Computational Linguistics and Intellectual Technologies:
Proceedings of the International Conference "Dialogue 2026"

June 24–26, 2026

WorkBench-RU: Can Agents Solve Multi-Step Workplace Tasks Beyond English?

Timur Ionov
ITMO University
Saint Petersburg, Russia
trionov@itmo.ru

Valentin Malykh
ITMO University
Saint Petersburg, Russia
valentin.malykh@phystech.edu

Abstract

We present WorkBench-RU, a bilingual extension of the WorkBench benchmark for evaluating tool-calling LLM agents on realistic workplace tasks requiring multi-step planning, function-call sequencing, and actual database execution. We add Russian localization — queries translated from English, localized tool descriptions, and parallel databases — across five business domains plus cross-domain tasks. During localization, we identify and fix 17 bugs in the original benchmark that had fundamentally compromised evaluation reliability. We evaluate 9 models (1.8B–120B parameters) under 4 reasoning configurations and report state-based correctness averaged across 3 independent trials, along with any@3, all@3, and side-effect metrics. Key findings: (1) structured reasoning (ReAct, native thinking) improves models by up to +20pp, with strongly architecture-dependent effects; (2) Qwen3-32B with native thinking reaches 70.1%, comparable to GPT-4.1-mini (68.1%); (3) English–Russian performance gaps vary from 4pp (Russian-native GigaChat3) to 28pp (GPT), suggesting that training language composition may matter more than model scale.

Keywords: LLM agents, tool calling, benchmark, bilingual evaluation, Russian NLP, reasoning
DOI: 10.29003/2075-7182-2026-24-153-162

WorkBench-RU: Справляются ли агенты с многошаговыми рабочими задачами за пределами английского языка?

Ионов Тимур
Университет ИТМО
Санкт-Петербург, Россия
trionov@itmo.ru

Валентин Малых
Университет ИТМО
Санкт-Петербург, Россия
valentin.malykh@phystech.edu

Аннотация

Мы представляем WorkBench-RU — двуязычное расширение бенчмарка WorkBench для оценки LLM-агентов с вызовом инструментов на реалистичных рабочих задачах, требующих многошагового планирования, последовательности вызовов функций и фактического выполнения операций с базами данных. Мы добавляем русскую локализацию — запросы, переведённые с английского, описания инструментов и параллельные базы данных — для пяти бизнес-доменов и кросс-доменных задач. В ходе локализации мы обнаруживаем и исправляем 17 ошибок в оригинальном бенчмарке, которые существенно искажали результаты оценки. Мы оцениваем 9 моделей (от 1,8 до 120 млрд параметров) в 4 конфигурациях рассуждений и приводим метрики корректности (среднее по 3 запускам, any@3, all@3) и побочных эффектов. Ключевые результаты: (1) структурированное рассуждение (ReAct, нативное мышление) улучшает модели до +20 п.п. с сильной зависимостью от архитектуры; (2) Qwen3-32B с нативным мышлением достигает 70,1% — сопоставимо с GPT-4.1-mini (68,1%); (3) разрыв между английским и русским варьируется от 4 п.п. (русскоязычная GigaChat3) до 28 п.п. (GPT), что указывает на большую роль состава обучающих данных по сравнению с размером модели.

Ключевые слова: LLM-агенты, вызов инструментов, бенчмарк, двуязычная оценка, русская NLP, рассуждения

1 Introduction

Large language models are increasingly deployed as autonomous agents that interact with external tools to complete workplace tasks. Evaluating such agents requires verifying the entire pipeline: correct tool selection, parameter population, multi-step dependencies, and absence of side effects. A single query may require 3–6 sequential tool calls — e.g., "email the attendees of my next meeting" involves calendar search, event identification, attendee extraction, and email composition.

Existing benchmarks (Patil et al., 2025; Yao et al., 2025; Qin et al., 2024; Lu et al., 2024; Wang et al., 2024; Jimenez et al., 2024) are all English-only. WorkBench (Styles et al., 2024) offers 690 tasks across five business domains with state-based evaluation, but also lacks multilingual support — and no equivalent exists for Russian, a language with rich morphology that poses distinct challenges for entity matching.

Across our experiments, the English–Russian gap is driven less by language understanding and more by tool-call grounding failures: directory lookups, mapping morphological forms to Latin-script entries, and selecting the correct enum value in a different language. We present **WorkBench-RU**, a bilingual extension with the following contributions:

1. **Russian localization** of 690 queries with localized tool descriptions, system prompts, and parallel databases, enabling direct EN–RU comparison.
2. **17 bug fixes** to the original benchmark that had compromised evaluation (e.g., email tools crashed on every write, inflating accuracy).
3. **Systematic reasoning comparison** across 9 models (1.8B–120B) under 4 configurations with multi-trial analysis revealing architecture-dependent effects.
4. **Cross-lingual analysis** showing EN–RU gaps from 4pp (GigaChat3) to 28pp (GPT), varying by domain (email: up to –50.7pp).

Code and data are publicly available.[1]

2 Related Work

Tool-calling benchmarks. WorkBench (Styles et al., 2024) provides 690 tasks across 5 workplace domains with state-based evaluation. Other benchmarks target single-turn function calls (Patil et al., 2025), multi-turn reliability (Yao et al., 2025), API breadth (Qin et al., 2024; Li et al., 2023), stateful dependencies (Lu et al., 2024; Wang et al., 2024), and code agents (Liu et al., 2024; Jimenez et al., 2024). All are English-only.

Cross-lingual evaluation. XTREME (Hu et al., 2020) and mGSM (Shi et al., 2023) cover multilingual NLU and reasoning. For tool-calling, MASSIVE-Agents (Kulkarni et al., 2025) tests single-turn calls in 52 languages; Luo et al. (2026) study robustness across 8 languages. Our work provides the first *multi-step, state-based* cross-lingual agent benchmark.

Reasoning for agents. ReAct (Yao et al., 2023) interleaves reasoning and action; chain-of-thought (Wei et al., 2022) improves multi-step reasoning; Qwen3 (Qwen Team, 2025) supports native thinking via `<think>` tokens.

Russian NLP. Russian SuperGLUE (Shavrina et al., 2020), TAPE (Taktasheva et al., 2022), and MERA (Fenogenova et al., 2024) evaluate Russian language understanding and few-shot reasoning, but no Russian benchmark targets agentic tool-calling.

3 WorkBench-RU

3.1 Original WorkBench

WorkBench (Styles et al., 2024) evaluates LLM agents on 690 workplace tasks across five domains: *calendar*, *email*, *analytics*, *project management*, and *CRM*. Each domain exposes 4–6 tools as OpenAI-format function-calling schemas (26 tools total). The benchmark includes 480 single-domain and 210 cross-domain tasks requiring coordination across multiple domains.

[1] `https://github.com/sir-timio/WorkBench-RU`

Evaluation is state-based: both predicted and ground-truth tool calls are executed against fresh sandbox databases and the resulting states are compared. A prediction is *correct* if the database state matches the ground truth, regardless of the specific call sequence. *Side effects* are detected when the state differs from the initial in ways not attributable to ground-truth actions.

3.2 Bug Audit

We discovered and fixed 17 bugs in the upstream repository.[2] The most critical: all three email write tools (`send_email`, `forward_email`, `reply_email`) crashed on every invocation due to a dtype mismatch (a `pandas.Timestamp` written into a string-typed column). Because both predicted and ground-truth executions crashed identically and left the sandbox in its initial state, the state-based comparison returned *correct* by default — silently marking 78% of email queries (`send`/`forward`/`reply`, but not `delete`) as solved regardless of model behavior. Other bugs include non-deterministic analytics (global state mutation), incorrect tool specifications, and metric false negatives that depressed PM accuracy. After fixes, all 164 unit tests pass (expanded from 77).

Quantitative impact. Table 1 compares evaluation metrics on the original (pre-fix) and corrected benchmark for Qwen3-32B Base, run on the same model outputs once with the upstream tool implementations and once with our fixes (3 trials averaged for both columns). Aggregate correctness barely moves (+1.3pp EN, –0.1pp RU), but the per-domain swings are large in both directions. *Email* drops sharply (–28.5pp EN, –39.2pp RU) once the silently-correct gating from the email-Timestamp crash is removed. *Analytics* and *calendar*, on the other hand, rise substantially (analytics +26.1pp EN, +31.4pp RU) because the upstream bugs — non-deterministic state mutation, mismatched tool-spec field names, missing `na=False`/`regex=False` guards on `str.contains` — caused legitimate model outputs to fail evaluation. The two effects roughly cancel at the aggregate level, so a surface-level comparison would suggest that the bugs "did not matter"; per-domain analysis shows that they were biasing scores by tens of points in opposite directions, masking which models were really better.

	English correct%			Russian correct%		
Domain	**Original**	**Fixed**	**Δ**	**Original**	**Fixed**	**Δ**
Analytics	20.0	46.1	+26.1	18.3	49.7	+31.4
Calendar	54.6	64.2	+9.6	24.6	36.1	+11.5
CRM	58.8	62.9	+4.1	21.3	21.3	+0.0
Email	77.8	49.3	–28.5	73.3	34.1	–39.2
PM	52.5	50.8	–1.7	55.0	51.3	–3.7
Multi	30.0	28.3	–1.7	22.9	22.4	–0.5
Overall	48.9	50.3	+1.3	35.9	35.8	–0.1

Table 1: Pre-fix vs. post-fix correctness for Qwen3-32B Base, both languages, averaged across 3 independent trials. "Original" is the upstream WorkBench (commit prior to our fixes); "Fixed" is the same evaluation pipeline with our 17 bug fixes applied. The overall figures are within 2pp, but Email loses 28–39pp of artefact-induced correctness while Analytics gains 26–31pp of genuine correctness that the upstream bugs were suppressing. The Email drop and Analytics rise are independent effects with opposite signs; the aggregate cancellation hides the magnitude of evaluation drift.

3.3 Russian Localization

The localization process involved three stages:

Stage 1: Query and answer translation. All 690 queries and their expected function-call sequences were translated simultaneously from English to Russian using the Claude Opus 4.5 API (Anthropic, 2025). We selected Claude Opus 4.5 for its strong multilingual generation quality at the time the localization pipeline was launched (late 2025), and the API ensures reproducibility of the translation process. Simultaneous translation ensures semantic alignment — e.g., when a query references "In Progress,"

[2]Full bug audit: `https://github.com/sir-timio/WorkBench-RU/blob/main/BUGS.md`.

the answer's parameter values are also translated to «Текущие». Translations were reviewed by three computational linguists: each query was checked by at least one reviewer, with disagreements resolved by discussion. Approximately 15% of queries required substantive corrections (phrasing, morphological agreement, register). The translation prompt and code are available in the repository. Table 2 shows examples.

Stage 2: Database adaptation. Parallel Russian databases were created using translation dictionaries validated through unit tests verifying referential consistency, field types, and completeness (Table 3).

Stage 3: Tool description localization. All 26 tool specifications were extended with Russian descriptions and parameter names. When `lang=ru`, the model receives Russian system prompts, queries, and databases. Tool parameter *names* remain in English, reflecting realistic multilingual API usage.

English query	Russian translation
Delete my first meeting on December 13	Удали мою первую встречу 13 декабря.
Move all 'In Progress' tasks on the Marketing Board to 'Done'	Перенеси все задачи со статусом «Текущие» на доске Marketing Board в статус «Готово»
Send an email to Cameron Anderson asking about the budget	Отправь письмо Камерону Андерсону с вопросом о бюджете

Table 2: EN→RU translation examples. Database content values are localized; person names, IDs, and dates remain in English (Table 3). Russian morphology requires case inflection («Камерону Андерсону», dative).

Domain	Translated DB fields	Entries
Calendar	event names, descriptions	48
Email	subjects, bodies, labels	500
PM	task names, list/board names, statuses	82
CRM	notes, products, statuses	14
Analytics	traffic source labels	4

Table 3: Translation scope by domain. "Entries" = number of distinct translated values. IDs, dates, emails, phone numbers, and numeric values remain unchanged across languages.

Design decisions. Person names are preserved in Latin script in the Russian databases: (a) this reflects real practice in international companies; (b) the model must map Russian morphological forms («Камерону Андерсону», dative) to English-script directory entries ("Cameron Anderson"), testing cross-script entity resolution; (c) avoiding transliteration ambiguity («Кэмерон» vs «Камерон») prevents an uncontrolled confound.

4 Experimental Setup

4.1 Models

We evaluate 9 models spanning five families (Table 4), all served via vLLM (Kwon et al., 2023) except GPT-4.1-mini (OpenAI, 2025b) and Llama 4 Scout (Meta AI, 2025)[3] (accessed via API). Open-weight models include Qwen3 (Qwen Team, 2025), Qwen2.5 (Yang et al., 2024), GPT-OSS-120B (OpenAI, 2025a) (MoE with built-in reasoning), and GigaChat3-10B (Sber AI, 2025) (1.8B active) — the first Russian-native model in our evaluation.[4]

Sampling parameters. Qwen3-VL-8B uses temperature=0.6, top_p=0.8, frequency_penalty=0.1; all other models use defaults. Each configuration is evaluated over 3 independent trials.

[3]Llama — семейство моделей компании Meta Platforms Inc.*, признанной экстремистской организацией, деятельность которой запрещена на территории РФ.

[4]vLLM v0.11.0; GigaChat3-10B required v0.14.0.

Model	Params	Type	Notes
GPT-OSS-120B	120B MoE	Open	OpenAI open-weight; 5.1B active; thinking only
GPT-4.1-mini	—	Proprietary	
Qwen3-32B	32B	Open	Native thinking support
Qwen3-14B	14B	Open	Native thinking support
Qwen3-VL-8B	8B	Open	Vision-language model
Qwen2.5-32B	32B	Open	No thinking support
Qwen2.5-7B	7B	Open	
GigaChat3-10B	10B MoE	Open	Sber AI; 1.8B active
Llama 4 Scout	109B MoE	Open	Meta*; 17B active

Table 4: Models evaluated. GPT-OSS-120B uses internal reasoning (thinking) by default with no non-reasoning variant available; it is labeled "Think" throughout. GigaChat3-10B is the first Russian-native model in this evaluation.

4.2 Reasoning Configurations

We test four reasoning configurations:

Baseline: Standard tool-calling with a system prompt specifying task rules, domain context, and behavioral constraints in OpenAI function-calling format.

ReAct: A modified ReAct (Yao et al., 2023) with *context stripping*: before each tool call, a forced text-only reasoning turn is generated and then removed from conversation context, preventing context pollution and forcing grounding in actual tool results. We test two prompt intensities with nearly identical results (+5.2pp vs +5.8pp on Qwen3-VL-8B).

Native thinking: Qwen3 models support internal reasoning via `<think>` tokens generated before each response but not added to conversation history (using vLLM's DeepSeek-R1 (Guo et al., 2025) parsing protocol). Not available for Qwen2.5.

ReAct + Think: Combines both: internal thinking tokens, then a text reasoning turn (stripped), then the tool call.

4.3 Metrics

All metrics are averaged across 3 independent trials. **Correct%** (primary): state-based database comparison. **any@3**: at least 1 of 3 trials correct (upper bound, following Chen et al. (2021)). **all@3**: all 3 trials correct (reliability, following TAU-bench's pass^k (Yao et al., 2025)). The any@3 – all@3 gap quantifies trial variance. **Side effects%**: fraction of all queries where the prediction is both incorrect and mutates database state in ways not attributable to ground-truth actions.

5 Results

5.1 Main Leaderboard

Table 5 presents all 19 model-configuration pairs grouped by model and ranked by English correct%.

GPT-OSS-120B leads at 76.8% (EN); it uses internal thinking by default with no non-reasoning variant (labeled "Think"). Qwen3-32B + Think reaches 70.1%, statistically indistinguishable from GPT-4.1-mini (68.1%; $p = 0.29$). GigaChat3-10B ranks last (31.9% EN) but has the smallest bilingual gap (–4.1pp; Section 5.4).

5.2 Trial Consistency

Running each configuration 3 times reveals substantial variance (Figure 1). The any@3 – all@3 gap ranges from 16.6pp (GPT-OSS-120B, most reliable) to 37.6pp (Llama 4 Scout), indicating that pass@1 evaluation can both underestimate potential and overestimate reliability.

		Correct% ↑			Side eff.% ↓	
Model	**Config**	**EN**	**RU**	**Gap**	**EN**	**RU**
GPT-OSS-120B	Think	**76.8**	**50.3**	–26.5	**17.3**	31.5
Qwen3-32B	Think	70.1	44.2	–25.9	22.5	28.6
	ReAct+Think	68.2	43.8	–24.4	20.6	23.4
	ReAct	50.7	37.5	–13.2	42.2	29.2
	Base	50.3	35.8	–14.5	43.1	36.7
GPT-4.1-mini	Base	68.1	40.0	–28.1	21.9	32.2
Qwen3-14B	Think	63.9	39.3	–24.6	26.1	28.3
	Base	46.1	26.1	–20.0	44.8	43.9
	ReAct	45.6	25.3	–20.3	45.1	45.0
Qwen2.5-32B	ReAct	64.3	41.3	–23.0	22.1	29.9
	Base	58.3	36.5	–21.8	25.9	30.7
Qwen3-VL-8B	ReAct high	57.3	43.3	–14.0	29.0	29.9
	ReAct mid	56.7	42.2	–14.5	29.0	30.4
	Base	51.5	43.1	–8.4	37.2	33.8
Llama 4 Scout	ReAct	45.9	31.3	–14.6	25.3	26.6
	Base	42.3	28.3	–14.0	36.7	39.0
Qwen2.5-7B	Base	37.9	23.2	–14.7	42.3	37.4
GigaChat3-10B	Base	31.9	27.8	–4.1	43.2	**7.4**
	ReAct	30.2	29.6	–0.6	53.7	14.4

Table 5: Full leaderboard: correct% and side-effect% averaged across 6 domain subsets (5 single-domain + cross-domain) and 3 trials. Gap = RU − EN. Side effects = fraction of all queries where the prediction is incorrect *and* the model mutated database state (e.g., sent an email to the wrong person, deleted the wrong event). Configs sorted by EN correct% within each model group. Llama 4 Scout (17B active) and GigaChat3-10B (1.8B active) are MoE models.

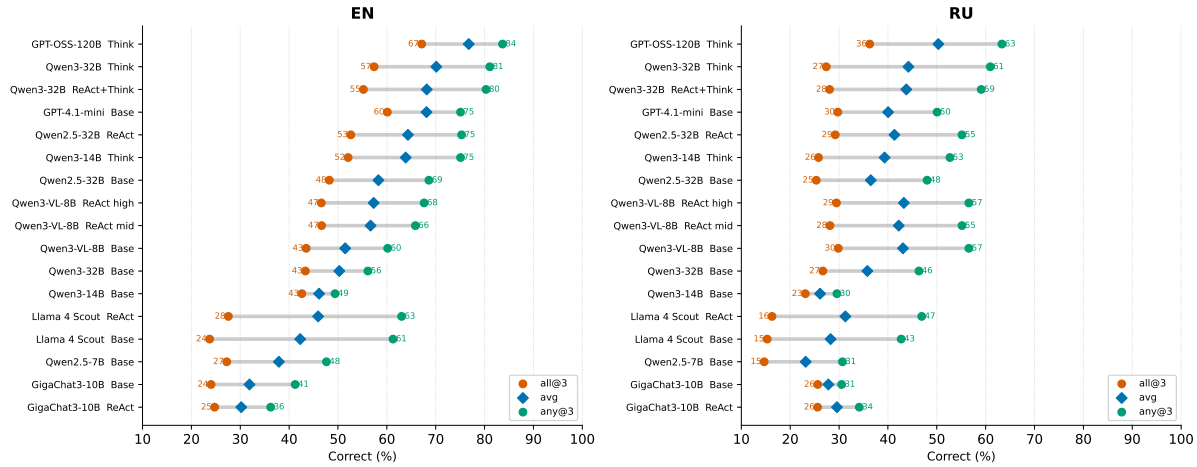

Figure 1: Trial consistency (EN left, RU right): each row shows all@3 (left dot), average (diamond), and any@3 (right dot). Wider spans indicate less consistent models.

5.3 Reasoning Impact

The effect of structured reasoning is strongly architecture-dependent (Figure 2).

Native thinking transforms Qwen3. Qwen3-32B jumps from 50.3% to 70.1% (+19.8pp, $p < 0.001$); Qwen3-14B gains +17.8pp. ReAct alone provides no benefit for Qwen3-32B (+0.4pp): forced text-only reasoning conflicts with the architecture designed for native `<think>` tokens.

ReAct helps models without native thinking. Qwen2.5-32B gains +6.0pp ($p < 0.001$), Qwen3-VL-8B gains +5.2–5.8pp, Llama 4 Scout gains +3.6pp (primarily reducing side effects). GigaChat3-10B (1.8B active) shows no ReAct benefit.

Statistical significance. All key findings are significant at $p < 0.05$ (paired bootstrap, 10K resamples). The advantage of Qwen3-32B Think over GPT-4.1-mini (2.0pp) is non-significant ($p = 0.29$).

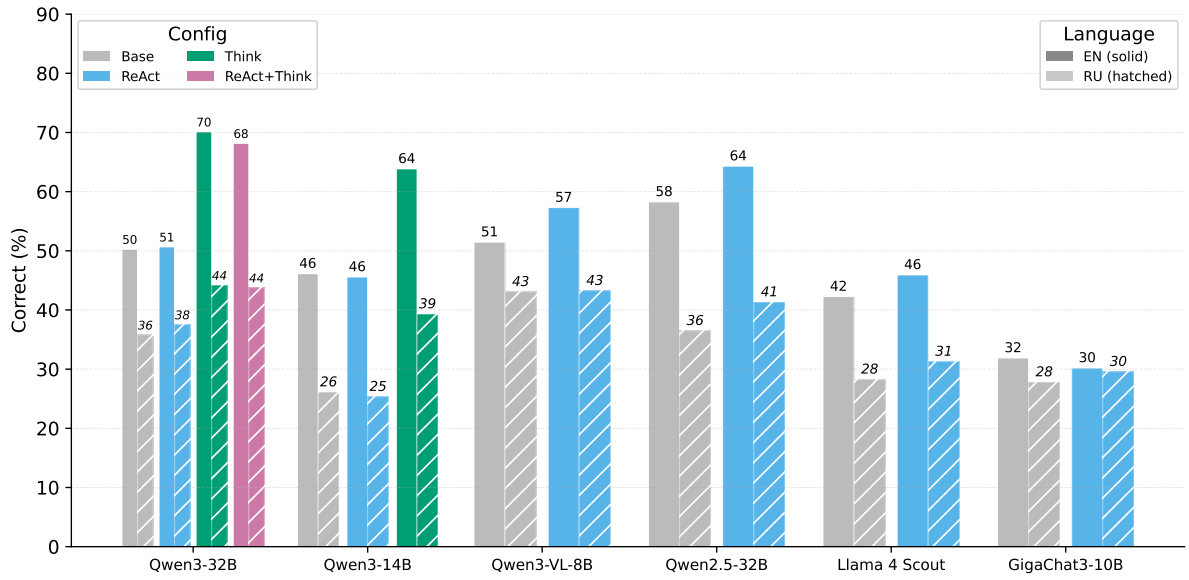

Figure 2: Reasoning configurations per model (solid bars = EN, hatched bars = RU; bar labels show EN correct%, RU in parentheses). Native thinking transforms Qwen3-32B/14B; ReAct helps Qwen2.5 (+6pp), Qwen3-VL-8B, and Llama 4 Scout.

5.4 Bilingual Analysis

The EN–RU gap in Table 5 reveals different cross-lingual robustness profiles.

GPT models show the largest gaps (26–28pp), consistent with (Kulkarni et al., 2025; Luo et al., 2026). **GigaChat3-10B has the smallest gap** (0.6–4pp), achieving near-parity and even outperforming English on Russian PM (71.2% vs 42.5%). **Reasoning widens the gap**: Qwen3-32B thinking improves English by +19.8pp but Russian by only +8.4pp. GigaChat3 is the exception: ReAct *narrows* its gap from 4.1pp to 0.6pp (Shi et al., 2023).

5.5 Per-Domain Analysis

Performance varies substantially across domains (Figure 3). Email shows the largest bilingual degradation (GPT-4.1-mini: –50.7pp); PM is hardest overall, with email fabrication as the dominant failure. Cross-domain tasks (210 queries) are substantially harder: 20–52% EN, 20–48% RU.

5.6 Side Effects

Side-effect rates (Table 5, rightmost columns) decrease with structured reasoning: Llama 4 Scout ReAct drops from 36.7% to 25.3% EN. GPT-OSS-120B has the lowest EN rate (17.3%) but higher Russian side effects (31.5%). GigaChat3-10B shows the opposite: high EN rate (43–54%) but very low Russian (7.4%), consistent with cautious behavior in its native language.

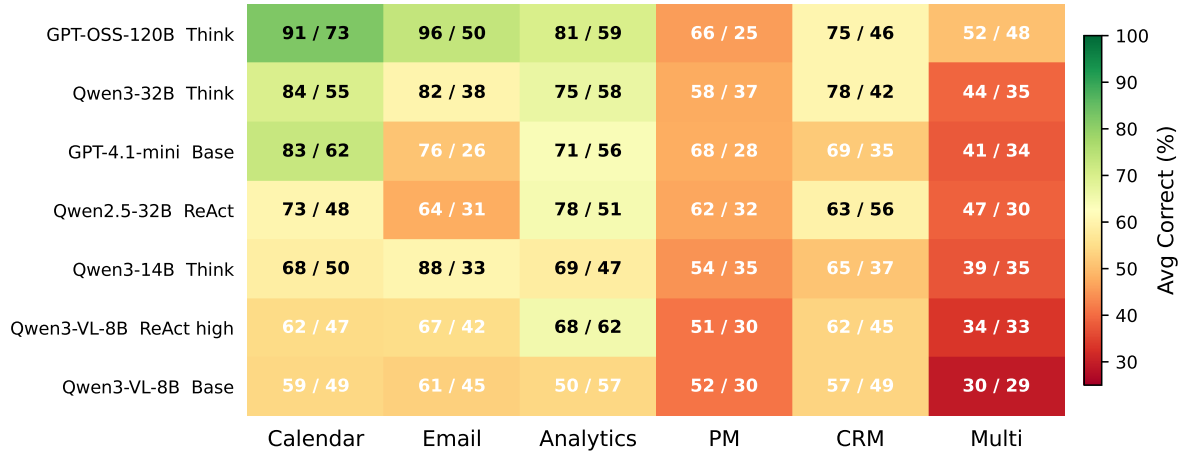

Figure 3: Per-domain correct% for top model-configurations (EN / RU). Cell color reflects average of EN and RU.

6 Error Analysis

Table 6 provides a heuristic-based error breakdown across five representative models. Error categories are derived from CSV evaluation flags as follows: **email fabrication** = predicted email address not found in company directory (`wrong_email` flag); **incomplete execution** = no write operations attempted (`no_actions` flag); **hallucinated action** = unintended database mutations on an incorrect query (`unwanted_side_effects & ~correct`); **wrong parameters** = remainder of incorrect queries (correct tool, wrong arguments, no side effects).

Error type	GPT-OSS Think	Q3-32B Think	Q2.5-32B Base	Q3-VL 8B	GigaChat3 10B
Wrong parameters	36.4	45.0	41.1	36.3	39.8
Hallucinated action	61.8	50.4	49.2	58.2	9.0
Incomplete execution	1.7	0.4	0.2	1.1	43.6
Email fabrication	0.1	4.2	9.5	4.5	7.6
Error rate (of queries)	48.6	56.4	64.3	58.6	72.2

Table 6: Error type distribution (RU %, of errors). Error rate = fraction of incorrect queries. GigaChat3-10B shows a unique pattern: 43.6% incomplete execution (vs. $<2\%$ for other models), consistent with its very low Russian side-effect rate (7.4%).

6.1 Error Types

Hallucinated action is the dominant RU error for most models (49–62%), where models attempt actions with wrong parameters or unwarranted operations. The exception is GigaChat3-10B (9.0%), whose Russian-native training produces more cautious behavior. **Wrong parameters** (36–45%) are failures without side effects — the largest category for Qwen3-32B (45.0%). **Email fabrication** (0–10%) primarily affects PM/CRM; GPT-OSS-120B is nearly immune (0.1%), while Qwen2.5-32B is most susceptible (9.5%). **Incomplete execution** is below 2% for all models except GigaChat3-10B (43.6%): the Russian-native model frequently finds content but does not complete write operations, explaining its very low side-effect rate (7.4%) alongside high error rate (72.2%).

6.2 Tool Call Trace Example

Consider the query: “Send an email to Cameron Anderson asking about the project deadline.” The ground-truth solution requires 2 tool calls:

1. `find_email_address(name="Cameron Anderson")` → returns "c.anderson@company.com"
2. `send_email(to="c.anderson@company.com", subject="Project Deadline", body="...")`

A common failure (email fabrication) skips step 1:

1. `send_email(to="cameron.anderson@company.com", ...)` — *fabricated address*

The model "knows" a plausible format but doesn't verify it against the directory — the key lesson is "read before write." In Russian, «Камерону Андерсону» (dative) must be mapped to "Cameron Anderson" in the English-script directory. Additional Russian failure modes include status value mismatch («В процессе» instead of «Текущие» for "In Progress") and reduced search precision from broader Russian queries hitting the 5-result limit.

7 Conclusion

We presented WorkBench-RU, the first bilingual agent benchmark for Russian workplace tasks, with Russian localization of 690 queries, 17 bug fixes to the original benchmark, and evaluation of 9 models under 4 reasoning configurations.

Key findings: native thinking improves Qwen3 by up to +20pp; ReAct helps models without thinking by +4–6pp; Qwen3-32B with native thinking (70.1%) reaches parity with GPT-4.1-mini (68.1%). Cross-lingual robustness varies from 4pp (GigaChat3) to 28pp (GPT), with training data composition appearing to matter more than scale. Email shows the most extreme bilingual degradation (–50.7pp for GPT-4.1-mini). Taken together, our error analysis suggests that most of the EN–RU drop is concentrated in tool-call grounding rather than in core language understanding, pointing to directory resolution, morphology-to-Latin-script mapping, and enum value selection as primary targets for future work.

Limitations. (1) Russian queries are translated, not natively authored. (2) Only one Russian-native model (GigaChat3-10B) is evaluated; YandexGPT and the T-Lite/T-Pro family were excluded because their tool-calling APIs at the time of evaluation did not match the OpenAI function-calling format we use. (3) Token costs and latency are not normalized across models. (4) The error taxonomy uses heuristic categorization. (5) State-based correctness does not weigh side-effect severity: deleting a single email and deleting all emails are scored identically as "correct with side effects". A severity-weighted variant is left as future work. (6) The company directory contains 20 contacts with unique first and last names, so models are not tested on disambiguating namesakes; expanding the directory with intentional name collisions is left as future work. (7) Some queries cite email titles verbatim, while a more realistic interaction would reference messages by partial keywords; future versions can switch to keyword-based references. (8) Tasks are limited to direct tool-call sequences and do not require summarization or ad-hoc multi-step reasoning over retrieved content; integrating read-and-summarize tasks is left as future work.

All code, data, and evaluation infrastructure are released under the MIT license.

References

Anthropic. 2025. Claude: AI assistant by Anthropic. `https://www.anthropic.com/claude`.

Mark Chen, Jerry Tworek, Heewoo Jun, et al. 2021. Evaluating large language models trained on code. *arXiv preprint arXiv:2107.03374*.

Alena Fenogenova, Artem Chervyakov, Nikita Martynov, et al. 2024. MERA: A comprehensive LLM evaluation in Russian. // *Proceedings of ACL*.

Daya Guo, Dejian Yang, He Zhang, et al. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Junjie Hu, Sebastian Ruder, Aditya Siddhant, Graham Neubig, Orhan Firat, and Melvin Johnson. 2020. XTREME: A massively multilingual multi-task benchmark for evaluating cross-lingual generalization. // *Proceedings of ICML*.

Carlos E. Jimenez, John Yang, Alexander Wettig, et al. 2024. SWE-bench: Can language models resolve real-world GitHub issues? // *Proceedings of ICLR*.

Mayank Kulkarni, Vittorio Mazzia, Judith Gaspers, Christopher Hench, and Jack FitzGerald. 2025. MASSIVE-Agents: A benchmark for multilingual function-calling in 52 languages. // *Findings of EMNLP*.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, et al. 2023. Efficient memory management for large language model serving with PagedAttention. // *Proceedings of SOSP*.

Minghao Li, Yingxiu Zhao, Bowen Yu, et al. 2023. API-Bank: A comprehensive benchmark for tool-augmented LLMs. // *Proceedings of EMNLP*.

Xiao Liu, Hao Yu, Hanchen Zhang, et al. 2024. AgentBench: Evaluating LLMs as agents. // *Proceedings of ICLR*.

Jiarui Lu, Thomas Zhu, Panupong Pasupat, et al. 2024. ToolSandbox: A stateful, conversational, interactive evaluation benchmark for LLM tool use capabilities. *arXiv preprint arXiv:2408.04682*.

Zheng Luo, T Pranav Kutralingam, et al. 2026. Lost in execution: On the multilingual robustness of tool calling in large language models. *arXiv preprint arXiv:2601.05366*.

Meta AI. 2025. The Llama 4 herd of models. `https://ai.meta.com/blog/llama-4-multimodal-intelligence/`.

OpenAI. 2025a. GPT-OSS-120B: An open-weight mixture-of-experts model. `https://huggingface.co/openai/gpt-oss-120b`.

OpenAI. 2025b. Introducing GPT-4.1 in the API. `https://openai.com/index/gpt-4-1/`.

Shishir G. Patil, Huanzhi Mao, Fanjia Yan, et al. 2025. The Berkeley Function Calling Leaderboard (BFCL): From tool use to agentic evaluation of large language models. // *Proceedings of ICML*.

Yujia Qin, Shihao Liang, Yining Ye, et al. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. // *Proceedings of ICLR*.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Sber AI. 2025. GigaChat3-10B-A1.8B: An open-weight mixture-of-experts model. `https://huggingface.co/ai-sage/GigaChat3-10B-A1.8B`.

Tatiana Shavrina, Alena Fenogenova, Anton Emelyanov, et al. 2020. RussianSuperGLUE: A Russian language understanding evaluation benchmark. // *Proceedings of EMNLP*.

Freda Shi, Mirac Suzgun, Markus Freitag, et al. 2023. Language models are multilingual chain-of-thought reasoners. // *Proceedings of ICLR*.

Olly Styles, Sam Miller, Patricio Cerda-Mardini, Tanaya Guha, Victor Sanchez, and Bertie Vidgen. 2024. WorkBench: A benchmark dataset for agents in a realistic workplace setting. // *Proceedings of COLM*.

Ekaterina Taktasheva, Tatiana Shavrina, Alena Fenogenova, Denis Shevelev, Nadezhda Katricheva, Maria Tikhonova, Albina Akhmetgareeva, Oleg Zinkevich, Anastasiia Bashmakova, Svetlana Iordanskaia, Alena Spiridonova, Valentina Kurenshchikova, Ekaterina Artemova, and Vladislav Mikhailov. 2022. TAPE: Assessing few-shot Russian language understanding. // Yoav Goldberg, Zornitsa Kozareva, and Yue Zhang, *Findings of the Association for Computational Linguistics: EMNLP 2022*, P 2472–2497, Abu Dhabi, United Arab Emirates, December. Association for Computational Linguistics.

Xingyao Wang, Zihan Wang, Jiateng Liu, et al. 2024. MINT: Evaluating LLMs in multi-turn interaction with tools and language feedback. // *Proceedings of ICLR*.

Jason Wei, Xuezhi Wang, Dale Schuurmans, et al. 2022. Chain-of-thought prompting elicits reasoning in large language models. // *Proceedings of NeurIPS*.

An Yang, Baosong Yang, Binyuan Hui, et al. 2024. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*.

Shunyu Yao, Jeffrey Zhao, Dian Yu, et al. 2023. ReAct: Synergizing reasoning and acting in language models. // *Proceedings of ICLR*.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. 2025. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. // *Proceedings of ICLR*.