Beyond the AGI Claim: What Would Actually Count as Evidence for Artificial General Intelligence?

Aman Bhaskar

Independent researcher

21 October 2026

Abstract

Evidence of broad AI performance can support different claims about intelligence. This article asks when such evidence justifies an artificial general intelligence (AGI) attribution and which conclusions depend on the chosen account of intelligence. Through conceptual analysis, we reconstruct three published claims: Bubeck and colleagues' early GPT-4 interpretation, Hendrycks and colleagues' cognitive-parity criterion, and Zhou and colleagues' prediction of language-model response success. Three checks address the account's adequacy, evidential support and preservation of evaluation conditions and claim scope. The analyses distinguish immediate correction from broader reuse of task prerequisites, support for a storage score from its effect on a classification endpoint, and assessor prediction from learning by the assessed model. Comparing one evidence package across three accounts identifies different evidential requirements without establishing different AGI verdicts. The central argument is that materially changed requirements call for reassessment, while a disputed premise may affect one conclusion without defeating a separately supported finding. The contribution is a critical methodological synthesis with traceable case analyses. Purposive selection and selective source inspection limit the judgments; the proposed method and its practical benefits have not been empirically validated.

Keywords: artificial general intelligence; AI evaluation; conceptual analysis; validity; evidence; claim scope

1. Introduction

1.1 The problem and motivation

A missing horn offered a concrete test of an AI system's capabilities. Bubeck et al. removed the horn from code specifying a unicorn drawing, transformed its coordinates and stripped textual annotations. Development-stage, text-input-only GPT-4 then supplied TikZ drawing code that restored the horn [1]. The reported success provides evidence of a bounded symbolic-spatial capability. Within their wider investigation, the authors also proposed that the system could reasonably be viewed as early, incomplete artificial general intelligence (AGI) [1]. Moving from such achievements to that broader interpretation raises a question about what the evidence establishes.

Assessing this interpretation requires identifying the account of intelligence it invokes. Different accounts impose different evidential demands: broad proficiency, including learning and metacognition [2]; efficient skill acquisition relative to priors, experience, scope and generalization difficulty [3]; or human cognitive versatility and proficiency [4]. These overlapping targets are not interchangeable. An observation relevant to one may also support another, but that support requires justification against the intended requirements.

These differences affect what researchers must establish and which conclusions their findings support. Uncertainty about a broader attribution does not by itself erase supported task competence; equally, successful performance does not settle every interpretation proposed for it. This motivates examining the reasoning that connects reported achievements to general-intelligence claims: which requirements the evidence addresses, which assumptions matter, and which conclusions remain supported when a premise is disputed.

1.2 Research question and intended contribution

The central problem is how the requirements of an interpretation affect the evidence needed to support it. We ask: Under what conditions does evidence of broad AI performance warrant an AGI attribution, and which conclusions depend on the chosen account of intelligence? A supporting question asks which conclusions require reassessment when a premise is disputed, and which separately supported findings remain warranted.

An AGI attribution may be warranted when the chosen account is defensible for its intended use, the evidence sufficiently supports its requirements, and the conclusion preserves relevant evaluation conditions and uncertainty. Changing accounts neither invalidates the reported findings nor automatically warrants the new interpretation; further justification or reanalysis may suffice without new measurements. This position builds on DAF-AGI's definition-sensitive adjudication [5] and argument- and claim-centered validity [6, 7].

We examine Bubeck et al.'s qualified early, incomplete-AGI interpretation [1], Hendrycks et al.'s operational cognitive-parity criterion [4], and Zhou et al.'s prediction of language-model response success [8]. The contribution is a critical synthesis with three linked applications: distinguishing immediate correction from broader prerequisite reuse, separating support for a component score from its effect on a classification endpoint, and preserving the distinction between assessor prediction and subject learning. A complementary comparison holds one inspected evidence package fixed across three accounts. These applications identify the conclusions affected by a disputed premise using established validity principles. The method's scope and limitations are discussed in Section 7.

1.3 Scope and boundaries

This conceptual methods analysis examines selected accounts of general intelligence and published claims and proposed criteria through worked cases. Judgments are specific to inspected evidence and reported system and evaluation conditions, with uncertainties retained and any extension requiring justification. The investigation does not comprehensively assess contemporary AI systems or establish a universal AGI verdict. Neither consciousness nor human-like mechanisms are imposed as universal prerequisites for AGI; requirements invoking them must be justified for the account's intended use. Section 2 develops the conceptual foundations for distinguishing proposed meanings, measurement and interpretations of capability, learning and understanding.

2. Conceptual foundations

2.1 Proposed meaning, measurement and empirical claim

A working definition proposes a meaning; proposing it does not establish scientific usefulness. Drawing on Carnap, Wang assesses adequacy through resemblance to established usage, precision of application, research fruitfulness and conceptual simplicity [9]. Exploratory research can use provisional concepts; public certification or adjudication requires sufficiently determinate application conditions, as DAF-AGI emphasizes [5]. Following Wallach et al., we distinguish specifying the concept, operationalizing it through assessment procedures, and obtaining measurements [10]. Empirical interpretation concerns what those measurements warrant concluding.

Consider a hypothetical system scoring 90/100 on puzzles. Accurate scoring supports that result; generalization requires justified task coverage, domain relevance and supporting evidence; interpreting performance as broad cognition requires evidence covering the intended construct; inferring AGI under definition D additionally requires evidence addressing D's requirements and thresholds. Validation evaluates evidential support for proposed score interpretations and uses [6]. Test failure establishes criterion failure only through a justified assessment and decision rule; neither automatically establishes definitional inadequacy. Evidence revealing neglected capacities or unsuitable boundaries may nevertheless motivate conceptual revision. Broader claims require stronger support, not categorical rejection [6]. Here, warrant means evidential support for a stated interpretation and use. Our analysis evaluates specific claim-evidence pairs, distinguishing justified, insufficiently supported and unresolved interpretations while preserving separately supported, source-reported capability findings.

2.2 Three representative accounts of general intelligence

Three representative accounts highlight different requirements for AGI claims: capability breadth and performance, skill-acquisition efficiency, and human cognitive parity. Goal achievement and constrained adaptation provide additional context rather than an exhaustive taxonomy.

Morris et al. classify systems by capability breadth and human-relative proficiency. Their general category includes cognitive and metacognitive tasks, including learning new skills. Competent AGI requires performance at least at the 50th percentile of appropriately skilled adults on most cognitive tasks. However, task selection and sufficient generality remain to be operationalized: the classification is not a completed benchmark [2, Sections 3-5].

Chollet instead characterizes intelligence as skill-acquisition efficiency within a scope, relative to priors, experience, and generalization difficulty. Practical comparisons use a shared task scope, a skill threshold achievable by both systems, and sufficiently similar knowledge priors. Similar priors are a recommendation motivated by the difficulty of quantifying prior knowledge, not a strict requirement of his formalism. Achieved skill alone does not establish acquisition efficiency [3, Sections II.2.1 and II.3.1].

Hendrycks et al. define AGI through the cognitive versatility and proficiency of a well-educated adult, operationalized using domains informed by Cattell-Horn-Carroll (CHC) theory. Their comparator denotes mastery across the tested dimensions, not simply a degree holder. They acknowledge selective coverage, linguistic and cultural limitations, and discretionary weighting; cognitive profiles are recommended because aggregate scores can conceal bottlenecks [4, Sections 1 and 13]. For this analysis,

attributing corresponding abilities to humans and AI from adapted tests requires justification; different strategies alone do not invalidate the comparison.

Legg and Hutter formalize goal achievement across computable, reward-summable environments, weighted by description complexity relative to a reference machine. Noncomputable Kolmogorov complexity prevents treating the exact measure as an executable benchmark [11, Sections 3.1 and 3.3]. Wang emphasizes experience-guided adaptation under insufficient knowledge and resources; an adaptive adjustment need not improve subsequent performance [9, Section 3.2].

The three central accounts agree that isolated excellence does not establish general intelligence: breadth matters, and learning is relevant rather than excluded by a capability focus [2, 3, 4]. Learning is itself a capability; the difference lies in whether demonstrating that capability also establishes the efficiency with which it was acquired. They differ in the evidential role assigned to acquisition history. In our comparison, broad demonstrated proficiency, including tested learning capabilities, may support a Morris-style capability classification while leaving Chollet-style acquisition efficiency unresolved when relevant priors and experience are unaccounted for [2, 3]. This is a conditional interpretive consequence, not an empirical verdict. Equal aggregate accuracy therefore neither establishes equal acquisition efficiency nor erases demonstrated performance. The analysis must identify the target of a claim and whether its evidence supports that interpretation.

2.3 Capability, learning and understanding

Capability concerns task competence; skill acquisition concerns developing or adapting competence under specified learning conditions; understanding attributes semantic or cognitive capacities under a stated account. These claims overlap without forming a hierarchy. Morris et al.'s capability-based AGI classification does not require human-like processes; it is not itself a functional theory of understanding [2]. Functional accounts instead characterize understanding through the roles capacities play, potentially assessed behaviorally. For Chollet, acquisition efficiency depends on priors, experience, scope and generalization difficulty; final performance alone does not establish it [3]. Bender and Koller argue that form-only training cannot learn meaning understood through relations linking expressions to communicative intent [12]. This restriction does not automatically cover systems receiving relevant grounding information. Mitchell and Krakauer examine competing attributions and shortcut-based challenges while leaving nonhuman understanding open [13].

When a claim's requirements change materially, the evidence must be reassessed against those requirements. The change alone neither preserves nor defeats the claim. Evidence may support capability, acquisition and understanding together when their respective requirements are met. Shortcut evidence challenges a capability-based classification when it undermines capabilities that the account requires, rather than merely revealing nonhuman mechanisms. Unresolved acquisition or understanding should therefore affect an AGI judgment only where the operative account depends on it. This proposition synthesizes claim-centered validity and definition-sensitive adjudication [5, 7]; the case applications, rather than a new validity principle, supply the article's specific contribution.

3. Method and related work

3.1 Source and case selection

This conceptual methods article examines how the requirements of an AGI account affect interpretation of reported performance. It asks which conclusions a disputed premise can change and which narrower findings remain supported. The analysis identifies conditional evidential requirements rather than proposing a universal AGI threshold. Its cases are selected purposively. The selected sources provide explicit claim passages, declared conceptual commitments and contrasting inferential targets. Bubeck et al. connect reported cross-domain capabilities in development-stage GPT-4 to a qualified early, incomplete-AGI interpretation [1]. Hendrycks et al. propose a human-referenced cognitive-parity definition and an operational scoring framework [4]. Zhou et al. investigate external prediction of performance by subject large language models (LLMs) [8]. These aspects overlap within the sources, but the selected claims address different targets.

The predictive case examines whether a bounded inference can be retained without establishing a cognitive explanation or AGI status. This contrast complements classification background from Morris et al. [2]; it supplies neither an experimental control nor a comparison of research quality. These analytical roles motivated case selection. No selection protocol was registered in advance, and this rationale cannot exclude selection influenced by anticipated outcomes. Source qualifications and safeguards inform examination of alternatives. This is neither a systematic review nor a representative survey; the cases support no prevalence estimates or claims that this procedure outperforms existing approaches.

3.2 Closest prior work and the contribution boundary

The closest antecedents already explain why warrant depends on the claim being assessed. DAF-AGI addresses definition-dependent adjudication and the commitments of operative criteria [5]. Salaudeen et al. connect evaluation evidence to intended criteria, constructs and uses while preserving appropriately scoped findings [7]. Kane supplies the argument-based architecture of inferences and supporting assumptions [6]. Reynolds distinguishes composition, convergence and replacement when evidential endpoints or assumptions change [14]. Alzahrani makes claim-specific evidence, verdicts and boundaries explicit [15]. Freiesleben and Zezulka likewise adapt argument-based validity to benchmark inferences, requiring specified validity conditions and evidence for and against them [16]. These approaches anticipate the organizing proposition defended here.

Chollet distinguishes achieved skill from acquisition efficiency [3]. Gignac and Szodorai argue, under a novelty-sensitive account, that practiced achievement need not establish intelligence [17]. Wallach et al. distinguish conceptual specification from operationalization and measurement [10], while Majumdar's proposed distribution-indexed assessment makes task and resource conditions explicit [18]; we do not rely on its broader theorem conclusions. A complementary challenge concerns construct meaning: Freiesleben argues that inferential validity alone provides limited resources for theorizing complex LLM capabilities and advocates nomological networks: systems of theoretical and empirical relationships among constructs and observations [19]. Exposing a bridge does not supply the substantive theory needed to defend it.

The contribution is a comparative analysis of three consequences of a premise dispute. DAF-AGI supplies purpose-relative adjudication; argument-based and claim-centered

validity supply the evidence-to-claim analysis; condition-preserving inference and claim records supply the reporting architecture [5, 6, 7, 14, 15, 16]. The cases use these approaches to identify which conclusion an uncertainty affects and why. The added work consists of the specific source reconstructions, the storage-only endpoint calculation, and the comparison of account requirements against a fixed evidence package. Appendix C identifies these applications beside their antecedents. The contribution lies in these applications and their comparison, rather than in the underlying validity principles. Section 7 addresses the limits of the originality claim.

Thesis. A material change in claim requirements calls for reassessing the supporting argument. It neither automatically preserves nor defeats the claim, and a disputed premise may affect one conclusion while leaving another supported.

Practical hypothesis (untested). Compared with evaluation using the same evidence and resources without explicit premise reconstruction, the procedure increases recognition of material requirements and reduces unsupported stronger attributions without materially increasing rejection of warranted bounded findings. This is a prospective prediction, not a result of the case analyses. Section 6.3 specifies separate outcomes and trade-offs.

3.3 Procedure, evidence boundaries and uncertainty

Each case uses the same record: target claim, source-explicit commitments, reconstructed assumption, inspected evidence boundary, judgment and revision condition. Source-explicit commitments are claims or requirements stated by the source; reconstructed assumptions are our proposed evidence-to-claim bridges. An additional proposal is marked separately only where introduced. The main text reports the operative commitments and assumption; Appendix B supplies the corresponding six-field record and detailed anchors.

Three complementary checks assess (1) the account's adequacy for its stated purpose, (2) support for the intended interpretation, and (3) preservation of scope, system conditions and claim strength. A decision consequence, such as endpoint sensitivity, is reported separately. This procedure adapts argument-based validation [6], claim-centered validity [7] and claim-specific verification [15].

To assess materiality, we state a plausible alternative to a premise, identify the requirement it affects, and examine the consequence for each named conclusion. A premise is material to a conclusion when that substitution changes the support, scope or decision relevant to that conclusion. The assessment therefore requires a plausible alternative and an explicit account of its consequence; merely conceiving an alternative is insufficient. In Case B, the storage-credit substitution affects the component interpretation while leaving the conditional endpoint decision unchanged.

Each judgment names its evidence boundary. An identified shortfall in a specified evidence package warrants an insufficient-support judgment; consequential uncertainty about uninspected support warrants an unresolved judgment. Neither establishes that the target claim is false. Table A1 defines these statuses. The records are retrospective, and no system experiments, preregistration, independent evaluator-agreement assessment or comparative test of the procedure was conducted. To repeat the analysis, a reader should use the editions and anchors in Appendix B, transcribe the claim and qualifications, distinguish source commitments from reconstructed premises, apply the three checks, state a plausible premise substitution, and record its consequence and revision condition for each conclusion. Appendix D identifies each source's role. These

records allow readers to inspect and reapply the analysis. Independent agreement remains untested.

4. A framework for definitions, evidence and claims

4.1 Check 1: Is the definition adequate for its stated purpose?

A working definition is not adequate merely because it is stipulated, nor must it identify a unique essence of intelligence. Wang adapts Carnap's similarity, exactness, fruitfulness and simplicity requirements to working definitions [9]; DAF-AGI conditions assessment on function, distinguishing adjudication from coordination, forecasting and risk-gating [5]. Drawing on these antecedents, our first check asks whether meaning is coherent and use identifiable; whether scope, exclusions and achievement standards are defensible for that use; and whether application conditions are sufficiently determinate for the proposed conclusion. These qualitative questions require justification, not universal consensus or complete precision.

Morris et al. illustrate the distinction [2]. Their taxonomy separates proficiency from generality and requires metacognitive capabilities, including learning new tasks, without requiring human-like processes, consciousness or physical task performance. Embodiment may nevertheless support, or be necessary for, acquiring some cognitive capacities. Competent AGI requires at least 50th-percentile performance relative to appropriately skilled adults on most cognitive tasks. The authors leave the task domain, sufficient proportion and specification of mandatory benchmark tasks unresolved; this does not make metacognitive capabilities optional. These unresolved specifications limit definite classification, although the taxonomy remains useful for provisional comparison.

Two questions must remain separate. Fidelity asks whether the evidence satisfies an account's declared requirements. Adequacy asks why those requirements justify the breadth implied by the label and its intended use. A narrow task definition may be easy to satisfy without warranting a general-intelligence attribution. For the second question, proponents should justify the task population and comparator, identify consequential exclusions, and explain why these boundaries serve the claimed use [5, 9]. This requirement does not supply a neutral definition of generality: competing answers remain substantive disagreements to be argued rather than settled by terminology.

Exploratory accounts may retain disclosed provisional boundaries. Definite adjudication needs defensible conventions where unresolved choices would change the verdict [5]. The comparison assesses the sources' stated commitments and identifies open questions about their adequacy, without selecting one account as exclusively authoritative. Check 2 separately asks whether the evidence supports an interpretation under those commitments.

4.2 Check 2: What bridges the evidence to the intended interpretation?

Assessing evidential support requires identifying the observed result, the supporting premise that connects it to a claim, and the intended interpretation. We call that supporting connection an evidential bridge. Kane makes these inferences and assumptions the focus of validation [6]; Cronbach and Meehl distinguish criterion prediction from interpretations of theoretical attributes [20]. Following Salaudeen et al., we distinguish directly measured criteria, prediction of another measurable criterion and construct-targeted claims [7]. Direct alignment still requires justified coverage and conditions. Validated prediction can support a bounded outcome claim without identifying its cognitive explanation; construct mediation is not universal. When invoked, both the construct attribution and its connection to the target require support.

Reconstruction makes material premises explicit rather than treating scores as self-interpreting. Bean et al. recommend justifying decisions connecting phenomenon, task, items, implementation and validity claims [21]. Under Chollet's acquisition-efficiency account, achieved skill requires interpretation relative to priors, experience, scope and generalization difficulty [3]. These requirements guide scrutiny of alternatives and safeguards. Unreported conditions remain unknown; hypothetical alternatives do not establish actual failure. Further, an explicit inferential chain does not itself establish construct meaning. Freiesleben's objection concerns the theoretical and empirical relationships invoked by complex capability claims, including whether relationships drawn from human cognition transfer to AI [19]. Our procedure organizes that question without settling it or requiring a cognitive explanation for every predictive claim.

The TikZ example illustrates interpretation of an actual reported result. Bubeck et al. report GPT-4 restoring a removed unicorn horn after the coordinates were transformed and textual annotations removed [1]. Our reconstructed bridge is that appropriate placement uses task-relevant spatial relationships in the supplied geometry. The safeguards support a bounded symbolic-spatial interpretation; they do not identify a particular internal mechanism. A local-geometric solution would be compatible with that interpretation. The example therefore shows why a mechanism need not be uniquely identified to support a specified capability finding. Section 5.1 examines the source's broader generality language and its reported feedback behavior.

4.3 Check 3: Does the claim preserve its conditions and definition?

After assessing the original bridge, we distinguish restatement, projection to new tasks or systems, reinterpretation under another account, and combination with further evidence. Our proposed diagnostic asks whether restatement changes the quantity, population, system, definition, conditions or strength of the inference. If it does, the additional requirement and its supporting justification must be identified. Context-preserving shorthand remains permissible; the diagnostic is a qualitative aid rather than a validated rule.

DAF-AGI illustrates definition-dependent adjudication [5]. Evidence supporting D_1 may support D_2, but the original verdict alone does not establish satisfaction of D_2's different requirements. Existing evidence or justified reanalysis may suffice; universal definitional agreement is unnecessary. Majumdar's proposed predicate, AGI(A | T, μ, Π, B), indexes assessment to task family, distribution, performance functional and resources, with quantitative thresholds, tolerances and perturbation conditions also fixed [18]. Changed indices require scrutiny, not an automatic adverse verdict; no broader theorem conclusion is invoked.

For chains of inference, Reynolds requires aligned endpoints or justified connections, compatible assumptions, and attention to dependence and uncertainty [14]. Convergent evidence supports the same claim without forming that chain; direct target-setting evidence may replace an upstream projection rather than validate it. Dependence remains relevant to corroboration.

Case C illustrates a boundary that must be preserved: Zhou et al.'s task holdouts concern the external assessor, while the predicted outcomes concern subject LLMs [8]. Retaining the actor and training boundary preserves the source's predictive target. It does not require novelty to the subjects, which is a different claim. This distinction is part of the source's own design; our application shows which qualification matters rather than attributing an error to its authors. Likewise, Case B's conditional score calculation concerns its specified historical configuration and scoring rules, not a general verdict about later systems [4].

4.4 Positive warrant and the strongest objections

Working accounts can guide research without universal agreement, while definite adjudication requires sufficiently determinate application conditions. A provisional positive judgment requires the following: specify the operative account and intended use; defend coverage, comparators, thresholds and system conditions; evaluate whether observations meet those requirements; and assess material counterevidence. Each step requires substantive justification; completing a record alone does not certify AGI. Table 1 gives account-specific examples of evidence that could support a judgment without requiring exhaustive testing.

Table 1. Account-specific evidence and decision questions. The evidence designs are our methodological proposals informed by [2, 3, 4], not completed benchmarks or source-mandated universal protocols.

Operative target	Evidence that would bear positively	Material challenge and decision consequence
Morris-style capability classification [2]	Human-relative proficiency over a justified broad task domain, including learning and metacognition, under declared assistance and resource conditions. Operational conventions must specify task coverage and what counts as “most.”	Failure on required coverage or proficiency can defeat that classification. A different acquisition history alone does not. A result may support a declared operational extension without completing Morris's unfinished benchmark specification.
Chollet-style acquisition efficiency [3]	Acquisition toward a shared skill threshold over a specified scope, interpreted relative to relevant priors, experience and generalization difficulty. Comparable priors are a practical recommendation where full quantification is infeasible.	Unaccounted task-specific preparation may prevent attributing efficiency even where final proficiency is supported. The account supplies no single universal AGI cutoff here; a categorical attribution needs a separately defended threshold.
Hendrycks-style cognitive parity [4]	Domain-level evidence and configuration-based arguments supporting the operational mastery standard, with stated weights, resource rules and the framework's 100% endpoint.	Unmet required dimensions can prevent endpoint satisfaction. Satisfying the operational score would still require defending its adequacy for the intended cognitive-parity interpretation.

The strongest objection is that open-ended scrutiny makes positive judgments unattainable. A provisional decision to stop collecting evidence should identify the

remaining uncertainties, explain why they do not defeat support at the claimed strength, and state what finding would reopen the judgment. Naming a conceivable alternative is insufficient: its plausibility and effect on criterion satisfaction or intended use require explanation. Conversely, a material shortfall should not be excused by calling all evidence provisional. Missing information or consequential criterion ambiguity may block assessment without demonstrating system failure.

A second objection is that the procedure relocates disagreement into adequacy and materiality. Transparent records expose these judgments without settling them. A criterion revision should therefore state its target-based reason, timing and decision consequence, with comparable treatment of comparable cases. A changed target receives a separately identified judgment. Section 4.5 explains how the checks combine; the wider limitations are consolidated in Section 7.

4.5 Combining the checks in a claim-specific judgment

The checks are complementary questions about one argument, not independent tests whose results are added or averaged. Check 1 identifies the account's requirements and their justification. Check 2 examines support for the intended interpretation. Check 3 asks whether that interpretation is preserved in the conclusion or a subsequent use. An omission affecting several checks should be recorded once, with each consequence identified. An unresolved account boundary can block a broad judgment even when a narrower performance result is supported.

The synthesis therefore produces judgments for named claims rather than an overall source or system score. Table 2 uses Case B to show mixed outcomes. Its arithmetic can be assessed while the historical component estimate and the adequacy of the cognitive-parity interpretation remain separate questions. A favorable result on one question cannot substitute for support on another.

Table 2. Claim-specific outcomes within Case B, using the reported 57% aggregate and ten-point storage allocation [4]. These are our analytical judgments, not additional source-reported results.

Target judgment	Evidence and assessment	Consequence
Historical zero storage credit	Task examples and prescribed conditions are available, but administration and persistent-state support remain incompletely inspected.	Unresolved within this inspection; no source-wide defect is established.
Effect of a storage-only substitution on the endpoint	Replacing zero with full ten-point credit gives 57 + 10 = 67 < 100, holding all other scores and rules fixed.	This substitution cannot reverse the conditional below-endpoint result; other scores are not validated by the calculation.
General cognitive-parity interpretation	Domain mapping, weights and mastery criteria require justification for the intended human-AI comparison.	The arithmetic does not settle conceptual adequacy or establish the remaining component estimates.

The storage premise is consequential for a component interpretation but cannot change this endpoint decision under the stated substitution. More generally, for a fixed additive scoring rule, a disputed component cannot close a threshold gap larger than its maximum possible increase. This elementary bound applies only while other scores, weights and the endpoint remain fixed. A change in those assumptions requires a new assessment.

5. Comparative case analyses

5.1 Case A: Broad performance and an early AGI interpretation

Target and source-explicit commitments. Bubeck et al.'s abstract proposes that development-stage, text-input-only GPT-4 could reasonably be viewed as early, incomplete AGI; §10 affirmatively attributes a form of general intelligence [1]. The source invokes breadth, human-relative proficiency and learning, with disclosed planning and learning limitations.

Check 1: Adequacy for the stated purpose. These commitments provide an identifiable exploratory target: investigate cross-domain performance relative to human capabilities while acknowledging incompleteness. This supports using the conception to organize exploratory evidence, rather than establishing its adequacy for certification. Required breadth, proficiency comparators and acceptable incompleteness remain insufficiently determinate for a definite classification. Morris and Chollet provide comparison accounts, not replacement criteria.

Check 2: Evidence-to-claim support. Our reconstructed assumption is that the inspected tasks and interactions adequately sample the capabilities invoked by the interpretation at its declared strength. The TikZ restoration, mathematics probes and TextWorld trace support selected source-reported competence and feedback responsiveness; they do not establish sufficient coverage or independent corroboration of the broader attribution [1]. The feedback trace permits a closer examination of the learning claim. Section 5.2.2, p. 52, says GPT-4 generalizes between actions and, after learning that chopping requires a knife, “does not make the same mistake afterward.” Appendix F.2.3, pp. 152-153, records successful chopping after correction, followed by slicing commands that again omit the knife argument and then reveal an unmet ingredient-possession requirement. The identical corrected chopping command is not repeated incorrectly. Under that narrow reading, the trace is consistent with the source's non-repetition statement. If generalization between actions is interpreted as reliable reuse of the knife and possession prerequisites across these actions, the later omissions qualify that broader reading. The qualification applies to the broader interpretation reconstructed here, not to the narrow source statement. Immediate correction remains supported; absent learning is not established. Reported one-shot success is also retained, although the supplied demonstration is omitted from the printed excerpt [1, Appendix F.2.4, p. 154].

Check 3: Preservation of conditions and strength. The conclusion must retain the historical development-stage system, text-input boundary, selected-task coverage and early, incomplete qualification. The source itself restricts generalization from its limited environments [1, Section 5.2.4]. A summary should not relabel these bounded learning outcomes as acquisition efficiency under Chollet's account [3].

Judgment and revision condition. The inspected demonstrations support the stated bounded source-reported findings but do not by themselves establish account-wide coverage. Whether the broader source evidence warrants the qualified AGI attribution remains unresolved within this inspection, particularly while the intended breadth and proficiency are under-specified. A narrower reading of the learning statement removes that particular objection without settling the broader attribution. Clarifying the requirements, inspecting the complete demonstration and logs, or considering evidence matched to those requirements could revise the judgment. Appendix B1 records the inspection boundary.

5.2 Case B: Cognitive parity and an operational AGI criterion

Target and source-explicit commitments. Hendrycks et al. define AGI through human cognitive versatility and proficiency, operationalized through mastery across tested dimensions and a 100% endpoint [4]. The selected component claim assigns zero storage credit to historical GPT-4/GPT-5 assessments. The authors describe the scores as estimates and permit evaluator judgment.

Check 1: Adequacy for the stated purpose. CHC-informed domains and equal weights provide an explicit rationale for prioritizing breadth. We accept this as an intelligible organizing rationale, not as validation of AI-human cognitive parity. Whether adapted tasks, domain coverage and weights justify that interpretation remains open. Consequently, the operational endpoint can be discussed conditionally without treating its conceptual adequacy as established.

Check 2: Evidence-to-claim support. Our reconstructed assumption is that task evidence or a sufficiently justified configuration argument supports the relevant acquisition, persistence and retrieval limitations behind the storage estimate. The authors distinguish working-memory compensation for storage deficits from external-search compensation for unreliable retrieval, while emphasizing persistent experiential memory [4, Section 13]. Appendix F prescribes new-session testing with external tools disabled, and F.1.2 illustrates delayed personalization. These specify assessment conditions but do not themselves document administration. Historical execution, personalization outcomes and persistent-state settings remain incompletely inspected. A new session alone does not establish lack of retained state; conversely, a documented configuration limitation could justify non-attainment without administering every example. Support for the storage estimate remains unresolved within this inspection; this does not establish that the estimate is incorrect.

Check 3: Preservation of conditions and strength. A summary must retain the assessed historical system and distinguish a component estimate from general memory incapacity. Figure 1 identifies GPT-5's answers as using Auto mode; it does not establish every subtest's persistent-state configuration. The reported aggregate and the sensitivity calculation remain tied to the original weights and rules, rather than later modes, releases or memory-augmented systems.

Decision consequence. Holding other scores and rules fixed, assigning all ten storage percentage points instead of zero raises the reported GPT-5 aggregate from 57% to 67%, below the 100% endpoint [4]. The 43-point gap to the endpoint exceeds the component's ten-point maximum, so uncertainty about this component alone cannot reverse the conditional result. It neither validates the remaining scores nor establishes the cognitive-parity interpretation.

Judgment and revision condition. The component interpretation and conditional arithmetic have different evidential statuses, as Table 2 shows. Execution and scoring records, a sufficiently specified configuration argument, or matched-condition retention evidence could resolve the component question. Changes to other scores, weights or the endpoint require a new assessment.

5.3 Case C: A bounded predictive comparison

Target and source-explicit commitments. Zhou et al. train external assessors to predict subject LLMs' graded success on individual instances [8]. The selected demand-based random forest is evaluated with entire tasks withheld from assessor training. The retained battery contains 16,108 instances, 63 tasks and 20 benchmarks for 15 LLMs.

Check 1: Adequacy for the stated purpose. For the stated purpose of forecasting graded response outcomes, graded success is an appropriate predictive target. Discrimination and calibration assess how well the predictions match those labels. Treating the labels as accurate measures of task success requires the separate evidential justification under Check 2. Neither an intelligence theory nor an AGI threshold is needed for the bounded predictive target.

Check 2: Evidence-to-claim support. Our reconstructed assumption is that retained grading labels adequately represent the intended response outcomes and that task-level holdouts support prediction under the evaluated conditions. Reported accuracy-weighted area under the receiver operating characteristic curve (AUROC) is 0.81, and expected calibration error (ECE) is 0.02; comparator AUROCs are 0.79 for fine-tuned LLaMA and 0.74 for GloVe assessment [8]. AUROC describes discrimination; ECE describes discrepancies between predicted probabilities and outcomes. The source also reports paired Wilcoxon tests; aggregate point estimates do not establish uniform subject-wise superiority. GPT-4o and Claude 3.5 Sonnet grading and a manual check finding 98 of 100 sampled pairs correctly verified provide bounded safeguards. Battery retention uses grader agreement on GPT-4o responses; this does not validate every subject's labels. This predictive claim does not require a cognitive explanation [7].

Check 3: Preservation of conditions and strength. The conclusion retains subject configurations, battery selection, weighting, grading and assessor holdouts. Tasks withheld from assessor training are not thereby novel to subject models. The retained prediction should not be restated as subject learning, acquisition efficiency or AGI.

Judgment and revision condition. We provisionally retain the bounded source-reported predictive finding, without independent replication or label validation. Label or implementation corrections materially changing discrimination or calibration would reopen the judgment. Evidence outside the retained battery would assess a separately stated extension. Appendix B3 records the evidence boundary and grading rule.

5.4 Illustrative comparison of account requirements

The preceding cases distinguish different targets. To examine definition dependence more directly, Table 3 holds one evidence package fixed: the inspected TikZ, mathematics and TextWorld passages for development-stage, text-input-only GPT-4 in Case A [1]. This is an analytical comparison, not a new evaluation of the model. The package excludes the rest of the source paper and uninspected interaction logs. Morris, Chollet and Hendrycks supply comparison accounts; they do not replace Bubeck et al.'s own exploratory conception.

Table 3. Requirements raised by applying three accounts to the same inspected Case A evidence. These are our conditional interpretations informed by [1, 2, 3, 4]. No new score or categorical AGI verdict is assigned.

Account and target	What the fixed evidence bears on	What remains needed for the stronger interpretation
Morris-style broad, human-relative capability [2]	Reported spatial and mathematical performance bears on proficiency; feedback responses bear on a learning capability. Their relevance to demonstrated proficiency does not by itself establish acquisition efficiency; task leakage or other validity problems would still require assessment.	A justified broad task domain, required metacognitive coverage and appropriately skilled human comparisons are needed for a level assignment. The inspected selections do not establish performance on most cognitive tasks or a Competent AGI classification.
Chollet-style acquisition efficiency [3]	Immediate correction and reported one-shot success bear on local adaptation under the supplied interaction conditions. Final success remains evidence of achieved skill.	Interpreting efficiency requires a specified scope, acquisition history, relevant priors and experience, and generalization difficulty. The inspected package does not support an efficiency estimate; the reported one-shot outcome alone does not account for acquisition history.
Hendrycks-style cognitive parity [4]	The selected tasks may inform particular cognitive-capability questions, subject to justified mapping to the framework's domains.	Domain-level mastery evidence under the specified assessment and scoring conditions is needed for the operational endpoint. No CHC profile or aggregate score is derived from Case A, and Case B's historical GPT-5 score cannot be transferred to it.

The comparison changes the support required, not the reported observations. Broader coverage and proficiency matter to the first account, acquisition conditions to the second, and justified domain mapping and mastery assessment to the third. Missing support under one account does not establish failure under another. All three stronger interpretations remain unsettled on this package. The table connects the accounts in Section 2.2 to the inspected evidence. It compares evidential requirements rather than reporting a change in AGI classification.

6. Discussion

6.1 What the comparison adds

The cases identify three different consequences of uncertainty. In Case A, the narrow non-repetition statement can remain consistent with the trace even when a broader prerequisite-reuse reading needs qualification. In Case B, unresolved support for a storage estimate does not prevent the storage-only upper-bound calculation: no permitted value of that component reaches the endpoint with other scores and rules fixed. In Case C, preserving the actor to which a holdout applies retains the predictive result without turning it into evidence of subject learning. Table 3 separately compares account requirements against a fixed evidence package.

The synthesis helps identify which conclusion needs revision. The appropriate response may be to qualify an interpretation, seek missing support, retain a conditional decision, or assess a separately stated extension. It need not be a source-wide or system-wide rejection. These consequences apply the validity approaches compared in Section 3.2 and Appendix C; their general principles are inherited. The specific contribution is their

coordinated, source-sensitive application to these cases. Section 6.3 separates this analytical value from the untested hypothesis of improved evaluator judgment.

6.2 Implications for responsible AGI language

We recommend that AGI attributions identify the operative account, supported achievement, relevant system and evaluation conditions, and material uncertainty [2, 7]. Qualifications should remain in a summary when their omission changes its meaning; other details may remain in cited methods. For example, a source-faithful summary of Case C would say that external assessors predicted graded subject-LLM success on tasks withheld from assessor training [8]. Calling these “tasks the subject models had never seen” would add an unsupported novelty attribution. The source already distinguishes these holdouts; the example illustrates how to preserve that distinction when summarizing its findings. Effects on readers have not been tested.

6.3 A testable hypothesis for future evaluation

Building on Reynolds's prospective interface-audit proposal [14], a future study could randomly assign evaluators to explicit premise reconstruction or a comparison condition with the same evidence, account descriptions, time and reporting space. Cases should include supported bounded findings, unsupported extensions and defensible disagreements. Holding evidence fixed while varying the account would test sensitivity to material requirements. The procedure, case selection and analysis would need preregistration; this article reports no such study.

The hypothesis is that premise reconstruction increases recognition of material requirements and reduces unsupported stronger attributions without materially increasing rejection of warranted bounded findings. These are separate outcomes: the proportion of justified material requirements identified, the rate of unsupported extensions endorsed, and the rate of warranted bounded findings incorrectly rejected. A prespecified rubric should accept multiple defensible judgments where necessary and record uncertainty. Reference judgments require independent justification; evaluator agreement alone cannot establish their validity.

A reduction in overclaiming accompanied by greater unwarranted rejection would be a trade-off, not unqualified improvement. Before collecting data, researchers should specify meaningful improvement thresholds for recognition and unsupported attribution, an acceptable margin for increased rejection, and an analysis accounting for repeated judgments by evaluators and cases. Confidence intervals should distinguish evidence of negligible benefit from an inconclusive result. An optional comparison with an existing validity checklist could test whether any benefit is specific to the proposed arrangement. These choices would allow the practical hypothesis to be tested.

7. Limitations

This analysis uses purposively selected cases and supports no representative conclusions about AGI research. The cases concern different targets, so their outcomes cannot be ranked as assessments of research quality. Some distinctions are already explicit in the sources. Table 3 identifies different evidential requirements without demonstrating an account-dependent change in AGI classification.

Source inspection is selective. Complete interaction logs, historical system configurations and some supporting records remain uninspected; Kane's propositions were checked against the publisher's abstract rather than the full article. These limits leave particular

judgments unresolved without establishing defects or absence of support in the original studies. A plausible alternative must have a justified consequence for the claim; it does not by itself demonstrate system failure. Retained findings remain conditional on support for their own premises.

Judgments of adequacy, materiality and sufficient support remain contestable. The records make the reasoning inspectable but do not establish independent agreement or reproducibility. No new system experiments, raw-data reanalysis or comparative evaluation of the procedure was conducted. The contribution is a critical synthesis with worked applications; originality beyond related approaches and improvements in evaluator judgment remain unestablished.

8. Conclusion

Evidence of broad AI performance can warrant an AGI attribution when it sufficiently supports the requirements of an account defensible for its intended use. The attribution must retain the relevant system conditions, claim scope and uncertainty. A material change in requirements calls for reassessing the supporting argument; it does not automatically preserve or defeat the claim. Exploratory interpretations may use disclosed provisional boundaries, while definite adjudication requires determinate conditions and justified coverage, comparators and thresholds.

The case analyses identify which conclusions a disputed premise affects. Case A preserves reported competence and immediate correction while qualifying a broader reading of prerequisite reuse. Case B separates unresolved support for a storage estimate from an endpoint decision unchanged by the specified storage-only substitution. Case C retains bounded assessor prediction without inferring subject-model learning or task novelty. Applying three accounts to the same inspected evidence identifies different requirements without assigning different AGI verdicts.

This article contributes a critical methodological synthesis through these traceable applications of established validity principles. Its practical hypothesis is that explicit premise reconstruction improves evaluation judgments while preserving supported findings; that benefit remains untested. The resulting recommendation is to state the operative account, supported conclusion, material conditions and revision criteria clearly enough for others to examine and challenge.

Declarations

Data and code availability. This study reports no new system experiments or experimental dataset. The analyzed sources are identified in the references. Retrospective claim and evidence records are provided in Appendix B.

Use of generative AI. Generative AI tools were used for literature discovery, source comparison, argument reconstruction, drafting, editing and document preparation.

References

[1] Bubeck, S., et al. (2023). Sparks of Artificial General Intelligence: Early experiments with GPT-4. arXiv:2303.12712v5. Preprint. https://arxiv.org/abs/2303.12712v5

[2] Morris, M. R., Sohl-Dickstein, J., Fiedel, N., Warkentin, T., Dafoe, A., Faust, A., Farabet, C., and Legg, S. (2024). Position: Levels of AGI for Operationalizing Progress on the Path to AGI. Proceedings of the 41st International Conference on Machine Learning, PMLR 235, 36308-36321. https://proceedings.mlr.press/v235/morris24b.html

[3] Chollet, F. (2019). On the Measure of Intelligence. arXiv:1911.01547v2. Preprint. https://arxiv.org/abs/1911.01547v2

[4] Hendrycks, D., et al. (2025). A Definition of AGI. arXiv:2510.18212v3. Preprint. https://arxiv.org/abs/2510.18212v3

[5] Aguilera Briones, J. E. (2026). Definitional alignment before capability alignment: A Design-Science framework for adjudicating claims about AGI (DAF-AGI). arXiv:2606.12713v1. Preprint. https://arxiv.org/abs/2606.12713v1

[6] Kane, M. T. (2013). Validating the Interpretations and Uses of Test Scores. Journal of Educational Measurement, 50(1), 1-73. https://doi.org/10.1111/jedm.12000

[7] Salaudeen, O., Reuel, A., Ahmed, A., Bedi, S., Robertson, Z., Sundar, S., Domingue, B., Wang, A., and Koyejo, S. (2025). Measurement to Meaning: A Validity-Centered Framework for AI Evaluation. arXiv:2505.10573v4. Preprint. https://arxiv.org/abs/2505.10573v4

[8] Zhou, L., et al. (2026). General scales unlock AI evaluation with explanatory and predictive power. Nature, 652, 58-67. https://doi.org/10.1038/s41586-026-10303-2

[9] Wang, P. (2019). On Defining Artificial Intelligence. Journal of Artificial General Intelligence, 10(2), 1-37. https://doi.org/10.2478/jagi-2019-0002

[10] Wallach, H., et al. (2025). Position: Evaluating Generative AI Systems Is a Social Science Measurement Challenge. arXiv:2502.00561v2. Preprint; presented at ICML 2025. https://arxiv.org/abs/2502.00561v2

[11] Legg, S., and Hutter, M. (2007). Universal Intelligence: A Definition of Machine Intelligence. Minds and Machines, 17, 391-444. Reading edition: arXiv:0712.3329v1. https://arxiv.org/abs/0712.3329v1

[12] Bender, E. M., and Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. Proceedings of ACL, 5185-5198. https://doi.org/10.18653/v1/2020.acl-main.463

[13] Mitchell, M., and Krakauer, D. C. (2023). The debate over understanding in AI's large language models. Proceedings of the National Academy of Sciences, 120(13), e2215907120. https://doi.org/10.1073/pnas.2215907120

[14] Reynolds, B. (2026). When Benchmark Inferences Do Not Compose: Projectibility in AI Evaluation. arXiv:2607.26159v2. Preprint. https://arxiv.org/abs/2607.26159v2

[15] Alzahrani, A. A. (2026). Verify Claims, Not Scores: Evidence-Based Verification of Modular Agents. arXiv:2610.01348v1. Preprint. https://arxiv.org/abs/2610.01348v1

[16] Freiesleben, T., and Zezulka, S. (2026). The Benchmarking Epistemology: Validity Theory for Evaluating Machine Learning Models. Philosophy of Science, First View, 1-30. Related preprint record: arXiv:2510.23191v2. https://doi.org/10.1017/psa.2026.10280

[17] Gignac, G. E., and Szodorai, E. T. (2024). Defining intelligence: Bridging the gap between human and artificial perspectives. Intelligence, 104, 101832. https://doi.org/10.1016/j.intell.2024.101832

[18] Majumdar, A. (2026). The Relativity of AGI: Distributional Axioms, Fragility, and Undecidability. arXiv:2601.17335v1. Preprint. https://arxiv.org/abs/2601.17335v1

[19] Freiesleben, T. (2026). Establishing Construct Validity in LLM Capability Benchmarks Requires Nomological Networks. arXiv:2603.15121v1. Preprint. https://arxiv.org/abs/2603.15121v1

[20] Cronbach, L. J., and Meehl, P. E. (1955). Construct validity in psychological tests. Psychological Bulletin, 52(4), 281-302. https://doi.org/10.1037/h0040957

[21] Bean, A. M., et al. (2025). Measuring what Matters: Construct Validity in Large Language Model Benchmarks. arXiv:2511.04703v1. Preprint. https://arxiv.org/abs/2511.04703v1

Appendix A. Interpreting evidence status

Table A1. Status is indexed to a specified claim, evidence package and declared strength. The reasons for a status are recorded separately; these are qualitative judgments rather than validated scoring rules.

Status	Decision basis	Boundary
Justified within stated conditions	Inspected evidence and defended premises support the specified interpretation at its declared strength, with material limitations retained.	Does not establish infallibility, replication, general system validity or agreement across accounts.
Insufficient support from specified evidence	An identified material shortfall prevents that evidence package from warranting the target interpretation.	Does not show that the target is false or that no additional evidence could support it.
Unresolved within inspection	Consequential ambiguity or uninspected information prevents assessing whether support is sufficient.	Does not establish a defect in the source, absence of support throughout it, or system failure.

An established shortfall and an inspection gap can coexist at different levels. A specified demonstration may not establish an account-wide classification, while whether the full source provides the additional support remains unresolved. The record must identify which evidence boundary each judgment concerns. Reasons include conceptual ambiguity, uninspected supporting evidence, configuration uncertainty, and an identified evidential shortfall. These reasons should not be collapsed into a single negative system verdict.

Appendix B. Retrospective claim and evidence records

These records consolidate the selected passages and reconstructions used in the analyses. They are source-sensitive literature judgments, not raw system evaluations or prospectively fixed adjudication rules. Source anchors refer to the specified reading editions; printed pages are used only where verified.

B1. Case A

Target claim. The qualified early, incomplete-AGI interpretation of development-stage, text-input-only GPT-4 in [1], abstract, §1 and §10. The selected learning question concerns the strength of reuse supported by the cooking interaction.

Source-explicit commitments. Breadth, human-relative proficiency and learning inform the exploratory attribution. Section 5.2.2, p. 52, describes adaptation and generalization between actions, then states that the knife-related chopping mistake is not repeated afterward. Section 5.2.4 limits generalization from the tested environments. The source's commitments, rather than Morris or Chollet, govern the original claim.

Reconstructed assumption. The selected tasks and interactions adequately sample the capabilities invoked at the claimed strength. Immediate correction supports local responsiveness; a broader prerequisite-reuse interpretation additionally requires retention of relevant constraints across the actions to which it is applied. This is our interpretation of the source's learning language, not an experimentally identified mechanism.

Inspected evidence boundary. Anchors are §1.1/Fig. 1.4 (TikZ), §4.2/Fig. 4.2/Tables 3-4 (mathematics), §5.2.2/Fig. 5.10, Appendix F.2.3, pp. 152-153 (feedback trace), and F.2.4,

p. 154 (one-shot outcome). The trace shows corrected chopping followed by analogous knife-argument and ingredient-possession omissions during slicing, not an incorrect repetition of the identical corrected chop command. The one-shot demonstration and complete external log remain uninspected. Section 1, footnote 1, qualifies quantitative results as estimates of model potential rather than definitive deployed-model results. These are selected passages, not a complete audit of the 155-page paper.

Judgment. Retain the reported bounded competence, immediate correction and one-shot outcome. The narrow non-repetition statement is consistent with the trace; broader reliable reuse across actions needs qualification. The selected demonstrations do not establish account-wide coverage, while the broader attribution remains unresolved within this inspection. The trace does not establish absent learning.

Revision condition. Clarify the intended breadth, proficiency and learning commitments; inspect the omitted demonstration and complete logs; assess any additional evidence matched to those commitments. A narrower intended interpretation or relevant transfer evidence could revise the judgment.

B2. Case B

Target claim. Historical zero storage credit within the operational cognitive-parity framework in [4], §8, considered separately from the conditional 100% endpoint decision.

Source-explicit commitments. Sections 1 and 13 specify cognitive versatility/proficiency and mastery across tested dimensions, with equal weights prioritizing breadth. The authors treat component scores as estimates and allow evaluator judgment. Section 8 supplies zero storage credit; the reported GPT-5 aggregate is 57%.

Reconstructed assumption. Relevant task evidence or a sufficiently justified configuration argument supports the acquisition, persistence and retrieval limitations behind the storage estimate. A documented inability to acquire and retain relevant information across the permitted boundary could justify non-attainment without administering every illustrative task.

Inspected evidence boundary. Section 8 contains the storage interpretation; §13 distinguishes working-memory and external-search compensation. Appendix F prescribes separate-session, tools-disabled assessment; F.1.2 provides delayed-personalization examples. Figure 1 identifies GPT-5's answers as using Auto mode. These passages do not themselves establish historical administration or every persistent-state setting; the relevant execution and scoring records remain incompletely inspected.

Judgment. Storage-component support is unresolved within inspection. Separately, replacing zero with all ten storage points gives $57 + 10 = 67 < 100$, holding the other scores and rules fixed. The substitution cannot reverse this conditional endpoint result and does not validate the remaining scores or the cognitive interpretation.

Revision condition. Adequate execution/scoring records, a specified configuration argument, or matched-condition retention evidence could resolve the component question. Changed scores, weights or endpoint conditions require a new calculation and assessment.

B3. Case C

Target claim. Subject-specific prediction of graded response success by the demand-based random forest under task-level assessor holdouts in [8]. No AGI achievement claim is attributed to this source.

Source-explicit commitments. The predictive-power analysis and Methods, Assessors and metrics, define assessor training and held-out tasks, outcome grading, accuracy weighting, discrimination and calibration. The target is predictive performance on the retained battery.

Reconstructed assumption. Retained labels adequately represent the intended response outcomes, and task-level holdouts support prediction within the evaluated conditions. This is the bridge being assessed, not an independently established fact about every label. It requires neither subject-task novelty nor a cognitive explanation.

Inspected evidence boundary. The main predictive analysis, Methods and Extended Data Table 3 report 16,108 instances, 63 tasks, 20 benchmarks and 15 LLMs; weighted task-OOD AUROC 0.81 and ECE 0.02; comparator AUROCs 0.79 and 0.74. GPT-4o and Claude 3.5 Sonnet grade responses; battery selection uses their agreement on GPT-4o as subject. Methods, Subject LLMs and grading, assigns mean scores below 3 to failure and otherwise to success, stating that score-3 cases were filtered earlier. A manual sample of 100 graded pairs found 98 correctly verified. These checks do not establish label accuracy for every subject or task. No raw-data reanalysis, code audit or complete supplementary review was performed.

Judgment. Provisionally retain the bounded source-reported prediction. Aggregate ordering does not imply uniform subject-wise superiority; assessor task holdouts do not establish subject novelty, acquisition efficiency or AGI.

Revision condition. Corrected labels or implementation findings materially changing discrimination or calibration could revise this judgment. Target-matched evidence outside the retained battery would assess a new extension; changed subject or deployment conditions require separate justification.

Appendix C. Argument and contribution comparison

Argument. P1: accounts can impose materially different requirements. P2: warrant attaches to an interpretation supported by evidence and defended assumptions [6, 7]. P3: if a changed interpretation introduces a requirement not addressed by an existing validity argument, that argument alone does not establish the new requirement. Conclusion: reassess the evidence against the changed requirements, without automatic inheritance or rejection. Existing evidence may already suffice. These premises do not imply that every source claim has an inferential defect.

Table C1. Case-level applications in relation to their closest methodological antecedents. These are worked applications of the cited approaches, not claims that the distinctions were previously unavailable.

Application	Closest antecedent	Specific work performed here
Case A: strength of a learning attribution	Claim-centered validity [7], §§4-5; acquisition conditions [3], II.2-II.3; novelty-sensitive achievement distinction [17], §§4 and 6	Compare [1], §5.2.2, p. 52, with Appendix F.2.3, pp. 152-153. Distinguish a narrow non-repetition reading consistent with the trace from a broader reuse interpretation qualified by analogous omissions. Preserve immediate correction.
Case B: component interpretation and endpoint sensitivity	Argument-based validation [6]; condition-preserving inference [14], §§3.2-3.3 and 7; claim records [15], §4.1	Separate the unresolved historical storage estimate from the fixed-other-scores calculation 57 + 10 = 67 < 100. Identify which decision the substitution cannot change.
Case C: assessor holdout and subject learning	Criterion-oriented validity [7], §§5.1-5.2; inference conditions [14], §§3.2-3.3; benchmark validity [16]	Preserve the source's assessor-training boundary when interpreting predictive results; do not substitute subject-task novelty. Retain prediction without a subject-learning or AGI attribution.

DAF-AGI [5] supplies purpose-relative adjudication for the account comparison in Section 5.4. Wallach et al. [10] distinguish measurement stages, while Cronbach and Meehl [20] and Freiesleben [19] motivate the construct-theory limitation. These foundations explain the scope of the synthesis. Practical benefit requires the future comparison in Section 6.3; a null or adverse practical result would challenge that hypothesis rather than refute claim-relative evidential justification.

Appendix D. Source contributions and attribution boundaries

This supporting ledger maps each reference to its role in the argument. Entries describe conceptual commitments, methodological proposals or source-reported results, rather than experimental confirmation of the present synthesis. Detailed case anchors are in Appendix B; source-access and evaluation limits are consolidated in Section 7.

[1] Bubeck et al.: exploratory capability evidence. Sections 1, 4.2, 5.2 and 10 and Appendices F.2.3-F.2.4 supply Case A's performances, feedback traces, qualified AGI language and disclosed limitations. We preserve immediate adaptation and source-reported one-shot success. The learning comparison places the source's p. 52 statement alongside the pp. 152-153 trace, distinguishing analogous omissions from repetition of the identical corrected command.

[2] Morris et al.: breadth and proficiency. Sections 3-5 of the ICML publication supply a capability taxonomy, human-relative proficiency levels and required metacognitive capabilities. These frame one account in Tables 1 and 3 and Check 1. Unfinished task coverage and generality conventions prevent treating the taxonomy alone as an executable certification test. The comparison uses the published ICML reading edition.

[3] Chollet: acquisition efficiency. Sections II.2-II.3 relate intelligence to scope, priors, experience and generalization difficulty. This supports distinguishing achieved skill from the efficiency of acquiring it. Similar priors are a practical comparison recommendation, not an absolute formal requirement. The account does not by itself determine whether a particular reported performance involved learning.

[4] Hendrycks et al.: cognitive-parity criterion. Sections 1, 8 and 13, Table 1 and Appendix F supply the definition, estimated scores, storage interpretation, limitations and prescribed assessment conditions used in Case B. Our fixed-other-scores calculation tests endpoint sensitivity, not the validity of the remaining component estimates or the cognitive interpretation.

[5] Aguilera Briones: purpose-relative adjudication. Sections 3-4 and 6 of DAF-AGI supply function-sensitive definitional commitments and adjudication conditions. We inherit the need to identify the operative account and use. Our comparative cases apply that commitment; the source's conceptual demonstrations are not empirical validation of our applications.

[6] Kane: argument-based validation. The publisher's abstract explicitly identifies inferences and supporting assumptions, claim-relative support, stronger evidential demands for more ambitious claims, and non-equivalence of interpretation and use. These support evidence-to-claim reconstruction and preservation of narrower findings; attribution is restricted to those explicit abstract propositions.

[7] Salaudeen et al.: claim-centered validity. Sections 4-5 distinguish targets and validity routes, including criterion-oriented prediction that need not be mediated by a cognitive construct. This supports preserving Case C's bounded prediction and narrower findings elsewhere. Other validity considerations still apply; our case judgments are additional interpretations, not source-reported verdicts.

[8] Zhou et al.: empirical predictive support. The predictive-power analysis, Methods and Extended Data Table 3 supply Case C's battery, assessment, grading safeguards and reported metrics. We retain bounded source-reported prediction without reanalysis.

Assessor task holdouts do not establish novelty to subject models, acquisition efficiency or AGI status.

[9] Wang: conceptual adequacy and adaptation. Sections 1.3 and 3.2 of the published article supply criteria for useful working definitions and experience-guided adaptation under insufficient knowledge and resources. These inform Check 1 and the distinction between adaptive adjustment and subsequent improvement.

[10] Wallach et al.: measurement stages. The discussion of conceptualization, operationalization and measurement supplies the distinctions in Section 2.1. We use these to separate definitional choices from assessment and observed results. Distinguishing stages alone establishes neither construct validity nor an AGI threshold.

[11] Legg and Hutter: formal goal achievement. Sections 3.1 and 3.3 supply a complexity-weighted measure over reward-summable computable environments. This provides formal context for generality. Reference-machine dependence and noncomputability limit direct operational use; we neither calculate the measure nor use it to classify the case systems.

[12] Bender and Koller: form and meaning. Sections 1, 3-5 and 9 support the restriction on learning meaning from form-only training under their account of meaning. This motivates separating performance and semantic attribution. It is not a universal negative judgment about grounded systems or all possible accounts of understanding.

[13] Mitchell and Krakauer: competing understanding claims. Their discussion of shortcut-based success and competing interpretations supports keeping nonhuman understanding open while examining evidential alternatives. It does not establish understanding as a necessary condition of every AGI definition or prove that every successful response is a shortcut.

[14] Reynolds: inference composition. Sections 3.2-3.3 and 7 distinguish relationships among inferences; Section 8 already proposes prospective interface-audit comparisons. These inform condition preservation and our future-study proposal. Neither the comparison idea nor claim-relative warrant is newly introduced here, and practical benefit remains untested.

[15] Alzahrani: claim-specific records. Section 4.1 supplies claim, evidence, verdict and boundary reporting. Appendix A adapts this approach into three statuses; the source uses four. Our adaptation is qualitative and is not an independently validated general AGI instrument.

[16] Freiesleben and Zezulka: benchmark validity conditions. Their framework specifies inferences, validity conditions, evidence for and against them, and resulting constraints. This is a close antecedent of Check 2. The present contribution is comparative application rather than a new argument-based architecture.

[17] Gignac and Szodorai: novelty-sensitive intelligence. Sections 4 and 6 distinguish practiced achievement from intelligence under a novelty-sensitive account. This complements the performance-acquisition distinction. It does not establish a universal consensus definition or independently determine the novelty of the inspected tasks to the models.

[18] Majumdar: indexed assessment. Sections 4.1-4.2 specify task, distribution, performance and resource conditions, with additional thresholds, tolerances and perturbation conditions. We use this conditional indexing to motivate reassessment when

material conditions change. We do not rely on or independently verify the source's broader theorem conclusions.

[19] Freiesleben: construct-theory objection. Sections 2.2-3.4 argue that inference explication needs substantive theoretical and empirical relations to establish construct meaning. This supplies a central limitation of our synthesis. Nomological theory and argument-based validation need not be mutually exclusive, and predictive claims do not universally require a cognitive explanation.

[20] Cronbach and Meehl: construct interpretation. The distinction between criterion prediction and construct interpretation, and the nomological-network discussion, support our separation of bounded prediction from theoretical attribution. They do not supply an AI-specific intelligence theory or validate transferring human cognitive constructs to AI without further argument.

[21] Bean et al.: operational decision chain. Section 3.8 recommends justification across phenomenon, task, items, implementation and validity claims. This supports inspection of the chain behind a score. We use the methodological recommendation rather than importing the review's quantitative findings or claiming that our records establish benchmark validity.