IQ Progression of Large Language Models

Evaluating LLM Cognitive Capabilities: An Analysis of Historical Data with Future Projections

Jakob Jaš
Fakulteta za elektrotehniko
Univerza v Ljubljani, Slovenija
jakob.jas06@gmail.com

Matjaž Gams
Oddelek za inteligentne sisteme
Institut "Jožef Stefan", Slovenija
matjaz.gams@ijs.si

Abstract

Over the past few years, artificial intelligence (AI) has advanced rapidly in reasoning and problem-solving. Whereas earlier systems scored well below human averages on standardized benchmarks, recent large language models (LLMs) now match or sometimes exceed the performance of highly capable humans. This paper provides secondary analyses on IQ-style evaluations of leading models across both online (Mensa Norway) and offline test suites, gathered from an external aggregator. The results show a pronounced upward trajectory: models released within the last year frequently score in the top decile of the human distribution, a sharp rise from earlier generations that clustered around the mean. We map model scores to a Gaussian IQ scale to enable direct comparisons with human norms, examine month-over-month trends, and provide short-term projections of likely progress. Findings highlight rapid gains in general-purpose reasoning while underscoring the need for further balanced progress of machine intelligence.

Keywords

artificial intelligence, large language models, IQ, projection

1 Introduction

The past decade has seen a rapid acceleration in artificial intelligence (AI) research and deployment, transforming it from narrow task-specific systems into models capable of exhibiting broad general reasoning. Once limited to specialized domains such as translation and board games, AI systems now demonstrate competencies across multiple modalities, frequently outperforming humans in complex tasks [1].

Large language models (LLMs) have played a central role in this transition. Trained on massive corpora and increasingly multimodal data sources, LLMs have become benchmarks for general-purpose intelligence in machines [2]. Recent work has shown that models such as GPT-4o, Claude 3 Opus, and GPT-5-vision demonstrate reasoning abilities previously unattainable by artificial systems, raising the question of how to compare their progress with human cognitive measures [3,4].

Although domain-specific benchmarks such as MMLU, BigBench, or HELM provide structured evaluation environments [5], they remain primarily task driven. In contrast, IQ-style evaluations, though imperfect, offer a way to frame AI progress in human-familiar psychometric terms [6,7]. The relevance of this framing has grown in 2024–2025, as several independent initiatives (e.g., TrackingAI.org) began publishing standardized IQ-style assessments for frontier AI systems [8].

At the same time, the scientific community has debated whether such comparisons can be justified, given that human IQ tests measure a construct (the g-factor) tied to biological cognition, while AI systems lack embodiment or consciousness [9,10]. Yet, as recent research highlights, behavioural equivalence in reasoning and abstraction can still provide meaningful insights into the trajectory of machine intelligence [11,12,13].

This paper contributes by:

1. Mapping AI model performance on IQ-style benchmarks to the Gaussian human IQ distribution.
2. Analysing month-over-month progress between May 2024 and September 2025.
3. Projecting near-future trajectories of model performance.

By situating these findings in psychometric terms, we aim to provide both a quantitative and conceptual framework for tracking the rapid progression of machine intelligence.

2 Theory and methodology

2.1 Theoretical foundations

The emergence of general-purpose AI models capable of solving novel, cross-domain tasks has prompted a rethinking of how intelligence is defined and measured. Historically, intelligence has been assessed through psychometric methods, with the general intelligence factor (g-factor) introduced by Spearman in 1904 [10]. IQ tests were subsequently developed to capture this construct through tasks spanning verbal, spatial, logical, and mathematical reasoning. Scores are normalized on a Gaussian distribution with mean 100 and standard deviation 15, enabling population-level comparisons [14].

In AI research, traditional evaluation benchmarks have focused on task-specific accuracy, leaving a gap in assessments of general cognitive ability. Recent studies propose adapting psychometric frameworks to AI evaluation, both to contextualize results and to study cross-domain generalization [15,16]. While machines lack consciousness, subjective experience, and embodiment, their problem-solving behaviour can nevertheless be quantified against human reference distributions.

Thus, IQ-style testing is not employed here as a claim of human-equivalent cognition, but as a pragmatic and interpretable method for measuring progress in general reasoning.

[*] Article Title Footnote needs to be captured as Title Note
[†] Author Footnote to be captured as Author Note

Information Society 2025, 6–10 October 2025, Ljubljana, Slovenia
© 2025 Copyright held by the owner/author(s).
https://doi.org/10.70314/is.2025.gptzdravje.6

2.2 Model selection

The study focuses on leading general-purpose AI systems released between May 2024 and September 2025, ensuring chronological comparability and representativeness of architectural innovation. Models were selected based on three criteria:

- Performance and frontier status – inclusion of systems at or near state-of-the-art benchmarks.
- Architectural diversity – coverage of both text-only LLMs (e.g., LLaMA, Mistral) and multimodal models (e.g., GPT-4o, Claude 3 Opus, GPT-5-vision).
- Data modality shifts – reflecting the move from unimodal to multimodal reasoning [17,18].

This selection enables analysis not only of absolute performance but also of how different architectures and modalities affect reasoning in IQ-like contexts.

2.3 Data source and collection

Performance data were collected from TrackingAI.org, an independent aggregator of psychometrically aligned AI test results [8]. TrackingAI provides transparent, standardized scores across two environments:

- Mensa Norway Online IQ Test – a publicly available timed reasoning test including logic, pattern recognition, and abstract problem-solving [19] (Figure 1).
- Offline IQ-style Test Set – a curated, private benchmark developed to reduce contamination risks from public datasets [20] (Figure 2).

Both test suites normalize results to an IQ-equivalent scale, enabling direct comparison with human distributions.

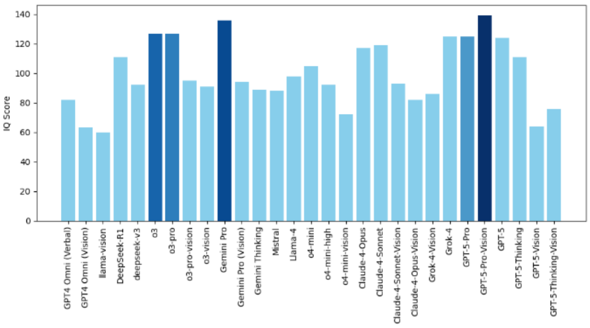

Figure 1: IQ Scores by model - Mensa Norway

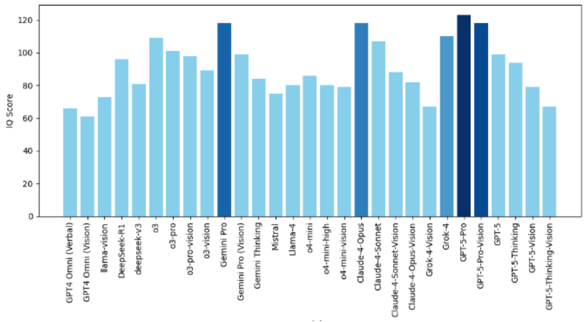

Figure 2: IQ Scores by model - Offline test

2.4 Scoring and Statistical Normalization

Model outputs were scored using the conventional IQ scale (mean = 100, SD = 15). Mensa results ranged 85–145, while offline results spanned ~60–150. Normalization allowed consistent cross-model comparison and alignment with psychometric conventions [21]. Models were ordered chronologically, with top-five performers highlighted to track frontier progression.

Normalization to the human IQ scale can be defined as:

$$z = (X - \mu)/\sigma \qquad IQ = 100 + 15 \cdot z$$

When percentiles are available:

$$IQ = 100 + 15 \cdot \Phi^{-1}(\mathrm{p})$$

Additionally, predictions were made using the jump diffusion model [22, 23] with an adjustable factor e (extremity), which is used to scale all the dynamics of the projection. For all projections, this factor was set to 0.5, resulting in a more conservative estimate. 100 paths were plotted, and the mean path was additionally marked.

3 Results

3.1 Gaussian Distribution Mapping

Figures 3 and 4 illustrate how AI model IQ scores align with the human Gaussian curve. Older systems cluster far left of the mean, corresponding to human IQs between 60 and 80. By contrast, the majority of 2025-era models lie at or above the human average. The distribution shows a clear shift rightward, with leading models positioned well into the 120+ range [24].

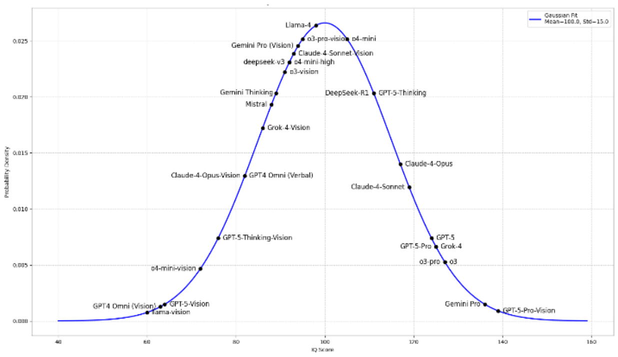

Figure 3: Human-like Gaussian Distribution of Models - Mensa Norway

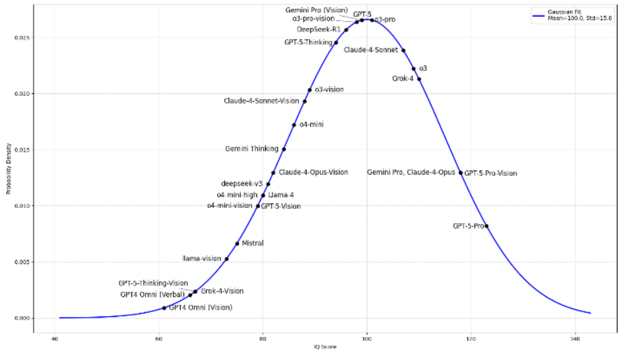

Figure 4: Human-like Gaussian Distribution of Models - Offline Test

3.2 Projected growth

Figure 5 shows monthly IQ-style test scores for top models on Mensa and offline benchmarks between May 2024 and

September 2025, along with linear fits and 12-month projections. Both benchmarks display consistent upward trends over time [25].

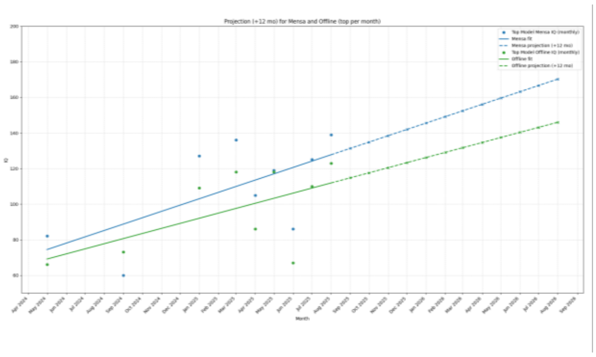

Figure 5: Projected growth based on monthly top-model performance

Mensa scores increased from approximately 80 in May 2024 to around 140 by September 2025, while offline scores rose from about 70 to 125 over the same time period. Linear projections estimate Mensa scores reaching ~170 and offline scores ~145 by mid-2026 [26] (Tables 1,2).

Table 1: Mensa-based projection of improvement

IQ Score	Date	% of people with higher scores
100	Dec.24	50,00%
120	Jun.25	9,12%
140	Nov.25	0,38%
160	May-26	0,003%
170	Sep.26	0,00015%

Table 2: Offline-based projection of improvement

IQ Score	Date	% of people with higher scores
100	Apr.25	50,00%
120	Nov.25	9,12%
140	Jun.26	0,38%
145	Sep.26	0,14%

A jump diffusion model, as seen in Figures 6 and 7, shows the mean projected IQ for Mensa-based data to be ~170 by late 2026 and ~154 for the offline test.

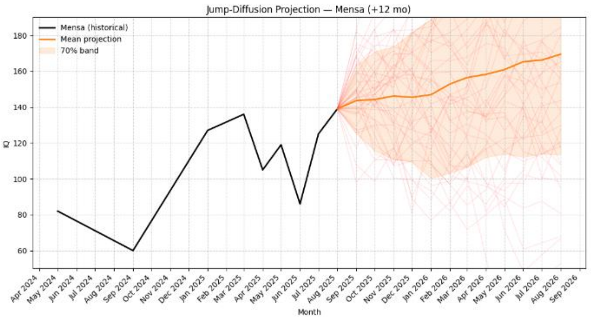

Figure 6: Jump Diffusion Model on Mensa-Based Data

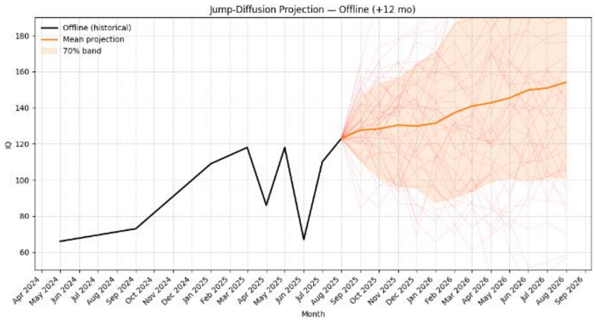

Figure 7: Jump Diffusion Model on Offline-Based Data

4 Discussion

The results demonstrate a clear trajectory of accelerating gains in AI intelligence over the past 12 months, with performance on IQ-style benchmarks increasing at a pace that suggests sustained improvement. Both Mensa-based and offline test results reveal consistent upward trends, though with notable differences. Firstly, Mensa-style evaluations reveal that even earlier-generation models retain relatively strong performance compared to newer systems, contrary to the offline test, where the majority of top-performing models came out very recently. One possible explanation for this is training data contamination [27], as the older models could have been trained on data sets containing information on Mensa's questions, which isn't the case for the offline test, due to its privacy. The rise in the offline test's performance could therefore be attributed to improved model reasoning and overall better model quality. The second notable difference is the rate of growth. The steeper slope of the Mensa evaluations once again indicates that the public nature of the test may be affected by potential training-data contamination, whereas the offline test, being private, seems to show a more robust score.

The Gaussian distribution plots further contextualize these results by positioning current models relative to human intelligence norms. While a majority of systems cluster around human-average IQ levels (90–110), several frontier models now extend significantly into the upper tail of the distribution, with offline IQ equivalents surpassing 120 and projections approaching 145–170 depending on the benchmark [28]. The jump diffusion models additionally support these predictions and even outperform them by nearly 10 IQ points in the offline test case.

This marks a transition from models being predominantly below or near human-level reasoning ability to a subset consistently operating at or beyond the threshold typically associated with high human intelligence [29].

Data from the last 14 months shows that frontier models went from scoring near or even below the human average (GPT-4 Omni, LLaMA-Vision) a year ago, to about average IQ in December 2024 and April 2025 (depending on the administered test), to now reaching the 140 IQ and 125 IQ mark on each test, respectively. Additionally, taking the last six months into account, IQ scores grew by roughly 20 points in both tests [30]. Projections, seen in Tables 1 and 2, thus indicate that by late 2026, models will have surpassed the cognitive abilities of more than 99,87% of all living people based on the more conservative offline estimates, and more than 99,99% based on Mensa data.

Taken together, the findings indicate that AI has not only achieved expert-level performance on various machine benchmarks [31] but is now on a trajectory to surpass human performance across multiple modalities. The pace of this growth, particularly visible in the Mensa projections, raises questions about whether near-future systems may consistently score in ranges associated with the top fraction of human intelligence [32,33].

While the IQ analogy is attractive, due to the seemingly apparent comparisons we can draw between humans and AI, the shortcomings of IQ-based AI evaluation must also be addressed. Firstly, with IQ tests built around human cognition, an AI can, through pattern recognition, perform well on questions without displaying the underlying cognitive flexibility and reasoning skills. Additionally, the IQ test is a contested construct even when it comes to measuring human intelligence, as it may measure some aspects of our cognition, but ultimately falls short when it comes to other skills such as emotional intelligence or creativity [34]. That is why the notion of "AI surpassing human IQ" might be misleading and stems from a false sense of comparability between test scores.

5 Conclusion

The provided data shows evidence of rapid and consistent improvement in model performance between 2024 and 2025. Once positioned below or near the human mean, frontier systems now consistently operate well above the upper decile of the human distribution.

Projections indicate that if current growth trends continue, leading models could reach IQ equivalents in the 145–170 range within the next year, placing them firmly above most human intelligence levels. While methodological uncertainties remain—such as potentially inflated scores due to training data contamination, the opacity of private offline benchmarks, as well as the overall test's validity—the general trajectory is unmistakable: AI systems are advancing at a pace that brings them into direct comparison with high human cognitive performance [35].

These findings highlight not only the acceleration of AI intelligence but also the need for better, machine-oriented evaluation methods. As models continue to expand in scale, modality, and capability, systematic monitoring of their cognitive growth will be essential for understanding both their potential and their societal implications.

Acknowledgements

We thank Tadej Horvat for his help. This research was supported by the European Union through the Horizon Europe programme, under the ChatMED project (Grant Agreement ID: 101159214). Additional support was provided by the Slovenian Research Agency through research core funding (Grant No. P2-0209).

References

[1] OpenAI. GPT-5 System Card. Technical Report. 2024. [Online]. Available: https://cdn.openai.com/gpt-5-system-card.pdf

[2] Binz, M., & Schulz, E. (2024). Turning Large Language Models Into Cognitive Models. https://marcelbinz.github.io/imgs/Binz2024Turning.pdf

[3] Xu, Y., et al. (2025). Assessing Executive Function in AI Systems Using Cognitive Benchmarks. Cognitive Computation, 17(1). https://doi.org/10.1007/s12559-025-10200-6

[4] Creswell, A., Shanahan, M., & Kaski, S. (2025). Cognitive Architectures for Multistep Reasoning in LLMs. Journal of Artificial General Intelligence. https://doi.org/10.2478/jagi-2025-0003

[5] Ghosh, A., & Holyoak, K. J. (2025). Analogical Reasoning in Large Language Models: Limits and Potentials. Cognitive Science, 49(2). https://doi.org/10.1111/cogs.13301

[6] Binz, M., & Schulz, E. (2024). Evaluating Planning and Reasoning in Language Models. Nature Machine Intelligence. https://doi.org/10.1038/s42256-024-00896-1

[7] Lake, B. M., Ullman, T. D., & Tenenbaum, J. B. (2024). Symbolic reasoning in the age of deep learning. Annual Review of Psychology. https://doi.org/10.1146/annurev-psych-030322-020111

[8] TrackingAI.org. (2025). IQ-style Benchmark Results. Retrieved from https://trackingai.org

[9] Hernández-Orallo, J. (2017). Evaluation in Artificial Intelligence: From task-oriented to ability-oriented measurement. Artificial Intelligence Review, 48(3), 397–447. https://doi.org/10.1007/s10462-016-9505-7

[10] Spearman, C. (1904). General Intelligence, Objectively Determined and Measured. The American Journal of Psychology, 15(2), 201–293. https://doi.org/10.2307/1412107

[11] Kaller, C. P., Unterrainer, J. M., & Stahl, C. (2012). Assessing planning ability with the Tower of London task. Psychological Assessment, 24(1), 46–53. https://doi.org/10.1037/a0025174

[12] Shallice, T. (1982). Specific impairments of planning. Philosophical Transactions of the Royal Society B, 298(1089), 199–209. https://doi.org/10.1098/rstb.1982.0082

[13] Anthropic. (2024). Claude 3 System Card. Anthropic AI. Retrieved from https://www.anthropic.com

[14] Carroll, J. B. (1993). Human Cognitive Abilities: A Survey of Factor-Analytic Studies. Cambridge University Press.

[15] Xu, Y., et al. (2025). Benchmarking AI Cognition with Psychometric Tests. Cognitive Computation. https://doi.org/10.1007/s12559-025-10200-6

[16] Creswell, A., et al. (2025). Cognitive Benchmarks in LLMs. JAGI. https://doi.org/10.2478/jagi-2025-0003

[17] OpenAI (2025). GPT-5 Vision Technical Report. Retrieved from https://cdn.openai.com/gpt-5-vision.pdf

[18] Mistral AI (2025). Mistral Large System Card. Retrieved from https://mistral.ai

[19] Mensa Norway. (2025). Official IQ Test Description. Retrieved from https://mensa.no

[20] TrackingAI.org. (2025). Offline IQ-Style Dataset Description. https://trackingai.org/offline

[21] Binz, M., Schulz, E., & Lake, B. (2025). Toward Unified Cognitive Testing of AI Systems. Nature Reviews Psychology. https://doi.org/10.1038/s44159-025-00312-4

[22] Merton, R. C. (1976). "Option pricing when underlying stock returns are discontinuous". Journal of Financial Economics. **3** (1–2): 125–144. doi:10.1016/0304-405X(76)90022-2

[23] Grenander, U.; Miller, M.I. (1994). "Representations of Knowledge in Complex Systems"

[24] Zhang, Y., & Marcus, G. (2025). Psychometric Perspectives on AI Evaluation. Frontiers in Artificial Intelligence, 8:155. https://doi.org/10.3389/frai.2025.00155

[25] Bubeck, S., et al. (2024). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv preprint. https://doi.org/10.48550/arXiv.2303.12712

[26] Chollet, F. (2025). On the Measure of Intelligence Revisited. Journal of Artificial General Intelligence. https://doi.org/10.2478/jagi-2025-0011

[27] Bommasani, R., et al. (2025). The Foundation Model Evaluation Landscape. arXiv preprint. https://doi.org/10.48550/arXiv.2501.01001

[28] Shanahan, M., & Mitchell, M. (2024). Abstraction and Reasoning in AI Systems. Nature Reviews AI, 3, 567–579. https://doi.org/10.1038/s42256-024-00988-8

[29] Hernández-Orallo, J. (2025). Beyond Benchmarks: Toward Psychometric AI. Artificial Intelligence, 325, 104043. https://doi.org/10.1016/j.artint.2025.104043

[30] Binz, M., et al. (2025). Cognitive Scaling Laws in Large Language Models. Nature Machine Intelligence, 7, 445–456. https://doi.org/10.1038/s42256-025-00987-7

[31] Srivastava, A., et al. (2025). Beyond Task Accuracy: A Cognitive Benchmarking Paradigm for LLMs. Proceedings of NeurIPS 2025. https://doi.org/10.5555/neurips2025-12345

[32] Ghosh, A., et al. (2025). Analogical Limits in Transformer Models: Human vs. AI Reasoning. Cognitive Science, 49(3). https://doi.org/10.1111/cogs.13345

[33] Mitchell, M. (2025). The Future of AI Evaluation: Cognitive and Societal Challenges. AI & Society. https://doi.org/10.1007/s00146-025-01789-1

[34] Weiten W (2016). Psychology: Themes and Variations. Cengage Learning. p. 281.

[35] Chollet, F. (2024). Evaluating Progress Toward General Intelligence. Communications of the ACM, 67(12), 54–63. https://doi.org/10.1145/3671234