

T-pro 2.0: An Efficient Russian Hybrid-Reasoning Model and Playground

Gen-T Team
T-Tech

Correspondence: anatolii.s.potapov@gmail.com

Abstract

We introduce *T-pro 2.0*, an open-weight Russian LLM for hybrid reasoning and efficient inference. The model supports direct answering and reasoning-trace generation, using a Cyrillic-dense tokenizer and an adapted EAGLE speculative-decoding pipeline to reduce latency. To enable reproducible and extensible research, we release the model weights, the *T-Wix* 500k instruction corpus, the *T-Math* reasoning benchmark, and the EAGLE weights on Hugging Face. These resources allow users to study Russian-language reasoning and to extend or adapt both the model and the inference pipeline. A public web demo exposes reasoning and non-reasoning modes and illustrates the speedups achieved by our inference stack across domains. T-pro 2.0 thus serves as an accessible open system for building and evaluating efficient, practical Russian LLM applications.

Demo: <https://t-pro2eagle.streamlit.app/>

 hf.co/collections/t-tech/t-pro-20

1 Introduction

Large Language Models (LLMs) have progressed from basic text generation to systems capable of multi-step reasoning and efficient inference. Recent foundation models show that reasoning-oriented training (DeepSeek-AI et al., 2025a; Yang et al., 2025) and improved decoding methods (Chen et al., 2023; Li et al., 2024e) can substantially boost both accuracy and speed.

In the Russian open-source space, progress remains limited. Most strong models are closed and accessible only through APIs (Mamedov et al., 2025; Zmitrovich et al., 2023), while open models are typically small adaptations of multilingual systems (Nikolich et al., 2024). There is no unified ecosystem for studying Russian-language reasoning: high-quality evaluation sets are scarce, and, to the best of our knowledge, there are currently few

public demos that let users compare direct answering and step-by-step reasoning, inspect inference-time optimizations, or observe how decoding speed impacts user experience.

To address these gaps, we introduce *T-pro 2.0*, an open-weight Russian LLM for hybrid reasoning and an interactive demo platform. The model supports two complementary modes—direct answering and explicit reasoning traces—enabling applications to balance speed and accuracy within a single deployed system.

Our training setup combines a Cyrillic-dense tokenizer derived from Qwen3 (Yang et al., 2025), large-scale instructional midtraining, supervised fine-tuning focused on both reasoning and non-reasoning, preference optimization, and an adaptation of EAGLE-style speculative decoding (Li et al., 2024e) to accelerate Russian-language inference. To sum up, our main contributions are:

- *T-pro 2.0*, an open-weight Russian hybrid-reasoning LLM with improved inference efficiency via an optimized Cyrillic tokenizer and EAGLE-style speculative decoding.
- *T-Wix*, the largest open Russian hybrid-reasoning SFT dataset to date ($\approx 500k$ samples) covering general instruction following, long-context tasks, and teacher-generated reasoning traces.
- *T-Math*, a benchmark of Russian high-school olympiad-level mathematics problems for curriculum-aligned reasoning evaluation.
- An interactive web demo that exposes *T-pro 2.0* as a research-oriented live system¹, enabling side-by-side comparison of reasoning and non-reasoning modes, running tasks from our datasets and benchmarks, and viewing telemetry for inference-time optimizations.

¹The [web demo video](#) is available on YouTube.

All model-related components (T-pro 2.0, EAGLE weights, and the T-Math benchmark) are released under the Apache-2.0 license, while the T-Wix corpus is released under the ODC-By open data license.

2 Related Work

The development of Russian LLMs primarily follows two tracks: monolingual pre-training and adaptation of multilingual models. Early decoder-only baselines like ruGPT (Kuratov and Arkhipov, 2019; Zmitrovich et al., 2023) and commercial systems such as YandexGPT² and GigaChat (Mamedov et al., 2025) focus on Russian-centric pre-training. While achieving promising results on Russian benchmarks, early versions face a capability gap compared to leading multilingual LLMs like Qwen (Yang et al., 2024) and Llama (Dubey et al., 2024).

To mitigate these limitations, *T-pro 1.0*³ adopts a continued pre-training strategy on large-scale Russian corpora, reaching state-of-the-art results on MERA (Fenogenova et al., 2024) among open Russian models. Its release aligns with a broader shift toward strengthening open-source Russian LLMs, alongside projects such as Saiga (Gusev, 2023), RuAdapt (Tikhomirov and Chernyshev, 2024), and Vikhr (Nikolich et al., 2024). These works emphasize the value of mitigating English-centric tokenizer limitations (Petrov et al., 2024) and extending pre-training on Russian data. This direction continues to grow: although YandexGPT-5-Lite⁴ is a fully pre-trained model rather than an adaptation, its recent open release further expands the set of publicly available Russian foundation models.

Russian Instruction Datasets. Existing Russian instruction datasets vary in provenance and domain coverage. Saiga (Gusev, 2023) applies self-instruct (Wang et al., 2023) pipelines producing `ru_turbo_saiga`, GrandMaster-PRO-MAX (Nikolich et al., 2024) aggregates sources across coding and general knowledge, and RuAdapt (Tikhomirov and Chernyshev, 2024) combines translated and native Russian samples. However, these datasets are usually small and contain few reasoning-intensive tasks.

²<https://ya.ru/ai/gpt>

³<https://huggingface.co/t-tech/T-pro-it-1.0>

⁴<https://huggingface.co/yandex/YandexGPT-5-Lite-8B-pretrain>

Efficient Inference. Speculative decoding accelerates autoregressive inference (Leviathan et al., 2023a). EAGLE (Li et al., 2024d) uses a lightweight head to generate draft token trees verified in parallel, achieving 2–3× speedup. Multi-Token Prediction (MTP) (Gloeckle et al., 2024) trains models to predict multiple tokens simultaneously and is deployed successfully in DeepSeek-V3 (DeepSeek-AI et al., 2025b). GigaChat models (Mamedov et al., 2025) also adopt MoE architecture for increased efficiency on training and inference stages. Speculative decoding remains underexplored for general-purpose Russian LLMs, with few publicly documented deployments.

3 T-pro 2.0

3.1 System and Demonstration Description

We provide a public web demo of T-pro 2.0 that exposes the model as an interactive hybrid-reasoning assistant and makes our inference optimizations directly observable. The service is stateless and does not store user prompts or completions. The interface supports multi-turn chat in Russian and English and side-by-side comparison with baseline models (by default Qwen3-32B-Instruct), allowing users to inspect both answers and reasoning traces under identical serving conditions. The demo currently supports text-only interactions and does not perform additional server-side content filtering beyond what is built into the underlying models.

Architecture. The demo is a single-page web application backed by a lightweight Python HTTP server. The server exposes a simple JSON API, attaches configuration options (model, decoding mode, generation parameters) received from the UI, and forwards requests to two serverless SGLang endpoints (Gu et al., 2024). Each endpoint runs on a single NVIDIA H100 GPU: one hosts T-pro 2.0 with an EAGLE-style speculative decoding pipeline (draft head + 32B verifier), and the other hosts the Qwen3-32B baseline with standard autoregressive decoding. The deployment is tuned for interactive use and supports around 20 concurrent users per model while keeping per-request latency low.

User interface and functionality. Figure 1 shows the main layout. The central comparison view presents parallel completions from two systems. For each side, users can independently choose between standard and reasoning modes.

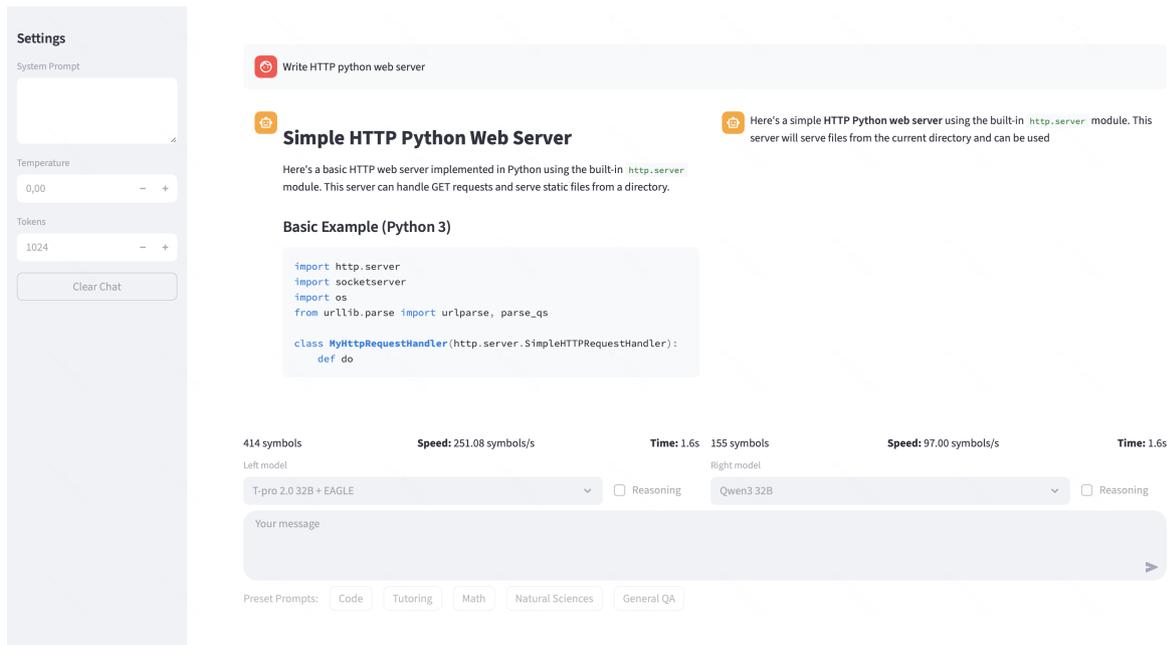


Figure 1: Screenshot of the system demo of the T-pro 2.0 EAGLE.

Outputs are streamed token by token, making differences in latency, verbosity, and reasoning structure directly visible. A control panel above the chat area lets users select models, toggle reasoning per model, and adjust decoding parameters such as temperature, maximum length, and sampling options. All decoding settings used for a given interaction are displayed in the UI, so that qualitative comparisons can be reproduced outside the demo. A typical interaction consists of selecting a preset prompt (or entering a custom query), choosing reasoning and generation settings, and launching both models to compare their outputs and telemetry.

To support systematic probing, the interface provides a small library of predefined prompts grouped by domain (general questions, math and science, code, etc.). Several presets are derived from our evaluation suites, including T-Math and other Russian reasoning benchmarks, so that users can quickly examine T-pro 2.0 on challenging tasks without reconstructing benchmark-style prompts.

Performance telemetry and usage patterns. A telemetry panel reports, for every request and model, the number of generated tokens, end-to-end latency, streaming throughput in tokens per second, and the acceptance ratio of speculative tokens for T-pro 2.0. Relating these statistics to the visible outputs illustrates how speculative decoding affects both accuracy and perceived responsiveness for short conversational turns and long reasoning

traces, complementing the benchmark results in Section 4.

3.2 Training recipe

This section describes the T-pro 2.0 training pipeline, integrating tokenizer adaptation, instructional midtraining, general post-training, and EAGLE-based speculative decoding. At all stages, we perform MinHash deduplication against benchmarks to prevent data leakage.

Cyrillic-dense tokenizer We address the systematic under-tokenization of Russian in multilingual models by replacing 34k low-frequency non-Cyrillic tokens in the Qwen3 (Yang et al., 2025) vocabulary with Cyrillic ones while keeping the total size fixed.

To build the expansion set, we extract 35.7k candidate tokens containing at least one Cyrillic character from four donor tokenizers (Qwen3, RuAdapt (Tikhomirov and Chernyshev, 2024), cl100k_base (OpenAI et al., 2024), MGPT (Shlitzhko et al., 2023)). For each candidate, we evaluate its decomposition under the current merge graph and iteratively add those merges required to make two-piece decompositions fully reachable. Four refinement passes make approximately 95% of candidates reachable. Tokens containing Cyrillic, pure Latin tokens, punctuation, and all 1–2-symbol units are preserved, while the 34k removed tokens are selected via log-smoothed frequency

scoring on the midtraining data mixture to maintain tokenization quality across all domains (see Appendix C, Table 11, for the domain distribution). Our modification yields substantial compression

T-pro	Qwen3	GigaChat [†]	Ruadapt-Qwen3
2.71	3.63	2.89	3.26

Table 1: Average tokens per word on Wikipedia for eight Cyrillic languages (ru, uk, be, bg, sr, mk, kk, ky). Lower values indicate more efficient segmentation of Cyrillic text. [†]Indicates <https://huggingface.co/ai-sage/GigaChat-20B-A3B-instruct> model.

gains: on Russian Wikipedia, the share of Russian words tokenized into at most two tokens rises from 38% to 60% (see Table 7), and Tables 1, 9 further demonstrates that this improvement generalizes consistently across eight Cyrillic languages, with all of them exhibiting shorter average segmentations under our tokenizer. A full set of tokenizer evaluation metrics is provided in Appendix B.

Instructional midtraining To adapt the *Qwen3-32B* model to the new dense Russian tokenizer and enhance reasoning, we employ an intermediate stage on 40B tokens drawn from curated open-source instructions, synthetic tasks, and parallel corpora. The mixture is dominated by Russian (49%) and English (36%) text; in terms of domains, it is dominated by Reasoning (34.6%), General QA (28.8%), and Math (16.2%), supplemented by grounded synthetic Question Answering (QA), code, and real user dialogues. The data mixture undergoes rigorous domain-specific local sensitive hashing (LSH) deduplication and InsTag-based semantic deduplication (Lu et al., 2024; Abbas et al., 2023) to balance diversity. To ensure high quality and stylistic consistency, all assistant responses are regenerated using Qwen3-235B-A22B teacher. Training utilizes a 32k context window, stabilizing the model for downstream supervised fine-tuning (SFT). Ablations show that the instruct-only mid-training outperforms mixtures retaining the fraction of raw pre-training data on reasoning tasks, improving ruAIME 2024 (T-Tech, 2025e) from 0.60 to 0.67. Separately, 8B-scale experiments confirm the tokenizer transition, with the T-pro tokenizer reaching a higher MERA (Fenogenova et al., 2024) macro-average (0.574) than the original Qwen3 tokenizer (0.560). Full details and ablations are provided in Appendix C.

Reward Model (RM) construction To support the T-pro 2.0 post-training pipeline, a dedicated reward model is trained (see Appendix F). The RM is initialized from Qwen3-32B with a scalar regression head and trained with a Bradley–Terry preference objective on sequences up to 32K tokens using Ulysses-style sequence parallelism. Synthetic preference data are generated using knockdown tournaments over completions from multiple instruct- and reasoning-oriented model groups of different scales, substantially reducing the number of pairwise evaluations relative to an exhaustive pairwise scheme. For each instruction, completion pairs are judged, pairs with positional bias are discarded, and transitive tournament relations are added to improve preference coverage. To assess downstream performance, we design an Arena-Hard Best-of- N benchmark based on the Δ_{BoN} (best@ N – worst@ N) metric, on which our RM outperforms existing open-source reward models; full details are provided in Appendix F.1.

General Post-Training The T-pro 2.0 post-training pipeline is implemented through general and reasoning SFT, and on-policy Direct Preference Optimization (DPO), with all filtering procedures detailed in Appendices D-E.

For the general part of the T-Wix SFT dataset, approximately 14M raw instructions from open-source corpora are reduced to 468k samples using deduplication, a multi-stage filtering pipeline, and domain/complexity balancing across six domains (Math, Code, Science, General Instruction, General Knowledge, Writing) and three difficulty tiers (School, Student, Professor). Each instruction is expanded with 8 candidate completions generated by Qwen3-235B-A22B or DeepSeek-V3 (DeepSeek-AI et al., 2025b) and then passed through an RM-guided selection step. The resulting mixture is low-noise, domain-balanced, and predominantly Russian, with approximately 10% English data retained to preserve bilingual competence.

For the reasoning component, approximately 30K samples are drawn from a 450k English pool covering general reasoning, mathematics, natural sciences, and code. After translation and deduplication, candidate solutions are generated by the Qwen3-235B-A22B teacher model and a midtraining student checkpoint and then filtered via RM-based rejection. For verifiable tasks, the highest-scoring factually correct teacher output is selected; for open-ended tasks, the shortest valid

trace among the teacher’s top RM-ranked candidates is chosen.

DPO is performed on 100k instructions sampled from the T-Wix dataset, with a 90/10 general-to-reasoning ratio. For each instruction, 16 on-policy completions are RM-scored, and one high-contrast preference pair (best vs. worst) is formed, so that observed failure modes are directly targeted and alignment is improved without the overhead of on-line RL.

Speculative Decoding We integrated a lossless EAGLE-based speculative decoding module into T-pro 2.0, where a lightweight draft model proposes candidate tokens that are verified by the frozen 32B target model (Leviathan et al., 2023b). The draft model consists of a single Llama-2-based decoder layer with an FR-Spec component (Zhao et al., 2025), trained on hidden-state reconstruction (smoothed L_1) and token distribution alignment (KL divergence) losses. During inference, we employ EAGLE-2’s dynamic draft-tree mechanism via SGLang. As shown in Tables 2 and 3, at temperature 0.8 the module achieves an average speedup of $1.85\times$ in standard mode, showing similar speedups for both standard and reasoning modes. STEM domains consistently outperform humanities ($1.99\times$ vs $1.62\times$), due to more predictable token distributions in technical content. See Appendix G for training pipeline details.

Benchmark	Speedup		Acceptance Length	
	Standard	Reasoning	Standard	Reasoning
ruMT-Bench ¹	1.79	1.69	3.31	3.10
ruAlpaca ²	1.61	1.57	2.94	2.85
ruCodeEval ³	2.15	1.84	3.93	3.34
T-Math ⁴	–	2.25	–	4.01
Average	1.85	1.83	3.39	3.33

Table 2: Aggregated T-pro-2.0-eagle performance at temperature 0.8. Comparison of relative speedup and average acceptance length for the standard and reasoning modes. ¹T-Tech (2025c), ²T-Tech (2025a), ³Fenogenova et al. (2024), ⁴T-Tech (2025g)

Domain Category	Speedup	Acceptance Length
STEM & Business [†]	1.99	3.57
Social & Humanities [‡]	1.62	2.88
Average	1.79	3.19

Table 3: Aggregated acceleration results on ruMMLU-Pro. [†]Includes Math, Chem, Eng, Bus, Phys, CS. [‡]Includes Econ, Bio, Health, Psych, Phil, Hist, Law.

4 Evaluation

4.1 Benchmarks

We evaluate along three axes: factual knowledge, dialogue, and reasoning capabilities.

Russian common-knowledge benchmarks

MERA, MaMuRAMu (Fenogenova et al., 2024), and ruMMLU-Pro (T-Tech, 2025f), targeting world knowledge and logical competence.

Dialogue benchmarks

Arena Hard Ru (T-Tech, 2025b), Arena Hard 2 (Li et al., 2024b,c), and WildChat Hard Ru (Kukushkin, 2024) (curated native Russian queries). WildChat uses o3-mini responses as baseline; DeepSeek-V3.1-Terminus (DeepSeek-AI et al., 2025b) serves as judge across all arenas and DeepSeek-V3.1 (DeepSeek-AI et al., 2025b) for WildChat.

Reasoning benchmarks

AIME 24/25 (Zhang and Math-AI, 2024) (Zhang and Math-AI, 2025), MATH-500 (Hendrycks et al., 2021), GPQA Diamond (Rein et al., 2023), Vikhr Math/Physics (Kuleshov et al., 2025), and LiveCodeBench v4_v5 (Jain et al., 2024). English benchmarks are professionally localized (ruAIME, ruMATH-500, ruGPQA, ruLCB (T-Tech, 2025d)). Vikhr benchmarks use Math-Verify scoring ⁵.

T-Math benchmark

We introduce T-Math—331 problems from All-Russian and Moscow olympiads (1998–2025), automatically extracted and human-verified. Details are provided in Appendix H.

4.2 Results

General knowledge and dialogue abilities

Table 4 shows the results on Russian general-knowledge and dialogue benchmarks. T-pro 2.0 performs consistently well across all evaluations, scoring 0.66 on MERA and 0.697 on ruMMLU-Pro. These numbers put it close to GPT-4o (0.714) and above Russian-adapted baselines such as RuadaptQwen3-32B-Instruct (0.652).

On dialogue tasks, the model reaches 91.1 on Arena Hard Ru and 72.6 on WildChat Hard Ru, outperforming all open-source systems and most proprietary ones. On Arena Hard 2, T-pro 2.0 scores 53.5 on Hard Prompts and 64.8 on Creative Writing, showing that it reliably follows instructions across different task types. These results directly reflect

⁵<https://github.com/huggingface/Math-Verify>

Model	MERA	MaMuRAMu	ruMMLU-Pro	Arena Hard Ru	WildChat Hard Ru	Arena Hard 2 HP	Arena Hard 2 CW
<i>Open Source Models (27B-32B class)</i>							
T-pro 2.0 (Ours)	0.66	0.851	0.697	91.1 / 90.36	72.6 / 76.4	53.5 / 46.2	64.2 / 62.8
Qwen3-32B	0.582	0.833	0.677	83.95 / 84.66	59.6 / 51.9	46.4 / 32.9	53.7 / 41.5
RuadaptQwen3-32B-Instruct ¹	0.574	0.823	0.652	68.4 / 64.76	41.5 / 39.4	13 / 14.2	19.4 / 12.7
Gemma 3 27B ²	0.577	0.808	0.665	82.66	58.4	23.5	74.7
DeepSeek-R1-Distill-Qwen-32B ³	0.508	0.787	0.537	34.07 / 22.83	12.1 / 8.7	7.2 / 7.2	5.9 / 3.5
<i>Open Source Larger Scale & Proprietary Models</i>							
DeepSeek-V3	0.677	0.875	0.736	92.67	66.8	45.8	59.9
DeepSeek-R1 ³	–	–	–	95.74	90.3	73.6	90.3
YandexGPT5-Pro ⁴	–	–	0.604	19.13	12.1	3.8	2.6
GigaChat 2 Max ⁵	0.67	0.864	0.649	61.44	10.1	8.5	27.1
o4-mini ⁶ (medium)	–	–	–	<u>95.63</u>	<u>74.4</u>	<u>67</u>	49.8
GPT-4o ⁷	<u>0.642</u>	<u>0.874</u>	<u>0.714</u>	85.14	41.4	20.0	44.2

¹Tikhomirov and Chernyshev (2024), ²Team et al. (2025), ³DeepSeek-AI et al. (2025a), ⁴<https://ya.ru/ai/gpt>, ⁵Mamedov et al. (2025), ⁶OpenAI (2025), ⁷OpenAI et al. (2024).

Table 4: Comparison of models on Russian language understanding and dialogue benchmarks. In Arena Hard 2, subsets are Hard Prompt (HP) and Creative Writing (CW). For entries reported as a/b , the first value corresponds to the *reasoning* setting and the second to the *non-reasoning* setting. o4-mini and DeepSeek-R1 are omitted from MERA as this benchmark does not support reasoning model mode, while YandexGPT5-Pro is omitted from MERA due to licensing restrictions.

Model	T-Math	ruAIME 2024	ruAIME 2025	ruMATH-500	ruGPQA Diamond	ruLCB	Vikhr Math	Vikhr Physics
<i>Open Source Models (27B-32B class)</i>								
T-pro 2.0 (Ours)	0.541	0.704	0.646	0.94	0.591	0.563	0.799	0.51
Qwen3-32B	0.529	0.706	0.625	0.938	0.606	0.537	0.809	0.531
RuadaptQwen3-32B-Instruct	0.444	0.575	0.450	0.450	0.591	0.500	0.528	0.337
Gemma 3 27B	0.208	0.248	0.231	0.860	0.439	0.261	0.548	0.276
DeepSeek-R1-Distill-Qwen-32B	0.254	0.510	0.402	0.898	0.631	0.493	0.462	0.286
<i>Open Source Larger Scale & Proprietary Models</i>								
DeepSeek-V3	0.278	0.319	0.285	0.882	0.657	0.444	0.613	0.367
DeepSeek-R1	0.619	0.800	0.800	0.972	0.763	0.69	0.864	0.469
YandexGPT5-Pro	0.13	0.062	0.046	0.682	0.354	0.265	0.372	0.252
GigaChat 2 Max	0.142	0.102	0.062	0.702	0.475	0.272	0.372	0.245
o4-mini (medium)	0.634	0.781	0.771	0.958	0.773	0.705	0.834	0.408
GPT-4o	0.106	0.090	0.069	0.766	0.510	0.131	0.372	0.296

Table 5: Comparison of models on Russian advanced reasoning benchmarks.

the structure of the T-Wix corpus, which mixes general instruction-following with long-context tasks and distilled reasoning traces from stronger teacher models.

Reasoning capabilities Table 5 summarizes performance on T-Math and localized reasoning benchmarks. On T-Math, the model scores 0.541, indicating strong performance on original olympiad-style Russian problems. On ruAIME 2024 and 2025 it reaches 0.704 and 0.646, sharply outperforming DeepSeek-V3 (0.319/0.285), GPT-4o (0.090/0.069) and all proprietary Russian models. Results on ruMATH-500 (0.94) and Vikhr Math (0.799) further confirm the model’s ability to perform mathematical reasoning in Russian under varied setups. These results also show that T-Math is a chal-

lenging benchmark that reveals meaningful performance differences that are obscured by translated or adaptation-based alternatives.

Crucially, the Russian-focused training does not hurt English performance. As shown in Table 23, T-pro 2.0 achieves 0.765 on AIME 2024, 0.966 on MATH-500, and 0.556 on LiveCodeBench, closely matching Qwen3-32B (0.808, 0.961, 0.546) and consistently outperforming other adapted models in the same class. These results indicate that the Cyrillic-focused tokenizer and Russian-centric training pipeline preserve robust English reasoning with minimal degradation. A detailed breakdown is provided in Appendix I.

5 Conclusion

We present *T-pro 2.0*, an open-weight Russian language model tailored for hybrid reasoning and efficient inference. The combination of a Cyrillic-dense tokenizer, a reasoning-focused midtraining stage, and an adapted EAGLE-style speculative decoding pipeline allows the model to deliver strong performance on Russian tasks without increasing model size and without notable degradation on English benchmarks. Along with the model, we release *T-Wix*, a large-scale SFT dataset enriched with reasoning traces, and *T-Math*, a benchmark designed to probe mathematical and analytical abilities in Russian.

These results point to two broader takeaways. First, careful, targeted adaptation of strong multilingual backbones remains a practical and reproducible route for building high-quality models for languages with limited resources. Second, tokenizer design and inference optimization are not optional details but key components for deploying reasoning-capable models beyond English. We expect the released models, datasets, evaluation code, and public demo to support research on Russian-language reasoning LLMs and to contribute to more transparent and consistent evaluation practices in this area.

Ethical Statement

Possible Misuse. Our work may enable generation of misleading, offensive, or otherwise harmful content if deployed without appropriate safeguards. We do not support applications that restrict access to information, facilitate disinformation, target individuals or groups, or automate harmful actions. To mitigate these risks, we apply toxicity and safety filtering during post-training and provide usage guidelines for responsible deployment.

Biases and Data Quality. The datasets used for pre-training and fine-tuning contain publicly available Russian and English text, which may include stereotypes, factual inaccuracies, or cultural biases. While automated filtering and manual checks are applied, residual biases may remain. We recommend additional evaluation when transferring the model to domains or communities that are underrepresented in the training data.

Human Subjects and Privacy. This work does not involve human-subjects research or the collection of personally identifiable information. All data

sources comply with their respective licenses and usage policies.

Limitations

Despite the strong performance of *T-pro 2.0*, several limitations should be acknowledged, which are planned to be addressed in future work.

Limited Multilingual Evaluation. The training recipe was solely focused on Russian with English data replay to preserve and transfer abilities. As a result, downstream performance was evaluated only on these two languages, and generation quality in other languages remains untested and may vary. Additionally, since the reasoning ability was trained via distillation without explicit language control, the language of reasoning traces may be occasionally different. While such cases are infrequent in practice, developing methods for explicit language control during reasoning, broader multilingual support and evaluation are left for future work.

Limited Agentic Capabilities No dedicated improvements for tool use or agentic behavior were incorporated into the model. Optimizations for function calling or complex multi-turn interactions were not performed, and as a result, performance in these areas is expected to be comparable to or slightly below that of the base Qwen3-32B model. Enhancements in this direction are prioritized for future development.

Offline-Only Reinforcement Learning Model alignment was conducted exclusively through offline DPO. Online reinforcement learning methods such as PPO (Schulman et al., 2017) or GRPO (Shao et al., 2024) were not employed. Although DPO is computationally efficient, the absence of interactive feedback may limit robustness and lead to performance degradation on out-of-domain tasks.

Unverified Long-Context Performance All training stages for *T-pro 2.0* were carried out with a fixed context window of 32k tokens, consistent with the base Qwen3-32B configuration. While support for up to 128k tokens is theoretically enabled via RoPE scaling, the model’s ability to maintain coherence and retrieve information over such extended contexts has not been empirically validated.

Reproducibility Issues Full reproducibility is restricted by the use of proprietary datasets in mid-training and the DPO stage. However, the curated

SFT dataset is being released to support and encourage further research, particularly in the development of high-quality Russian-language language models.

Author Contributions

- *Project Lead and Technical Direction:* Anatolii Potapov
- *Training Pipelines Team:* German Abramov, Pavel Gein
- *Post-training Team:* Dmitrii Stoianov, Olga Tsymboi, Danil Taranets, Ramil Latypov, Almaz Dautov, Dmitry Abulkhanov
- *Inference Team:* Vladislav Kruglikov, Nikita Surkov
- *Evaluation Team:* Mikhail Gashkov, Viktor Zelenkovskiy, Artem Batalov, Aleksandr Medvedev

References

2025. Turbo-alignment. <https://github.com/turbo-llm/turbo-alignment>. GitHub repository.
- Amro Abbas, Kushal Tirumala, Dániel Simig, Surya Ganguli, and Ari S. Morcos. 2023. *Semdedup: Data-efficient learning at web-scale through semantic deduplication*. *arXiv preprint arXiv:2303.09540*.
- Berk Atil, Sarp Aykent, Alexa Chittams, Lisheng Fu, Rebecca J. Passonneau, Evan Radcliffe, Guru Rajan Rajagopal, Adam Sloan, Tomasz Tudrej, Ferhan Ture, Zhe Wu, Lixinyu Xu, and Breck Baldwin. 2025. *Non-determinism of "deterministic" llm settings*. *Preprint*, arXiv:2408.04667.
- Shuai Bai, Keqin Chen, Xuejing Liu, Jialin Wang, Wenbin Ge, Sibao Song, Kai Dang, Peng Wang, Shijie Wang, Jun Tang, Humen Zhong, Yuanzhi Zhu, Mingkun Yang, Zhaohai Li, Jianqiang Wan, Pengfei Wang, Wei Ding, Zheren Fu, Yiheng Xu, and 8 others. 2025. *Qwen2.5-vl technical report*. *Preprint*, arXiv:2502.13923.
- Akhiad Bercovich, Itay Levy, Izik Golan, Mohammad Dabbah, Ran El-Yaniv, Omri Puny, Ido Galil, Zach Moshe, Tomer Ronen, Najeeb Nabwani, Ido Shahaf, Oren Tropp, Ehud Karpas, Ran Zilberstein, Jiaqi Zeng, Soumye Singhal, Alexander Bukharin, Yian Zhang, Tugrul Konuk, and 114 others. 2025. *Llama-nemotron: Efficient reasoning models*. *Preprint*, arXiv:2505.00949.
- RALPH ALLAN BRADLEY and MILTON E. TERRY. 1952. *Rank analysis of incomplete block designs: The method of paired comparisons*. *Biometrika*, 39(3-4):324–345.
- Charlie Chen, Sebastian Borgeaud, Geoffrey Irving, Jean-Baptiste Lespiau, Laurent Sifre, and John Jumper. 2023. *Accelerating large language model decoding with speculative sampling*. *Preprint*, arXiv:2302.01318.
- Yang Chen, Zhuolin Yang, Zihan Liu, Chankyu Lee, Peng Xu, Mohammad Shoeybi, Bryan Catanzaro, and Wei Ping. 2025. *Acereason-nemotron: Advancing math and code reasoning through reinforcement learning*. *arXiv preprint arXiv:2505.16400*.
- Peng Cui and Mrinmaya Sachan. 2025. *Investigating the zone of proximal development of language models for in-context learning*. *Preprint*, arXiv:2502.06990.
- DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu, Qihao Zhu, Shirong Ma, Peiyi Wang, Xiao Bi, Xiaokang Zhang, Xingkai Yu, Yu Wu, Z. F. Wu, Zhibin Gou, Zhihong Shao, Zhuoshu Li, Ziyi Gao, and 181 others. 2025a. *Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning*. *Preprint*, arXiv:2501.12948.
- DeepSeek-AI, Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, Damai Dai, Daya Guo, Dejian Yang, Deli Chen, Dongjie Ji, Erhang Li, Fangyun Lin, Fucong Dai, and 181 others. 2025b. *Deepseek-v3 technical report*. *Preprint*, arXiv:2412.19437.
- Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, and 1 others. 2024. *The Llama 3 herd of models*. *arXiv preprint arXiv:2407.21783*.
- Alena Fenogenova, Artem Chervyakov, Nikita Martynov, Anastasia Kozlova, Maria Tikhonova, Albina Akhmetgareeva, Anton Emelyanov, Denis Shevelev, Pavel Lebedev, Leonid Sinev, Ulyana Isaeva, Katerina Kolomeytseva, Daniil Moskovskiy, Elizaveta Goncharova, Nikita Savushkin, Polina Mikhailova, Anastasia Minaeva, Denis Dimitrov, Alexander Panchenko, and Sergey Markov. 2024. *Mera: A comprehensive llm evaluation in russian*. *arXiv preprint arXiv:2401.04531*.
- Fabian Gloeckle, Badr Youbi Idrissi, Baptiste Rozière, and 1 others. 2024. *Better & faster large language models via multi-token prediction*. *arXiv preprint arXiv:2404.19737*.
- Shiyang Gu and 1 others. 2024. *Sglang: Efficient serving of llms with speculative decoding and continuous batching*. GitHub repository.
- Igor Gusev. 2023. *Saiga: Instruction-tuned russian llama models*. <https://huggingface.co/IlyaGusev/saiga>.
- Dan Hendrycks and 1 others. 2021. *Measuring mathematical problem solving with the math dataset*. *arXiv preprint arXiv:2103.03874*.

- Hugging Face. 2025. [Open r1: A fully open reproduction of deepseek-r1](#).
- Shawn Im and Sharon Li. 2025. [Can dpo learn diverse human values? a theoretical scaling law](#). *Preprint*, arXiv:2408.03459.
- Sam Ade Jacobs, Masahiro Tanaka, Chengming Zhang, Minjia Zhang, Reza Yazdani Aminadabi, Shuaiwen Leon Song, Samyam Rajbhandari, and Yuxiong He. 2024. [System optimizations for enabling training of extreme long sequence transformer models](#). In *Proceedings of the 43rd ACM Symposium on Principles of Distributed Computing, PODC '24*, page 121–130, New York, NY, USA. Association for Computing Machinery.
- Naman Jain, King Han, Alex Gu, Wen-Ding Li, Fanjia Yan, Tianjun Zhang, Sida Wang, Armando Solar-Lezama, Koushik Sen, and Ion Stoica. 2024. [Livecodebench: Holistic and contamination free evaluation of large language models for code](#). *Preprint*, arXiv:2403.07974.
- Diederik P. Kingma and Jimmy Ba. 2017. [Adam: A method for stochastic optimization](#). *Preprint*, arXiv:1412.6980.
- Alexander Kukushkin. 2024. [wildchat-hard-ru](https://github.com/kuk/wildchat-hard-ru). <https://github.com/kuk/wildchat-hard-ru>. GitHub repository.
- Ilya Kuleshov, Pavel Ilin, Nikolay Kompanets, Ksenia Sycheva, and Aleksandr Nikolich. 2025. [Doom: Difficult olympiads of math](#). ArXiv preprint.
- Yuri Kuratov and Alexey Arkipov. 2019. [Adaptation of deep bidirectional multilingual transformers for russian language](#). *arXiv preprint arXiv:1912.11283*.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023a. [Fast inference from transformers via speculative decoding](#). In *Proceedings of the 40th International Conference on Machine Learning (ICML '23)*, pages 19274–19286.
- Yaniv Leviathan, Matan Kalman, and Yossi Matias. 2023b. [Fast inference from transformers via speculative decoding](#). *arXiv preprint arXiv:2211.17192*.
- Ming Li, Yong Zhang, Zhitao Li, Jiuhai Chen, Lichang Chen, Ning Cheng, Jianzong Wang, Tianyi Zhou, and Jing Xiao. 2024a. [From quantity to quality: Boosting llm performance with self-guided data selection for instruction tuning](#). *Preprint*, arXiv:2308.12032.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Tianhao Wu, Banghua Zhu, Joseph E Gonzalez, and Ion Stoica. 2024b. [From crowdsourced data to high-quality benchmarks: Arena-hard and benchbuilder pipeline](#). *arXiv preprint arXiv:2406.11939*.
- Tianle Li, Wei-Lin Chiang, Evan Frick, Lisa Dunlap, Banghua Zhu, Joseph E. Gonzalez, and Ion Stoica. 2024c. [From live data to high-quality benchmarks: The arena-hard pipeline](#).
- Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024d. [Eagle-2: Faster inference of language models with dynamic draft trees](#).
- Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2024e. [Eagle: Speculative sampling requires rethinking feature uncertainty](#). *arXiv preprint arXiv:2401.15077*.
- Yuhui Li, Fangyun Wei, Chao Zhang, and Hongyang Zhang. 2025. [Eagle-3: Scaling up inference acceleration of large language models via training-time test](#). *arXiv preprint arXiv:2503.01840*.
- Chris Yuhao Liu, Liang Zeng, Jiakai Liu, Rui Yan, Jujie He, Chaojie Wang, Shuicheng Yan, Yang Liu, and Yahui Zhou. 2024. [Skywork-reward: Bag of tricks for reward modeling in llms](#). *arXiv preprint arXiv:2410.18451*.
- Chris Yuhao Liu, Liang Zeng, Yuzhen Xiao, Jujie He, Jiakai Liu, Chaojie Wang, Rui Yan, Wei Shen, Fuxiang Zhang, Jiacheng Xu, Yang Liu, and Yahui Zhou. 2025a. [Skywork-reward-v2: Scaling preference data curation via human-ai synergy](#). *arXiv preprint arXiv:2507.01352*.
- Yantao Liu, Zijun Yao, Rui Min, Yixin Cao, Lei Hou, and Juanzi Li. 2025b. [Pairjudge rm: Perform best-of-n sampling with knockout tournament](#). *arXiv preprint arXiv:2501.13007*. In progress work.
- Zihe Liu, Jiashun Liu, Yancheng He, Weixun Wang, Jiaheng Liu, Ling Pan, Xinyu Hu, Shaopan Xiong, Ju Huang, Jian Hu, Shengyi Huang, Johan Obando-Ceron, Siran Yang, Jiamang Wang, Wenbo Su, and Bo Zheng. 2025c. [Part i: Tricks or traps? a deep dive into rl for llm reasoning](#). *Preprint*, arXiv:2508.08221.
- Keming Lu, Hongyi Yuan, Zheng Yuan, Runji Lin, Junyang Lin, Chuanqi Tan, Chang Zhou, and Jingren Zhou. 2024. [#instag: Instruction tagging for analyzing supervised fine-tuning of large language models](#). In *The Twelfth International Conference on Learning Representations*.
- Saumya Malik, Valentina Pyatkin, Sander Land, Jacob Morrison, Noah A. Smith, Hannaneh Hajishirzi, and Nathan Lambert. 2025. [Rewardbench 2: Advancing reward model evaluation](#). *Preprint*, arXiv:2506.01937.
- Valentin Mamedov, Evgenii Kosarev, Gregory Leyletner, Ilya Shchuckin, Valeriy Berezovskiy, Daniil Smirnov, Dmitry Kozlov, Sergei Averkiev, Lukyanenko Ivan, Aleksandr Proshunin, Ainur Israfilova, Ivan Baskov, Artem Chervyakov, Emil Shakirov, Mikhail Kolesov, Daria Khomich, Daria Latortseva, Sergei Porkhun, Yury Fedorov, and 14 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, Vienna, Austria. Association for Computational Linguistics.

- Aleksandr Nikolich, Konstantin Korolev, Sergei Bratchikov, Igor Kiselev, and Artem Shelmanov. 2024. Vikhr: The family of open-source instruction-tuned large language models for russian. *arXiv preprint arXiv:2405.13929*.
- OpenAI, :, Aaron Hurst, Adam Lerer, Adam P. Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, AJ Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, and 401 others. 2024. *Gpt-4o system card*. *Preprint*, arXiv:2410.21276.
- OpenAI. 2025. Introducing o3 and o4-mini. <https://openai.com/index/o3-o4-mini-system-card/>. Accessed: 2025-05-15.
- Aleksandar Petrov, Emanuele La Malfa, Philip Torr, and Adel Biber. 2024. Language model tokenizers introduce unfairness between languages. *arXiv preprint arXiv:2305.15425*.
- Qwen, :, An Yang, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chengyuan Li, Dayiheng Liu, Fei Huang, Haoran Wei, Huan Lin, Jian Yang, Jianhong Tu, Jianwei Zhang, Jianxin Yang, Jiayi Yang, Jingren Zhou, and 25 others. 2025. *Qwen2.5 technical report*. *Preprint*, arXiv:2412.15115.
- Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. 2024. *Direct preference optimization: Your language model is secretly a reward model*. *Preprint*, arXiv:2305.18290.
- David Rein and 1 others. 2023. Gpqa: A google-proof q&a benchmark for large language models. *arXiv preprint arXiv:2311.16452*.
- John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. 2017. *Proximal policy optimization algorithms*. *Preprint*, arXiv:1707.06347.
- Zhihong Shao, Peiyi Wang, Qihao Zhu, Runxin Xu, Junxiao Song, Xiao Bi, Haowei Zhang, Mingchuan Zhang, Y. K. Li, Y. Wu, and Daya Guo. 2024. *Deepseekmath: Pushing the limits of mathematical reasoning in open language models*. *Preprint*, arXiv:2402.03300.
- Oleh Shliachko, Alena Fenogenova, Maria Tikhonova, Vladislav Mikhailov, Anastasia Kozlova, and Tatiana Shavrina. 2023. *mgpt: Few-shot learners go multilingual*. *Preprint*, arXiv:2204.07580.
- Jacob Mitchell Springer, Sachin Goyal, Kaiyue Wen, Tanishq Kumar, Xiang Yue, Sadhika Malladi, Graham Neubig, and Aditi Raghunathan. 2025. *Overtrained language models are harder to fine-tune*. In *Forty-second International Conference on Machine Learning*.
- Yang Sui, Yu-Neng Chuang, Guanchu Wang, Jiamu Zhang, Tianyi Zhang, Jiayi Yuan, Hongyi Liu, Andrew Wen, Shaochen Zhong, Na Zou, Hanjie Chen, and Xia Hu. 2025. *Stop overthinking: A survey on efficient reasoning for large language models*. *Preprint*, arXiv:2503.16419.
- T-Tech. 2025a. ru-alpaca-eval. <https://huggingface.co/datasets/t-tech/ru-alpaca-eval>. Hugging Face Datasets; accessed 2025.
- T-Tech. 2025b. ru-arena-hard. <https://huggingface.co/datasets/t-tech/ru-arena-hard>. Hugging Face Datasets; accessed 2025.
- T-Tech. 2025c. ru-mt-bench. <https://huggingface.co/datasets/t-tech/ru-mt-bench>. Hugging Face Datasets; accessed 2025.
- T-Tech. 2025d. ru-reasoning-benchmarks. <https://huggingface.co/collections/t-tech/ru-reasoning-benchmarks>. Hugging Face Datasets; accessed 2025.
- T-Tech. 2025e. ruaime-2024. <https://huggingface.co/datasets/t-tech/ruAIME-2024>. Hugging Face Datasets; accessed 2025.
- T-Tech. 2025f. rummlu-pro. <https://huggingface.co/datasets/t-tech/ruMMLU-pro>. Hugging Face Datasets; accessed 2025.
- T-Tech. 2025g. T-math. <https://huggingface.co/datasets/t-tech/T-math>. Hugging Face Datasets; accessed 2025.
- Sijun Tan, Siyuan Zhuang, Kyle Montgomery, William Yuan Tang, Alejandro Cuadron, Chengguang Wang, Raluca Popa, and Ion Stoica. 2025. *Judgebench: A benchmark for evaluating LLM-based judges*. In *The Thirteenth International Conference on Learning Representations*.
- Gemma Team, Aishwarya Kamath, Johan Ferret, Shreya Pathak, Nino Vieillard, Ramona Merhej, Sarah Perrin, Tatiana Matejovicova, Alexandre Ramé, Morgane Rivière, Louis Rouillard, Thomas Mesnard, Geoffrey Cideron, Jean bastien Grill, Sabela Ramos, Edouard Yvinec, Michelle Casbon, Etienne Pot, Ivo Penchev, and 197 others. 2025. *Gemma 3 technical report*. *Preprint*, arXiv:2503.19786.
- Mikhail Tikhomirov and Daniil Chernyshev. 2024. *Facilitating large language model russian adaptation with learned embedding propagation*. *arXiv preprint arXiv:2412.21140*.
- Tianchun Wang, Zichuan Liu, Yuanzhou Chen, Jonathan Light, Weiyang Liu, Haifeng Chen, Xiang Zhang, and Wei Cheng. 2025a. *On the effect of sampling diversity in scaling llm inference*. *Preprint*, arXiv:2502.11027.

- Yizhong Wang, Yeganeh Kordi, Swaroop Mishra, and 1 others. 2023. Self-instruct: Aligning language models with self-generated instructions. *arXiv preprint arXiv:2212.10560*.
- Zhilin Wang, Jiaqi Zeng, Olivier Delalleau, Hoo-Chang Shin, Felipe Soares, Alexander Bukharin, Ellie Evans, Yi Dong, and Oleksii Kuchaiev. 2025b. [HelpSteer3-Preference: Open human-annotated preference data across diverse tasks and languages](#). *Preprint*, arXiv:2505.11475.
- Wei Xiong, Jiarui Yao, Yuhui Xu, Bo Pang, Lei Wang, Doyen Sahoo, Junnan Li, Nan Jiang, Tong Zhang, Caiming Xiong, and Hanze Dong. 2025. [A minimalist approach to llm reasoning: from rejection sampling to reinforce](#). *Preprint*, arXiv:2504.11343.
- An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, Chang Gao, Chengen Huang, Chenxu Lv, Chujie Zheng, Dayiheng Liu, Fan Zhou, Fei Huang, Feng Hu, Hao Ge, Haoran Wei, Huan Lin, Jialong Tang, and 41 others. 2025. [Qwen3 technical report](#). *Preprint*, arXiv:2505.09388.
- An Yang, Baosong Yang, Binyuan Hui, and 1 others. 2024. Qwen2 technical report. *arXiv preprint arXiv:2407.10671*.
- Xiang Yue, Tuney Zheng, Ge Zhang, and Wenhui Chen. 2024. [Mammoth2: Scaling instructions from the web](#). In *Advances in Neural Information Processing Systems*, volume 37, pages 90629–90660. Curran Associates, Inc.
- Yifan Zhang and Team Math-AI. 2024. American invitational mathematics examination (aime) 2024.
- Yifan Zhang and Team Math-AI. 2025. American invitational mathematics examination (aime) 2025.
- Weilin Zhao, Tengyu Pan, Xu Han, Yudi Zhang, Ao Sun, Yuxiang Huang, Kaihuo Zhang, Weilun Zhao, Yuxuan Li, Jie Zhou, Hao Zhou, Jianyong Wang, Zhiyuan Liu, and Maosong Sun. 2025. [FR-spec: Accelerating large-vocabulary language models via frequency-ranked speculative sampling](#). In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3909–3921, Vienna, Austria. Association for Computational Linguistics.
- Dmitry Zmitrovich, Alexander Abramov, Andrey Kalmykov, Maria Tikhonova, Ekaterina Taktasheva, Danil Astafurov, Mark Baushenko, Artem Snegirev, Vitalii Kadulin, Sergey Markov, Tatiana Shavrina, Vladislav Mikhailov, and Alena Fenogenova. 2023. A family of pretrained transformer language models for russian. *arXiv preprint arXiv:2309.10931*.

A Released Resources and Licenses

All released components use permissive, research-friendly licenses. The T-pro 2.0 model, its EA-GLUE draft weights, and the T-Math benchmark are distributed under Apache-2.0, allowing broad academic and commercial use. The T-Wix 500k corpus is released under ODC-By. Full license details appear in Table 6.

B Tokenizer adaptation statistics

Russian and English corpora. Table 7 reports tokenization statistics for Russian and English on both Wikipedia and our in-domain SFT corpus (T-Wix). For Russian, the Cyrillic-dense T-pro 2.0 tokenizer substantially reduces the average number of tokens per word and increases the share of words segmented into at most two tokens, while English compression is essentially unchanged.

Table 8 extends this analysis to eight Cyrillic languages using Wikipedia. In all cases the new tokenizer reduces tokens per word, with particularly large gains for Kyrgyz, which is poorly served by generic multilingual tokenizers.

Comparison with other Cyrillic-rich tokenizers. Finally, Table 9 compares T-pro 2.0 against several strong Cyrillic-focused baselines. T-pro 2.0 achieves the lowest tokens-per-word for seven out of eight languages (ru, uk, be, bg, sr, mk, ky) and remains competitive on Kazakh, demonstrating that our tokenizer design is competitive with specialized alternatives.

C Instructional midtraining

We employ an intermediate *instructional midtraining* stage between generic large-scale pre-training and downstream alignment. Starting from the publicly available Qwen3-32B dense model (Yang et al., 2025), already pre-trained on $\sim 36T$ tokens, we perform continual training on 40B tokens of instruction-style data. The goals of this stage are: (i) adapt the model to a denser, Russian-centric tokenizer, (ii) learn useful representations for new subword units, and (iii) further strengthen Russian language and reasoning skills without degrading the base model’s capabilities.

Training setup. Midtraining uses a 4M-token global batch and 40B total tokens ($\approx 9,750$ steps). We train with AdamW using a peak learning rate of 1×10^{-5} and cosine decay to 1×10^{-6} , with 100 warmup batches. Data are formatted in the same

chat-style schema and packed up to a 32K context window without a length curriculum. Training is performed in bf16 with FSDP full-shard and activation checkpointing. Remaining hyperparameters are listed in Table 10.

Midtraining datamix All midtraining samples are in instruction format. The datamix combines (i) public instruction datasets from the Hugging Face Hub, (ii) web and forum data (e.g., question-answer style threads), (iii) real user–assistant dialogues, and (iv) synthetic instructional data and reasoning traces grounded in pre-training corpora (books, Common Crawl, code) via a WebInstruct-style pipeline (Yue et al., 2024). Compared to the SFT stage, the midtraining datamix is intentionally *larger and less curated*: we trade some noise for broader coverage of tasks and domains. All instructions are derived from public sources; internal data are used only as anonymized targets in dialogue-style responses.

Table 11 reports the category-level breakdown of the 40B-token corpus. In terms of language, the corpus is predominantly Russian and English: roughly 49% of tokens are Russian (19.6B), 36% English (14.4B), 9.3% code (3.7B) and 5.5% come from parallel Russian–English data (2.2B).

InsTag deduplication. To control redundancy while preserving diversity across sources, we apply #INSTAG-based deduplication (Lu et al., 2024) *independently within each component* of the datamix (reasoning, general QA, code, etc.). The tagger is applied only to user utterances; all tags from a multi-turn sample are unioned into a single tag set. We then perform exact-match and semantic deduplication at the tag level, followed by greedy diversity sampling over tagged samples. This procedure gives macro-level control over the balance between different categories instead of deduplicating the raw pool as a whole. On large components such as reasoning and general QA, only about 10–30% of the raw candidates are retained (the remaining 70–90% are discarded), whereas for smaller, less repetitive sources we keep 80–90% of samples. For each retained sample, the final assistant turn is regenerated with a stronger teacher, Qwen3-235B, which improves answer quality and stylistic consistency while keeping the original user input and context intact.

Ablations on datamix design We compare two variants of the midtraining corpus, both trained

Resource	Type	Location	License
T-pro 2.0	Model	https://huggingface.co/t-tech/T-pro-it-2.0	Apache-2.0
EAGLE weights	Model component	https://huggingface.co/t-tech/T-pro-it-2.0-eagle	Apache-2.0
T-Math	Benchmark dataset	https://huggingface.co/datasets/t-tech/T-math	Apache-2.0
T-Wix 500k	Instruction corpus	https://huggingface.co/datasets/t-tech/T-wix	ODC-By

Table 6: Released resources and licenses.

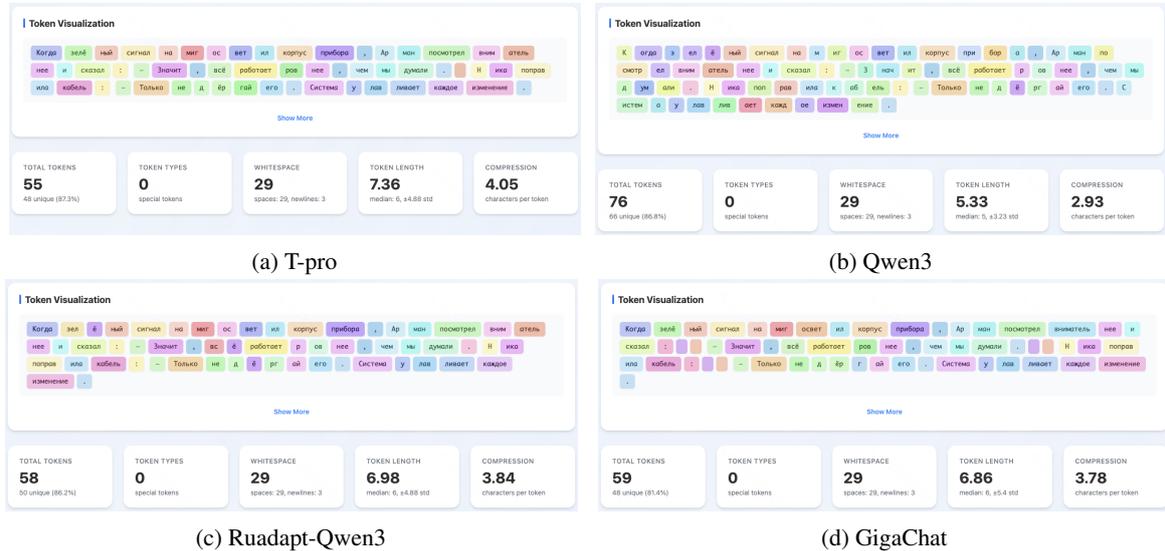


Figure 2: Qualitative comparison of Russian tokenization. A 220-character text is tokenized by T-pro 2.0, the original Qwen3 tokenizer, and other Cyrillic-optimized models. T-pro 2.0 encodes the text into just 55 tokens compared to 76 for Qwen3, demonstrating superior compression efficiency.

Corpus	Tokenizer	tok/ word	1 tok (%)	≤ 2 tok (%)	> 2 tok (%)
<i>Russian</i>					
ruWiki	Qwen3	3.12	20.3	38.2	61.8
ruWiki	T-pro 2.0	2.38	28.7	60.1	39.9
T-Wix	Qwen3	2.70	31.8	52.4	47.6
T-Wix	T-pro 2.0	2.26	39.3	65.5	34.5
<i>English</i>					
enWiki	Qwen3	1.68	61.2	83.7	16.3
enWiki	T-pro 2.0	1.68	61.1	83.7	16.3

Table 7: Tokenization density statistics for Russian and English on Wikipedia and our SFT corpus (T-Wix). We compare the original Qwen3 and Cyrillic-dense T-pro 2.0 tokenizers. Columns show: average tokens per word (tok/ word), percentage of words tokenized into exactly 1 token, at most 2 tokens, and more than 2 tokens.

Lang	Tokens/Word		% Words (≤ 2 tok)	
	Qwen3	T-pro	Qwen3	T-pro
ru	3.12	2.38	38.20	60.13
uk	3.70	2.80	31.17	45.79
be	3.97	2.94	30.15	41.36
bg	2.99	2.35	43.42	59.60
sr	3.26	2.62	37.65	51.79
mk	3.04	2.41	42.42	57.19
kk	4.60	3.07	15.30	37.69
ky	4.35	3.09	21.27	39.93

Table 8: Tokenization density on Wikipedia for Cyrillic languages. *Tokens/Word*: average tokens per word; *% Words (≤ 2 tok)*: percentage of words tokenized into at most 2 tokens.

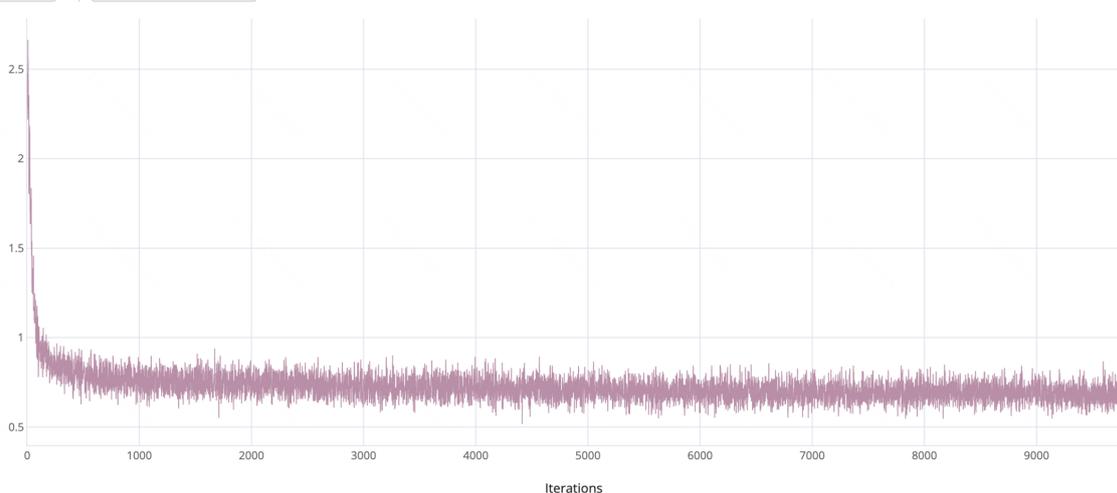


Figure 3: Midtraining training loss as a function of optimization steps. Loss drops steeply during the first $\sim 1k$ steps ($\sim 4B$ tokens) and then gradually plateaus, indicating that most adaptation to the new tokenizer happens early in the run.

Lang	T-pro	GigaChat [†]	Ruadapt-Qwen3	gpt-oss
ru	2.38	2.49	2.43	2.70
uk	2.80	3.09	3.29	2.92
be	2.94	3.32	3.54	3.03
bg	2.35	2.58	2.50	2.56
sr	2.62	2.97	3.07	2.73
mk	2.41	2.67	2.70	2.59
kk	3.07	2.67	4.60	3.11
ky	3.09	3.33	3.97	3.17
Avg	2.71	2.89	3.26	2.85

Table 9: Tokenization density (tokens/word) on Wikipedia for T-pro and other Cyrillic-dense tokenizers. Lower is better. [†]Indicates <https://huggingface.co/ai-sage/GigaChat-20B-A3B-instruct> model.

for 40B tokens with identical optimization settings (Table 10):

- **Pre-train + instruct:** mixture including generic pre-training-style data (Common Crawl, Wikipedia, code) alongside instruction-formatted examples.
- **Instruct-only:** the same instruction pool but without additional raw pre-training sources, i.e., all examples follow an explicit instruction–response schema.

The instruct-only variant thus allocates more of the 40B-token budget to high-quality instruction data, whereas the mixed variant spends a fraction of tokens on generic web/code continuation. We evaluate both models on a suite of Russian and multilingual math/reasoning benchmarks, including ru-

Hyperparameter	Value
Global batch size (tokens)	4M (128 seq \times 32K)
Total tokens	40B
Steps	$\approx 9,750$
Max context length	32K
Optimizer	AdamW
Adam betas	(0.9, 0.95)
Adam ϵ	10^{-8}
Weight decay	10^{-6}
LR schedule	cosine
Peak / min LR	$1 \times 10^{-5} / 1 \times 10^{-6}$
Warmup	100 batches
Gradient clipping	max norm 1.0
Precision	BF16
Parallelism	FSDP full-shard, act. checkpointing

Table 10: Midtraining optimization setup.

Category	Share	Tokens (B)
Reasoning	34.5%	13.8
General	29.3%	11.7
Math	16.3%	6.5
Real chat	5.5%	2.2
IF	5.0%	2.0
Grounded QA synth	3.8%	1.5
Code	2.8%	1.1
Forum	1.7%	0.7
Summarization	0.7%	0.3
ICL	0.4%	0.2

Table 11: Midtraining datamix by category (40B tokens total).

AIME’24/25, ruMATH500, ruGPQA, ruLCB, T-Math, and Arena-style pairwise comparisons. All evaluations are zero-shot; AIME-style metrics are computed as avg@8 over 30 problems, and other benchmarks are run once due to computational

cost.

Table 12 shows a representative subset of metrics. Across most math and reasoning benchmarks, the instruct-only datamix outperforms or matches the mixed variant despite using the same token budget, and even early checkpoints from the instruct-only run are ahead of the mixed model. This is consistent with recent evidence that heavily pre-trained models are harder to adapt via continual pre-training (Springer et al., 2025), especially when the additional data distribution differs from downstream tasks.

Benchmark	PT+I	I-only
ruAIME'24	0.60	0.67
ruAIME'25	0.47	0.63
ruMATH500	0.93	0.94
ruGPQA	0.58	0.66
ruLCB	0.53	0.55
T-Math	0.49	0.50
Arena hard (think)	43.7	44.5
Arena wildchat ru (think)	55.0	55.1

Table 12: Ablation on midtraining datamix design (zero-shot). “PT+I” denotes the pre-training+instruct mixture; “I-only” uses only instruction-formatted data.

We did not run additional ablations such as training on original (non-regenerated) answers or disabling InsTag deduplication. In the first case, instruction data come from heterogeneous sources with uneven answer quality and formats, and we found it undesirable to train on such out-of-distribution completions. In the second case, skipping deduplication would require regenerating answers for a much larger pool of raw samples, significantly increasing computational cost; we leave this exploration for future work.

Tokenizer adaptation and MERA results A key objective of midtraining is to adapt the model to a new, denser tokenizer for Russian without degrading downstream quality. To quantify the impact of the tokenizer choice, we train two 8B models on the same midtraining datamix with identical optimization hyperparameters, differing only in the tokenizer (original Qwen3 vs. T-pro 2.0).

Table 13 reports MERA scores for these two variants. The T-pro 2.0 tokenizer attains a macro-average score comparable to the original one (0.574 vs. 0.560), with only small per-task differences in both directions. In other words, replacing the tokenizer with a denser Cyrillic segmentation does not degrade general Russian-language performance

on MERA, which is the primary design goal of midtraining.

Task	Qwen tokenizer	T-pro 2.0 tokenizer
USE	0.198	0.191
MaMuRaMu	0.784	0.796
ruWorldTree	0.966/0.966	0.966/0.966
ruCodeEval	0.173/0.45/0.585	0.454/0.689/0.756
RCB	0.557/0.479	0.564/0.47
MathLogicQA	0.710	0.731
ruOpenBookQA	0.897/0.897	0.922/0.923
RWSD	0.446	0.250
CheGeKa	0.30/0.368	0.31/0.384
LCS	0.102	0.096
PARUS	0.868	0.912
MultiQ	0.381/0.517	0.344/0.478
ruMultiAr	0.402	0.400
ruTiE	0.788	0.798
ruModAr	0.515	0.627
AVG	0.560	0.574

Table 13: MERA scores for an 8B model with the original Qwen3 tokenizer and the Cyrillic-dense T-pro 2.0 tokenizer. Bold marks the better value per row.

The midtraining loss curve in Figure 3 further illustrates the adaptation process. Training loss decreases sharply over the first \approx 1k steps (\approx 4B tokens) before plateauing, indicating that a substantial token budget at a relatively high learning rate is required to adapt the model to the new tokenizer beyond what the smaller, lower-LR SFT budget alone could provide.

D T-Wix SFT dataset

D.1 General part of T-Wix

The general part of the dataset consists of 468k diverse prompts collected from open-source data and high-quality translations of English-language datasets, subsequently deduplicated. The dataset is assembled to enhance the model’s capabilities in coding, mathematics, dialogue, and other competencies expected from a modern LLM.

D.1.1 Data Preparation

First and foremost, a corpus of 14M instructions (mostly in English) is compiled from various open-source datasets. To select the most useful samples, a data filtering pipeline is developed. It consists of several consecutive stages aimed at deduplication and ensuring high thematic, qualitative, and complexity diversity of the SFT dataset. We also perform deduplication against benchmark datasets to ensure that no benchmark examples leak into the training corpus.

LSH and Embedding-Based Deduplication. At the initial stage, simple deduplication is performed using locality-sensitive hashing (LSH) and embedding-based similarity search to eliminate duplicated samples originating from different open-source datasets.

Thematic Tag Filtering. To ensure thematic balance, the #INSTAG-based filtering approach (Lu et al., 2024) is employed. The pipeline uses a trained tagging model to extract thematic tags from each instruction.

For the present work, the tagger is trained using the Qwen2.5-7B (Qwen et al., 2025) model on multilingual data with a context length of up to 32k tokens, allowing tagging in both Russian and English, including long-context data. This modification substantially reduces translation overhead, as tagging and filtering can be applied directly to multilingual raw data without prior translation into English.

To improve thematic balance, an additional domain-balancing stage—“Domain & Complexity Balancing”—is included, as the tagger primarily produces low-level thematic annotations.

Domain & Complexity Balancing. In addition to fine-grained thematic filtering, a higher-level balance is introduced across major knowledge domains and difficulty levels within each domain. To achieve this, six domains — *Math, Code & Programming, Science, General Instruct, General Knowledge, Writing* — and three complexity levels — *School, Student, Professor* — are defined.

Using large-scale LLM-assisted annotation, approximately 14M samples are automatically labeled with both domain and complexity tags. Subsequently, the dataset is balanced across domains and further normalized by difficulty within each domain to regulate the model’s output capabilities and ensure a uniform skill distribution.

This stage enables finer control over the resulting model’s generalization behavior, preventing over-representation of specific topics or difficulty levels.

Reward Model Filtering. In the subsequent stage, samples are filtered according to prompt quality using scores from the Reward Model (RM) described in F. For each of the datasets comprising the 14M instructions, an RM score is computed, and the bottom 10% of samples with the lowest scores within each dataset are filtered out.

This step effectively removes “noisy” or low-quality samples that could negatively impact downstream model performance, preserving only high-quality and instructionally meaningful examples.

Instruction Following Difficulty Filtering. A further filtering stage based on Instruction Following Difficulty (IFD) scores is incorporated, following the approach introduced in (Li et al., 2024a). These scores quantify the difficulty a language model faces in following a given instruction. For the present work, IFD scores are computed relative to a midtraining checkpoint to reflect the model’s actual instruction-following capability. Samples with excessively high IFD values (>1.0) are discarded as overly complex or ambiguous, while those with very low IFD scores (<0.7) are filtered out as trivially simple.

This selective filtering makes it possible to retain the most challenging and instructionally rich examples—those that contribute most to improving the model’s instruction-following ability—while removing both overly simple and excessively difficult samples.

Multilingual Filtering and Translation. The multilingual setup enables filtering to be conducted directly on mixed-language raw data, reducing both the cost and time associated with preliminary translation. Only the final curated dataset is translated into Russian to ensure cross-lingual consistency.

Rejection Sampling and Generation. High dataset quality is further ensured through the use of top LLMs and rejection sampling. Each final training completion is produced using DeepSeek-V3 and Qwen-235B-A22B models, generating 8 candidate responses per instruction. These candidates are then filtered using RM scores to select the highest-quality outputs.

This approach not only eliminates translation artifacts present in the raw multilingual data but also results in substantially higher-quality responses compared to the original samples, thereby improving the overall consistency and instructional value of the dataset.

Overall, the combined multistage filtering pipeline ensures that the final SFT dataset is diverse, balanced, and composed of high-quality, instructionally valuable samples, free from data leakage and redundancy.

This approach allows the training process to remain balanced across domains (e.g., code and

math) without bias toward any particular category.

D.2 Long Context

To enhance the model’s ability to process extended inputs, a dedicated long-context dataset is constructed.

A diverse collection of long texts is selected from publicly available data for the pre-training phase, covering various domains such as *education*, *technology*, *business*, *scientific literature*, and *fiction*. The dataset is distributed across multiple context lengths ranging from 8k to 32k tokens, enabling the model to learn robustly across different input sizes. Using DeepSeek-V3 and Qwen-235B-A22B, a variety of prompts and responses are generated for each text, encompassing summarization, open- and closed-domain QA tasks, as well as reasoning-oriented datasets.

The resulting long-context dataset increment constitutes about 1% of the total SFT training data in samples and 7.7% in tokens, providing valuable coverage for instruction tuning under extended context conditions.

D.3 Parallel Corpora

To maintain strong English proficiency, parallel corpora are added to the SFT dataset — that is, instructional samples presented in English alongside their Russian counterparts.

A series of experiments on the language share within the dataset shows that an optimal ratio is approximately 10% English data relative to the total SFT mix.

D.4 Reasoning part of T-Wix

To enhance reasoning capabilities in Large Language Models (LLMs) for the Russian language, a high-quality, reasoning-focused dataset is constructed through a targeted distillation pipeline. Rather than maximizing data volume, the pipeline prioritizes instructional value and appropriate task difficulty, ensuring that each retained sample provides meaningful learning potential for the target student model. The process starts from a broad collection of English-language reasoning instructions, which are subsequently translated into Russian, deduplicated, and carefully balanced across domains to support diverse and robust reasoning behaviors.

Initial Pool Generation and Deduplication. The initial pool of data is constructed from approx-

imately 450k high-quality English-language reasoning instructions, drawn from established open-source datasets (e.g., Open-R1 (Hugging Face, 2025), Nvidia/AceReason-Math (Chen et al., 2025), Nvidia/Nemotron (Bercovich et al., 2025)), covering general knowledge, mathematics, natural sciences, and code generation.

Domain Distribution:

- 60% general knowledge and open-ended reasoning (to establish fluent, structured reasoning in Russian),
- 10% verifiable mathematics (e.g., arithmetic, algebra),
- 10% open-ended mathematics (e.g., proofs, conceptual explanations),
- 15% natural sciences (physics, chemistry, biology),
- 5% code-related reasoning.

After the initial collection, these English-language instructions are translated into Russian. As in the general part of T-Wix, deduplication is applied to eliminate near-duplicates and ensure sample uniqueness.

Reward-Based Completion Evaluation. To mitigate the stochasticity inherent in LLM generation (Wang et al., 2025a; Atil et al., 2025), 8 diverse completions are generated per instruction by both:

- The teacher model (Qwen3-235B-A22B),
- The student model (midtraining checkpoint).

This yields 16 completions per instruction, which are independently scored by a trained reward model (RM). The inclusion of student generations enables direct assessment of the model’s current reasoning capability on each task, while the teacher generations provide high-quality reference behaviors. This multi-generation approach provides a more robust statistical picture of the model’s performance on each instruction.

Statistical Filtering Based on Reward Stability. Instructions exhibiting high variance in RM scores across generations are discarded, as they reflect ambiguous or unstable evaluation signals. Additionally, instructions for which the student model consistently receives very low RM scores—even

with low variance—are excluded, as they lie beyond the student’s current learning capacity and are unlikely to support effective knowledge transfer (Xiong et al., 2025; Liu et al., 2025c).

Mean Reward-Based Selection Within the Zone of Proximal Development. To operationalize the pedagogical principle of the zone of proximal development (ZPD) (Cui and Sachan, 2025), the RM scores for the 8 teacher and 8 student completions per instruction are aggregated by computing their respective means. The average reward of the teacher responses and the average reward of the student responses are then used to estimate the reasoning gap.

Samples are selected based on the absolute difference between these mean rewards. A small difference indicates that the student already performs comparably to the teacher (suggesting limited learning potential), whereas a very large difference implies that the task lies beyond the student’s capabilities (making distillation ineffective). Only samples with a moderate gap in mean RM scores—neither too small nor too large—are retained. This ensures that the selected samples are challenging enough to drive improvement, yet sufficiently within reach for successful knowledge transfer.

Final Completion Selection. For the final training targets, teacher-generated completions are selected as follows:

- For *verifiable* instructions (e.g., mathematical problems), factually incorrect completions are first filtered out; among the remaining correct ones, the completion with the highest RM score is chosen.
- For *open-ended* instructions, the shortest reasoning trace among the top-3 RM-ranked teacher completions is selected. This encourages the student to learn concise and non-redundant reasoning patterns (Sui et al., 2025).

This pipeline yields a high-quality reasoning dataset of approximately 30k samples, consisting of 90% Russian and 10% English instructions, consistent with the overall language strategy of T-Wix. By design, the dataset emphasizes stable, diverse, and pedagogically optimal reasoning traces in Russian across multiple domains, effectively balancing task difficulty with learning potential.

D.5 T-Wix dataset analytics

The final distribution of data in the SFT dataset (T-Wix) are presented in Figure 4. The total size is 500k samples.

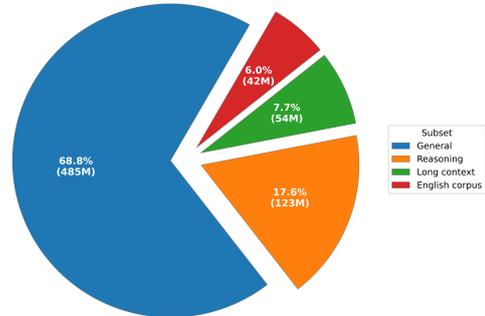


Figure 4: Token distribution in T-Wix. Token counts were computed using tiktoken with o200k base tokenizer.

D.6 SFT Training Recipe

The SFT stage took 9 hours on 4 nodes with 8xH100 GPUs for T-pro 2.0, using gradient checkpointing and FSDP, as well as packing samples into 32k-token contexts without truncation. After a series of experiments, the optimal fine-tuning hyperparameters were selected, as described in Table 14.

Hyperparameter	Value
Global batch size (samples)	32
Max context length	32k
Number of training epochs	2
Optimizer	Adam (Kingma and Ba, 2017)
Adam betas	(0.9, 0.95)
Adam ϵ	10^{-12}
Learning rate	1e-6
Learning rate scheduler	cosine
Warmup ratio	0.1
Gradient clipping	max norm 2.0
Precision	BF16

Table 14: Hyperparameters used for SFT training.

E Preference tuning

To enhance alignment beyond supervised fine-tuning, an on-policy Direct Preference Optimization (DPO) procedure is applied. Recent work shows that on-policy preference optimization offers more stable and reliable alignment gains than off-policy alternatives, as it learns directly from the model’s own generative distribution and therefore avoids distribution shift while targeting realistic error modes (Rafailov et al., 2024; Im and Li, 2025).

For each instruction, the SFT-trained model produces 16 completions. All candidates are scored using the RM described in App. F, and preference pairs are constructed by selecting the highest- and lowest-scoring completions. This contrastive selection yields stable and informative training pairs by removing low-signal, ambiguous comparisons.

The DPO dataset is constructed from filtered SFT data (T-Wix). A total of 100k preference pairs is formed, consisting of:

- 90k sampled from the General SFT part,
- 10k sampled from the Reasoning SFT part.

In addition, cross-subset augmentation is applied to enrich preference diversity: 4k samples from the General subset are paired with reasoning-style reformulations, while 6k samples from the Reasoning subset are converted into general-style instructions. This yields a smoother distribution of reasoning complexity without altering the intended emphasis of each subset.

The resulting on-policy DPO stage improves the model’s alignment, coherence, and reasoning structure while preserving broad general-purpose capabilities.

E.1 Preference Training Recipe

The DPO stage required 28 hours of training on 4 nodes with 8×H100 GPUs. The training was carried out using sequence parallelism, which enabled efficient distribution of computation across devices. The hyperparameters listed in Table 15 were identified as optimal.

Hyperparameter	Value
Global batch size (samples)	128
Max context length	32k
Number of training epochs	1
Optimizer	AdamW
Adam betas	(0.9, 0.95)
Adam ϵ	10^{-12}
Weight decay	0.01
Learning rate	1e-7
Learning rate scheduler	cosine
Warmup ratio	0.05
Gradient clipping	max norm 2.0
Precision	BF16
Loss type	DPO
DPO beta	0.5

Table 15: Hyperparameters used for preference tuning

F Reward Model

Tournament-Based Synthetic Preference Data Generation To construct a high-quality reward model, it is essential to obtain reliable preference data —pairs of model completions ranked according to their relative quality. Direct pairwise annotation across all available completions, however, is computationally expensive and inefficient. To address this, similar to the knockout-tournament method introduced by Liu et al. (2025b), we propose a tournament-based preference generation approach that substantially reduces the number of required comparisons while preserving the informativeness of the resulting preference signal.

Each tournament comprises n participants, randomly sampled from the pool of available models. For each instruction, every model generates a completion, and the tournament bracket is constructed according to model category —for instance, small-scale models (7B–13B) are paired against models of similar scale, and reasoning-oriented models compete within the same subclass. This grouping strategy ensures that comparisons are made between models of comparable generative quality, encouraging the reward model to learn fine-grained distinctions rather than relying on trivial cases where one output is clearly superior (e.g., when a large model is compared with a small size model, a comparison might yield an obvious outcome —the larger model would consistently produce more coherent and contextually appropriate responses, leaving little room for the reward model to learn subtle differences).

Each round of the tournament consists of a single instruction and the corresponding completions generated by the competing models. An external LLM, not participating in the tournament, is employed as judge to determine the preferred completion for each matchup. To avoid positional bias, each pair of completions is evaluated in both possible orders, and samples exhibiting positional bias are excluded from the final training set.

At the completion of each single-elimination tournament with n participants, a total of $\frac{n}{2} \log_2 N$ preference pairs are obtained. This result comes from the hierarchical structure of the tournament: in each round, half of the remaining participants compete, producing $\frac{n}{2}$ new pairwise outcomes (both direct and transitive). Because a tournament with n participants requires $\log_2 N$ rounds to determine a winner, the total number of inferred prefer-

ence pairs accumulates to $\frac{n}{2} \log_2 N$.

Each round contributes the same number of new known preferences because every winner’s new victory also establishes transitive relationships over all opponents defeated in earlier rounds. For instance, if player A beats player B in the final, it is implied that A outperform every player that B previously defeated. Consequently, even though only $n - 1$ matches are directly played, the tree-like transitive structure allows many additional indirect comparisons to be inferred.

This process produces preference set dense enough to capture comparative information among many participants, yet far more efficient than exhaustively comparing every possible pair (which would require $\frac{n(n-1)}{2}$ comparisons). This tournament-based approach yields an informative preference dataset while significantly reducing annotation complexity.

Reward Model Training The reward model is based on Qwen3-32B (Yang et al., 2025) with a regression head to produce a single preference score for each completion. Training follows the Bradley–Terry (BRADLEY and TERRY, 1952) formulation, which models the probability of one completion being preferred over another as a logistic function of their respective scores. All training is conducted with a maximum sequence length of 32k tokens, leveraging Ulysses sequence parallelism (Jacobs et al., 2024) to efficiently support long-context optimization. Data preprocessing, batching, and distributed training are managed through the TurboAlignment library (tur, 2025).

Evaluation For intrinsic evaluation, we adapt RewardBench 2 (Malik et al., 2025) to Russian by translating the original benchmark and report standard leaderboard metrics. For downstream evaluation, we additionally construct a Best-of- N selection benchmark on top of the Arena-Hard-RU instruction set to assess the reward model under realistic generation scenarios. In this setting, the base model produces N candidate completions per instruction, and the reward model selects the highest-scoring (best@ N) and lowest-scoring (worst@ N) outputs. These selections are then evaluated using Arena-Hard, allowing us to measure the alignment between reward-model rankings and externally validated quality. We further report the Δ_{BoN} metric (best@ $N - \text{worst@}N$) to quantify discriminative capacity. Although our model performs compara-

bly to existing open-source reward models on the translated RewardBench 2, it demonstrates a clear advantage on our Best-of- N Arena-Hard benchmark. As shown in Table 16, our model obtains the highest Δ_{BoN} score, reflecting the strongest separation between high- and low-quality completions.

RM-model	best@8↑	worst@8↓	Δ_{BoN}
Qwen3-32B-RM (Ours)	92.69 (-0.99)	70.48 (+2.34)	22.21
Llama-3.3-Nemotron-70B-Reward-Multilingual ¹	85.93 (-1.93)	84.91 (+1.85)	1.02
Skywork-Reward-Gemma-2-27B ²	89.05 (-1.6)	74.35 (+2.07)	14.70
Skywork-Reward-V2-Llama-3.1-8B ³	90.49 (-1.43)	77.31 (+1.77)	13.18
Llama-3.1-Tulu-3-70B-SFT-RM-RB2 ⁴	87.37 (-1.86)	78.47 (+1.76)	8.90

Table 16: Best-of- N ($N = 8$) evaluation on Arena-Hard-RU. We report win rates for the highest- (best@8) and lowest-scoring (worst@8) completions selected by each reward model, and their difference $\Delta_{\text{BoN}} = \text{best@8} - \text{worst@8}$, which measures discriminative capacity. ¹Wang et al. (2025b), ²Liu et al. (2024), ³Liu et al. (2025a), ⁴Malik et al. (2025)

Prompt selection Furthermore, in the process of synthetic data generation, we evaluated a range of prompting strategies derived from the JudgeBench (Tan et al., 2025). Empirical analysis indicates that the Google Vertex prompt yields superior evaluation quality in different benchmarks (see Table 18), particularly on RewardBench 2 (RU). This improvement underscores the sensitivity of LLM-based evaluators to prompt design and highlights the importance of selecting domain-appropriate judging configurations for reliable preference data generation.

F.1 Reward Model Analysis

Our experiments reveal that DeepSeek-V3 (DeepSeek-AI et al., 2025b) demonstrates superior judgment capabilities in open-domain and conversational (chat) tasks, whereas Qwen3-235B-A22B exhibits stronger performance in mathematical, code-related and other domains (see Table 17).

Ablation on Transitive Samples. An ablation study was conducted to evaluate the contribution of transitive preference pairs. Removing transitive samples led to a consistent degradation across all evaluation metrics (see Table 17), suggesting that inferred pairwise relationships enrich the preference signal and improve the model’s generalization to unseen instructions. Conversely, adding additional transitive samples beyond the first closure continued to yield marginal but positive improvements.

Category	Model	RewardBench 2 (RU)					
		Fact.	Focus	Math	Prec. IF	Safety	Total
Comparison with Existing RMs	Qwen3-32B-RM (Ours)	0.66	0.87	0.62	0.42	0.89	0.69
	Skywork-Reward-V2-Llama-3.1-8B	0.68	0.88	0.65	0.45	0.79	0.69
	Skywork-Reward-Gemma-2-27B	0.69	0.88	0.64	0.40	0.92	0.7
	Llama-3.1-Tulu-3-70B-SFT-RM-RB2	0.72	0.74	0.69	0.41	0.76	0.66
	Llama-3.3-Nemotron-70B-Reward-Multilingual	0.73	0.85	0.62	0.41	0.86	0.69
Judge Model Ablation	Qwen-3-RM-8B-DeepSeek-V3	0.478	0.756	0.598	0.341	0.736	0.581
	Qwen-3-RM-8B-Qwen3-235B-A22B	0.467	0.324	0.688	0.350	0.840	0.533
Transitive Samples Ablation	Qwen3-8B-RM w/o transitive	0.467	0.324	0.688	0.350	0.840	0.533
	Qwen3-8B-RM w/ transitive	0.505	0.453	0.704	0.413	0.860	0.587

Table 17: Evaluation results on RewardBench 2 (RU). We compare our Qwen3-32B-RM against existing reward models (top), analyze the impact of different judge models for preference annotation (middle), and study the effect of tournament-derived transitive preference samples during training (bottom). Bold indicates best performance within each category.

Prompt	RewardBench 2 (RU)					
	Fact.	Focus	Math	Prec. IF	Safety	Total
Skywork	0.636	0.782	0.834	0.394	0.881	0.706
Arena Hard	0.653	0.638	0.762	0.349	0.899	0.660
Google Vertex	0.741	0.846	0.830	0.549	0.915	0.776
Prometheus 2	0.600	0.622	0.743	0.432	0.790	0.637
Chat-Eval	0.667	0.781	0.795	0.478	0.831	0.710

Table 18: Assessing the role of prompt selection in RewardBench 2 (RU).

Length Sensitivity and Distribution Effects. A further observation concerns the length distribution between chosen and rejected completions. RewardBench 2 (RU) exhibits a substantial drop in evaluation quality when the distribution becomes skewed—specifically, when longer or shorter completions dominate. This imbalance appears to induce a length-based bias in the reward model, leading it to systematically favor responses of a particular size rather than quality.

For instance, Qwen3-235B-A22B as a judge displayed a pronounced length bias, consistently preferring longer completions regardless of their semantic quality. This highlights the importance of maintaining a balanced length distribution during preference data generation and tournament construction to prevent undesirable inductive shortcuts in the reward model.

G Speculative Decoding Implementation

To mitigate the sequential latency of autoregressive generation, we integrate an EAGLE-based speculative decoding module (Li et al., 2024e) into T-pro 2.0. This setup employs a lightweight draft model to propose candidate tokens in parallel, which are subsequently verified by the target model to ensure

the output distribution remains identical to standard decoding (Leviathan et al., 2023a).

Architecture and Objective. Our draft model utilizes a single decoder layer augmented with an FR-Spec component (Zhao et al., 2025), based on the Llama 2 architecture and implemented via SGLang (Gu et al., 2024). Unlike standard approaches that replicate the full target architecture, this model approximates essential hidden-state dynamics. The training objective combines a smoothed L_1 loss (MAE and MSE) for hidden state reconstruction with KL divergence to align the draft token distribution with the target model.

Data and Training Pipeline. We evaluated three data pipelines: offline labeling (I/O bound), chunked streaming (network bound), and online labeling. We adopted Online Labeling for the final setup. Although this increases HBM footprint by requiring the frozen target model to reside in memory, it yields the highest Tensor Core utilization.

Training was performed on a single node with $8 \times H100$ GPUs. The frozen verifier used Tensor Parallelism, while the EAGLE draft model utilized Distributed Data Parallelism. Full training hyperparameters are listed in Table 19.

Deployment and Results. Deployed via SGLang using EAGLE-2’s dynamic draft tree (Li et al., 2024d), the system achieves significant latency reductions. Table 20 highlights speedups up to $2.28 \times$ on reasoning tasks (T-Math) and consistent gains on ruMT-Bench. Table 21 illustrates domain-specific performance on ruMMLU-Pro, where Math and Engineering domains show the highest acceptance lengths (~ 3.7) and speedups

Hyperparameter	Value
Hardware	8×H100 (80GB)
Verifier parallelism	TP=2
Draft model parallelism	DDP (world size=8)
Batch size	32
Learning rate	3e-5
Number of epochs	4
Learning rate scheduler	cosine
Warmup steps	100
Weight decay	0.01
Optimizer	AdamW
Data type	BF16
TF32	enabled

Table 19: Hyperparameters used for EAGLE draft model training.

($\sim 2.0\times$). Future work will focus on draft model quantization and integrating EAGLE 3 (Li et al., 2025).

Benchmark	Temp.	Mode	Speedup	Acceptance Length
ruMT-Bench	0	No Think	2.05	3.55
	0	Think	1.86	3.37
	0.8	No Think	1.79	3.31
	0.8	Think	1.69	3.10
ruAlpaca	0	No Think	1.78	3.23
	0	Think	1.77	3.20
	0.8	No Think	1.61	2.94
	0.8	Think	1.57	2.85
ruCodeEval	0	No Think	2.26	4.09
	0	Think	2.07	3.76
	0.8	No Think	2.15	3.93
	0.8	Think	1.84	3.34
T-Math	0	Think	2.28	4.14
	0.8	Think	2.25	4.01

Table 20: Performance metrics for T-pro-2.0-eagle across different benchmarks, temperatures, and reasoning modes. Comparison of Speedup and Acceptance Length with and without Eagle.

H T-Math benchmark

T-Math⁷ is a Russian math reasoning benchmark constructed from high-school olympiad problems. It contains 331 tasks drawn from the All-Russian School Olympiad and the Moscow Olympiad in mathematics over the period 1998–2025. All items are single-answer problems with numeric gold solutions, which makes the benchmark suitable for automatic evaluation of long-chain mathematical reasoning.

Problem statements and ground-truth answers are extracted from PDF collections using the

⁷<https://huggingface.co/datasets/t-tech/T-math>

Domain	Speedup	Accept. Length	TPS	
			w/o Eagle	w/ Eagle
Biology	1.68	3.00	108.22	181.86
Business	2.00	3.63	107.83	216.49
Computer Sci.	1.89	3.37	107.99	204.22
Economics	1.72	3.07	108.26	185.80
Engineering	2.00	3.60	106.96	214.37
Health	1.67	2.98	108.29	181.00
History	1.52	2.72	108.15	164.17
Law	1.51	2.69	108.03	163.17
Math	2.06	3.70	107.88	221.96
Philosophy	1.62	2.88	108.37	175.29
Physics	1.96	3.50	107.60	210.60
Psychology	1.65	2.85	108.38	179.03
Chemistry	2.04	3.66	107.56	219.20

Table 21: Performance metrics for T-pro-2.0-eagle across ruMMLU- Pro domains (Temperature 0.8, Thinking mode, Batch size=1).

Qwen2.5-VL-72B-Instruct (Bai et al., 2025) model. The raw pool is then filtered with an LLM-based checker to discard (i) tasks requiring multiple answers, (ii) problems without a unique correct answer, (iii) theorem-style questions where the main goal is to prove a statement, (iv) tasks whose solutions are non-numeric, and (v) items that cannot be solved without access to auxiliary figures. Next, medium-difficulty tasks on which Qwen3-8B achieves near-perfect pass@16 are removed to focus the benchmark on genuinely challenging instances. Finally, both the question texts and the verifiable answers are manually reviewed against the original olympiad sources. Evaluation uses a standardized answer format (final answer wrapped in `\boxed{}`) and the `math_verify` library⁸ to compare predicted and reference expressions.

Table 24 reports pass@1 scores for several strong reasoning models. Although frontier systems such as o4-mini-high, DeepSeek-R1 and Gemini 2.5 Pro achieve competitive performance, the benchmark remains far from saturated, with none of the models exceeding 0.75 pass@1.

I Additional Evaluations

As shown in Table 23, the model preserves strong English reasoning ability despite being primarily optimized for Russian. Within the 27B–32B class, it remains closely aligned with the Qwen3-32B baseline: on MATH-500 it slightly improves accuracy, and on AIME 2024/2025 and GPQA the margins stay narrow. Performance is also competitive

⁸<https://github.com/huggingface/Math-Verify>

#	Problem statement (translated from Russian for readability) ⁶	Answer
1	Combinatorics / logic. In a tournament there are 20 players and 10 referees. After each game, the participants of that game take a photograph together with the referee. After the tournament it turned out that for some people it is impossible to determine whether they are a player or a referee (based only on the set of photos they appear in). What is the maximum possible number of such people?	2
2	Number theory / arithmetic constructions. Using any number of coins of denominations 1, 2, 5 and 10 roubles, together with (free) parentheses and the four arithmetic operations, construct an expression whose value is 2009, while spending as little money as possible. In the answer, write the minimum possible total value of the coins used (i.e., the minimum amount of money you need to “spend”).	23
3	Geometry, olympiad level. In triangle ABC with side lengths $AB = 3$, $BC = 4$, $CA = 5$, we mark pairs of points on its sides: points C_1 and C_2 on side AB , points A_1 and A_2 on side BC , and points B_1 and B_2 on side CA . Inside triangle ABC there is a point P such that triangles PA_1A_2 , PB_1B_2 and PC_1C_2 are congruent and equilateral. Find the area of the convex hexagon with vertices $A_1, A_2, B_1, B_2, C_1, C_2$. If necessary, round your answer to two decimal places.	3.34

Table 22: Example problems from the T-Math benchmark. Statements are translated from the original Russian for readability; see the dataset for the original wording and full benchmark specification.

Model	AIME 2024	AIME 2025	MATH-500	GPQA Diamond	LCB
<i>Open Source Models (27B-32B class)</i>					
T-pro 2.0 (Ours)	0.765	0.679	0.966	0.641	0.556
Qwen3-32B	0.808	0.725	<u>0.961</u>	0.668	0.546
RuadaptQwen3-32B-Instruct	0.692	0.604	0.948	0.596	0.489
Gemma 3 27B	0.260	0.221	0.882	0.515	0.246
DeepSeek-R1-Distill-Qwen-32B	0.706	0.573	0.950	0.621	0.572
<i>Open Source Larger Scale & Proprietary Models</i>					
DeepSeek-V3	0.52	0.285	0.942	0.655	0.405
DeepSeek-R1	0.914	0.875	0.983	0.813	0.770
YandexGPT5-Pro	0.117	0.090	0.776	0.434	0.272
GigaChat 2 Max	0.110	0.058	0.742	0.449	0.272
o4-mini (medium)	<u>0.800</u>	<u>0.819</u>	<u>0.974</u>	<u>0.783</u>	<u>0.757</u>
GPT-4o	0.098	0.065	0.762	0.545	0.246

Table 23: Comparison of models on English advanced reasoning benchmarks.

Model	pass@1
o4-mini-high	0.73
DeepSeek-R1-0528	<u>0.71</u>
Gemini-2.5-Pro	0.70
Claude Sonnet 4	0.56
T-pro 2.0	<u>0.54</u>
Qwen3-32B	0.53

Table 24: Pass@1 accuracy on the T-Math benchmark (331 problems).

with reasoning-distilled systems such as DeepSeek-R1-Distill-Qwen-32B, outperforming them on several metrics. Overall, these results indicate that the Cyrillic-focused tokenizer and our training pipeline do not meaningfully degrade English performance, maintaining robust cross-lingual generalization with minimal loss on advanced benchmarks.