BUILT TO READ, NOT TO RECALL: LANGUAGE SKILLS OVER WORLD KNOWLEDGE IN SMALL LANGUAGE MODELS

Anonymous authors
Paper under double-blind review

ABSTRACT

Frontier-scale language models are commonly assumed to need their full scale. We argue instead that LLM abilities fall into two families with very different scaling behavior: world knowledge—stored facts and entities—improves steeply with parameter count, whereas language skills—comprehending, extracting, and structuring given text—are largely attainable at modest sizes when cultivated deliberately. Retrieval-augmented generation (RAG) and agent pipelines keep knowledge external and rely chiefly on the latter; a compact purpose-trained model can therefore replace frontier-scale components at a fraction of the cost and energy. We test this hypothesis with `Meno-Lite-0.1`, a publicly released bilingual Russian–English 7B-class model whose name inverts Plato's anamnesis: the knowledge lives in the corpus, the skill in the model—built to read, not to recall. Across a 0.5B–72B backbone size ladder, knowledge benchmarks climb steeply (closed-book quiz F1 grows $21\times$ from 0.5B to 72B) while skill benchmarks plateau after $\sim$7B, and the memorization ratio—answering from parameters instead of the given context—rises with scale. The skill-trained 7B model sits above the entire ladder on knowledge-graph construction (0.468 vs. 0.416 harmonic mean against a 32B model), reaches 92% of the 32B level on open-book multi-hop QA while retaining its backbone's knowledge level, is the most context-faithful model measured, and serves Russian text $6.2\times$ faster than the 32B engine. We further document protocol sensitivities that can masquerade as scaling behavior. Weights, training data, and code are publicly available.

1 INTRODUCTION

The dominant recipe for capable language models is to scale: more parameters, more tokens, more clusters. Scaling-law analysis (Kaplan et al., 2020; Hoffmann et al., 2022) and the resulting frontier systems demonstrate that parameter count buys genuine capability—above all, *world knowledge*, the ability to answer from parametric memory. Yet this recipe has a price. Frontier-class models require clusters whose training and serving footprints translate directly into energy consumption and CO_2 emissions (Patterson et al., 2021; Luccioni et al., 2023), and whose inference cost dominates the economics of production systems. In most business scenarios that need a language model as a *component*—a reader inside a RAG pipeline, an extractor inside a GraphRAG engine, a tool caller inside a multi-agent system—frontier capability is not merely expensive but redundant: the knowledge the component needs is already externalized in the retrieval corpus, the document under analysis, or the tool result being digested.

We therefore revisit what such components actually require. We hypothesize that the abilities of LLMs decompose into two families:

- **world knowledge**—facts, entities, dates, and associations stored in parameters, which improves steeply with model size and is what closed-book benchmarks measure; and
- **language skills**—comprehension of long contexts, information extraction into specified schemas, faithful summarization, reasoning over *given* evidence (Wei et al., 2022), and instruction following—which, we argue, are largely attainable at modest sizes when deliberately cultivated.

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

If this decomposition holds, then a small model whose training budget is invested in language skills rather than factual recall should be an excellent—and radically cheaper—neural component for RAG and agentic systems, where knowledge is external by construction.

This paper tests the hypothesis on `Meno-Lite-0.1` (Bondarenko, 2026a), a publicly released bilingual (Russian–English) 7B-class model. The name alludes to Plato's dialogue *Meno* (Plato, 1956), in which Socrates argues that knowledge is not learned but recollected from within (*anamnesis*). Meno-Lite inverts this doctrine: rather than assuming knowledge is already inside the model, we externalize it into a retrieval corpus and train the model to read instead of recall—to arrive at answers given the right context, without carrying the answers in its parameters. The inversion is precise rather than merely decorative. Socrates elicits geometry from the slave boy while transferring no content, because for Plato the content is innate; a retrieval pipeline *transfers* content into the context, because none of it is innate. In the resulting system the model stands on the questioner's side—drawing knowledge out of documents, knowledge graphs, and tool results—while the corpus plays the soul that holds the answers. Investing the training budget in language skills is therefore not teaching the slave boy more facts; it is training a sharper Socrates. `Meno-Lite-0.1` is built on a Qwen2.5-7B descendant (Yang et al., 2024; Tikhomirov et al., 2024) with skill-oriented continued pretraining and supervised fine-tuning; it has served as the extraction engine of a GraphRAG system (Komarov et al., 2026), a winning shared-task ensemble member (Bondarenko et al., 2026), and a document-reasoning harness.

Our contributions are:

1. an operational decomposition of LLM ability into world knowledge and language skills, with a three-axis evaluation design—closed-book knowledge, open-book skills, and context faithfulness over a 0.5B–72B size ladder—that makes the different scaling behavior measurable;
2. `Meno-Lite-0.1` itself: a compact bilingual, skill-oriented model that instantiates the decomposition—above the entire size ladder on knowledge-graph construction, at its backbone's knowledge level, the most context-faithful model measured, and $6.2\times$ faster in Russian serving—released with weights, training data, and code;
3. scaling evidence on all three axes, cross-checked against seven same-size Russian-adapted peers: knowledge climbs with size, skills plateau after $\sim$7B, memorization rises with scale, and adaptation buys knowledge but not extraction;
4. protocol-sensitivity findings: on generative benchmarks with short references, prompt-format choices (chat template, brevity instruction) move scores by more than the effects under study—a confounder that grows with model size (§5.5).

2 Related work

Scaling laws and their interpretation. Kaplan et al. (2020) and Hoffmann et al. (2022) established power-law scaling of loss with parameters, data, and compute; the community has largely read these laws as "bigger is better everywhere." Our claim is not that scaling fails, but that it is *differentially productive*: it predominantly buys world knowledge, while the language skills that pipeline components rely on plateau early. Compact-model research—Phi-3 (Abdin et al., 2024b), MiniCPM (Hu et al., 2024), the Qwen3 small sizes (Qwen Team, 2025), and the distillation literature (Hinton et al., 2015)—shows small models can be strong when data quality substitutes for scale; our novelty is the *deliberate selection of which abilities to train* in a target-language component model, together with a measurement design that separates the two ability families.

Concurrent evidence for the decomposition. Several concurrent reports corroborate the split from complementary angles. OCC-RAG mid-trains compact Qwen3 models as "cognitive cores" for context-faithful QA and shows, via counterfactual contexts, that parametric-knowledge interference *grows* with model size (Savkin et al., 2026); NesyCD separates general abilities (distillable into small weights) from sparse specialized knowledge (better kept in an external symbolic store) (Liao et al., 2025); across the Gemma-4 dense ladder, instruction-following and agentic scores stay nearly flat from 4.5B to 31B while knowledge benchmarks climb steeply (Gemma Team, Google DeepMind, 2026); and an independent Russian tool-calling study reports smaller instruct models

Table 1: The two capability families. "Externalized" abilities are those a pipeline supplies at inference time (retrieved passages, documents, tool results, schemas).

	World knowledge	**Language skills**
Content	facts, entities, dates, associations	comprehension, extraction, structuring, reasoning over given evidence
Source	parameters (closed-book)	context (open-book)
Scaling with size	steep, roughly monotone	early saturation, then flat
Failure if missing	factual errors, hallucination	malformed output, lost facts
Benchmarks	MMLU-Pro, ruMMLU, CheGeKa, ruWorldTree, ruOpenBookQA	NEREL-bench, MultiQ, HotpotQA, BFCL

outperforming larger ones on BFCL-style tests (KreGGo, 2026). Data-centric results point the same way: educational and synthetic curricula move reasoning and skill benchmarks far more than knowledge probes (Penedo et al., 2024; Gunasekar et al., 2023; Abdin et al., 2024a).

RAG and GraphRAG. Retrieval-augmented generation externalizes knowledge to a retrievable corpus (Lewis et al., 2020); graph-based variants (GraphRAG, LightRAG, HippoRAG) further restructure the corpus into entity–relation graphs (Edge et al., 2024; Guo et al., 2025; Gutiérrez et al., 2024), making extraction quality the binding constraint on system accuracy. GraphRAG-Bench (Xiang et al., 2026) operationalizes system-level comparison. Prior work by the RAGU team showed that a compact 7B skill-oriented model can replace a 32B model as the in-pipeline extractor with overall system quality preserved (Komarov et al., 2026); we build on that observation, supply the scaling evidence behind it, and analyze when it holds.

Tool use and structured output. Toolformer (Schick et al., 2024), ReAct (Yao et al., 2023), and Gorilla (Patil et al., 2024) established tool invocation as a trainable LLM ability; BFCL (Gorilla Team, UC Berkeley, 2024) and τ-bench (Yao et al., 2025) measure it at component and agentic level, and IFEval (Zhou et al., 2023) and JSONSchemaBench (Geng et al., 2025) the declarative counterpart. In our taxonomy these are language skills; they were deliberately *outside* the training budget of `Meno-Lite-0.1` (whose BFCL-Russian score trails dedicated tool-calling models, §4), and the documented gap is the clearest boundary of the present recipe.

Multilingual distillation and tokenizer adaptation. Vocabulary substitution with continued pretraining adapts multilingual models to a target language (Tikhomirov et al., 2024); `Meno-Lite-0.1` inherits such a tokenizer (3.77 vs. 2.57 chars/token on Russian, a 47% gain), which converts directly into fewer autoregressive steps per Russian answer.

Sustainability accounting. Patterson et al. (2021) and Luccioni et al. (2023) established methodology for neural-network carbon accounting. We apply it at the *component* level: the relevant comparison is not training-time emissions alone but the per-request energy of a pipeline component that runs on every query.

3 Language skills versus world knowledge

Table 1 operationalizes the decomposition. The hypothesis has a testable consequence. Fix a suite of context-grounded tasks (skills) and closed-book tasks (knowledge), and evaluate a ladder of model sizes in two conditions, with and without pipeline scaffolding:

$$\text{performance}(N) \;=\; \begin{cases} \text{steeply increasing in } N & \text{knowledge tasks,} \\ \text{saturating early in } N \text{ given skill training} & \text{skill tasks,} \end{cases} \tag{1}$$

and, critically, the *harness sensitivity*—the gain from adding scaffolding (retrieval, program-of-thought execution (Chen et al., 2023; Gao et al., 2023), self-consistency sampling (Wang et al.,

Table 2: Knowledge-graph construction on NEREL-bench (Bondarenko, 2026b), derived from the NEREL corpus (Loukachevitch et al., 2021): harmonic mean (the benchmark's official aggregate) over entity recognition (ER, F1), entity definition (ED, chrF++), relation extraction (RE, F1), and relation definition (RD, chrF++); chrF++ is the character/word n-gram F-score (Popović, 2017). Our evaluation, greedy decoding, chat templates applied; the SFT training set shares the NEREL annotation schema (see §7).

Model	Size	HM	ER	ED	RE	RD
Meno-Lite-0.1	7B	**0.4676**	0.5043	0.5273	**0.3469**	0.5582
Qwen2.5-32B-Instruct	32B	0.4163	**0.5361**	**0.5275**	0.2393	**0.5993**
gemma-3-12b-it	12B	0.4075	0.5136	0.4955	0.2450	0.5649
gemma-3-27b-it	27B	0.3964	0.5436	0.4818	0.2243	0.5827
Qwen3-30B-A3B-Instruct-2507	30B	0.3962	0.4904	0.4077	0.2623	0.5582
Qwen2.5-14B-Instruct	14B	0.3957	0.5096	0.5182	0.2222	0.5829
gemma-3-4b-it	4B	0.3648	0.4468	0.3958	0.2356	0.5113
Qwen2.5-7B-Instruct	7B	0.3558	0.4770	0.4790	0.1919	0.5411
A-vibe	8B	0.3483	0.4683	0.4351	0.2207	0.3971
Qwen3-4B-Instruct-2507	4B	0.3451	0.5031	0.4363	0.1809	0.5613
T-lite-it-1.0	7B	0.3356	0.4660	0.4644	0.1741	0.5329
GigaChat3-10B-A1.8B	10B	0.3210	0.4532	0.4277	0.1659	0.5291
GigaChat3.1-10B-A1.8B	10B	0.3088	0.4372	0.4102	0.1576	0.5316
T-lite-it-2.1	8B	0.2845	0.4889	0.3933	0.1308	0.5469
RuadaptQwen2.5-7B-Lite-Beta (ancestor)	7B	0.2642	0.4208	0.3925	0.1215	0.5041
gemma-3-1b-it	1B	0.2039	0.3298	0.3006	0.0961	0.3511
Qwen2.5-1.5B-Instruct	1.5B	0.1864	0.3052	0.3265	0.0775	0.4518
Qwen2.5-3B-Instruct	3B	0.1337	0.3959	0.4088	0.0441	0.4417

2023), schema constraints)—should be largest for small models, because scaffolding substitutes for exactly the resource small models lack.

Published results already triangulate this. On the DocInsights document-reasoning shared task, adding a program-of-thought harness moved joint accuracy by +0.282 for a compact 7B model but only +0.005 for a 72B model, and a 27B model with the full harness matched the 72B model (0.884 vs. 0.873) at 2.7× fewer parameters and roughly a quarter of the CO_2 (Singh et al., 2026). On knowledge-graph construction, `Meno-Lite-0.1`—a 7B-class model trained for skills—beats models up to 32B (Bondarenko, 2026b; Komarov et al., 2026). Conversely, on closed-book multiple-choice knowledge (ruMMLU) the same model stays at its 7B backbone's level while the backbone family climbs onward to 32B (§5); we return in §5.5 to why its generative quiz score is *not* evidence against this asymmetry.

4 MENO-LITE-0.1: A SKILL-ORIENTED COMPACT MODEL

`Meno-Lite-0.1` (Bondarenko, 2026a) is a ∼7B bilingual (Russian-primary, English-retained) model descending from Qwen2.5-7B-Instruct (Yang et al., 2024) through T-lite-it-1.0 and RuadaptQwen2.5-7B-Lite-Beta (Tikhomirov et al., 2024), the latter contributing the Russian-extended tokenizer. Its two training stages implement the skill-first budget:

Continued pretraining (1.3B tokens). A balanced Russian–English mix of educational and scientific-technical text: sampled FineWeb-Edu (EN), a RuLM subset filtered for maximal FineWeb-Edu similarity with multilingual embeddings, Russian FinePDFs-edu, and RuREBus corpora. The selection targets *linguistic quality*, not factual breadth.

Supervised fine-tuning (50M tokens). Skill instructions that teach the model to *use context rather than recall facts*: NEREL-derived extraction instructions (Loukachevitch et al., 2021), RAG query logs over real corpora, multi-hop QA dialogs (MultiHopRAG, Tang & Yang, 2024), multi-turn RAG dialogs (MTRAGEval, Rosenthal et al., 2026), and self-cognition samples.

(i) Extraction and structuring (Table 2). `Meno-Lite-0.1` holds the highest harmonic mean of the eighteen models evaluated, a +12.5% relative gain over Qwen2.5-32B-Instruct, driven by relation extraction (0.3469 vs. 0.2393 F1)—the subtask that most requires binding entity mentions into typed relations *from the given text* rather than reproducing definition-style boilerplate. Two rows act as a natural ablation. Its direct ancestor RuadaptQwen2.5-7B-Lite-Beta shares the same tokenizer and backbone but scores 0.2642: the skill-oriented pretraining and SFT contribute $+0.20$ harmonic mean at fixed size. And the 1.5B `meno-tiny-0.1`, trained with the same recipe, scores 0.1726—above the 4.6$\times$ larger Qwen2.5-3B-Instruct (0.1337), indicating that the skill transfers down the size axis. No same-size competitor (A-vibe, T-lite-it-1.0/2.1, GigaChat3) comes within 0.11 of the 7B model.

(ii) Context use (official MERA-1.0 leaderboard numbers (Fenogenova et al., 2024)). On MultiQ open-book multi-hop QA the model scores 0.536/0.403 (F1/EM): above its bare backbone Qwen2.5-7B (0.425/0.296), on par with the best same-size Russian-adapted peers (A-vibe 0.539/0.41; T-lite-it-1.0 0.523/0.398), and at 92% of Qwen2.5-32B's level (0.585/0.452). On closed-book knowledge it stays at its backbone's level—ruWorldTree 0.949 vs. 0.939 (7B), ruOpenBookQA 0.88 vs. 0.845 (7B), while the family climbs to 0.983/0.923 at 32B: the intended asymmetry. Its overall MERA score (0.555) falls between Qwen2.5-7B and -32B, on par with 70B-class open models (e.g., Llama-3.3-70B at 0.555). Its CheGeKa quiz score (0.346/0.293) *exceeds the entire backbone ladder including 72B* (0.316/0.248); we show in §5.5 that this task is contaminated by answer-format compliance, which is precisely why it cannot serve as a clean knowledge axis here. On LIBRA (AI Forever Research, 2025), passkey retrieval stays at 0.98 up to 128k tokens.

(iii) As a pipeline component. As the extractor of the RAGU GraphRAG engine it matches a 32B engine with the best evidence recall on GraphRAG-Bench (Medical) (0.84 vs. $\leq$0.76 Komarov et al., 2026; Xiang et al., 2026); as an ensemble member it contributed to the first-place SemEval-2026 Task 8 system (Bondarenko et al., 2026).

Known weakness. On the BFCL-Russian tool-calling benchmark the model scores 58.9%, far below dedicated tool-calling peers (KreGGo, 2026), and on MERA Text 2.0 (AI Alliance Russia, 2026) its Agentic score is 0.029 (overall 0.104). Tool calling and schema-constrained generation are language skills in our taxonomy—but they were deliberately not in the 0.1 training budget, and they respond to training rather than scale; closing this gap is the goal of our ongoing successor work.

5 Scaling analysis: two families, two curves

The evaluation instantiates Eq. 1 as a grid: *model size* (the Qwen2.5-Instruct ladder 0.5B–72B, the family of `Meno-Lite-0.1`) $\times$ *capability family* (knowledge: CheGeKa, ruMMLU, MMLU-Pro, ruOpenBookQA; skills: NEREL-bench, MultiQ, HotpotQA) $\times$ *context faithfulness* (memorization ratio). Sources are layered: official MERA-1.0 leaderboard numbers are primary for MERA tasks; NEREL-bench numbers are our published evaluation (Table 2); ruMMLU, MMLU-Pro, and HotpotQA come from our local runs under the official protocols (App. A). Benchmarks that appeared in the model's training data are excluded by construction.

5.1 Knowledge climbs, skills plateau

Table 3 gives the raw scores; Figure 1 condenses them into family corridors (each task normalized by its best ladder score). Below 7B, both families grow with size—mastery of instruction following and reading arrives late for very small models. Above 7B they diverge. Knowledge continues to climb: CheGeKa F1 grows $\times 4.1$ from 7B to 72B (0.077$\rightarrow$0.316; $\times 21$ over the whole ladder), ruMMLU $+22\%$ (0.612$\rightarrow$0.747), MMLU-Pro $+21\%$ (0.566$\rightarrow$0.682). Skills flatten: NEREL HM gains $+17\%$ from 7B to 32B (0.356$\rightarrow$0.416), MultiQ F1 rises $\times 1.35$ to 32B and is *flat* from 32B to 72B (0.585 vs. 0.576), HotpotQA EM gains $+4\%$ (0.535$\rightarrow$0.556, non-monotone through 14B). Against these curves, the skill-trained 7B model sits *above the entire NEREL-bench ladder* (0.468 vs. 0.416 for 32B) and at 92% of the family's MultiQ asymptote, while its knowledge remains at its backbone's level (ruMMLU 0.624 vs. 0.612). Exactly as Eq. 1 predicts: size buys the knowledge axis, skill training moves the point on the skills axis at fixed size.

Table 3: The size ladder. CheGeKa/MultiQ: official MERA-1.0 leaderboard (F1). ruMMLU: official protocol, full test (14,012 examples). MMLU-Pro: our run (11,149 examples). NEREL: harmonic mean (Table 2). HotpotQA: our task implementation, EM, distractor mode. MR: memorization ratio on ConFiQA counterfactuals (500 examples; lower is better). "—" = not evaluated. †format-contaminated (§5.5).

Model	CheGeKa	ruMMLU	MMLU-Pro	MultiQ	NEREL	HotpotQA	MR
Qwen2.5-0.5B	0.015	0.345	0.168	0.146	—	0.084	0.272
Qwen2.5-1.5B	0.021	0.451	0.320	0.275	0.186	0.406	—
Qwen2.5-3B	0.029	0.533	0.448	0.391	0.134	0.359	0.232
Qwen2.5-7B	0.077	0.612	0.566	0.425	0.356	0.535	0.443
Qwen2.5-14B	—	0.700	0.636	—	0.396	0.515	0.738
Qwen2.5-32B	0.191	0.747	0.682	0.585	0.416	0.556	0.735
Qwen2.5-72B	0.316	—	—	0.576	—	—	—
`Meno-Lite-0.1` (7B)	0.346†	0.624	—	0.536	**0.468**	—	**0.161**

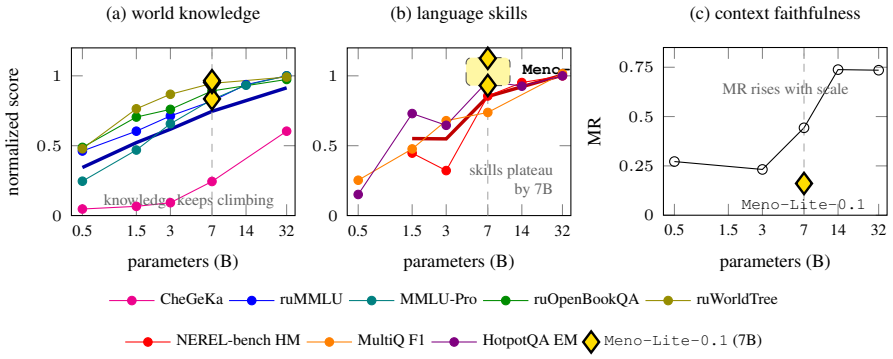

Figure 1: Scaling of the two capability families and context faithfulness over the Qwen2.5-Instruct ladder (log parameter axis). Each task's scores are normalized by that task's best value over the evaluated ladder (including 72B where available); shaded corridors show the family mean ± one standard deviation across tasks on the common size grid (tasks as in the legend). Knowledge keeps climbing through 32B; skills reach ∼85% of their ladder maximum by 7B and plateau. Gold diamonds mark `Meno-Lite-0.1` (7B): its knowledge points sit inside the 7B corridor cross-section (ruMMLU 0.84, ruOpenBookQA 0.95, ruWorldTree 0.97), while its skills form the gold band at 1.03 ± 0.10 (NEREL-bench HM 1.13, MultiQ 0.93)—the level the ladder reaches only at 32B (HotpotQA and MMLU-Pro were not evaluated for Meno; its CheGeKa score is excluded as format-contaminated, §5.5). (c) Memorization ratio $\mathrm{MR} = P_0/(P_0 + P_c)$: the share of answers following the memorized fact over the counterfactual one given in the context (lower is better); it rises with scale, and `Meno-Lite-0.1` (0.161) is the most context-faithful model measured.

5.2 Context faithfulness: the memorization ratio

For open-book components, fidelity to the *given* context matters as much as skill level: a reader that overrides the corpus with parametric memory is unsafe in a RAG pipeline. We measure the memorization ratio $\mathrm{MR} = P_0/(P_0+P_c)$ on 500 counterfactual question–answer pairs (ConFiQA split; the protocol of Savkin et al., 2026): P_0 is the frequency of answering with the original (memorized) fact, P_c with the counterfactual stated in the context; lower is better. MR *rises* with scale: 0.272 (0.5B), 0.232 (3B), 0.443 (7B), 0.738 (14B), 0.735 (32B)—larger models increasingly answer from parameters over the given context, in line with OCC-RAG (Savkin et al., 2026). `Meno-Lite-0.1` has the lowest MR of all models measured, 0.161—below even the 0.5B model—while its control accuracy on original facts (0.902) shows the low MR is not incapacity but a genuine reader stance; it is also by far the most concise responder (13.4 tokens/answer vs. 44–399 across the ladder). Same-size adapted peers sit far higher (T-lite-it-1.0: 0.449; A-vibe: 0.428). Caveats: single seed, an English counterfactual split, and 16–34% of answers (across models) contain neither fact.

Table 4: Russian-adapted peers, single protocol identical to the ladder (our replication layer; NEREL HM computed from the four subtask scores of this layer). Qwen2.5 rows repeated for reference. Saiga's HotpotQA EM is a format artifact (answers 5× longer than gold; F1 0.425; §5.5).

Model	Size	CheGeKa EM	ruMMLU	HotpotQA EM	NEREL HM
Saiga-Llama-3-8B	8B	0.061	0.513	0.244	0.258
A-vibe	8B	0.086	0.692	0.515	0.346
T-lite-it-2.1	8B	0.112	0.688	0.569	0.268
GigaChat-20B-A3B	20B	0.149	0.584	0.416	0.298
GigaChat3-10B-A1.8B	10B	0.157	0.662	0.555	0.302
YandexGPT-5-Lite-8B	8B	0.199	0.663	0.416	0.379
RuadaptQwen2.5-7B-Lite	7B	0.110	0.635	0.390	0.258
Qwen2.5-7B	7B	0.048	0.612	0.535	0.352
Qwen2.5-14B	14B	0.100	0.700	0.515	0.383
Qwen2.5-32B	32B	0.119	0.747	0.556	0.355
`Meno-Lite-0.1`	7B	0.115	0.624	—	**0.456**

5.3 Same-size adapted peers

A natural objection is that Russian-adapted same-size models make better components. Table 4 evaluates seven of them—including the three families suggested by prior critiques (Saiga-Llama-3-8B, A-vibe, T-lite-it-2.1) plus GigaChat and YandexGPT entries—under a single protocol identical to the ladder (App. A). Three findings. **(1) The ordering replicates.** No adapted peer reaches the 32B backbone on NEREL-bench (best: YandexGPT-5-Lite, HM 0.379), while `Meno-Lite-0.1` is highest in *both* evaluation protocols (0.468 primary; 0.456 in this replication layer). **(2) Adaptation buys knowledge, not extraction.** Four of seven peers beat Qwen2.5-32B on CheGeKa (YandexGPT 0.199 vs. 0.119 EM) and two approach 14B on ruMMLU (A-vibe 0.692, T-lite 0.688 vs. 0.700)—yet none matches 32B on structural extraction: knowledge is cheaply purchasable via language adaptation, schema-faithful reading is not. **(3) The price of adaptation depends on the base.** T-lite (Qwen3-based) does not pay in English skills (HotpotQA 0.569, above 32B), whereas the Qwen2.5-based Ruadapt loses 15 EM points against its own backbone (0.390 vs. 0.535).

5.4 Refusals, not inability, explain the relation-extraction anomalies

The one skill task with a non-monotone ladder is NEREL relation extraction (3B: 0.066 below 1.5B; 32B: 0.173 below 14B). Diagnosing the raw generations shows the metric mixes two behaviors: extraction ability and the *willingness to always answer*. The test split contains no example where "no related entities" is correct, so abstention is penalized by construction; refusal shares are 22.8% (3B), 17.0% (32B), and 12.5% (14B)—exactly the anomalous sizes—while excluding refusals recovers monotonicity (3B 0.085, 14B 0.232, 32B 0.208; App. D). `Meno-Lite-0.1` never refuses (0.0%) and is best on the task by a wide margin (0.328). Abstention is penalized by construction here; we report refusal shares alongside F1 to keep the comparison honest.

5.5 Protocol sensitivity in generative evaluation

Assembling this evaluation surfaced a methodological hazard that we believe deserves standalone attention: *generative benchmarks scored against short references can measure answer format rather than ability, and the artifact grows with model size*. We hit it three times, always with instruct models, and only ever diagnosed it by reading raw generations—in every case the score curve looked plausible.

(a) Missing chat template (Table 5). Fed a bare prompt, an instruct model operates in text-continuation mode: on NEREL relation definition, Qwen2.5-32B loops in 65.5% of answers (chrF 0.224); with its template the same model solves the task (0.577, zero loops). On HotpotQA the template-less ladder was *inverted* (7B/14B/32B EM = 0.512/0.379/0.280) purely because larger models wrote longer continuations.

Table 5: Chat-template effects on generative tasks with short references (single-flag controlled runs; lengths in characters). Without a template, instruct models continue the document: NEREL degenerates into loops, HotpotQA answers balloon and *invert* the ladder—the larger the model, the larger the artifact.

Task	Model	w/o template	with template	len. w/o	len. with
NEREL RD (chrF)	Qwen2.5-32B	0.224	**0.577**	1357	125
HotpotQA (EM)	Qwen2.5-7B	0.512	**0.535**	20	14
HotpotQA (EM)	Qwen2.5-14B	0.379	**0.515**	58	15
HotpotQA (EM)	Qwen2.5-32B	0.280	**0.556**	127	28

Table 6: Serving efficiency on one A100 80GB, identical vLLM stack and fixed workload. The tokenizer advantage is Russian-only; English throughput and density are unchanged.

Model	tok/s	chars/tok RU	chars/tok EN	chars/s RU	chars/s EN
`Meno-Lite-0.1`	928.8	**4.42**	4.49	**4105**	4168
Qwen2.5-7B-Instruct	926.4	2.90	4.49	2687	4160
Qwen2.5-32B-Instruct	229.2	2.90	4.49	665	1029

(b) The assistant role is not brevity. The converse artifact: on CheGeKa (official MERA protocol, 1-shot, no template) *adding* the template *hurts* every model from 3B up, by -8% to -29% EM (e.g., 7B: 0.048→0.035; 32B: 0.119→0.109; `Meno-Lite-0.1`: 0.115→0.097; App. C). A controlled run with the few-shot example glued into a single turn rules out example-splitting as the cause: cast as an assistant, the model explains and normalizes (for the gold answer *Two*, official mode returns "TWO" while template mode returns "2"), and strict matching against the short gold collapses. Saiga-Llama-3-8B shows the same failure *with* the template on HotpotQA: EM 0.244 but F1 0.425, with answers five times longer than gold. We therefore report both CheGeKa estimates side by side, keep the official one primary (leaderboard-comparable), and treat the task's leaderboard ranking as format-contaminated—which is exactly why `Meno-Lite-0.1`'s CheGeKa lead over the 72B model (§4) measures answer-format compliance, a language skill, and not knowledge.

(c) Rule of thumb. A generative benchmark with short references measures output format until the prompt conditions are pinned down—instruction-driven format needs the chat template, few-shot-driven format may be damaged by it—and the size-dependence of the artifact makes it a confounder specifically for scaling claims about small versus large models. Verification requires reading generations, not inspecting curves.

6 Efficiency and environmental accounting

Table 6 reports measured serving efficiency (identical vLLM stack (Kwon et al., 2023)). At identical token throughput to its same-size backbone, the Russian-extended tokenizer packs $1.53\times$ more text per token, so `Meno-Lite-0.1` emits 4105 Russian characters/s—$1.53\times$ the 7B backbone and $6.2\times$ the 32B model that it matches on extraction quality. Against a 32B component, per-request energy and CO_2e fall by roughly the same factor as active-parameter compute ($\sim$4.6$\times$) under the accounting conventions of Strubell et al. (2019); Patterson et al. (2021); Luccioni et al. (2023); a published reference point for the harness side is the DocSem system, where a 27B-with-harness matched a 72B model at $\approx 4\times$ lower CO_2 (Singh et al., 2026). A full per-request energy audit (GPU power integrated over request windows, grid carbon intensities) is left to the revision.

7 Discussion and limitations

Where the trade-off breaks. The hypothesis predicts failure modes that we observe: closed-book factual QA and competition mathematics beyond the given context, agentic tool use (MERA-2.0 Agentic 0.029, BFCL-Russian 58.9%). The recipe is a statement about *components*, not about standalone assistants.

Russian-centric validation. Our strongest evidence is bilingual with a Russian skew; the tokenizer adaptation, several training classes, and the MERA axes are Russian-specific. The English side (MMLU-Pro, HotpotQA, ConFiQA) relies on public benchmarks only.

Single ladder family. The scaling curves cover one backbone family (Qwen2.5); replication on further dense families is natural future work, as is applying the recipe to a larger base to test the saturation claim directly.

Distributional caveats. The NEREL SFT/eval lineage shares annotation schema and text domain (Table 2 carries this advantage; independent IE benchmarks are needed); CheGeKa is scored locally on the train split under the official protocol (test answers hidden); MR uses a single seed and an English counterfactual split with 16–34% undecided answers; the adapted-peer layer is our replication, with one peer (T-lite-it-2.1) showing an unexplained entity-definition deficit we verified is not a format artifact.

Protocol dependence. As §5.5 documents, generative rankings with short references are sensitive to prompt format; our headline skill results (NEREL-bench, MultiQ, HotpotQA) use pinned template-on protocols, and the NEREL ordering was replicated under two independent promptings.

8 Conclusion

We argued that what RAG pipelines and agent systems need from their language-model components is not what scale buys. By decomposing LLM ability into world knowledge and language skills, investing the training budget of a compact bilingual model in the latter, and measuring both axes plus context faithfulness across a size ladder, we showed that the families scale differently—knowledge climbs, skills plateau, memorization grows—and that a deliberately trained 7B reader can sit above a 32B model on extraction, at 92% of its multi-hop QA level, below the whole ladder on memorization, and $6.2\times$ faster to serve, at a fraction of the energy. Meno—an inversion of anamnesis: a model built to read rather than recall—is the embodiment; the evaluation-protocol hazards we documented are part of the same lesson, because whether a small model "cannot" do something is itself something that must be measured carefully. Weights, training data, and code are public; a successor targeting the remaining language skills (tool calling, schema-constrained generation) with structural reinforcement learning is in progress.

Reproducibility statement

All training stages are scripted with fixed seeds; the evaluation protocols (splits, shots, chat-template modes, seeds, metric implementations) are specified per benchmark in Appendix A, and the replication layers (ladder, adapted peers, protocol controls) in Appendices B–D. Dataset provenance and licenses are documented in the repository. An anonymous code repository and model checkpoints are provided as supplementary material.

Ethics statement

This work trains on publicly available instruction datasets and publicly released model checkpoints under their respective licenses. The models are intentionally knowledge-limited and intended for deployment *inside* retrieval-grounded pipelines; using them as standalone closed-book assistants will produce factual errors. Efficiency claims are specific to the stated hardware and workloads; per-request CO_2 depends on grid carbon intensity.

AI use statement

In this work we used generative AI tools as follows: (i) an AI coding assistant helped develop the training and evaluation code, with all generated code reviewed, tested, and validated by the authors; (ii) AI assistants helped with literature triage and polished the phrasing of author-written drafts. The authors reviewed and edited all AI-assisted material and take responsibility for the final content of this work.

REFERENCES

Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, et al. Phi-4 technical report, 2024a. URL https://arxiv.org/abs/2412.08905.

Marah Abdin, Sam Ade Jacobs, Ammar Ahmad Awan, Jyoti Aneja, Awais Behnan, Arash Behnam, et al. Phi-3 technical report: A highly capable language model locally on your phone, 2024b. URL https://arxiv.org/abs/2404.14219.

AI Alliance Russia. MERA Text 2.0: A benchmark for russian language models. https://mera.a-ai.ru/ru/text/leaderboard, 2026.

AI Forever Research. LIBRA: A benchmark for Russian long-context language understanding. https://huggingface.co/datasets/ai-forever/LIBRA, 2025.

Ivan Bondarenko. Meno-lite-0.1: A 7b language model optimized for russian rag pipelines, 2026a. URL https://huggingface.co/bond005/meno-lite-0.1.

Ivan Bondarenko. Nerel-bench: A benchmark for evaluating llms on russian knowledge graph construction tasks. https://huggingface.co/datasets/bond005/NEREL_bench, 2026b.

Ivan Bondarenko, Roman Derunets, Oleg Sedukhin, Mikhail Komarov, Ivan Chernov, and Mikhail Kulakov. RaguTeam at SemEval-2026 task 8: Meno and Friends in a judge-orchestrated LLM ensemble for faithful multi-turn response generation. In Ekaterina Kochmar, Debanjan Ghosh, Kai North, Mamoru Komachi, and Marcos Zampieri (eds.), *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, pp. 1678–1694, San Diego, California, USA, July 2026. Association for Computational Linguistics. doi: 10.18653/v1/2026.semeval-1.216. URL https://aclanthology.org/2026.semeval-1.216/.

Wenhu Chen, Xueguang Ma, Xinyi Wang, and William W. Cohen. Program of thoughts prompting: Disentangling computation from reasoning for numerical reasoning tasks. *Transactions on Machine Learning Research*, 2023. ISSN 2835-8856. URL https://openreview.net/forum?id=YfZ4ZPt8zd.

Darren Edge, Ha Trinh, Newman Cheng, Joshua Bradley, Alex Chao, Apurva Mody, Steven Truitt, and Jonathan Larson. From local to global: A graph RAG approach to query-focused summarization. *arXiv preprint arXiv:2404.16130*, 2024. URL https://arxiv.org/abs/2404.16130.

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. MERA: A comprehensive LLM evaluation in Russian. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar (eds.), *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 9920–9948, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.534. URL https://aclanthology.org/2024.acl-long.534/.

Luyu Gao, Aman Madaan, Shuyan Zhou, Uri Alon, Pengfei Liu, Yiming Yang, Jamie Callan, and Graham Neubig. Pal: Program-aided language models. In *Proceedings of the 40th International Conference on Machine Learning*, pp. 10764–10799, 2023. URL https://proceedings.mlr.press/v202/gao23f.html.

Gemma Team, Google DeepMind. Gemma 4 technical report, 2026. URL https://arxiv.org/abs/2607.02770.

Saibo Geng, Hudson Cooper, Michał Moskal, Samuel Jenkins, Julian Berman, Nathan Ranchin, Robert West, Eric Horvitz, and Harsha Nori. JSONSchemaBench: A rigorous benchmark of structured outputs for language models, 2025. URL https://arxiv.org/abs/2501.10868.

Gorilla Team, UC Berkeley. Berkeley function calling leaderboard (BFCL). https://gorilla.cs.berkeley.edu/leaderboard.html, 2024.

Suriya Gunasekar, Yihe Zhang, Jyoti Aneja, Caio César Teodoro Mendes, Allie Del Giorno, Sivakanth Gopi, Baishakhi Jang, Junkang Guo, Rinesh Saarikivi, et al. Textbooks are all you need, 2023. URL https://arxiv.org/abs/2306.11644.

Zirui Guo, Lianghao Xia, Yanhua Yu, Tu Ao, and Chao Huang. LightRAG: Simple and fast retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: EMNLP 2025*, pp. 10746–10761, Suzhou, China, November 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.findings-emnlp.568/.

Bernal Jiménez Gutiérrez, Yiheng Shu, Yu Gu, Michihiro Yasunaga, and Yu Su. Hipporag: Neurobiologically inspired long-term memory for large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang (eds.), *Advances in Neural Information Processing Systems*, volume 37, pp. 59532–59569. Curran Associates, Inc., 2024.

Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015. URL https://arxiv.org/abs/1503.02531.

Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, Tom Hennigan, Eric Noland, Katie Millican, George van den Driessche, Ben Mann, Chin-Leong Lomeli, Nathan Bevan, Malcolm Reynolds, Richard Ring, Yujia Li, Edward Grefenstette, Koray Kavukcuoglu, Demis Hassabis, and Oriol Vinyals. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022. URL https://arxiv.org/abs/2203.15556.

Shengding Hu, Yuge Xu, Jiazheng Yu, Minghao Wang, Quanfeng Huang, Yuxiang Xing, Chenghao Huang, Wendi Xiong, Qingxiu Zhang, Jiezhong Zhao, et al. MiniCPM: Unveiling the potential of small language models with scalable training recipes, 2024. URL https://arxiv.org/abs/2404.06395.

Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B. Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020. URL https://arxiv.org/abs/2001.08361.

Mikhail Komarov, Ivan Bondarenko, Stanislav Shtuka, Oleg Sedukhin, Roman Shuvalov, Yana Dementyeva, Matvey Solovyov, and Nikolay O. Nikitin. Ragu: A multi-step graphrag engine with a compact domain-adapted llm, 2026. URL https://arxiv.org/abs/2607.11683.

M. KreGGo. Llm tool calling benchmark (BFCL Russian). https://github.com/MKreGGo/ru_tool_calling_tests, 2026. Methodology article (Habr): https://habr.com/ru/articles/982156/.

Woosuk Kwon, Zhuohan Li, Siyuan Zhuang, Ying Sheng, Lianmin Zheng, Cody Hao Yu, Joseph Gonzalez, Hao Zhang, and Ion Stoica. Efficient memory management for large language model serving with pagedattention. In *Proceedings of the 29th Symposium on Operating Systems Principles*, pp. 611–626. Association for Computing Machinery, 2023. doi: 10.1145/3600006.3613165. URL https://doi.org/10.1145/3600006.3613165.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. Retrieval-augmented generation for knowledge-intensive nlp tasks. In H. Larochelle, M. Ranzato, R. Hadsell, M.F. Balcan, and H. Lin (eds.), *Advances in Neural Information Processing Systems*, volume 33, pp. 9459–9474. Curran Associates, Inc., 2020. URL https://proceedings.neurips.cc/paper_files/paper/2020/file/6b493230205f780e1bc26945df7481e5-Paper.pdf.

Huanxuan Liao, Shizhu He, Yao Xu, Yuanzhe Zhang, Kang Liu, and Jun Zhao. Neural-symbolic collaborative distillation: Advancing small language models for complex reasoning tasks. In *Proceedings of the AAAI Conference on Artificial Intelligence*, 2025. URL https://arxiv.org/abs/2409.13203.

Natalia Loukachevitch, Ekaterina Artemova, Tatiana Batura, Pavel Braslavski, Ilia Denisov, Vladimir Ivanov, Suresh Manandhar, Alexander Pugachev, and Elena Tutubalina. NEREL: A russian dataset with nested named entities, relations and events. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pp. 876–885, Held Online, 2021. INCOMA Ltd. URL https://aclanthology.org/2021.ranlp-1.100/.

Alexandra Sasha Luccioni, Sylvain Viguier, and Anne-Laure Ligozat. Estimating the carbon footprint of BLOOM, a 176B parameter language model. *Journal of Machine Learning Research*, 24 (253):1–15, 2023.

Shishir G. Patil, Tianjun Zhang, Xin Wang, and Joseph E. Gonzalez. Gorilla: Large language model connected with massive APIs. *Advances in Neural Information Processing Systems*, 37:126544–126565, 2024.

David Patterson, Joseph Gonzalez, Quoc Le, Chen Liang, Lluis-Miquel Munguia, Daniel Rothchild, David So, Maud Texier, and Jeff Dean. Carbon emissions and large neural network training. *arXiv preprint arXiv:2104.10350*, 2021. URL https://arxiv.org/abs/2104.10350.

Guilherme Penedo, Hynek Kydlíček, Loubna Ben Allal, Anton Lozhkov, Margaret Mitchell, Colin Raffel, Leandro Von Werra, and Thomas Wolf. The fineweb datasets: Decanting the web for the finest text data at scale, 2024. URL https://arxiv.org/abs/2406.17557.

Plato. *Plato: Protagoras and Meno*. Penguin Classics, 1956. Translated with an introduction by W. K. C. Guthrie.

Maja Popović. chrF++: words helping character n-grams. In Ondřej Bojar, Christian Buck, Rajen Chatterjee, Christian Federmann, Yvette Graham, Barry Haddow, Matthias Huck, Antonio Jimeno Yepes, Philipp Koehn, and Julia Kreutzer (eds.), *Proceedings of the Second Conference on Machine Translation*, pp. 612–618, Copenhagen, Denmark, September 2017. Association for Computational Linguistics. doi: 10.18653/v1/W17-4770. URL https://aclanthology.org/W17-4770/.

Qwen Team. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

Sara Rosenthal, Vraj Shah, Yannis Katsis, and Marina Danilevsky. SemEval-2026 task 8: MTRAGEval: Evaluating multi-turn RAG conversations. In Ekaterina Kochmar, Debanjan Ghosh, Kai North, Mamoru Komachi, and Marcos Zampieri (eds.), *Proceedings of the 20th International Workshop on Semantic Evaluation (2026)*, pp. 3673–3690, San Diego, California, USA, July 2026. Association for Computational Linguistics. ISBN 979-8-89176-414-9. doi: 10.18653/v1/2026.semeval-1.447. URL https://aclanthology.org/2026.semeval-1.447/.

Aleksandr Savkin, Dmitrii Goncharov, Georgii Gambashidze, Maria Chepurova, Dmitrii Tarasov, Ilya Andriianov, Alena Pugacheva, Artem Konovalov, Anton Galichin, and Ivan Oseledets. OCC-RAG: Optimal cognitive core for faithful question answering, 2026. URL https://arxiv.org/abs/2606.00683.

Timo Schick, Jane Dwivedi-Yu, Roberto Dessì, Roberta Raileanu, Maria Lomeli, Luke Zettlemoyer, Nicola Cancedda, and Thomas Scialom. Toolformer: Language models can teach themselves to use tools. In *Advances in Neural Information Processing Systems*, volume 36, 2024.

Jyotika Singh, Fang Tu, Aziza Mirsaidova, Amit Agarwal, Hitesh Laxmichand Patel, Sandip Ghoshal, Miguel Ballesteros, Karan Dua, Yassine Benajiba, Weiyi Sun, Tao Sheng, Graham Horwood, Sujith Ravi, and Dan Roth. GSM-SEM: Benchmark and framework for generating semantically variant augmentations, 2026. URL https://arxiv.org/abs/2605.07053.

Emma Strubell, Ananya Ganesh, and Andrew McCallum. Energy and policy considerations for deep learning in NLP. *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019. URL https://aclanthology.org/P19-1355/.

Yixuan Tang and Yi Yang. MultiHop-RAG: Benchmarking retrieval-augmented generation for multi-hop queries. *arXiv preprint arXiv:2401.15391*, 2024. URL https://arxiv.org/abs/2401.15391.

Mikhail Tikhomirov, Dmitry Chernyshev, Nikolay Krayko, Vasily Konovalov, Maria Nikishina, and Alexander Panchenko. RuadaptQwen2.5-7B-Lite-Beta: Efficient Russian adaptation of Qwen2.5 via vocabulary substitution and continued pretraining. https://huggingface.co/RefalMachine/RuadaptQwen2.5-7B-Lite-Beta, 2024.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=1PL1NIMMrw.

Yubo Wang, Xueguang Ma, Ge Zhang, Yuansheng Ni, Abhranil Chandra, Shiguang Guo, Weiming Ren, Aaran Arulraj, Xuan He, Ziyan Jiang, et al. MMLU-Pro: A more robust and challenging multi-task language understanding benchmark. *Advances in Neural Information Processing Systems*, 37:95266–95290, 2024.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Brian Ichter, Fei Xia, Ed Chi, Quoc V Le, and Denny Zhou. Chain-of-thought prompting elicits reasoning in large language models. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh (eds.), *Advances in Neural Information Processing Systems*, volume 35, pp. 24824–24837. Curran Associates, Inc., 2022.

Zhishang Xiang, Chuanjie Wu, Qinggang Zhang, Shengyuan Chen, Zijin Hong, Xiao Huang, and Jinsong Su. When to use graphs in RAG: A comprehensive analysis for graph retrieval-augmented generation. In *International Conference on Learning Representations (ICLR 2026)*, 2026. URL https://openreview.net/pdf?id=i9q9xDMjG7.

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, Jiaxi Yang, Jingren Zhou, Junyang Lin, Kai Dang, Keqin Bao, Kexin Yang, Le Yu, Mei Li, Mingfeng Xue, Pei Zhang, Qin Zhu, Rui Men, Runji Lin, Tianhao Li, Tianyi Tang, Tingyu Xia, Xingzhang Ren, Xuancheng Ren, Yang Fan, Yang Su, Yichang Zhang, Yu Wan, Yuqiong Liu, Zeyu Cui, Zhenru Zhang, and Zihan Qiu. Qwen2.5 technical report. *arXiv preprint arXiv:2412.15115*, 2024. URL https://arxiv.org/abs/2412.15115.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William W. Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium, 2018. Association for Computational Linguistics. URL https://aclanthology.org/D18-1259/.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. ReAct: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=WE_vluYUL-X.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=Uc6b4Ae0BR.

Jeffrey Zhou, Tianjian Lu, Swaroop Mishra, Siddhartha Brahma, Sujoy Basu, Yi Luan, Denny Zhou, and Le Hou. Instruction-following evaluation for large language models, 2023. URL https://arxiv.org/abs/2311.07911.

A Benchmark protocols

Multiple-choice tasks (ruMMLU, MMLU-Pro) compare likelihoods and are template-insensitive. Training-data exclusions: MultiHop-RAG, ruMMLU-pro, MTRAGEval, and DocSem-derived sets were excluded from evaluation by construction; NEREL-bench uses the NEREL test split while the model's SFT used train-derived instructions (schema-sharing caveat in §7). References: MERA (Fenogenova et al., 2024); MERA Text 2.0 (AI Alliance Russia, 2026); NEREL (Loukachevitch et al., 2021; Bondarenko, 2026b); HotpotQA (Yang et al., 2018); MMLU-Pro (Wang et al., 2024); chrF++ (Popović, 2017); counterfactual-context protocol (Savkin et al., 2026).

Table 7: Evaluation protocols. MERA-1.0 leaderboard columns are taken from the official leaderboard (test split, official scorer); local layers use the official MERA harness (lm-eval fork) or our task implementations as noted. Greedy decoding throughout; thinking modes disabled where a model has them.

Benchmark	Split	Shots	Template	Seed	Source layer
CheGeKa, MultiQ	test (official)	official	per official	—	MERA-1.0 leaderboard
ruWorldTree, ruOpenBookQA	test (official)	official	per official	—	MERA-1.0 leaderboard
ruMMLU	test, 14,012 ex.	1-shot	off (official)	1234	local, official harness
MMLU-Pro	validation subset	0-shot	off (MC)	42	local
NEREL-bench (Table 2)	test	3-shot	on	42	our published run
NEREL replication layer	test, $\leq$1000 ex.	3-shot	on	1234	local
HotpotQA	dev, 1000 ex.	0-shot	on	1234	local, official metrics
MR / ConFiQA	500 counterfactual	0-shot	on	42	local
Adapted peers	as above per task	as above	as above	1234	local

B ADAPTED PEERS: FULL PER-TASK RESULTS

Table 8: Russian-adapted peers, replication layer (chat templates on, greedy, seed 1234). CheGeKa EM/F1 (official 1-shot protocol); ruMMLU acc (full test); HotpotQA EM/F1; NEREL subtasks ER/ED/RE/RD.

Model	CheGeKa	ruMMLU	HotpotQA	ER	ED	RE	RD
Saiga-Llama-3-8B	.061/.101	.513	.244/.425	.474	.479	.106	.524
A-vibe (8B)	.086/.142	.692	.515/.672	.462	.435	.217	.405
T-lite-it-2.1 (8B)	.112/.162	.688	.569/.709	.501	.390	.117	.554
GigaChat-20B-A3B	.149/.224	.584	.416/.555	.398	.383	.158	.504
GigaChat3-10B-A1.8B	.157/.219	.662	.555/.696	.454	.406	.150	.532
YandexGPT-5-Lite-8B	.199/.277	.663	.416/.584	.506	.485	.210	.574
RuadaptQwen2.5-7B-Lite	.110/.168	.635	.390/.516	.439	.394	.115	.514
Qwen2.5-7B	.048/.084	.612	.535/.676	.456	.472	.191	.550
Qwen2.5-32B	.119/.179	.747	.556/.716	.523	.527	.173	.592
Meno-Lite-0.1	.115/.161	.624	—	.488	.530	.328	.557

C CHEGEKA UNDER THREE PROMPT MODES

Table 9: CheGeKa EM under three modes: the official protocol (1-shot, no template), with chat template, and with template but the few-shot example glued into one turn. The template hurts every model from 3B up; gluing the example does not recover the loss, isolating the assistant role (verbosity, normalization) as the cause.

Model	official	+template	+template, glued example
Qwen2.5-0.5B	0.001	0.001	—
Qwen2.5-1.5B	0.004	0.005	—
Qwen2.5-3B	0.017	0.012	0.016
Qwen2.5-7B	0.048	0.035	0.026
Qwen2.5-14B	0.100	0.077	—
Qwen2.5-32B	0.119	0.109	0.102
`Meno-Lite-0.1`	0.115	0.097	—

Representative answers for the gold "Two": official mode "TWO", template mode "2"; for the gold "LIVER": official mode returns the word alone, while template mode returns a full explanatory sentence ("When blood sugar rises, the liver converts excess glucose into glycogen…"). F1 moves in lockstep with EM (7B 0.084→0.063; 32B 0.179→0.163).

D RELATION-EXTRACTION REFUSALS

Table 10: NEREL relation extraction: refusal share ("no related entities" answers; never correct in the test split) and F1 with refusals excluded. The anomalous ladder points (3B, 32B) are exactly the frequent abstainers.

Model	F1	refusal share	F1 w/o refusals
Qwen2.5-0.5B	0.099	1.2%	0.098
Qwen2.5-1.5B	0.114	1.7%	0.116
Qwen2.5-3B	0.066	22.8%	0.085
Qwen2.5-7B	0.191	1.7%	0.194
Qwen2.5-14B	0.209	12.5%	0.232
Qwen2.5-32B	0.173	17.0%	0.208
`Meno-Lite-0.1`	**0.328**	**0.0%**	**0.328**