A Coherence-Based Measure of AGI

Fares Fourati
KAUST
fares.fourati@kaust.edu.sa

Abstract

Recent approaches to evaluating *Artificial General Intelligence* (AGI) typically summarize a system's capability using the arithmetic mean of its proficiencies across multiple cognitive domains. While simple, this implicitly assumes compensability: exceptional performance in some areas can offset severe deficiencies in others. Genuine general intelligence, however, requires coherent sufficiency: balanced competence across all essential faculties. We introduce a coherence-based measure of AGI that integrates the generalized mean over a continuum of compensability exponents. This yields an *area-under-the-curve* (AUC) metric spanning arithmetic, geometric, and harmonic regimes, quantifying how robust an evaluated capability remains as compensability assumptions become stricter. Unlike the arithmetic mean, which rewards specialization, the AUC penalizes imbalance and exposes bottlenecks that constrain performance. To illustrate the framework, we apply it to cognitive profiles derived from the Cattell–Horn–Carroll (CHC) model, showing how coherence-based aggregation highlights imbalances that are obscured by arithmetic averaging. As a second, independent example, we apply the same methodology to a set of 17 heterogeneous benchmarks, demonstrating how coherence-based evaluation can reveal unevenness even in narrower task collections. These examples show that the proposed approach offers a principled, interpretable, and stricter foundation for measuring progress toward AGI.

1 Introduction

Artificial General Intelligence (AGI) is often described as the ultimate goal of Artificial Intelligence (AI) research, the creation of systems with intelligence comparable to that of humans. Yet, despite decades of progress, measuring advancement toward this goal remains elusive, largely because both *intelligence* and *generality* lack rigorous operational definitions [Chollet, 2019]. While individual capabilities can be evaluated, their integration into a single indicator of AGI is far less straightforward.

Recent work by Hendrycks et al. [2025] builds on the informal notion that "AGI is an AI that can match or exceed the cognitive versatility and proficiency of a well-educated adult," their framework grounds evaluation in human psychometrics through the Cattell–Horn–Carroll (CHC) theory of cognitive abilities [Carroll, 1993, McGrew, 2009, 2023a], the dominant empirical model of human intelligence. By decomposing cognition into ten broad domains (e.g., reasoning, memory, perception, and speed) and averaging performance across them, they define a unified *AGI score* as the arithmetic mean of domain proficiencies.

While simple and appealing, this formulation, for example, implicitly assumes a degree of *compensability*: that exceptional performance in some faculties can offset severe deficiencies in others. As a result, an AI system could appear "general" while failing entirely in critical domains such as reasoning or memory. This assumption conflicts with both psychometric evidence and systems theory. In human cognition, abilities are interdependent, reasoning depends on working memory, perception constrains abstraction, and learning relies on durable long-term memory [Carroll, 1993, McGrew, 2009, 2023a]. Empirical studies in cognitive psychology consistently find that extreme imbalance

Preprint.

across faculties is associated with functional impairment rather than high general intelligence [Cattell, 1987]. Likewise, in complex engineered systems, overall capability is limited by the weakest component, a "bottleneck" or limiting-factor dynamic [Keeney and Raiffa, 1993, Kitano, 2004], also acknowledged by Hendrycks et al. [2025].

We argue that general intelligence should instead reflect *coherent sufficiency*: consistently high competence across all essential faculties, such that no domain falls to a low level that would preclude genuine generality. This principle resonates across disciplines, psychometrics emphasizes coherence among subtests Carroll [1993], McGrew [2009], systems theory highlights limiting-factor dynamics Kitano [2004], and multi-criteria decision analysis formalizes non-compensatory aggregation Keeney and Raiffa [1993]. Collectively, these perspectives converge on the view that true generality arises from robustness and systemic balance rather than isolated excellence.

To formalize this intuition, we extend the arithmetic aggregation to the continuous family of *generalized means* [Bullen, 2013], parameterized by a compensability exponent p. When $p = 1$, the measure reduces to the arithmetic mean; as p decreases, it increasingly penalizes imbalance across cognitive domains, smoothly transitioning toward non-compensatory evaluation. The resulting curve over p (e.g., Figure 2) provides a concise diagnostic of coherence: models with flatter, higher curves exhibit more uniformly distributed competence and greater robustness to compensability assumptions.

Integrating this curve over a specified range of p yields an *area-under-the-curve* (AUC) score, expressed as a percentage, that summarizes the stability of a system's performance under increasingly strict compensability assumptions. This construction preserves continuity with prior arithmetic-mean definitions while introducing a coherence-based and theoretically grounded measure of generality.

Our contributions are threefold:

1. **Conceptual:** We identify *compensability* as a fundamental limitation of existing AGI measurement practices and situate it within broader theories of cognitive coherence, systems interdependence, and limiting-factor dynamics.
2. **Methodological:** We propose evaluating AGI through the full AGI_p curve, characterizing how performance behaves across compensability regimes, and through its integral, an AUC-based indicator that captures the coherence and robustness of a system's capabilities.
3. **Illustrative:** We validate the framework in two independent settings: (i) CHC-style domain profiles for GPT-4 and GPT-5, and (ii) a heterogeneous suite of 17 benchmark scores for Gemini 3 Pro, GPT-5.1, Gemini 2.5 Pro, and Claude Sonnet 4.5. In both analyses, coherence-based aggregation exposes capability imbalances and limiting factors that arithmetic means obscure, demonstrating the practical value of the approach.

By reframing AGI evaluation around coherence rather than compensation, this work provides a stricter, more interpretable, and more principled criterion for assessing genuine progress toward general intelligence.

2 Background

2.1 Psychometric Models of General Intelligence

Human intelligence is widely understood as a structured hierarchy of interdependent abilities rather than a single undifferentiated trait. The Cattell–Horn–Carroll (CHC) theory of cognitive abilities [Carroll, 1993, McGrew, 2009, 2023a] provides the dominant empirical framework for modeling this structure. CHC organizes cognition into a three-stratum hierarchy: a broad *general intelligence* factor (g) at the apex, a set of *broad abilities* (e.g., fluid reasoning, crystallized knowledge, memory, and processing speed) beneath it, and dozens of *narrow abilities* at the base. Over the past three decades, nearly all major human IQ tests have adopted the CHC approach for test construction and interpretation [Keith and Reynolds, 2010, McGrew, 2023b].

The CHC framework thus operationalizes *general intelligence* not as monolithic competence, but as coherent sufficiency across multiple cognitive domains. An individual's overall intelligence score reflects consistent adequacy across reasoning, memory, perception, and other core faculties, with extreme deficiencies in any one domain typically resulting in significant functional impairment. This interdependence of abilities motivates extending the same principle to artificial systems.

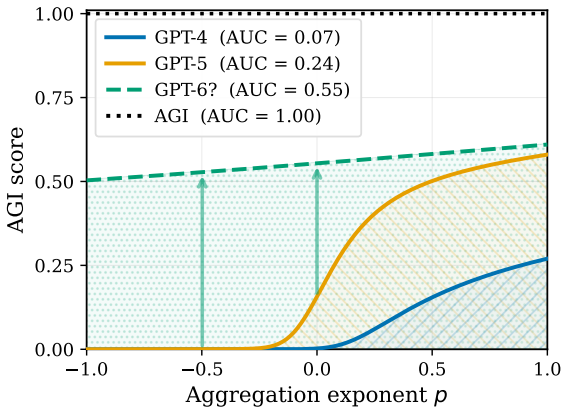

Figure 1: Comparison of model performance across aggregation exponents p. Curves show AGI_p values; shaded regions indicate the AUC.

2.2 CHC-Inspired AGI Evaluation

Most AI benchmarks to date have assessed narrow capabilities in isolation, such as ImageNet for visual recognition [Deng et al., 2009], MATH for mathematical reasoning [Hendrycks et al., 2021], PIQA for commonsense inference [Bisk et al., 2020]. Such benchmarks measure *specialization*, not *generality*. To address this, Hendrycks et al. [2025] proposed the first psychometrically grounded definition of AGI, explicitly mapping CHC's cognitive domains to artificial analogues. Their framework decomposes machine cognition into ten broad abilities—General Knowledge, Reading and Writing, Mathematics, On-the-Spot Reasoning, Working Memory, Long-Term Memory (Storage and Retrieval), Visual Processing, Auditory Processing, and Speed—and evaluates proficiency in each through adapted psychometric tasks.

These domain scores are aggregated using the **arithmetic mean**, producing a single AGI Score between 0% and 100%. For instance, Hendrycks et al. [2025] report scores of 27% for GPT-4 [Achiam et al., 2023] and 58% for GPT-5, illustrating both rapid improvement and persistent gaps relative to human-level cognition. However, this additive aggregation assumes, at least implicitly, that cognitive abilities are independent and somehow compensatory, i.e, that exceptional performance in one domain can hide complete absence in another. As a result, the arithmetic mean can overstate overall competence, producing deceptively high composite scores for systems that appear broadly capable while still exhibiting catastrophic deficits in essential faculties such as long-term memory storage or cross-modal reasoning.

Table 1: Domain-level AGI scores reported by Hendrycks et al. [2025]. K = General Knowledge; RW = Reading and Writing; M = Mathematics; R = On-the-Spot Reasoning; WM = Working Memory; MS = Long-Term Memory Storage; MR = Long-Term Memory Retrieval; V = Visual Processing; A = Auditory Processing; S = Speed.

Model	**K**	**RW**	**M**	**R**	**WM**	**MS**	**MR**	**V**	**A**	**S**
GPT-4 (2023)	80	60	40	0	20	0	40	0	0	30
GPT-5 (2025)	90	100	100	70	50	0	40	40	60	30
AGI	100	100	100	100	100	100	100	100	100	100

2.3 Compensability in Complex Systems

In multi-criteria decision theory [Keeney and Raiffa, 1993], *compensatory aggregation* methods assume that trade-offs between dimensions are permissible—strong performance in one criterion can offset weakness in another. By contrast, *non-compensatory aggregation* models treat dimensions as jointly constraining, such that poor performance on any critical dimension limits the overall outcome. This framing parallels the *limiting-factor principle* and analyses of robustness in complex engineered

systems [Kitano, 2004], both of which emphasize that system-level functioning can be bounded by its weakest component.

In psychometrics, interdependence among cognitive abilities has been repeatedly observed. Factor-analytic studies within the CHC framework reveal a pervasive *positive manifold*—individuals who perform well in one broad ability tend to perform well across others [Carroll, 1993, McGrew, 2009, 2023a]. This pattern implies mutual reinforcement among abilities rather than independence. Furthermore, empirical work shows that limitations in core processes such as working memory or processing speed can constrain higher-level reasoning and learning performance [Fry and Hale, 2000]. Accordingly, simple averages across domains may obscure such dependencies, overstating general competence when specific foundational faculties are weak.

2.4 Motivation for a Coherent-Based AGI Metric

Given this background, an AGI metric should not merely *sum* capabilities but ensure their coherent sufficiency, a minimal level of competence across all essential faculties. An AI that fails completely in long-term memory or perception cannot be considered "general," regardless of its average score. In contrast, non-compensatory aggregation naturally enforces balance by penalizing uneven or "brittle" cognitive profiles. The *generalized mean* offers a principled mathematical family that interpolates between compensatory and non-compensatory regimes through a single parameter p. This property motivates our proposed measure in the following section, which redefines general intelligence as balanced, interdependent competence across all CHC-inspired cognitive domains.

3 A Coherence-Based Measure of General Intelligence

Evaluating progress toward AGI requires aggregating performance across multiple capabilities that jointly determine a system's generality. Let

$$s_i \in [0, 100], \qquad i = 1, \ldots, n,$$

denote the normalized proficiencies of a model across n benchmark dimensions. These may represent cognitive domains, task clusters, or heterogeneous evaluations drawn from diverse modalities. Our framework makes no assumptions about the underlying ontology: each s_i need only reflect a meaningful ability that contributes to general-purpose competence.

A widely used aggregation rule, adopted for example in the AGI definition of Hendrycks et al. [2025], is the arithmetic mean:

$$\mathrm{AGI}_1 = \frac{1}{n} \sum_{i=1}^{n} s_i. \tag{1}$$

This formulation, denoted as AGI_1, assumes that strengths in some faculties can compensate for weaknesses in others, a property known as *compensability*. Although convenient, such additive aggregation is misaligned with both psychometric evidence and limiting-factor dynamics in complex systems. Severe deficiencies in essential faculties, e.g., reasoning, memory, or perception, often impose system-level bottlenecks that cannot be compensated by strengths elsewhere, causing AGI_1 to substantially overestimate holistic capability.

3.1 Generalized Means and Compensability

To relax this compensatory assumption, we generalize Eq. (1) to generalized-mean (power-mean) family [Bullen, 2013], parameterized by a compensability exponent $p \in \mathbb{R}$, where each value of p corresponds to a distinct *coherence regime* for general intelligence, controlling the degree to which strengths can offset weaknesses:

$$\mathrm{AGI}_p = \begin{cases} \left(\frac{1}{n} \sum_{i=1}^{n} \max(s_i, \varepsilon)^p \right)^{1/p}, & p \neq 0, \\ \left(\prod_{i=1}^{n} \max(s_i, \varepsilon) \right)^{1/n}, & p = 0, \end{cases} \tag{2}$$

where $\varepsilon > 0$ is a small stability constant (set to 10^{-6} in all experiments) that prevents numerical collapse when any $s_i = 0$ and can be viewed as a minimal detectable competence or a noise floor in domain measurement.

The exponent p determines the allowed degree of compensability:

- $p = 1$: arithmetic mean (strongly compensatory),
- $p = 0$: geometric mean (moderately non-compensatory),
- $p = -1$: harmonic mean (strongly non-compensatory),
- $p \to -\infty$: minimum operator (strict bottleneck).

As p decreases, AGI_p shifts from measuring *average proficiency* to measuring *coherent sufficiency*, requiring balanced competence across all essential dimensions. This reflects the intuition that general intelligence is not characterized by isolated peaks but by the absence of critical weaknesses.

The limiting case $p \to -\infty$ yields a fully non-compensatory measure equal to the weakest domain score, but is often too sparse to be practical: the overall score changes only when the minimum dimension improves. Even the harmonic mean ($p = -1$) can collapse toward zero on uneven profiles. Conversely, permissive regimes such as $p = 1$ can mask catastrophic deficits, yielding inflated scores despite near-zero performance in key areas.

These considerations motivate restricting attention to a moderate range, such as $p \in [-1, 1]$, which enforces meaningful non-compensatory behavior while preserving sensitivity and avoiding degeneracy. Within this interval, the AGI_p curve provides a concise diagnostic of coherence: models with flatter, higher curves exhibit more uniformly distributed competence and greater robustness to compensability assumptions.

For a single scalar summary of progress toward AGI, one may integrate AGI_p over the chosen p-range, yielding a measure that captures both robustness and coherence.

3.2 An Integrated Measure of Coherence: $\mathrm{AGI}_{\mathrm{AUC}}$

To summarize robustness across compensability regimes, we define the $\mathrm{AGI}_{\mathrm{AUC}}$ metric:

$$\mathrm{AGI}_{\mathrm{AUC}} = \frac{1}{p_{\max} - p_{\min}} \int_{p_{\min}}^{p_{\max}} \mathrm{AGI}_p \, dp. \tag{3}$$

Throughout this work, we set $p_{\min} = -1$ and $p_{\max} = 1$, spanning the range from strongly non-compensatory to strongly compensatory aggregation. The integral can be evaluated numerically using the composite trapezoidal rule over a uniform grid in p.

The AGI_p family highlights the characteristic "jaggedness" of modern AI systems: high proficiency in some domains coexists with catastrophic failures in others.

$\mathrm{AGI}_{\mathrm{AUC}}$ measures *coherence* and *stability across compensability regimes*. Systems with high AGI_1 but low $\mathrm{AGI}_{\mathrm{AUC}}$ rely on narrow excellence and lack balanced, integrated capability. In contrast, a system approaching general intelligence would display a uniformly high AGI_p curve across the entire range of p, indicating robustness even under stricter non-compensatory evaluation. Thus, $\mathrm{AGI}_{\mathrm{AUC}}$ provides a principled and coherence-sensitive summary of general intelligence, capturing both the breadth and the structural balance of a system's abilities.

4 Results on the CHC-Based Domain Scores

Setup. We apply the proposed framework to the published AGI domain scores of Hendrycks et al. [2025] for GPT-4 [Achiam et al., 2023] and GPT-5 (2025). For completeness, these baseline scores are reproduced in Table 1. The floor constant was fixed at $\varepsilon = 10^{-6}$, and the compensability interval was set to $p \in [-1, 1]$. We report AGI_{-1}, $\mathrm{AGI}_{-0.5}$, AGI_0, $\mathrm{AGI}_{0.5}$, and AGI_1, along with their corresponding AGI_p curves and the aggregate $\mathrm{AGI}_{\mathrm{AUC}}$ measure. To assess external alignment, we also compare the resulting $\mathrm{AGI}_{\mathrm{AUC}}$ values with performance on challenging intelligence benchmarks.

The analysis in this section builds directly on the ten normalized domain scores reported in Hendrycks et al. [2025]. In Appendix A, we document the origin of these values and compute several alternative aggregates, including weighted and unweighted geometric means. We keep this extended discussion in the appendix to maintain focus on the primary contribution of this paper: demonstrating how coherence-based aggregation provides a more principled framework for assessing AGI.

Table 2: Key scores per model (in %). AGI_1 corresponds to the arithmetic mean (p=1), $\mathrm{AGI}_{0.5}$ to the quasi-arithmetic mean (p=0.5), AGI_0 to the geometric mean (p=0), $\mathrm{AGI}_{-0.5}$ to the inverse-quasi mean (p= $-$ 0.5), AGI_{-1} to the harmonic mean (p= $-$ 1), and $\mathrm{AGI}_{\mathrm{AUC}}$ to the area under the AGI_p curve. Values are expressed as percentages relative to the ideal AGI reference model (= 100%).

Model	$\mathbf{AGI_1}$	$\mathbf{AGI_{0.5}}$	$\mathbf{AGI_0}$	$\mathbf{AGI_{-0.5}}$	$\mathbf{AGI_{-1}}$	$\mathbf{AGI_{AUC}}$ (ours)
GPT-4 (2023)	27	16	0	0	0	**7**
GPT-5 (2025)	58	50	16	0	0	**24**
AGI	100	100	100	100	100	100

Results. Figure 2 and Table 2 summarize how model performance varies under different compensability assumptions. When $p = 1$, the arithmetic mean (AGI_1) reproduces the values reported in Hendrycks et al. [2025]. Under this compensatory definition, GPT-5 appears to have made substantial progress over GPT-4 (58% vs. 27%), suggesting meaningful gains in overall capability and surpassing the halfway mark toward human-level proficiency.

However, as compensability decreases, this apparent advantage diminishes sharply. The geometric mean (AGI_0) collapses to near zero for both GPT-4 and GPT-5, reflecting the persistence of minimal proficiency in several cognitive domains. This pattern indicates that both systems rely heavily on a subset of strong faculties rather than demonstrating balanced competence across domains. Consequently, the arithmetic mean tends to overstate progress by averaging across highly uneven abilities, producing an inflated signal of movement toward AGI.

The area-under-the-curve measure, $\mathrm{AGI}_{\mathrm{AUC}}$, integrates performance across $p \in [-1, 1]$ to capture overall coherence. GPT-5 achieves a roughly threefold improvement over GPT-4 (24% vs. 7%), reflecting greater but still limited robustness. The AGI_p curves in Figure 2 exhibit steep declines for negative p values, indicating that weaknesses in domains such as long-term memory and adaptive reasoning continue to act as system-wide bottlenecks. A system approaching genuine generality would instead maintain a flatter and higher AGI_p curve across the full range of p, indicating resilience under increasingly strict coupling regimes. In the speculative case of "GPT-6?", shown in In Figure 2, we simulate the effect of mitigating GPT-5's principal bottleneck in long-term memory storage (MS; see Table 1) by increasing its domain score from 0% to 30%, while keeping all other domain scores unchanged. Both AGI_p and $\mathrm{AGI}_{\mathrm{AUC}}$ increase substantially, suggesting that even modest improvements in the weakest faculties can yield disproportionate gains in overall coherence as captured by the proposed metrics.

These results motivate the use of $\mathrm{AGI}_{\mathrm{AUC}}$ as a more reliable summary metric. Strict aggregations such as AGI_{-1} or $\mathrm{AGI}_{-0.5}$ are excessively punitive, collapsing uneven but meaningful progress to near-zero values, whereas optimistic measures such as AGI_1 or $\mathrm{AGI}_{0.5}$ exaggerate progress by ignoring structural dependencies. By integrating across compensability regimes, $\mathrm{AGI}_{\mathrm{AUC}}$ jointly captures both *breadth* and *balance*, yielding a stable and interpretable indicator of holistic capability. It neither collapses for incomplete systems nor inflates scores for narrow specialists, making it a robust and coherence-sensitive measure of emerging general intelligence.

Alignment with external benchmarks. We compare the AGI_p metrics with challenging external reasoning benchmarks. The ARC-AGI-2 benchmark [Chollet, 2019, Chollet et al., 2025], which emphasizes out-of-distribution reasoning and cross-domain abstraction, reveals a clear calibration effect. While GPT-5 achieves an arithmetic-mean score of $\mathrm{AGI}_1 = 58\%$, suggesting near-human generality, its coherence-aware score of $\mathrm{AGI}_{\mathrm{AUC}} = 24\%$ aligns much more closely with ARC-AGI-2 results, where GPT-5 (Pro) reaches approximately 18% and GPT-5 (High) about 10%. Similarly, the BIG-Bench Extra Hard evaluations [Kazemi et al., 2025], which probe broader reasoning proficiency across linguistically diverse tasks, report a GPT-4 score of roughly 6%, closely matching its $\mathrm{AGI}_{\mathrm{AUC}} = 7\%$, unlike the arithmetic mean of 27%.

This correspondence supports the interpretation that $\mathrm{AGI}_{\mathrm{AUC}}$ more faithfully captures *functional coherence*, the capacity to sustain competence across diverse cognitive domains, whereas the arithmetic mean AGI_1 tends to overstate progress by rewarding isolated peaks of specialization. Accordingly, $\mathrm{AGI}_{\mathrm{AUC}}$ serves as a more conservative indicator of emerging generality, aligning closely with independent measures of out-of-distribution reasoning such as ARC-AGI-2 and BIG-Bench Hard.

5 Beyond CHC-Domain Evaluation

While the previous section focuses on the ten CHC-inspired cognitive domains defined by Hendrycks et al. [2025], the coherence-based framework developed in this work is not tied to that taxonomy. As demonstrated in Appendix B, we further evaluate the same generalized-mean family AGI_p and the integrated coherence metric $\mathrm{AGI}_{\mathrm{AUC}}$ on a heterogeneous set of 17 benchmarks drawn from the publicly released *Gemini 3 Pro Model Evaluation Report* [DeepMind, 2025], for GPT-5.1, Gemini 3 Pro, Claude Sonnet 2.5 Pro, and Claude Sonnet 4.5 [Anthropic, 2025]. These benchmarks span symbolic reasoning, scientific problem solving, mathematical tasks, coding and agentic behavior, multimodal abstraction, and long-context understanding, illustrating the flexibility of the framework across evaluation settings.

The extended analysis reveals patterns that closely mirror those observed in the CHC-domain setting. Aggregation with $p = 1$ (i.e., the arithmetic mean) again yields optimistic assessments of overall capability by masking imbalances across tasks, whereas negative-p aggregation exposes persistent bottlenecks. Models with similar average performance diverge sharply in coherence: their AGI_p curves show steep declines for $p < 0$, and $\mathrm{AGI}_{\mathrm{AUC}}$ provides a clearer and more stable ordering of holistic competence. These results reinforce that coherence captures whether a system's abilities form a balanced, general-purpose profile rather than a collection of narrow skill peaks.

At the same time, while coherence-based aggregation provides a substantially more principled and informative summary than the arithmetic mean, any benchmark-derived score must ultimately be interpreted in the context of the benchmark suite it aggregates. To meaningfully quantify "progress toward AGI," the benchmarks themselves must span the full spectrum of essential cognitive faculties. Expanding benchmark coverage—across continual learning, long-horizon reasoning, memory formation, planning, and other fundamental abilities, remains an important direction for the community. Within any given benchmark set, however, our coherence-based aggregation offers a robust and conservative measure of generality: it penalizes uneven skill profiles, highlights hidden bottlenecks, and rewards the balance and robustness that are core to general intelligence.

6 Discussion

The results indicate that the arithmetic mean used in prior definitions of AGI (AGI_1) is a fragile indicator of generality. It rewards specialization and conceals structural weaknesses. Under coherence-aware formulations (AGI_0, AGI_{-1}, and $\mathrm{AGI}_{\mathrm{AUC}}$), the same systems score far lower, revealing that contemporary models remain unevenly developed across core cognitive domains and skills.

The current CHC-derived score is stricter yet still optimistic. The coherence-based framework reveals that current estimates of general intelligence in frontier models are likely inflated by additive aggregation. While GPT-5 achieves an AGI_1 (arithmetic mean) of 58%, suggesting near-human breadth, its coherence-adjusted $\mathrm{AGI}_{\mathrm{AUC}}$ falls to only 24%. Even this value should be regarded as an *upper bound* on general competence. As shown in Appendix A, the ten constituent cognitive scores from Hendrycks et al. [2025] themselves display substantial inflation when aggregated arithmetically. When recalculated using geometric or unweighted means, several domains collapse: auditory processing drops from roughly 60% to near 0%, visual processing from 40% to near 0%, processing speed from 30% to near 0%, and on-the-spot reasoning from 70% (arithmetic) to 19% (weighted geometric) or 5% (unweighted). These discrepancies expose a systematic overestimation of generality due to compensatory aggregation, where strong subskills mask critical deficiencies elsewhere. Accordingly, the nominal $\mathrm{AGI}_{\mathrm{AUC}}$ of 24% should be interpreted as conservative yet still optimistic, given (potential) inflation at the domain-score level.

Rethinking progress toward AGI. The divergence between AGI_1 and coherence-based metrics highlights a key distinction between *breadth of performance* and *integration of capability*. Improvements in select domains inflate arithmetic averages but contribute little to genuine generality if other faculties remain weak. This mirrors refinements in human psychometrics, where single-factor IQ models evolved into hierarchical frameworks [Carroll, 1993, McGrew, 2009, 2023a] to capture both domain-specific and global ability. AI evaluation requires a similar shift, from measuring *average proficiency* to assessing *functional sufficiency* across interdependent cognitive dimensions.

Toward coherence-based evaluation. The generalized-mean formulation offers a continuous lens for exploring compensability assumptions. Across this spectrum, $\mathrm{AGI}_{\mathrm{AUC}}$ emerges as the most stable and informative single measure: it integrates performance across optimistic and strict regimes, balancing sensitivity to weaknesses with tolerance for partial unevenness. Systems that achieve high AGI_1 but low AGI_0 or $\mathrm{AGI}_{\mathrm{AUC}}$ should thus be interpreted as *specialized* rather than coherent generalists. Conversely, genuine progress toward AGI would manifest as a flatter, consistently high AGI_p curve, indicating resilience under increasingly strict coupling of abilities. Future benchmarks and model cards should therefore report both AGI_1 (average proficiency) and coherence-oriented measures such as $\mathrm{AGI}_{\mathrm{AUC}}$ to disentangle breadth from balance.

On metric pluralism. Although coherence-based measures expose the structural limitations of arithmetic averaging, no single scalar, including $\mathrm{AGI}_{\mathrm{AUC}}$, can fully capture general intelligence. Each aggregate reflects a complementary facet of capability: the arithmetic mean highlights breadth, negative-p regimes reveal bottlenecks, and the full AGI_p curve provides a visual diagnostic of how performance behaves under varying compensability assumptions. Rather than serving as a single leaderboard number, the continuum over p forms a multidimensional assessment toolkit. External benchmarks such as ARC-AGI-2 [Chollet et al., 2025] and BIG-Bench Extra Hard [Kazemi et al., 2025] further complement this analysis, and can themselves be integrated using the same generalized-mean framework to yield coherence-aware summaries of performance.

Interpretive implications. Low curves or AUC-based scores do not imply an absence of generalization, but rather highlight fragility when tasks demand multi-domain coordination. The coherence perspective thus bridges psychometric and computational traditions, emphasizing integration and stability over isolated domain success. The goal of proposing a stricter, coherence-based measure is not to diminish the remarkable capabilities of current GPT-class systems, but to sharpen the lens through which progress toward AGI is assessed. True general intelligence should demonstrate balanced competence in reasoning, abstraction, generalization, planning, and memory, the capacities that underpin adaptability and understanding beyond pattern replication. By tightening the definition of generality, this framework aims to steer progress toward integrated, cross-domain coherence, an intelligence that not only performs well, but *reasons, learns, and plans* as a unified system.

Adaptive evaluation framework. A central advantage of the proposed coherence-based measure is its generality: it applies uniformly across heterogeneous benchmark families without committing to a fixed cognitive taxonomy. Our formulation allows for any collection of tasks, whether CHC-aligned domains (as in Section 4), symbolic reasoning suites, multimodal assessments, or mixed benchmark portfolios (as in Section 5), to be aggregated using the same generalized-mean family AGI_p. The metric depends only on normalized task scores, not on assumptions about human-inspired structure or domain boundaries. As demonstrated in Appendix B, this enables seamless integration of diverse evaluations (e.g., Humanity's Last Exam [Phan et al., 2025] and ARC-AGI-2 [Chollet et al., 2025]), producing coherence-based summaries that remain practical even as benchmarks evolve. By grounding assessment in task-level constructs while maintaining a taxonomy-agnostic aggregation rule, the framework is inherently adaptable: new benchmarks, modalities, or task definitions can be incorporated without modifying the underlying notion of general intelligence.

7 Conclusion

The arithmetic-mean definition of general intelligence, while simple and interpretable, conceals the structural imbalance of current AI systems. By generalizing this formulation to a continuum of compensability exponents, we reveal that apparent progress toward AGI is often confined to narrow strengths rather than coherent, system-wide competence. The proposed family of AGI_p measures, and in particular the integrated $\mathrm{AGI}_{\mathrm{AUC}}$, provides a stricter and more interpretable framework for assessing generality.

Our results show that coherence, the balanced sufficiency of abilities across domains, is a more faithful indicator of general intelligence than arithmetic breadth alone. Genuine progress toward AGI will thus manifest not as higher averages, but as flatter and elevated AGI_p curves reflecting resilient, interdependent capabilities. Incorporating coherence-based metrics such as $\mathrm{AGI}_{\mathrm{AUC}}$ into future benchmarks can align empirical evaluation with the intended meaning of "general" in AGI.

References

François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

Dan Hendrycks, Dawn Song, Christian Szegedy, Honglak Lee, Yarin Gal, Erik Brynjolfsson, Sharon Li, Andy Zou, Lionel Levine, Bo Han, Jie Fu, Ziwei Liu, Jinwoo Shin, Kimin Lee, Mantas Mazeika, Long Phan, George Ingebretsen, Adam Khoja, Cihang Xie, Olawale Salaudeen, Matthias Hein, Kevin Zhao, Alexander Pan, David Duvenaud, Bo Li, Steve Omohundro, Gabriel Alfour, Max Tegmark, Kevin McGrew, Gary Marcus, Jaan Tallinn, Eric Schmidt, and Yoshua Bengio. A definition of agi, 2025. URL `https://arxiv.org/abs/2510.18212`.

John Bissell Carroll. *Human cognitive abilities: A survey of factor-analytic studies*. Number 1. Cambridge university press, 1993.

Kevin S McGrew. Chc theory and the human cognitive abilities project: Standing on the shoulders of the giants of psychometric intelligence research, 2009.

Kevin S McGrew. Carroll's three-stratum (3s) cognitive ability theory at 30 years: Impact, 3s-chc theory clarification, structural replication, and cognitive–achievement psychometric network analysis extension. *Journal of Intelligence*, 11(2):32, 2023a.

Raymond Bernard Cattell. *Intelligence: Its structure, growth and action*, volume 35. Elsevier, 1987.

Ralph L Keeney and Howard Raiffa. *Decisions with multiple objectives: preferences and value trade-offs*. Cambridge university press, 1993.

Hiroaki Kitano. Biological robustness. *Nature Reviews Genetics*, 5(11):826–837, 2004.

Peter S Bullen. *Handbook of means and their inequalities*, volume 560. Springer Science & Business Media, 2013.

Timothy Z Keith and Matthew R Reynolds. Cattell–horn–carroll abilities and cognitive tests: What we've learned from 20 years of research. *Psychology in the Schools*, 47(7):635–650, 2010.

Kevin S. McGrew. Carroll's three-stratum (3s) cognitive ability theory at 30 years: Impact and clarification. *Journal of Intelligence*, 2023b.

Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Proc. CVPR*, 2009.

Dan Hendrycks, Collin Burns, Saurav Kadavath, Akul Arora, Steven Basart, Eric Tang, Dawn Song, and Jacob Steinhardt. Measuring mathematical problem solving with the math dataset. *arXiv preprint arXiv:2103.03874*, 2021.

Yonatan Bisk, Rowan Zellers, Jianfeng Gao, Yejin Choi, et al. Piqa: Reasoning about physical commonsense in natural language. In *Proceedings of the AAAI conference on artificial intelligence*, volume 34, pages 7432–7439, 2020.

Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

Astrid F Fry and Sandra Hale. Relationships among processing speed, working memory, and fluid intelligence in children. *Biological psychology*, 54(1-3):1–34, 2000.

Francois Chollet, Mike Knoop, Gregory Kamradt, Bryan Landers, and Henry Pinkard. Arc-agi-2: A new challenge for frontier ai reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

Mehran Kazemi, Bahare Fatemi, Hritik Bansal, John Palowitch, Chrysovalantis Anastasiou, Sanket Vaibhav Mehta, Lalit K Jain, Virginia Aglietti, Disha Jindal, Peter Chen, et al. Big-bench extra hard. *arXiv preprint arXiv:2502.19187*, 2025.

DeepMind. Gemini 3 pro model evaluation – approach, methodology & results. Technical report, DeepMind, 2025. URL `https://storage.googleapis.com/deepmind-media/gemini/gemini_3_pro_model_evaluation.pdf`. Accessed: 2025-11-27.

Anthropic. Claude sonnet 4.5 — system card. Web page, 2025. URL https://www.anthropic.com/claude-sonnet-4-5-system-card. Accessed: 2025-11-27.

Long Phan, Alice Gatti, Ziwen Han, Nathaniel Li, Josephina Hu, Hugh Zhang, Chen Bo Calvin Zhang, Mohamed Shaaban, John Ling, Sean Shi, et al. Humanity's last exam. *arXiv preprint arXiv:2501.14249*, 2025.

David Rein, Betty Li Hou, Asa Cooper Stickland, Jackson Petty, Richard Yuanzhe Pang, Julien Dirani, Julian Michael, and Samuel R Bowman. Gpqa: A graduate-level google-proof q&a benchmark. In *First Conference on Language Modeling*, 2024.

Mislav Balunović, Jasper Dekoninck, Ivo Petrov, Nikola Jovanović, and Martin Vechev. Matharena: Evaluating llms on uncontaminated math competitions. *arXiv preprint arXiv:2505.23281*, 2025.

Xiang Yue, Tianyu Zheng, Yuansheng Ni, Yubo Wang, Kai Zhang, Shengbang Tong, Yuxuan Sun, Botao Yu, Ge Zhang, Huan Sun, et al. Mmmu-pro: A more robust multi-discipline multimodal understanding benchmark. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 15134–15186, 2025.

Kaixin Li, Ziyang Meng, Hongzhan Lin, Ziyang Luo, Yuchen Tian, Jing Ma, Zhiyong Huang, and Tat-Seng Chua. Screenspot-pro: Gui grounding for professional high-resolution computer use. In *Proceedings of the 33rd ACM International Conference on Multimedia*, pages 8778–8786, 2025.

Zirui Wang, Mengzhou Xia, Luxi He, Howard Chen, Yitao Liu, Richard Zhu, Kaiqu Liang, Xindi Wu, Haotian Liu, Sadhika Malladi, et al. Charxiv: Charting gaps in realistic chart understanding in multimodal llms. *Advances in Neural Information Processing Systems*, 37:113569–113697, 2024.

Kairui Hu, Penghao Wu, Fanyi Pu, Wang Xiao, Yuanhan Zhang, Xiang Yue, Bo Li, and Ziwei Liu. Video-mmmu: Evaluating knowledge acquisition from multi-discipline professional videos. *arXiv preprint arXiv:2501.13826*, 2025.

The Terminal-Bench Team. Terminal-bench: A benchmark for ai agents in terminal environments, Apr 2025. URL https://github.com/laude-institute/terminal-bench.

John Yang, Kilian Lieret, Carlos E. Jimenez, Alexander Wettig, Kabir Khandpur, Yanzhe Zhang, Binyuan Hui, Ofir Press, Ludwig Schmidt, and Diyi Yang. Swe-smith: Scaling data for software engineering agents, 2025. URL https://arxiv.org/abs/2504.21798.

Victor Barres, Honghua Dong, Soham Ray, Xujie Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment, 2025. URL https://arxiv.org/abs/2506.07982.

Lukas Haas, Gal Yona, Giovanni D'Antonio, Sasha Goldshtein, and Dipanjan Das. Simpleqa verified: A reliable factuality benchmark to measure parametric knowledge. *arXiv preprint arXiv:2509.07968*, 2025.

Tyler A Chang, Catherine Arnett, Abdelrahman Eldesokey, Abdelrahman Sadallah, Abeer Kashar, Abolade Daud, Abosede Grace Olanihun, Adamu Labaran Mohammed, Adeyemi Praise, Adhikarinayum Meerajita Sharma, et al. Global piqa: Evaluating physical commonsense reasoning across 100+ languages and cultures. *arXiv preprint arXiv:2510.24081*, 2025.

A Normalized Cognitive Capacity Tables (Percentage Form)

This appendix provides a detailed quantitative analysis of model performance across the ten cognitive domains defined in the *A Definition of AGI* framework [Hendrycks et al., 2025]. All subdomain values are drawn directly from Hendrycks et al. [2025] and are normalized to express the percentage of human-level proficiency attained within each subdomain. Subdomain weights follow the fixed allocations specified in that framework, reflecting each domain's theoretical contribution to overall general intelligence; these weights may be refined in future extensions of the methodology.

Each subdomain carries a maximum weight that reflects its intended influence on the composite score. Subdomain raw results (drawb directly from [Hendrycks et al., 2025]) are normalized as:

$$\text{Normalized Score } (\%) = \frac{\text{Raw Score}}{\text{Weight}} \times 100,$$

so that 100% corresponds to human-equivalent proficiency within that subdomain.

Four domain-level aggregates. To represent both breadth and robustness, we report:

- **AM** — Unweighted arithmetic mean (breadth of capability).
- **WAM** — Weighted arithmetic mean using subdomain weight shares.
- **GM** — Unweighted geometric mean (robustness without weighting).
- **WGM** — Weighted geometric mean using subdomain weights.

For geometric means, let $p_i = \text{score}_i/100 \in [0, 1]$, $\tilde{w}_i = w_i / \sum_j w_j$, and use a stabilizer $\varepsilon = 10^{-6}$ to avoid collapse when any $p_i = 0$:

$$\text{GM} = 100 \times \Big(\prod_i \max(p_i, \varepsilon) \Big)^{1/n}, \qquad \text{WGM} = 100 \times \prod_i \max(p_i, \varepsilon)^{\tilde{w}_i}.$$

A.1 General Knowledge (K)

Table 3 shows high crystallized knowledge for both models, with GPT-5 improving cultural coverage to 50 and achieving strong robustness (GM/WGM 87.1). GPT-4's zero in culture drops geometric aggregates, exposing imbalance despite an 80.0 AM.

Table 3: General Knowledge (K): Normalized Subdomains with Weights and Aggregates

Model	Common	Science	Soc. Sci.	History	Culture	**AM**	**WAM**	**GM**	**WGM**
Weight	20	20	20	20	20	–	–	–	–
GPT-4	100	100	100	100	0	80.0	80.0	6.3	6.3
GPT-5	100	100	100	100	50	90.0	90.0	87.1	87.1

A.2 Reading and Writing (RW)

As detailed in Table 4, GPT-4's uneven orthography/usage yields a brittle GM (2.2). GPT-5 is uniformly strong across all subskills, reflected in perfect aggregates.

Table 4: Reading and Writing (RW): Normalized Subdomains with Weights and Aggregates

Model	Letters	Reading	Writing	Usage	**AM**	**WAM**	**GM**	**WGM**
Weight	10	30	30	30	–	–	–	–
GPT-4	0	67	100	33	50.0	60.0	2.2	16.0
GPT-5	100	100	100	100	100.0	100.0	100.0	100.0

A.3 Mathematical Ability (M)

Table 5 indicates GPT-4's failures in geometry/calculus drive near-zero GM/WGM, while GPT-5 closes all gaps, consistent with stronger symbolic and quantitative abstraction.

Table 5: Mathematical Ability (M): Normalized Subdomains with Weights and Aggregates

Model	Arithmetic	Algebra	Geometry	Prob.	Calculus	**AM**	**WAM**	**GM**	**WGM**
Weight	20	20	20	20	20	–	–	–	–
GPT-4	100	50	0	50	0	40.0	40.0	0.3	0.3
GPT-5	100	100	100	100	100	100.0	100.0	100.0	100.0

A.4 On-the-Spot Reasoning (R)

In Table 6, GPT-5 improves in deduction, ToM, and planning, but adaptive reasoning remains at 0. Low GM/WGM (5.5/19.0) indicate fragile, non-transferable reasoning.

Table 6: On-the-Spot Reasoning (R): Normalized Subdomains with Weights and Aggregates

Model	Deduction	Induction	ToM	Planning	Adapt.	**AM**	**WAM**	**GM**	**WGM**
Weight	20	40	20	10	10	–	–	–	–
GPT-4	0	0	0	0	0	0.0	0.0	0.0	0.0
GPT-5	100	50	100	100	0	70.0	70.0	5.5	19.0

A.5 Working Memory (WM)

As seen in Table 7, capacity is dominated by text; GPT-5 shows partial progress in visual and cross-modal maintenance, but GM/WGM remain low, signaling a bottleneck.

Table 7: Working Memory (WM): Normalized Subdomains with Weights and Aggregates

Model	Textual	Auditory	Visual	Cross-Mod.	**AM**	**WAM**	**GM**	**WGM**
Weight	20	20	40	20	–	–	–	–
GPT-4	100	0	0	0	25.0	20.0	0.0	0.0
GPT-5	100	0	25	50	43.8	40.0	1.9	3.2

A.6 Long-Term Memory Storage (MS)

Table 8 confirms no durable storage: all subdomains are 0 for both models. This undermines coherence and constrains non-compensatory aggregates.

Table 8: Long-Term Memory Storage (MS): Normalized Subdomains with Weights and Aggregates

Model	Assoc.	Meaningful	Verbatim	**AM**	**WAM**	**GM**	**WGM**
Weight	40	30	30	–	–	–	–
GPT-4	0	0	0	0.0	0.0	0.0	0.0
GPT-5	0	0	0	0.0	0.0	0.0	0.0

A.7 Long-Term Memory Retrieval (MR)

As shown in Table 9, partial fluency is offset by hallucination-prone retrieval, leaving GM/WGM near zero. Retrieval weaknesses reinforce the storage gap in Table 8.

Table 9: Long-Term Memory Retrieval (MR): Normalized Subdomains with Weights and Aggregates

Model	Fluency	Halluc. Avoid.	**AM**	**WAM**	**GM**	**WGM**
Weight	60	40	–	–	–	–
GPT-4	67	0	33.5	40.2	0.1	0.3
GPT-5	67	0	33.5	40.2	0.1	0.3

A.8 Visual Processing (V)

Table 10 indicates progress for GPT-5 in perception/generalization, but reasoning/spatial remain at 0. Low GM/WGM reflect this limitation.

Table 10: Visual Processing (V): Normalized Subdomains with Weights and Aggregates

Model	Percep.	Gen.	Reason.	Spatial	**AM**	**WAM**	**GM**	**WGM**
Weight	40	30	20	10	–	–	–	–
GPT-4	0	0	0	0	0.0	0.0	0.0	0.0
GPT-5	50	67	0	0	29.3	40.1	0.1	1.1

A.9 Auditory Processing (A)

In Table 11, GPT-5 reaches strong speech recognition and partial voice discrimination, but lacks phonetic/rhythmic/musical competence. The GM/WGM confirms incomplete auditory integration.

Table 11: Auditory Processing (A): Normalized Subdomains with Weights and Aggregates

Model	Phonetic	Speech Rec.	Voice	Rhyth.	Musical	**AM**	**WAM**	**GM**	**WGM**
Weight	10	40	30	10	10	–	–	–	–
GPT-4	0	0	0	0	0	0.0	0.0	0.0	0.0
GPT-5	0	100	67	0	0	33.4	60.1	0.0	1.4

A.10 Processing Speed (S)

Table 12 shows unchanged speed profiles across models: strong document-centric throughput (Re/Wr/Num) but near-zero simple/choice reaction and inspection time. GM/WGM (0.01) indicate real-time responsiveness remains a bottleneck.

Table 12: Processing Speed (S): Normalized Subdomains with Weights and Aggregates

Model	PS-S	PS-C	Re	Wr	Num	SRT	CRT	IT	CS	PF	**AM**	**WAM**	**GM**	**WGM**
Weight	10	10	10	10	10	10	10	10	10	10	–	–	–	–
GPT-4	0	0	100	100	100	0	0	0	0	0	30.0	30.0	0.01	0.01
GPT-5	0	0	100	100	100	0	0	0	0	0	30.0	30.0	0.01	0.01

A.11 Discussion

AM captures domain breadth, **WAM** preserves the weight structure adopted by Hendrycks et al. [2025], while **GM** and **WGM** emphasize robustness and balance across subskills. Across Tables 3–12, the geometric aggregates consistently expose residual brittleness in multimodal perception, long-term memory (storage and retrieval), and adaptive reasoning—even where GPT-5's AM/WAM approach human-level performance. This divergence between arithmetic and geometric summaries indicates that apparent gains are concentrated in a subset of faculties (e.g., crystallized knowledge and text-centric skills), with limited cross-modal integration and durable memory.

Benchmark	Description	Gemini 3 Pro	Gemini 2.5 Pro	Claude Sonnet 4.5	GPT-5.1
Humanity's Last Exam	Academic reasoning	**37.5**	21.6	13.7	26.5
ARC-AGI-2	Visual reasoning puzzles	**31.1**	4.9	13.6	17.6
GPQA Diamond	Scientific knowledge	**91.9**	86.4	83.4	88.1
AIME 2025	Mathematics (no tools)	**95.0**	88.0	87.0	94.0
MathArena Apex	Challenging math contest problems	**23.4**	0.5	1.6	1.0
MMMU-Pro	Multimodal understanding & reasoning	**81.0**	68.0	68.0	76.0
ScreenSpot-Pro	Screen understanding	**72.7**	11.4	36.2	3.5
CharXiv Reasoning	Information synthesis from charts	**81.4**	69.6	68.5	69.5
Video-MMMU	Video-based knowledge acquisition	**87.6**	83.6	77.8	80.4
Terminal-Bench 2.0	Agentic terminal coding	**54.2**	32.6	42.8	47.6
SWE-Bench Verified	Agentic coding (single attempt)	76.2	59.6	**77.2**	76.3
t^2-bench	Agentic tool use	**85.4**	54.9	84.7	80.2
FACTS Benchmark Suite	Grounding, parametric, MM, search	**70.5**	63.4	50.4	50.8
SimpleQA Verified	Parametric knowledge	**72.1**	54.5	29.3	34.9
MMLU	Multilingual QA	**91.8**	89.5	89.1	91.0
Global PIQA	Commonsense reasoning (100 languages)	**93.4**	91.5	90.1	90.9
MRCR v2 (8-needle)	Long-context (128k avg)	**77.0**	58.0	47.1	61.6

Table 13: Comparison of leading large-language and multimodal models on diverse reasoning, multimodal, agentic, and long-context benchmarks. Scores are taken from the publicly released *Gemini 3 Pro Model Evaluation Report* [DeepMind, 2025].

Methodologically, this paper uses the ten normalized domain scores reported by Hendrycks et al. [2025] as a shared empirical basis. We note, however, that these scores inherit several debatable design choices (e.g., potential conflation across subdomains and relatively low weights for adaptive planning). The contrast between AM/WAM and GM/WGM within Tables 3–12 further shows that arithmetic aggregation can *inflate* overall impressions of capability relative to coherence-oriented measures.

Therefore, the $\mathrm{AGI}_{\mathrm{AUC}}$ value of approximately 24% for GPT-5 should be interpreted as an *upper bound* on current general competence. As demonstrated across domains in Appendix A, even the ten constituent cognitive scores display inherent inflation when converted from weighted arithmetic means to geometric means or unweighted aggregates. For instance, in auditory processing, the reported weighted mean of roughly 60% drops to 33% under unweighted averaging and collapses near zero when using geometric means (weighted or unweighted). A similar pattern holds for processing speed (from 30% to near zero under geometric averaging), visual processing (from 40% weighted to 29% unweighted and again near zero geometrically), and the working- and long-term memory domains, which remain effectively null regardless of weighting. Even on-the-spot reasoning, nominally at 70% by arithmetic mean, falls sharply to 19% (weighted geometric) or 5% (unweighted geometric), revealing fragile compositional performance masked by additive aggregation. These discrepancies underscore that the nominal 24% $\mathrm{AGI}_{\mathrm{AUC}}$ should be viewed as conservative yet still optimistic, given the inflation present at the domain-score level.

In the main paper, we therefore focus on the *aggregation principle* rather than revising the underlying measurements. The reported domain scores are treated as reasonable, though potentially optimistic, proxies for current model ability, and are used to demonstrate how coherence-aware aggregation (culminating in $\mathrm{AGI}_{\mathrm{AUC}}$) provides a more faithful summary of systemic competence. Future work should revisit the domain taxonomy, explore nonlinear scaling and adaptive weightings based on empirical dependencies, and develop diagnostic tools to identify which deficiencies most constrain $\mathrm{AGI}_{\mathrm{AUC}}$, ultimately guiding targeted architectural or training interventions.

B Beyond the CHC-domain Scores

To illustrate that the coherence-based framework extends beyond the CHC taxonomy, we apply the same aggregation methodology to the 17 percentage-based benchmarks reported in Table 13, treating each as an independent capability axis. These benchmarks are drawn from the publicly released

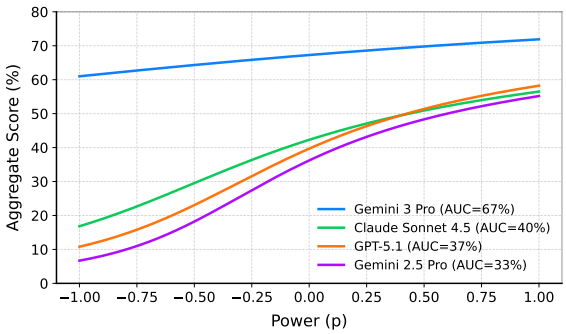

Figure 2: Comparison of model performance across aggregation exponents p. Curves show AGI_p values derived from the 17 benchmarks in Table 13.

Gemini 3 Pro Model Evaluation Report [DeepMind, 2025], which compares Gemini 3 Pro against GPT-5.1, Gemini 2.5 Pro, and Claude Sonnet 4.5 [Anthropic, 2025].

These benchmarks span a broad spectrum of reasoning [Phan et al., 2025, Chollet et al., 2025, Rein et al., 2024, Balunović et al., 2025], multimodal perception [Yue et al., 2025, Li et al., 2025, Wang et al., 2024, Hu et al., 2025], agentic behavior [Team, 2025, Yang et al., 2025, Barres et al., 2025], knowledge and grounding [Haas et al., 2025, Chang et al., 2025], and long-context understanding.

Let $b_i \in [0, 100]$ denote a model's score on benchmark i for $i = 1, \ldots, 17$. We compute AGI_p using the same continuous family of compensability exponents $p \in [-1, 1]$ introduced in Eq. (2), with stability parameter $\varepsilon = 10^{-6}$. We then evaluate the corresponding AGI_p curves and compute the integrated coherence score $\mathrm{AGI}_{\mathrm{AUC}}$ (Eq. 3) using the trapezoidal rule.

Results. The 17-benchmark evaluation produces a pattern that closely mirrors the CHC-domain analysis. Positive-p aggregation (including the arithmetic mean at $p = 1$) consistently overstates overall capability by masking unevenness across domains, whereas negative-p aggregation sharply penalizes weak dimensions, revealing the true coherence profile of each model. Across the p-sweep, $\mathrm{AGI}_{\mathrm{AUC}}$ yields a more stable and discriminative ordering than the arithmetic mean: models with similar averages often diverge substantially in coherence, exposing persistent bottlenecks that single-number averages fail to capture.

Notably, Gemini 3 Pro exhibits the highest coherence across the full range of p values. Its AGI_p curve remains consistently above those of the other models, reflecting improvements not only in overall performances but also in the hardest benchmarks, precisely the regime where coherence penalties matter most. This demonstrates that Gemini 3 Pro's gains are broad rather than concentrated, leading to a substantially stronger AUC score (67%).

These findings confirm that the coherence-based evaluation introduced in this work is not dependent on the CHC taxonomy. Instead, it generalizes naturally to heterogeneous benchmark suites, offering a unified and domain-agnostic measure of holistic capability. By rewarding balanced competence and penalizing systemic weaknesses, $\mathrm{AGI}_{\mathrm{AUC}}$ provides a conservative and robust indicator of emerging general intelligence.

It is important to note that any aggregate score, whether derived from 17 benchmarks or hundreds, must be interpreted in light of the benchmark suite it summarizes. No numerical percentage can meaningfully reflect "progress toward AGI" unless the underlying tasks collectively span all essential cognitive faculties. When key abilities such as continual learning, long-term memory, or autonomous planning are missing, the resulting aggregate reflects performance only within the tested scope. A system with a bottleneck in any essential faculty cannot be described as achieving a given fraction of AGI, but rather as exhibiting unequal development across capabilities. Within any benchmark suite, however, coherence-based aggregation remains the most informative and principled evaluation method: it highlights imbalances, exposes hidden bottlenecks, and rewards the balance and robustness that are central to general intelligence.