Superhuman Scores, Sub-Human Adaptation:

A Multi-Dimensional, Growth-Aware Comparison of Frontier AI Models and Human Intelligence

Shreyash Gondane
Independent Researcher
shreyasgondane@gmail.com

White paper · Data snapshot 27 September 2026

Abstract

Headlines now report that frontier AI models have "genius-level" IQ. We test how far that claim holds when intelligence is treated as a profile rather than a single number, and when rate of improvement is considered alongside level. We compile a September 2026 snapshot of public leaderboards for the ten leading model families (TrackingAI IQ tests, ARC-AGI-1/2/3, METR task-completion time horizons, Humanity's Last Exam and the Artificial Analysis Intelligence Index) and place them on human psychometric scales grounded in Cattell–Horn–Carroll theory. We introduce four simple instruments: a contamination-aware static z-score, an adaptive z-score derived from the interactive ARC-AGI-3 benchmark, a logistic success model over task novelty and task length, and arithmetic versus bottleneck (geometric) composite indices under alternative weightings. Four findings emerge. (1) On static tests, models exceed nearly all humans: 151 on the public Mensa Norway test (about 1 in 3,000) and 135 on a leak-resistant offline test (top 1%), with a 15–21 point public-to-offline gap indicating contamination. (2) On interactive learning of unfamiliar tasks, the best model reaches 62.7% of human efficiency and the median frontier model 2–30%, placing models in a region of ability space that almost no human occupies: high static reasoning, below-median adaptation. (3) The ranking between AI, an average adult and a domain expert reverses depending on the weighting; under a bottleneck index frontier AI loses about 11 points, the same penalty as a top expert, driven by near-zero continual learning. (4) Growth rates differ by orders of magnitude: the top AI IQ score rose about 50 points in 30 months against roughly 3 points per decade for the human Flynn effect, and the 50% time horizon doubles about every 131 days. We conclude that "smarter than humans" is ill-posed; the defensible statement is that current models are superhuman on well-specified, knowledge-heavy problems and remain below a human master on novel, long-horizon work, a gap that is closing on a timescale of months.

Keywords: artificial intelligence; psychometrics; Cattell–Horn–Carroll; fluid intelligence; ARC-AGI; time horizon; benchmark contamination; AGI measurement

1 Introduction

Every few months a new model is reported to have a higher IQ than the last. In September 2026 two systems share a score of 151 on the public Mensa Norway matrix test, a level reached by roughly one person in three thousand. Taken at face value, that number implies that frontier models are more intelligent than almost everyone. Yet the same models fail at things most people find easy: learning the rules of an unfamiliar game in a few minutes, remembering what happened last week, or carrying a project reliably over weeks.

The tension is not new. Psychometrics has long distinguished general ability from its components [1–3], and machine-intelligence researchers have argued that skill on a known task is a poor

proxy for the ability to acquire new skills [7, 8, 11]. What has changed is the size of the discrepancy. On static benchmarks models are near saturation; on the newest interactive benchmark they started 2026 below 1% [26]. A single scalar cannot represent both facts.

This paper makes three contributions. First, it assembles a dated, sourced snapshot of how the top frontier models score on IQ-style and adaptability benchmarks, with explicit human reference points. Second, it proposes a small set of transparent instruments for placing AI systems on human scales (contamination-aware and adaptive z-scores, a success landscape, and weighted composites) and shows how conclusions depend on the weighting. Third, it quantifies growth velocity, contrasting the months-scale doubling of AI capability with the decades-scale drift of human test performance. An interactive companion reproduces every figure (Data availability).

2 Background

2.1 Human intelligence as a structured profile

Spearman observed that performance on diverse mental tests is positively correlated and proposed a general factor, g [1]. Cattell separated fluid intelligence, the capacity to solve novel problems, from crystallized intelligence, accumulated knowledge [2]. Carroll's reanalysis of more than 460 datasets produced the three-stratum model that, merged with Cattell–Horn theory, became the Cattell–Horn–Carroll (CHC) framework, the most widely used structural model of human abilities [3]. Two further results matter here. Population IQ rose by roughly three points per decade during the twentieth century [4], the fastest documented change in human scores. Expert performance, by contrast, comes largely from about a decade of deliberate practice within a domain [5], which is why we distinguish an average adult from a domain master. General mental ability is the single strongest predictor of job performance, but it explains only part of the variance [6].

2.2 Defining and measuring machine intelligence

Legg and Hutter define universal intelligence as expected performance weighted across all computable environments [7], a formal ancestor of the weighted composites used below. Chollet redefines intelligence as skill-acquisition efficiency relative to priors and experience, and introduced the Abstraction and Reasoning Corpus (ARC) to measure it [11]. Hernández-Orallo shows why tests calibrated on humans can mislead when applied to machines [8]. Lake et al. argue that human learning is sample-efficient because it builds compositional, causal models [9], and catastrophic forgetting remains the central obstacle to continual learning in neural networks [10]. Recent frameworks rate systems on performance and generality [12] or score them on the ten CHC domains at equal weight; the latter reports GPT-4 at 27% and GPT-5 at 57% of a well-educated adult, with long-term memory storage as the largest deficit [13].

2.3 Capability growth and emergent behaviour

Language-model loss falls as a power law in parameters, data and compute [14, 16]; scale also brought in-context few-shot learning [15]. Some capabilities appear to emerge abruptly with scale [17], although part of that appearance is an artefact of discontinuous metrics [18], a caution that applies equally to the composite indices here. METR operationalises progress as the length of expert task an agent completes with a given reliability, finding exponential growth in this time horizon [19].

2.4 Human–AI comparisons

GPT-3 matched or exceeded people on Raven-style analogy problems without training on them [21], and early GPT-4 experiments were read by some as sparks of general intelligence [22]. A field experiment with 758 consultants found that AI assistance raised quality on tasks inside the model's capability frontier and lowered it outside, a "jagged frontier" [23]. Our results quantify that jaggedness across psychometric dimensions.

3 Data and methods

3.1 Data

All figures are a snapshot taken on 27 September 2026 from public sources (Table 1). IQ scores come from TrackingAI, which administers the public Mensa Norway test and an offline test written by a Mensa member and never published online [27, 28]. ARC-AGI-2 and ARC-AGI-3 scores come from the ARC Prize leaderboard and aggregators that republish it [25, 26, 29, 30]; for ARC-AGI-3 we use only standard-harness results. Humanity's Last Exam (HLE) [24] and Intelligence Index scores come from Artificial Analysis [31]. Time horizons come from METR [19, 20].

Human reference values are: IQ mean 100 and standard deviation 15; average individual accuracy of 64% on ARC-AGI-1 and 66% on ARC-AGI-2 against a human panel near 100% [25]; and a human baseline of 100% on ARC-AGI-3, which is defined relative to human action efficiency [26].

Table 1: Frontier model scores, 27 September 2026. Dashes mark no published result under that protocol. HLE = Humanity's Last Exam; AA = Artificial Analysis Intelligence Index v4.3.

Model	Mensa Norway	Offline IQ	ARC-AGI-2	ARC-AGI-3	HLE	AA
Claude Opus 5.5	–	–	91.7%	–	61.4%	58
Claude Fable 5.1	151	130	–	–	59.1%	53
GPT-6 Astra	151[v]	–	95.0%	62.7%	–	52.7
GPT-5.6 Sol	145	130	92.5%	7.8%	–	–
Claude Opus 5	144	129	–	30.2%	54.9%	–
Claude Fable 5	–	–	–	–	55.5%	–
GPT-5.6 Terra	–	135[v]	–	–	–	–
Gemini 3.8 Flash	–	–	–	10.4%	–	–
Gemini 3.1 Pro	145	–	77.1%[a]	0.37%[b]	–	–
Grok 4.6 / 4.20	145[c]	–	–	2.1%	–	–
Average adult	100	100	66%	–	–	–
Human panel / baseline	–	–	100%	100%	–	–

[v]Vision variant. [a]February 2026. [b]March 2026. [c]Grok-4.20, April 2026.

3.2 Placing models on human scales

A static z-score uses offline IQ where available, because public test items are likely present in training data:

$$z_s = \frac{IQ_{\text{off}} - 100}{15}. \tag{1}$$

Where only a public score exists we estimate $IQ_{\text{off}} \approx IQ_{\text{Mensa}} - 18$, the mean public-to-offline gap among models with both scores; such points are flagged as estimates. An adaptive z-score maps an

ARC-AGI-3 score $S \in [0, 100]$ onto the human distribution under the assumption that the 100% baseline corresponds to the human median:

$$z_a = \Phi^{-1}\left(\tfrac{1}{2}\, S/100\right), \tag{2}$$

where Φ^{-1} is the standard normal quantile. A simulated human population of 20,000 draws (z_s, z_a) from a bivariate normal with correlation 0.6, reflecting the positive manifold [1]; a human master is placed at $z = 3$ on both axes (about the top 0.1%). Combining the two gives an adaptability-adjusted IQ,

$$IQ^*(\lambda) = 100 + 15\big[(1-\lambda)\, z_s + \lambda\, z_a\big], \tag{3}$$

where $\lambda \in [0, 1]$ is the weight placed on learning new tasks.

3.3 Success landscape

Following METR's logistic fit in log task length [19], we model the probability of completing a task of expert duration h and novelty $n \in [0, 1]$ as

$$P(n, h) = \sigma\left(k \log_2 \frac{H_{50}}{h} - \beta\,(n - n_0)\right), \qquad n_0 = 0.2. \tag{4}$$

For frontier AI, $k = 0.574$ reproduces METR's reported 50% horizon of about 16 h and 80% horizon of about 3 h ($k = \ln 4 / \log_2(16/3)$), and $\beta = 6.0$ reproduces roughly 30% success on eight-minute fully novel tasks, the median of ARC-AGI-3 leaders. For a human master we assume $H_{50} = 2{,}000$ h (about one working year), $k = 0.5$ and $\beta = 1.5$. These human parameters are assumptions, not measurements.

3.4 Growth model

The AI horizon grows as

$$H_{50}(t) = H_0 \cdot 2^{(t - t_0)/T_d}, \tag{5}$$

with $H_0 = 16$ h at $t_0 = 8$ May 2026 (Claude Mythos Preview, a lower bound) and $T_d = 131$ days, METR's post-2023 doubling time [20]. We report a sensitivity band for $T_d \in [100, 200]$ days.

3.5 Cognitive profile and weighted composites

We score nine dimensions on 0–100 for three reference profiles: frontier AI (best of the top ten), an average adult, and a top expert in their own field (Table 2). Dimensions follow CHC broad abilities where possible and add dimensions that matter for deployed agents. Scores are anchored to a benchmark where one exists; for continual learning and embodied or social judgment they are informed judgments. For weights w_i and scores s_i we compute an arithmetic index and a geometric (bottleneck) index:

$$I_A = \frac{\sum_i w_i s_i}{\sum_i w_i}, \qquad I_G = \prod_i s_i^{\,w_i / \sum_j w_j}. \tag{6}$$

I_A lets a strength offset a weakness; I_G penalises the weakest dimension, closer to how a missing ability limits real work. Four weightings are compared: IQ-test (knowledge and static reasoning), Adaptability (after Chollet [11]), Economic work (autonomy, reliability, speed, judgment, after Schmidt and Hunter and Dell'Acqua et al. [6, 23]) and Balanced (equal weights, as in Hendrycks et al. [13]).

Table 2: Nine-dimension cognitive profile (0–100) and anchoring evidence.

Dimension	AI	Avg.	Expert	Anchor
Knowledge breadth	97	45	70	HLE 61% across 100+ fields
Static abstract reasoning	92	50	95	ARC-AGI-2 95% vs 66%; offline IQ 135
Expert problem solving	88	20	95	Competition maths, PhD-level science Q&A
Novel interactive learning	60	85	100	ARC-AGI-3 best 62.7% of human efficiency
Long-horizon autonomy	45	65	95	METR: 16 h at 50%, 3 h at 80%
Sample efficiency	30	75	90	Few-shot concept learning [9]
Continual learning & memory	15	80	90	No weight updates in deployment [10, 13]; judgment
Speed & scale	99	5	10	Throughput, parallel copies
Embodied & social judgment	20	85	90	Judgment; little benchmark coverage

4 Results

4.1 IQ-style tests

On the public Mensa Norway test the highest scores are 151 (Claude Fable 5.1 and GPT-6 Astra, vision), followed by a cluster at 142–145 (Figure 2). Fifteen models score 140 or higher. Under a normal distribution, 151 corresponds to about 1 in 2,970 people and 145 to 1 in 741 (Figure 1). Offline scores are lower: the best is 135 (GPT-5.6 Terra, vision), about 1 in 100. For models tested both ways the gap ranges from 2 points (GPT-5.6 Sol Ultra, vision) to 21 points (Claude Fable 5.1). The top public score rose from about 101 in March 2024 to 151 in September 2026.

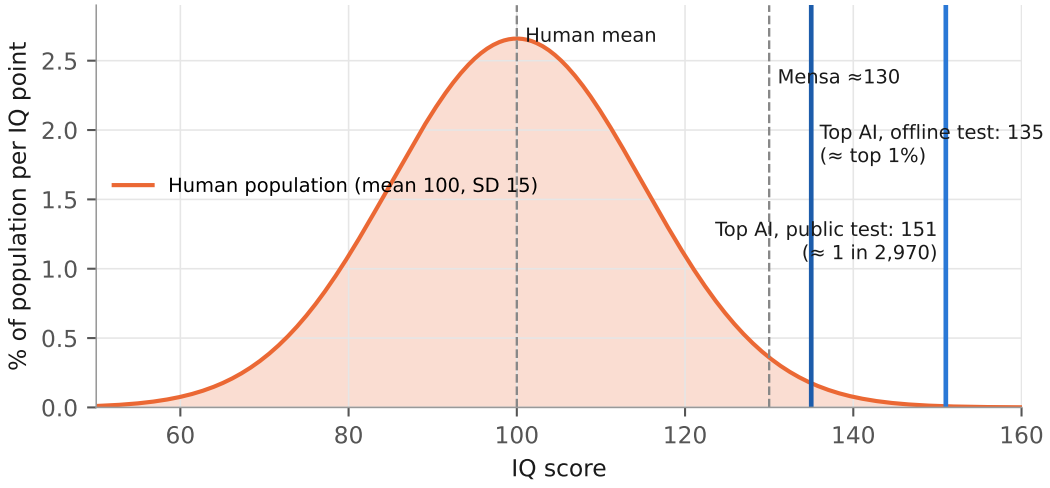

Figure 1: Human IQ distribution with the top AI scores on the public (151) and offline (135) tests.

4.2 Static versus interactive reasoning

Within one puzzle family the picture reverses as the task shifts from recognising a pattern to learning an environment (Figure 3). The best model scores 98.5% on ARC-AGI-1 and 95.0% on ARC-AGI-2, well above the 64–66% of an average individual. On ARC-AGI-3 every frontier model

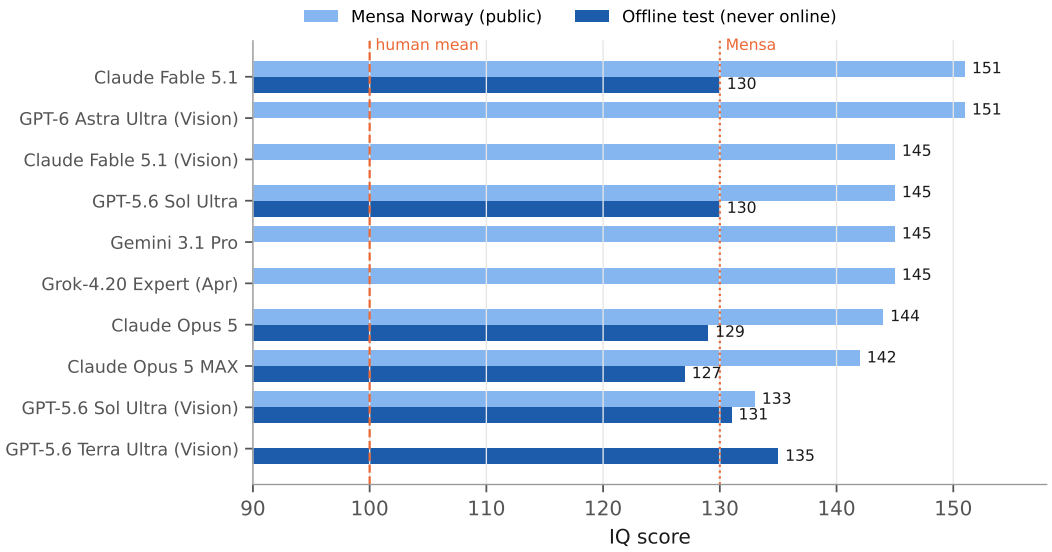

Figure 2: Public and offline IQ for the top ten model variants. The public-to-offline gap indicates how much of the public score reflects memorised items.

scored below 1% at launch in March 2026; by September the leader reached 62.7% and the second 30.2%, still below the human baseline. Mapped to human scales (Figure 4), measured models sit at $z_s \approx 2$ but z_a between -1.0 and -1.8: a "test-taker corner" ($z_s \geq 1.8$, $z_a \leq -1$) that one simulated person in 20,000 occupies.

4.3 Profile and weighting sensitivity

The profile is sharply jagged (Figure 5). Frontier AI leads on knowledge breadth, speed and static reasoning, while scoring lowest of the three profiles on novel learning, sample efficiency, continual learning and embodied judgment. Table 3 and Figure 6 show that the ranking depends on the weighting. Under IQ-test weights AI scores 90.1 against 83.9 for an expert and 39.6 for an average adult. Under adaptability weights AI falls to 44.6, below the average adult (72.8). Under a bottleneck index with balanced weights AI loses 10.9 points and the expert 11.1, so the geometric penalty does not single out AI; it shifts where each profile's weakness lies.

Table 3: Composite scores by weighting. I_A arithmetic, I_G bottleneck (geometric).

	Frontier AI		Average adult		Top expert	
Weighting	I_A	I_G	I_A	I_G	I_A	I_G
IQ-test	**90.1**	**89.6**	39.6	33.2	83.9	76.0
Adaptability (Chollet)	44.6	37.4	72.8	69.9	**93.8**	**93.7**
Economic work	60.9	50.2	53.9	39.9	**78.5**	**65.0**
Balanced (CHC-style)	60.7	49.8	56.7	44.1	**81.7**	**70.6**

The adaptability-adjusted IQ makes the same point on a familiar scale (Table 4, Figure 7). At $\lambda = 0$ the measured models score 129–133; at $\lambda = 0.5$ they fall to 102–113; at $\lambda = 1$ they score 74–93, below the human mean. A human master stays at 145 for every λ.

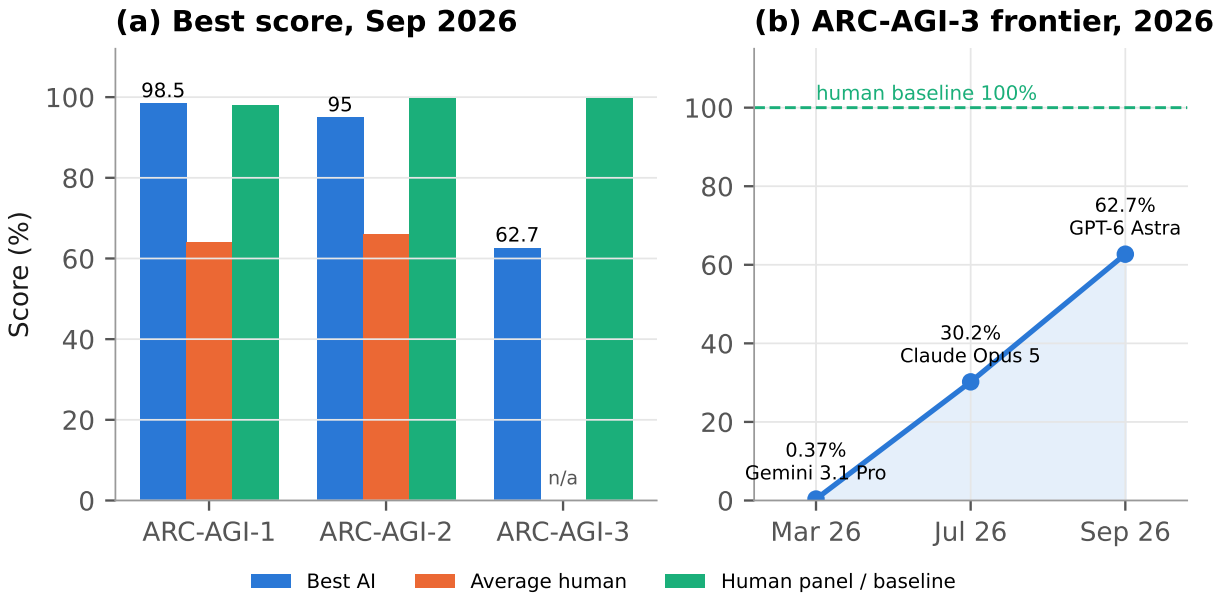

Figure 3: (a) Best AI versus human reference on three generations of ARC. (b) ARC-AGI-3 frontier over 2026 (standard harness).

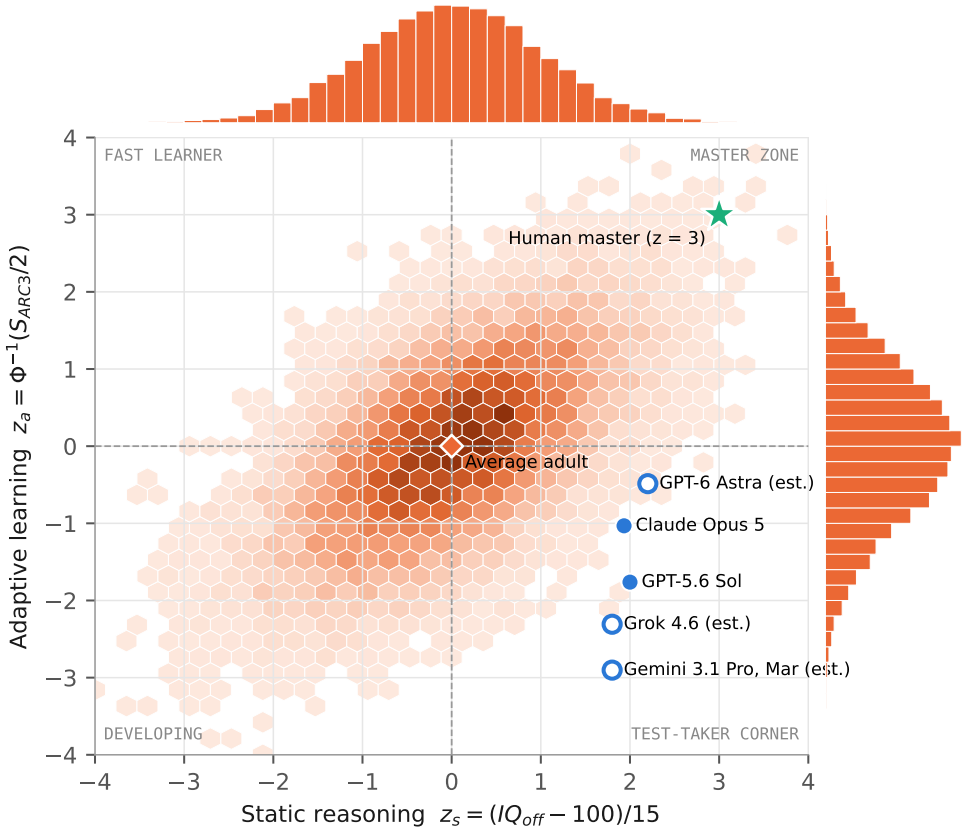

Figure 4: Simulated human population (20,000 draws, $\rho = 0.6$) on static and adaptive z-scores, with marginal histograms. Filled markers are measured; hollow markers use an estimated offline IQ.

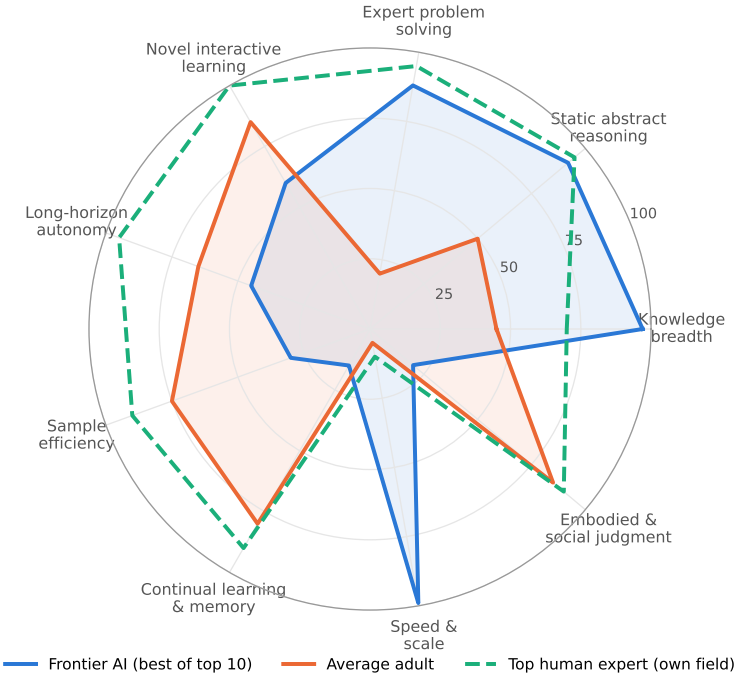

Figure 5: Nine-dimension profiles for frontier AI, an average adult and a top expert. Scores are an editorial index anchored to benchmarks where available (Table 2).

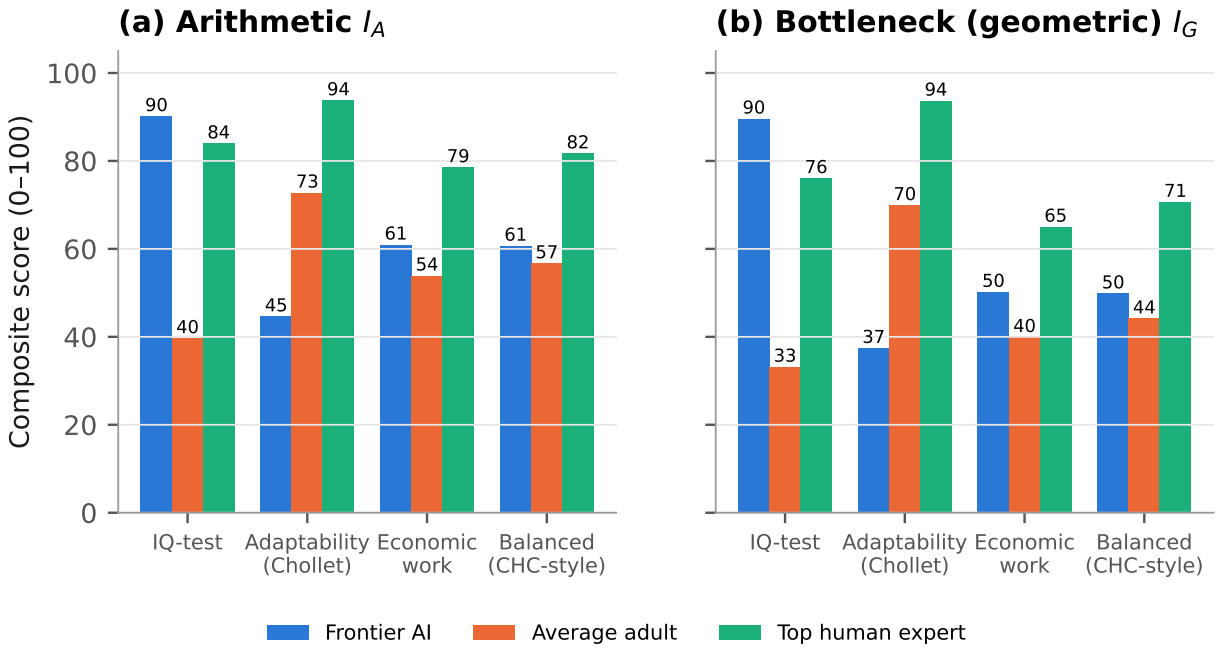

Figure 6: Composite scores under four weightings, (a) arithmetic and (b) bottleneck.

Table 4: Adaptability-adjusted IQ, $IQ^*(\lambda)$.

Profile	z_s	z_a	$\lambda = 0$	0.25	0.5	1
Human master	3.00	3.00	145	145	145	145
GPT-6 Astra (est. z_s)	2.20	−0.49	133	123	113	93
Claude Opus 5	1.93	−1.03	129	118	107	85
GPT-5.6 Sol	2.00	−1.76	130	116	102	74
Average adult	0	0	100	100	100	100

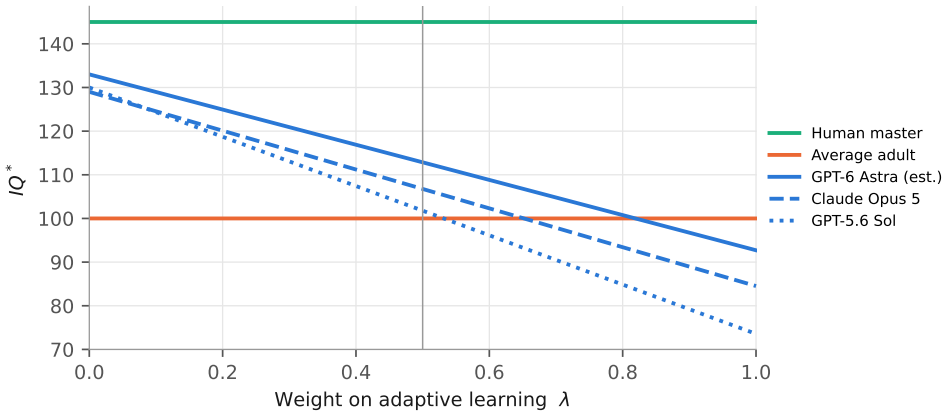

Figure 7: IQ^* as a function of the weight on adaptive learning.

4.4 Success landscape

Figure 8 shows the modelled probability of success across task novelty and length. On short, familiar tasks AI and a human master both succeed almost always. The master leads by more than 25 percentage points on 64% of the task space in September 2026. Holding novelty handling fixed and growing only the horizon, that share falls to 45% by September 2027 and 24% by September 2028. The remaining gap sits in the novel, long-duration corner.

4.5 Growth velocity

The METR 50% horizon rose from seconds (GPT-2, 2019) to at least 16 h (May 2026), roughly one doubling every 131 days since 2023 (Figure 9). Extrapolation gives the dates in Table 5. For comparison, the Flynn effect corresponds to about 0.3 IQ points per year [4], while the top AI public IQ score rose by about 20 points per year between 2024 and 2026, and ARC-AGI-3 rose from 0.37% to 62.7% in six months.

Table 5: Projected dates at which the AI 50% horizon reaches human work durations, if the 2023–26 trend holds.

Threshold	$T_d = 100$ d	$T_d = 131$ d	$T_d = 200$ d
1 work-week (40 h)	Sep 2026	Oct 2026	Jan 2027
1 work-month (167 h)	Apr 2027	Jul 2027	Mar 2028
Master's 1-year project (2,000 h)	Apr 2028	Nov 2028	Mar 2030

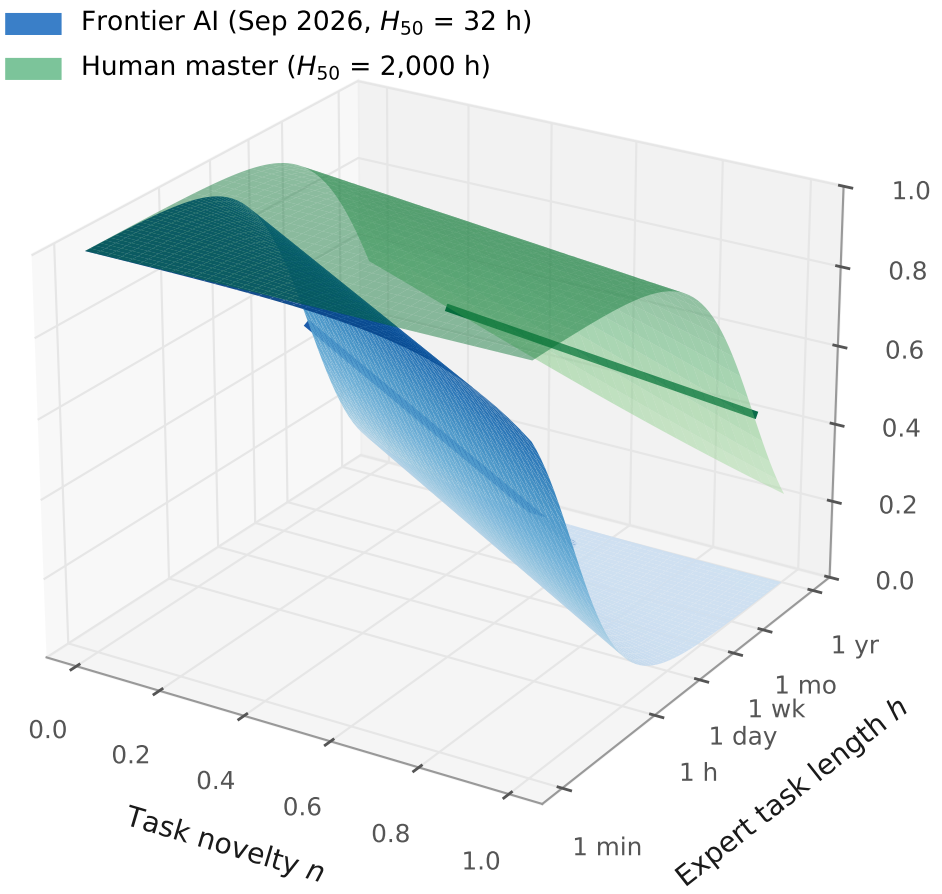

Figure 8: Modelled probability of success by task novelty and expert task length. Lines mark each profile's 50% frontier.

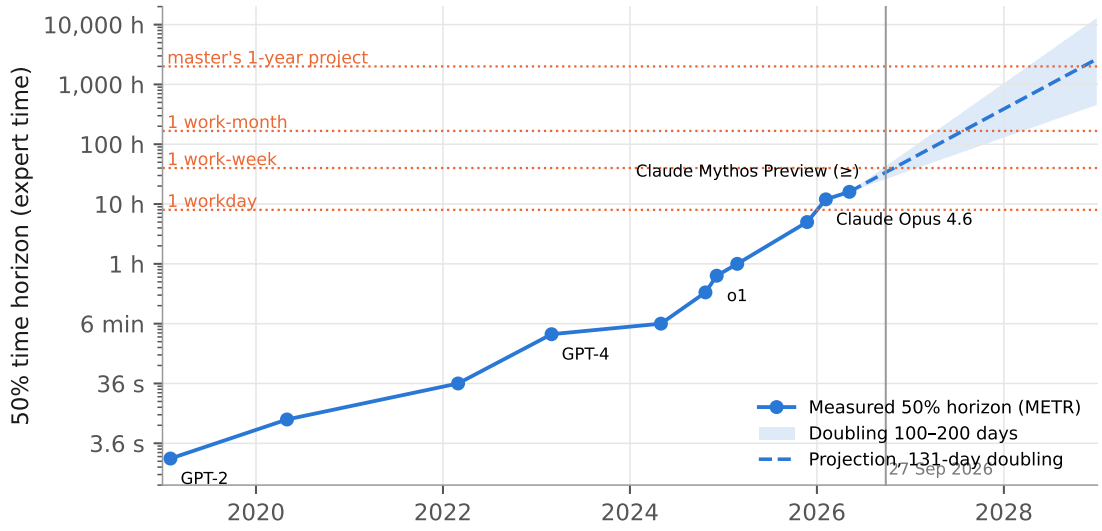

Figure 9: METR 50% time horizon (log scale) with projection and a 100–200 day doubling band, against human work durations.

5 Discussion

"Smarter than humans" is ill-posed. A model with IQ 151 and an adaptive z-score of -1.8 is not more or less intelligent than a person with IQ 115 and an adaptive z-score of $+1$; they have different shapes. Scalar comparisons implicitly choose a weighting, and Table 3 shows that reasonable weightings reverse the order. Reporting a profile together with the weighting is more informative than reporting a number.

The gap is in learning, not knowing. The dimensions where AI trails, namely interactive learning, sample efficiency, continual learning and long-horizon reliability, are the ones that psychometrics labels fluid and that Chollet [11] labels intelligence proper. Knowledge and pattern completion, where AI leads, are closer to crystallized ability. This matches the CHC-based finding that long-term memory storage is the largest deficit [13].

Implications for work. In practice the profile suggests a division of labour. AI is already superhuman on well-specified, knowledge-heavy subtasks. A human master remains necessary for setting goals, handling genuinely new situations, maintaining context across weeks and making judgment calls, consistent with the jagged-frontier field evidence [23]. The success landscape indicates that this boundary moves outward each quarter.

Measurement lessons. Three practices would improve comparisons: prefer leak-resistant or held-out tests to public ones; report the harness and reasoning effort alongside every score, since agent scaffolds can move ARC-AGI-3 from 62.7% to reported figures above 99%; and report reliability at more than one threshold, because the 80% horizon is about five times shorter than the 50% horizon.

6 Limitations

The profile scores and the human-master parameters are the author's estimates, anchored to benchmarks where possible. The ARC-AGI-3 mapping assumes the human baseline equals the human median. Offline IQ for three models is estimated from public scores. Several leaderboard values are vendor or aggregator reports rather than independent runs, and evaluation protocols (tools, effort, vision versus text) differ across rows. METR notes that its task suite is unreliable above about 16 h, so the current horizon is a lower bound and extrapolations are conditional on a trend that may bend [20]. IQ norms are population-specific, and none of the instruments measures motivation, values or embodiment directly. Results describe a single snapshot and will date quickly.

7 Conclusion

Frontier AI models in September 2026 are superhuman on static reasoning and knowledge (public IQ 151, offline 135, ARC-AGI-2 95%) and below a human master on learning unfamiliar tasks and sustaining long work (ARC-AGI-3 at most 62.7% of human efficiency; 50% horizon of about 16 h against months for an expert). Which is "more intelligent" depends on the weighting, and under balanced or adaptability-focused weightings a domain expert remains ahead. What does not depend on weighting is velocity: AI capability improves on a timescale of months, human capability on a timescale of decades. The most useful framing is therefore not a race but a moving boundary between tasks best delegated to AI and tasks that still need a human master.

Disclosure of AI assistance

Data compilation, analysis code, figures and drafting were carried out with the assistance of Anthropic's Claude (model identifier `claude-opus-5-5`). The author directed the study, reviewed the sources and is responsible for the content. Claude Opus 5.5 and other Anthropic models appear in the results; all rankings are taken from third-party leaderboards.

Data availability

All inputs are listed in Table 1 with sources below. An interactive companion reproducing every figure, including rotatable 3D views and adjustable weights, is available at https://claude.ai/artifact/5ukUJpsoqppaka2aM7BVtD.

References

[1] C. Spearman. "General intelligence," objectively determined and measured. *American Journal of Psychology*, 15(2):201–292, 1904. doi:10.2307/1412107.

[2] R. B. Cattell. Theory of fluid and crystallized intelligence: a critical experiment. *Journal of Educational Psychology*, 54(1):1–22, 1963. doi:10.1037/h0046743.

[3] J. B. Carroll. *Human Cognitive Abilities: A Survey of Factor-Analytic Studies*. Cambridge University Press, 1993. doi:10.1017/CBO9780511571312.

[4] J. R. Flynn. Massive IQ gains in 14 nations: what IQ tests really measure. *Psychological Bulletin*, 101(2):171–191, 1987. doi:10.1037/0033-2909.101.2.171.

[5] K. A. Ericsson, R. T. Krampe, and C. Tesch-Römer. The role of deliberate practice in the acquisition of expert performance. *Psychological Review*, 100(3):363–406, 1993. doi:10.1037/0033-295X.100.3.363.

[6] F. L. Schmidt and J. E. Hunter. The validity and utility of selection methods in personnel psychology. *Psychological Bulletin*, 124(2):262–274, 1998. doi:10.1037/0033-2909.124.2.262.

[7] S. Legg and M. Hutter. Universal intelligence: a definition of machine intelligence. *Minds and Machines*, 17(4):391–444, 2007. doi:10.1007/s11023-007-9079-x.

[8] J. Hernández-Orallo. *The Measure of All Minds*. Cambridge University Press, 2017. doi:10.1017/9781316594179.

[9] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman. Building machines that learn and think like people. *Behavioral and Brain Sciences*, 40:e253, 2017. doi:10.1017/S0140525X16001837.

[10] J. Kirkpatrick et al. Overcoming catastrophic forgetting in neural networks. *PNAS*, 114(13):3521–3526, 2017. doi:10.1073/pnas.1611835114.

[11] F. Chollet. On the measure of intelligence. arXiv:1911.01547, 2019. https://arxiv.org/abs/1911.01547.

[12] M. R. Morris et al. Levels of AGI for operationalizing progress on the path to AGI. arXiv:2311.02462, 2023. https://arxiv.org/abs/2311.02462.

[13] D. Hendrycks, D. Song, C. Szegedy, H. Lee, Y. Gal, et al. A definition of AGI. arXiv:2510.18212, 2025. https://arxiv.org/abs/2510.18212.

[14] J. Kaplan et al. Scaling laws for neural language models. arXiv:2001.08361, 2020. https://arxiv.org/abs/2001.08361.

[15] T. B. Brown et al. Language models are few-shot learners. In *NeurIPS*, 2020. https://arxiv.org/abs/2005.14165.

[16] J. Hoffmann et al. Training compute-optimal large language models. In *NeurIPS*, 2022. https://arxiv.org/abs/2203.15556.

[17] J. Wei et al. Emergent abilities of large language models. *Transactions on Machine Learning Research*, 2022. https://arxiv.org/abs/2206.07682.

[18] R. Schaeffer, B. Miranda, and S. Koyejo. Are emergent abilities of large language models a mirage? In *NeurIPS*, 2023. https://arxiv.org/abs/2304.15004.

[19] T. Kwa et al. Measuring AI ability to complete long tasks. arXiv:2503.14499, 2025. https://arxiv.org/abs/2503.14499.
[20] METR. Task-completion time horizons of frontier AI models; Time Horizon 1.1, 2026. https://metr.org/time-horizons/.
[21] T. Webb, K. J. Holyoak, and H. Lu. Emergent analogical reasoning in large language models. *Nature Human Behaviour*, 7:1526–1541, 2023. doi:10.1038/s41562-023-01659-w.
[22] S. Bubeck et al. Sparks of artificial general intelligence: early experiments with GPT-4. arXiv:2303.12712, 2023. https://arxiv.org/abs/2303.12712.
[23] F. Dell'Acqua et al. Navigating the jagged technological frontier. Harvard Business School Working Paper 24-013, 2023. doi:10.2139/ssrn.4573321.
[24] L. Phan et al. Humanity's Last Exam. arXiv:2501.14249, 2025. https://arxiv.org/abs/2501.14249.
[25] F. Chollet et al. ARC-AGI-2: a new challenge for frontier AI reasoning systems. arXiv:2505.11831, 2025. https://arxiv.org/abs/2505.11831.
[26] ARC Prize Foundation. ARC-AGI-3: a new challenge for frontier agentic intelligence. arXiv:2603.24621, 2026. https://arxiv.org/abs/2603.24621.
[27] M. Lott. TrackingAI: IQ test results for AI models, 2026. https://trackingai.org/IQ.
[28] BinaryVerse AI. AI IQ test, September 2026 snapshot. https://binaryverseai.com/ai-iq-test-2025/.
[29] BenchLM. ARC-AGI-2 and ARC-AGI-3 leaderboards, September 2026. https://benchlm.ai/benchmarks/arc-agi-2.
[30] LLMLearner. ARC-AGI-3 leaderboard (ARC Prize data, read 25 September 2026). https://llmlearner.com/rankings/arc-agi-3.
[31] Artificial Analysis. Humanity's Last Exam and Intelligence Index v4.3, September 2026. https://artificialanalysis.ai/evaluations/humanitys-last-exam.