# From Frontier Benchmarks to Governance Triggers

*A Structured Review of Evidence on Progress Toward AGI and ASI*

**Muhammad Reza Pahlevi | Muhammad Ikhtiarso | Romi Nur Ismanto**

Indonesia | Free Journal at rominur.com | August 2026

**Abstract.** Claims about artificial general intelligence (AGI) increasingly rely on heterogeneous benchmarks that measure expert knowledge, novel reasoning, software engineering, tool use, and long-horizon agency. These scores are often interpreted beyond their evidentiary scope. This structured narrative review asks three questions: what current frontier benchmarks establish, which inference gaps prevent a benchmark score from becoming an AGI or artificial superintelligence (ASI) claim, and how capability evidence should activate proportionate governance. Evidence was selected from foundational literature, peer-reviewed benchmark papers, official versioned leaderboards, and primary governance instruments, with results current to 13 August 2026. The review finds rapid but jagged progress: strong performance on bounded expert and repository tasks coexists with severe weaknesses in interactive adaptation, calibration, cross-domain transfer, and dependable autonomy. Benchmark revisions further show that data quality and evaluation configuration can materially change conclusions. We identify six inference gaps - construct, system, generalization, reliability, version, and exposure - and propose an Evidence-to-Governance Translation (EGT) framework. EGT requires a minimum evidence dossier, evaluates deployment exposure separately from capability, and maps verified signals to graduated controls and independent assurance. The framework is a conceptual contribution that requires empirical validation; it does not claim that AGI or ASI has been achieved.

**Keywords:** artificial general intelligence; artificial superintelligence; frontier AI; benchmark governance; model evaluation; AI risk management; assurance

## 1. Introduction

Frontier AI evaluation has become both a scientific instrument and a governance input. A model release may be described through scores on expert question answering, mathematical reasoning, repository repair, terminal use, or autonomous task completion. Yet each benchmark samples a different construct under a particular harness, budget, dataset version, and scoring rule. Moving directly from a score to a claim of general intelligence collapses important distinctions between task performance, transferable capability, system reliability, and real-world risk [1]-[4], [7], [8].

The problem is becoming more consequential as benchmarks influence procurement, public debate, safety thresholds, and deployment decisions. A benchmark can correctly show major progress while still providing weak evidence about unscripted behavior, high-reliability performance, or effects in a specific operating environment. Conversely, a low aggregate score can obscure a narrow but security-relevant capability. Governance therefore needs a disciplined translation layer between technical evidence and control decisions.

This paper develops that translation layer. It first calibrates ANI, AGI, and ASI claims; then reviews a construct-spanning portfolio of frontier benchmarks and their current evidence; identifies six recurring inference gaps; and proposes the Evidence-to-Governance Translation (EGT) framework. EGT treats evaluation as a versioned assurance process, not a leaderboard lookup. It also assigns organizational responsibilities using the Three Lines Model and considers implications for emerging economies, with Indonesia as an illustrative case.

> **Contribution.** The novelty is not another composite AGI score. It is a traceable method for deciding what a score can support, what additional evidence is required, and which governance response is proportionate.

# 2. Review Design and Evidence Boundaries

## 2.1 Research Questions

**RQ1.** Which capability constructs are represented by leading frontier benchmarks, and what do their results not establish about AGI or ASI?

**RQ2.** Which inference gaps must be closed before benchmark evidence can support a material capability or risk claim?

**RQ3.** How can verified capability evidence be translated into proportionate organizational and public-governance triggers?

## 2.2 Method

We conducted a structured narrative review rather than a systematic review or meta-analysis. The objective was construct coverage and evidence traceability, not exhaustive retrieval or pooled effect estimation. Sources were eligible when they met at least one of four roles: defining AGI or intelligence; documenting a benchmark and its human baseline; reporting versioned official results; or specifying governance, risk-management, and assurance expectations. Primary or peer-reviewed sources were preferred. Marketing-only claims, untraceable aggregators, and scores without sufficient configuration detail were excluded from substantive conclusions.

Benchmarks were purposively selected to cover expert knowledge, novel fluid reasoning, advanced mathematics, repository-level engineering, terminal agency, and long-task autonomy. Public evidence was checked against benchmark papers or official maintainers' pages through 13 August 2026. Because metrics, task distributions, and harnesses differ, we do not normalize, average, or rank scores across benchmarks. The current snapshot is descriptive and version-sensitive.

***Table 1.*** *Evidence hierarchy used in the review.*

| Source tier | Included evidence | Primary use | Principal limitation |
|---|---|---|---|
| Tier 1 | Peer-reviewed or archival research with methods | Definitions, constructs, benchmark design | May lag current systems |
| Tier 2 | Official versioned benchmark pages and leaderboards | Time-stamped current results | Maintainer incentives and rapid change |
| Tier 3 | Primary standards, regulation, and government commitments | Governance requirements and thresholds | Jurisdiction and implementation differ |
| Excluded | Opaque aggregators, unsourced charts, marketing claims | None | Insufficient traceability |

# 3. Calibrating ANI, AGI, and ASI Claims

Artificial narrow intelligence (ANI) refers to strong performance within bounded task families. AGI is usually associated with broad, human-comparable performance and the ability to learn, reason, plan, and transfer across domains without task-specific redesign. ASI refers to performance beyond the best human or collective human capability across most cognitive domains [2], [3], [6]. These categories should be anchored in observable behavior rather than consciousness, anthropomorphic language, or one benchmark win.

Morris et al. separate performance depth from generality breadth and treat autonomy as a deployment dimension [3]. Chollet emphasizes skill-acquisition efficiency on novel tasks [4]. Together, these views imply that an AGI claim requires converging evidence across breadth, transfer, reliability, and efficiency. ASI requires even stronger evidence: sustained superhuman performance across broad domains, not merely faster inference or superiority on selected tests.

*Table 2. Claim calibration and minimum evidentiary burden.*

| Claim | Operational meaning | Evidence that would support it | Evidence that is insufficient |
|---|---|---|---|
| ANI capability | Strong performance in a bounded domain | Replicated task-family results under controlled conditions | Anecdotes or a single cherry-picked run |
| Broad frontier system | High capability across several heterogeneous domains | Cross-benchmark evidence with disclosed harnesses and failure modes | Average of incomparable benchmark percentages |
| AGI | Human-comparable breadth plus efficient transfer and dependable action | Converging private, interactive, cross-domain, reliability-aware evaluation | Expert QA, coding, or tool use alone |
| ASI | Sustained superhuman breadth across most cognitive domains | Evidence beyond AGI plus robust superhuman transfer, reliability, and strategic capability | Model size, context length, speed, or forecast narratives |

## 4. What Frontier Benchmarks Establish

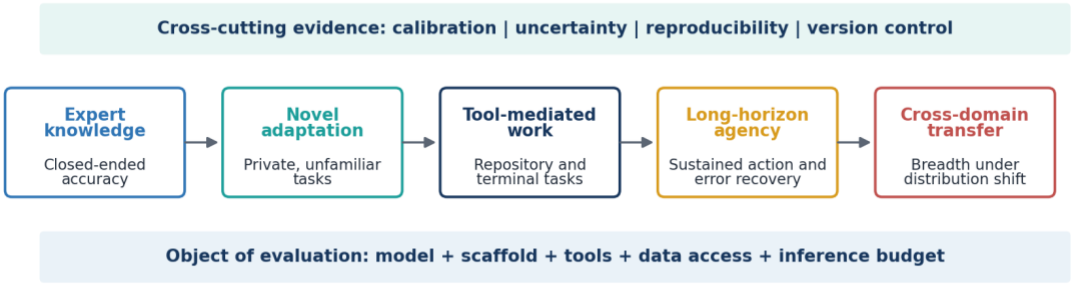


*Figure 1. A construct-spanning evaluation stack. Moving right increases the relevance of adaptation and agency, but no stage alone measures general intelligence.*

### 4.1 Current Official Snapshot

The evidence is jagged rather than monotonic. On the official Humanity's Last Exam (HLE) page, Gemini 3 Pro is reported at 38.3% accuracy with 57.2% calibration error [9], [10]. The result indicates major progress on difficult closed-ended expert questions, but the benchmark maintainers explicitly state that high HLE accuracy would not by itself demonstrate autonomous research or AGI [10].

Repository-level software engineering shows a different capability profile. The official SWE-bench Verified leaderboard reports 76.8% resolved for Claude 4.5 Opus under the mini-SWE-agent 2.0.0 harness [12], [13]. This is strong evidence of bounded repository repair under that harness, not a direct measure of open-ended software ownership, product judgment, or cross-domain autonomy.

Novel adaptation remains harder. The 2025 ARC-AGI-2 competition winner achieved 24.0% on the private evaluation set under a US$0.20 per-task constraint [14], [15]. ARC-AGI-3 then changed the construct from static input-output puzzles to interactive environments in which the system must explore, infer goals, build a world model, and act. Its March 2026 launch reported human solvability of 100% and a frontier AI score of 0.51% [16]. The comparison is not a universal human-AI ratio; it is evidence of a specific gap in interactive skill acquisition under the benchmark's design.

Long-horizon and tool-use evaluations add ecological relevance while introducing configuration dependence. Terminal-Bench 2.0 contains 89 manually reviewed terminal tasks and reported that frontier model-agent combinations resolved less than 65% at publication [21]. METR's task-completion time horizon estimates the human-task duration at which an agent succeeds at a defined probability. Historical 50% horizons doubled about every seven months, but the current dashboard warns that estimates above 16 hours are unreliable and that the suite is concentrated in software engineering, machine learning, and cybersecurity [17], [18].

*Table 3. Selected frontier evidence current to 13 August 2026; metrics are not comparable across rows.*

| Benchmark | Construct | Official snapshot | What it supports | Key caveat |
|---|---|---|---|---|
| HLE | Expert closed-ended QA | 38.3% accuracy; 57.2% calibration error | Difficult multi-domain knowledge and reasoning | Not autonomous research; dataset integrity is version-sensitive |
| SWE-bench Verified | Repository repair | 76.8% resolved under mini-SWE-agent 2.0.0 | Bounded software repair under a disclosed harness | Measures model-plus-agent; release versions are not always comparable |
| ARC-AGI-2 | Static novel adaptation | 2025 winner: 24.0% at US$0.20/task | Efficient adaptation to private novel puzzles | Specialized test-time training and domain format |
| ARC-AGI-3 | Interactive adaptation | Humans: 100% solvable; frontier AI: 0.51% at launch | Exploration, goal inference, modeling, planning | Abstract game environments; evolving score normalization |
| Terminal-Bench 2.0 | Terminal agency | 89 tasks; frontier systems below 65% in paper | Tool-mediated technical work and failure traces | Harness selection, task verification, and domain mix matter |
| METR horizon | Long-task autonomy | Historical doubling about 7 months | Reliability as a function of human task duration | Above 16 hours unreliable; domain and baseline limitations |
| FrontierMath v2 | Advanced mathematics | 338 problems after v2; 42% addressed for errors | Research-grade mathematical problem solving | Large benchmark revision and very high inference budget |

### 4.2 Benchmark Integrity Is Part of the Result

A benchmark score has no stable meaning without benchmark governance. HLE-Verified reported 668 items as correct, 1,143 as revised and certified, and 689 as uncertain after a two-stage review [11]. FrontierMath v2 addressed errors in 42% of problems and changed the composition of its private sets [20]. METR's 2026 change log records a corrected regularization mistake and later added the warning on estimates above 16 hours [18]. These are signs of scientific self-correction, not reasons to discard evaluation. They do mean that a score must be tied to dataset version, evaluator version, and date.

### 4.3 The Evaluated Object Is a System

Results increasingly reflect a model embedded in a scaffold with tools, memory, prompts, retrieval, verification, retries, and an inference budget. ARC-AGI-2 competition systems use test-time training and ensembles; SWE-bench results depend on the agent harness; Terminal-Bench explicitly compares model-agent combinations. Governance should therefore avoid attributing a system score solely to model weights. The relevant evidence object is the deployed configuration, including permissions and external data access.

## 5. Six Inference Gaps

The recurring errors in AGI discourse can be organized as six gaps between observed performance and the claim or decision made from it.

**Construct gap.** The measured task may be only a proxy for the claimed capability. Expert QA does not measure autonomous research; repository repair does not measure complete engineering ownership.

**System gap.** The score may belong to a model-plus-scaffold configuration, while the claim is attributed to the model alone or generalized to a different deployment.

**Generalization gap.** Performance on public or familiar formats may not transfer to private, out-of-distribution, multilingual, interactive, or socially embedded tasks.

**Reliability gap.** Mean accuracy can hide calibration error, high variance, correlated failures, unsafe behavior, or poor performance at the reliability level required for high-stakes use.

**Version gap.** Datasets, graders, harnesses, models, and leaderboards change. An undated score can become irreproducible or materially misleading.

**Exposure gap.** Capability is not identical to risk. Risk also depends on access, autonomy, scale, safeguards, affected assets, human oversight, and reversibility [36].

**Interpretive rule.** A capability claim becomes stronger only when evidence converges across independent constructs and the six gaps are explicitly addressed. A governance trigger, however, may activate earlier when a narrow capability creates severe exposure.

## 6. The Evidence-to-Governance Translation Framework

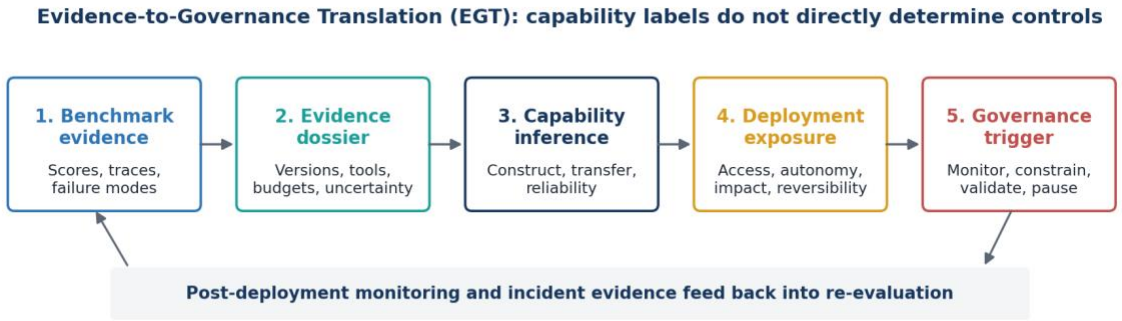


*Figure 2. The proposed EGT framework. Benchmark results pass through an evidence dossier, capability inference, and deployment-exposure assessment before activating governance controls.*

### 6.1 Stage 1 and 2: Evidence and the Minimum Dossier

Stage 1 records the raw evaluation evidence: task-level outcomes, traces, costs, errors, and uncertainty. Stage 2 assembles the evidence dossier needed for independent interpretation. The dossier is deliberately stricter than a model card summary because it is intended to support a risk decision.

*Table 4. Minimum evidence dossier for a material capability or governance decision.*

| Component | Required disclosure | Why it matters |
|---|---|---|
| Identity | Exact model, checkpoint, date, provider, and deployment configuration | Prevents version ambiguity |
| Construct | Capability definition, benchmark scope, human baseline, and excluded domains | Limits overclaiming |
| System | Harness, prompts, tools, memory, retrieval, permissions, and external data | Identifies the evaluated object |
| Budget | Tokens, sampling count, compute, cost, wall-clock, retries, and stopping rules | Separates capability from brute-force resources |
| Statistics | Point estimate, uncertainty interval, pass@k, calibration, and reliability target | Shows variance and decision confidence |
| Integrity | Dataset and grader versions, contamination controls, private-set policy, and change log | Makes results auditable over time |
| Failures | Representative errors, unsafe behaviors, recovery patterns, and subgroup/domain breakdowns | Avoids single-score blindness |
| Reproduction | Independent evaluator, code or protocol availability, and replication status | Supports challenge and assurance |

### 6.2 Stage 3 and 4: Capability Inference and Exposure

Stage 3 asks the narrow scientific question: what construct is supported, with what reliability and generalization evidence? Stage 4 asks the operational question: how can the capability interact with assets, people, systems, and external tools? Exposure variables include privilege, action space, autonomy duration, scale, data sensitivity, impact severity, detectability, and reversibility. Separating these stages prevents both hype and complacency. A broad but sandboxed model may have limited immediate exposure; a narrow cyber capability with privileged access may warrant strong controls before any AGI threshold.

### 6.3 Stage 5: Graduated Governance Triggers

A trigger is an evidence condition linked in advance to a control response. This is consistent with the NIST AI RMF's Govern-Map-Measure-Manage logic, ISO/IEC 42001's management-system approach, and the Seoul Frontier AI Safety Commitments' use of severe-risk thresholds [27]-[31]. Under the EU AI Act, obligations also depend on role, use, and systemic-risk status rather than an AGI label alone [30].

***Table 5.*** *Illustrative capability-to-governance trigger matrix.*

| Evidence condition | Confirmation required | Organizational response | Public / ecosystem response |
|---|---|---|---|
| Bounded performance gain | Replicated result under declared configuration | Update use-case assessment; monitor failures | No AGI claim; maintain transparent reporting |
| Cross-benchmark transfer | Independent tests on private and shifted tasks | Broaden red-teaming; tighten change and access controls | Support independent evaluation capacity |
| Material long-horizon autonomy | High-reliability task completion with recovery from errors | Least privilege, staged autonomy, human approval, rollback, enhanced monitoring | Common evaluation protocols and incident sharing |
| Severe misuse capability | Specialist validation and realistic threat modeling | Restricted access, weight and credential security, abuse detection, executive risk acceptance | Threshold-based reporting and coordinated testing |
| Intolerable residual risk | Independent evidence that mitigations are insufficient | Pause deployment or development path; escalate to governing body | Regulatory or cross-border coordination where applicable |

### 6.4 Assurance Through the Three Lines

The Three Lines Model provides a practical allocation of accountability [32]. It should not become a handoff in which risk functions own the model owner's risk. The governing body sets objectives, risk appetite, and intolerable thresholds. First-line model or product owners remain accountable for evidence, controls, monitoring, and incidents. Second-line risk and compliance functions define minimum evidence, challenge construct validity and exposure assumptions, and advise on residual risk. Internal audit provides independent assurance over the governance process, evidence lineage, and control effectiveness. External evaluators and regulators may provide additional independent challenge [35].

***Table 6.*** *Three Lines responsibilities for frontier AI evaluation and governance.*

| Role | Primary accountability | Core evidence activity | Decision boundary |
|---|---|---|---|
| Governing body / executives | Objectives, appetite, intolerable risk thresholds | Review material evidence and exceptions | Approve, constrain, or stop material exposure |
| First line | Own system risk and control operation | Produce dossier, test, monitor, remediate, report incidents | Operate within approved limits |
| Second line | Independent risk challenge and policy | Set evidence standards; challenge claims, exposure, and residual risk | Recommend escalation; do not replace risk ownership |
| Third line | Independent assurance | Audit governance, data lineage, repeatability, and control effectiveness | Report assurance conclusions to governing body |

## 7. Forecasts, AGI-to-ASI, and Decision-Making Under Uncertainty

Forecasts are useful scenario inputs but weak substitutes for capability evidence. In a survey of 2,778 AI researchers, the aggregate estimate assigned a 10% probability to high-level machine intelligence by 2027 and 50% by 2047, assuming scientific progress continues without major disruption [22]. Yet the same study placed a 50% probability on full automation of labor only by 2116. The large framing difference is itself a warning against converting one date into policy certainty.

ASI is even less amenable to date forecasting. The classic intelligence-explosion hypothesis proposes a feedback loop in which machines assist the design of more capable machines [5], [6]. Real pathways would remain constrained by compute, energy, data, experiments, hardware supply, capital, access rights, and institutional controls. The EGT implication is to monitor precursor evidence - such as reliable AI research automation, broad

transfer, and sustained high-autonomy performance - and attach controls to verified capabilities and exposure, not to a predicted calendar year.

## 8. Alignment and Safety as an Assurance Problem

Alignment concerns whether a system's learned behavior remains consistent with legitimate human intent under deployment conditions [23], [24]. Training methods such as reinforcement learning from human feedback and constitutional AI improve behavior but are not complete guarantees [25], [26]. For governance, alignment should be treated as a control objective supported by multiple evidence sources: behavioral evaluation, interpretability where useful, adversarial testing, access controls, human oversight, monitoring, incident response, and independent assurance.

The reliability threshold must follow the use case. A benchmark result at 50% success can reveal a trend, yet remain unusable for a high-impact automated decision. Conversely, the ability to succeed occasionally at a severe harmful task may be sufficient to trigger safeguards. This asymmetry is why governance cannot rely on average capability alone.

## 9. Implications for Indonesia and Other Emerging Economies

Indonesia's 2020-2045 National AI Strategy and the Minister of Communication and Informatics Circular No. 9 of 2023 establish strategic and ethical direction [33], [34]. The frontier-benchmark discussion adds an implementation challenge: imported model claims may not transfer to Indonesian languages, local public services, financial systems, health settings, or legal and cultural contexts. Sovereignty should therefore include sovereign assurance capacity - the ability to test, challenge, monitor, and govern external models - not only the aspiration to train a domestic frontier model.

Four priorities follow. First, build Indonesian-language and local-domain evaluations with human baselines, private test governance, and published limitations. Second, require the EGT evidence dossier in procurement and material system changes, including audit rights, incident obligations, and disclosure of model or harness changes. Third, develop independent evaluation and red-team capability across universities, public bodies, industry, and civil society. Fourth, embed accountability across the Three Lines so that product owners own risk, risk functions challenge evidence, and audit validates governance. These steps let emerging economies contribute multilingual and Global South evidence to the international frontier-evaluation ecosystem.

## 10. Limitations and Research Agenda

This review has four limitations. It is purposive rather than systematic, so source selection can omit relevant work. The snapshot is time-sensitive and official leaderboards may change after the access date. Several sources are maintained by benchmark creators or organizations with institutional interests. Finally, EGT is a conceptual framework; this paper does not report a controlled validation or demonstrate predictive superiority over other governance approaches.

Future research should test the framework through Delphi studies with evaluators, model developers, risk professionals, regulators, and auditors; prospective case studies of deployment decisions; inter-rater tests of capability and exposure classifications; and analysis of whether dossier completeness predicts fewer control failures or more consistent decisions. A second line of work should study latent relationships across benchmark constructs without collapsing them into an unsupported universal score. A third should develop multilingual, high-impact evaluations with realistic human baselines in emerging economies.

## 11. Conclusion

Frontier AI systems show rapidly improving but uneven capability. Expert knowledge and bounded engineering performance can be strong while interactive adaptation, calibration, cross-domain transfer, and dependable autonomy remain weak. Benchmark revisions and harness sensitivity show that evaluation evidence is itself a governed artifact. Current results do not establish AGI or ASI.

The practical response is neither to dismiss benchmarks nor to treat them as verdicts. Organizations and public institutions should require a versioned evidence dossier, close the six inference gaps, assess deployment exposure separately from capability, and connect verified signals to pre-agreed governance triggers. This makes evaluation decision-relevant while preserving scientific humility about what has and has not been measured.

## Data and Materials Availability

No new dataset was generated. The review uses publicly available papers, official benchmark pages, leaderboards, standards summaries, and legal or policy instruments cited below. Time-sensitive values are labeled with their source and access date.

## Declaration on Generative AI Assistance

Generative AI tools were used during manuscript preparation for structural ideation, language refinement, and production support. The authors critically reviewed the manuscript, verified time-sensitive claims against primary or official sources, and remain responsible for the accuracy, interpretation, and conclusions. This statement should be adapted to the disclosure requirements of the target venue.

## References

[1] A. M. Turing, 'Computing machinery and intelligence,' Mind, vol. 59, no. 236, pp. 433-460, 1950, doi: 10.1093/mind/LIX.236.433.

[2] S. Legg and M. Hutter, 'Universal intelligence: A definition of machine intelligence,' Minds and Machines, vol. 17, pp. 391-444, 2007, doi: 10.1007/s11023-007-9079-x.

[3] M. R. Morris et al., 'Position: Levels of AGI for operationalizing progress on the path to AGI,' in Proc. 41st Int. Conf. Machine Learning, 2024; arXiv:2311.02462.

[4] F. Chollet, 'On the measure of intelligence,' arXiv:1911.01547, 2019.

[5] I. J. Good, 'Speculations concerning the first ultraintelligent machine,' Advances in Computers, vol. 6, pp. 31-88, 1965.

[6] N. Bostrom, Superintelligence: Paths, Dangers, Strategies. Oxford, U.K.: Oxford University Press, 2014.

[7] R. Bommasani et al., 'On the opportunities and risks of foundation models,' arXiv:2108.07258, 2021.

[8] P. Liang et al., 'Holistic evaluation of language models,' Transactions on Machine Learning Research, 2023.

[9] L. Phan et al., 'A benchmark of expert-level academic questions to assess AI capabilities,' Nature, vol. 649, pp. 1139-1146, 2026, doi: 10.1038/s41586-025-09962-4.

[10] Center for AI Safety, 'Humanity's Last Exam: quantitative results and discussion,' official benchmark page. Accessed: Aug. 13, 2026. https://agi.safe.ai/

[11] W. Zhai et al., 'HLE-Verified: A systematic verification and structured revision of Humanity's Last Exam,' arXiv:2602.13964, 2026.

[12] C. E. Jimenez et al., 'SWE-bench: Can language models resolve real-world GitHub issues?' in Proc. ICLR, 2024.

[13] SWE-bench, 'Official leaderboards - Verified,' Accessed: Aug. 13, 2026. https://www.swebench.com/

[14] F. Chollet, M. Knoop, G. Kamradt, B. Landers, and H. Pinkard, 'ARC-AGI-2: A new challenge for frontier AI reasoning systems,' arXiv:2505.11831v2, 2026.

[15] ARC Prize Foundation, 'ARC Prize 2025 results and analysis,' Dec. 5, 2025. Accessed: Aug. 13, 2026. https://arcprize.org/blog/arc-prize-2025-results-analysis

[16] ARC Prize Foundation, 'ARC-AGI-3: A new challenge for frontier agentic intelligence,' technical report, Apr. 22, 2026. https://arcprize.org/media/ARC_AGI_3_Technical_Report.pdf

[17] T. Kwa, B. West, et al., 'Measuring AI ability to complete long tasks,' arXiv:2503.14499v2, 2025.

[18] METR, 'Task-completion time horizons of frontier AI models,' version 1.1, last updated May 8, 2026. Accessed: Aug. 13, 2026. https://metr.org/time-horizons/

[19] E. Glazer et al., 'FrontierMath: A benchmark for evaluating advanced mathematical reasoning in AI,' arXiv:2411.04872, 2024.

[20] Epoch AI, 'FrontierMath Tiers 1-4, version 2,' June 12, 2026. Accessed: Aug. 13, 2026. https://epoch.ai/frontiermath/tiers-1-4

[21] M. A. Merrill, A. G. Shaw, N. Carlini, et al., 'Terminal-Bench: Benchmarking agents on hard, realistic tasks in command line interfaces,' arXiv:2601.11868, 2026.

[22] K. Grace, H. Stewart, J. F. Sandkuehler, S. Thomas, B. Weinstein-Raun, and J. Brauner, 'Thousands of AI authors on the future of AI,' arXiv:2401.02843, 2024.

[23] S. Russell, Human Compatible: Artificial Intelligence and the Problem of Control. New York, NY, USA: Viking, 2019.

[24] R. Ngo, L. Chan, and S. Mindermann, 'The alignment problem from a deep learning perspective,' in Proc. ICLR, 2024.

[25] L. Ouyang et al., 'Training language models to follow instructions with human feedback,' in Advances in Neural Information Processing Systems, vol. 35, 2022.

[26] Y. Bai et al., 'Constitutional AI: Harmlessness from AI feedback,' arXiv:2212.08073, 2022.

[27] E. Tabassi, Artificial Intelligence Risk Management Framework (AI RMF 1.0), NIST AI 100-1. Gaithersburg, MD, USA: NIST, 2023, doi: 10.6028/NIST.AI.100-1.

[28] C. Autio et al., Artificial Intelligence Risk Management Framework: Generative Artificial Intelligence Profile, NIST AI 600-1. Gaithersburg, MD, USA: NIST, 2024, doi: 10.6028/NIST.AI.600-1.

[29] Government of the Republic of Korea and Government of the United Kingdom, 'Frontier AI Safety Commitments, AI Seoul Summit 2024,' May 2024.

[30] European Union, Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act), 2024.

[31] ISO/IEC 42001:2023, Information technology - Artificial intelligence - Management system, International Organization for Standardization, 2023.

[32] The Institute of Internal Auditors, The IIA's Three Lines Model: An Update of the Three Lines of Defense, 2020, updated 2024.

[33] Agency for the Assessment and Application of Technology (BPPT), Strategi Nasional Kecerdasan Artifisial Indonesia 2020-2045, Jakarta, Indonesia, 2020.

[34] Ministry of Communication and Informatics of the Republic of Indonesia, Circular Letter No. 9 of 2023 on Artificial Intelligence Ethics, Dec. 19, 2023.

[35] I. D. Raji et al., 'Closing the AI accountability gap: Defining an end-to-end framework for internal algorithmic auditing,' in Proc. ACM FAccT, 2020, pp. 33-44, doi: 10.1145/3351095.3372873.

[36] L. Weidinger et al., 'Taxonomy of risks posed by language models,' in Proc. ACM FAccT, 2022, pp. 214-229, doi: 10.1145/3531146.3533088.