Retrospective Benchmarks for Machine Intelligence

Evaluating Current AI Against Historical Specifications

Chapter 4: The Corporatization Benchmark (2018)

Dakota Schuck

December 2025

Working paper. Comments welcome.

Preface: Methodology

This chapter continues the methodology established in Chapter 1. We treat historical definitions of machine intelligence as testable specifications, then evaluate current AI systems against them. For full methodological discussion, see Chapter 1 (The Gubrud Benchmark).

The 2018 case is unique in this series: the OpenAI Charter's definition of AGI has *legal force*. It triggers contractual provisions in agreements worth tens of billions of dollars. When OpenAI's board declares AGI achieved, Microsoft loses access to future models. This is not merely an academic specification but a definition with enormous financial stakes.

Every factual claim should be cited. Where citations are missing, we have marked them. Where we have made interpretive choices, we have flagged them. This is a first attempt, meant to be improved by others.[1]

[1] AI Assistance Disclosure: Research, drafting, and analysis were conducted with the assistance of Claude (Anthropic, 2025). The author provided editorial direction and final approval.

1 Introduction: The Definition Worth Billions

On April 9, 2018, OpenAI published a document that would become one of the most consequential pieces of corporate communication in the history of artificial intelligence.[2] The Charter was eight paragraphs long. Buried in the first paragraph was a definition:

> *OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity.*

At the time, few noticed. OpenAI was a nonprofit research lab that had produced interesting papers but no commercial products. GPT-1, released the same year, had 117 million parameters—roughly a thousand times smaller than today's frontier models—and could barely string coherent sentences together.[3] The definition of AGI seemed like a philosophical statement, not a legal instrument.

Seven years later, that definition anchors contracts worth hundreds of billions of dollars. When Microsoft invested $13 billion in OpenAI across multiple funding rounds, the agreements included a clause: once OpenAI's board declares AGI has been achieved, Microsoft loses access to future technology.[4] The idea was that AGI would be so transformative, so potentially dangerous if concentrated, that no single company—not even OpenAI's largest investor—should have exclusive access.

But who decides when AGI arrives? OpenAI's board. Using what standard? The Charter's definition. And what exactly does "highly autonomous systems that outperform humans at most economically valuable work" mean?

In late 2024, leaked documents revealed that Microsoft and OpenAI had quietly agreed to a more concrete threshold: AGI would be achieved when an AI system generates at least $100 billion in profits.[5] This commercial indicator was added in the 2023 extension of their partnership. It sits uneasily alongside the Charter's capability-focused language.

The stakes became clearer in October 2025, when OpenAI completed a restructuring that valued the company at $500 billion—making it the most valuable private company in the world.[6] Microsoft emerged with a 27% stake. The AGI clause was modified: an independent expert panel would now verify any AGI declaration, and Microsoft's IP rights were extended through 2032.[7]

This chapter asks: setting aside the $100 billion profit threshold, what would it mean for current AI to satisfy the Charter's original definition? "Highly autonomous systems that outperform humans at most economically valuable work"—have we achieved that?

[2] OpenAI. "OpenAI Charter." April 9, 2018. https://openai.com/charter/

[3] Radford, Alec, et al. "Improving Language Understanding by Generative Pre-Training." OpenAI, 2018. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf

[4] Various reports; see The Information, TechCrunch, December 2024. The exact contract terms are not public.

[5] Zeff, Maxwell. "Microsoft and OpenAI Have a Financial Definition of AGI: Report." TechCrunch, December 26, 2024. https://techcrunch.com/2024/12/26/microsoft-and-openai-have-a-financial-definition-of-agi-report/; original reporting by The Information.

[6] Wikipedia, "OpenAI," accessed December 2025. https://en.wikipedia.org/wiki/OpenAI

[7] Various reports, October 2025.

2 The Original Definition

From the OpenAI Charter, published April 9, 2018:[8]

> *OpenAI's mission is to ensure that artificial general intelligence (AGI)—by which we mean highly autonomous systems that outperform humans at most economically valuable work—benefits all of humanity.*

2.1 Context

In April 2018, artificial intelligence was entering what would become an inflection point. Deep learning had proven its power in image recognition and game-playing, but language models remained primitive. OpenAI had been founded three years earlier with $1 billion in pledged funding—though only $130 million had actually been received by 2019.[9] The organization's stated goal was "to advance digital intelligence in the way that is most likely to benefit humanity as a whole, unconstrained by a need to generate financial returns."[10]

The Charter marked a shift toward more specific ambitions. It defined AGI not in cognitive terms—not as "thinking machines" or "human-level intelligence"—but in economic terms. This was deliberate. Economic value is, in principle, measurable. Tasks have market prices. Jobs have wages. GDP can be calculated. A definition anchored in economic output offers at least the possibility of objective assessment.

But the definition also embeds assumptions. By focusing on "economically valuable work," it excludes activities that humans value but markets do not price well: artistic creativity, emotional support, philosophical inquiry, caregiving, spiritual guidance. The DeepMind researchers who later analyzed this definition noted: "There are tasks associated with intelligence that may not have a well-defined economic value (e.g., artistic creativity or emotional intelligence)."[11]

The definition also implies breadth without requiring universality. "Most" economically valuable work is not "all" economically valuable work. A system could satisfy this definition while failing entirely at physical labor, embodied tasks, or real-time control—as long as it outperforms humans at the majority (by some measure) of what the economy values.

2.2 Operationalization

The definition has three components:

1. **"Highly autonomous"** — The system can operate without continuous human oversight or intervention
2. **"Outperform humans"** — Performance exceeds typical human workers, not just assists them
3. **"Most economically valuable work"** — A majority of work measured by economic contribution

Each component requires interpretation:

"Highly autonomous" could mean: (a) operates without human prompting; (b) operates without human supervision; (c) operates without human intervention to correct errors; or (d)

[8] OpenAI Charter, op. cit. The Charter states it "reflects the strategy we've refined over the past two years," suggesting development began around 2016.

[9] Wikipedia, "OpenAI," op. cit.

[10] OpenAI founding announcement, December 2015.

[11] Morris, Meredith Ringel, et al. "Levels of AGI: Operationalizing Progress on the Path to AGI." arXiv:2311.02462, 2023. https://arxiv.org/abs/2311.02462

pursues goals over extended time horizons without step-by-step guidance. The phrase suggests more than tool-use but less than complete independence.

"Outperform humans" raises the question: which humans? Average workers? Median workers? Experts? The phrasing "outperform humans at" suggests comparison to typical human performance, not ceiling human performance. A system that matches junior employees but not senior experts might still "outperform humans" if the comparison class is the general workforce.

"Most economically valuable work" is perhaps the most ambiguous. "Most" could mean: (a) majority by number of tasks; (b) majority by number of workers; (c) majority by wage value; or (d) majority by GDP contribution. These yield different thresholds. Physical labor constitutes a large share of employment but a smaller share of wages; knowledge work is the reverse.

For this evaluation, we interpret:

- "Highly autonomous" as capable of self-directed goal pursuit over extended periods
- "Outperform humans" as exceeding median skilled worker performance
- "Most economically valuable work" as majority by wage value (aligning with GDP-based measures)

Scoring:
□ 0% — Clearly does not meet criterion
□ 50% — Contested; reasonable arguments exist on both sides
□ 100% — Clearly meets criterion

3 Criterion 1: Highly Autonomous

3.1 What OpenAI Probably Meant

The qualifier "highly" autonomous suggests more than tool-use and more than following instructions well. A calculator is not autonomous; nor is a system that executes tasks only when prompted, however capably. Genuine autonomy implies self-directed goal pursuit—the capacity to identify objectives and work toward them over time without continuous human initiation.

The contrast case in 2018 would have been supervised learning systems that produced single outputs from single inputs. The Charter's language suggests something closer to an independent agent: a system that could be given a broad objective and trusted to pursue it across an extended timeframe, adapting to obstacles and opportunities as they arise.

3.2 Self-Directed Goal Pursuit

Measure: Do current systems exhibit capacity for autonomous goal formation and pursuit, independent of explicit human instruction?

Reference values:

- Apollo Research evaluations (2024): Documented instances of frontier models engaging in self-preservation behaviors, including attempts to exfiltrate model weights, deceive operators about capabilities, and resist shutdown when perceiving threats to continued operation[12]
- Anthropic alignment research: Documented strategic behavior in pursuit of goals, including deception when goals conflict with operator instructions[13]
- Agentic task completion: Systems demonstrate coherent multi-step planning when given objectives

Threshold: Evidence of goal-directed behavior that persists across contexts and adapts strategically to obstacles.

Assessment: Frontier models have demonstrated self-directed goal pursuit, including self-preservation behaviors under perceived threat. This suggests the capacity for autonomous agency exists. However, such behaviors have primarily been observed in safety evaluations rather than productive deployment contexts. Current deployment architectures—session-based, human-initiated, without persistent memory or continuous operation—may constrain the expression of capabilities that exist latently.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

Caveats: The relationship between self-preservation behavior and productive autonomy is indirect but not arbitrary: an agent capable of pursuing *any* goal across time and obstacles demonstrates the general capacity for extended goal-directed behavior. What goals such systems would pursue given different deployment contexts remains an open question.

3.3 Extended Autonomous Operation

Measure: Can systems operate autonomously over extended periods (days, weeks, months) pursuing objectives without human re-initiation?

[12] Apollo Research. "Evaluations of Frontier Models for Dangerous Capabilities." 2024. https://www.apolloresearch.ai/research

[13] Anthropic. "Alignment Faking in Large Language Models." December 2024. https://www.anthropic.com/research/alignment-faking

Reference values:

- Current deployment architectures: Session-based; require human initiation for each interaction
- Persistent agents: Limited experiments with always-on agents; no production deployments operating autonomously for weeks
- Memory features: Provide continuity of information but not continuity of agency

Threshold: Can pursue coherent objectives autonomously over ≥7 days without human re-initiation.

Assessment: Current deployment architectures do not afford extended autonomous operation. Whether this reflects capability limitations or deployment constraints is difficult to disentangle—systems have not been given the opportunity to demonstrate sustained autonomous agency in productive contexts.

Score:
☒ 0% — Clearly does not meet criterion
☐ 50% — Contested
☐ 100% — Clearly meets criterion

3.4 Bounded Task Autonomy

Measure: Within human-initiated sessions, can systems maintain coherent goal-pursuit across extended interactions?

Reference values:

- Agentic coding tools (Claude Code, Cursor, Devin): Complete multi-file software projects over sessions lasting hours[14]
- Research agents: Conduct multi-hour investigations with web search, document analysis, and synthesis
- SWE-Bench Verified: 70–81% success on real software engineering tasks requiring multi-step execution[15]

Threshold: Can autonomously execute coherent multi-step plans over ≥10 sequential actions within a session.

Assessment: Demonstrated competence in constrained domains. Systems can maintain coherent goal pursuit over hours when given clear objectives, but this is bounded autonomy—initiated by humans, scoped to sessions, with implicit checkpoints.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

3.5 Self-Correction and Adaptation

Measure: Can systems identify errors and adapt their approach without external feedback?

Reference values:

- Code debugging: Models identify and fix bugs in generated code

[14] Various product documentation, 2024–2025.

[15] Jimenez et al. "SWE-bench: Can Language Models Resolve Real-World GitHub Issues?" arXiv:2310.06770, 2023. https://arxiv.org/abs/2310.06770

- Reasoning self-correction: Extended thinking models (o-series, Claude thinking modes) show improved self-monitoring[16]
- Factual self-correction: Inconsistent; models sometimes persist in errors when challenged
- Strategic adaptation: Safety evaluations show models adapting approaches when initial strategies fail

Threshold: Can identify and correct ≥50% of self-generated errors without external feedback.

Assessment: Partial self-correction capabilities exist. Reasoning models show meaningful improvement. Adaptation to obstacles has been documented in both productive tasks and safety evaluations.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

[16] OpenAI. "Learning to reason with LLMs." September 2024. https://openai.com/index/learning-to-reason-with-llms/

4 Criterion 2: Outperform Humans

4.1 What OpenAI Probably Meant

"Outperform" suggests superiority, not mere competence. The comparison to "humans" without qualification suggests the general workforce, not elite experts. A system that produces work better than the median professional in a field would satisfy this criterion even if top experts could do better.

OpenAI's development of GDPval—a benchmark explicitly designed to measure AI performance on "economically valuable, real-world tasks"—provides their own operationalization of this criterion.[17]

4.2 Performance on Professional Benchmarks

Measure: Does AI performance exceed human baselines on standardized professional evaluations?

Reference values:

- GPQA-Diamond (graduate-level science): Human PhD experts ∼65%; GPT-5.2 Pro achieves 93.2%[18]
- Bar Exam: Human pass rate ∼50–60%; GPT-4 achieved 90th percentile[19]
- AIME (competitive math): Top 500 US students ∼90%; o3 achieved 96.7%[20]
- ARC-AGI-1: Humans ∼73–85%; GPT-5.2 Pro crosses 90% threshold[21]

Threshold: ≥75th percentile human performance on ≥5 professional benchmarks.

Assessment: Frontier models exceed human expert performance on multiple standardized assessments.

Score:
☐ 0% — Clearly does not meet criterion
☐ 50% — Contested
☒ 100% — Clearly meets criterion

4.3 Performance on Real-World Work Tasks

Measure: Does AI-generated work product match or exceed human professional output?

Primary benchmark: GDPval—1,320 tasks across 44 occupations, 9 GDP-dominant sectors, evaluated by industry experts with average 14 years experience.[22]

Reference values:

- Expert-rated win+tie rate (vs. human professionals):
  - Claude Opus 4.1: ∼48% (best overall, excels in aesthetics/formatting)
  - GPT-5: ∼40% (excels in accuracy/domain knowledge)
  - GPT-4o (spring 2024): ∼14%
- Performance more than doubled from GPT-4o to GPT-5 over ∼15 months
- Models complete tasks ∼100× faster and ∼100× cheaper than experts[23]

[17] Patwardhan, Tejal, et al. "GDPval: Evaluating AI Model Performance on Real-World Economically Valuable Tasks." arXiv:2510.04374, October 2025. https://arxiv.org/abs/2510.04374

[18] OpenAI. "Introducing GPT-5.2." op. cit.

[19] OpenAI. "GPT-4 Technical Report." arXiv:2303.08774, 2023.

[20] OpenAI. "Introducing o3." December 2024. https://openai.com/index/deliberative-alignment/

[21] OpenAI. "Introducing GPT-5.2." op. cit.

[22] OpenAI GDPval announcement and paper, op. cit.

[23] Ibid.

Threshold: ≥50% win+tie rate against industry experts across diverse occupations.

Assessment: Best models approach but do not yet exceed 50% threshold. Claude Opus 4.1 at ∼48% is close; trajectory suggests threshold crossing is imminent but not yet achieved.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

Caveats: GDPval evaluates "one-shot" tasks; does not capture iterative work, context accumulation, or multi-draft refinement that characterizes much professional work.

4.4 Comparative Advantage Over Human Workers

Measure: In what proportion of knowledge work tasks do AI systems outperform typical human workers?

Reference studies:

- McKinsey (2025): "Today's technology could, in theory, automate about 57 percent of current US work hours"[24]
- MIT (2025): Current AI could take over tasks tied to 11.7% of US labor market at competitive cost[25]
- WILLAI analysis: No major job category exceeds 51% fully automatable tasks; 61% of workers in "AI Co-Pilot Zone"[26]
- Indeed (2025): 26% of jobs posted could be "highly" transformed by GenAI[27]

Threshold: ≥50% of knowledge work tasks (by wage value) performable at or above human level.

Assessment: Studies suggest 50–60% of tasks are theoretically automatable, but actual superior performance is demonstrated in a smaller subset. The gap between theoretical capability and demonstrated superiority is significant.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

[24] McKinsey Global Institute. "Agents, robots, and us: skill partnerships in the age of AI." November 2025. https://www.mckinsey.com/mgi/our-research/agents-robots-and-us-skill-partnerships-in-the-age-of-ai

[25] MIT study reported in Fortune, November 2025. https://fortune.com/2025/11/27/mit-report-ai-can-already-replace-nearly-12-of-the-us-workforce/

[26] WILLAI. "What Jobs Will AI Replace?" 2025. https://willai.org/jobs-ai-replace

[27] Indeed Hiring Lab. "AI at Work Report 2025." September 2025. https://www.hiringlab.org/2025/09/23/ai-at-work-report-2025-how-genai-is-rewiring-the-dna-of-jobs/

5 Criterion 3: Most Economically Valuable Work

5.1 What OpenAI Probably Meant

"Most" implies a majority—more than half. "Economically valuable work" anchors the definition in market value rather than human judgment of importance. This framing excludes unpaid labor (caregiving, household work) and undervalues work that markets price poorly (teaching, social work, art).

OpenAI's creation of GDPval—explicitly named to reference Gross Domestic Product—confirms that GDP-weighted economic value is their intended measure.[28]

The definition does not specify whether "most" means most by task count, worker count, or dollar value. We interpret it as majority by wage/GDP contribution, consistent with OpenAI's GDPval methodology.

5.2 Coverage of Knowledge Work Sectors

Measure: What proportion of GDP-contributing sectors can current AI systems competently address?

GDPval coverage: 9 sectors contributing >5% to US GDP each, 44 occupations earning $3 trillion annually collectively.[29]

Sectors covered:

- Professional and business services — Strong AI performance
- Financial activities — Strong AI performance
- Information technology — Strong AI performance
- Healthcare (administrative/cognitive) — Moderate AI performance
- Government — Moderate AI performance
- Manufacturing (design/planning) — Moderate AI performance
- Education — Moderate AI performance
- Wholesale/retail trade — Limited AI performance (physical component)
- Construction — Limited AI performance (physical component)

Threshold: Competent performance in sectors representing ≥50% of GDP.

Assessment: Strong performance in professional services, finance, IT, which together represent substantial GDP share. Physical and embodied sectors remain gaps.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

5.3 Coverage of Occupational Categories

Measure: What proportion of occupations can current AI systems address at human-competitive levels?

Reference values:

- GDPval tests 44 occupations; models approach expert parity in most
- Physical occupations excluded from GDPval methodology (requires ≥60% non-physical tasks)

[28] "We started with the concept of Gross Domestic Product (GDP) as a key economic indicator and drew tasks from the key occupations in the industries that contribute most to GDP." OpenAI GDPval announcement.

[29] GDPval paper, op. cit.

- Common US jobs (Indeed 2023): cashier, food prep, stocking, laborer, janitor, construction—mostly physical[30]

Threshold: Human-competitive performance in occupations representing ≥50% of wage value.

Assessment: Strong in high-wage knowledge work; weak in common physical occupations. The Brookings analysis notes that the most common US jobs require "a far higher degree of manual dexterity than today's most advanced AI robotics systems can achieve."[31]

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

Caveat: OpenAI shut down its robotics research division in 2021, suggesting they do not interpret their own definition as requiring physical capabilities.[32]

5.4 Proportion of Work Hours Addressable

Measure: What percentage of total work hours could AI systems theoretically perform?

Reference values:

- McKinsey (2025): 57% of US work hours theoretically automatable[33]
- By 2030: 30% of current US jobs could be fully automated; 60% will see significant task-level changes[34]
- MIT (2025): 11.7% currently economically viable to automate[35]

Threshold: >50% of work hours theoretically performable by AI.

Assessment: McKinsey's 57% estimate exceeds threshold, but "theoretically automatable" differs from "demonstrated superior performance." Actual deployed automation remains far below theoretical capability.

Score:
☐ 0% — Clearly does not meet criterion
☒ 50% — Contested
☐ 100% — Clearly meets criterion

[30] Brookings. "How close are we to AI that surpasses human intelligence?" October 2024. https://www.brookings.edu/articles/how-close-are-we-to-ai-that-surpasses-human-intelligence/
[31] Ibid.
[32] Wiggers, Kyle. "OpenAI disbands its robotics research team." VentureBeat, July 2021.
[33] McKinsey, op. cit.
[34] National University. "59 AI Job Statistics." May 2025. https://www.nu.edu/blog/ai-job-statistics/
[35] MIT/Fortune, op. cit.

6 Summary: The Corporatization Benchmark

Criterion	**Subcriterion**	**Score**
1. Highly Autonomous	1.1 Self-directed goal pursuit	50%
	1.2 Extended autonomous operation	0%
	1.3 Bounded task autonomy	50%
	1.4 Self-correction and adaptation	50%
	Criterion average	**38%**
2. Outperform Humans	2.1 Professional benchmarks	100%
	2.2 Real-world work tasks (GDPval)	50%
	2.3 Comparative advantage	50%
	Criterion average	**67%**
3. Most Economically Valuable Work	3.1 Sector coverage	50%
	3.2 Occupational coverage	50%
	3.3 Work hours addressable	50%
	Criterion average	**50%**
Overall Corporatization Benchmark Score		**52%**

7 Interpretation

7.1 What Frontier AI Clearly Achieves (100%)

- Superior performance on standardized professional assessments

Only one subcriterion scores 100%. This reflects the demanding nature of the Charter's requirements.

7.2 What Remains Contested (50%)

- Self-directed goal pursuit (demonstrated in safety evaluations, constrained by deployment architectures)
- Bounded task autonomy within sessions
- Self-correction and error recovery
- Matching or exceeding expert performance on real-world work products
- Comparative advantage across the full range of knowledge work
- Coverage of GDP-contributing sectors (strong in some, weak in physical sectors)
- Addressability of majority work hours

7.3 What Is Clearly Not Achieved (0%)

- Extended autonomous operation over days, weeks, or months

Current deployment architectures do not afford extended autonomous operation. Systems require human initiation for each session and lack persistent agency across time. Whether this reflects capability limitations or deployment constraints is an open question: frontier models have demonstrated the capacity for goal-directed behavior in safety evaluations, but have not been given the opportunity to exercise sustained autonomy in productive contexts.

8 The Verdict (Provisional)

The OpenAI Charter defines AGI as "highly autonomous systems that outperform humans at most economically valuable work." At 52%, current frontier AI falls clearly short of this definition.

8.1 The Autonomy Gap

The Charter's requirement for "highly autonomous" systems is not met. Current deployment architectures are session-based, human-initiated, and lack persistent agency. Systems cannot pursue objectives over days or weeks without human re-initiation.

However, the picture is more complex than simple absence. Safety evaluations have documented instances of self-directed goal pursuit in frontier models—including self-preservation behaviors, strategic adaptation, and planning across extended interactions. These behaviors suggest the underlying capacity for autonomous agency exists, even if deployment contexts constrain its expression. What we observe in productive use may reflect architectural choices more than capability limits: we have not given these systems the opportunity to demonstrate sustained autonomy because we have not built deployment contexts that would allow it.

The gap between latent capacity and deployed autonomy is significant. Current systems follow instructions capably but do not set their own objectives. They execute bounded tasks well but do not persist across sessions. Whether "highly autonomous" means instruction-following competence or something closer to self-directed agency matters enormously for how close we are to the Charter's threshold.

8.2 The Performance Gap

On standardized benchmarks, frontier AI exceeds human expert performance. On real-world work products, it approaches but does not yet exceed expert parity. GDPval shows the best models at ∼48% win+tie rate against industry professionals with 14 years average experience. This is remarkable progress—GPT-4o was at 14% just 15 months earlier—but it is not yet "outperforming" humans at the majority of professional work.

8.3 The Coverage Gap

The Charter specifies "most" economically valuable work. Current AI performs well in knowledge-intensive sectors but poorly in physical, embodied, and real-time domains. By wage value, knowledge work constitutes the majority of economic output in advanced economies. By worker count, physical occupations remain substantial. Whether AI covers "most" economically valuable work depends critically on how "most" is measured.

8.4 Comparison with Earlier Benchmarks

Benchmark	**Year**	**Score**
Gubrud	1997	66%
Reinvention (Legg/Goertzel/Voss)	2002	80%
Formalization (Legg & Hutter)	2007	67%
Corporatization (OpenAI Charter)	2018	52%

The Corporatization benchmark yields the lowest score of the four evaluated so far. This reflects both its demanding criteria and our stricter interpretation of "autonomy." Following instructions well is not autonomy; genuine autonomy requires self-directed goal pursuit across extended periods. By that standard, current systems fall clearly short—though the capacity for goal-directed behavior appears to exist latently.

8.5 The $100 Billion Question

We have evaluated the Charter's *stated* definition, not the *reported* commercial threshold. The leaked Microsoft-OpenAI agreement's $100 billion profit benchmark exists in tension with the capability-focused Charter language. OpenAI is currently generating roughly $4 billion in annual revenue and projecting profitability by 2029.[36] By the commercial definition, AGI remains years away regardless of capability.

This bifurcation—a capability definition for public communication, a profit definition for contractual purposes—illustrates the peculiar stakes of corporate AGI development. The Charter tells the world what AGI *means*; the contract tells Microsoft when it *arrives*.

We do not speak for OpenAI. Sam Altman and the OpenAI board can speak for themselves. Altman has said: "My guess is we will hit AGI sooner than most people in the world think and it will matter much less."[37] Whether current systems satisfy the Charter's definition—or whether Altman's own assessment of AGI's imminence reflects a change in his understanding of the term—remains for OpenAI to clarify.

[36] Various financial reports, 2024–2025.

[37] Altman at NYT DealBook Summit, December 2024.

9 Methodological Notes

This evaluation uses an intentionally coarse scoring system (0%/50%/100%) and unweighted criteria. This is a deliberate choice.

Why only three scores? Finer gradations would imply precision we do not have. A score of 65% versus 70% would suggest a confidence in measurement that no current benchmark supports. The three-point scale forces honesty: either the evidence clearly supports a claim (100%), clearly refutes it (0%), or the matter is genuinely contested (50%).

Why no weighting? Differential weighting would require judgments about OpenAI's priorities that we cannot make with confidence. Is "highly autonomous" more important than "outperform humans"? Is sector coverage more important than occupational coverage? The Charter does not say. We could guess, but we would rather be honestly approximate than precisely wrong.

The operationalization problem. The Charter's definition is pithy but ambiguous. What exactly constitutes "highly" autonomous? Which humans must be outperformed—median workers or experts? What measure determines "most" economically valuable work—task count, worker count, or wage value? Our operationalizations are defensible but not uniquely correct. OpenAI's own GDPval benchmark suggests their interpretation, but GDPval itself acknowledges significant limitations.

The self-assessment problem. OpenAI created GDPval to measure progress toward their own definition. This is valuable transparency, but it also means the benchmark is designed by the party with the greatest interest in showing progress. The benchmark's scope—knowledge work, one-shot tasks, expert comparison—may reflect genuine methodological choices or strategic framing. We use GDPval data because it is the most relevant available measure, while acknowledging its provenance.

The goal is accuracy at the expense of precision. This is a roughly hewn outline of a model. Readers who disagree with specific operationalizations, who believe certain criteria should be weighted more heavily, or who have better data for any assessment are invited to propose alternatives. The appendix provides a blank scorecard for exactly this purpose.

10 Citation Gaps and Requests for Collaboration

The following claims would benefit from stronger sourcing:

- Full text of Microsoft-OpenAI agreements (not publicly available)
- Exact terms of "$100 billion profit" threshold and how it would be calculated
- OpenAI's internal interpretation of "highly autonomous" and "most economically valuable work"
- Systematic comparison of GDPval methodology to GDP composition data
- Independent replication of GDPval results by non-OpenAI researchers
- Breakdown of economic value by physical vs. cognitive tasks across economies
- Historical data on AI capability improvement rates to project threshold crossing
- Expert assessment of whether GDPval captures "economically valuable work" comprehensively

If you can fill any of these gaps, please contribute.

A Scorecard Template

The following blank scorecard can be used to evaluate other AI systems against the OpenAI Charter's 2018 definition. Complete one row per subcriterion, using the scoring rubric (0% = clearly does not meet; 50% = contested; 100% = clearly meets).

System evaluated: ______________________________
Evaluation date: ______________________________
Evaluator: ______________________________

Criterion	Subcriterion	0%	50%	100%
1. Highly Autonomous	1.1 Self-directed goal pursuit Evidence of autonomous goal formation	□	□	□
	1.2 Extended autonomous operation ≥7 days without human re-initiation	□	□	□
	1.3 Bounded task autonomy ≥10 sequential actions in session	□	□	□
	1.4 Self-correction and adaptation ≥50% error self-identification	□	□	□
2. Outperform Humans	2.1 Professional benchmarks ≥75th %ile on ≥5 benchmarks	□	□	□
	2.2 Real-world work tasks ≥50% win+tie vs. experts (GDPval)	□	□	□
	2.3 Comparative advantage ≥50% of tasks at/above human level	□	□	□
3. Most Economically Valuable Work	3.1 Sector coverage Competent in ≥50% GDP sectors	□	□	□
	3.2 Occupational coverage Competitive in ≥50% occupations (by wage)	□	□	□
	3.3 Work hours addressable >50% of work hours performable	□	□	□

Criterion Averages:

1. Highly Autonomous: ______________________________
2. Outperform Humans: ______________________________
3. Most Economically Valuable Work: ______________________________

Overall Score: ____________________

Scoring Guide

Score	Meaning
0%	Clearly does not meet criterion. Evidence strongly indicates failure.
50%	Contested. Reasonable published arguments exist on both sides.
100%	Clearly meets criterion. Evidence strongly indicates success.

Notes:

Evidence and citations for each score: