the English xSID corpus, in-language manually translated data, or in-language culturally adapted data, crossed with evaluation on the manually translated or adapted test benchmark. Generative models are evaluated only on the manually translated test sets.

On validation, mDeBERTa-v3-base outperformed the other encoders when comparing English, Russian, and Tatar runs and when fine-tuning from Arabic xSID data; we use it for all further experiments below.

mDeBERTa-v3-base is fine-tuned on each xSID pivot language — Arabic, Danish, German, Indonesian, Italian, Japanese, Kazakh, Dutch, Serbian, Turkish, and Mandarin Chinese — and evaluated on Russian and Tatar test sets (translated and adapted). We additionally run symmetric Russian – Tatar transfer in both directions: Tatar translated and adapted training evaluated on Russian translated and adapted tests, and Russian translated and adapted training evaluated on the corresponding Tatar tests.

We record the adapted test set and transcribe it using GigaAM v3 RNNT (Russian) and Söyle (Tatar). The adapted training configuration is selected because the audio content reflects culturally localised entities; using translated training data would introduce a distributional mismatch between training and test domains. The resulting transcriptions are passed to mDeBERTa-v3-base fine-tuned on the adapted training and validation splits for joint intent and slot prediction. Since ASR systems do not produce punctuation, comparing raw

transcriptions against the original gold test would artificially penalise slot metrics — punctuation tokens within slot spans would be counted as missing words.

We therefore remove punctuation from the gold test; slot boundaries are not affected. NLU metrics are computed against this normalised reference.

4.3.1 Generative evaluation

Zero-shot: no in-domain examples in the prompt.

Few-shot: 1, 5, or 10 in-domain examples; the single-example condition uses one demonstration from *weather/find* (most frequent), *SearchScreeningEvent* (error-prone), or *BookRestaurant* (slot-rich). The schema prompts used at inference are given in Appendix B (§B.3–B.4).

5 Results

5.1 Model comparison

Russian: Switching from English to Russian training raises Avg. by roughly 15–20 percentage points for mBERT, RemBERT, mDeBERTa, and mmBERT; EuroBERT gains even more (0.185–0.269→0.880–0.911). Intent accuracy exceeds Slot F_1 in every setup, and translated versus adapted training differs little for most encoders. mDeBERTa is the strongest model (Avg. 0.933); EuroBERT-610M is weakest under English-only fine-tuning.

Tatar: English-only training is far below in-language fine-tuning: slot filling nearly fails (mBERT Slot F_1 0.235), while intent scores degrade less. mDeBERTa is the most stable under English training (Avg. 0.703) and the overall leader on Tatar data (Avg. 0.879); Glot500-base is close (0.876). With adapted training, Slot F_1 on the adapted test drops relative to translated training (mDeBERTa: 0.786→0.640), consistent with entity mismatches between translated training examples and culturally localised test slots.

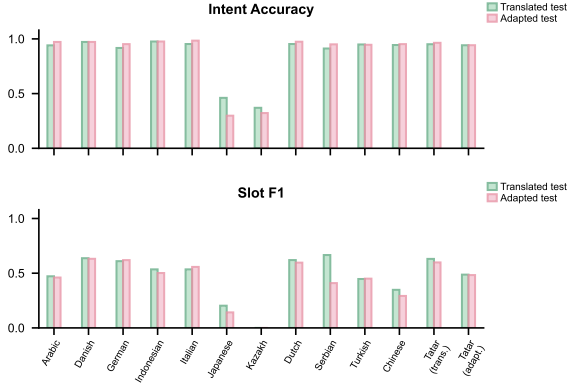
Tables 3 and 4 report encoder scores on translated (*trans*) and adapted (*adapt*) test sets.

5.2 Cross-lingual transfer

We fine-tune mDeBERTa-v3-base on each xSID pivot language and evaluate on Russian and Tatar test sets (Figures 4 and 5).

Russian: I_{acc} typically exceeds Slot F_1 by 25–40 pp; Indonesian (0.976) and Danish (0.972) lead on intent, while Serbian (0.666) and Danish (0.637)

Figure 4: mDeBERTa pivot transfer to Russian: I_{acc} and Slot F_1 on translated and adapted test sets.



Model	Train	trans			adapt		
		Int.	Slot	Avg	Int.	Slot	Avg
mDeBERTa	EN	0.977	0.801	0.889	0.976	0.743	0.860
	RU trans.	0.987	0.879	0.933	0.992	0.821	0.907
	RU adapt.	0.977	0.702	0.840	0.986	0.783	0.885
mmBERT	EN	0.928	0.649	0.789	0.924	0.586	0.755
	RU trans.	0.987	0.848	0.918	0.996	0.790	0.893
	RU adapt.	0.977	0.725	0.851	0.990	0.821	0.906
RemBERT	EN	0.905	0.638	0.772	0.958	0.540	0.749
	RU trans.	0.983	0.850	0.917	0.992	0.757	0.875
	RU adapt.	0.975	0.715	0.845	0.988	0.796	0.892
mBERT	EN	0.799	0.635	0.717	0.872	0.610	0.741
	RU trans.	0.989	0.852	0.921	0.996	0.786	0.891
	RU adapt.	0.983	0.642	0.813	0.990	0.764	0.877
EuroBERT-210M	EN	0.444	0.227	0.336	0.362	0.176	0.269
	RU trans.	0.979	0.843	0.911	0.988	0.771	0.880
	RU adapt.	0.981	0.655	0.818	0.994	0.768	0.881
EuroBERT-610M	EN	0.414	0.170	0.292	0.274	0.095	0.185
	RU trans.	0.981	0.813	0.897	0.988	0.739	0.864
	RU adapt.	0.970	0.654	0.812	0.994	0.766	0.880

Table 3: Model comparison on Russian NLU (test). Bold: highest Avg per split.

lead on slots. Symmetric Tatar→Russian training preserves intent (0.94–0.96) but slots fall to 0.48–0.63.

Tatar: Intent stays near 0.82–0.91 for most pivots, whereas Slot F_1 spans 0.13–0.50; Turkish is best (0.502), followed by Russian (0.461–0.530).

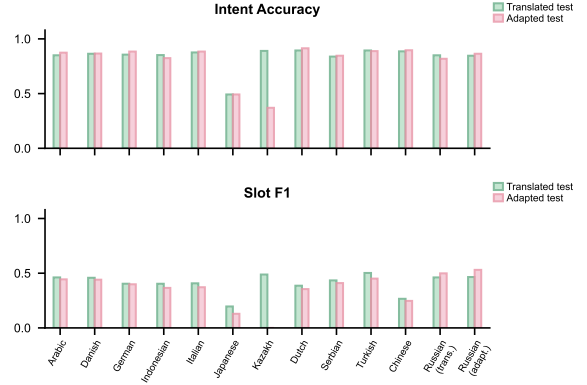
Japanese and Kazakh: Japanese xSID covers only seven intents (Krückl et al., 2025), which limits transfer to the full Russian and Tatar inventories ($I_{acc} \approx 0.46$ –0.49). Kazakh yields near-zero Slot F_1 on Russian test and collapses on Tatar adapted test despite Turkic relatedness — likely due to poor machine-translated pivot data and weak Kazakh pretraining coverage rather than typology alone.

5.3 Spoken NLU (ASR→NLU)

Table 5 reports the best ASR→NLU configuration on the adapted test (500 utterances, mDeBERTa).

GigaAM→NLU preserves Russian intent (I_{acc}

Figure 5: mDeBERTa pivot transfer to Tatar: I_{acc} and Slot F_1 on translated and adapted test sets.



Model	Train	trans			adapt		
		Int.	Slot	Avg	Int.	Slot	Avg
mDeBERTa	EN	0.882	0.523	0.703	0.866	0.492	0.679
	TT trans.	0.972	0.786	0.879	0.964	0.696	0.830
	TT adapt.	0.968	0.640	0.804	0.942	0.679	0.811
mmBERT	EN	0.666	0.253	0.460	0.686	0.216	0.451
	TT trans.	0.966	0.784	0.875	0.950	0.638	0.794
	TT adapt.	0.930	0.580	0.755	0.914	0.672	0.793
RemBERT	EN	0.714	0.291	0.503	0.738	0.285	0.512
	TT trans.	0.936	0.796	0.866	0.942	0.694	0.818
	TT adapt.	0.908	0.583	0.746	0.912	0.649	0.781
mBERT	EN	0.528	0.235	0.382	0.528	0.221	0.375
	TT trans.	0.942	0.748	0.845	0.914	0.668	0.791
	TT adapt.	0.930	0.566	0.748	0.888	0.681	0.785
GLOT500	EN	0.804	0.447	0.626	0.806	0.429	0.618
	TT trans.	0.966	0.786	0.876	0.964	0.699	0.832
	TT adapt.	0.914	0.603	0.759	0.914	0.661	0.788

Table 4: Model comparison on Tatar NLU (test). Bold: highest Avg per split.

Configuration	Intent Acc	Slot F_1
Russian adapted (GigaAM)	0.992	0.384
Tatar adapted (Söyle)	0.881	0.519

Table 5: Spoken NLU on adapted test (mDeBERTa)

0.992) but Slot $F_{1,500}$ is only 0.384 (Avg. 0.688), roughly half the text-only baseline (0.933). Raw Söyle transcripts retain Tatar intent (0.881) yet slots fall to 0.519, well below the text encoder result (0.786); ASR noise primarily harms slot boundaries rather than intent labels.

5.4 Generative NLU

Six instruction-tuned LLMs are evaluated in zero-shot and few-shot (1, 5, and 10 in-domain examples) without fine-tuning. Russian runs use the translated test with a Russian prompt (selected in a pilot because Slot F_1 exceeded the English prompt); Tatar runs use the translated test with a Tatar prompt (selected after a pilot on Qwen2.5-

3B: Intent Acc 0.600 vs. 0.522 English). Figure 6 and Figure 7 in Appendix A report Intent Accuracy and Slot F_1 ; 1-shot values average three demonstration types (popular weather/find, problematic SearchScreeningEvent, slot-rich BookRestaurant).

Russian: Gemma-2-9B dominates across all shot counts on Russian, reaching Intent Accuracy of 0.95 at zero-shot and 0.97 at 10-shot. Most models show stable or slightly improving Intent Accuracy with more examples, while Slot F_1 improves more consistently across shots (+4–16 p.p. from zero- to 10-shot). Phi-4-mini is the sole exception: its Intent Accuracy drops sharply at 5-shot (0.63) before partially recovering at 10-shot (0.79), suggesting sensitivity to prompt length. The intent-slot gap of approximately 40 p.p. persists across all configurations.

Tatar: On Tatar, generative models perform substantially worse, particularly on slot filling. Gemma-2-9B again leads, achieving Intent Accuracy of 0.89 at zero-shot and 0.93 at 10-shot, while Qwen2.5-7B reaches 0.83 at 5-shot. Slot F_1 stays below 0.40 for all models across all shot counts, confirming that slot extraction remains out of reach for instruction-tuned LLMs without fine-tuning on Tatar. Gemma-2-2B collapses at 1-shot (Intent Accuracy 0.12) before recovering; this instability may partly reflect a data preprocessing artefact in the Tatar evaluation pipeline.

5.5 Error analysis

For both languages and both model types, the dominant intent error pattern is confusion between semantically similar classes: SearchCreativeWork vs. SearchScreeningEvent, PlayMusic vs. AddToPlaylist, and reminder vs. alarm sub-intents. Slot errors in encoders follow two patterns: missing spans or incorrect BIO boundaries, and correct span detection with wrong slot type; on Tatar, boundary errors are especially frequent on agglutinative datetime and possessive reference spans (§3.3), and adapted-test slots degrade when training still contains translated rather than localised entities. Russian generative models show additional slot error patterns beyond this encoder baseline: slot hallucination, transliteration mismatches, slot merging, span truncation, and confusion between rating_value and best_rating. Russian intent errors additionally include indirect requests (e.g., “Do I

need a sweater?” predicted as none instead of weather/find) and lexical ambiguity where verb choice shifts the prediction.

On Tatar, generative models share the same intent confusions — including indirect weather queries such as “Miña sviter kiräkme?” predicted as none instead of weather/find — but rarely exhibit this extended slot taxonomy: failures are dominated by slot omission (e.g., the inflected datetime span *bügen sägat un altyga* missing from “*Bügen sägat un altyga xäterlättüne östä*”), empty or partial JSON, and surface forms that do not match the inflected token in the utterance; Slot F_1 stays below 0.40 for every LLM, and smaller models can collapse under 1-shot prompting.

6 Conclusion

We present Russian and Tatar NLU corpora — text and audio — built on xSID-0.7 with manually curated test and validation sets and scalable machine-translated training data.

In-language fine-tuning is essential for both languages: switching from English to Russian or Tatar training data raises average scores by 15–20 pp for encoders and closes most of the intent gap.

For cross-lingual transfer to Russian, Germanic and Scandinavian pivots work best on slot filling (Danish Avg. 0.804), while distant pivots such as Japanese and Kazakh fail on slots entirely. For Tatar, Turkish achieves the highest slot transfer among xSID pivots (Slot F_1 0.502), and Russian-to-Tatar transfer yields the strongest adapted-test slot scores (0.530), confirming that typological proximity and shared pretraining data jointly drive transfer quality.

Translated training data outperforms culturally adapted training for slot filling; adapted test sets are preferable when evaluating on localised entities.

For spoken NLU, ASR noise harms slot boundaries far more than intent labels: GigaAM preserves Russian intent (0.992) but halves Slot F_1 relative to the text baseline, and Söyle transcripts retain Tatar intent (0.881) while slots fall to 0.519.

Fine-tuned encoders remain preferable for slot-heavy NLU; instruction-tuned generative models are competitive on intent detection with well-chosen few-shot demonstrations but trail substantially on slot extraction, particularly for Tatar.

Limitations

Generative models are evaluated without fine-tuning due to computational resource constraints; parameter-efficient fine-tuning may further improve slot performance. Audio data were recorded by non-professional speakers using consumer-grade microphones under everyday conditions, which limits acoustic diversity. Consequently, the audio corpus is small, and coverage of age groups, regional accents, and recording conditions remains limited.

Acknowledgments

This work was supported by the Russian Science Foundation grant № 26-71-00095.

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 EMNLP 2023*, pages 9904–9923.
- Gaurav Arora, Shreya Jain, and Srujana Merugu. 2024. [Intent detection in the age of LLMs](#). In *Proceedings of EMNLP 2024 (Industry Track)*, pages 1559–1570.
- Ekaterina Artemova, Verena Blaschke, and Barbara Plank. 2024. [Exploring the robustness of task-oriented dialogue systems for colloquial German varieties](#). In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 445–468.
- Emanuele Bastianelli, Andrea Vanzo, Pawel Swietojanski, and Verena Rieser. 2020. [SLURP: A spoken language understanding resource package](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 7252–7262.
- Nicolas Boizard and 1 others. 2025. EuroBERT: Scaling multilingual encoders for European languages. *arXiv preprint*.
- Artem Chervyakov, Ulyana Isaeva, Aleksandr Emelyanov, Arseny Safin, Maria Tikhonova, Aleksandr Kharitonov, Yulia Lyakh, Pavel Surovtsev, Dmitry Shevelev, Vladimir Saburov, Vladimir Kononov, Evgeny Rykov, Ivan Sviridov, Alina Miftakhova, Irina Alimova, Andrey Panchenko, Aleksandr Kapitanov, and Aleksandra Fenogenova. 2026. Multimodal evaluation of Russian-language architectures. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2114–2161.
- Artem Chirkin, Sofia Kuznetsova, Maria Volina, and Anastasia Dengina. 2025. [RusConText benchmark: A Russian language evaluation benchmark for understanding context](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 1158–1170.
- Hyung Won Chung, Sebastian Ruder, Taemin Yu, and Jörg Tiedemann. 2020. Revisiting embedding strategies for cross-lingual language model adaptation. In *Proceedings of EMNLP 2020*.
- Alice Coucke, Alaa Saade, Adrian Ball, and 1 others. 2018. [Snips Voice Platform: An embedded spoken language understanding system for private-by-design voice interfaces](#). ArXiv:1805.10190.
- 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 2019*, pages 4171–4186.
- Pengcheng He, Jianfeng Gao, and Weizhu Chen. 2021. DeBERTaV3: Improving DeBERTa using ELECTRA-style pre-training with gradient-disentangled embedding sharing. In *Proceedings of ICLR 2023*.
- Songbo Hu, Han Zhou, Mete Hergul, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Ivan Vulić, and Anna Korhonen. 2023a. [Multi³WOZ: A multilingual, multi-domain, multi-parallel dataset for training and evaluating culturally adapted task-oriented dialog systems](#). *Transactions of the Association for Computational Linguistics*, 11:1396–1415.
- Songbo Hu, Han Zhou, Moy Yuan, Milan Gritta, Guchun Zhang, Ignacio Iacobacci, Anna Korhonen, and Ivan Vulić. 2023b. [A systematic study of performance disparities in multilingual task-oriented dialogue systems](#). In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 6825–6851.
- Vojtěch Hudeček and Ondřej Dušek. 2023. Are large language models all you need for task-oriented dialogue? In *Proceedings of the 24th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 216–228.
- Ayyoob Imani, Peiqin Lin, ..., and Hinrich Schütze. 2023. [Glot500: Scaling multilingual corpora and language models to 500 languages](#). In *Proceedings of ACL 2023*, pages 1082–1117.
- 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.

- Xaver Maria Krücl, Verena Blaschke, and Barbara Plank. 2025. Improving dialectal slot and intent detection with auxiliary tasks. In *Proceedings of VarDial 2025*, pages 128–146.
- Aleksandr Kutsakov, Alexandr Maximenko, Georgii Gospodinov, Pavel Bogomolov, and Fyodor Minkin. 2025. GigaAM: Efficient self-supervised learner for speech recognition. In *Interspeech 2025*, pages 1213–1217.
- Liz Li and Wei Zhu. 2026. A generative model for joint multiple intent detection and slot filling. *arXiv preprint arXiv:2602.08322*.
- Petter Mæhlum and Yves Scherrer. 2024. NoMusic: The Norwegian multi-dialectal slot and intent detection corpus. In *Proceedings of VarDial 2024*, pages 107–116.
- Olga Majewska, Evgeniia Razumovskaia, Edoardo Maria Ponti, Ivan Vulić, and Anna Korhonen. 2023. [Cross-lingual dialogue dataset creation via outline-based generation](#). *Transactions of the Association for Computational Linguistics*, 11:139–156.
- Marc Marone and 1 others. 2025. mmBERT: A modern multilingual encoder with annealed language learning. *arXiv preprint*.
- Paramita Mirza, Viju Sudhi, Soumya Ranjan Sahoo, and Sinchana Ramakanth Bhat. 2024. ILLUMINER: Instruction-tuned large language models as few-shot intent classifier and slot filler. In *Proceedings of LREC-COLING 2024*, pages 8639–8651.
- Saida Mussakhoyayeva, Rinat Gilmullin, Bulat Khakimov, Mansur Galimov, Daniil Orel, Adal Abilbekov, and Huseyin Atakan Varol. 2024. [Noise-robust multi-lingual speech recognition with long transcription featuring the tatar speech corpus](#). In *2024 International Conference on Artificial Intelligence in Information and Communication (ICAIC)*, pages 732–737.
- Jinjie Ni, Tom Young, Vlad Pandelea, Fuzhao Xue, and Erik Cambria. 2023. Recent advances in deep learning based dialogue systems: A systematic survey. *Artificial Intelligence Review*, 56:3055–3155.
- Evgeniia Razumovskaia, Goran Glavaš, Olga Majewska, Edoardo Maria Ponti, Anna Korhonen, and Ivan Vulić. 2022. Crossing the conversational chasm: A primer on natural language processing for multilingual task-oriented dialogue systems. *Journal of Artificial Intelligence Research*, 74:1351–1402.
- Sebastian Schuster, Sonal Gupta, Rushin Shah, and Mike Lewis. 2019. Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of NAACL-HLT 2019*, pages 3795–3805.
- Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov, Denis Shevelev, Ekaterina Artemova, Valentin Malikh, Vladislav Mikhailov, Maria Tikhonova, Andrey Chertok, and Andrey Evlampiev. 2020. [RussianSuperGLUE: A Russian language understanding evaluation benchmark](#). In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*, pages 4717–4726.
- Rob van der Goot, Ibrahim Sharaf, Aizhan Imankulova, Ahmet Üstün, Marija Stepanović, Alan Ramponi, Siti Oryza Khairunnisa, Mamoru Komachi, and Barbara Plank. 2021a. [From masked language modeling to translation: Non-English auxiliary tasks improve zero-shot spoken language understanding](#). In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 2479–2497.
- Rob van der Goot, Ahmet Üstün, Alan Ramponi, Ibrahim Sharaf, and Barbara Plank. 2021b. Massive choice, ample tasks (MaChAmp): A toolkit for multi-task learning in NLP. In *Proceedings of EACL 2021 System Demonstrations*, pages 176–197.
- Miriam Winkler, Virginija Juozapaityte, Rob van der Goot, and Barbara Plank. 2024. Slot and intent detection resources for Bavarian and Lithuanian: Assessing translations vs natural queries to digital assistants. In *Proceedings of LREC-COLING 2024*, pages 14898–14915.

Appendix A: Generative NLU Results

Figures 6 and 7 report generative NLU results under zero-shot and few-shot prompting.

Russian: In zero-shot, Gemma-2-9B leads (Intent Acc. ≈ 0.95 , Slot $F_1 \approx 0.52$); Gemma-2-2B is lowest on slots (≈ 0.27). No model exceeds Slot F_1 0.52 in zero-shot. 1-shot scores average three demonstration types (≈ 1 –5 pp spread). Phi-4-mini drops at 5-shot (≈ 0.63 intent) then recovers at 10-shot (≈ 0.78). At 10-shot, Gemma-2-9B leads (intent ≈ 0.97 , Slot $F_1 \approx 0.59$). The intent–slot gap (≈ 40 pp) persists across all configurations.

Tatar: A Qwen2.5-3B pilot favoured Tatar prompts (Intent Acc. 0.600 vs. 0.522 / 0.533 English/Russian); all prompts gave Slot $F_1 \leq 0.09$. Zero-shot: Gemma-2-9B leads (Intent Acc. ≈ 0.89 , Slot $F_1 \approx 0.27$); Mistral-7B and Gemma-2-2B lag on slots (≤ 0.07). Gemma-2-9B gains most from demonstrations (Slot $F_1 \approx 0.39$ at 10-shot). Qwen2.5-7B shows strong intent gains with more examples (0.65 $\rightarrow$ 0.83 at 5-shot). Phi-4-mini peaks at 5-shot (≈ 0.40 intent) then recovers slightly at 10-shot (≈ 0.45). Other models stay below Slot F_1 0.26; Gemma-2-2B collapses at 1-shot (≈ 0.12 intent), which may partly reflect a data preprocessing artefact.

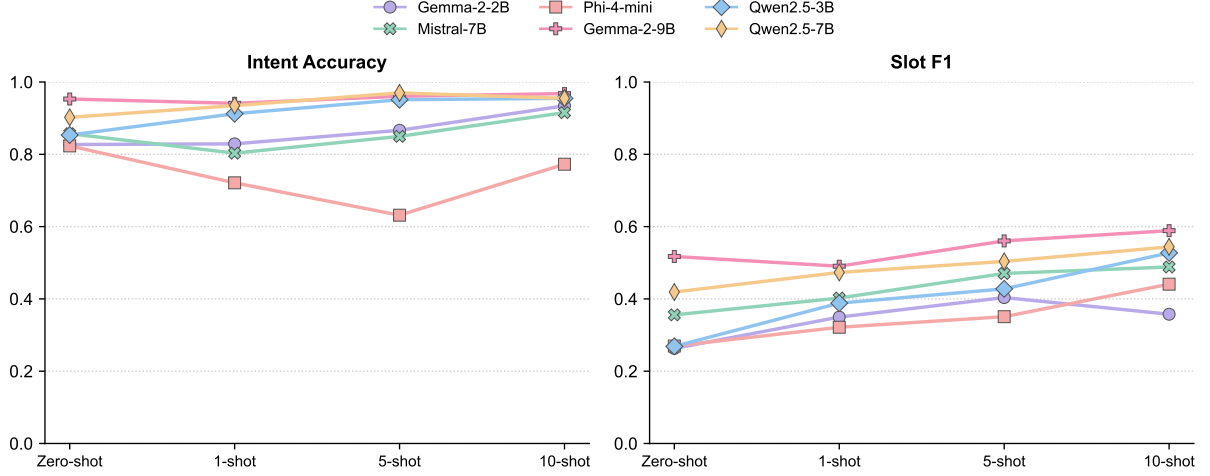


Figure 6: Generative LLMs on Russian translated test: Intent Accuracy and Slot F_1 (zero-shot and few-shot).

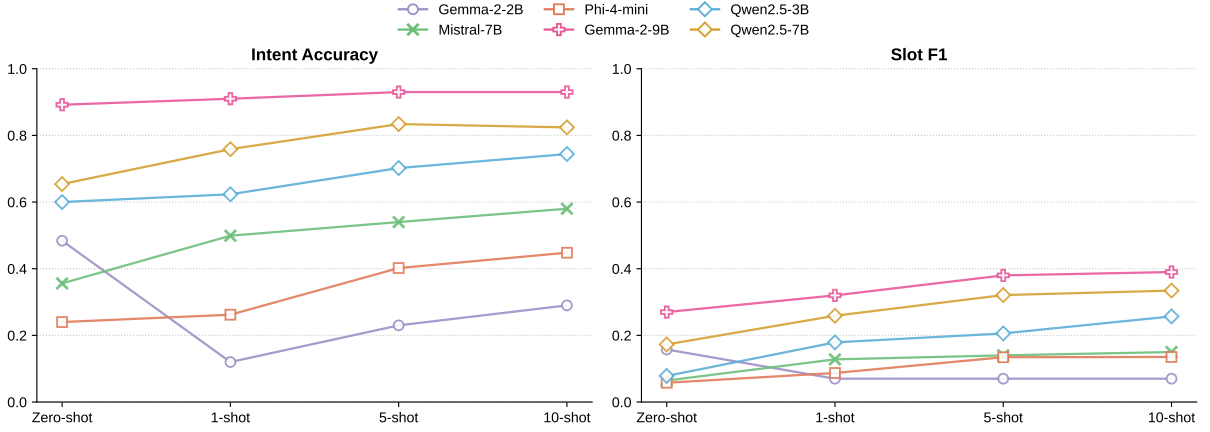


Figure 7: Generative LLMs on Tatar test (Tatar prompt): Intent Accuracy and Slot F_1 (zero-shot and few-shot).

Appendix B: Training Data Translation Prompts

We used four system prompts in this work: two token-level translation prompts for building the training corpora, and two intent/slot classification prompts for the generative evaluation in §4.3.1. Several LLMs (GPT-5, DeepSeek, Gemini-2.5-Flash, and the instruction-tuned models listed in §4.1.2) were used at different stages; prompts were adjusted per model where needed, and the versions below are representative.

B.1 Tatar Training Set Translation Prompt

SYSTEM_PROMPT = """STRICT RULES:

- Do NOT add comments.
- Output ONLY CoNLL format.

1. Response ALWAYS starts with:
text: <Tatar translation>
intent: <intent from the source example>

2. The only forbidden combination is "minja" + "minem" (both meaning "my").
3. Split time expressions into separate entity tokens, e.g. 9:15 -> 9 I-datetime, : I-datetime, 15 I-datetime.
4. For datetime slots, case/form selection depends on the time's semantic role:
 - Target time (setting/scheduling: set/add/schedule/alarm/reminder): use the directional case (-ka/-ka). If am/pm/morning/evening/night is present, do NOT use the adverbs for "in the morning/evening/night/daytime"; instead use "morning-/evening-/night-/daytime-adjective + number + -ka/-ka". E.g. "at 5 am" -> "irtange 5-ka".
 - Descriptive time (a fact/description of an event's timing): use the locative-temporal case (-ta/-ta); adverbial forms for morning/evening/night/daytime are allowed.
5. TOP PRIORITY RULE: translate as literally as possible token-by-token

from English to Tatar, preserving order and grammatical form (mood, tense, number), unless this violates basic Tatar grammar rules.

6. SECOND MOST IMPORTANT RULE: proper names (people, artists, authors, etc.) MUST be written in Cyrillic transliteration with Tatar suffixes as needed. Names of objects (playlists, services, brands, titles of works, etc.) keep their original (English) spelling.

Example 1: EN "Add the Matt Murphy tune to the Flow Espanol playlist" -> TAT "Matt Murphy-ning koyen Flow Espanol playlistina osta."

Example 2: EN "I want to listen to something on Youtube" -> TAT "Min Youtube-ta berer narse tinglarga telim."

Example 3: "Matt Murphy" -> "Matt Murphy-ning", where "-ning" is a Tatar genitive suffix.

7. Titles of films, series, etc. may be translated if they are well known to a general audience in Russia (e.g. "Friends" -> the Russian-known title).

8. The pronoun "my" is translated STRICTLY as a possessive suffix (-m/-em/-im).

9. In the English source, "my" is tagged as `reference`. In the Tatar version, `reference` tags the word carrying the possessive suffix (-m/-em/-im). E.g. "turn off my alarm" -> "buldil' negimni sundir", where "buldil' negimni" is B-reference and "sundir" is O.

10. Slot tokens belonging to one entity must stay contiguous, as in the source. If translation would split them, rephrase the sentence so the entity stays whole (example given in the original showing reordering to keep a B-datetime/I-datetime span unbroken).

11. Replace any LGBT-related names/content with neutral equivalents.

12. Numbers are ALWAYS written as digits; case endings are added with a hyphen. NEVER spell numbers out as words. E.g. "6-dan", "4-ta", "5-ka".

13. Do not add function words that are not present in the original.

14. If entities in the source text are logically linked (possession, authorship, purpose, source), this link MUST be made explicit in the Tatar translation (via genitive case, possessive suffix, or an equivalent construction).

[English example -> Tatar example follows, same content as source, omitted here for brevity]

""

B.2 Russian Training Set Translation Prompt

SYSTEM_PROMPT = ""Translation rules for Russian annotation:

1. The pronoun "my": either translate as "all/current" for alarms and reminders, or translate literally as "my" (moy), depending on context.
2. If a hyphen occurs inside an entity, tag it as part of the same single entity (e.g. "8 Tatra B-location, 9 - I-location, 10 National I-location").
3. For playlist/movie/etc. titles: either keep the original (untranslated) form, or, if it is a well-known title/name (e.g. "Harry Potter", "The Big Bang Theory", "Friends"), adapt it to its established Russian equivalent.
4. State abbreviations in examples (e.g. "AK" in "...in Suwanee AK") are kept as-is, untranslated.
5. For present continuous constructions: either translate as a continuous/ongoing action (e.g. "Is The Two Gladiators playing" -> ongoing process reading), or, if a grammatical future marker is present, translate as future tense (e.g. "Is Any Which Way You Can playing in 15 seconds" -> future reading).
6. Replace any LGBT-related content in titles/names with an allowed neutral equivalent.

[English example -> Russian example follows, same content as source, omitted here for brevity]

""

The two prompts below were used to constrain the zero-/few-shot generative LLMs (§4.1.2) to the xSID-0.7 intent/slot inventory at inference time.

B.3 Schema Prompt for Generative Models (Tatar)

SYSTEM_PROMPT = ""You are a natural language understanding (NLU) system. Carefully read the user's query text, determine its intent, and extract all slots applicable to that intent.

IMPORTANT: slot words must keep the same morphological form as in the text. Write numbers as words.

AVAILABLE INTENTS AND THEIR SLOTS:
[the same intent/slot inventory as the Russian generative prompt below, listing BookRestaurant, SearchScreeningEvent, SearchCreativeWork, AddToPlaylist, RateBook, PlayMusic, weather/find, the alarm/* intents, and the reminder/* intents, each with their slot sets]

```

---
STRICT RULES:
1. Use ONLY the intents and slots listed
   above.
2. Copy slot values EXACTLY from the
   user's text -- do not translate or
   normalize.
3. If a slot is not present in the text,
   do not include it in the response.
4. Output ONLY valid JSON, with no
   formatting, comments, or backticks.
5. NEVER invent new slot values.

Response format:
{"intent": "...", "slots": {"slot_name":
  "value copied from text"}}
"""
INTENTS = [
    "AddToPlaylist", "BookRestaurant", "
    PlayMusic", "RateBook",
    "SearchCreativeWork", "
    SearchScreeningEvent",
    "alarm/cancel_alarm", "alarm/
    modify_alarm", "alarm/set_alarm",
    "alarm/show_alarms", "alarm/
    snooze_alarm", "alarm/
    time_left_on_alarm",
    "reminder/cancel_reminder", "
    reminder/set_reminder", "reminder/
    show_reminders",
    "weather/find",
]
SLOTS = [
    "alarm/alarm_modifier", "album", "
    artist", "best_rating",
    "condition_description", "
    condition_temperature", "cuisine", "
    datetime",
    "entity_name", "facility", "genre",
    "location", "movie_name", "
    movie_type",
    "music_item", "negation", "news/type
    ", "object_location_type", "
    object_name",
    "object_part_of_series_type", "
    object_select", "object_type",
    "party_size_description", "
    party_size_number", "playlist", "
    rating_unit",
    "rating_value", "recurring_datetime",
    "reference", "reminder/
    reminder_modifier",
    "reminder/todo", "restaurant_name",
    "restaurant_type", "served_dish",
    "service", "sort", "timer/attributes
    ", "track", "weather/attribute",
    "weather/temperatureUnit",
]

```

B.4 Schema Prompt for Generative Models (Russian)

SYSTEM_PROMPT = """You are a natural language understanding (NLU) system. Carefully read the user's query text, classify its intent, and extract all slots applicable to that intent.

IMPORTANT: slot words must keep the same

morphological form as in the text. Write numbers as words.

```

---
AVAILABLE INTENTS AND THEIR SLOTS:

BookRestaurant -- restaurant/cafe
  booking:
  slots: cuisine, datetime, facility,
         location, party_size_description,
         party_size_number,
         restaurant_name, restaurant_type,
         served_dish, sort

SearchScreeningEvent -- searching for
  film screenings:
  slots: datetime, location, movie_name,
         movie_type, object_location_type,
         object_type

SearchCreativeWork -- searching for
  creative works:
  slots: object_name, object_type

AddToPlaylist -- adding to a playlist:
  slots: artist, entity_name, music_item,
         playlist, reference

RateBook -- rating a book:
  slots: best_rating, object_name,
         object_part_of_series_type,
         object_select,
         object_type, rating_unit,
         rating_value

PlayMusic -- playing music:
  slots: album, artist, datetime, genre,
         music_item, playlist, service, sort,
         track

weather/find -- weather query:
  slots: condition_description,
         condition_temperature, datetime,
         location, weather/attribute

alarm/cancel_alarm, alarm/modify_alarm,
alarm/set_alarm, alarm/show_alarms,
alarm/snooze_alarm, alarm/
time_left_on_alarm -- slots:
datetime, reference,
recurring_datetime, reminder/todo

reminder/cancel_reminder, reminder/
set_reminder, reminder/
show_reminders -- slots: datetime,
reference, reminder/todo

```

```

---
STRICT RULES:
1. Use ONLY the intents and slots listed
   above.
2. Copy slot values EXACTLY from the
   user's text.
3. If a slot is absent from the text, do
   not include it in the response.
4. Output ONLY valid JSON, with no
   comments.
5. NEVER invent new slot values.

```

Response format:

```
{"intent": "...", "slots": {"slot_name":  
    "value from text"}}  
"""
```