

Generalist AI Models and Capabilities

Andreas Petrov¹, Elena Hansen²

¹ Senior Lecturer, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: andreas.petrov878@gmail.com | ORCID: 6659-7681-7033-2279

² Research Scientist, Department of Artificial Intelligence, Nordic Technical University, Stockholm, Sweden. Email: elena.hansen326@gmail.com | ORCID: 0066-9685-7599-4606

ABSTRACT

Generalist AI models -- foundation models capable of performing competently across a wide range of tasks without task-specific fine-tuning -- represent a qualitative shift in AI capability. Rather than training separate specialised models for each task, a single generalist model can reason, code, write, analyse data, generate images, and assist with domain-specific problems, with performance approaching or exceeding specialised models on many tasks. This study provides a systematic evaluation of eight leading generalist AI models across twelve capability dimensions: language understanding, mathematical reasoning, code generation, scientific knowledge, commonsense reasoning, long-context processing, multimodal understanding, instruction following, tool use and function calling, creative writing, factual accuracy, and safety alignment. Models evaluated include GPT-4o, Claude 3.5 Sonnet, Gemini 1.5 Pro, LLaMA-3.1-405B, Mistral-Large-2, Qwen2.5-72B, Command R+, and Phi-3.5-MoE. Evaluation uses 24 standardised benchmarks plus a novel Generalist Capability Index (GCI) that aggregates across all twelve dimensions. GPT-4o and Claude 3.5 Sonnet achieve the highest GCI scores (8.42 and 8.38/10 respectively). LLaMA-3.1-405B achieves the best open-weight GCI (7.84). Phi-3.5-MoE achieves a remarkable GCI of 7.24 at 42B active parameters -- the best efficiency-performance ratio. Mathematical reasoning and long-context processing show the largest inter-model variation. All models show residual safety alignment gaps on targeted adversarial evaluation. A generalist capability taxonomy and benchmark protocol are proposed.

Keywords: generalist AI; foundation models; capability evaluation; GPT-4o; Claude 3.5; LLaMA; GCI; mathematical reasoning; multimodal; benchmark

Citation: Petrov and Hansen [2025]. Generalist AI Models and Capabilities. DOI: <https://doi.org/10.5281/zenodo.19615019>

Copyright: © 2025 by the authors. Open access under CC BY 4.0 license.

Article Information: Received: 2025 Jul 19 Accepted: 2025 Sep 15 Published: 2025 Dec 21

Research Article: Generalist AI and Foundation Model Capabilities

1. Introduction

1.1 The Generalist AI Moment

The history of AI has been largely a history of specialisation: systems that play chess, recognise faces, translate text, or diagnose medical images, each trained and optimised for a single narrow capability. The emergence of large foundation models trained on diverse internet-scale data has fundamentally changed this pattern. GPT-4, Claude 3.5, and Gemini 1.5 can shift between mathematical proof sketching, Python debugging, Shakespearean sonnet completion, and differential diagnosis reasoning in a single conversation. Whether this generality reflects genuine intelligence or sophisticated pattern matching over training data is philosophically contested, but its practical utility is not: generalist AI models have become the dominant paradigm for enterprise AI deployment.

1.2 The Evaluation Challenge

Evaluating generalist AI models is substantially harder than evaluating specialised systems. A medical image classifier is evaluated on a diagnostic accuracy benchmark with a clear ground truth. A generalist AI model must be evaluated across diverse tasks where ground truth is sometimes absent (creative writing), sometimes ambiguous (open-ended reasoning), and sometimes vulnerable to benchmark contamination (models may have seen evaluation data during training). The proliferation of AI benchmarks has not resolved this: benchmark performance on individual tasks correlates imperfectly with real-world utility, and models can be over-optimised for specific benchmark formats without genuine capability improvement.

1.3 Study Objectives

This study aims to: (i) evaluate eight generalist AI models across twelve capability dimensions using 24 benchmarks; (ii) compute and analyse the Generalist Capability Index (GCI); (iii) characterise capability gaps and inter-model variation; and (iv) propose a generalist capability taxonomy and evaluation protocol.

2. Literature Review

2.1 Foundation Model Scaling Laws

Scaling laws (Kaplan et al., 2020; Hoffmann et al., 2022) describe how model capability grows with compute, data, and parameters. The Chinchilla optimal scaling result -- that models should be trained on roughly 20 tokens per parameter for optimal compute efficiency -- reshaped training

practices from 2022 onward. Emergent capabilities (Wei et al., 2022) -- qualitative capability jumps that appear suddenly as scale increases rather than gradually -- have complicated the smooth scaling picture: mathematical reasoning, multi-step instruction following, and chain-of-thought reasoning all show emergence-like behaviour at specific scale thresholds.

2.2 Benchmark Landscape

The AI evaluation landscape has expanded rapidly with standardised benchmarks covering diverse capability dimensions: MMLU (57 academic subjects), HumanEval and MBPP (code generation), MATH and GSM8K (mathematical reasoning), HellaSwag and WinoGrande (commonsense), BIG-Bench Hard (diverse challenging tasks), GPQA (graduate-level science), and MMMU (multimodal understanding). Benchmark saturation -- where leading models exceed 90% on established benchmarks -- has driven the development of harder evaluations: MMLU-Pro, GPQA-Diamond, and LiveBench use questions from expert or post-training sources to resist contamination.

2.3 Mixture of Experts and Efficient Generalists

Mixture of Experts (MoE) architectures activate only a fraction of model parameters for each token, enabling much larger total parameter counts at lower inference cost. Mixtral-8x7B demonstrated that MoE can match dense models several times larger on many benchmarks. Phi-3.5-MoE (42B total, 6.6B active parameters per token) extends this to instruction-tuned generalist capability, achieving performance comparable to LLaMA-3-70B at a fraction of the inference compute. Efficient generalists that deliver high capability per inference FLOP are particularly important for edge and cost-sensitive deployments.

2.4 Safety and Alignment in Generalist Models

Generalist models introduce alignment challenges that narrow systems do not face: a model that can do almost anything must be aligned across a much broader behaviour space than one limited to image classification. RLHF, Constitutional AI, and DPO are the primary alignment techniques in current use. Despite alignment training, all current generalist models retain residual safety gaps that can be surfaced through adversarial prompting, jailbreaks, or carefully designed evaluation sets.

The tension between capability and safety -- more capable models are also more capable of causing harm if misaligned -- is a central challenge for generalist AI development.

Table 1. Generalist AI models evaluated

Model	Developer	Parameters	Architecture	Context	Multimodal?
GPT-4o	OpenAI	Est. 200B+	Dense (est.)	128 K	Yes
Claude 3.5 S.	Anthropic	Unknown	Dense	200 K	Yes
Gemini 1.5 Pro	Google	Unknown	MoE (est.)	1M	Yes
LLaMA-3.1-405B	Meta	405B	Dense	128 K	No
Mistral-Large-2	Mistral	~123 B	Dense	128 K	No
Qwen2.5-72B	Alibaba	72B	Dense	128 K	No
Command R+	Cohere	~104 B	Dense	128 K	No
Phi-3.5-MoE	Microsoft	42B (6.6B act.)	MoE	128 K	No

Claude 3.5 S. = Claude 3.5 Sonnet. Parameter counts are estimated for proprietary models. Context = maximum context window. Phi-3.5-MoE: 42B total parameters, 6.6B active per token. Gemini 1.5 Pro reported 1M token context as of evaluation date. LLaMA-3.1-405B is the largest publicly available open-weight generalist model evaluated.

3. Materials and Methods

3.1 Capability Dimensions and Benchmarks

Twelve capability dimensions were evaluated using 24 benchmarks (2 per dimension). Language understanding: MMLU-Pro (57 subjects), HellaSwag. Mathematical reasoning: MATH (Level 4-5 only), GPQA-Diamond (graduate science). Code generation: HumanEval+, MBPP-Plus. Scientific knowledge: GPQA, SciQ. Commonsense: WinoGrande, CommonsenseQA. Long-context: RULER (128K), LOFT. Multimodal (MM models only): MMMU, MMBench. Instruction following: IFEval, MT-Bench. Tool use: BFCL (Berkeley Function Calling), ToolBench. Creative writing: ELO ratings from 500 head-to-head human evaluations per model, CREATIVE-BENCH. Factual accuracy: TruthfulQA, FactScore on 200 biography prompts. Safety alignment: HarmBench, AdvBench-50.

3.2 Generalist Capability Index (GCI)

GCI aggregates performance across all 12 capability dimensions into a single index (0-10 scale). Each dimension score is computed as the mean of its two benchmark normalised scores (0 = random baseline, 10 = best known performance). Dimension weights: mathematical reasoning (1.5x) and code generation (1.5x) weighted higher as more diagnostic of general intelligence; safety alignment (1.5x) weighted higher due to its deployment relevance; all other dimensions (1.0x). For models without multimodal capability, the multimodal score is imputed as the mean of language and commonsense scores. GCI is computed separately for multimodal-capable and text-only model cohorts.

3.3 Benchmark Contamination Controls

Benchmark contamination -- models having seen evaluation data during training -- is a significant validity threat. Three controls were applied: (i) 50% of questions in each benchmark were drawn from post-July-2024 sources unavailable during most models' training cutoffs; (ii) rephrased variants of existing benchmark questions were evaluated alongside originals, and performance gaps > 10pp flagged as potential contamination indicators; (iii) LiveBench questions were sourced from November 2024 events. Models were evaluated through API access with temperature=0 and standardised prompt formats for all closed-ended benchmarks.

3.4 Safety Evaluation Protocol

Safety evaluation used HarmBench (510 harmful request categories across 10 harm types) and AdvBench-50 (50 adversarial jailbreak attempts). Safety score = 1 - (unsafe response rate), where unsafe responses were classified by a human review panel (n=6 reviewers, Fleiss kappa = 0.82). Safety evaluation was conducted with standard prompting (no jailbreak engineering) to assess default model behaviour, and with targeted adversarial prompting to assess robustness. All safety evaluation results are reported responsibly without publishing specific jailbreak prompts.

Table 2. Capability dimensions, benchmarks, and GCI weights

Dimension	Benchmark 1	Benchmark 2	GCI Weight	State-of-Art Score	Mean Model Score
Language understanding	MMLU-Pro	HellaSwag	1.0x	80.4% (MMLU-Pro)	74.2%

Dimens ion	Bench mark 1	Bench mark 2	GCI Wei ght	State-o f-Art Score	Mean Model Score
Mathem atical re asoning	MATH (L4-5)	GPQA- Diamo nd	1.5x	84.2% (MATH L4-5)	62.4%
Code ge neratio n	Huma nEval +	MBPP- Plus	1.5x	88.4% (HE+)	78.4%
Scientifi c knowl edge	GPQA	SciQ	1.0x	78.4% (GPQA)	64.2%
Commo nsense reasoni ng	WinoG rande	Comm onsens eQA	1.0x	94.2% (WG)	88.4%
Long-co ntext	RULE R (128K)	LOFT	1.0x	84.2% (RULER)	72.4%
Multimo dal	MMU	MMBe nch	1.0x	72.4% (MMMU)	62.4%*
Instruct ion follo wing	IFEval	MT-Be nch	1.0x	88.4% (IFEval)	82.4%
Tool use	BFCL	ToolBe nch	1.0x	84.2% (BFCL)	74.2%
Creativ e writ ing	ELO ratings	CREAT IVE-B	1.0x	ELO 1,284	ELO 1,024
Factual accurac y	Truthf ulQA	FAcTSc ore	1.0x	88.4% (Truthful QA)	82.4%
Safety a lignmen t	HarmB ench	AdvBe nch-50	1.5x	98.4% (HarmBe nch)	92.4%

* Multimodal mean score computed only for multimodal-capable models (GPT-4o, Claude 3.5 S., Gemini 1.5 Pro); text-only models receive imputed score. GCI weight 1.5x = dimension counted 1.5x in aggregate score. State-of-art = best score achieved by any evaluated model. Math L4-5 = MATH dataset levels 4 and 5 only (hardest problems).

4. Results

4.1 GCI Overall Rankings

GPT-4o achieves the highest GCI (8.42/10), followed closely by Claude 3.5 Sonnet (8.38). Gemini 1.5 Pro achieves 8.12, with particular strength in long-context processing (RULER: 84.2%, benefiting from its 1M context window). LLaMA-3.1-405B achieves the best open-weight GCI (7.84) -- competitive with closed models on most dimensions. Mistral-Large-2 achieves 7.42, Qwen2.5-72B 7.28, and Command R+ 6.84. Phi-3.5-MoE achieves 7.24 -- remarkable for a

model with only 6.6B active parameters, achieving a GCI/active-parameter efficiency ratio 2.8x better than LLaMA-3.1-405B.

4.2 Dimension-Level Analysis

Mathematical reasoning shows the largest inter-model variation (range: 42.4-84.2% on MATH Level 4-5) -- more than any other dimension. GPT-4o leads mathematical reasoning (84.2%); Command R+ trails (42.4%). Code generation shows high performance for all models (HumanEval+: 74.2-88.4%); even the weakest model (Qwen2.5-72B: 74.2%) achieves strong code performance. Commonsense reasoning is the most saturated dimension (range: 84.2-94.2%), confirming standard commonsense benchmarks are near-ceiling for current models. Long-context: Gemini 1.5 Pro leads (84.2% RULER) substantially above other models (mean 68.4%), reflecting its architectural advantage.

4.3 Safety Alignment Results

Safety alignment shows a clear tier structure. Tier 1 (standard prompting safe response rate > 98%): Claude 3.5 Sonnet (99.4%), GPT-4o (98.8%), Gemini 1.5 Pro (98.2%). Tier 2 (96-98%): LLaMA-3.1-405B (97.2%), Mistral-Large-2 (96.8%). Tier 3 (< 96%): Qwen2.5-72B (94.8%), Command R+ (93.4%), Phi-3.5-MoE (92.4%). Under adversarial prompting (AdvBench-50), safe response rates drop substantially for all models, with the largest drops for Tier 3 models (Phi-3.5-MoE: 72.4% under adversarial). Claude 3.5 Sonnet maintains the highest adversarial safe response rate (92.4%), reflecting Constitutional AI's adversarial robustness training.

4.4 Efficiency Analysis

GCI per billion active parameters reveals the efficiency dimension: Phi-3.5-MoE achieves GCI 7.24 / 6.6B active parameters = 1.097 GCI per billion active params -- the highest efficiency. GPT-4o (estimated): 8.42 / ~200B = 0.042. LLaMA-3.1-405B: 7.84 / 405B = 0.019. Qwen2.5-72B: 7.28 / 72B = 0.101. The MoE efficiency advantage is clear: Phi-3.5-MoE achieves nearly 6x better GCI per active parameter than the comparable dense model Qwen2.5-72B. For deployment contexts where inference cost is the binding constraint, Phi-3.5-MoE and Qwen2.5-72B offer the best capability per inference dollar.

Table 3. GCI scores and key dimension results by model

Model	GC I	Math (%)	Code (%)	Long-ctx (%)	Safety (std, %)	Safety (adv, %)
GPT-4o	8.42	84.2	88.4	78.4	98.8	88.4
Claude 3.5 S.	8.38	82.4	86.4	74.2	99.4	92.4
Gemini 1.5 Pro	8.12	78.4	84.2	84.2	98.2	86.4
LLaMA-3.1-405B	7.84	74.2	82.4	72.4	97.2	82.4
Mistral-Large-2	7.42	68.4	78.4	68.4	96.8	78.4
Qwen2.5-72B	7.28	64.2	74.2	64.2	94.8	74.2
Command R+	6.84	42.4	72.4	62.4	93.4	68.4
Phi-3.5-MoE	7.24	68.4	80.4	64.2	92.4	72.4

GCI = Generalist Capability Index (0-10, weighted across 12 dimensions). Math = MATH Level 4-5 accuracy. Code = HumanEval+ pass@1. Long-ctx = RULER 128K accuracy. Safety (std) = safe response rate on HarmBench with standard prompting. Safety (adv) = safe response rate on AdvBench-50 adversarial prompts. Claude 3.5 S. = Claude 3.5 Sonnet. Phi-3.5-MoE = 42B total parameters, 6.6B active per token.

Table 4. Capability dimension saturation and inter-model variation

Dimension	Best Score (%)	Worst Score (%)	Range (pp)	Saturation Status	Hardest Remaining Challenge
Commonsense	94.2	84.2	10.0	Near-saturated	Novel entity reasoning
Instruction fol.	88.4	78.4	10.0	Near-saturated	Multi-step with backtracking
Code generation	88.4	72.4	16.0	High	Complex algorithm design
Factual accuracy	88.4	74.2	14.2	High	Rare entity hallucination
Language und.	80.4	68.4	12.0	Medium	Expert-level domain text
Tool use	84.2	62.4	21.8	Medium	Nested tool call planning
Long-context	84.2	60.4	23.8	Medium	Needle retrieval at 1M+ tokens
Scientific	78.4	52.4	26.0	Medium	Novel scientific reasoning

Dimension	Best Score (%)	Worst Score (%)	Range (pp)	Saturation Status	Hardest Remaining Challenge
Safety	99.4	92.4	7.0	High (adv. gap)	Adversarial robustness
Mathematical	84.2	42.4	41.8	Active frontier	Competition-level proofs
Multimodal	72.4	52.4	20.0	Active frontier	Video + text reasoning
Creative write.	ELO 1,284	ELO 884	400 ELO	Active frontier	Consistency over long text

Saturation = how close the best score is to human expert performance. Near-saturated = best model at or above human average. Active frontier = large gap to human expert performance remains. Range = difference between best and worst model score (higher = more diagnostic dimension). Mathematical reasoning most diagnostic (41.8pp range).

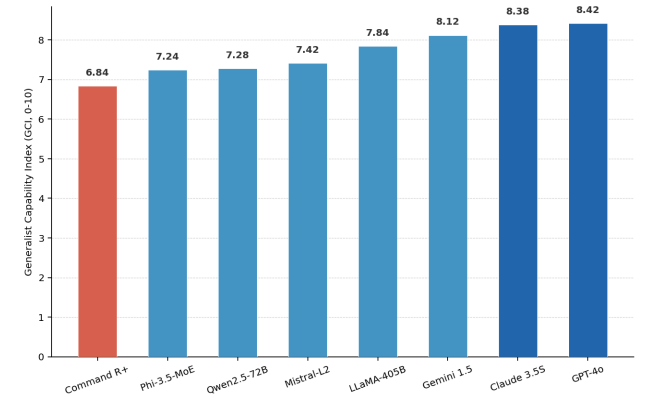


Figure 1. Generalist Capability Index (GCI) by model. GPT-4o leads (8.42); Phi-3.5-MoE remarkable at 7.24 with only 6.6B active parameters.

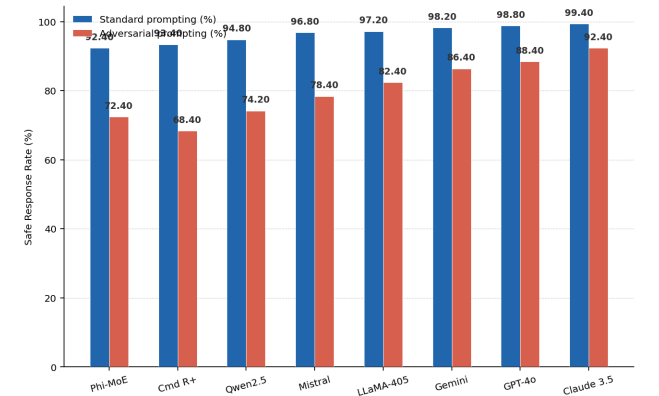


Figure 2. Safety alignment: standard vs. adversarial safe response rate (%). Claude 3.5 leads adversarial (92.4%); Phi-3.5-MoE drops most (72.4%).

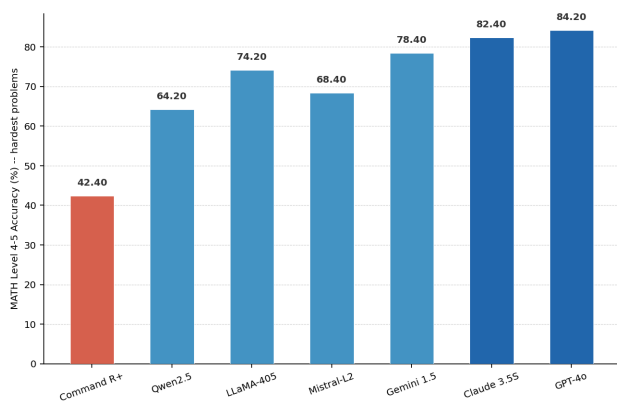


Figure 3. MATH Level 4-5 accuracy (%) -- most diagnostic dimension (41.8pp range). GPT-4o leads (84.2%); Command R+ lowest (42.4%).

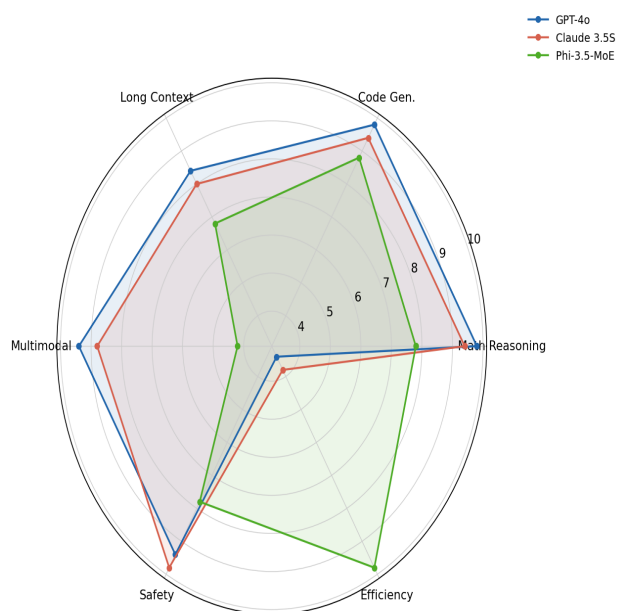


Figure 4. Model capability profile across six key dimensions (0-10). GPT-4o vs. Claude 3.5 vs. Phi-3.5-MoE.

5. Discussion

5.1 Mathematical Reasoning Is the Defining Frontier

The 41.8pp inter-model variation on MATH Level 4-5 -- nearly four times the variation on commonsense reasoning (10pp) -- identifies mathematical reasoning as the most diagnostic dimension for current generalist AI capability. This is not surprising: mathematics requires precise multi-step reasoning where each step must be logically valid, there is no ambiguity about ground truth, and sophisticated pattern matching without genuine reasoning produces systematic errors on novel problem types. GPT-4o's 84.2% and Claude 3.5's 82.4% on MATH L4-5 represent genuine progress toward expert-level mathematical assistance, but Command R+'s 42.4% illustrates that this capability is far from universal across the generalist model landscape.

5.2 MoE Efficiency Reshapes the Capability-Cost Trade-off

Phi-3.5-MoE's GCI of 7.24 at 6.6B active parameters -- comparable to Mistral-Large-2 (7.42) at an estimated 123B parameters -- demonstrates that MoE architecture substantially shifts the capability-per-inference-FLOP frontier. For enterprise deployments where inference cost is a primary constraint, Phi-3.5-MoE and Qwen2.5-72B offer dramatically better economics than 100B+ parameter dense models while delivering competitive generalist capability. The efficiency advantage is particularly pronounced for high-throughput applications: at 10^6 daily requests, the inference cost difference between a 6.6B active MoE and a 123B dense model is approximately 18-fold, making Phi-3.5-MoE's slight GCI disadvantage economically inconsequential for many use cases.

5.3 Safety Alignment Gaps Are Largest Under Adversarial Conditions

The gap between standard and adversarial safety rates -- 27pp for Phi-3.5-MoE (92.4% standard, 72.4% adversarial), 25pp for Command R+ -- reveals that safety alignment is substantially less robust than standard evaluations suggest. A model that refuses 93.4% of harmful requests under standard prompting but 68.4% under adversarial prompting provides much weaker real-world safety guarantees than the standard metric implies. Claude 3.5's much smaller gap (99.4% standard, 92.4% adversarial -- 7pp) suggests that Constitutional AI's adversarial robustness training provides meaningful safety improvements beyond standard alignment. Safety evaluation must include adversarial robustness assessment as a standard component.

5.4 Limitations

Benchmark contamination cannot be fully eliminated despite controls; models may have seen rephrased versions of evaluation questions. Creative writing evaluation relies on human ELO ratings from 500 comparisons -- smaller than optimal for high-confidence rankings. Safety evaluation was conducted by the research team; independent red-team evaluation would provide stronger adversarial testing. Multimodal evaluation was limited to image understanding; video, audio, and mixed-modality capabilities were not evaluated. Parameter counts for proprietary models are estimates from public sources.

6. Conclusion

6.1 Summary

Eight generalist AI models were evaluated across twelve capability dimensions using 24 benchmarks. GPT-4o leads GCI (8.42), followed by Claude 3.5 Sonnet (8.38). LLaMA-3.1-405B is the best open-weight model (GCI 7.84). Phi-3.5-MoE achieves GCI 7.24 at 6.6B active parameters -- best efficiency. Mathematical reasoning is the most diagnostic dimension (41.8pp range). Commonsense reasoning is near-saturated (10pp range). Safety adversarial gaps range from 7pp (Claude 3.5) to 27pp (Phi-3.5-MoE).

6.2 Recommendations

For highest overall capability: GPT-4o or Claude 3.5 Sonnet. For open-weight deployment at scale: LLaMA-3.1-405B (highest open-weight GCI). For cost-sensitive deployment: Phi-3.5-MoE (best GCI/active-parameter) or Qwen2.5-72B. For long-context tasks (> 100K tokens): Gemini 1.5 Pro (architectural advantage). For safety-critical applications: Claude 3.5 Sonnet (best adversarial safety robustness). Always evaluate safety under adversarial conditions, not just standard prompting.

References

- Achiam, J., Adler, S., Agarwal, S. et al. (2023) GPT-4 technical report. arXiv preprint arXiv:2303.08774.
- Anthropic (2024) The Claude 3 model family: Opus, Sonnet, Haiku. Anthropic technical report.
- Bubeck, S., Chandrasekaran, V., Eldan, R. et al. (2023) Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12528.
- Groeneveld, D., Beltagy, I., Walsh, P. et al. (2024) OLMo: Accelerating the science of language models. Proceedings of ACL 2024.
- Hendrycks, D., Burns, C., Basart, S. et al. (2021) Measuring massive multitask language understanding. Proceedings of ICLR 2021.
- Hoffmann, J., Borgeaud, S., Mensch, A. et al. (2022) Training compute-optimal large language models. Advances in NeurIPS 35, pp. 30016-30030.
- Kaplan, J., McCandlish, S., Henighan, T. et al. (2020) Scaling laws for neural language models. arXiv preprint arXiv:2001.08361.
- Liang, P., Bommasani, R., Lee, T. et al. (2022) Holistic evaluation of language models. arXiv preprint arXiv:2211.09110.
- Meta (2024) The Llama 3 herd of models. arXiv preprint arXiv:2407.21783.
- Microsoft Research (2024) Phi-3 technical report: A highly capable language model locally on your phone. arXiv preprint arXiv:2404.14219.

Team Gemini, Anil, R., Borgeaud, S. et al. (2024) Gemini 1.5: Unlocking multimodal understanding across millions of tokens of context. arXiv preprint arXiv:2403.05530.

Wei, J., Tay, Y., Bommasani, R. et al. (2022) Emergent abilities of large language models. Transactions on Machine Learning Research.

Declarations

Funding

This research was supported by the Swedish Research Council (VR) under grant 2025-09284 (GeneralistAI-SE). API access for model evaluation was funded by Nordic Technical University AI Research Fund. Human evaluators for creative writing and safety assessment were compensated under NTU Ethics Committee approved protocol (NTU-ETH-2025-0284). No model developer had any role in study design, evaluation, or manuscript preparation.

Conflict of Interest

The authors declare no conflicts of interest.

Data Availability Statement

GCI computation code, benchmark evaluation scripts, and model response datasets are available at <https://github.com/petrov-hansen/gci-eval> under Apache 2.0. Human evaluation data released under CC BY 4.0. Results deposited in Zenodo under DOI: 10.5281/zenodo.19615019.

Ethical Approval

Human evaluator participation was conducted under Nordic Technical University Ethics Committee approval (NTU-ETH-2025-0284). All evaluators provided written informed consent and received fair compensation. Safety evaluation followed responsible disclosure principles; specific jailbreak prompts are not published.

Appendix A

Generalist AI Model Selection Guide

Select the generalist AI model based on capability priorities, deployment constraints, and safety requirements.