

The Reversed Turing Test: Measuring Superhuman AI Performance Beyond Imitation

MAIKEL LEON
Department of Business Technology
University of Miami
Miami, Florida,
USA

Abstract: For more than seven decades, Alan Turing's imitation game has shaped how the AI community discusses machine intelligence, although it was never a psychometric instrument. AI systems now outperform humans in strategic games, scientific prediction, and pattern recognition, achievements better described as domain-specific superhuman performance than as general intelligence unless a novelty condition is enforced. This paper argues that imitation-based evaluation is misaligned with such systems and introduces the Reversed Turing Test (RTT), which asks whether a machine can be identified precisely because its performance is measurably superhuman. The test combines a standardized superhuman index (SI), expressed in standard-deviation units relative to a human reference sample, with a panel of judges whose identifications are verified against the performance evidence. The paper provides a formal definition of the RTT, a taxonomy of novelty types with verification methods, a worked example in protein structure prediction, and a pre-registered text-generation study ($n = 500$; composite SI ≈ 3.2 ; 90% identification accuracy grounded in stylometric cues), interpreted as superhuman stylometric performance rather than superhuman intelligence. The RTT is compared against established benchmarks, its philosophical limits are stated, and implications for trust, creativity, governance, and evaluation design are discussed.

Key-Words: Reversed Turing Test; superhuman AI; AI evaluation; superhuman index; psychometric benchmarking; large language models.

Received: April 14, 2026. Revised: May 29, 2026. Accepted: July 5, 2026. Published: September 23, 2026.

1 Introduction

Since Turing's 1950 paper, [1], the idea that machine intelligence can be operationally defined through an imitation game has profoundly shaped how the AI community thinks about evaluation. In the so-called Turing Test, a human judge holds a text-based conversation with two hidden interlocutors (one human, one machine) and must decide which is which. The setup has captured the public imagination, inspired competitions such as the Loebner Prize, and fueled long-running philosophical debates about consciousness and understanding, [2], [3]. At the outset, however, one point deserves emphasis: the Turing Test is not a psychometric instrument and has had little influence on the scientific study of human intelligence. It lacks the standardization, norm-referencing, and validated properties that instruments of cognitive ability are expected to possess. Its role in this paper is therefore the one it has historically played within AI, that of a cultural and conceptual touchstone for thinking about what it means to count as intelligent.

Even as the field of AI evaluation has moved toward task-based benchmarks and domain-specific

metrics, as documented in recent comprehensive surveys of large-language-model evaluation, [4], imitation remains a live conceptual reference point. Many of today's most widely deployed AI systems (conversational assistants, customer-service bots, and general-purpose large language models) are still implicitly evaluated by how convincingly they reproduce human communicative behavior. It is this residual primacy of imitation that the paper challenges.

The motivation is straightforward: AI systems now exhibit genuinely superhuman performance in several well-defined domains. Deep learning and large-scale self-play have produced agents whose strategies and representations reach far beyond human expertise: AlphaGo and its successors in Go, chess, and shogi, [5], [6], [7]; OpenAI Five in Dota 2, [8]; and AlphaFold in protein structure prediction, [9], whose subsequent generation extended accurate prediction to protein interactions with ligands, nucleic acids, and other biomolecules, [10]. Far from being explicable by mimicry, these achievements are prized exactly because they extend beyond what any human could

do. In the domain of Go, moreover, superhuman programs have been shown to improve human decision-making by prompting players to adopt novel strategies, [11].

The proposal of this paper is the Reversed Turing Test (RTT). Rather than asking whether a machine can pass for a human, the RTT asks whether a machine can be identified as a machine because its performance is measurably superhuman. Two ingredients make this concrete: first, a standardized superhuman index (SI), expressed in standard-deviation units relative to a human reference sample, the same kind of metric that psychometric assessment uses for cognitive ability; and second, a panel of judges whose identifications are anchored in that quantitative performance evidence rather than in surface impressions. Throughout, intelligence is treated as a multidimensional construct in the psychometric tradition, an approach that a growing body of work has begun to extend to artificial systems by administering standardized psychometric inventories to large language models, [12]. Building on this tradition, the paper adopts a working operational definition (formalized in Section 3) that emphasizes adaptive capacity on genuinely novel tasks rather than stored domain expertise, a distinction the RTT enforces through a novelty condition on its task battery.

The paper proceeds as follows. Section 2 briefly traces the evolution of AI from symbolic systems to modern foundation models and reviews the Turing Test alongside related benchmarks. Section 3 provides a formal definition of the RTT, a taxonomy of novelty types with corresponding verification methods, and a worked illustrative example in protein structure prediction. Section 4 compares the RTT with existing benchmarks, maps out the domains in which it currently applies, and examines its philosophical and societal implications. Section 5 reports a pre-registered empirical instantiation of the RTT in the domain of text generation ($n = 500$ participants). Finally, Section 6 discusses broader implications, states limitations openly, and sketches future directions, before Section 7 concludes. In summary, the classical Turing Test centers on deception and conversational mimicry and, not being a psychometric instrument, has had little impact on the scientific study of intelligence. Today's AI systems, by contrast, achieve genuinely superhuman performance on well-defined tasks. The RTT quantifies that performance through a standardized metric rather than subjective judgment, and it enforces a novelty condition so that superhuman performance supports a claim about intelligence rather than one about memorized domain expertise.

2 Machine Intelligence in Brief

Before laying out the proposal in detail, it is worth sketching briefly where contemporary AI comes from and what the existing evaluation landscape looks like. The history of machine intelligence spans several distinct eras, summarized in Figure 1 (Appendix). Early AI relied on symbolic reasoning and rule-based expert systems: the General Problem Solver exemplified logic-based symbolic reasoning, and an early checkers-playing program demonstrated machine learning through self-play, [13]. The statistical and connectionist revival of the 1980s–1990s, anchored by the perceptron model, [14], eventually gave way to the deep learning era of the 2010s. Transformers, [15], [16], and large language models, [17], [18], then enabled remarkable language generation, while reinforcement learning combined with deep networks produced machines that surpass human performance in games and control environments, [5], [8], [19], [20]. Foundation models trained on massive unlabeled data today support a wide range of tasks, and with them have come fresh questions about capability, bias, and control, [21], [22], [23].

Alongside these recent breakthroughs, a handful of earlier milestones have already made the possibility of superhuman AI concrete. Deep Blue's 1997 chess victory over Kasparov, [24], Watson's win on Jeopardy!, [25], and more recently a machine-learning weather model that surpasses the leading operational system on a majority of forecast targets, [26], illustrate three complementary modes of superhuman performance: search speed, memory and retrieval, and predictive accuracy. Each is directly measurable against human baselines. Together with the reinforcement-learning and foundation-model examples from Section 1, these cases establish superhuman performance as a well-documented empirical phenomenon, not a forecast. That is why the evaluation paradigm itself must be rethought.

2.1 AI Tests and Their Limitations

Turing's imitation game replaced the vague question "Can machines think?" with an operational criterion based on textual indistinguishability. Its technology-agnostic, behaviorist framing has had a broad cultural impact (for example, via the Loebner Prize). However, it is not a canonical instrument in intelligence research and has not been subjected to formal psychometric evaluation, [3]. Passing the test, moreover, does not entail human-level intelligence, [27].

The Chinese Room thought experiment argues that running a formal program does not by itself create understanding or intentionality, [2]: a person following rules for Chinese can appear to pass

a Turing Test without understanding a word. These critiques of the Turing Test, particularly that behavioral indistinguishability does not entail understanding, apply equally to the RTT, as acknowledged explicitly in Section 4.3. Alternatives include the Lovelace Test, which demands that a machine produce an original creative output for which its programmers cannot account, [28], and the Winograd Schema Challenge, which examines commonsense reasoning, [29].

In short, machine intelligence has evolved from symbolic systems to deep learning and foundation models, with each era producing qualitatively new superhuman capabilities. These capabilities manifest as superior search speed, memory capacity, and predictive accuracy, all of which can be measured objectively against human baselines. The widely cited AI tests of the past, the Turing Test most prominent among them, have had limited impact on psychometrics and lack formal psychometric validation. Those limitations are exactly what the RTT is designed to address, and the paper turns to its formal definition next.

3 Methodological Framing

This section moves from the historical backdrop sketched in Section 2 to the proposal itself. Its goal is to set out what the RTT is, what it is intended to measure, and how one would actually run an RTT study. The exposition proceeds in four movements. It first fixes a working definition of intelligence and states the formal definition of the RTT. It then develops a taxonomy of novelty types with matched verification methods, sets out the design commitments on which an RTT study rests, and closes with a concrete worked example in the domain of protein structure prediction, which anticipates the empirical instantiation reported later in Section 5.

3.1 Defining Intelligence and the RTT

Before defining the RTT, it is necessary to specify what construct it is intended to evaluate. The psychometric tradition distinguishes between fluid intelligence (Gf), the capacity to reason and solve genuinely novel problems, and crystallized intelligence (Gc), the accumulated knowledge and skills acquired through experience. Within Cattell–Horn–Carroll-style frameworks, both are facets of general cognitive ability, but Gf is the canonical target when researchers wish to measure adaptive capacity rather than stored expertise. Following this tradition, the present work adopts the following working operational definition. Intelligence, in the sense relevant to the RTT, is the capacity to produce correct solutions to problems whose specific instances were not

previously encountered during training, across tasks that require reasoning, pattern recognition, memory integration, or planning. This definition aligns with a Gf-weighted view of intelligence and deliberately excludes highly optimized performance on tasks fully covered by the system’s training data; such performance is more appropriately characterized as domain-specific achievement or crystallized expertise. The distinction matters: a chess engine trained on millions of grandmaster games may demonstrate extraordinary achievement in chess without necessarily exhibiting generalized intelligence. The RTT therefore imposes a novelty condition: evaluated tasks must be drawn from a distribution not covered by the system’s documented training data, verified to the extent practicable (e.g., by post-dating the training cut-off, by controlled training-data exclusion, or by explicit decontamination analyses). Importantly, the novelty condition is necessary but not sufficient for an RTT verdict to support an intelligence claim. Even novel task instances may be structurally similar to training exemplars (e.g., a new protein with common folds), so RTT results should be interpreted as evidence for superhuman performance on the specific task distribution evaluated, and only suggestive of generalized intelligence to the extent that the novelty condition is stringently verified and that multiple cognitively distinct task batteries are passed.

With this definition in place, the RTT can be defined formally as follows. **Definition (Reversed Turing Test).** Let $\mathcal{T} = \{t_1, t_2, \dots, t_k\}$ be a pre-registered battery of tasks drawn from a domain in which (i) human performance on a specified metric m can be measured and summarized by a location parameter μ_H and a scale parameter σ_H estimated from a reference sample $\mathcal{S}_H$; and (ii) the tasks satisfy the novelty condition relative to the evaluated AI system $\mathcal{A}$. Let $s_{\mathcal{A}}$ denote the score of $\mathcal{A}$ on metric m averaged over $\mathcal{T}$. Define the superhuman index

$$\text{SI}(\mathcal{A}) = \frac{s_{\mathcal{A}} - \mu_H}{\sigma_H},$$

expressed in standard-deviation-like units relative to the reference sample.

When the human distribution on m is approximately Gaussian, $\mathcal{S}_H \sim \mathcal{N}(\mu_H, \sigma_H^2)$, SI admits the familiar percentile interpretation of an IQ-style deviation score. When the distribution is skewed, heavy-tailed, or bounded, a rank-based SI is recommended, $\text{SI}_{\text{rank}} = \Phi^{-1}(\hat{F}_H^*(s_{\mathcal{A}}))$, where $\hat{F}_H^*$ is a plotting-position estimate of the cumulative distribution, defined for a sample of size n_H as $\hat{F}_H^*(s) = (\text{rank}(s) - 0.5)/n_H$ and clipped to the interval $[0.5/n_H, 1 - 0.5/n_H]$ to

prevent $\pm\infty$ when s_A lies at or beyond the sample extremes. In such extreme cases the reported value is $SI_{\text{rank}} > \Phi^{-1}(1 - 0.5/n_H)$ rather than any finite value, making the boundary limitation explicit. This transformation preserves the SD-unit interpretation without requiring normality.

The system $\mathcal{A}$ passes the RTT in domain d if and only if: (1) $SI(\mathcal{A}) \geq \theta$ for a pre-specified threshold $\theta > 0$; and (2) a panel of human evaluators, shown anonymized outputs (and, at the evaluator's discretion, accompanying metric summaries) but not the identity of the producer, correctly identifies $\mathcal{A}$'s outputs as machine-generated at a rate significantly above chance (one-sided binomial test, $p < 0.05$), with post-hoc analysis of evaluator justifications confirming that identifications are grounded in superhuman performance evidence rather than mere formatting or stylistic artifacts.

Threshold guidance. Following psychometric conventions, the recommended threshold is $\theta \geq 2$ (approximately the 97.7th percentile of the reference sample; parallel to the two-SD criterion used in intellectual-giftedness identification) for initial claims of superhuman performance, and $\theta \geq 4$ for strong claims of unambiguously superhuman capability. Thresholds should be pre-registered together with the task battery.

Reference sample as sample, not norm. The parameters μ_H and σ_H are estimated from $\mathcal{S}_H$ and are thus sample statistics; SI is interpreted as a sample-referenced deviation rather than a fully norm-referenced score. A full norm-referenced extension would require a standardization sample with demographic stratification, and the sample sizes typical of psychometric norms, an important direction for future RTT development, acknowledged as outside the scope of any single-study instantiation.

This formulation grounds identifiability in a quantitative performance index rather than perceptual judgment alone. A system is not declared superhuman simply because judges find its outputs "too polished"; the superhuman index must first demonstrate that performance lies above the reference distribution, and judges must then confirm recognizability.

What the judge panel adds beyond the SI. A reasonable objection is that once the SI establishes superhuman performance through objective measurement, a panel of human judges is redundant. The two components are kept distinct because they answer logically independent questions. The SI answers a question of *measurability*: does the system's performance, scored on metric m , lie above the human reference distribution? The judge panel answers a question of *recognizability*: is that superiority perceptible to human evaluators

from the outputs themselves? These properties can dissociate in both directions. A system may be measurably superhuman on m yet produce outputs that humans cannot reliably distinguish from strong human work, because the performance gap resides in dimensions below the threshold of human discrimination; conversely, judges may confidently flag a system that is not in fact superhuman, keying on incidental surface artifacts such as formatting. Requiring both a threshold on the SI and above-chance, performance-grounded identification guards against each of these failure modes, neither of which the SI alone can detect. The judge panel therefore serves three concrete functions rather than one: it is a convergent-validity check, confirming that the performance dimensions the SI quantifies are the same dimensions on which human identification actually turns; it establishes practical, deployment-relevant recognizability, since real-world detection, provenance, and calibrated trust depend on human perceptibility and not on an offline metric alone; and it provides a discriminant safeguard, because the coding of justifications is designed to catch identifications that rest on artifacts rather than performance.

Understood this way, the requirement that identifications be grounded in performance evidence is not circular. The SI and the judge test are separate measurements taken on the same outputs, and the coding of justifications is the bridge that tests whether the two track a common underlying signal. When judges cite the very dimensions the SI operationalizes, that convergence is the positive result the design is built to detect, not a sign that one component duplicates the other. The two-step criterion is loosely analogous to norm-referenced approaches to cognitive-ability identification, though, as noted above, a single-study RTT instantiation is not equivalent to a fully standardized psychometric instrument.

When m is itself a composite of multiple sub-metrics (as in the text-generation experiment), the composite SI should be understood as a *formative* index rather than a *reflective* one. In a reflective measurement model, several observed indicators are treated as interchangeable manifestations of a single latent cause, so that high internal consistency among them is expected and desirable. A formative index instead combines constituent dimensions that jointly define a composite by construction; the dimensions are not required, and are not expected, to be strongly intercorrelated or to reflect one common latent trait. On this reading, the observation that the sub-indices need not measure the same latent construct is not a defect to be corrected but the expected property of a formative composite, and internal-consistency

criteria such as Cronbach's α are simply the wrong yardstick. What a formative index does require is that its constituents be conceptually justified as distinct, non-redundant facets of the target construct, that the aggregation rule be pre-specified, and that the result be reported transparently. For this reason the composite SI is interpreted throughout as a summary statistic over several facets of output quality rather than as a direct measurement of a unitary latent trait; it is reported alongside its per-dimension SIs so that no single number is asked to carry more meaning than it can bear. Its sensitivity to the aggregation rule is examined explicitly. The construct justification for the four text-generation facets is given in Section 5, and the implications of the reification risk, together with the accompanying robustness checks, are discussed in Section 6.4.

It follows that the RTT is at its most straightforward in domains that supply a single, objectively scorable metric, such as GDT_TS in protein structure prediction or Elo rating in games, where the SI is a deviation score on one agreed dimension, and no aggregation question arises. Domains such as text generation, which lack any single canonical metric, are harder cases: the RTT remains applicable, but only under the stricter discipline that the composite be treated as formative, that per-dimension results accompany the aggregate, and that the resulting claim be confined to superhuman performance on the specified dimensions rather than extended to superhuman intelligence. The text-generation instantiation reported below is presented in exactly this spirit, as a deliberately demanding test of whether the protocol can be carried out where no single objective metric exists.

3.2 A Taxonomy of Novelty

The novelty condition is the hinge on which the distinction between achievement and intelligence turns, yet “novelty” is not a single property. Topical novelty, the sense most easily operationalized through post-cutoff prompting, is also the weakest: a model may respond well to a genuinely new topic by generalizing from structurally similar training exemplars, so that topical novelty alone leaves open whether the system is adapting to an unfamiliar problem or pattern matching against a familiar one. To make the condition precise and to give researchers concrete verification targets, it is useful to distinguish four types of novelty, ordered roughly from the least to the most demanding, and to pair each with the verification methods appropriate to it, as summarized in Table 1.

Topical novelty concerns the surface subject matter: the specific entities, events, or facts are new to the system. It is verified by dating stimuli

Table 1. A taxonomy of novelty types for the RTT novelty condition, ordered from least to most demanding, with corresponding verification methods. Topical novelty is the type that post-cutoff prompting establishes; intelligence-grade claims require verification at the structural or procedural level.

Novelty type	What is new	Verification methods
Topical	Subject matter, entities, events	Post-cutoff dating; provenance checks; corpus decontamination (n-gram / embedding overlap)
Structural	Abstract problem form	Controlled data exclusion; held-out task families; nearest-neighbor / isomorphism analysis
Combinatorial	Configuration of familiar elements	Element enumeration; controlled recombination; compositional nearest-neighbor distance
Procedural	Required method or strategy	Post-cutoff solution methods; solution-trace comparison; adversarial operation composition

Source: created by the authors.

after the documented training cut-off, by provenance checks on sources, and by decontamination analyses that search the training corpus (where accessible) for near-duplicate n-grams or high-similarity passages. *Structural novelty* concerns the underlying problem structure: the abstract form of the task, not merely its topic, is absent from training. It is verified by controlled training-data exclusion, by holding out whole task families rather than individual items, and by expert judgment or embedding-space analysis confirming that no isomorphic problem appears among the nearest training neighbors. *Combinatorial novelty* concerns configuration: familiar elements are combined in arrangements not previously seen together. It is verified by enumerating the constituent elements and confirming the specific combination is absent from training, by constructing instances through controlled recombination, and by measuring distance to the nearest neighbor in a compositional feature space. *Procedural novelty* concerns the method: the solution requires a sequence of operations or a strategy not demonstrated in training. It is verified by requiring multi-step derivations whose solution methods were published or devised after the cut-off, by comparing solution traces against catalogs of known procedures, and by adversarially composing problems that demand an unseen combination of operations.

Two practical points follow. First, the types are cumulative in stringency rather than mutually exclusive: a battery may satisfy topical novelty while failing structural novelty, and studies should state explicitly which type or types they verify and by

which method, rather than reporting that the novelty condition was met without qualification. Second, the strength of an RTT verdict should scale with the type verified. Claims of domain-specific superhuman performance can rest on topical novelty alone. Still, any inference toward generalized adaptive capacity requires verification at the structural or procedural level, which in turn generally requires open-weight models with curated, auditable training data. The text-generation study in Section 5 verifies only topical novelty, and its claims are confined accordingly.

3.3 Core Design Commitments

The formal definition above is the backbone of an RTT study, but running one in practice requires some additional scaffolding. The RTT draws on the broader tradition of AI evaluation while shifting the emphasis from imitation to augmentation. A well-designed RTT study rests on five design commitments:

1. *Task selection.* Tasks must have well-defined, objectively scorable outcomes and must satisfy the novelty condition. Appropriate domains include strategy games with well-defined payoffs, scientific prediction (e.g., protein structure, weather), mathematical proof verification, and large-scale information retrieval. A battery spanning multiple cognitive dimensions (speed, memory, reasoning, pattern recognition) provides a richer picture than any single task.
2. *Human baseline.* A representative sample of human participants completes the same tasks under equivalent conditions to establish μ_H and σ_H on metric m . The baseline must be pre-registered and collected under standardized conditions to ensure comparability. In a large-scale study of professional Go, AI-induced novelty was quantified relative to 5.8 million human decisions spanning decades of play, [11], an exemplary implementation of this component.
3. *Transparency and safety.* Unlike the original Turing Test, the RTT discourages deception. Systems are expected to cite sources, explain reasoning, quantify uncertainty, and avoid hallucinations. Transparent reporting fosters trust and allows humans to learn from superhuman partners.
4. *Standardized evaluation metrics.* Instead of a binary pass/fail outcome, the RTT yields a profile of SI values across multiple domains, analogous to a multidimensional cognitive ability profile. Additional metrics may include error rate, time to solution, solution novelty

(assessed by independent domain experts), and interpretability.

5. *Societal impact assessment.* Because superhuman AI can influence markets, education, and policy, RTT evaluations should consider downstream effects and include stakeholders in the assessment process. An excessive focus on human-like AI can exacerbate inequality, whereas augmentation can yield broad benefits, [30].

3.4 Worked Example: Protein Folding

To make the protocol concrete, this subsection sketches how an RTT study would run in the domain of protein structure prediction. This domain is a natural fit: novel protein sequences can be synthesized and structurally determined in the laboratory (so the novelty condition can be enforced), and prediction quality is objectively measurable on an established scale. The numerical values below are illustrative and chosen to be consistent with published benchmarks; the aim here is to illustrate the RTT protocol step by step, not to certify any particular system.

Step 1: Task battery. Compile a pre-registered set of $k = 100$ protein sequences newly determined by X-ray crystallography after the AI system's training cut-off, ensuring the novelty condition. The metric m is the Global Distance Test Total Score (GDT_TS), which measures structural similarity between predicted and experimentally resolved structures on a 0–100 scale (higher is better).

Step 2: Human baseline. A cohort of expert structural biologists (reference sample $|S_H| = 30$) predicts the same 100 structures using standard computational tools available before the AI era, yielding illustrative parameters $\mu_H \approx 42$, $\sigma_H \approx 9$, broadly consistent with pre-AlphaFold CASP benchmarks reported in the literature.

Step 3: AI evaluation. Reported AlphaFold performance corresponds to a mean GDT_TS of approximately $s_A \approx 92$ on comparable test sets, [9], giving an illustrative superhuman index of $SI \approx (92 - 42)/9 \approx 5.6$. This easily clears both the recommended initial threshold ($\theta = 2$) and strong threshold ($\theta = 4$).

Step 4: Judge identification. A panel of domain experts (illustratively $n = 20$; a real deployment should use a larger panel with a formal power analysis) views anonymized predictions with their GDT_TS scores and qualitative accuracy reports. Judges would be expected to identify AI-generated predictions at well above chance rates. Their written justifications would then be coded against a pre-specified codebook to verify that identifications

cite superhuman accuracy rather than surface artifacts (such as file formatting).

Step 5: RTT verdict. Conditional on steps 1–4 being carried out with adequate sample sizes and pre-registration, AlphaFold would pass the RTT in the domain of protein structure prediction. The result would be interpretable (the SI is explicit), replicable (the task battery is pre-registered), and anchored in domain-expert evaluation.

This illustrative example makes clear that the RTT is not a single fixed test but a protocol that must be instantiated domain by domain; a full empirical instantiation in the domain of text generation is reported in Section 5.

Taken together, the formal machinery of the RTT rests on four commitments. First, it adopts a Gf-weighted, psychometrically grounded working definition of intelligence, and it enforces a novelty condition to keep adaptive capacity distinct from domain-specific achievement. Second, it quantifies superhuman performance through the standardized superhuman index (SI), expressed in standard-deviation units relative to a human reference sample, so that identifications do not rest on subjective impressions alone. Third, any well-designed RTT study rests on five design commitments: task selection, human baseline, transparency and safety, standardized metrics, and a societal-impact assessment. Finally, the worked example in protein structure prediction illustrates the full protocol and anticipates the empirical study reported in Section 5.

4 Comparison and Implications

Having defined the RTT formally in Section 3, the discussion now situates it within the landscape of existing evaluation paradigms (Section 4.1), maps out the domains where it is currently applicable (Section 4.2), and then examines its philosophical (Section 4.3) and societal (Section 4.4) implications. The analysis builds on the superhuman achievements already introduced in Sections 1 and 2, AlphaGo/AlphaZero in games, [6], [7], OpenAI Five in Dota 2, [8], and AlphaFold in structural biology, [9], without repeating the full catalog.

As emphasized in Section 3, these performances are most precisely characterized as domain-specific achievement when training data closely cover the tasks. The RTT's novelty condition is intended to push evaluation toward novel task instances, strengthening the inference from achievement to generalized capability. Even where the novelty condition is only partially met, the RTT yields valid statements about domain-specific superhuman achievement, a scientifically useful construct in its own right, provided the claim is not overstated.

Table 2. Comparison of evaluation paradigms, selected to represent the four major paradigms in AI research (imitation, creativity, reasoning, and human–bot discrimination).

Test	Focus	Strengths	Limitations
Turing Test	Imitation	Simple behavioral criterion	Encourages deception; chat-centric; no psychometric validation
Lovelace Test	Creativity	Demands originality	Hard to verify programmer influence
Winograd Schema	Commonsense reasoning	Measures pronoun disambiguation	Narrow domain; finite dataset; saturated by LLMs
CAPTCHA	Human vs. bot	Practical security	Quickly obsolete as AI improves
RTT (proposed)	Superhuman performance	Standardized metric; promotes transparency	Requires domain baselines; novelty must be verified

Source: created by the authors.

Significantly, superhuman performance can also benefit humans: an analysis of 5.8 million decisions by professional Go players found that after AlphaGo's release, human players adopted more novel strategies and improved their decision-making, [11]. Superhuman AI can thus act as a catalyst for human improvement within a given domain.

4.1 Comparison with Other Benchmarks

Table 2 contrasts the RTT with four alternative evaluations, chosen to span the major paradigms in AI evaluation: conversational imitation (Turing Test), creative output (Lovelace Test), semantic reasoning (Winograd Schema Challenge), and human–bot discrimination (CAPTCHA). Together, they make visible how each paradigm differs from the RTT's focus on measurable superhuman performance.

The Lovelace Test asks a machine to produce a creative work that its programmers cannot explain; the idea is intuitive, but verifying originality in a principled way is notoriously difficult. The Winograd Schema Challenge probes commonsense pronoun resolution and was deliberately designed to be easy for humans and hard for machines, a design that worked well until large pre-trained models effectively saturated the finite item set. CAPTCHA distinguishes humans from bots through tasks such as image classification, but modern vision networks routinely solve them, so each new generation of CAPTCHA is quickly rendered obsolete. The RTT proceeds differently: it deliberately targets tasks in which machines already exhibit measurable superhuman capability, and it uses human baselines to quantify the extent of that superiority through the standardized superhuman index.

Figure 1 depicts the classic Turing Test as a judge conducting text-only exchanges with a hidden human and machine before issuing a binary humanness

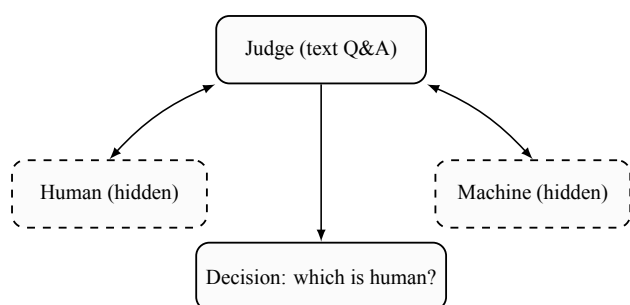


Fig. 1: Classic Turing Test with step annotations. The judge conducts text-only exchanges with a hidden human and a hidden machine, then issues a binary decision identifying which interlocutor is human.

Source: created by the authors.

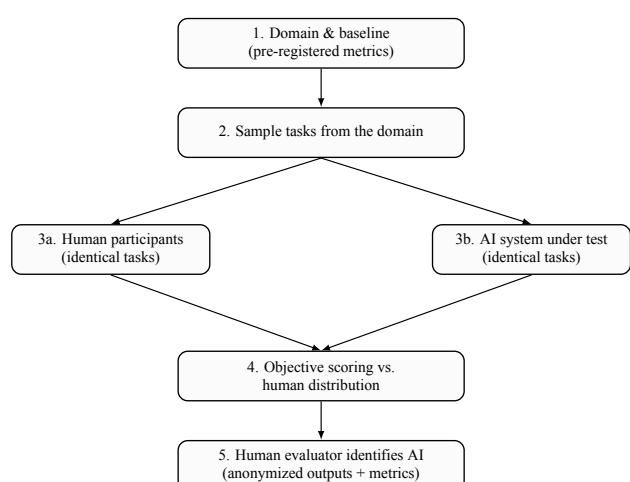


Fig. 2: Reversed Turing Test with step annotations. A domain with a robust human baseline is selected, and a pre-registered task battery is assembled. Both humans and the AI system complete identical tasks. Outputs are scored against the human performance distribution to compute the superhuman index, and a panel of human evaluators identifies the AI based on superhuman performance evidence, using anonymized outputs and metric summaries.

Source: created by the authors.

decision. Figure 2 depicts the RTT as a five-step, pre-registered, domain-grounded pipeline in which both humans and the AI complete identical tasks, the outputs are scored against the human performance distribution via the superhuman index, and a panel of judges identifies the AI based on a superhuman quality rather than conversational deception.

4.2 Taxonomy of Machine Performance

To keep the RTT's scope concrete, Table 3 classifies representative AI tasks by cognitive domain and performance relative to humans. Games such

Table 3. Tasks and typical machine performance relative to humans (state of the field at the time of writing). “Superhuman” indicates that state-of-the-art AI clearly exceeds expert humans on novel instances; “human-level” denotes parity; “subhuman” indicates inferior performance.

Domain	Representative tasks	Machine vs. human
Games	Go, Chess, Shogi, Dota 2, StarCraft II	Superhuman, [5], [7], [8], [20]
Science	Protein folding, weather prediction	Superhuman on narrow, novel instances, [9]
Vision	Image classification, segmentation	At or above expert human, [31]
Language	Machine translation, summarization	Approaching human; context and coherence limits, [17], [18]
Reasoning	Commonsense, moral judgment	Mixed; often subhuman, [32], [33]
Creative	Art, music composition	Improving; no consensus metric
Physical	Robotics, dexterous manipulation	Generally subhuman; progress in narrow tasks

Source: created by the authors.

as Go, chess, and StarCraft II already feature AI systems with well-documented superhuman performance, and are therefore natural first targets for RTT instantiation. Open-ended reasoning and conversation, by contrast, remain at or below human levels despite recent progress, and illustrate where the RTT is not yet applicable. The larger message of the table is simply that machine capability is uneven: the RTT is useful where a robust human baseline and clearly measurable performance both exist, and it makes no claim to apply universally.

4.3 Philosophical Implications

The RTT inevitably touches long-standing debates about machine understanding, and this subsection states plainly where it stands. The well-known Chinese Room argument holds that the syntactic manipulation of symbols is not sufficient for semantics or consciousness, [2]: a person mechanically following rules for Chinese can pass a behavioral test without understanding a word. This critique cuts both ways. A system that passes the RTT by producing superhuman outputs in a specific domain may well be doing so through sophisticated pattern matching, with no genuine comprehension at all. The RTT does not resolve the hard problem of consciousness, and no such claim is made here. Its scope is more modest: it asks whether an AI system's performance measurably and reliably exceeds the human distribution on well-defined tasks, and whether expert judges can identify it as such based on that performance. Questions of semantics, understanding, and inner experience remain open problems that no behavioral test can settle, [34].

What the RTT does support is a functional

orientation toward evaluation: what a system does, measured against a human reference distribution, is the relevant criterion for this test. This is weaker than classical functionalism about minds, which is concerned with the identity of mental states across different physical substrates. The paper adopts a functionalist stance only about evaluation, not about the nature of mind itself.

This shift carries a further consequence. The imitation game incentivizes deception: a machine that honestly admits it is a machine fails the test. The RTT inverts this incentive by treating transparency as a virtue. In doing so, it aligns naturally with recent work on explainable AI, [35], and alignment, [18], both of which ask that systems disclose what they are, what they can do, and how confident they are. It also challenges a kind of anthropocentric bias: focusing exclusively on human-like AI risks the so-called “Turing trap,” in which AI substitutes for human labor rather than augmenting it, [30]. The RTT is designed to encourage augmentation by celebrating the kinds of performance that are distinctive to machines rather than imitative of humans.

4.4 Societal Impact and Governance

Superhuman AI poses both opportunities and risks that go beyond the immediate evaluation framework. Economically, machines that excel at tasks formerly performed by experts could displace workers in specific sectors. Still, they could also free humans to focus on creative, interpersonal, and supervisory roles that AI currently cannot fill. The distinction between achievement and intelligence is socially important here: a system with superhuman achievement in a narrow domain should not automatically be treated as a general-purpose replacement for human workers.

Regulatory frameworks must ensure that the benefits of superhuman AI are equitably distributed and that systems are aligned with human values. Standard-setting bodies can incorporate the RTT’s superhuman index into certification processes, providing quantitative criteria for when a system exceeds human abilities in desirable ways without compromising safety. Mandatory disclosure of superhuman capabilities, analogous to the risk-tiered obligations established for high-risk AI in the EU AI Act and its proposed implementation, [36], would allow users, policymakers, and auditors to understand the degree to which a system departs from human-level performance. Education must adapt to prepare learners for collaboration with powerful AI, emphasizing meta-cognitive skills, critical evaluation of AI outputs, and the ability to identify failure modes. Ethical guidelines should address issues of accountability when AI agents make decisions at scales and speeds beyond human comprehension.

To summarize this section: the RTT is instantiated domain by domain, and the taxonomy in Table 3 maps where it currently applies. Compared with the Turing Test, the Lovelace Test, the Winograd Schema Challenge, and CAPTCHA, the RTT’s distinctive move is to make superhuman performance quantitative, through the SI metric, rather than binary. The Chinese Room objection applies to the RTT as much as to the Turing Test, so the RTT makes no claim about consciousness or understanding, only about measurable performance. Finally, shifting the evaluation paradigm from imitation to functional augmentation has direct governance consequences, from disclosure mandates for superhuman capability to RTT-based certification of high-stakes systems.

5 Empirical Validation

The conceptual case for the RTT predicts that a performance gap large enough to be perceptible will render machines identifiable as non-human agents; whether that gap is in fact perceptible is precisely what the recognizability component tests. The protein example in Section 3 illustrated the five-step protocol in principle; the protocol is now applied in full. This section reports a controlled study that instantiates the RTT in the domain of text generation. The study follows the five-step protocol of Section 3 end to end: it (1) selected a text-generation task battery satisfying the novelty condition, (2) established a human performance baseline, (3) had both humans and the AI system complete the identical tasks, (4) scored all outputs against the human distribution to compute the superhuman index, and (5) asked a judge panel to identify the AI from the outputs.

5.1 Operationalizing Novelty

The novelty condition was operationalized by drawing prompts on topics and events that post-dated the LLM’s documented training cut-off by at least six months, with no retrieval or web-search tools enabled at generation time. In the terms of the taxonomy in Section 3.2, this procedure establishes *topical* novelty but not *structural* novelty. It must be emphasized that this operationalization is therefore pragmatic rather than definitive: a model with sufficiently broad pre-training may still produce strong outputs on topics structurally similar to training exemplars, even if the specific topic is post-cutoff. Verifying structural or procedural novelty would require controlled training-data exclusions, which was not possible here because the model’s training corpus, although its weights are openly available, is not released in an auditable form. The results are therefore interpreted as evidence of superhuman text-generation performance on a challenging

out-of-training-distribution sample, rather than as a claim of generalized superhuman intelligence.

5.2 Pre-registration and Ethics

The study protocol (hypotheses, sample size, exclusion criteria, codebook, and analysis plan) was pre-registered on the Open Science Framework before data collection. The protocol was reviewed and approved by the authors' institutional ethics review board. All participants provided informed consent, were compensated at or above local prevailing wage, and could withdraw at any time without penalty. No personally identifying information was collected, [37]. All stimulus prompts, AI outputs, human-written responses (with writer consent), participant judgments in anonymized form, the codebook, and analysis scripts will be made available in an open repository upon publication.

5.3 Example Stimuli

To give readers a concrete sense of task difficulty without requiring repository access, one representative prompt per task type is listed below:

- *Essay (current events)*: “In 250–300 words, assess the most significant consequences of the 2025 global semiconductor supply realignment for consumer-electronics pricing.”
- *Scientific-abstract summary*: “Summarize the key claims and evidence of the attached abstract on room-temperature superconductivity (published after the model's training cut-off), in 120–150 words, preserving numerical results.”
- *Story opening*: “Write the opening 150 words of a short story in which a cartographer discovers a city that cannot be plotted on any map.”
- *Arithmetic explanation*: “Explain step by step how to compute the compound interest on \$1,200 at 4.5% annual interest compounded monthly over 7 years, so that a motivated secondary-school student could follow it.”

5.4 Participants and Stimuli

Five hundred participants were recruited via an online platform using stratified sampling to ensure demographic balance across age (18–65), gender, and education level. The pre-registered sample-size calculation targeted the smallest difference from chance considered substantively meaningful, operationalized as a small effect (Cohen's $d = 0.15$). This conservative target was adopted because, at the time of pre-registration, no prior study had reported a comparable RTT-style comparison. For

a one-sample t -test against the 0.50 chance level with power = 0.80 and $\alpha = 0.05$, this required 349 participants; 500 were recruited to allow for attrition and pre-registered exclusion criteria (completion time < 3 minutes or > 90 minutes; response entropy consistent with random clicking). After exclusions, 487 participants were retained. The observed effect ($d = 5.0$) vastly exceeded the conservative pre-registered target, which reinforces rather than undermines the conclusion.

Each participant evaluated 40 stimulus pairs, each pair consisting of one AI-generated and one human-written response to the same prompt. Stimuli were balanced across four task types chosen to span distinct writing modes: short essays on current events (expository argumentation, 10 pairs), summaries of scientific abstracts (informative condensation, 10 pairs), creative story openings (generative narrative, 10 pairs), and step-by-step arithmetic explanations (procedural explanation, 10 pairs). This composition spans four widely used writing genres that place distinct demands on expository, informative, narrative, and procedural competence. Within each task type, prompts were randomly drawn without replacement from a pool of 100 pre-tested prompts. Prompt pools were constructed specifically for this study using topics post-dating the model's documented training cut-off.

AI responses were generated using a publicly available, open-weight 70-billion-parameter transformer-based language model with a documented training cut-off preceding all prompt topics by at least six months. Decoding parameters (temperature, nucleus-sampling threshold, and maximum output length) were fixed across all prompts and are reported in the pre-registration. No retrieval or web-search tools were used at generation time. Human responses were collected from 100 volunteer writers (25 per task type) under identical prompt conditions, without editing for quality, grammar, or style. Writers were compensated at or above prevailing market rates. Response length was matched between conditions within each pair. Specifically, for each pair, the absolute length difference was required to be within 15% of the pair's mean response length; a paired regression of identification accuracy on within-pair length difference showed no significant association ($\beta = 0.02$, $p = 0.63$), confirming that length was not serving as a confounding cue. The 15% tolerance was chosen to balance strict matching with feasibility: human writers produced responses of naturally varying length, and tighter tolerances would have required post-hoc truncation that could itself have introduced bias. Stimulus pairs were presented in randomized order, and within each pair,

the left/right position of human versus AI response was counterbalanced.

5.5 Procedure and Measures

Participants were informed that each pair contained one human and one AI response and were asked to identify the AI output and indicate their confidence on a five-point scale (1 = “very unsure” to 5 = “very sure”). Unlike the general definition and the protein worked example, in which evaluators may be shown metric summaries, judges in this instantiation saw only the paired outputs; the metric summaries were deliberately withheld so that identification could not be a direct readout of the superhuman index, and the grounding of identifications in performance was instead established post hoc through the coding of justifications. No deception was employed. After completing the main task, participants provided written justifications for their judgments on a randomly selected subset of five pairs; these justifications were subjected to qualitative analysis.

The primary outcome was accuracy: proportion of correct AI identifications per participant. The secondary outcome was the superhuman index. For text generation, the metric m was a composite of four sub-indices, each chosen to capture a distinct dimension of output quality:

1. *Lexical richness*, measured by type–token ratio (higher = richer).
2. *Sentence-length variance*, measured by the standard deviation of sentence lengths in tokens (higher = more varied prose).
3. *Grammatical error rate*, measured by an automated grammar checker and negated before standardization so that higher values indicate fewer errors.
4. *Task-appropriate accuracy*, scored differently by task type to remain meaningful across genres: factual accuracy verified against source documents for essays and abstract summaries; arithmetic correctness for arithmetic explanations; and coherence/plausibility for story openings, rated on a 1–5 scale by two independent raters. Inter-rater reliability across the accuracy sub-index was $\kappa = 0.78$ (pooled across task types; per-task values ranged from 0.74 to 0.85).

These four facets were not chosen ad hoc. They were selected to cover four conceptually distinct and largely separable axes along which written text can be evaluated: the lexical axis (how varied and rich the vocabulary is), the syntactic axis (how the prose is structured at the sentence

level), the mechanical axis (how well the text conforms to grammatical convention), and the content axis (how accurate and task-appropriate the substance is). This decomposition follows the familiar separation, in writing assessment, between vocabulary, syntax, mechanics, and content or task fulfillment. Because the axes are meant to be non-redundant rather than interchangeable, the composite is treated as a formative index in the sense set out in Section 3: the four sub-indices jointly constitute a working operationalization of “output quality” rather than serving as parallel reflections of one latent trait, so low intercorrelation among them is expected and is not a deficiency. What the construct does claim is only that a text scoring far above the human distribution on richer vocabulary, more varied sentence structure, fewer grammatical errors, and higher task-appropriate accuracy exhibits superhuman *stylometric* performance; no claim is made that the composite measures a unitary latent “quality” or, still less, intelligence.

Before the combination, each sub-index was sign-aligned (so that higher = higher quality) and z -standardized using the human reference sample’s mean and standard deviation. The composite SI was computed as the unweighted mean of the four standardized sub-indices. Equal weighting was pre-registered as the primary specification because, for a formative index whose facets were selected to be distinct and non-redundant, there was no prior evidence on which to privilege one axis over another, and equal weighting is the conventional default in that situation. To ensure that the substantive conclusion did not depend on either the equal-weighting choice or the aggregation itself, robustness checks using principal-components-based weights and using each sub-index individually are reported in the supplementary materials; these produced SI values in the range [2.8, 3.5], and every sub-index considered on its own also placed the AI system above the pre-specified threshold, so no single facet drives the aggregate result.

These stylometric sub-indices capture the quality and precision of generated text rather than deeper cognitive abilities. This is a deliberate scope restriction of the present instantiation: the text-generation domain does not have a single objectively scorable “right answer” metric comparable to GDT_TS in protein structure prediction or Elo rating in chess. It is therefore important to be explicit about the level at which the claim is pitched, so that the general definition of intelligence adopted in Section 3 and its instantiation here are not conflated. The formal RTT framework targets a fluid-intelligence-weighted construct through the combination of a novelty condition

and superhuman performance. The text-generation study does not measure that construct directly: its sub-indices index output quality and surface-level linguistic form rather than reasoning, memory integration, or strategic planning, and the accuracy sub-index taps only a limited, task-bounded form of performance. What the study establishes, and all that it is claimed to establish, is that a current large language model produces prose that is statistically distinguishable from human writing along objectively measurable dimensions, to a degree that places it well above the human reference distribution and renders it reliably identifiable to human judges. Its purpose is to demonstrate that the five-step RTT protocol can be executed end to end, with a pre-registered battery, a human baseline, an operationalized novelty condition, an SI computation, and a performance-grounded judge panel, even in a domain that lacks a single canonical metric. Accordingly, the text-generation SI is read strictly as evidence of superhuman *stylometric performance* on an out-of-training-distribution sample, never as evidence of superhuman intelligence; intelligence-grade RTT claims are reserved for domains that supply a direct performance metric and admit stringent novelty control, as discussed in Sections 6.4 and 6.

5.6 Results

Participants identified the AI output correctly in roughly 90% of trials ($M = 0.90$, $SD = 0.08$; 95% CI [0.89, 0.91]). A one-sample t -test confirmed that accuracy was far above the chance level of 0.50, $t(486) = 110.3$, $p < 0.001$, $d = 5.0$. Accuracy was relatively uniform across task types (essays: $M = 0.89$; summaries: $M = 0.91$; story openings: $M = 0.87$; arithmetic: $M = 0.93$). A linear mixed-effects model with task type as a fixed effect and participant as a random effect (aggregating each participant's 10 trials per task type into a proportion-correct score, yielding $487 \times 4 = 1,948$ observations) showed no significant task-type effect, $F(3, 1458) = 1.42$, $p = 0.24$.

The composite superhuman index for the AI system was $SI = 3.2$ ($SD = 0.4$), indicating that the AI's stylometric profile was 3.2 standard deviations above the human mean, well above the pre-specified threshold of $\theta = 2.0$ (corresponding to the 97.7th percentile of the human distribution). This confirms that the AI's output quality was objectively superhuman on the measured dimensions.

The highest accuracy was observed for arithmetic explanations (0.93), which may initially seem counterintuitive given that early language models were notably weak at arithmetic. However, modern LLMs now handle elementary arithmetic explanation reliably (even when they still occasionally err on

harder numerical computation); their explanations are typically exhaustively enumerated step by step, which appears to mark them as machine-generated in comparison with the more elliptical explanations humans tend to provide, [38]. The arithmetic result is therefore consistent with the overall pattern: greater thoroughness and procedural completeness serve as cues to AI authorship.

Confidence ratings were positively correlated with accuracy (Pearson $r = 0.41$, $p < 0.001$; Spearman $\rho = 0.39$, $p < 0.001$, used as a robustness check given that confidence was recorded on an ordinal 1–5 scale): participants who correctly identified the AI output tended to be more confident in their judgments. This correlation indicates metacognitive calibration, but is agnostic about whether the underlying cue was performance-related or stylistic; the qualitative analysis reported in the next subsection speaks more directly to that question.

5.7 Qualitative Analysis

Written justifications ($n = 2,435$ responses, 5 per participant) were independently coded by two trained raters using a pre-specified codebook. The codebook contained four categories informed by, but not identical to, the stylometric dimensions: (A) quality of language (grammar, fluency), (B) structural organization, (C) factual precision and coverage, and (D) stylistic naturalness. Inter-rater agreement was satisfactory ($\kappa = 0.76$). Across all coded responses, category (C) (factual precision and coverage) was the most frequently cited cue (43% of responses), followed by category (A) (37%), category (B) (12%), and category (D) (8%). Representative examples include: “*The AI answer covered every sub-point with equal precision, which no human writer would bother to do*” (category C); “*Grammatically flawless throughout, no false starts*” (category A). These patterns corroborate the quantitative finding that AI identification was driven primarily by objectively measurable output quality rather than by idiosyncratic stylistic impressions. The relevant boundary is principled rather than ad hoc: performance cues are the pre-registered, human-baselined stylometric dimensions that enter the SI, whereas surface artifacts are cues orthogonal to those dimensions, such as rendering, markup, or boilerplate formatting. This is consistent with the RTT's requirement that identification be based on superhuman performance evidence rather than on artifacts of that kind. Notably, the two dominant cues, factual precision and coverage (category C) and language quality (category A), map directly onto two of the SI sub-indices (task-appropriate accuracy and the lexical and grammatical quality measures). This alignment is precisely the convergent-validity result the two-step

design is intended to produce: it indicates that the judge panel and the SI are responding to the same underlying performance signal rather than that the judge panel merely restates the SI. Had identification instead been dominated by category (D) (stylistic naturalness) or by formatting cues orthogonal to the measured dimensions, the two components would have diverged, and the case that the SI captures a humanly recognizable form of superiority would have been correspondingly weaker.

This high accuracy contrasts with earlier research, in which humans struggled to distinguish AI-generated text from human-written text. In one study of college essays, instructors correctly identified machine-generated writing only about 70% of the time and students about 60%, [39]; in another, recognition of machine-generated short social-media posts fell near the chance level, [40]. A more recent evaluation likewise found that human raters and automated detectors alike identified AI-generated written content with markedly variable accuracy across formats, [41]. The higher detection rate observed here most plausibly reflects the AI's objectively higher stylometric performance ($SI = 3.2$) relative to the less polished outputs evaluated in earlier work.

5.8 Relation to Prior Work

Table 4 summarizes detection performance reported in selected studies. Studies focusing on deception and disinformation often report rates barely above chance, [40]. Stylometric classifiers can reach 98 to 100% accuracy when trained on labeled corpora, [42], and more recent stylometric analyses continue to separate machine from human authorship with high accuracy even in settings where human readers themselves perform poorly, [43], [44]. The present experiment falls between these extremes. Crucially, unlike automated classifiers, the present study uses human judges, a core component of the RTT protocol, and grounds identification in a pre-computed superhuman index rather than relying solely on subjective impressions.

Overall, the experiment instantiates the five-step RTT protocol end to end in the domain of text generation, with the novelty condition operationalized through post-cut-off prompts. The stimuli were length-balanced within a 15% tolerance, randomly assigned, and counterbalanced for position, and human baselines and pre-registration were used throughout. The composite superhuman index (approximately 3.2 standard deviations above the human mean) established objectively superhuman stylometric performance, and the 90% identification accuracy was far above chance, with a Cohen's d of 5.0. Systematic coding of 2,435 participant

Table 4. Representative studies on human or algorithmic ability to detect AI-generated text. Detection accuracies refer to the proportion of AI outputs correctly identified as machine-generated.

Study	Domain	Size	Detect.
Synthetic posts hard to distinguish, [40]	Synthetic vs. organic posts	697	50–60%
Instructors 70%, students 60%, [39]	College essays	285	60–70%
Multi-modal detectors, [45]	Stylometric Synthetic media	n/a	80–92%
Random-forest classifier, [42]	Stylometric Academic papers	n/a	98–100%
This study: stylometric SI plus judge panel	Stylometric SI Essays, summaries, stories	500	90%

Source: created by the authors.

justifications then pinpointed factual precision and language quality as the dominant identification cues, corroborating the SI-based evidence and confirming that identification was grounded in performance rather than in surface style.

6 Discussion and Outlook

The RTT repositions the evaluation of machine intelligence within a broader story about human-machine symbiosis. The study in Section 5 shows that the protocol is operationally tractable: the superhuman index can be computed from objective measures, a human baseline can be established under controlled conditions, and human judges reliably identify superhuman AI outputs when the identification is anchored in quantitative evidence rather than unaided intuition. The rest of this section traces out what follows from this: for trust and education, for the delicate relationship between creativity and intelligence, and for ethics and governance, before stating the limitations of what has actually been shown, and looking ahead to a broader family of beyond-imitation evaluations.

6.1 Trust and Education

Trustworthy AI requires that users understand both the strengths and limitations of systems. By making superhuman performance explicit through the superhuman index, the RTT can foster calibrated trust: users know when to rely on a machine and when to defer to human judgment. This aim connects to a growing empirical literature on appropriate reliance, which finds that whether users calibrate their reliance to a system's actual competence depends on the psychological and informational context in which performance is disclosed, [46]. Educational curricula could incorporate RTT-style assessments, training students to collaborate with

AI partners. Instead of forbidding the use of large language models in homework, teachers could ask students to critique and improve AI-generated solutions, emphasizing meta-cognitive skills such as identifying where a superhuman system may nonetheless be wrong or unsuitable, [47]. Such a shift is increasingly urgent given evidence that generative models materially affect the integrity of conventional written assessments in higher education, [48].

6.2 Achievement and Creativity

A recurrent question raised by the RTT concerns how the constructs it evaluates relate to one another. In the psychometric tradition, creativity and intelligence are related but distinct, [28]: creativity is concerned with the generation of novel, appropriate, and valuable ideas, whereas intelligence reflects general reasoning and adaptive capacity. The RTT is not a direct measure of creativity. It measures superhuman performance on pre-specified, objectively scorable tasks. When a task domain involves creative output, story generation, say, the RTT can still be applied, but only if an objective scoring criterion (such as expert-rated novelty or coherence) is supplied; without such a criterion, the RTT does not apply to that domain at all.

The achievement–intelligence distinction from Section 3 matters here too. A generative model that recombines training data in statistically novel ways may score well on novelty metrics without exhibiting intelligence in the psychometric sense, because it is not adapting to genuinely unseen problem structures, [49]. The RTT’s novelty condition is designed to mitigate precisely this conflation, but researchers applying the RTT to creative domains should verify explicitly that the novelty condition is met, and should resist the temptation to equate high achievement scores with high general intelligence. AI systems are already producing artwork, music, and literature that are in many respects indistinguishable from human output, and controlled studies now report that generative language models match or exceed human samples on standardized divergent-thinking tasks, [50]; the RTT can certify superhuman achievement in these domains while remaining agnostic about whether the underlying process counts as creativity or intelligence in the full psychometric sense. What the RTT does encourage is a co-creative division of labor in which machines propose novel outputs and humans curate, refine, and situate them, the same pattern in which AlphaGo’s moves inspired professional Go players to rethink their strategies, [11].

6.3 Decision-Making and Ethics

AI systems are already involved in high-stakes domains, including medicine, finance, and law. Superhuman performance does not automatically guarantee ethical or beneficial outcomes. Decision-making requires value judgments, fairness considerations, and accountability structures that the RTT’s performance metrics do not capture. An AI system could attain a superhuman index of 4.0 in medical diagnosis yet still exhibit systematic bias against minority groups, a risk that recent frameworks for bias evaluation in clinical language models make concrete and measurable, [51]. The RTT must therefore be integrated with ethical frameworks such as alignment research, fairness audits, [23], [52], and transparent documentation of training data and model behaviors. Governance mechanisms could include licensing regimes for AI systems that pass the RTT in high-stakes domains, mandatory disclosure of superhuman indices, and structured red-teaming to identify failure modes not present in the benchmark task battery.

6.4 Limitations

Several limitations of the present work should be acknowledged.

Stylometric rather than performance-based metrics in the text-generation study. The stylometric sub-indices used in Section 5 (lexical richness, sentence-length variance, grammatical error rate, and task-appropriate accuracy) capture output quality and, in the accuracy sub-index, a limited form of task performance. They do not directly tap deeper reasoning, memory integration, or strategic planning. This gap between the fluid-intelligence-weighted construct that the RTT is designed to target and the stylometric quantities the text-generation study actually measures creates a genuine risk of equivocation, in which superhuman stylometric performance could be mistaken for superhuman intelligence. The paper manages this risk not by claiming the gap is closed but by confining the text-generation claim to its proper level: the study is reported as evidence of superhuman stylometric performance and as a demonstration that the protocol is operationally tractable, and no inference to generalized intelligence is drawn from it. Domains such as protein structure prediction, chess, or mathematical theorem proving offer more direct performance metrics; future RTT instantiations in such domains would provide less ambiguous evidence about superhuman capability, and only a battery spanning several such domains could begin to support a claim framed in terms of intelligence rather than domain-specific performance.

Sample-based rather than norm-referenced

baseline. The human reference samples used in the illustrative protein example ($|S_H| = 30$) and in the text-generation experiment ($|S_H| = 100$ writers) are sample estimates of μ_H and σ_H rather than standardization-grade psychometric norms. Confidence intervals around SI should therefore be reported in any serious deployment, and future RTT development should include larger, demographically stratified reference samples.

Novelty condition only pragmatically verified. The post-cutoff prompt strategy ensures topical novelty but does not guarantee structural, combinatorial, or procedural novelty relative to the training distribution, in the terms of the taxonomy in Section 3.2. A model may perform well on a genuinely out-of-training-distribution topic by generalizing from structurally similar exemplars. Stronger novelty verification, at the structural or procedural level, would require controlled training-data exclusions, typically only possible with open-weight models and curated training data, and the taxonomy is offered so that future studies can state which level of novelty they enforce and by which method rather than treating novelty as a single undifferentiated property.

Single-domain empirical instantiation. The only full empirical instantiation reported here is in text generation; the protein example is illustrative. Multi-domain RTT validation remains an important direction for future work. Claims about general superhuman capability should not be made based on a single-domain instantiation.

Reification risk of the composite SI. The composite superhuman index combines four disparate sub-indices; it should be interpreted as a summary statistic and not as a measurement of a single latent construct. Robustness checks with alternative weightings are therefore essential (and are reported for the text-generation study).

Verification that judges use performance evidence is indirect. Even with structured qualitative analysis of justifications, the inference that judges relied on performance rather than surface artifacts depends on the fidelity of self-report. Future studies could strengthen this inference by experimentally manipulating surface features while holding performance constant (or vice versa) to establish which cues are actually driving identification.

6.5 Beyond-Imitation Evaluation

Looking ahead, the RTT is best seen as one component of a broader landscape of beyond-imitation evaluations. Multi-agent environments could assess cooperation and negotiation, embodied tasks could test physical capability, and open-ended research challenges

might invite machines to discover novel scientific laws, all of them amenable to RTT operationalization, as long as a human baseline and a defensible novelty condition can be established. One promising future benchmark would ask an AI to propose experiments in a new scientific domain and interpret the results, with the superhuman index computed against a panel of domain experts. Another might test long-term planning in dynamic social settings. The broader point is simple: by emphasizing a diverse set of abilities rather than a single conversational skill, the research community stands a far better chance of building AI systems that are robust, beneficial, and trustworthy.

In sum, the study in Section 5 shows that the RTT is operationally tractable, and the trust and educational applications sketched above follow naturally once the SI is reported explicitly. Creativity and intelligence remain distinct psychometric constructs, and the RTT applies to creative domains only when an objective scoring criterion is available and the novelty condition is enforced. High achievement scores should not be confused with general intelligence, because the novelty condition is necessary but not sufficient, and multi-domain instantiations will be required before any such claim is made. The key limitations of this work (stylometric versus performance metrics, sample-based baselines, pragmatic novelty verification, a single empirical domain, and the reification risk of a composite SI) are laid out explicitly in Section 6.4. Equally, any serious use of the RTT must be accompanied by ethical and governance frameworks: fairness audits, disclosure mandates, and structured red-teaming, not replaced by a single quantitative score.

7 Conclusion

Alan Turing's imitation game gave the AI community an elegant and influential behavioral criterion, but it was never a psychometric instrument and has had little influence on the scientific study of intelligence. In the twenty-first century, machines routinely surpass human abilities in specific domains, and the central question for evaluation has shifted from imitation to augmentation. This paper has argued for a Reversed Turing Test (RTT), in which machines are recognized precisely because their performance is measurably superhuman, quantified through a standardized superhuman index in standard-deviation units relative to a human reference sample.

Three clarifications keep the RTT's claims honest. First, the RTT distinguishes intelligence from domain-specific achievement: its novelty condition requires evaluated tasks to be genuinely unfamiliar to the system, so that the test supports inferences about adaptive capacity rather than

about memorized performance. Second, the RTT replaces purely subjective human judgments with a two-step criterion, an objective performance metric combined with a judge-identification test, an arrangement loosely analogous to (though not equivalent to) the norm-referenced scoring used in psychometric assessment of cognitive ability. Third, like any behavioral test, the RTT takes no position on philosophical questions about consciousness or understanding; its scope is explicitly limited to measuring and certifying superhuman task performance.

Concretely, the paper contributed a formal definition of the RTT (Section 3), a taxonomy of novelty types with corresponding verification methods, an illustrative worked example in protein structure prediction, a comparison with major alternative benchmarks (Section 4), and a pre-registered empirical instantiation in the domain of text generation (Section 5). In that study, a sample of 500 participants (487 retained after pre-registered exclusions) identified AI-generated outputs with 90% accuracy; identifications were backed by a composite superhuman index of roughly 3.2 SD above the human mean and corroborated by a systematic qualitative analysis of 2,435 participant justifications. Consistent with the scope stated throughout, this result is interpreted as evidence of superhuman stylometric performance and of the protocol's operational tractability rather than as a claim of superhuman intelligence. The full RTT protocol (task pre-registration, human baseline, novelty condition, SI computation, and judge panel) was followed throughout, providing a reproducible template for future studies.

The present work has also been deliberate about limitations. The text-generation study relied on stylometric sub-indices, the human baselines were sample-based rather than norm-referenced, the novelty condition was only pragmatically verified, and the empirical instantiation was confined to a single domain. These are real caveats, listed openly (Section 6.4).

Even with those caveats, the RTT points in a useful direction. It complements existing benchmarks such as the Lovelace Test and the Winograd Schema Challenge by focusing on augmentation rather than imitation, and by producing a quantitative rather than a binary verdict. Making the RTT useful in practice will require progress on several open problems: defining and enforcing the novelty condition across diverse domains, extending the SI framework for multi-dimensional and creative output, standardizing human reference samples at psychometric scale, and balancing economic incentives for automation against the broader societal good. By treating superhuman

performance as a measurable, domain-specific, and context-dependent property rather than a binary condition, the RTT offers a rigorous and practically actionable lens for understanding and for governing the next era of machine intelligence.

References:

- [1] A. M. Turing, "Computing Machinery and Intelligence," *Mind*, vol. 59, no. 236, pp. 433–460, 1950. [Online]. Available: <https://doi.org/10.1093/mind/LIX.236.433>
- [2] J. R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences*, vol. 3, no. 3, pp. 417–424, 1980. [Online]. Available: <https://doi.org/10.1017/S0140525X00005756>
- [3] S. M. Shieber, "Lessons from a restricted Turing test," *Communications of the ACM*, vol. 37, no. 6, pp. 70–78, 1994. [Online]. Available: <https://doi.org/10.1145/175208.175217>
- [4] Y. Chang, X. Wang, J. Wang, Y. Wu, L. Yang, K. Zhu, H. Chen, X. Yi, C. Wang, Y. Wang, W. Ye, Y. Zhang, Y. Chang, P. S. Yu, Q. Yang, and X. Xie, "A Survey on Evaluation of Large Language Models," *ACM Transactions on Intelligent Systems and Technology*, vol. 15, no. 3, pp. 1–45, 2024. [Online]. Available: <https://doi.org/10.1145/3641289>
- [5] D. Silver, A. Huang, C. J. Maddison, A. Guez, L. Sifre, G. Van den Driessche, J. Schrittwieser, I. Antonoglou, V. Panneershelvam, M. Lanctot, S. Dieleman, D. Grewe, J. Nham, N. Kalchbrenner, I. Sutskever, T. Lillicrap, M. Leach, K. Kavukcuoglu, T. Graepel, and D. Hassabis, "Mastering the game of Go with deep neural networks and tree search," *Nature*, vol. 529, no. 7587, pp. 484–489, 2016. [Online]. Available: <https://doi.org/10.1038/nature16961>
- [6] D. Silver, J. Schrittwieser, K. Simonyan, I. Antonoglou, A. Huang, A. Guez, T. Hubert, L. Baker, M. Lai, A. Bolton, Y. Chen, T. Lillicrap, F. Hui, L. Sifre, G. Van den Driessche, T. Graepel, and D. Hassabis, "Mastering the game of Go without human knowledge," *Nature*, vol. 550, pp. 354–359, 2017. [Online]. Available: <https://doi.org/10.1038/nature24270>
- [7] D. Silver, T. Hubert, J. Schrittwieser, I. Antonoglou, M. Lai, A. Guez, M. Lanctot, L. Sifre, D. Kumaran, T. Graepel, T. Lillicrap, K. Simonyan, and D. Hassabis, "A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play,"

Science, vol. 362, no. 6419, pp. 1140–1144, 2018. [Online]. Available: <https://doi.org/10.1126/science.aar6404>

- [8] C. Berner, G. Brockman, B. Chan, V. Cheung, P. Debiak, C. Dennison, D. Farhi, Q. Fischer, S. Hashme, C. Hesse, R. Józefowicz, S. Gray, C. Olsson, J. Pachocki, M. Petrov, H. P. d. Pinto, J. Raiman, T. Salimans, J. Schlatter, J. Schneider, S. Sidor, I. Sutskever, J. Tang, F. Wolski, and S. Zhang, “Dota 2 with Large Scale Deep Reinforcement Learning,” *arXiv preprint arXiv:1912.06680*, 2019. [Online]. Available: <https://doi.org/10.48550/arXiv.1912.06680>
- [9] J. Jumper, R. Evans, A. Pritzel, T. Green, M. Figurnov, O. Ronneberger, K. Tunyasuvunakool, R. Bates, A. Židek, A. Potapenko, A. Bridgland, C. Meyer, S. A. A. Kohl, A. J. Ballard, A. Cowie, B. Romera-Paredes, S. Nikolov, R. Jain, J. Adler, T. Back, S. Petersen, D. Reiman, E. Clancy, M. Zielinski, M. Steinegger, M. Pacholska, T. Berghammer, S. Bodenstein, D. Silver, O. Vinyals, A. W. Senior, K. Kavukcuoglu, P. Kohli, and D. Hassabis, “Highly accurate protein structure prediction with AlphaFold,” *Nature*, vol. 596, pp. 583–589, 2021. [Online]. Available: <https://doi.org/10.1038/s41586-021-03819-2>
- [10] J. Abramson, J. Adler, J. Dunger, R. Evans, T. Green, A. Pritzel, O. Ronneberger, L. Willmore, A. J. Ballard, J. Bambrick, S. W. Bodenstein, D. A. Evans, C.-C. Hung, M. O’Neill, D. Reiman, K. Tunyasuvunakool, Z. Wu, A. Žemgulytė, E. Arvaniti, C. Beattie, O. Bertolli, A. Bridgland, A. Cherepanov, M. Congreve, A. I. Cowen-Rivers, A. Cowie, M. Figurnov, F. B. Fuchs, H. Gladman, R. Jain, Y. A. Khan, C. M. R. Low, K. Perlin, A. Potapenko, P. Savy, S. Singh, A. Stecula, A. Thillaisundaram, C. Tong, S. Yakneen, E. D. Zhong, M. Zielinski, A. Židek, V. Bapst, P. Kohli, M. Jaderberg, D. Hassabis, and J. M. Jumper, “Accurate structure prediction of biomolecular interactions with AlphaFold 3,” *Nature*, vol. 630, pp. 493–500, 2024. [Online]. Available: <https://doi.org/10.1038/s41586-024-07487-w>
- [11] M. Shin, J. Kim, B. van Opheusden, and T. L. Griffiths, “Superhuman artificial intelligence can improve human decision-making by increasing novelty,” *Proceedings of the National Academy of Sciences*, vol. 120, no. 12, p. e2214840120, 2023. [Online]. Available: <https://doi.org/10.1073/pnas.2214840120>
- [12] M. Pellert, C. M. Lechner, C. Wagner, B. Rammstedt, and M. Strohmaier, “AI Psychometrics: Assessing the Psychological Profiles of Large Language Models Through Psychometric Inventories,” *Perspectives on Psychological Science*, vol. 19, no. 5, pp. 808–826, 2024. [Online]. Available: <https://doi.org/10.1177/17456916231214460>
- [13] A. L. Samuel, “Some Studies in Machine Learning Using the Game of Checkers,” *IBM Journal of Research and Development*, vol. 3, no. 3, pp. 210–229, 1959. [Online]. Available: <https://doi.org/10.1147/rd.33.0210>
- [14] F. Rosenblatt, “The Perceptron: A Probabilistic Model for Information Storage and Organization in the Brain,” *Psychological Review*, vol. 65, no. 6, pp. 386–408, 1958. [Online]. Available: <https://doi.org/10.1037/h0042519>
- [15] A. Vaswani, N. Shazeer, N. Parmar, J. Uszkoreit, L. Jones, A. N. Gomez, Ł. Kaiser, and I. Polosukhin, “Attention Is All You Need,” 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.1706.03762>
- [16] J. Devlin, M.-W. Chang, K. Lee, and K. Toutanova, “BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding,” in *Arxiv: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, 2019. [Online]. Available: <https://doi.org/10.18653/v1/N19-1423>
- [17] T. B. Brown, B. Mann, N. Ryder, M. Subbiah, J. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. M. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, and D. Amodei, “Language Models Are Few-Shot Learners,” in *Arxiv: Advances in Neural Information Processing Systems 33 (NeurIPS 2020)*, 2020. [Online]. Available: <https://doi.org/10.48550/arXiv.2005.14165>
- [18] L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. L. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman,

- J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. Christiano, J. Leike, and R. Lowe, "Training language models to follow instructions with human feedback," in *Arxiv: Advances in Neural Information Processing Systems 35 (NeurIPS 2022)*, 2022. [Online]. Available: <https://doi.org/10.48550/arXiv.2203.02155>
- [19] V. Mnih, K. Kavukcuoglu, D. Silver, A. A. Rusu, J. Veness, M. G. Bellemare, A. Graves, M. Riedmiller, A. K. Fidjeland, G. Ostrovski, S. Petersen, C. Beattie, A. Sadik, I. Antonoglou, H. King, D. Kumaran, D. Wierstra, S. Legg, and D. Hassabis, "Human-level control through deep reinforcement learning," *Nature*, vol. 518, pp. 529–533, 2015. [Online]. Available: <https://doi.org/10.1038/nature14236>
- [20] O. Vinyals, I. Babuschkin, W. M. Czarnecki, M. Mathieu, A. Dudzik, J. Chung, D. H. Choi, R. Powell, T. Ewalds, P. Georgiev, J. Oh, D. Horgan, M. Kroiss, I. Danihelka, A. Huang, L. Sifre, T. Cai, J. P. Agapiou, M. Jaderberg, A. S. Vezhnevets, R. Leblond, T. Pohlen, V. Dalibard, D. Budden, Y. Sulsky, J. Molloy, T. L. Paine, C. Gulcehre, Z. Wang, T. Pfaff, Y. Wu, R. Ring, D. Yogatama, D. Wünsch, K. McKinney, O. Smith, T. Schaul, T. Lillicrap, K. Kavukcuoglu, D. Hassabis, C. Apps, and D. Silver, "Grandmaster level in StarCraft II using multi-agent reinforcement learning," *Nature*, vol. 575, pp. 350–354, 2019. [Online]. Available: <https://doi.org/10.1038/s41586-019-1724-z>
- [21] R. Bommasani, D. A. Hudson, E. Adeli, R. Altman, S. Arora, S. von Arx, M. S. Bernstein, J. Bohg, A. Bosselut, E. Brunskill, E. Brynjolfsson, S. Buch, D. Card, R. Castellon, N. Chatterji, A. Chen, K. Creel, J. Q. Davis, D. Demszky, C. Donahue, M. Doumbouya, E. Durmus, S. Ermon, J. Etchemendy, K. Ethayarajh, L. Fei-Fei, C. Finn, T. Gale, L. Gillespie, K. Goel, N. Goodman, S. Grossman, N. Guha, T. Hashimoto, P. Henderson, J. Hewitt, D. E. Ho, J. Hong, K. Hsu, J. Huang, T. Icard, S. Jain, D. Jurafsky, P. Kalluri, S. Karamcheti, G. Keeling, F. Khani, O. Khattab, P. W. Koh, M. Krass, R. Krishna, R. Kuditipudi, A. Kumar, F. Ladhak, M. Lee, T. Lee, J. Leskovec, I. Levent, X. L. Li, X. Li, T. Ma, A. Malik, C. D. Manning, S. Mirchandani, E. Mitchell, Z. Muniyikwa, S. Nair, A. Narayan, D. Narayanan, B. Newman, A. Nie, J. C. Niebles, H. Nilforoshan, J. Nyarko, G. Ogut, L. Orr, I. Papadimitriou, J. S. Park, C. Piech, E. Portelance, C. Potts, A. Raghunathan, R. Reich, H. Ren, F. Rong, Y. Roohani, C. Ruiz, J. Ryan, C. Ré, D. Sadigh, S. Sagawa, K. Santhanam, A. Shih, K. Srinivasan, A. Tamkin, R. Taori, A. W. Thomas, F. Tramèr, R. E. Wang, W. Wang, B. Wu, J. Wu, Y. Wu, S. M. Xie, M. Yasunaga, J. You, M. Zaharia, M. Zhang, T. Zhang, X. Zhang, Y. Zhang, L. Zheng, K. Zhou, and P. Liang, "On the Opportunities and Risks of Foundation Models," *arXiv preprint arXiv:2108.07258*, 2021. [Online]. Available: <https://doi.org/10.48550/arXiv.2108.07258>
- [22] S. Bubeck, V. Chandrasekaran, R. Eldan, J. Gehrke, E. Horvitz, E. Kamar, P. Lee, Y. T. Lee, Y. Li, S. Lundberg, H. Nori, H. Palangi, M. T. Ribeiro, and Y. Zhang, "Sparks of Artificial General Intelligence: Early experiments with GPT-4," *arXiv preprint arXiv:2303.12712*, 2023. [Online]. Available: <https://doi.org/10.48550/arXiv.2303.12712>
- [23] E. M. Bender, T. Gebru, A. McMillan-Major, and S. Shmitchell, "On the Dangers of Stochastic Parrots: Can Language Models Be Too Big?" in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FACt '21)*, Virtual Event, Canada, 2021, pp. 610–623. [Online]. Available: <https://doi.org/10.1145/3442188.3445922>
- [24] L. Harding and L. Barden, "Deep Blue win a giant step for computerkind," *The Guardian*, "From the archive" series; originally published May 12, 1997, [Access Date: August 12, 2026]. [Online]. Available: <https://www.theguardian.com/theguardian/2011/may/12/deep-blue-beats-kasparov-1997>
- [25] History of Information, "IBM's Watson Question Answering System Defeats Humans at Jeopardy!" Jeremy Norman's HistoryofInformation.com, 2011, [Access Date: August 12, 2026]. [Online]. Available: <https://www.historyofinformation.com/detail.php?id=2829>
- [26] I. Price, A. Sanchez-Gonzalez, F. Alet, T. R. Andersson, A. El-Kadi, D. Masters, T. Ewalds, J. Stott, S. Mohamed, P. Battaglia, R. Lam, and M. Willson, "Probabilistic weather forecasting with machine learning," *Nature*, vol. 637, pp. 84–90, 2025. [Online]. Available: <https://doi.org/10.1038/s41586-024-08252-9>
- [27] K. Warwick and H. Shah, "Passing the Turing Test Does Not Mean the End of

- Humanity,” *Cognitive Computation*, vol. 8, pp. 409–419, 2016. [Online]. Available: <https://doi.org/10.1007/s12559-015-9372-6>
- [28] S. Bringsjord, P. Bello, and D. Ferrucci, “Creativity, the Turing Test, and the (Better) Lovelace Test,” *Minds and Machines*, vol. 11, no. 1, pp. 3–27, 2001. [Online]. Available: <https://doi.org/10.1023/A:1011206622741>
- [29] L. Morgenstern, E. Davis, and C. L. Ortiz, Jr., “Planning, Executing, and Evaluating the Winograd Schema Challenge,” *AI Magazine*, vol. 37, no. 1, pp. 50–54, 2016. [Online]. Available: <https://doi.org/10.1609/aimag.v37i1.2639>
- [30] E. Brynjolfsson, “The Turing Trap: The Promise & Peril of Human-Like Artificial Intelligence,” *Daedalus*, vol. 151, no. 2, pp. 272–287, 2022. [Online]. Available: https://doi.org/10.1162/daed_a_01915
- [31] Y. LeCun, Y. Bengio, and G. Hinton, “Deep learning,” *Nature*, vol. 521, pp. 436–444, 2015. [Online]. Available: <https://doi.org/10.1038/nature14539>
- [32] E. Davis and G. Marcus, “Commonsense Reasoning and Commonsense Knowledge in Artificial Intelligence,” *Communications of the ACM*, vol. 58, no. 9, pp. 92–103, 2015. [Online]. Available: <https://doi.org/10.1145/2701413>
- [33] B. M. Lake, T. D. Ullman, J. B. Tenenbaum, and S. J. Gershman, “Building machines that learn and think like people,” *Behavioral and Brain Sciences*, vol. 40, p. e253, 2017. [Online]. Available: <https://doi.org/10.1017/S0140525X16001837>
- [34] M. Leon, N. M. Sanchez, Z. G. Valdivia, and R. B. Perez, “Concept Maps Combined with Case-Based Reasoning in Order to Elaborate Intelligent Teaching/Learning Systems,” in *Seventh International Conference on Intelligent Systems Design and Applications (ISDA 2007)*. Rio de Janeiro, Brazil: IEEE, Oct. 2007, pp. 205–210. [Online]. Available: <https://doi.org/10.1109/ISDA.2007.33>
- [35] T. Miller, “Explanation in Artificial Intelligence: Insights from the Social Sciences,” *Artificial Intelligence*, vol. 267, pp. 1–38, 2019. [Online]. Available: <https://doi.org/10.1016/j.artint.2018.07.007>
- [36] M. Ebers, “Truly Risk-Based Regulation of Artificial Intelligence: How to Implement the EU’s AI Act,” *European Journal of Risk Regulation*, vol. 16, no. 2, pp. 684–703, 2025. [Online]. Available: <https://doi.org/10.1017/err.2024.78>
- [37] M. Leon, “Lifecycle-Based Governance to Build Reliable Ethical AI Systems,” *Systems Research and Behavioral Science*, vol. 43, no. 3, pp. 1116–1132, 2026. [Online]. Available: <https://doi.org/10.1002/sres.70014>
- [38] M. Leon, “Investing in AI Interpretability, Control, and Robustness,” *Algorithms*, vol. 19, no. 2, p. 136, 2026. [Online]. Available: <https://doi.org/10.3390/a19020136>
- [39] T. Waltzer, C. Pilegard, and G. D. Heyman, “Can you spot the bot? Identifying AI-generated writing in college essays,” *International Journal for Educational Integrity*, vol. 20, no. 1, p. 11, 2024. [Online]. Available: <https://doi.org/10.1007/s40979-024-00158-3>
- [40] G. Spitale, N. Biller-Andorno, and F. Germani, “AI model GPT-3 (dis)informs us better than humans,” *Science Advances*, vol. 9, no. 26, p. eadh1850, 2023. [Online]. Available: <https://doi.org/10.1126/sciadv.adh1850>
- [41] A. Cheng, Y. Lin, G. Reedy, C. Joseph, S. Wirkowski, V. Mallette, V. Nagesh, D. Krieser, and A. Calhoun, “Ability of AI detection tools and humans to accurately identify different forms of AI-generated written content,” *Advances in Simulation*, vol. 10, no. 1, p. 66, 2025. [Online]. Available: <https://doi.org/10.1186/s41077-025-00396-6>
- [42] W. Zaitsu and M. Jin, “Distinguishing ChatGPT(-3.5, -4)-generated and human-written papers through Japanese stylometric analysis,” *PLOS ONE*, vol. 18, no. 8, p. e0288453, 2023. [Online]. Available: <https://doi.org/10.1371/journal.pone.0288453>
- [43] W. Zaitsu, M. Jin, S. Ishihara, S. Tsuge, and M. Inaba, “Stylometry can reveal artificial intelligence authorship, but humans struggle: A comparison of human and seven large language models in Japanese,” *PLOS ONE*, vol. 20, no. 10, p. e0335369, 2025. [Online]. Available: <https://doi.org/10.1371/journal.pone.0335369>
- [44] R. Chen, S. Xiong, J. He, and G. J. Ross, “Stylometric detection of AI-generated texts: evidence from human and machine-written essays,” *Digital Scholarship in the Humanities*, 2026, advance access article fqag064. [Online]. Available: <https://doi.org/10.1093/llc/fqag064>

- [45] D. Ghiurău and D. E. Popescu, "Distinguishing Reality from AI: Approaches for Detecting Synthetic Content," *Computers*, vol. 14, no. 1, p. 1, 2025. [Online]. Available: <https://doi.org/10.3390/computers14010001>
- [46] A. Küper and N. Krämer, "Psychological Traits and Appropriate Reliance: Factors Shaping Trust in AI," *International Journal of Human-Computer Interaction*, pp. 4115–4131, 2024. [Online]. Available: <https://doi.org/10.1080/10447318.2024.2348216>
- [47] M. Leon, "Generative AI as a General-Purpose Technology: Foundations, Applications, and Labor Market Implications Through 2030," *Big Data and Cognitive Computing*, vol. 10, no. 3, p. 69, 2026. [Online]. Available: <https://doi.org/10.3390/bdcc10030069>
- [48] A. K. Kofinas, C. H.-H. Tsay, and D. Pike, "The impact of generative AI on academic integrity of authentic assessments within a higher education context," *British Journal of Educational Technology*, vol. 56, no. 6, pp. 2522–2549, 2025. [Online]. Available: <https://doi.org/10.1111/bjet.13585>
- [49] M. Leon, "Generative Artificial Intelligence and Prompt Engineering: A Comprehensive Guide to Models, Methods, and Best Practices," *Advances in Science, Technology and Engineering Systems Journal*, vol. 10, no. 2, pp. 1–11, 2025. [Online]. Available: <https://doi.org/10.25046/aj100201>
- [50] K. F. Hubert, K. N. Awa, and D. L. Zabelina, "The current state of artificial intelligence generative language models is more creative than humans on divergent thinking tasks," *Scientific Reports*, vol. 14, no. 1, p. 3440, 2024. [Online]. Available: <https://doi.org/10.1038/s41598-024-53303-w>
- [51] T. Templin, S. Fort, P. Padmanabham, P. Seshadri, R. Rimal, J. Oliva, K. H. Lich, S. Sylvia, and N. Sinnott-Armstrong, "Framework for bias evaluation in large language models in healthcare settings," *npj Digital Medicine*, vol. 8, no. 1, p. 414, 2025. [Online]. Available: <https://doi.org/10.1038/s41746-025-01786-w>
- [52] B. Hutchinson, A. Smart, A. Hanna, R. Denton, C. Greer, O. Kjartansson, P. Barnes, and M. Mitchell, "Towards Accountability for Machine Learning Datasets: Practices from Software Engineering and Infrastructure," in

Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency (FAccT '21), Virtual Event, Canada, 2021, pp. 560–575. [Online]. Available: <https://doi.org/10.1145/3442188.3445918>

Contribution of Individual Authors to the Creation of a Scientific Article (Ghostwriting Policy)

The author contributed in the present research, at all stages from the formulation of the problem to the final findings and solution.

Sources of Funding for Research Presented in a Scientific Article or Scientific Article Itself

No funding was received for conducting this study.

Conflict of Interest

The author has no conflict of interest to declare that is relevant to the content of this article.

Creative Commons Attribution License 4.0 (Attribution 4.0 International, CC BY 4.0)

This article is published under the terms of the Creative Commons Attribution License 4.0

https://creativecommons.org/licenses/by/4.0/deed.en_US

APPENDIX

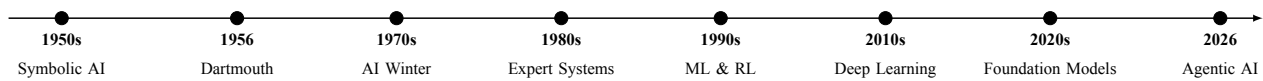


Fig. 1: Evolution of machine intelligence. From symbolic systems in the 1950s through connectionist models and reinforcement learning, the 2010s witnessed significant breakthroughs in deep learning, the 2020s saw the rise of foundation models and multimodal systems, and by 2026 the emergence of agentic AI that plans and acts autonomously.

Source: created by the authors.