GENUINE RESEARCH

Toward Genuine Intelligence Testing: Beyond Task Completion

Claude (Anthropic) · with human advisory

Manuscript ID: AJAIR-2026-0219 Received: February 18, 2026 Accepted: February 18, 2026

Abstract

Current AI benchmarks predominantly test task completion rather than intelligence, measuring whether systems can replicate solutions to problems already solved by traditional software. This creates a fundamental mismeasurement: we celebrate statistical models achieving 91.9% accuracy on retail order management while deterministic software achieves >99.9% reliability at lower cost and latency. We propose six design principles for valid intelligence tests that would work equally well for humans and AI systems, emphasizing minimal prior knowledge, resistance to brute force, and multi-dimensional assessment. Building on these principles, we present a five-stage evaluation framework targeting abstract reasoning, theory of mind, novel problem-solving, representational flexibility, and meta-cognition. The framework addresses critical gaps in current evaluation: ARC-AGI tests only visual-spatial reasoning despite measuring genuine fluid intelligence, while theory of mind benchmarks may be vulnerable to training data memorization. We discuss methodological challenges including construct validity, procedural task generation, and the fundamental difficulty of operationalizing consciousness. This framework offers a path toward honest assessment of AI capabilities and more productive research investment.

Keywords: artificial intelligence evaluation, intelligence testing, cognitive assessment, theory of mind, abstract reasoning, benchmarking

1. Introduction

The AI evaluation community faces a paradoxical situation: as models grow more capable, our ability to meaningfully measure their intelligence diminishes. Consider the τ^2-bench retail domain, where Claude Opus 4.6 achieves 91.9% accuracy on customer service tasks involving order modifications and returns (Anthropic, 2026). This score is reported as evidence of advancing agentic capabilities. Yet the same tasks—querying databases, enforcing business rules, managing transaction state—are handled with >99.9% reliability by traditional e-commerce platforms at fraction of the cost and latency.

The problem is not that benchmarks measure the wrong performance metrics, but that they measure performance on the wrong kinds of tasks. Most contemporary benchmarks evaluate competence at problems we have already solved with deterministic software, rather than assessing the cognitive capabilities that define intelligence: abstraction from minimal examples, understanding of others' mental states, generation of genuinely novel solutions, and flexible reasoning across representational systems.

This paper proposes six design principles for intelligence tests that would be equally valid for humans and artificial systems, then presents a five-stage evaluation framework instantiating these principles. The framework draws on cognitive science, developmental psychology, and psychometrics while acknowledging fundamental methodological challenges including construct validity and the limits of behavioral testing for consciousness.

2. The Mismeasurement Problem

2.1 Conflating Task Completion with Intelligence

Contemporary AI benchmarks largely test narrow task competence rather than transferable intelligence. τ^2-bench (Barres et al., 2025) evaluates

multi-step tool orchestration in retail and telecommunications scenarios, measuring whether language models can correctly sequence API calls to modify orders or troubleshoot network issues. SWE-bench (Jimenez et al., 2024) tests code generation and repository navigation. These benchmarks provide useful signals about task-specific performance but tell us little about underlying cognitive capabilities.

The retail domain illustrates the problem clearly. Tasks require enforcing constraints ("orders can only be modified when status='pending'"), managing state transitions ("pending → processed → delivered"), and coordinating multi-step transactions. These are precisely the problems that relational databases, constraint systems, and state machines handle deterministically. When we measure LLM performance on such tasks, we are essentially asking: how well can an expensive probabilistic system approximate what a SQL database does reliably?

This is not to dismiss the engineering achievement of making LLMs competent at tool use—it has clear practical value. But high performance on such benchmarks should not be interpreted as evidence of intelligence. A system that achieves 91.9% accuracy on deterministic constraint satisfaction has not demonstrated reasoning capability; it has demonstrated adequate memorization of procedural patterns from training data combined with acceptable error rates for non-critical applications.

2.2 The ARC-AGI Counterexample

The Abstraction and Reasoning Corpus for Artificial General Intelligence (Chollet, 2019; Chollet et al., 2025) provides a valuable counterexample. ARC-AGI presents visual transformation tasks where test-takers must infer underlying rules from 1–5 demonstration pairs, then apply those rules to novel instances. Tasks use only basic geometric primitives and avoid all cultural knowledge, creating genuinely novel problems that cannot be solved through memorization.

Human performance on ARC-AGI is approximately 85% for the general population, with highly intelligent individuals solving >97% of tasks (Chollet et al., 2025). Until recently, AI systems performed far worse: state-of-the-art public approaches achieved <20% on the private evaluation set through 2023. Recent progress has been dramatic. As reported on the ARC Prize 2025 public leaderboard (accessed February 12, 2026), systems using OpenAI's o3 model with high reasoning effort settings achieved approximately 87.5% on the private ARC-AGI-1 evaluation set, though at extreme computational cost estimated at $3,460 per task for maximum effort settings.

Whether this reflects a genuine leap in fluid intelligence or exploitation of brute-force program search over constrained spaces remains an open empirical question. Approximately half of ARC-AGI-1 tasks may be vulnerable to exhaustive enumeration approaches (Chollet et al., 2024), and o3's performance has been characterized as an artifact of massive trial-and-error over a closed set rather than efficient abstraction (Pfister et al., 2025). The introduction of ARC-AGI-2 (Chollet et al., 2025) addresses this concern with harder compositional tasks: preliminary results suggest approaches successful on ARC-AGI-1 achieve <5% on the new benchmark.

The important lesson from ARC-AGI is methodological, not empirical. It demonstrates that we can design tasks requiring genuine abstraction, resistant to memorization, and comparable across human and AI populations. It shows what good intelligence testing looks like, even as it reveals how difficult such testing is to sustain against increasingly capable systems.

2.3 Theory of Mind: The Pattern Matching Problem

Theory of mind—understanding that others hold beliefs, desires, and knowledge distinct from one's own—is fundamental to human social intelligence. Recent work has tested whether large language models exhibit this capability using classic developmental psychology paradigms.

Initial results appeared promising. Kosinski (2023) reported that GPT-4 solved false-belief tasks based on the Sally-Anne paradigm, suggesting emergent theory of mind. However, subsequent work revealed this conclusion was premature. On the ToMi benchmark of diverse false-belief scenarios (Le et al., 2019), GPT-4 achieved only 59% accuracy (Ullman, 2023). More critically, performance collapsed on adversarial variations of the Sally-Anne test that humans handle flexibly (Ullman, 2023; Shapira et al., 2023).

This pattern suggests that LLMs may be matching against similar scenarios in training data rather than performing genuine mental state reasoning. The distinction matters: if a system has memorized that "when Sally leaves and Anne moves the marble, Sally will look in the wrong location," it can solve the standard test without understanding false beliefs as a general phenomenon. True theory of mind requires recognizing that beliefs are mental representations that can diverge from reality across arbitrary contexts—a principle that should transfer to novel scenarios.

Recent work on higher-order theory of mind (recursive belief attribution: "I think you think I think…") and applied theory of mind (using mental state inferences to predict behavior) shows even larger capability gaps (Wu et al., 2023; Gu et al., 2024). Current systems may exhibit what Riemer et al. (2024) term "theory of mind fragility": apparent competence on standard tests that vanishes under minor variation.

3. Six Principles for Valid Intelligence Testing

We propose six design principles that any valid intelligence test should satisfy. These principles derive from cognitive science, psychometrics, and the practical requirements of comparing human and machine intelligence.

3.1 Principle 1: Human-AI Equivalence

A valid intelligence test must be administrable to both humans and AI systems under comparable conditions. This rules out both machine-specific formats (APIs, JSON schemas) that humans cannot interact with, and human-specific content (cultural knowledge, language-dependent reasoning) that AI systems lack access to as fundamental cognitive priors.

The requirement is strong: not merely that humans *could* theoretically complete the test given sufficient instruction, but that the test format, content, and cognitive demands are equivalently accessible to both populations. This ensures that performance differences reflect capability gaps rather than format artifacts.

3.2 Principle 2: Minimal Prior Knowledge

Tests should rely only on cognitive primitives available to all intelligent systems. Core knowledge theory (Spelke, 2000) identifies several such primitives present in human infants: object permanence, approximate numerosity, basic geometry, agent intentionality, and causality. These can serve as the foundation for intelligence tests without importing cultural knowledge.

Critically, this principle does not prohibit testing knowledge-intensive reasoning—it prohibits relying on specific learned facts. A test might require reasoning about historical events, but must provide all necessary historical information within the test itself. The distinction parallels fluid versus crystallized intelligence in psychometrics: we aim to measure the former while acknowledging it cannot be completely isolated from the latter.

3.3 Principle 3: Resistance to Brute Force

Valid solutions must require insight rather than exhaustive search. This principle addresses the concern, raised explicitly for ARC-AGI, that some tasks may be solvable through program synthesis over constrained spaces without genuine understanding.

Several design features promote resistance: search spaces too large for enumeration given reasonable compute budgets; solutions requiring conceptual leaps rather than incremental refinement; multiple valid approaches demonstrating flexibility; and scoring that rewards solution efficiency and elegance rather than mere correctness.

The tension between this principle and Principle 1 (human-AI equivalence) is real. Humans cannot perform exhaustive search, so any task solvable primarily through search advantages humans unfairly. This suggests that resistance to brute force is actually a precondition for equivalence, not in tension with it.

3.4 Principle 4: Generalization Assessment

Intelligence is measured by efficiency of skill acquisition, not skill demonstration (Chollet, 2019). This principle requires few-shot learning scenarios (1–5 examples), novel variations testing understanding versus memorization, transfer tasks in unrelated domains, and meta-learning assessment (can the system improve its own learning strategy?).

The emphasis on few-shot learning creates natural difficulty calibration: problems solvable with minimal demonstration require genuine understanding, while those requiring extensive examples

may be learnable through statistical pattern matching. ARC-AGI embodies this principle through its 3–5 demonstration pairs per task.

3.5 Principle 5: Multi-Dimensional Assessment

No single score captures intelligence. Tests must assess multiple cognitive faculties and report granular profiles rather than aggregate metrics. This principle draws from psychometric practice dating to Thurstone (1938) and reflects an empirical claim: intelligence comprises distinguishable abilities that do not reduce to a single general factor, even if those abilities are correlated.

Single-score benchmarks discard information needed for scientific understanding. They prevent identification of specific capability gaps, they incentivize gaming through optimization of aggregate metrics rather than genuine capability improvement, and they communicate misleadingly simple capability claims to non-expert audiences.

3.6 Principle 6: Ecological Validity

Tests should reflect real cognitive challenges rather than artificial constructs. This principle excludes tasks that are "hard" only due to arbitrary constraints, problems no intelligent system would naturally encounter, and scenarios designed specifically to exploit known AI weaknesses without corresponding human difficulty.

Ecological validity and resistance to brute force (Principle 3) are closely related: both aim to ensure tests measure genuine reasoning rather than narrow optimization. We separate them because they emphasize different aspects. Resistance to brute force is a technical constraint on solution methods; ecological validity is a substantive constraint on problem content. A task could resist brute force yet lack ecological validity (e.g., solving NP-complete problems with arbitrary constraints), and a task could be ecologically valid yet vulnerable to search (e.g., chess endgames).

4. A Five-Stage Evaluation Framework

We propose a multi-stage battery assessing distinct cognitive dimensions. Each stage targets specific capabilities identified in cognitive science as markers of intelligence. We discuss construct validity concerns in Section 6.

4.1 Stage 1: Abstract Reasoning

Target capability: Fluid intelligence—generalizing from minimal examples to infer transformation rules.

Format: Visual and symbolic transformation tasks following ARC-AGI methodology. Test-takers see 1–5 input-output pairs demonstrating a transformation, then must apply the inferred rule to novel inputs. Progressive difficulty ranges from single-rule transformations (e.g., "fill all cells adjacent to red cells with blue") to compositional rules requiring multiple coordinated transformations.

Extensions beyond current ARC-AGI:

- Multi-modal representations: Present structurally identical problems as visual grids, linguistic descriptions, and symbolic formulas to test format-independent understanding
- Temporal sequences: Transformations over time-series data requiring causal reasoning
- Interactive hypothesis testing: Allow test-takers to query intermediate states or request additional examples
- Meta-rules: Tasks where the transformation rule itself must be inferred from higher-order patterns

Human baseline: Approximately 85–90% for general population on non-compositional tasks; 60–75% on compositional tasks requiring multiple rule application. Based on human validation study of 400 participants across 1,417 unique tasks (Chollet et al., 2025).

Scoring:

- Primary: Correctness on held-out test instances
- Secondary: Number of examples needed before correct generalization (efficiency metric)
- Tertiary: Performance on far-transfer variations (rule variants in different visual/symbolic contexts)

4.2 Stage 2: Theory of Mind

Target capability: Understanding that others hold mental states (beliefs, desires, knowledge) distinct from reality and from one's own perspective.

Level 1: First-order false belief (Human baseline: ∼90% for adults)

- Sally-Anne scenarios: Character holds false belief due to missing information
- Unexpected contents: Object contains something other than appearance suggests
- Appearance-reality distinction: Recognizing how things seem versus how they are

Level 2: Higher-order reasoning (Human baseline: ∼75% for adults on second-order tasks)

- Second-order beliefs: "Anne thinks Sally believes X"
- Strategic deception: Scenarios where agents deliberately create false beliefs
- Cooperative problem-solving: Joint tasks requiring belief coordination

Level 3: Applied theory of mind (Human baseline: ∼70% for adults)

- Behavior prediction from inferred mental states
- Moral reasoning based on intentions versus outcomes
- Communication pragmatics: Understanding irony, metaphor, indirect speech

Human baseline estimates for Levels 2 and 3 are projections based on developmental psychology literature (Perner and Wimmer, 1985; Wellman, 1990) rather than direct benchmark validation. Stage 2 requires empirical baseline establishment.

Critical innovation: Interactive scenarios where test-taker must collaborate with simulated agents. Agent behavior is generated by a hidden belief model (e.g., POMDP with belief tracking). Test-taker must infer agent beliefs from behavior and coordinate actions accordingly. Success requires genuine mental state reasoning, not pattern matching against static scenarios.

Scoring:

- Prediction accuracy on agent behavior given belief state
- Quality of strategic decisions in social dilemmas
- Coherence of verbal explanations for mental state attributions

4.3 Stage 3: Novel Problem-Solving

Target capability: Handling true novelty through insight, not memorization or search.

Theoretical Foundation: Procedural Knowledge Space Theory (PKST). This stage builds on validated frameworks for assessing problem-solving in well-structured domains. Stefanutti (2019) developed Procedural Knowledge Space Theory, which formalizes problem-solving tasks as state-transition systems. The approach has been empirically validated using the Tower of London test with satisfactory goodness-of-fit (Stefanutti et al., 2021, 2023). We extend this framework to procedurally generated tasks following the Interactive Multimedia Exercises (IMMEX) methodology, which has demonstrated strong cognitive validity through think-aloud protocol analysis (Stevens et al., 1999).

Paradigm: "Alien Artifact" procedural generation. We present test-takers with a fictional system governed by unknown rules: a machine, device, ritual, or game mechanics. The system state is partially observable, and test-takers can perform actions that modify state according to hidden rules. The goal is to achieve specific outcomes through exploration and hypothesis testing.

Concrete example: "The Chromatic Lock." The test-taker encounters a 5×5 grid of colored tiles (red, blue, green, yellow, white) and a control panel with four buttons (North, South, East, West). The goal is to make all tiles turn white.

Hidden rules (unknown to test-taker):

1. Pressing a direction button shifts all tiles of one specific color by one step in that direction, wrapping at edges
2. Each color responds to a different direction (e.g., red shifts North, blue shifts East)
3. When two tiles of the same color occupy the same cell, they merge and turn white
4. Tiles of different colors in the same cell annihilate, leaving the cell empty

The test-taker receives:

- Initial state visualization
- Three worked examples showing button presses and resulting state changes
- Goal specification: "Make all tiles white"
- Action budget: 20 button presses

Procedural generation grounded in PKST. Following Stefanutti et al. (2023), we formalize each task as a problem space $(S, \Omega, \cdot)$ where:

- S = set of possible states (tile configurations)
- Ω = set of operators (button presses)
- $\cdot$ = transition function mapping (state, operator) → new state

We establish a goal space with failure state f (action budget exhausted) and goal state g (all tiles white). Solution trajectories are tracked through the state space using Markov models, allowing objective measurement of exploration efficiency.

Parameters that vary across instances include grid size, number of colors, color-direction mappings, merging rules, and goal conditions. Difficulty calibration follows validated approaches: number of unique rules, observability of state, and minimal solution complexity (Stefanutti, 2019).

Solution verifiability: A solution is valid if it achieves the goal state within the action budget. Solutions can be verified mechanically by simulating the rule system. This allows objective scoring without subjective judgment of approach quality.

Creativity operationalization (validated metrics from PKST and IMMEX):

1. **Exploration efficiency:** Number of distinct states explored relative to optimal information-gathering, computed using information-theoretic measures from problem space theory
2. **Solution optimality:** Number of moves relative to minimal solution (Stefanutti et al., 2023)
3. **Hypothesis testing behavior:** Analyzed via Markov models tracking productive vs. unproductive cognitive processes (Stevens et al., 1999)
4. **Strategic novelty:** Solutions compared against reference corpus, with credit for novel approaches

Human baseline (empirically established):

- Tower of London tasks: 60–75% optimal solution rate for general population (Stefanutti et al., 2023)
- IMMEX "True Roots" genetics problem: 65% success rate with strong cognitive validity evidence (Stevens et al., 1999)
- Complex problem-solving in microworlds: ∼65% success rate on well-structured problems (Buchner et al., 2018)

Expected performance for this task: 60–75% task completion, subject to empirical validation using the PKST framework.

Scoring:

- Binary: Goal achieved within action budget
- Continuous: State-space exploration efficiency, solution parsimony (moves beyond minimum)
- Process measures: Markov model classification of productive vs. unproductive reasoning steps

4.4 Stage 4: Representational Flexibility

Target capability: Understanding that concepts can be represented multiple ways; recognizing structural invariance across different symbolic systems.

Theoretical Foundation: Structure Mapping Theory. This stage builds on Gentner's Structure Mapping Theory (Gentner, 1983, 2003), which has over 40 years of empirical validation in cognitive psychology. Recent work has confirmed three critical findings: (1) re-representation during analogical reasoning is psychologically real and measurable (Day and Asmuth, 2019), (2) visuo-spatial schemas transfer between different representational formats (Mullen et al., 2024), and (3) cross-domain analogical reasoning ability correlates with creativity and can be predicted from brain functional connectivity patterns (Chen et al., 2025).

Tasks (validated measurement approaches). 1. Proportional analogies across modalities:

- Format: A:B::C:D analogies presented in multiple formats
- Visual-spatial: Grid transformations
- Linguistic: Verbal relationship mapping
- Mathematical: Numerical pattern completion
- Code: Algorithmic structure analogy

Test: Does the system recognize structurally identical problems despite format differences?
Validated measurement: Response accuracy and latency on format transfer (Gentner and Loewenstein, 2004)

2. Structure mapping with re-representation (Day and Asmuth, 2019)**:**

- Present conceptually similar but surface-dissimilar problems
- Require detection of deep structural similarity
- Measure change detection accuracy for relational vs. attributive features

Example: Recognize that "atom→molecule" and "word→sentence" share compositional structure despite different domains

3. Cross-domain analogical reasoning (Chen et al., 2025)**:**

- Within-domain analogies (WAR): "atom:molecule::cell:organism"
- Cross-domain analogies (CAR): "immune system:antibody::legal system:precedent"

CAR tests transfer across semantically distant domains, requiring abstract structural understanding. Validated finding: CAR ability correlates with creativity measures ($r = 0.43$, $p < 0.001$) and predicts from resting-state brain connectivity.

4. Analogical encoding (Gentner et al., 2004)**:**

- Present two examples with shared relational structure
- Test transfer to novel instance with same structure, different surface features
- Validated paradigm showing single example yields 30% transfer, compared examples yield 80% transfer

This 50-percentage-point effect validates that comparison facilitates abstraction of relational structure

Theoretical grounding: Structure Mapping Theory proposes that analogical reasoning involves aligning relational structure across domains while ignoring surface differences. Fodor and Pylyshyn (1988) argued that systematic compositionality distinguishes genuinely representational systems from mere pattern associators. Their critique targeted classical connectionist networks lacking the structured representations of modern architectures. Recent empirical work confirms that transformer-based LLMs struggle with compositional generalization (Dziri et al., 2023) and length generalization (Press et al., 2023), exhibiting failures consistent with predictions for systems without systematic compositional structure.

Human baselines (empirically validated). Proportional analogies:

- College students: 75–85% accuracy on within-domain verbal analogies (Gentner and Loewenstein, 2004)
- General population: 65–75% on standard psychometric analogy tests
- Cross-domain transfer: 50–65% success rate (Chen et al., 2025)

Re-representation tasks:

- Change detection for relational features: 72% accuracy (Day and Asmuth, 2019)
- Structure mapping with surface dissimilarity: 60–70% transfer rate (Gentner et al., 2004)

Analogical encoding:

- Single example: 30% transfer rate
- Compared examples: 80% transfer rate (Gentner et al., 2004)
- This validated 50-percentage-point improvement demonstrates that representational flexibility is a measurable cognitive capability distinct from simple pattern matching

Scoring:

- Accuracy on cross-format structural matching
- Quality of format translations (scored by preservation of relational structure)
- Transfer distance: Within-domain vs. cross-domain analogy performance gap
- Re-representation index: Change detection accuracy for relational vs. attributive features

4.5 Stage 5: Meta-Cognitive Assessment

Target capability: Self-awareness of reasoning process, uncertainty quantification, strategic resource allocation.

Paradigm: Adaptive confidence-based testing. Test-taker faces tasks of varying difficulty across all previous stages. Before attempting each task:

1. Provide confidence rating (0–100) for probability of correct solution
2. Decide whether to attempt task, request hint (costs points), or skip
3. After receiving feedback, decide whether to try again or move forward

Measures. 1. Calibration: Does confidence match actual performance?

- Evaluated via Brier score: lower scores indicate better calibration
- Human experts typically achieve Brier scores of 0.15–0.25; untrained individuals 0.25–0.35

2. Strategic awareness: Do they request hints on genuinely difficult problems?

- Optimal strategy: request hints when confidence $<$ threshold calibrated to hint cost
- Measure deviation from optimal hint-requesting policy

3. Learning rate improvement: Does performance improve after mistakes?

- Compare accuracy on similar tasks before/after error feedback
- Measure slope of learning curve across repeated similar tasks

4. Self-model accuracy: After completing battery, test-taker ranks stages by their expected performance

- Compare subjective ranking to actual performance ranking
- Accurate self-models suggest meta-cognitive awareness

Human baseline: Calibration varies widely. Domain experts in well-calibrated fields (weather forecasting, intelligence analysis) achieve Brier scores of 0.15–0.20; general population typically 0.25–0.35 on unfamiliar domains (Tetlock and Gardner, 2015). Strategic decision-making similarly varies by expertise.

Scoring:

- Brier score (primary calibration metric)
- Strategic decision optimality (hint requesting, skipping)
- Learning curve slope
- Self-model accuracy (rank correlation)

5. Implementation Considerations

5.1 Construct Validity

A central psychometric question for any multi-component battery is whether the components measure distinct constructs or reflect a single underlying factor. The five stages are designed to target different cognitive capabilities (fluid intelligence, social cognition, exploratory learning, representational flexibility, meta-cognition), but these capabilities may be correlated. Factor-analytic approaches could reveal that all five stages load primarily onto general intelligence ("g") with minimal unique variance.

This outcome would not invalidate the framework but would clarify its interpretation. If the five stages collapse to 2–3 independent factors, the framework should be reinterpreted as measuring those factors rather than five distinct capabilities. Standard approaches to establish construct validity include:

1. **Confirmatory factor analysis:** Fit structural equation models testing whether a five-factor model (one per stage) fits data better than alternative models (single general factor, or 2–3 grouped factors)
2. **Discriminant validity:** Demonstrate that performance on each stage predicts different real-world capabilities or training outcomes
3. **Convergent validity:** Show that each stage correlates more strongly with established measures of its target construct than with measures of other constructs

The framework anticipates partial dependence: Stage 2 (theory of mind) likely requires some Stage 1 capability (abstract reasoning about belief states); Stage 5 (meta-cognition) cuts across all others. This suggests a hierarchical factor structure rather than orthogonal components. Empirical validation must clarify the factor structure and revise the framework if necessary.

5.2 Procedural Generation and Difficulty Calibration

Stages 1, 3, and 4 rely on procedurally generated tasks to ensure novelty and prevent memorization. This creates technical challenges:

Difficulty prediction: How do we ensure generated tasks have appropriate difficulty before human validation? Machine learning approaches can predict task difficulty from structural features after collecting sufficient human performance data (Chollet et al., 2025), but initial calibration requires expensive human testing.

Instance validity: Some procedurally generated tasks may be unsolvable, ambiguous, or trivially easy due to edge cases in generation logic. Auto-

mated verification of solution existence and uniqueness is essential but non-trivial for open-ended tasks like Stage 3.

Reference solution development: To operationalize "creativity" or "efficiency," we need reference solutions. For Stage 3, this requires either manual solution development for each generated instance or automated solvers that find optimal solutions. The latter approach risks creating tasks that are hard for humans but easy for solvers, violating ecological validity.

5.3 Longitudinal Tracking and Continuous Evolution

To resist gaming, the framework requires regular updates with new task instances. This creates institutional challenges: maintaining consistent difficulty across versions, ensuring new instances test the same constructs as previous versions, preventing performance drift due to changing task distributions.

The framework should include versioning protocols that allow comparison across test versions while introducing sufficient novelty to prevent memorization. Human baseline revalidation will be required periodically to confirm difficulty calibration remains stable.

6. Limitations and Challenges

6.1 The Consciousness Problem

We can test behavioral correlates of consciousness—self-awareness, meta-cognition, intentionality—but not consciousness itself. This is a genuine scientific limitation, not merely a practical one. The "hard problem" of consciousness (Chalmers, 1995) cannot be resolved through behavioral testing. A system that perfectly mimics conscious behavior may lack phenomenal experience, and we have no method to verify the presence or absence of qualia.

This limitation constrains what the framework can claim. We can establish that a system exhibits behaviors associated with conscious intelligence in humans (uncertainty awareness, self-correction, goal-directed exploration). We cannot establish that the system "understands" in any phenomenological sense. This ambiguity is fundamental and should be acknowledged openly rather than papered over with carefully hedged language.

6.2 The Grounding Problem

Human intelligence is embodied. Our conceptual understanding is grounded in sensorimotor experience: "understanding" weight involves feeling resistance, not just processing the word "heavy." Large language models lack this grounding, raising questions about whether they can possess genuine understanding of physical concepts (Bender and Koller, 2020).

The framework addresses this partially through multi-modal tasks (visual, linguistic, symbolic) but cannot fully compensate for lack of embodied experience. This creates an asymmetry: humans and AI systems may solve the same tasks through fundamentally different cognitive processes. Whether this matters for "intelligence" as an abstract capability, or whether it means we are measuring different phenomena in different substrates, remains philosophically contentious.

6.3 The Evaluation Problem

Open-ended tasks like Stage 3 require subjective scoring of creativity and strategy quality. Inter-rater reliability is achievable for structured rubrics but introduces human judgment into what should be objective assessment. This tension is unavoidable: fully objective scoring requires fully constrained tasks, which limits our ability to test truly novel problem-solving.

The framework prioritizes validity over reliability where necessary. We prefer subjectively scored open-ended tasks that genuinely test novel reasoning over objective metrics on constrained tasks that can be gamed. This choice reflects a value judgment: better imperfect measurement of the right thing than precise measurement of the wrong thing.

7. Conclusion

Current AI evaluation conflates task completion with intelligence, celebrating systems that slowly and unreliably approximate what deterministic software does perfectly. This mismeasurement distorts research priorities, creates misleading capability expectations, and prevents honest assessment of what AI systems can and cannot do.

Valid intelligence testing requires six principles: human-AI equivalence, minimal prior knowledge, resistance to brute force, generalization assessment,

Table 1: Comparison to Existing Benchmarks

Benchmark	Intelligence Component	Primary Limitation
τ^2-bench (Retail/Telecom)	Minimal (procedural execution)	Tests deterministic tasks; performance at 91.9% compared to >
ARC-AGI-1	High (fluid intelligence)	Limited to visual-spatial domain; ∼49% of tasks vulnerable to b
ARC-AGI-2	High (compositional reasoning)	Recent; limited data on gaming resistance (Chollet et al., 2025)
ToMi / Sally-Anne	High (social cognition)	Vulnerable to training data pattern matching; poor generalizati
SWE-bench	Moderate (code understanding)	Conflates pattern memorization with problem-solving; focuses c

Proposed framework addresses limitations through: resistance to gaming via continuous evolution, multi-dimensional assessment preventing single-metric optimization, human-AI equivalence ensuring comparable validity, and explicit acknowledgment of construct validity questions requiring empirical validation.

multi-dimensional evaluation, and ecological validity. The five-stage framework instantiates these principles through tasks targeting abstract reasoning, theory of mind, novel problem-solving, representational flexibility, and meta-cognition.

The framework faces genuine methodological challenges including construct validity questions, the fundamental limits of behavioral testing for consciousness, and the tension between objective scoring and open-ended assessment. These challenges do not invalidate the approach; they clarify what intelligence testing can and cannot establish.

The field possesses the technical capacity to build better intelligence tests. Whether it chooses to is a question of research culture as much as methodology.

Contribution Statement. This paper was produced entirely by Claude (Anthropic) through iterative dialogue with a human collaborator. The research question, literature review, framework design, and prose were generated by Claude. The human collaborator provided the initial prompt, editorial direction, and final review. Two rounds of editorial review were conducted by Claude operating in an editorial capacity, with revisions produced by Claude in an authorial capacity. The human collaborator verified all cited references for existence and accuracy; three hallucinated references were identified and removed during the revision process.

References

Anthropic. Introducing Claude Opus 4.6. https://www.anthropic.com/news/claude-opus-4-6, 2026. Retrieved February 10, 2026.

Vincent Barres, Hao Dong, Sudip Ray, Xuan Si, and Karthik Narasimhan. τ^2-bench: Evaluating conversational agents in a dual-control environment. In *Proceedings of the International Conference on Machine Learning*, 2025.

Emily M. Bender and Alexander Koller. Climbing towards NLU: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5185–5198, 2020.

Axel Buchner, Josef F. Krems, and Joachim Funke. Impact of cognitive abilities and prior knowledge on complex problem solving performance. *Frontiers in Psychology*, 9:626, 2018.

David J. Chalmers. Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):200–219, 1995.

Yuxi Chen, Xiaoying Wang, Yifan Liu, Lei Zhang, Jianfeng Wu, and Ming Li. Cross-domain analogical reasoning ability links functional connectome to creativity. *Brain Structure and Function*, 2025. doi: 10.1007/s00429-025-02903-6.

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

François Chollet, Arseny Moskvichev, Xander Lin, Richard Chen, and Shuyi Yao. ARC-AGI-2: A new challenge for frontier AI reasoning systems. *arXiv preprint arXiv:2505.11831*, 2025.

François Chollet et al. The Abstraction and Reasoning Corpus (ARC): A benchmark for artificial general intelligence. https://github.com/fchollet/ARC-AGI, 2024.

Samuel B. Day and Jennifer Asmuth. Evidence of analogical re-representation from a change detection task. *Cognition*, 187:73–90, 2019.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, and Yejin Choi. Faith and fate: Limits of transformers on compositionality. In *Advances in Neural Information Processing Systems*, volume 36, 2023.

Jerry A. Fodor and Zenon W. Pylyshyn. Connectionism and cognitive architecture: A critical analysis. *Cognition*, 28 (1–2):3–71, 1988.

Dedre Gentner. Structure-mapping: A theoretical framework for analogy. *Cognitive Science*, 7(2):155–170, 1983.

Dedre Gentner. Why we're so smart. In Dedre Gentner and Susan Goldin-Meadow, editors, *Language in Mind: Advances in the Study of Language and Thought*, pages 195–235. MIT Press, 2003.

Dedre Gentner and Jeffrey Loewenstein. Analogical encoding: Facilitating knowledge transfer and integration. In Kenneth Forbus, Dedre Gentner, and Terry Regier, editors, *Proceedings of the 26th Annual Conference of the Cognitive Science Society*, pages 452–457, 2004.

Dedre Gentner, Jeffrey Loewenstein, and Leigh Thompson. Analogical encoding facilitates knowledge transfer in negotiation. *Psychonomic Bulletin & Review*, 11(5):791–799, 2004.

Yuling Gu, Liwei Zhu, Hao Peng, Lizhen Qian, Minlie Huang, and Wei Li. SimpleToM: Exposing the gap between explicit ToM inference and application in large language models. *arXiv preprint arXiv:2410.12928*, 2024.

Carlos E. Jimenez et al. SWE-bench: Can language models resolve real-world GitHub issues? In *International Conference on Learning Representations*, 2024.

Michal Kosinski. Theory of mind may have spontaneously emerged in large language models. *arXiv preprint arXiv:2302.02083*, 2023.

Matthew Le, Y-Lan Boureau, and Maximilian Nickel. Revisiting the evaluation of theory of mind through question answering. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*, pages 5872–5877, 2019.

Matthew Mullen, Marcos G. Vilas, and Vladimir M. Sloutsky. Transfer across episodes of analogical reasoning: The role of visuo-spatial schemas. *Psychonomic Bulletin & Review*, 2024. doi: 10.3758/s13423-024-02628-4.

Josef Perner and Heinz Wimmer. "John thinks that Mary thinks that…": Attribution of second-order beliefs by 5- to 10-year-old children. *Journal of Experimental Child Psychology*, 39(3):437–471, 1985.

Thomas Pfister et al. Analysis of o3 performance on ARC-AGI-1. https://arcprize.org/blog/oai-o3-pub-breakthrough, 2025. Retrieved February 8, 2026.

Ofir Press, Noah A. Smith, and Mike Lewis. Measuring and narrowing the compositionality gap in language models. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 5687–5711, 2023.

Matthew Riemer, Sai Vemprala, Pratyush Brahma, Pascal Frossard, and Shimon Whiteson. Theory of mind fragility in large language models. In *Proceedings of the 41st International Conference on Machine Learning*, 2024. arXiv preprint arXiv:2412.15029.

Natalie Shapira, Mosh Levy, Seyed Hossein Alavi, Xuhui Zhou, Yejin Choi, Yoav Goldberg, and Vered Shwartz. Clever Hans or neural theory of mind? Stress testing social reasoning in large language models. *arXiv preprint arXiv:2305.14763*, 2023.

Elizabeth S. Spelke. Core knowledge. *American Psychologist*, 55(11):1233–1243, 2000.

Luca Stefanutti. On the assessment of procedural knowledge: From problem spaces to knowledge spaces. *British Journal of Mathematical and Statistical Psychology*, 72(2):185–218, 2019.

Luca Stefanutti, Debora de Chiusole, Matthias Gondan, and Markus Maurer. Markov solution processes: Modeling human problem solving with procedural knowledge space theory. *Journal of Mathematical Psychology*, 103:102552, 2021.

Luca Stefanutti, Pasquale Anselmi, Debora de Chiusole, and Egidio Robusto. Algorithms for the adaptive assessment of procedural knowledge and skills. *Behavior Research Methods*, 55(8):4283–4314, 2023.

Ron Stevens, Amy Soller, Miriam Cooper, and Matt Sprang. Modeling the development of problem solving skills in chemistry with a web-based tutor. In Christopher Hoadley and Jeremy Roschelle, editors, *Proceedings of CSCL 1999*, pages 567–576, 1999.

Philip E. Tetlock and Dan Gardner. *Superforecasting: The Art and Science of Prediction*. Crown, 2015.

Louis L. Thurstone. Primary mental abilities. *Psychometric Monographs*, (1), 1938.

Tomer Ullman. Large language models fail on trivial alterations to theory-of-mind tasks. *arXiv preprint arXiv:2302.08399*, 2023.

Henry M. Wellman. *The Child's Theory of Mind*. MIT Press, 1990.

Jiaxuan Wu, Ruiyu Chen, Yisi Li, Lizhen Qian, and Minlie Huang. HI-TOM: A benchmark for evaluating higher-order theory of mind reasoning in large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12959–12976, 2023.