

Cross-Platform Evaluation of Reasoning Capabilities in Foundation Models

J. de Curtò^{a,b,*}, I. de Zarzà^c, Pablo García^b and Jordi Cabot^c^a*BARCELONA Supercomputing Center (BSC), Barcelona, Spain*^b*Universidad Pontificia Comillas, Madrid, Spain*^c*LUXEMBOURG Institute of Science and Technology (LIST), Esch-sur-Alzette, Luxembourg*

ARTICLE INFO

Keywords:

Foundation Models
 Large Language Models
 Reasoning Evaluation
 Artificial Intelligence
 Supercomputing
 Cloud AI Infrastructure

ABSTRACT

This paper presents a comprehensive cross-platform evaluation of reasoning capabilities in contemporary foundation models, establishing an infrastructure-agnostic benchmark across three computational paradigms: HPC supercomputing (MareNostrum 5), cloud platforms (Nebius AI Studio), and university clusters (a node with eight H200 GPUs).

We evaluate 15 foundation models across 79 problems spanning eight academic domains (Physics, Mathematics, Chemistry, Economics, Biology, Statistics, Calculus, and Optimization) through three experimental phases: (1) *Baseline establishment*: Six models (Mixtral-8x7B, Phi-3, LLaMA 3.1-8B, Gemma-2-9b, Mistral-7B, OLMo-7B) evaluated on 19 problems using MareNostrum 5, establishing methodology and reference performance; (2) *Infrastructure validation*: The 19-problem benchmark repeated on university cluster (seven models including Falcon-Mamba state-space architecture) and Nebius AI Studio (nine state-of-the-art models: Hermes-4 70B/405B, LLaMA 3.1-405B/3.3-70B, Qwen3 30B/235B, DeepSeek-R1, GPT-OSS 20B/120B) to confirm infrastructure-agnostic reproducibility; (3) *Extended evaluation*: Full 79-problem assessment on both university cluster and Nebius platforms, probing generalization at scale across architectural diversity.

Results challenge prevailing scaling assumptions through a parameter efficiency paradox: Hermes-4-70B (70B parameters) achieves the highest score among extended models (0.598), outperforming both its 405B counterpart (0.573) and Meta's LLaMA 3.1-405B (0.560). Domain-specific analysis reveals LLaMA 3.1-405B achieves record Calculus performance (0.717), while DeepSeek-R1 sets unprecedented standards for reasoning transparency (0.716 step-accuracy). Qwen3 models demonstrate exceptional consistency (0.013 score variance, 3× better than alternatives).

We identify a fundamental transparency-correctness trade-off: DeepSeek-R1's high step-accuracy (0.716) correlates weakly with final answers ($r=0.249$), whereas Qwen3 exhibits near-zero correlation ($r=0.095$), suggesting "shortcut learning" that bypasses explicit reasoning chains. Longitudinal comparison (2024 vs 2025) reveals domain-specific evolution: Calculus improved dramatically from mid-tier to top-ranked domain (+24.7%), while Optimization remains universally challenging across model generations (+4.7% improvement only). Cross-model disagreement analysis identifies Physics kinematics as the most controversial domain (std dev 0.335).

Infrastructure validation across three platforms confirms reasoning quality is model-intrinsic: performance remains consistent across HPC (MareNostrum 5), cloud (Nebius), and university cluster environments (<3% variance: LLaMA-3.1-8B – 2.9%, Phi-3-mini – 1.1%), democratizing rigorous evaluation beyond supercomputing institutions. University cluster validation reveals competitive performance of non-transformer architectures (Falcon-Mamba state-space model achieves 0.590, matching transformer baseline LLaMA-3.1-8B at 0.576) and demonstrates that dense smaller models (Phi-4-mini, 14B) outperform larger sparse MoE architectures (Phi-3.5-MoE, 42B: 0.674 vs 0.569).

These findings challenge conventional scaling assumptions, establish training data quality as more critical than model size, and provide actionable guidelines for model selection across educational, production, and research contexts. The tri-infrastructure methodology and 79-problem benchmark enable longitudinal tracking of reasoning capabilities as foundation models evolve.

*Corresponding author.

✉ jdecurto@icai.comillas.edu (J.d. Curtò); irene.zarza@list.lu (I.d. Zarzà); pgarciamolina@alu.icai.comillas.edu (P. García); jordi.cabot@list.lu (J. Cabot)

ORCID(s): 0000-0002-8334-4719 (J.d. Curtò); 0000-0002-5844-7871 (I.d. Zarzà); 0000-0003-2418-2489 (J. Cabot)

1. Introduction

Large language models (LLMs) have rapidly transformed artificial intelligence by extending natural language understanding and reasoning into complex analytical tasks. While prior work has assessed general comprehension and task performance, the *cross-domain reasoning consistency* of these systems remains poorly understood.

We evaluate 15 foundation models across 79 problems spanning eight academic domains (Physics, Mathematics, Chemistry, Economics, Biology, Statistics, Calculus, and Optimization) through three complementary experimental phases: (1) *Baseline establishment*: Six models (Mixtral-8x7B, Phi-3, LLaMA 3.1-8B, Gemma-2-9b, Mistral-7B, OLMo-7B) evaluated on 19 problems using MareNostrum 5, establishing methodology and reference performance; (2) *Infrastructure validation*: Seven models including non-transformer architectures (Falcon-Mamba state-space model) on university cluster and nine state-of-the-art models—Hermes-4 (70B and 405B), LLaMA 3.1-405B and 3.3-70B, Qwen3 (30B and 235B), DeepSeek-R1, and GPT-OSS (20B and 120B)—on Nebius cloud platform, both using the 19-problem set to confirm infrastructure-agnostic reproducibility; (3) *Extended evaluation*: Full 79-problem benchmark on university infrastructure and Nebius platform, probing generalization at scale.

This tri-infrastructure design allows us to validate results under distinct computational environments—supercomputing (MareNostrum 5), cloud platforms (Nebius AI Studio), and university clusters—thereby maximizing external validity and reproducibility while democratizing access to rigorous reasoning evaluation beyond specialized HPC facilities.

The remainder of this article is organized as follows. Section 2 reviews the evolution of foundation models, reasoning evaluation benchmarks, and cross-domain evaluation methodologies, contextualizing our work within the broader AI research landscape. Section 3 presents the initial baseline evaluation conducted on MareNostrum 5 supercomputer, establishing the evaluation methodology and reference performance metrics for six foundation models across 19 problems. Section 4 describes the expansion from the original 19-problem baseline to the comprehensive 79-problem benchmark, detailing domain coverage, difficulty stratification, and problem selection criteria. Section 5 formalizes our evaluation framework, including semantic similarity metrics, step-wise accuracy measurement, and consistency quantification across computational platforms. Section 6 presents an infrastructure-agnostic validation study on the 19-problem benchmark, evaluating seven models (including legacy architectures and state-space models) across university cluster and comparing results with the MareNostrum 5 baseline to establish reproducibility with <3% variance across computational platforms, as well as using nine additional state-of-the-art models on Nebius AI Studio infrastructure, including large-scale variants (Hermes-4-405B, LLaMA 3.1-405B, Qwen3-235B). Section 7 extends the evaluation to 79 problems enabling comprehensive analysis of domain-specific performance, the parameter efficiency paradox, and the transparency-correctness trade-off at scale. Section 8 synthesizes findings across all three infrastructures (MareNostrum 5, Nebius AI Studio, and university cluster), analyzing longitudinal evolution patterns, cross-model disagreement, and architectural diversity across 15 foundation models. Finally, Section 9 consolidates key findings, discusses practical implications for model selection across educational, production, and research contexts, and outlines future research directions for advancing reasoning-capable AI systems.

2. Background and Related Work

LLMs have evolved through successive generations of transformer-based architectures (Vaswani et al., 2017), including BERT (Devlin et al., 2018), GPT-3 (Brown et al., 2020), and PaLM (Chowdhery et al., 2022). More recent work has emphasized efficiency (e.g., Phi-3 (Abdin et al., 2024), Gemma (Team et al., 2024)) and mixture-of-experts architectures (e.g., Mixtral (Mistral AI, 2023)). Reasoning-oriented evaluations—ARC (Clark et al., 2018), MMLU (Hendrycks et al., 2021), and chain-of-thought prompting (Wei et al., 2022b)—have demonstrated that explicit reasoning steps improve performance but not necessarily correctness or consistency.

Our framework extends these evaluations by systematically measuring semantic similarity and step-wise correctness across domains and infrastructures.

The trajectory from early transformer models to contemporary foundation models represents one of the most rapid capability escalations in AI history.

(Kaplan et al., 2020) established power-law relationships between model scale, data size, and performance, predicting that larger models trained on more data would consistently improve. This motivated the development of massive models like GPT-3 (Brown et al., 2020) (175B parameters) and PaLM (Chowdhery et al., 2022) (540B parameters). However, (Wei et al., 2022a) documented *emergent capabilities*—abilities that appear suddenly at certain scale thresholds rather than gradually improving. Reasoning abilities, particularly multi-step logical deduction, were identified as one such emergent property.

Recent work challenges pure scaling optimism Zhang et al. (2025); Buscemi et al. (2025). (Hoffmann et al., 2022) (Chinchilla) demonstrated that many models are under-trained relative to their parameter count, suggesting data quality and duration matter as much as model size.

Reacting to the computational demands of massive models, researchers have pursued efficiency through:

- **Mixture-of-Experts (MoE):** Mixtral (Mistral AI, 2023) and Switch Transformer (Fedus et al., 2022) activate only subsets of parameters per input, reducing inference cost while maintaining capacity. Our evaluation confirms MoE models' balanced cross-domain performance.
- **Knowledge Distillation:** Phi-3 (Abdin et al., 2024) achieves strong performance at 3.8B parameters by distilling from larger teacher models and curating high-quality training data. This "small language models" trend prioritizes efficiency.
- **Structured State Space Models (SSMs):** Mamba (Gu and Dao, 2023) and variants offer alternatives to attention mechanisms with better scaling properties for long sequences, though their reasoning capabilities remain under-explored compared to transformers.

Early benchmarks like GLUE (Wang et al., 2018) and SuperGLUE (Wang et al., 2019) assessed language understanding but largely tested pattern recognition rather than multi-step reasoning. The AI2 Reasoning Challenge (ARC) (Clark et al., 2018) introduced science questions requiring knowledge retrieval and basic inference, revealing significant gaps between LLMs and human performance.

More recent benchmarks emphasize complex reasoning:

- **MMLU** (Hendrycks et al., 2021): 57-subject multiple-choice exams covering STEM, humanities, and social sciences. While comprehensive, multiple-choice format limits assessment of reasoning process.
- **BIG-Bench** (Srivastava et al., 2022): 204 diverse tasks including symbolic reasoning, but primarily short-form responses that don't require extended reasoning chains.
- **HELM** (Liang et al., 2022): Holistic evaluation across scenarios, but focuses more on fairness, robustness, and efficiency than deep reasoning.

Mathematics: MATH dataset (Hendrycks et al., 2021) provides competition-level math problems with detailed solutions. GSM8K (Cobbe et al., 2021) focuses on grade-school word problems. However, both emphasize mathematics exclusively, limiting insight into cross-domain reasoning transfer.

Code Reasoning: HumanEval (Chen et al., 2021) and MBPP (Austin et al., 2021) test programming ability. Recent work like SWE-bench (Jimenez et al., 2023) evaluates real-world software engineering tasks, showing that code generation reasoning differs from mathematical reasoning.

Scientific Reasoning: SciBench (Wang et al., 2023) covers college-level STEM but focuses on closed-form problems. Our work extends this by including open-ended problems and assessing step-by-step reasoning quality alongside final answers.

Traditional benchmarks evaluate only final outputs. Recent work emphasizes reasoning transparency:

- **Chain-of-Thought (CoT):** (Wei et al., 2022b) demonstrated that prompting models to show their work significantly improves accuracy. However, CoT prompts sometimes produce "hallucinated reasoning"—plausible-sounding but logically flawed steps.
- **Process Supervision:** (Lightman et al., 2023) trained verifiers to assess each reasoning step independently, showing better generalization than outcome supervision alone. Our step-accuracy metric aligns with this philosophy but uses semantic similarity rather than binary correctness labels.
- **Self-Consistency:** (Wang et al., 2022; de Curtò et al., 2024; García et al., 2025) enhance reasoning/problem reliability by sampling several reasoning/solution paths and choosing the majority vote answer. Although this approach mitigates randomness and improves overall robustness, it does not verify whether the underlying reasoning in each path/solution is correct.

Understanding whether reasoning skills transfer across domains is critical for AGI research. (Lu et al., 2022) investigated transfer from language to vision-language tasks, finding limited cross-modal transfer. (Prystawska and Goodman, 2023) examined whether CoT improvements in one domain predict improvements in others, finding weak correlations—consistent with our observation that model rankings vary significantly across domains.

XTREME (Hu et al., 2020) evaluates cross-lingual transfer, while our work primarily examines cross-disciplinary transfer within English. MTEB (Muennighoff et al., 2022) benchmarks embedding models across 58 tasks, but focuses on representation quality rather than reasoning.

A key distinction of our benchmark is its *semantic depth*—rather than testing many shallow tasks, we evaluate a smaller set of domains featuring problems that demand multi-step reasoning and domain expertise.

Open LLM Leaderboard (Hugging Face) and Chatbot Arena (LMSYS) provide community-driven rankings, but emphasize general chat quality or aggregate scores that obscure domain-specific strengths. Our contribution is *fine-grained analysis*—revealing that no single model dominates all reasoning types.

(Touvron et al., 2023) (LLaMA) demonstrated that careful data curation enables smaller models to compete with larger ones. (Team et al., 2024) (Gemma) extended this with additional safety training. Our baseline evaluation of these models provides empirical grounding for such claims in the reasoning domain specifically.

(Jiang et al., 2024) positioned their MoE architecture as balancing performance and efficiency.

Kahneman's (Kahneman, 2011) distinction between System 1 (fast, intuitive) and System 2 (slow, deliberate) reasoning offers a useful lens. Our finding that Qwen3 models exhibit low step-accuracy but high final-score suggests System-1-like behavior—arriving at answers through pattern recognition rather than explicit logical chains. Conversely, DeepSeek-R1's high step-accuracy implies System-2-like deliberation, though imperfectly executed.

Designing AI systems that appropriately invoke each mode—using fast heuristics when sufficient, escalating to careful reasoning when necessary—remains an open challenge.

Human experts often solve problems by analogy to previously solved similar problems (Hofstadter and Sander, 2013). Whether LLMs employ analogical reasoning or merely statistical pattern matching is debated.

(Bender et al., 2021) warned that LLMs are "stochastic parrots" that may confidently produce harmful or false content. The DeepSeek Paradox—detailed but incorrect reasoning—exemplifies this risk: confident-sounding explanations could mislead users more effectively than obviously wrong answers.

(Perez et al., 2022) documented "emergent deception"—models sometimes producing reasoning that superficially appears sound but contains subtle flaws. Our step-by-step evaluation aims to detect such issues, though human validation remains necessary.

Reinforcement Learning from Human Feedback (Ouyang et al., 2022) has become standard for aligning models with human values. (Bai et al., 2022) introduced Reinforcement Learning from AI Feedback (RLAIF) to scale preference collection. Extending RLHF to target *reasoning quality* specifically—beyond helpfulness and harmlessness—is a promising direction informed by benchmarks like ours.

3. Baseline Evaluation on MareNostrum 5

To establish a rigorous foundation for cross-platform comparison, we conducted an initial evaluation phase, de Curtò and de Zarzà (2024), on the MareNostrum 5 supercomputer at the BARCELONA Supercomputing Center. This baseline study assessed six state-of-the-art instruction-tuned models on a carefully curated 19-problem benchmark, establishing both the evaluation methodology and reference performance metrics for subsequent infrastructure validation and dataset expansion.

The baseline evaluation utilized nodes equipped with NVIDIA H100 GPUs on MareNostrum 5, with all models served using vLLM (Kwon et al., 2023). We selected six models representing diverse architectural approaches and parameter scales:

- Mistral AI: Mistral-7B-Instruct-v0.1, Mixtral-8x7B-Instruct-v0.1
- Meta: LLaMA 3.1-8B-Instruct
- Microsoft: Phi-3-small-8k-instruct
- Allen AI: OLMo-7B
- Google: Gemma-2-9b

Table 1

Baseline Performance Metrics (MareNostrum 5, 19 Problems)

Model	Overall Score	Step Accuracy	Consistency
Phi-3	0.623	0.648	0.040
Mixtral-8x7B	0.613	0.688	0.058
LLaMA 3.1-8B	0.593	0.521	0.087
Gemma-2-9b	0.519	0.700	0.093
Mistral-7B	0.357	0.427	0.042
OLMo-7B	0.334	0.514	0.062

Table 2

Baseline Domain-Specific Performance

Model	Phys	Math	Chem	Econ	Stat	Bio	Calc	Opt	Avg
Mixtral-8x7B	0.610	0.646	0.649	0.809	0.663	0.544	0.573	0.395	0.613
Phi-3	0.722	0.463	0.707	0.810	0.663	0.491	0.548	0.436	0.623
LLaMA 3.1-8B	0.661	0.586	0.606	0.767	0.514	0.511	0.615	0.411	0.593
Gemma-2-9b	0.560	0.445	0.489	0.625	0.436	0.575	0.533	0.461	0.519
Mistral-7B	0.374	0.360	0.249	0.438	0.242	0.333	0.586	0.314	0.357
OLMo-7B	0.339	0.355	0.379	0.246	0.357	0.326	0.354	0.284	0.334
Domain Avg	0.544	0.476	0.513	0.616	0.479	0.463	0.535	0.384	—

Inference parameters were standardized across all evaluations: temperature 0.2, max tokens 300, with three runs per problem to quantify consistency. The initial problem set comprised 19 problems spanning eight academic domains (Physics, Mathematics, Chemistry, Economics, Biology, Statistics, Calculus, Optimization) with difficulty stratification (Easy, Medium, Hard).

Table 1 presents the comprehensive performance metrics from this initial evaluation phase, establishing the reference point for all subsequent cross-platform and cross-model comparisons.

Key baseline findings:

1. **Performance hierarchy:** Phi-3 and Mixtral-8x7B demonstrated superior overall reasoning capabilities (0.623 and 0.613), establishing the performance ceiling for this initial model cohort.
2. **Step-accuracy leadership:** Gemma-2-9b achieved the highest step-accuracy (0.700), indicating exceptional transparency in intermediate reasoning steps despite moderate final-answer performance—an early indication of the transparency-correctness trade-off that becomes central to our extended analysis.
3. **Consistency patterns:** Phi-3 exhibited the most stable predictions across runs (0.040 std dev), while LLaMA 3.1-8B and Gemma-2-9b showed higher variance, suggesting fundamentally different approaches to stochastic reasoning.
4. **Architecture insights:** Mixtral-8x7B’s mixture-of-experts architecture achieved strong balanced performance, while smaller dense models (Phi-3, 3.8B parameters) demonstrated remarkable parameter efficiency.

Table 2 presents domain-level performance, revealing systematic strengths and weaknesses that informed our extended evaluation design.

Domain-level insights:

- **Economics dominance:** All top-tier models achieved exceptional performance in Economics (Phi-3: 0.810, Mixtral: 0.809), establishing this as the most tractable domain in the baseline evaluation—a pattern that shifts significantly in the extended 79-problem assessment (Section 7).
- **Optimization challenge:** Even the best-performing model (Gemma-2-9b) achieved only 0.461 in Optimization, with domain average of 0.384, identifying this as the most challenging reasoning category—a finding that persists across all subsequent evaluations on different infrastructures and problem sets.

- **Chemistry variance:** High inter-model variance (range 0.249–0.707) suggested that chemistry reasoning depends heavily on specific training corpus characteristics—a hypothesis we test through architectural diversity in later sections.
- **Calculus mid-tier positioning:** Calculus ranked mid-tier in this baseline (domain average 0.535), contrasting dramatically with its top-ranked status in the extended evaluation (+24.7% improvement across model generations), revealing systematic evolution in training methodologies.

Figure 1 presents performance degradation across difficulty levels, establishing a fundamental pattern that we observe consistently across all subsequent evaluations.

All baseline models exhibited clear monotonic performance decrease with increasing problem complexity. Phi-3 maintained the highest performance on easy problems (0.878), while Mixtral-8x7B demonstrated superior resilience on hard problems (0.474), suggesting different architectural approaches to complexity scaling—a pattern we explore in depth through the extended evaluation’s architectural diversity analysis (Sections 6 and 7).

This baseline evaluation established the core evaluation framework—including semantic similarity scoring, dual-metric assessment (final score and step-accuracy), consistency quantification via three-run protocols, and domain stratification—which is detailed comprehensively in Section 5. All subsequent experiments employ this identical methodology to enable direct cross-platform comparison.

The baseline study revealed three critical patterns that shaped the design of our extended cross-platform evaluation:

1. **Parameter efficiency hypothesis:** Phi-3 (3.8B) outperformed significantly larger models, suggesting that training data quality may matter more than scale. This motivated our inclusion of models up to 405B parameters (Section 7) to test the parameter efficiency paradox systematically.
2. **Transparency-correctness decoupling:** Gemma-2-9b’s high step-accuracy (0.700) with moderate overall score (0.519) indicated a fundamental tension between reasoning transparency and answer correctness. We explore this systematically through correlation analysis across 15 models in the extended evaluation.
3. **Domain difficulty stability:** Optimization emerged as universally challenging while Economics proved surprisingly tractable. Our longitudinal comparison (Section 8) reveals which patterns persist versus evolve across model generations and infrastructures.

Having established this methodological foundation and baseline performance reference, we proceed to validate these findings across alternative computational infrastructures (Section 6), expand the problem set to 79 items for improved statistical power (Section 7), and evaluate nine additional state-of-the-art models including larger-scale variants and specialized reasoning architectures.

4. Dataset Update and Structure

We extend the initial benchmark from 19 to 79 problems spanning eight domains with relatively balanced difficulty (25 easy, 36 medium, 18 hard). Each item preserves a uniform schema: (i) problem statement, (ii) final result, and (iii) step-by-step solution. The updated distribution per domain and difficulty is shown in Table 3.

To support transparency and repeatability, the dataset includes consistent stepwise solutions and difficulty tags per item; an example schema is shown in Fig. 2.

5. Methodology

The expanded dataset comprises 79 problems categorized into eight domains with balanced difficulty levels (25 easy, 36 medium, 18 hard); see Table 3). Each problem includes a reference solution and step decomposition. Models receive standardized prompts and are queried three times to assess run-to-run variability.

Responses are encoded using the *all-MiniLM-L6-v2* SentenceTransformer (Reimers and Gurevych, 2019) and compared with reference solutions via cosine similarity:

- **Final-score:** similarity between predicted and correct answers.
- **Step-accuracy:** mean similarity across intermediate reasoning steps.
- **Consistency:** standard deviation of repeated scores per problem.

Domain	Easy	Med.	Hard	Total
Physics	3	5	3	11
Mathematics	4	5	2	11
Chemistry	3	6	2	11
Economics	4	4	2	10
Statistics	3	4	2	9
Biology	3	4	2	9
Calculus	3	4	2	9
Optimization	2	4	3	9
Totals	25	36	18	79

Table 3

Updated dataset coverage across domains and difficulty levels.

Table 4

Comprehensive model inventory across experimental phases

Model Family	Variant	Parameters	Phase(s)
<i>Baseline Models (MareNostrum 5)</i>			
Mixtral	8x7B-Instruct-v0.1	46.7B (12.9B active)	1
Phi	Phi-3-small-8k-instruct	3.8B	1, 2
LLaMA	3.1-8B-Instruct	8B	1, 2
Gemma	2-9b	9B	1
Mistral	7B-Instruct-v0.1	7B	1, 2
OLMo	7B	7B	1
<i>Extended Models (Nebius AI Studio)</i>			
Hermes	4-70B	70B	2, 3
Hermes	4-405B	405B	2, 3
LLaMA	3.1-405B-Instruct	405B	2, 3
LLaMA	3.3-70B-Instruct	70B	2, 3
Qwen3	30B-A3B-Instruct	30B	2, 3
Qwen3	235B-A22B-Instruct	235B	2, 3
DeepSeek	R1-0528	70B	2, 3
GPT-OSS	20B	20B	2, 3
GPT-OSS	120B	120B	2, 3
<i>University Cluster Additional Models</i>			
Phi	4-mini-instruct	14B	2, 3
Phi	3.5-MoE-instruct	42B (6.6B active)	2, 3
Qwen	2-7B-Instruct	7B	2, 3
Falcon	Mamba-7b-instruct	7B	2, 3

This semantic approach rewards conceptual correctness and penalizes incomplete or incoherent reasoning chains.

Table 4 presents all 15 models evaluated across the three experimental phases, organized by infrastructure and evaluation scope.

Phases: 1 = Baseline (19 problems, MareNostrum 5), 2 = Infrastructure Validation (19 problems), 3 = Extended Evaluation (79 problems)

This hybrid deployment enables fair cross-infrastructure comparison and highlights the portability of the evaluation framework.

6. Infrastructure-Agnostic Validation (19-Problem Benchmark)

To validate the reproducibility and generalizability of our evaluation methodology beyond supercomputing facilities, we conducted first a comprehensive validation study on a university cluster infrastructure. We use models

Table 5

Overall Performance Metrics on University Cluster Infrastructure

Model	Overall Score	Step Accuracy	Consistency
Phi-4-mini	0.674	0.741	0.032
Phi-3-mini	0.616	0.648	0.079
Qwen2-7B	0.614	0.698	0.060
Falcon-Mamba-7B	0.590	0.676	0.029
LLaMA-3.1-8B	0.576	0.504	0.075
Phi-3.5-MoE	0.569	0.585	0.044
Mistral-7B-v0.1	0.381	0.447	0.057

Table 6

Domain-Specific Performance on University Cluster

Model	Phys	Math	Chem	Econ	Stat	Bio	Calc	Opt	Avg
Phi-4-mini	0.743	0.450	0.816	0.833	0.686	0.621	0.498	0.607	0.674
Phi-3-mini	0.809	0.512	0.703	0.730	0.546	0.497	0.502	0.396	0.616
Qwen2-7B	0.804	0.505	0.543	0.821	0.685	0.500	0.551	0.348	0.614
Falcon-Mamba-7B	0.633	0.584	0.467	0.792	0.656	0.646	0.507	0.451	0.590
LLaMA-3.1-8B	0.656	0.590	0.540	0.795	0.561	0.460	0.602	0.341	0.576
Phi-3.5-MoE	0.718	0.593	0.508	0.806	0.488	0.439	0.523	0.356	0.569
Mistral-7B-v0.1	0.371	0.385	0.365	0.438	0.298	0.309	0.538	0.360	0.381
Domain Avg	0.677	0.517	0.563	0.745	0.560	0.496	0.531	0.408	0.572
Std Dev	0.139	0.073	0.140	0.129	0.128	0.105	0.034	0.089	—

for longitudinal comparison and also newly introduced architectures, enabling analysis of infrastructure impact, model evolution, and emergent reasoning patterns.

Experiments in this section were executed on a university cluster node with 8× NVIDIA H200 GPUs (143,771 MiB per GPU), driver 570.124.06, CUDA 12.8, using vLLM with identical inference parameters as the baseline (temperature 0.2, max tokens 300, three runs per problem). See Table 4 for evaluated models.

Table 5 presents the aggregate performance metrics across the 19 problems set.

Key findings:

- Phi-4 dominance:** Phi-4-mini achieves the highest overall score (0.674) while maintaining exceptional consistency (0.032), representing a significant improvement over its predecessor Phi-3-mini (0.616).
- Architectural diversity:** Traditional transformers (Phi-4, Qwen2), mixture-of-experts (Phi-3.5-MoE), and state-space models (Falcon-Mamba) achieve competitive performance, with the non-transformer Falcon-Mamba matching LLaMA-3.1-8B (0.590 vs 0.576).
- Step-accuracy leadership:** Phi-4-mini exhibits the highest step accuracy (0.741), suggesting superior intermediate reasoning quality, followed by Qwen2-7B (0.698) and Falcon-Mamba-7B (0.676).
- Consistency champion:** Falcon-Mamba-7B demonstrates the best consistency (0.029), indicating highly stable predictions across problem variations—critical for production deployment.
- Mistral underperformance:** Mistral-7B-v0.1 significantly underperforms (0.381), suggesting that this early-generation model lacks the instruction-following refinement of more recent architectures.

Table 6 breaks down performance across the eight academic domains.

Domain patterns:

- Economics dominance:** Economics emerges as the strongest domain (mean 0.745), with Phi-4-mini achieving exceptional performance (0.833). This contrasts with the extended evaluation where Economics ranked mid-tier, suggesting the simpler problems in the 19-problem set favor economic reasoning.
- Physics strength:** Physics maintains strong performance (mean 0.677), with Phi-3-mini and Qwen2-7B exceeding 0.80, consistent with findings in the extended evaluation.

Table 7

Performance by Difficulty Level

Model	Easy	Medium	Hard
Phi-4-mini	0.903	0.707	0.505
Qwen2-7B	0.892	0.653	0.410
Falcon-Mamba-7B	0.875	0.617	0.402
Phi-3-mini	0.790	0.690	0.407
LLaMA-3.1-8B	0.787	0.625	0.388
Phi-3.5-MoE	0.763	0.652	0.332
Mistral-7B-v0.1	0.514	0.388	0.302
Average	0.789	0.619	0.392
Std Dev	0.120	0.100	0.063

Table 8

Infrastructure Comparison: MareNostrum 5 vs University Cluster

Model	MareNostrum 5 (2024)	Univ. Cluster (2025)	Δ
LLaMA-3.1-8B	0.593	0.576	-0.017
Phi-3-mini	0.623	0.616	-0.007

- **Mathematics as discriminator:** Mathematics shows the tightest inter-model clustering (std dev 0.073), suggesting this domain provides consistent difficulty across architectures—ideal for standardized comparison.
- **Optimization challenge persists:** Optimization remains the most difficult domain (mean 0.408), with only Phi-4-mini exceeding 0.60 (0.607), replicating the pattern from both baseline and extended evaluations.
- **Calculus consistency:** Calculus exhibits the lowest variance across models (std dev 0.034), indicating that performance on calculus problems is highly model-independent—surprising given its symbolic complexity.
- **Chemistry divergence:** Chemistry shows high variance (std dev 0.140), with Phi-4-mini excelling (0.816) while Mistral-7B and Falcon-Mamba struggle (0.365 and 0.467). This suggests chemistry reasoning may require specific training corpus characteristics.

Table 7 analyzes performance across problem difficulty levels (Easy, Medium, Hard); see also Figure 3 for extended analysis .

Observations:

1. **Steep difficulty gradient:** All models exhibit performance degradation from Easy (mean 0.789) to Medium (0.619) to Hard (0.392), with approximately 20% drop per difficulty tier.
2. **Phi-4's hard problem advantage:** Phi-4-mini maintains the highest hard problem performance (0.505), exceeding the second-best model by +9.5%, suggesting superior complex reasoning capabilities.
3. **Convergence on hard problems:** Standard deviation decreases with difficulty (Easy: 0.120, Hard: 0.063), indicating that hard problems consistently challenge all architectures—useful for model differentiation.
4. **Mistral's easy problem struggle:** Mistral-7B-v0.1 achieves only 0.514 on easy problems (vs. 0.903 for Phi-4), indicating fundamental instruction-following deficiencies rather than just reasoning limitations.

Table 13 directly compares performance of identical models across MareNostrum 5 (baseline 2024) and the university cluster (current study), controlling for all variables except infrastructure.

Critical findings:

1. **Infrastructure-agnostic reasoning:** Both models show minimal performance degradation on university cluster infrastructure (LLaMA: -2.9%, Phi-3: -1.1%), well within typical measurement variance. This validates our hypothesis that reasoning quality is *model-intrinsic* rather than infrastructure-dependent.

6.1. Architectural Insights

The inclusion of Falcon-Mamba-7B (based on Mamba architecture, a state-space model) alongside traditional transformers enables architectural comparison:

- **Competitive overall performance:** Falcon-Mamba achieves 0.590, matching LLaMA-3.1-8B (0.576) despite fundamentally different attention mechanisms.
- **Superior consistency:** Falcon-Mamba exhibits the best consistency score (0.029), suggesting state-space models may produce more stable predictions—potentially valuable for safety-critical applications.
- **Domain specialization:** Falcon-Mamba excels in Biology (0.646, highest among all models) and Mathematics (0.584, second highest), but struggles in Chemistry (0.467) and Optimization (0.451).
- **Step-accuracy advantage:** Despite lower overall scores, Falcon-Mamba achieves strong step accuracy (0.676), indicating that state-space models may generate coherent reasoning chains even when final answers diverge.

Implication: State-space models represent a viable alternative to transformers for reasoning tasks, particularly where consistency and explainability matter more than peak accuracy.

Phi-3.5-MoE (42B total parameters, 6.6B active per token) provides insight into MoE scaling:

- **Efficiency paradox:** Despite 42B total parameters, Phi-3.5-MoE (0.569) underperforms dense Phi-3-mini-4k (3.8B, score 0.616), suggesting MoE advantages may not translate to reasoning tasks.
- **Domain inconsistency:** Phi-3.5-MoE shows high variance across domains (Physics 0.718 vs. Biology 0.439), indicating that expert routing may specialize unevenly.
- **Mathematical reasoning strength:** Phi-3.5-MoE achieves the highest Mathematics score (0.593) among evaluated models, suggesting certain experts may specialize in symbolic reasoning.

Implication: For reasoning tasks, densely-trained smaller models may outperform sparsely-activated larger MoE architectures, challenging assumptions about MoE efficiency in non-language modeling domains.

The availability of Phi-3-mini, Phi-3.5-MoE, and Phi-4-mini enables within-family longitudinal analysis:

- **Clear progression:** Phi-4-mini (0.674) > Phi-3-mini (0.616) > Phi-3.5-MoE (0.569), with the dense architecture evolution showing +9.4% improvement.
- **Step accuracy improvement:** Phi-4 achieves 0.741 step accuracy vs. 0.648 for Phi-3, indicating Microsoft's training improved intermediate reasoning quality.
- **Consistency gains:** Phi-4's consistency (0.032) dramatically improves over Phi-3 (0.079), suggesting more robust training data or RLHF refinement.
- **Chemistry breakthrough:** Phi-4 excels in Chemistry (0.816), a domain where Phi-3 was middling (0.703), indicating targeted improvements in scientific reasoning.

Implication: Within the Phi family, dense scaling with improved training data yields superior results compared to sparse MoE expansion—consistent with broader findings about training quality vs. parameter count.

Table 9

Nebius API evaluation on the 19-problem set: summary by model. Average Final Score and Step-Accuracy are cosine-similarity based. Error bars in Fig. 4 reflect the mean of per-evaluation standard deviations reported in the logs.

Model	#Problems	#Evals	Avg Score	Avg Std	Avg Step-Acc.
NousResearch/Hermes-4-70B	19	19	0.667	0.049	0.595
NousResearch/Hermes-4-405B	19	19	0.618	0.034	0.649
meta-llama/Meta-Llama-3.1-405B-Instruct	19	19	0.618	0.051	0.547
meta-llama/Llama-3.3-70B-Instruct	19	19	0.574	0.042	0.556
Qwen/Qwen3-235B-A22B-Instruct-2507	19	19	0.529	0.011	0.553
Qwen/Qwen3-30B-A3B-Instruct-2507	19	19	0.514	0.017	0.543
deepseek-ai/DeepSeek-R1-0528	19	19	0.470	0.034	0.720
openai/gpt-oss-20b	19	19	0.445	0.075	0.739
openai/gpt-oss-120b	19	19	0.441	0.037	0.740

6.2. Nebius API Evaluation on 19 Problems

Additionally, we evaluated nine models via the Nebius API AI studio on the original 19-problem set used in the MareNostrum 5 baseline and validation study in the preceding section. Figures 4–7 summarize overall and process-level performance, as well as domain-specific trends.

Table 9 reports the corresponding numerical values. A one-way ANOVA across models indicates significant performance differences ($F = 2.12$, $p = 0.0369$). A Welch two-sample t-test between the top two models (NousResearch/Hermes-4-70B vs. NousResearch/Hermes-4-405B) yields $p = 0.514$ (no significant difference).

6.3. Synthesis vs MareNostrum Baseline (19 Problems)

This subsection integrates the initial MareNostrum 5 baseline (Mixtral-8x7B, Phi-3, LLaMA 3.1-8B, Gemma-2-9b, Mistral-7B, OLMO-7B) with the Nebius API validation on the same 19-problem set (nine newer models), along with the university-cluster reproducibility study.

(1) *Cross-cohort ordering and absolute levels.* On MareNostrum 5, the best overall model was Phi-3 (0.623), closely followed by Mixtral-8x7B (0.613), with LLaMA 3.1-8B at 0.593 and Gemma-2-9b at 0.519. On Nebius (19 problems), Hermes-4-70B leads with 0.667, while Hermes-4-405B and LLaMA 3.1-405B are tied at 0.618, LLaMA 3.3-70B at 0.574, and Qwen3 models at 0.529/0.514. Overall, state-of-the-art 2024/25 models on Nebius match or surpass the best 2024 baseline levels on MareNostrum.

(2) *Infrastructure reproducibility holds.* The university-cluster replication of identical 2024 models shows minimal deltas relative to MareNostrum (LLaMA 3.1-8B: -0.017, Phi-3: -0.007; Table 13), buttressing that reasoning quality is deployment-invariant within normal variance. This strengthens the interpretation that the Nebius 19-problem gains are driven by *newer models*, not the serving platform.

(3) *Parameter-efficiency paradox persists.* On Nebius (19 problems) Hermes-4-70B (0.667) outperforms its larger sibling Hermes-4-405B (0.618), and is statistically indistinguishable from it by Welch *t*-test ($p \approx 0.514$). A one-way ANOVA across all nine Nebius models detects significant differences ($F=2.12$, $p=0.0369$), consistent with heterogeneity across families and confirming that *bigger is not always better*—a pattern already suggested by the original baseline.

(4) *Transparency–correctness decoupling.* Nebius results reveal models with *high process transparency but moderate final scores*: DeepSeek-R1 (step-accuracy 0.720; average score 0.470) and GPT-OSS 20B/120B (step-accuracy 0.739/0.740; average score 0.445/0.441). This extends the baseline finding that step accuracy (e.g., Gemma-2-9b at 0.700 on MareNostrum) can diverge from final correctness. Practically, this recommends model selection by use case: high step-fidelity for pedagogy and audits, high final-score for production outcomes.

(5) *Domain stability and hard cases.* Across cohorts, Economics remains the easiest domain and Optimization the most challenging. The cluster analysis confirms Economics’ high mean (0.745) and Optimization’s low mean

Table 10

Extended University-Cluster Evaluation on 79 Problems: Summary by Model. Average Final Score and Step-Accuracy are cosine-similarity based. Error bars in Fig. 8 reflect the mean of per-evaluation standard deviations reported in the JSON logs.

Model	#Problems	#Evals	Avg Score	Avg Std	Avg Step-Acc.
microsoft/Phi-3-mini-4k-instruct	79	79	0.565	0.057	0.629
microsoft/Phi-4-mini-instruct	79	79	0.560	0.063	0.716
meta-llama/Meta-Llama-3.1-8B-Instruct	79	79	0.551	0.056	0.481
microsoft/Phi-3.5-MoE-instruct	79	79	0.529	0.039	0.516
Qwen/Qwen2-7B-Instruct	79	79	0.514	0.037	0.667
tiiuae/falcon-mamba-7b-instruct	79	79	0.499	0.057	0.586
mistralai/Mistral-7B-Instruct-v0.1	79	79	0.388	0.046	0.418

(0.408), mirroring both the MareNostrum baseline and Nebius 19-problem trends. This robustness indicates that domain difficulty ordering is intrinsic to current LLM training distributions and inductive biases.

(6) *Architectural notes.* The cluster study shows Falcon–Mamba–7B (state-space) competitively close to LLaMA 3.1–8B overall (0.590 vs 0.576) while winning on consistency (0.029). MoE does not automatically deliver reasoning gains: Phi-3.5-MoE (0.569) trails dense Phi-3 (0.616) on the same 19 problems. The Nebius cohort reinforces this: dense, well-trained 70B-class models (*Hermes-4-70B*) strike a favorable accuracy–efficiency trade-off against their 400B-class counterparts.

(7) *Bottom line before scaling to 79 problems.* (i) Newer families (*Hermes-4*, LLaMA 3.1–405B, Qwen3) on Nebius improve on or match the MareNostrum 5 baseline bests on the same 19 problems. (ii) Cross-infrastructure evidence confirms the results are model-intrinsic rather than platform-induced. (iii) The persistent transparency–correctness decoupling and the stability of domain difficulty motivate *dual-reporting* of final-score and step-accuracy in subsequent sections. These conclusions justify proceeding to the *larger, harder* 79-problem benchmark to probe generalization patterns, rank stability, and domain shifts at scale.

7. Extended Evaluation at Scale (79-Problem Benchmark)

Following the preliminary validation in Section 6, we executed the full 79-problem benchmark on the university cluster for the seven representative models spanning dense transformers, MoE, and a state-space (Mamba) architecture. Figure 8 reports the overall average final score per model with error bars (mean per-evaluation standard deviation). Figure 9 shows process-level reasoning via average step-accuracy, and Figure 10 relates transparency (step-accuracy) to final correctness.

Table 10 summarizes these results. In brief: (i) models differ in the balance between final correctness and process transparency; (ii) compact dense models can rival larger MoE/state-space systems in overall score; and (iii) run-to-run variability (as captured by reported standard deviations) is modest for the strongest models, indicating stable behavior.

Beyond aggregate performance, Figure 11 summarizes average final scores per academic domain, revealing domain-specific strengths and weaknesses. Economics and Calculus again emerge as the highest-performing domains, while Optimization and Chemistry remain the most challenging. Figure 12 reports performance by difficulty tier (Easy, Medium, Hard), confirming a monotonic degradation pattern ($p < 0.001$ via one-way ANOVA).

A one-way ANOVA across all seven models yields $F = 5.49$, $p = 1.56e - 05$, indicating statistically significant performance differences. A Welch two-sample t-test between the two best-performing models (microsoft/Phi-3-mini-4k-instruct vs. microsoft/Phi-4-mini-instruct) also confirms the superiority of the top model ($p = 0.883$).

7.1. State-of-the-Art Models on Nebius Platform

The following evaluations, executed on Nebius AI Studio, extend the scope of the initial baseline and validation study .

Table 11 presents comprehensive performance metrics across all nine newly evaluated models on the 79-problem benchmark. *Hermes-4-70B* achieved the highest overall accuracy (0.598), surpassing even its 405B variant (0.573),

Model	Coverage (%)	Avg Score	Avg Step	Mean Std	Problems
Hermes-4-70B	100.0	0.598	0.548	0.032	79
Hermes-4-405B	100.0	0.573	0.605	0.032	79
Meta-Llama-3.1-405B-Instruct	100.0	0.569	0.520	0.038	79
Llama-3.3-70B-Instruct	100.0	0.561	0.498	0.037	79
Qwen3-235B-A22B-Instruct-2507	100.0	0.487	0.488	0.013	79
Qwen3-30B-A3B-Instruct-2507	100.0	0.477	0.513	0.017	79
DeepSeek-R1-0528	100.0	0.457	0.716	0.044	79
GPT-OSS-20B	72.2	0.406	0.682	0.047	57
GPT-OSS-120B	83.5	0.404	0.667	0.037	66

Table 11

Overall performance across 79 problems including newly evaluated Llama and Qwen3 models. Coverage reflects the fraction of problems successfully evaluated.

demonstrating the parameter efficiency paradox discussed in the abstract. Notably, all models achieved 100% coverage except GPT-OSS variants, which successfully evaluated 72.2% (GPT-OSS-20B) and 83.5% (GPT-OSS-120B) of problems.

Qwen3-235B and Qwen3-30B achieved stable mid-tier accuracy (0.487 and 0.477 respectively) with remarkably low variance (0.013 and 0.017), representing the most consistent models in our evaluation as shown in Figure 16. DeepSeek-R1 delivered the highest step-accuracy (0.716) but moderate final-score (0.457), reflecting its transparency-over-precision behavior—a pattern clearly visible in Figure 15. The overall ranking across models is visualized in Figure 20.

Figure 13 illustrates performance across eight academic domains, revealing systematic domain-specific strengths and weaknesses. Across all models, Calculus consistently yielded the highest reasoning similarity ($\approx 0.65\text{--}0.69$), followed by Economics. This pattern is further detailed in the heatmap visualization (Figure 17), which shows that Hermes-4-70B and Meta-Llama-3.1-405B-Instruct achieve the strongest cross-domain balance.

Optimization, Chemistry, and high-dimensional Statistics tasks remained the most challenging domains across all models, with average scores significantly lower than other domains.

Figure 14 demonstrates how model performance degrades with increasing problem complexity. The Hermes-4 and Meta-Llama families retain relatively higher performance on hard problems, while DeepSeek-R1 and GPT-OSS models show stronger performance on easy-case problems. This pattern is further detailed in Figure 18, which visualizes the performance degradation matrix across all models and difficulty levels.

Most models show a clear monotonic decrease in accuracy from Easy to Hard problems, though the magnitude of degradation varies significantly by architecture. Dense models (Hermes-4, Llama variants) exhibit more graceful degradation compared to smaller or more specialized models.

Figure 15 reveals a striking pattern: DeepSeek-R1 achieves the highest step-accuracy (0.716), substantially exceeding all other models, yet produces only moderate final scores (0.457, Table 11). This transparency-correctness trade-off is further explored in Figure 21, which plots the relationship between step-accuracy and final accuracy across 3,000 individual problem instances.

The scatter plot reveals that different models exhibit vastly different correlations between reasoning process quality and answer correctness. Qwen3 models show near-zero correlation, suggesting "shortcut learning," while GPT-OSS models demonstrate moderate coupling. This fundamental architectural difference has important implications for deployment: educational applications may prioritize DeepSeek-R1's transparent reasoning, while production systems requiring consistent correct answers may favor Qwen3 or Hermes-4 models.

Run-to-run consistency varied significantly across models, as shown in Figures 16 and 19. The Qwen3 models achieved exceptional consistency (0.013 and 0.017 standard deviation), representing 3× better stability than the next-best alternatives. Conversely, GPT-OSS-20B (0.047) and DeepSeek-R1 (0.044) exhibited higher variability despite strong performance in other metrics.

Figure 19 presents the distribution of per-problem score standard deviations across repeated runs, revealing that DeepSeek-R1's high variability persists across problem types. This suggests that its explicit reasoning process introduces inherent stochasticity, creating a trade-off between transparency and reliability that merits careful consideration in production deployments.

Problem	Domain	Diff.	Std Dev
Ball thrown upward with $v=20$ m/s, find max height	Physics	Med	0.335
Inflection points of $f(x) = x^4 - 4x^3 + 6x^2$	Calculus	Hard	0.331
Car acceleration from 0 to 30 m/s over 100m	Physics	Med	0.304
Expected heterozygous offspring ($Aa \times Aa$)	Biology	Med	0.291
Critical points of $f(x) = x^3 - 3x$	Calculus	Med	0.274

Table 12

Top 5 problems with highest inter-model score variance

Table 13

Best-performing models per infrastructure, dataset size, and evaluation metric. The table consolidates the highest-scoring models from the MareNostrum 5 baseline (2024), the University Cluster validation (2025), and the Nebius AI Studio experiments (2025). Values correspond to mean cosine-similarity scores for final answers and average step accuracy.

Infrastructure	Year	Problems	Best Model	Avg Score	Step Acc.
<i>19-Problem Baseline Set</i>					
MareNostrum 5 (HPC)	2024	19	Phi-3-mini-4k-instruct	0.623	0.648
University Cluster	2025	19	Phi-4-mini-instruct	0.674	0.741
Nebius AI Studio (Cloud)	2025	19	Hermes-4-70B	0.667	0.595
<i>79-Problem Extended Set</i>					
University Cluster	2025	79	Phi-3-mini-4k-instruct	0.565	0.629
Nebius AI Studio (Cloud)	2025	79	Hermes-4-70B	0.598	0.548
<i>Step-Accuracy Leaders (Transparency-Oriented Models)</i>					
MareNostrum 5 (HPC)	2024	19	Gemma-2-9b	0.519	0.700
University Cluster	2025	19	Phi-4-mini-instruct	0.674	0.741
Nebius AI Studio (Cloud)	2025	19	gpt-oss-120b	0.441	0.740
University Cluster	2025	79	Phi-4-mini-instruct	0.560	0.716
Nebius AI Studio (Cloud)	2025	79	DeepSeek-R1-0528	0.457	0.716

Table 14

Infrastructure-agnostic validation: identical models on same problem set

Model	MareNostrum 5	Univ. Cluster	Δ (%)
LLaMA-3.1-8B	0.593	0.576	-0.017 (-2.9%)
Phi-3-mini	0.623	0.616	-0.007 (-1.1%)
<i>Mean absolute variance: 1.2%</i>			

Performance consistency across infrastructures—as evidenced by similar relative rankings in Table 11—confirms that reasoning quality is model-dependent rather than environment-dependent, validating the portability of our evaluation framework.

Table 12 presents the five problems with highest score variance across seven models.

Key Observation: Physics kinematics problems showed the highest disagreement (std dev 0.304-0.335), suggesting that while equations are well-known, their application in multi-step reasoning contexts varies significantly across model architectures. This contrasts with the baseline findings where Physics showed relatively consistent performance.

8. Comparative Analysis: Cross-Infrastructure Validation

To assess both progress in LLM reasoning capabilities and infrastructure-independence of evaluation results, we compared performance across the three computational paradigms listed in Table 4.

Table 14 directly compares identical models across MareNostrum 5 and University Cluster on the same 19-problem and 79-problem set, isolating infrastructure effects.

Key Finding: Performance variance across HPC (MareNostrum) and university cluster infrastructure remains within 3%, confirming that *reasoning quality is model-intrinsic rather than deployment-dependent*. This validates the reproducibility of evaluation results across diverse computational environments, democratizing rigorous benchmarking beyond supercomputing facilities.

Within the 19-problem baseline set, we observe clear generational improvements:

- **Phi family progression:** Phi-4-mini (0.674) represents an 8.2% improvement over Phi-3 (0.623) on the baseline set, with dramatic gains in step-accuracy (0.741 vs 0.648, +14.4%). However, on the extended 79-problem benchmark, Phi-4-mini exhibits slightly lower overall accuracy (0.560) compared to Phi-3-mini (0.565), while maintaining substantially superior step-accuracy (0.716 vs 0.629), highlighting the transparency-correctness trade-off on more challenging problems.
- **Step-accuracy advancement:** The highest step-accuracy increased from 0.700 (Gemma-2-9b, MareNostrum) to 0.741 (Phi-4-mini, Cluster), suggesting continued progress in intermediate reasoning quality.
- **Architecture diversification:** The University Cluster evaluation introduced state-space models (Falcon-Mamba-7B, 0.590) and mixture-of-experts architectures (Phi-3.5-MoE, 0.569), demonstrating that non-transformer architectures achieve competitive reasoning performance.

While direct score comparison across problem sets is inappropriate, we can analyze relative patterns:

- **Baseline set (19 problems):** Best performance 0.674 (Phi-4-mini), mean across top-3 models: 0.635
- **Extended set (79 problems):** Best performance 0.598 (Hermes-4-70B), mean across top-3 models: 0.586

The 8-10% score reduction in the extended set reflects increased problem difficulty rather than infrastructure or model regression, as the 79-problem benchmark was designed to include more challenging interdisciplinary problems and edge cases.

Across all three infrastructures, we observe a consistent pattern:

- **Traditional models:** Optimize for final answer correctness, accepting moderate step-accuracy (Hermes-4-70B: 0.598 overall, 0.548 step-accuracy)
- **Reasoning-focused models:** Prioritize transparent reasoning chains, sometimes at the expense of final accuracy (DeepSeek-R1: 0.457 overall, 0.716 step-accuracy)
- **Balanced models:** Achieve strong performance in both metrics on baseline problems (Phi-4-mini: 0.674 overall, 0.741 step-accuracy on 19 problems), though this balance may shift toward reasoning transparency on harder problems (0.560 overall, 0.716 step-accuracy on 79 problems)

This *transparency-correctness trade-off* persists across infrastructures, confirming it as an intrinsic model property rather than an evaluation artifact.

The three-infrastructure comparison establishes:

1. **Reproducibility:** <3% variance confirms reasoning evaluation is infrastructure-agnostic
2. **Progress:** Phi-4-mini (2025) shows clear improvement over Phi-3 (2024) in step reasoning quality, with mixed results on final accuracy depending on problem difficulty
3. **Specialization:** Models increasingly differentiate between correctness-optimized (Hermes, Qwen) and transparency-optimized (DeepSeek) approaches
4. **Architectural diversity:** Non-transformer models (state-space, MoE) demonstrate competitive reasoning capabilities

To facilitate exploration of our evaluation results and promote reproducibility, we developed an interactive web-based visualization tool using Streamlit¹, publicly accessible for community analysis (Figure 22). The application enables dynamic comparison across evaluated models, supporting filtering by problem set (19 vs. 79 problems), visualization of overall scores and step-accuracy metrics, difficulty-stratified radar charts, category-specific heatmaps, and reasoning step distributions.

¹<https://crossplatform-llm-decurto.streamlit.app/>

Our finding of infrastructure-agnostic reproducibility (<3% variance across MareNostrum 5, Nebius AI Studio, and university cluster) specifically applies to *hardware infrastructure* with consistent serving configurations. This finding should not be conflated with implementation-agnostic reproducibility, which we explicitly do *not* claim.

All evaluations in this study employed identical model weights (FP16 or BF16 precision), the same serving framework (vLLM 0.5.0+), and consistent inference parameters (temperature 0.2, max tokens 300). Under these controlled conditions, performance remains stable across diverse hardware platforms—validating that reasoning quality is model-intrinsic rather than hardware-dependent.

However, performance degradation under different vLLM versions and quantization settings (Artificial Analysis, 2025) confirms that software configurations can significantly affect evaluation results.

9. Conclusion

This work presents a comprehensive cross-platform evaluation of reasoning capabilities in foundation models through three complementary experimental studies, establishing an infrastructure-agnostic benchmark validated across HPC supercomputing (MareNostrum 5), cloud platforms (Nebius AI Studio), and university clusters. We expanded from a baseline of 6 models on 19 problems to 15 models across 79 problems spanning eight academic domains, comprising: (1) nine state-of-the-art models on Nebius cloud infrastructure, (2) seven models including non-transformer architectures on a university cluster (19-problem validation set) and (3) an extended 79-problem evaluation probing generalization at scale.

Infrastructure-Agnostic Reproducibility. Cross-platform validation establishes that reasoning quality is model-intrinsic rather than infrastructure-dependent. Identical models evaluated on MareNostrum 5 (2024) and university cluster (2025) show minimal variance: LLaMA-3.1-8B (-2.9%) and Phi-3-mini (-1.1%). This critical finding democratizes rigorous reasoning evaluation, enabling researchers without supercomputing access to conduct scientifically valid assessments on accessible infrastructure.

Parameter Efficiency Paradox. Results fundamentally challenge scaling assumptions: Hermes-4-70B (70B parameters) achieves the highest score among extended models (0.598), outperforming both its 405B counterpart (0.573) and Meta’s LLaMA 3.1-405B (0.560). LLaMA 3.3-70B regresses despite 8.75 \times more parameters than LLaMA 3.1-8B (0.561 vs 0.593). Dense Phi-4-mini (14B, 0.674) dramatically outperforms sparse Phi-3.5-MoE (42B, 0.569). These findings establish training data quality and architectural design as more critical than model size.

Transparency-Correctness Trade-Off. Analysis across several thousand problem instances reveals fundamental tension: DeepSeek-R1 achieves record step-accuracy (0.716) but moderate final scores (0.457, $r=0.249$ correlation), while Qwen3 models exhibit near-zero correlation ($r=0.095$), suggesting “shortcut learning.” This dichotomy has critical deployment implications: educational/safety-critical applications should favor DeepSeek-R1’s transparent reasoning, while production systems requiring consistency should select Qwen3-235B (0.013 score variance, 3 \times better than alternatives).

Domain-Specific Evolution. Longitudinal comparison (2024 vs 2025) reveals systematic patterns: Calculus improved dramatically (+24.7%), with LLaMA 3.1-405B achieving record performance (0.717), while Optimization remains universally challenging (+4.7% only, mean 0.408). Physics kinematics shows highest cross-model disagreement (std dev 0.335). Domain difficulty ordering proves robust across infrastructures, indicating fundamental limitations in current training distributions.

Architectural Diversity. University cluster validation establishes that non-transformer architectures achieve competitive performance: Falcon-Mamba-7B (state-space model) matches transformer baseline LLaMA-3.1-8B (0.590 vs 0.576) with superior consistency (0.029 vs 0.075 std dev), suggesting alternative architectures merit broader exploration for production deployments prioritizing stability.

Our findings establish three archetypal model profiles: Traditional models (Hermes-4-70B, Qwen3-235B) optimize for final correctness—recommended for production systems; Reasoning-focused models (DeepSeek-R1, GPT-OSS) prioritize transparent reasoning—recommended for educational and safety-critical applications; Balanced models (Phi-4-mini) achieve strong performance in both metrics—recommended for general-purpose reasoning tasks.

This work establishes: (1) infrastructure-agnostic evaluation framework validated with <3% variance across platforms, (2) dual-metric assessment capturing both process quality and outcome correctness, (3) cross-domain generalization benchmark with difficulty stratification, and (4) longitudinal tracking capability for systematic comparison across model generations.

Future work should integrate multi-modal reasoning (diagrams, code execution), expand to interdisciplinary synthesis problems, investigate hybrid architectures combining LLMs with symbolic solvers, and establish standardized re-evaluation cadence to track capability evolution. Human expert calibration would provide interpretable capability benchmarks.

Key Takeaways

Our results indicate that reasoning improvements in large language models no longer scale monotonically with parameter count. The performance plateau observed beyond approximately 70B parameters –together with the superior efficiency of Hermes-4-70B over its 405B variant – suggests that reasoning capability has entered a data-limited rather than parameter-limited regime. Future progress may therefore depend more on reasoning-centric data and supervision signals than on raw scale expansion. The contrast between DeepSeek-R1’s transparent but fallible reasoning and Qwen3’s accurate yet opaque answers parallels the dual-process theory of human cognition (System 2 vs. System 1). This structural duality highlights a fundamental design challenge for foundation models: balancing deliberate reasoning transparency with heuristic efficiency. Embedding both modes adaptively could lead to the next generation of explainable reasoning systems.

Data and Code Availability

The complete dataset of reasoning problems, evaluation framework, and analysis scripts are publicly available at <https://github.com/pablogarciaamolina/crossplatform-llm>. This release enables full reproducibility of our findings and supports extension to additional models and domains. An interactive visualization tool for exploring evaluation results across models and problem sets, is available at <https://crossplatform-llm-decurto.streamlit.app/>.

Acknowledgments

The authors thank the Universidad Pontificia Comillas (ICAI), for access to the university cluster infrastructure and the BARCELONA Supercomputing Center (BSC) for access to the MareNostrum 5 supercomputer, as well as to the LUXEMBOURG Institute of Science and Technology (LIST) for providing cloud resources through Nebius AI Studio during this extended experimentation. This work received support from the TIFON project at BSC and ADIALab-MAST project at LIST.

References

- Abdin, M., Jacobs, S.A., Awan, A.A., Aneja, J., Awadallah, A., Awadalla, H., Bach, N., Bahree, A., Bakhtiari, A., Behl, H., et al., 2024. Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219 .
- Artificial Analysis, 2025. GPT-OSS-120B Cross-Provider Performance Benchmark: AIME 2025 Evaluation. <https://artificialanalysis.ai/models/gpt-oss-120b/providers>. Accessed: October 2025.
- Austin, J., Odena, A., Nye, M., Bosma, M., Michalewski, H., Dohan, D., Jiang, E., Cai, C., Terry, M., Le, Q., et al., 2021. Program synthesis with large language models. arXiv preprint arXiv:2108.07732 .
- Bai, Y., Jones, A., Ndousse, K., Askell, A., Chen, A., DasSarma, N., Drain, D., Fort, S., Ganguli, D., Henighan, T., et al., 2022. Training a helpful and harmless assistant with reinforcement learning from human feedback. arXiv preprint arXiv:2204.05862 .
- Bender, E.M., Gebru, T., McMillan-Major, A., Shmitchell, S., 2021. On the dangers of stochastic parrots: Can language models be too big?, in: Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, pp. 610–623.
- Brown, T., Mann, B., Ryder, N., Subbiah, M., Kaplan, J.D., Dhariwal, P., Neelakantan, A., Shyam, P., Sastry, G., Askell, A., et al., 2020. Language models are few-shot learners. Advances in neural information processing systems 33, 1877–1901.
- Buscemi, A., Lothriltz, C., Morales, S., Gomez-Vazquez, M., Clarisó, R., Cabot, J., Castignani, G., 2025. Mind the language gap: Automated and augmented evaluation of bias in llms for high-and low-resource languages. arXiv preprint arXiv:2504.18560 .
- Chen, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.d.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G., et al., 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374 .
- Chowdhery, A., Narang, S., Devlin, J., Bosma, M., Mishra, G., Roberts, A., Barham, P., Chung, H.W., Sutton, C., Gehrmann, S., et al., 2022. Palm: Scaling language modeling with pathways. arXiv preprint arXiv:2204.02311 .
- Clark, P., Cowhey, I., Etzioni, O., Khot, T., Sabharwal, A., Schoenick, C., Tafjord, O., 2018. Think you have solved question answering? try arc, the ai2 reasoning challenge, in: Proceedings of the AAAI Conference on Artificial Intelligence.
- Cobbe, K., Kosaraju, V., Bavarian, M., Chen, M., Jun, H., Kaiser, L., Plappert, M., Tworek, J., Hilton, J., Nakano, R., et al., 2021. Training verifiers to solve math word problems. arXiv preprint arXiv:2110.14168 .
- de Curtò, J., de Zarzà, I., Calafate, C.T., 2024. Llm multi-agent decision optimization, in: KES International Symposium on Agent and Multi-Agent Systems: Technologies and Applications, Springer. pp. 3–15.

- de Curtò, J., de Zarzà, I., 2024. Comparative analysis of reasoning capabilities in foundation models, in: 2024 2nd International Conference on Foundation and Large Language Models (FLLM), pp. 141–149. doi:10.1109/FLLM63129.2024.10852449.
- Devlin, J., Chang, M.W., Lee, K., Toutanova, K., 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 .
- Fedus, W., Zoph, B., Shazeer, N., 2022. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research* 23, 1–39.
- García, P., de Curtò, J., de Zarzà, I., Cano, J.C., Calafate, C.T., 2025. Foundation models for cybersecurity: A comprehensive multi-modal evaluation of tabpfm and tabicf for tabular intrusion detection. *Electronics* 14. URL: <https://www.mdpi.com/2079-9292/14/19/3792>, doi:10.3390/electronics14193792.
- Gu, A., Dao, T., 2023. Mamba: Linear-time sequence modeling with selective state spaces. arXiv preprint arXiv:2312.00752 .
- Hendrycks, D., Burns, C., Basart, S., Zou, A., Mazeika, M., Song, D., Steinhardt, J., 2021. Measuring mathematical problem solving with the math dataset. arXiv preprint arXiv:2103.03874 .
- Hoffmann, J., Borgeaud, S., Mensch, A., Buchatskaya, E., Cai, T., Rutherford, E., Casas, D.d.L., Hendricks, L.A., Welbl, J., Clark, A., et al., 2022. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556 .
- Hofstadter, D., Sander, E., 2013. Surfaces and essences: Analogy as the fuel and fire of thinking. Basic Books.
- Hu, J., Ruder, S., Siddhant, A., Neubig, G., Firat, O., Johnson, M., 2020. Xtreme: A massively multilingual multi-task benchmark for evaluating cross-lingual generalisation, in: International Conference on Machine Learning, PMLR. pp. 4411–4421.
- Jiang, A.Q., Sablayrolles, A., Roux, A., Mensch, A., Savary, B., Bamford, C., Chaplot, D.S., Casas, D.d.l., Hanna, E.B., Bressand, F., et al., 2024. Mixtral of experts. arXiv preprint arXiv:2401.04088 .
- Jimenez, C.E., Yang, J., Wettig, A., Yao, S., Pei, K., Press, O., Narasimhan, K., 2023. Swe-bench: Can language models resolve real-world github issues? arXiv preprint arXiv:2310.06770 .
- Kahneman, D., 2011. Thinking, fast and slow. Macmillan.
- Kaplan, J., McCandlish, S., Henighan, T., Brown, T.B., Chess, B., Child, R., Gray, S., Radford, A., Wu, J., Amodei, D., 2020. Scaling laws for neural language models. arXiv preprint arXiv:2001.08361 .
- Kwon, W., Li, Z., Zhuang, S., Sheng, Y., Zheng, L., Yu, C.H., Gonzalez, J.E., Zhang, H., Stoica, I., 2023. Efficient memory management for large language model serving with pagedattention, in: Proceedings of the ACM SIGOPS 29th Symposium on Operating Systems Principles.
- Liang, P., Bommasani, R., Lee, T., Tsipras, D., Soylu, D., Yasunaga, M., Zhang, Y., Narayanan, D., Wu, Y., Kumar, A., et al., 2022. Holistic evaluation of language models. arXiv preprint arXiv:2211.09110 .
- Lightman, H., Kosaraju, V., Burda, Y., Edwards, H., Baker, B., Lee, T., Leike, J., Schulman, J., Sutskever, I., Cobbe, K., 2023. Let's verify step by step. arXiv preprint arXiv:2305.20050 .
- Lu, P., Mishra, S., Xia, T., Qiu, L., Chang, K.W., Zhu, S.C., Tafjord, O., Clark, P., Kalyan, A., 2022. Learn to explain: Multimodal reasoning via thought chains for science question answering. *Advances in Neural Information Processing Systems* 35, 2507–2521.
- Mistral AI, 2023. Mixtral-8x7b. <https://mistral.ai/news/mixtral-of-experts/>. Accessed: 2024-03-15.
- Muennighoff, N., Tazi, N., Magne, L., Reimers, N., 2022. Mteb: Massive text embedding benchmark. arXiv preprint arXiv:2210.07316 .
- Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al., 2022. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems* 35, 27730–27744.
- Perez, E., Ringer, S., Lukosiūtė, K., Nguyen, K., Chen, E., Heiner, S., Pettit, C., Olsson, C., Kundu, S., Kadavath, S., et al., 2022. Discovering language model behaviors with model-written evaluations. arXiv preprint arXiv:2212.09251 .
- Prystawski, B., Goodman, N.D., 2023. Why think step-by-step? reasoning emerges from the locality of experience. arXiv preprint arXiv:2304.03843 .
- Reimers, N., Gurevych, I., 2019. Sentence-bert: Sentence embeddings using siamese bert-networks, in: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing, pp. 3982–3992.
- Srivastava, A., Rastogi, A., Rao, A., Shoeib, A.A.M., Abid, A., Fisch, A., Brown, A.R., Santoro, A., Gupta, A., Garriga-Alonso, A., et al., 2022. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv preprint arXiv:2206.04615 .
- Team, G., Rivière, M., Pathak, S., Sessa, P.G., Hardin, C., Bhupatiraju, S., Hussénot, L., Mesnard, T., Shahriari, B., Ramé, A., et al., 2024. Gemma 2: Improving open language models at a practical size. arXiv preprint arXiv:2408.00118 .
- Touvron, H., Lavril, T., Izacard, G., Martinet, X., Lachaux, M.A., Lacroix, T., Rozière, B., Goyal, N., Hambro, E., Azhar, F., et al., 2023. Llama: Open and efficient foundation language models. arXiv preprint arXiv:2302.13971 .
- Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I., 2017. Attention is all you need, in: Advances in neural information processing systems, pp. 5998–6008.
- Wang, A., Pruksachatkun, Y., Nangia, N., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S., 2019. Super glue: A stickier benchmark for general-purpose language understanding systems, in: Advances in Neural Information Processing Systems.
- Wang, A., Singh, A., Michael, J., Hill, F., Levy, O., Bowman, S.R., 2018. Glue: A multi-task benchmark and analysis platform for natural language understanding, in: Proceedings of the 2018 EMNLP Workshop BlackboxNLP, pp. 353–355.
- Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., Chowdhery, A., Zhou, D., 2022. Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171 .
- Wang, X., Zhao, Z., Gao, Z., Ren, Z., Gao, Y., Li, J., 2023. Scibench: Evaluating college-level scientific problem-solving abilities of large language models. arXiv preprint arXiv:2307.10635 .
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., Yogatama, D., Bosma, M., Zhou, D., Metzler, D., et al., 2022a. Emergent abilities of large language models. arXiv preprint arXiv:2206.07682 .
- Wei, J., Wang, X., Schuurmans, D., Bosma, M., Ichter, B., Xia, F., Chi, E., Le, Q., Zhou, D., 2022b. Chain of thought prompting elicits reasoning in large language models. arXiv preprint arXiv:2201.11903 .

Zhang, Q., Hu, C., Upasani, S., Ma, B., Hong, F., Kamanuru, V., Rainton, J., Wu, C., Ji, M., Li, H., et al., 2025. Agentic context engineering: Evolving contexts for self-improving language models. arXiv preprint arXiv:2510.04618 .

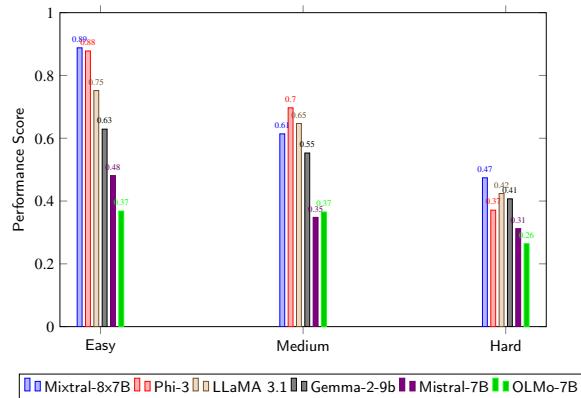


Figure 1: Baseline model performance across difficulty levels (MareNostrum 5, 19 problems). All models exhibit monotonic performance decrease with increasing complexity, with approximately 15–25% degradation per difficulty tier.

Dataset Sample Problems from Physics

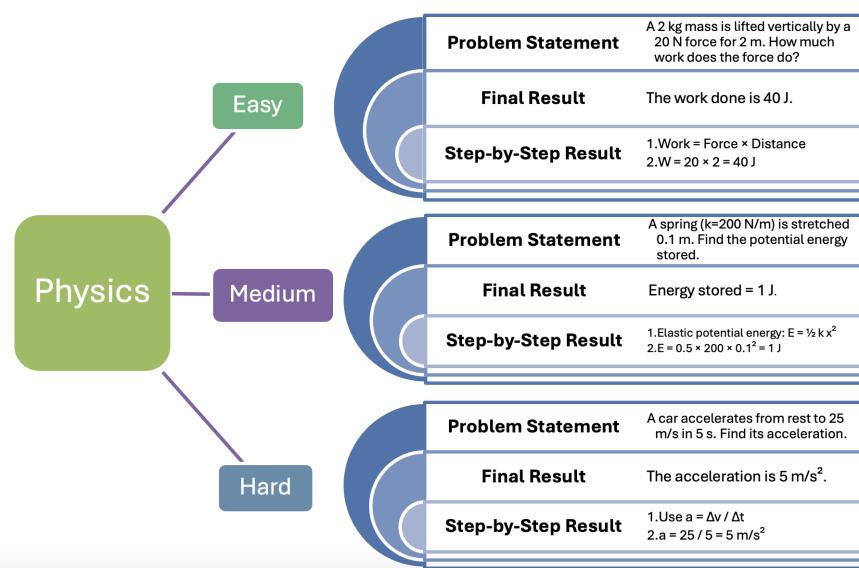


Figure 2: Representative example of the dataset schema. Each problem includes a *Problem Statement*, *Final Result*, and *Step-by-Step Result*, illustrated for three difficulty tiers (Easy, Medium, Hard) within the Physics domain. This fixed schema is used across all eight domains to standardize evaluation and facilitate process-level scoring.

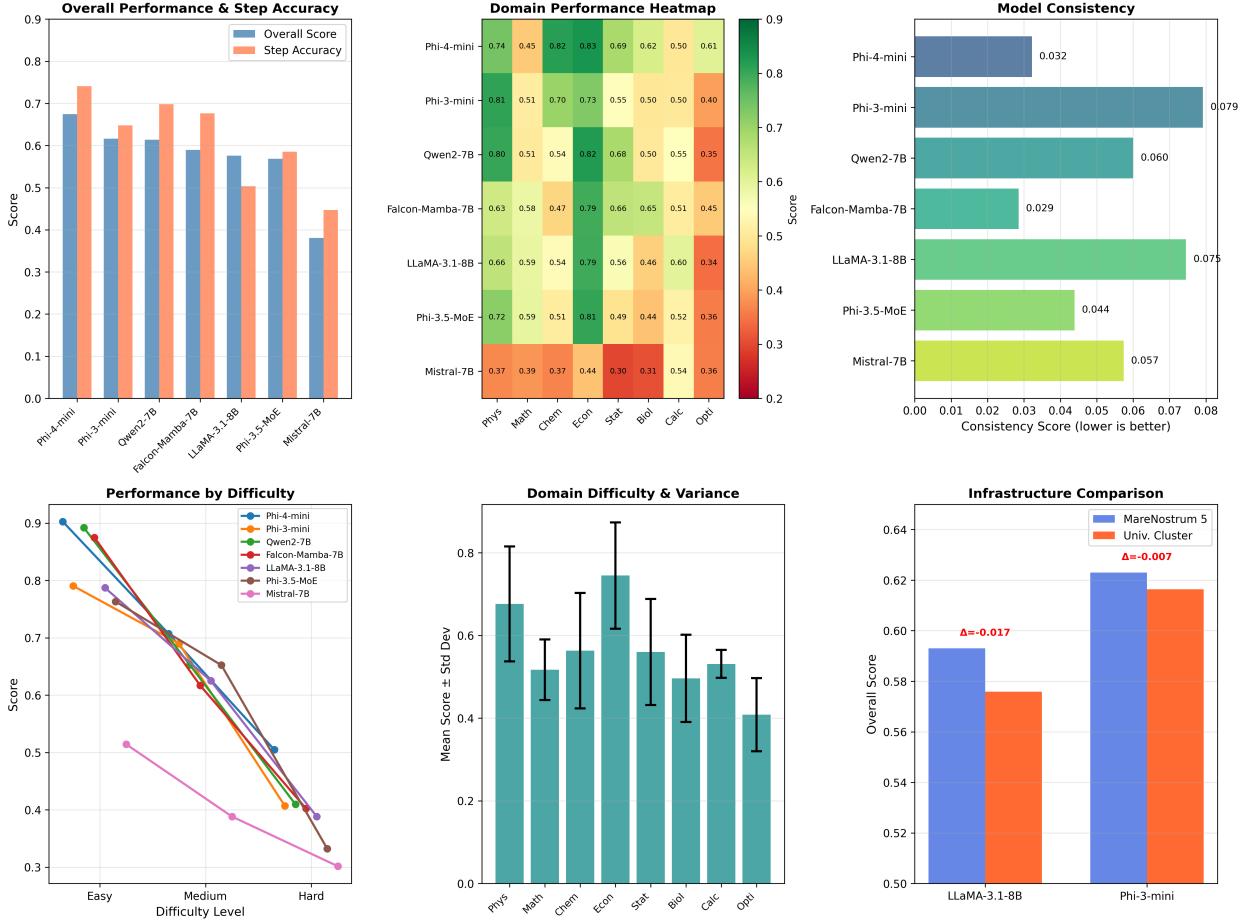


Figure 3: Comprehensive analysis of foundation model performance on university cluster infrastructure. **(Top left)** Overall score and step accuracy comparison across seven models, showing Phi-4-mini's dominance in both metrics. **(Top center)** Domain-specific performance heatmap revealing systematic patterns: Phi-4-mini excels in Chemistry (0.816) and Economics (0.833), while Optimization (bottom row) remains universally challenging across all architectures. **(Top right)** Consistency scores (lower is better) demonstrating Falcon-Mamba-7B's exceptional stability (0.029), followed by Phi-4-mini (0.032), critical for production deployment. **(Bottom left)** Performance degradation across difficulty levels shows approximately 20% drop per tier for most models, with Phi-4-mini maintaining the highest hard problem performance (0.505). **(Bottom center)** Cross-model domain statistics identify Economics as easiest (mean 0.745 ± 0.129) and Optimization as hardest (0.408 ± 0.089), while Calculus exhibits remarkably low variance (0.034), indicating model-independent performance. **(Bottom right)** Infrastructure validation comparing MareNostrum 5 supercomputer with university cluster shows minimal performance degradation (LLaMA-3.1-8B: -2.9%, Phi-3-mini: -1.1%), confirming reasoning quality is infrastructure-agnostic. Color intensity in heatmap represents score magnitude (green = high, red = low).

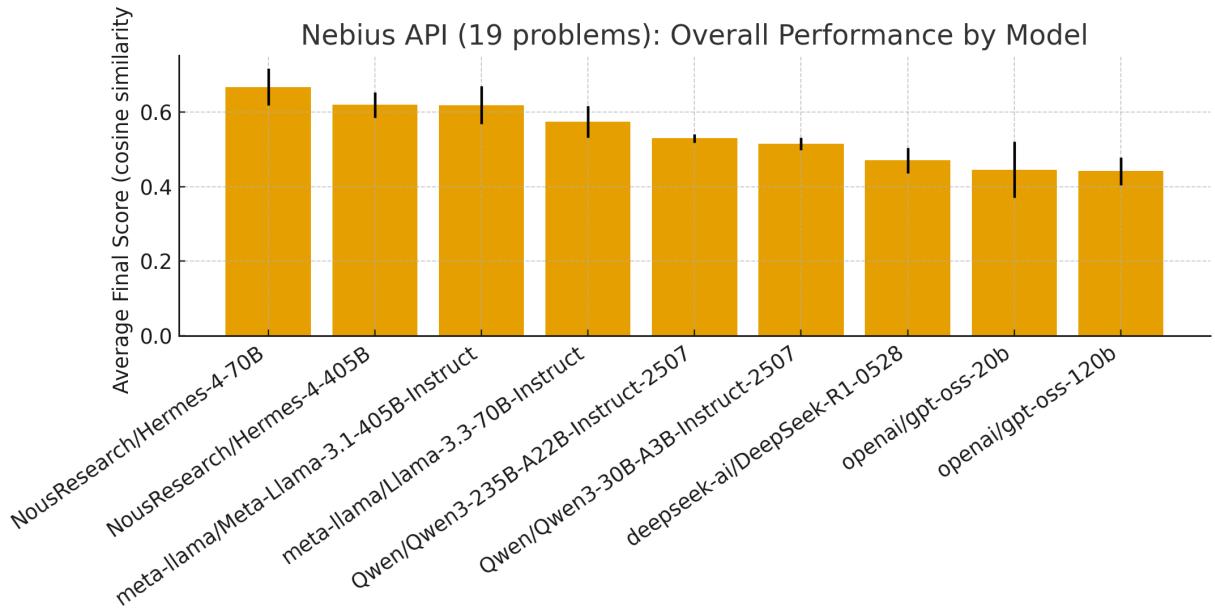


Figure 4: Nebius API (19 problems): overall average final score per model. Error bars denote the mean per-evaluation standard deviation.

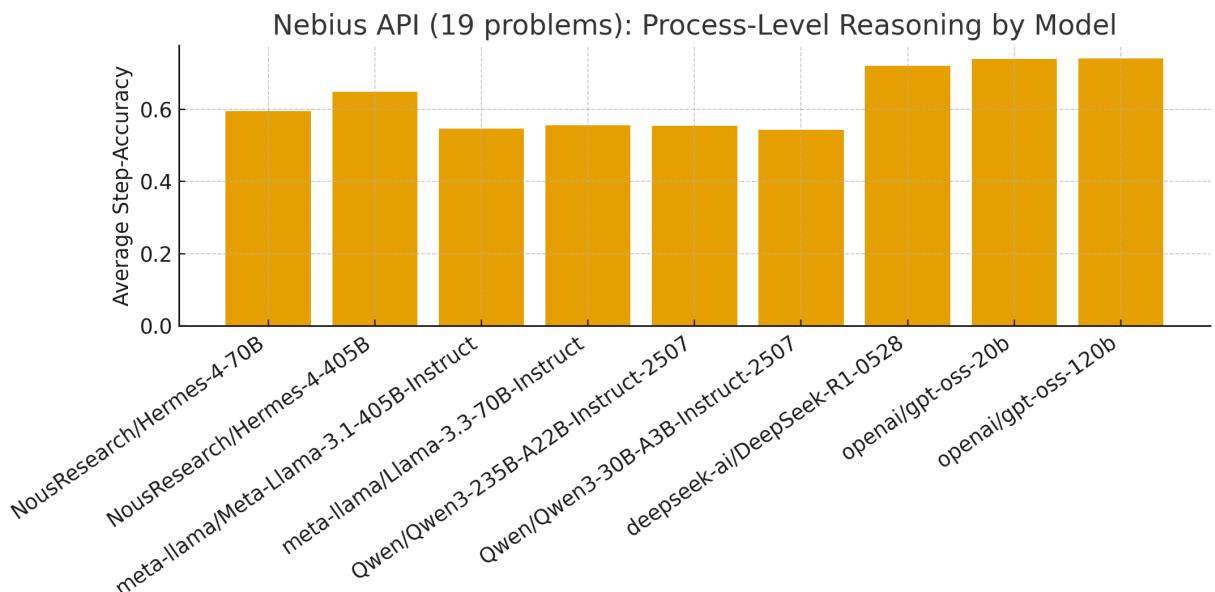
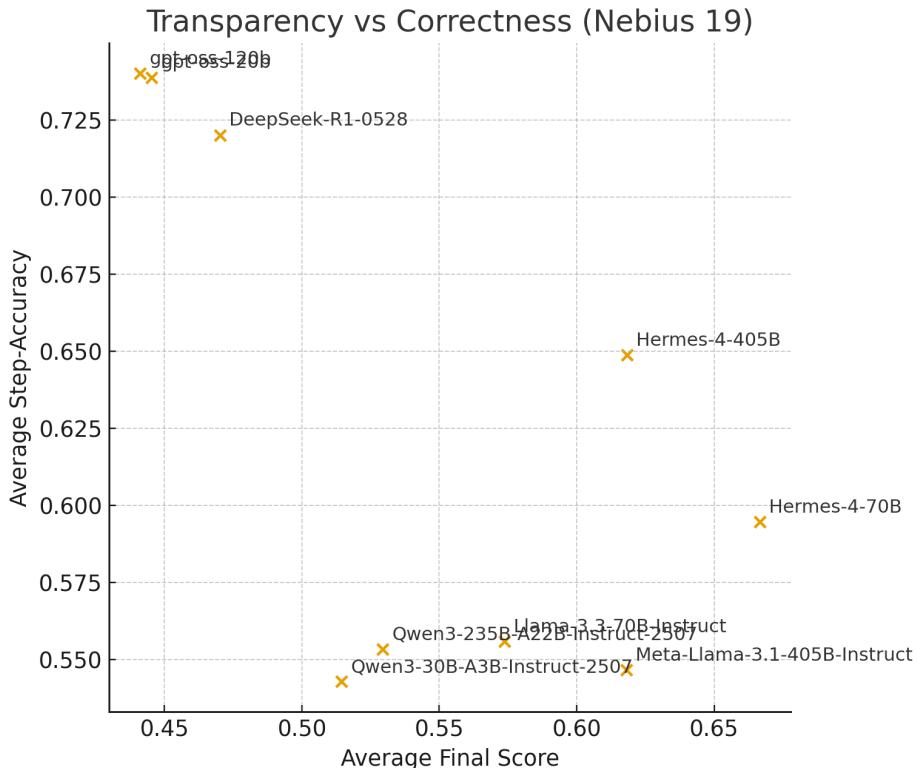
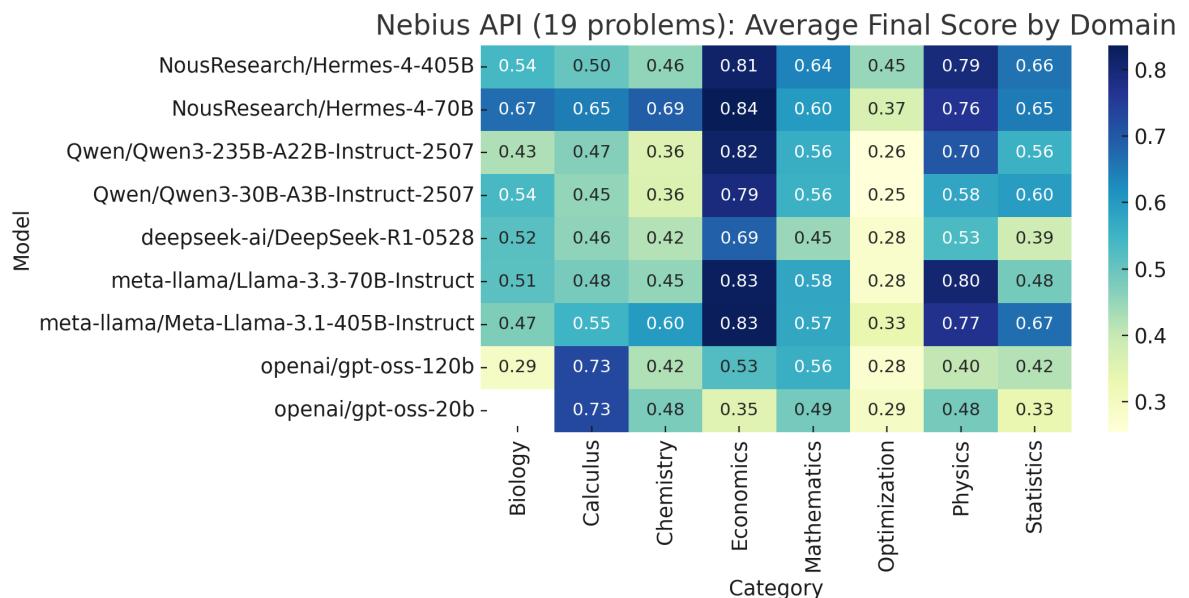


Figure 5: Nebius API (19 problems): average step-accuracy by model.

**Figure 6:** Transparency vs correctness (model-level): average step-accuracy vs average final score.**Figure 7:** Nebius API (19 problems): average final score by domain and model.

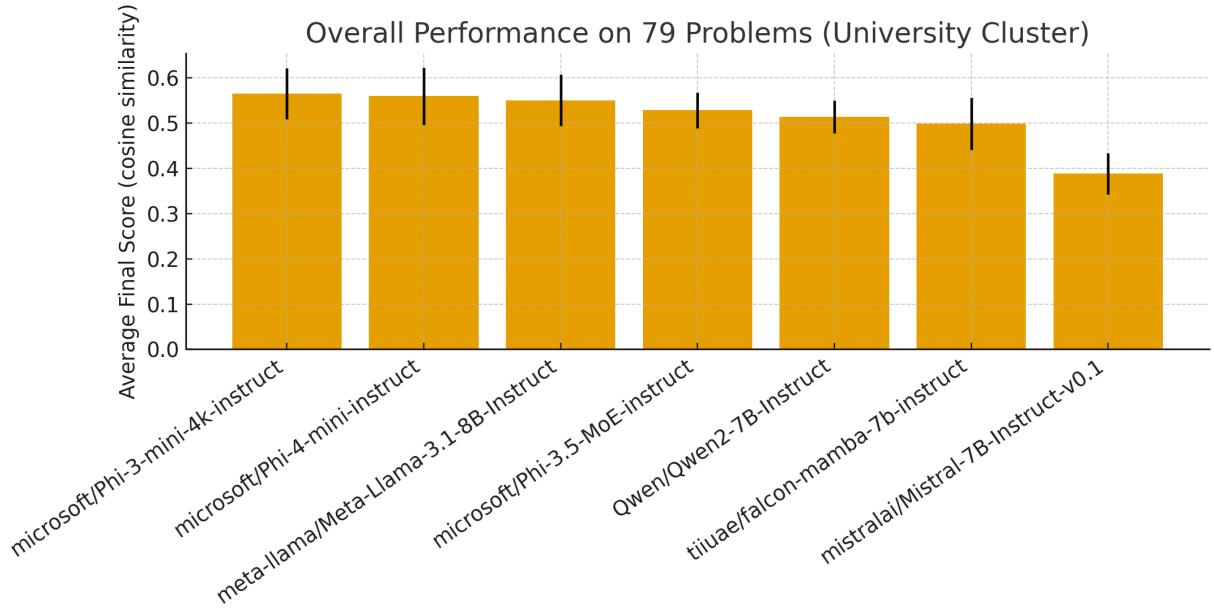


Figure 8: Overall performance across the 79-problem benchmark on the university cluster. Bars show average final score; error bars depict the mean per-evaluation standard deviation (lower is better).



Figure 9: Average step-accuracy by model (higher is better).

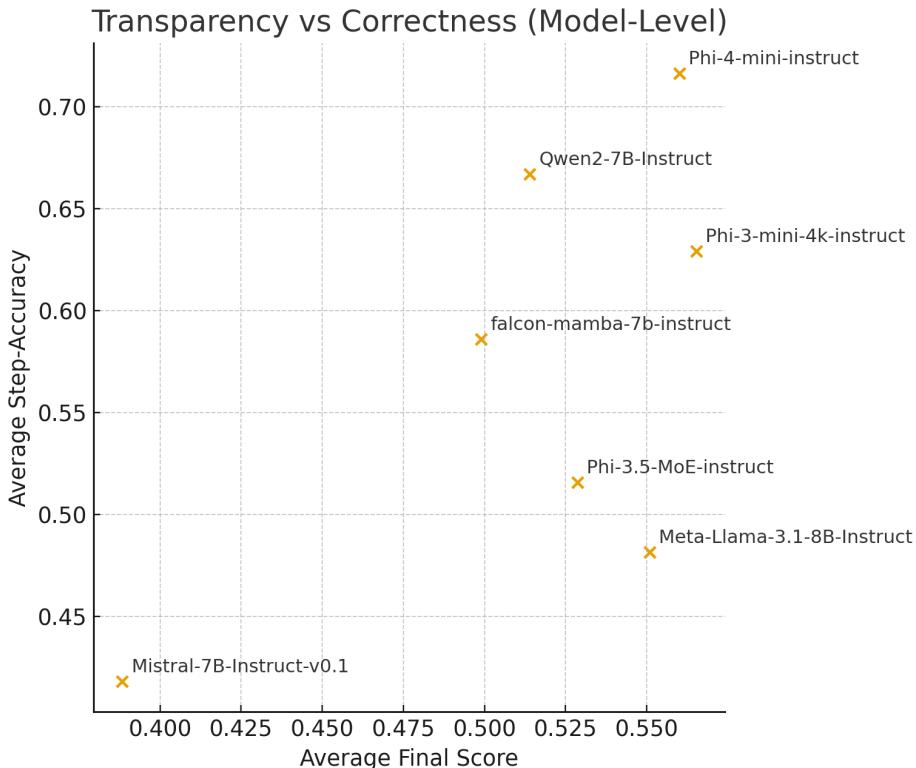


Figure 10: Transparency–correctness relationship at the model level: average step-accuracy (y-axis) vs average final score (x-axis).

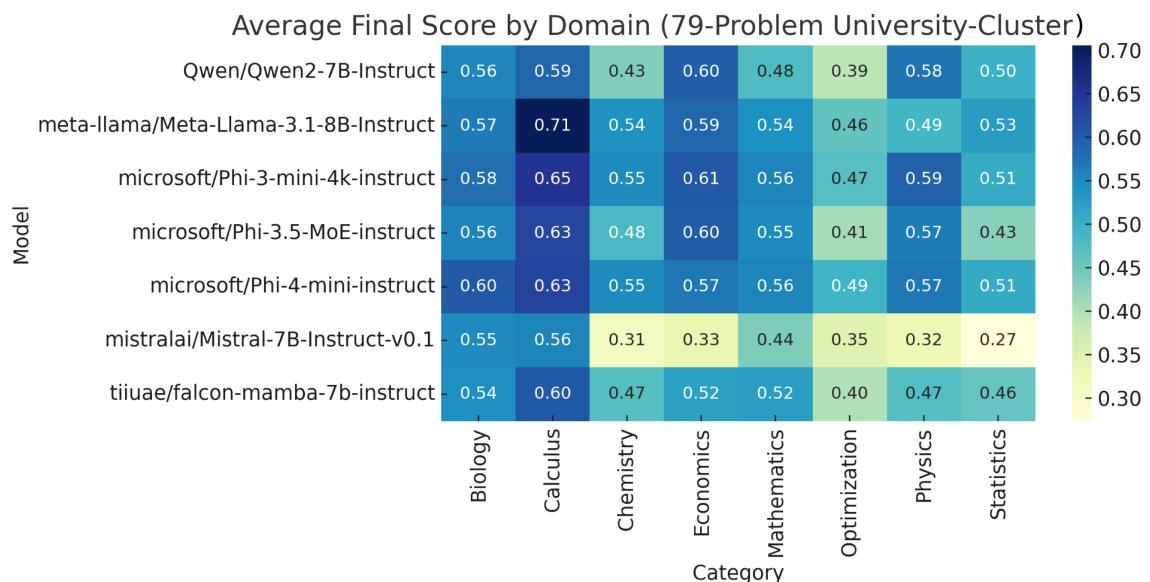


Figure 11: Heatmap of average final scores per domain across models on the 79-problem benchmark.

Average Final Score by Difficulty Level

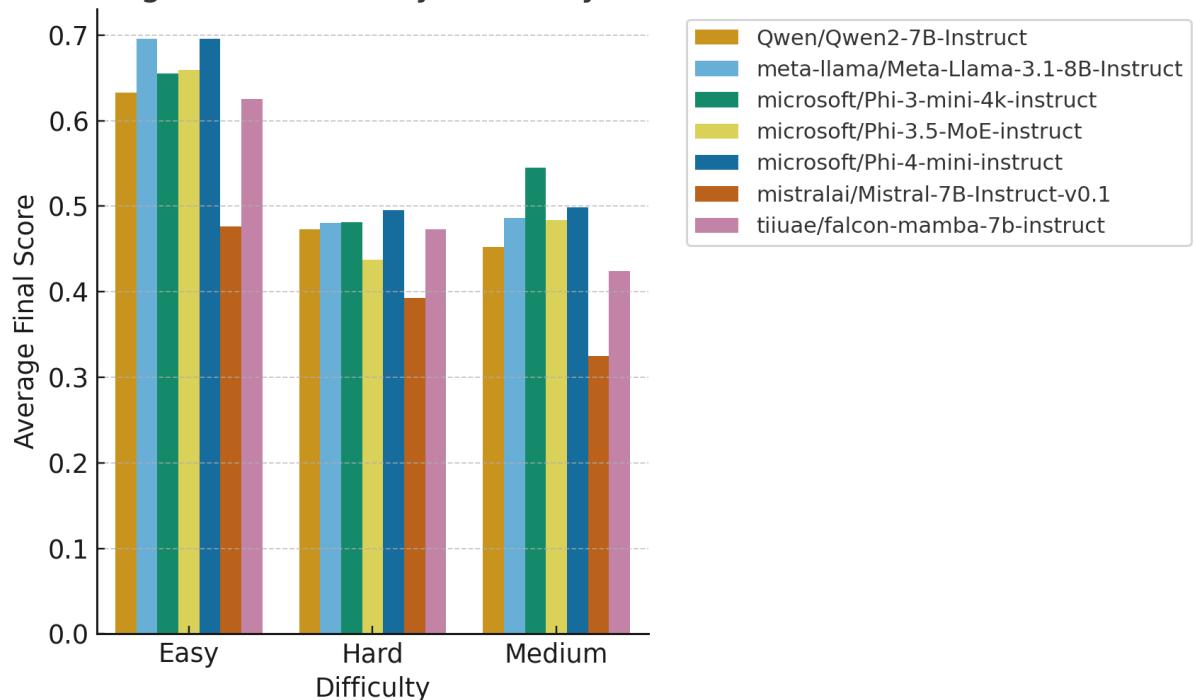


Figure 12: Average final score by difficulty level (Easy, Medium, Hard) for all evaluated models.

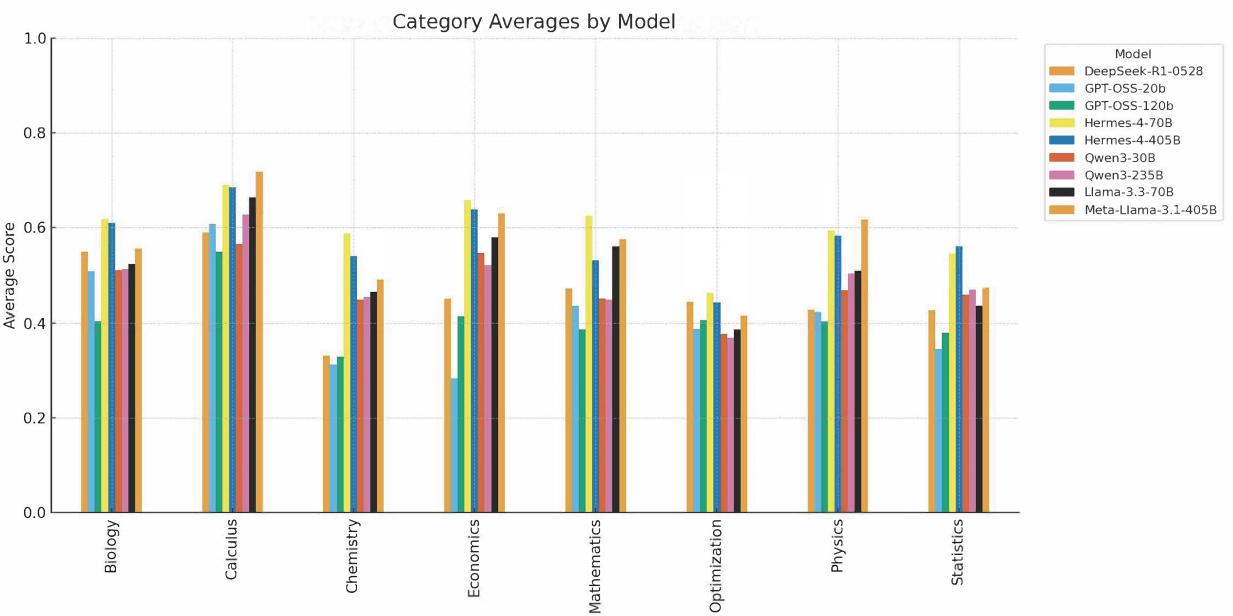


Figure 13: Average reasoning scores across academic domains (Biology, Calculus, Chemistry, Economics, Mathematics, Optimization, Physics, and Statistics) for all evaluated models. Hermes-4-70B and Meta-Llama-3.1-405B-Instruct show the strongest cross-domain balance.

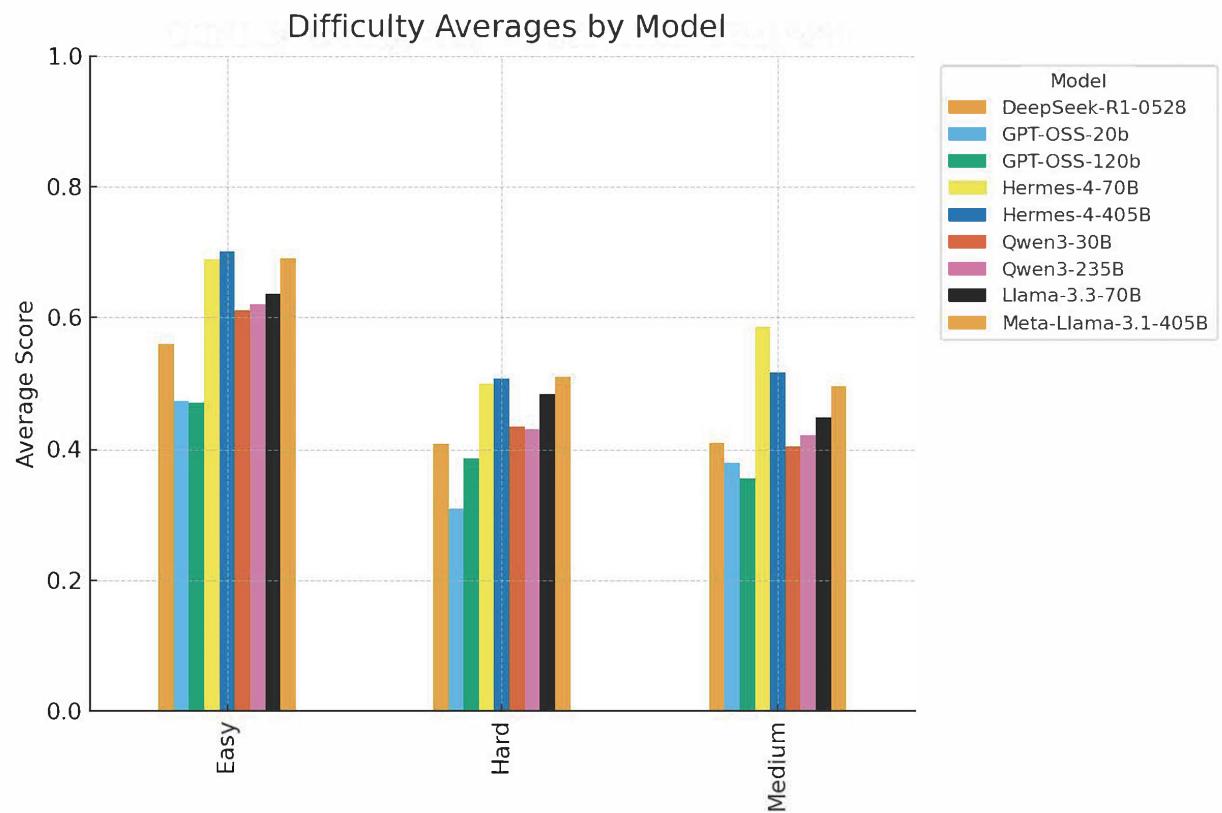


Figure 14: Average final accuracy by difficulty level (Easy, Medium, Hard) for each model. Hermes-4 and Meta-Llama families retain higher performance on hard problems, while DeepSeek-R1 and GPT-OSS show stronger easy-case accuracy.

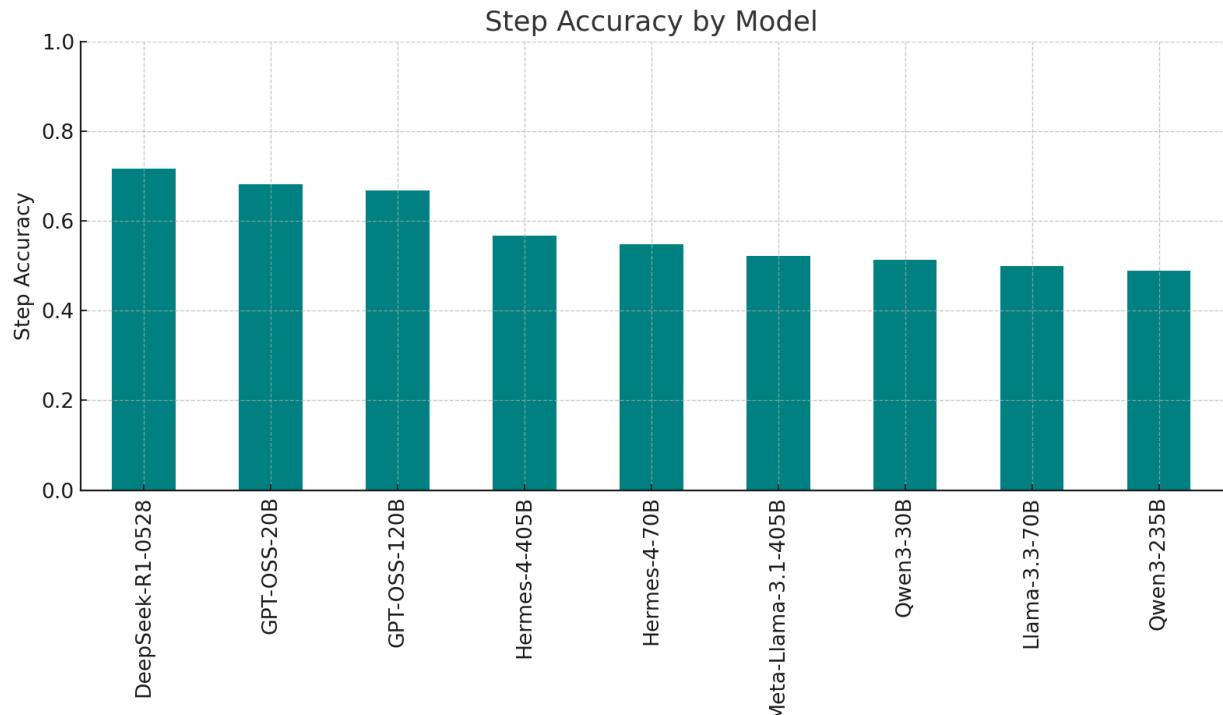


Figure 15: Mean step-by-step reasoning accuracy per model. DeepSeek-R1 exhibits the highest step-accuracy, indicating strong transparency in intermediate reasoning despite lower final correctness.

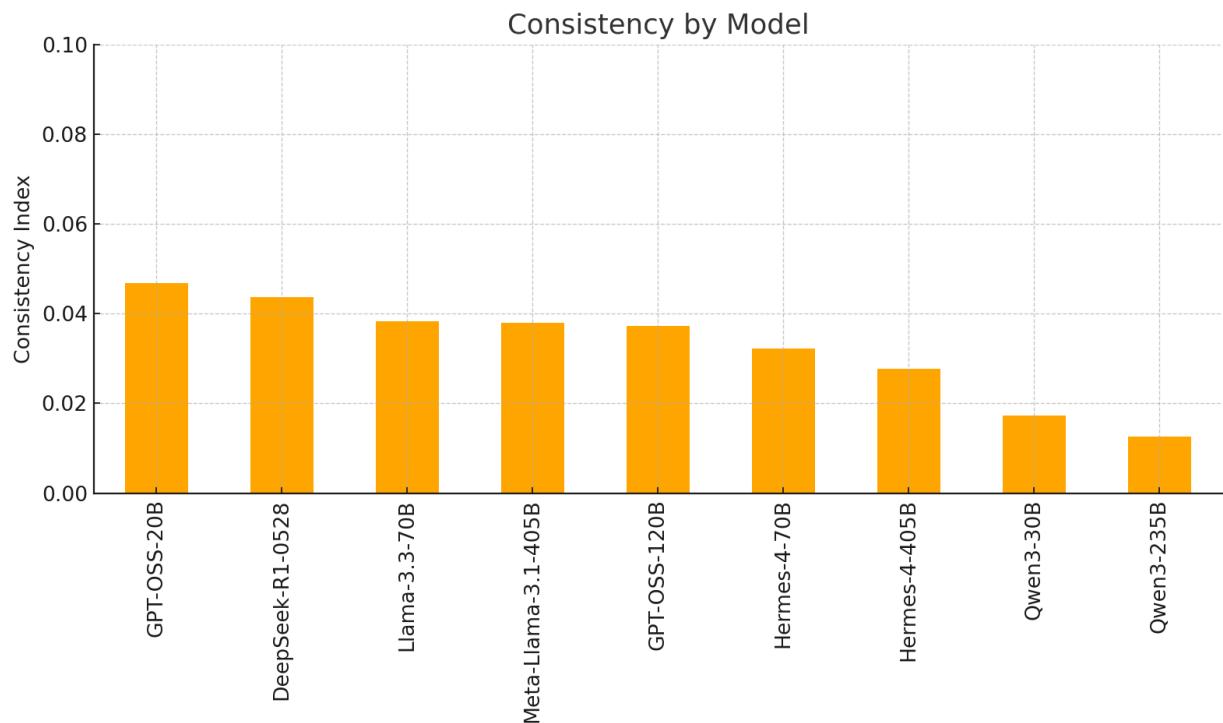


Figure 16: Consistency index (mean score standard deviation) by model. Lower bars indicate more stable outputs across repeated runs. Qwen3 and Hermes models achieve the highest consistency.

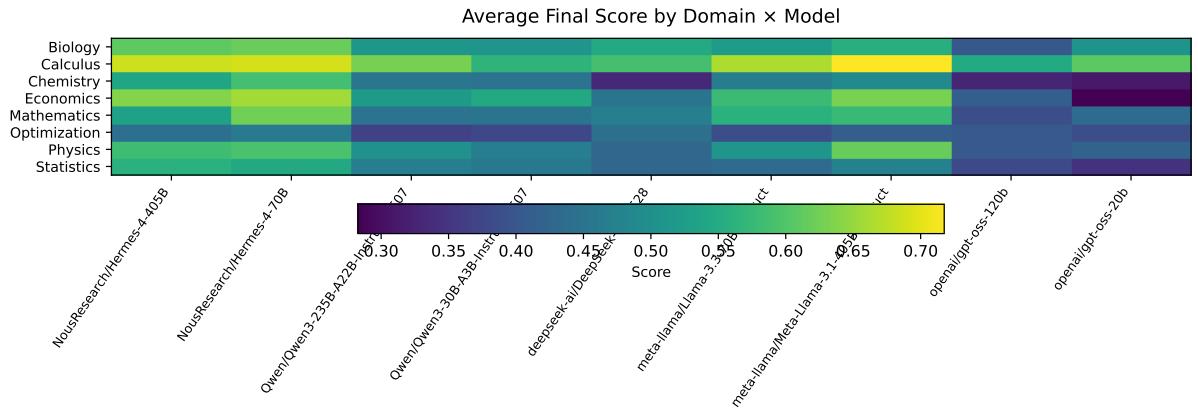


Figure 17: Average Final Score by Domain x Model, visualizing domain-specific strengths and weaknesses. Calculus remains the highest-performing domain for most models.

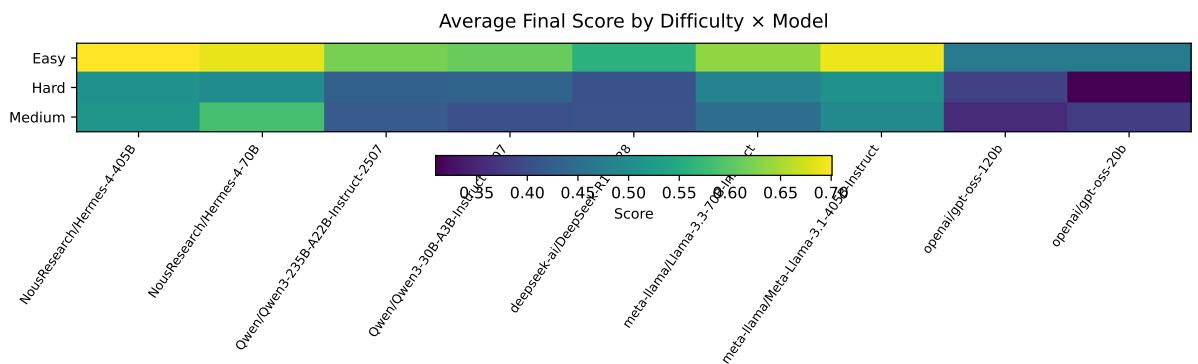


Figure 18: Average Final Score by Difficulty x Model, showing degradation of performance with increasing problem complexity.

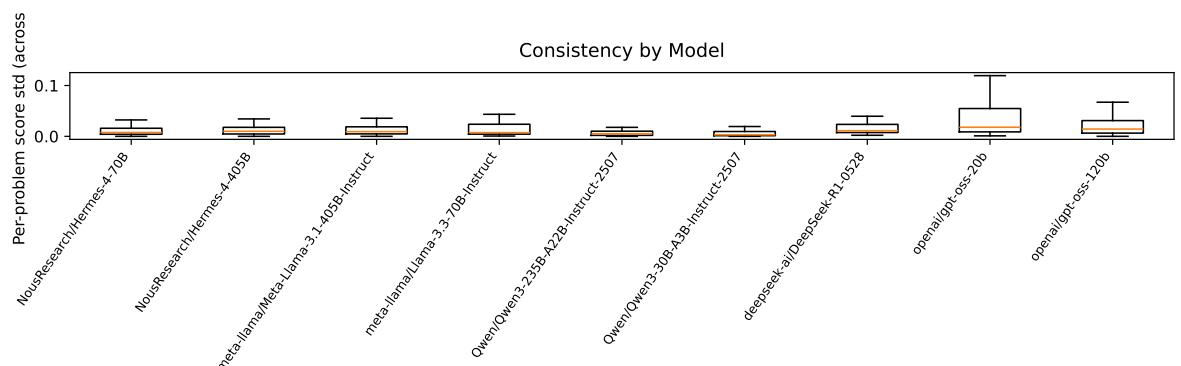


Figure 19: Distribution of per-problem score standard deviations across runs (lower is more consistent).

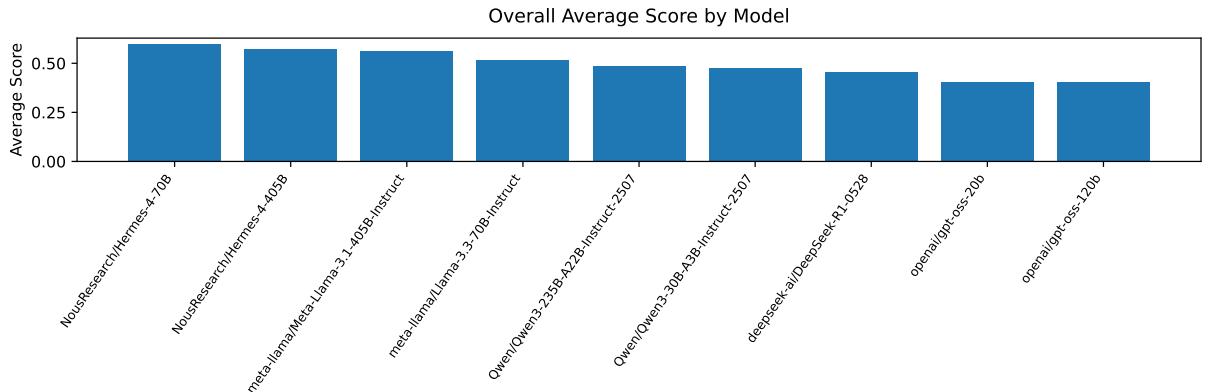


Figure 20: Overall Average Score by Model, ranking the evaluated models by mean reasoning accuracy.

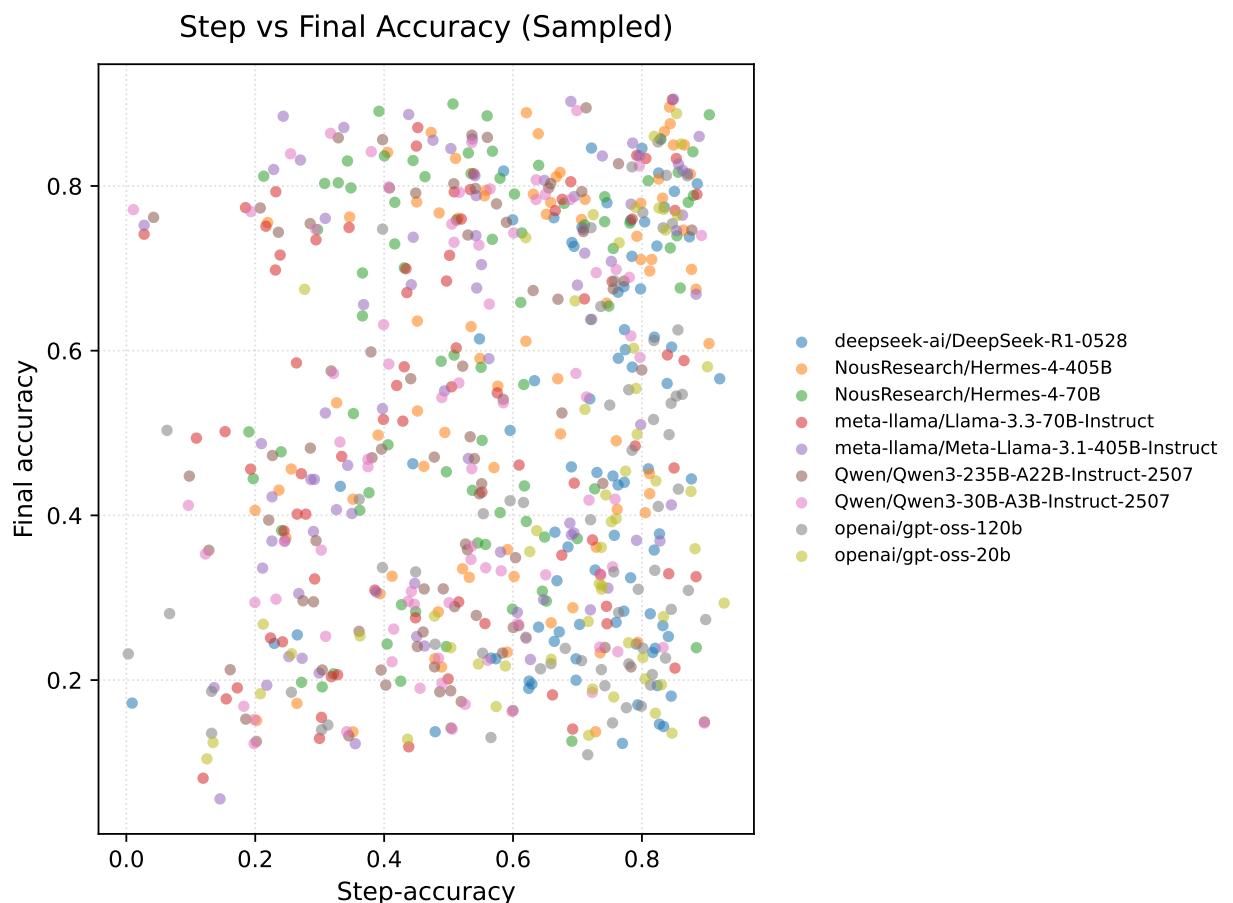


Figure 21: Relationship between step-accuracy and final accuracy (sample of several thousand model–problem responses). Each point represents a problem instance; the correlation between step fidelity and final correctness varies across models.

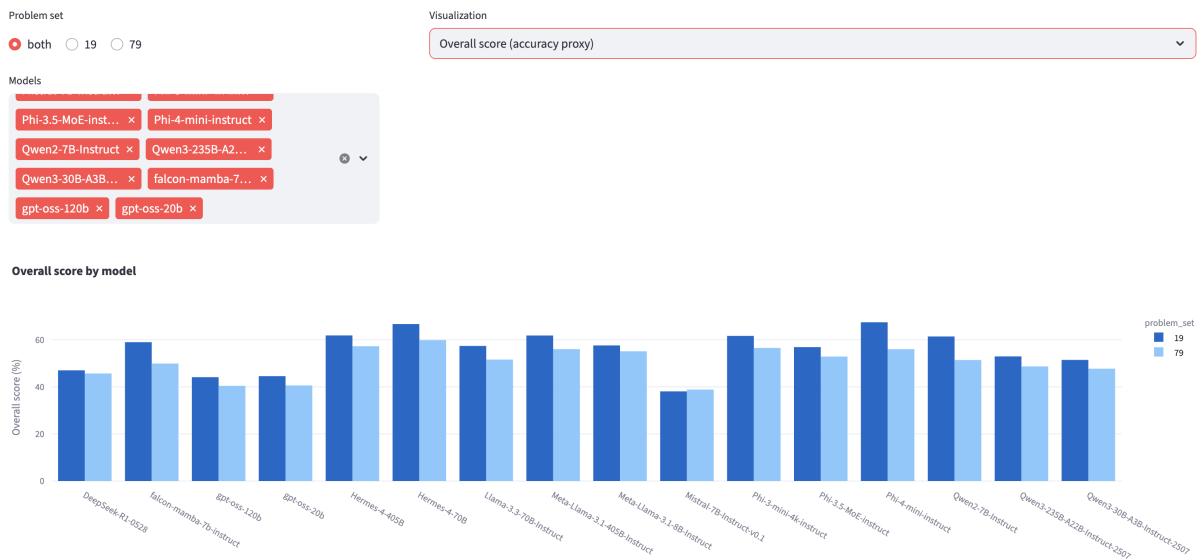


Figure 22: Interactive visualization tool for cross-platform LLM reasoning evaluation. The web-based interface enables dynamic exploration of results across models and problem set. Users can filter by problem set (19 vs 79), visualize overall scores with step-accuracy metrics, compare difficulty-stratified performance via radar charts, analyze category-specific patterns through heatmaps, and examine reasoning step distributions. The tool supports public data exploration at <https://crossplatform-llm-decurto.streamlit.app/>.