The Quest for True Machine Intelligence: An Overview of Progress and Pitfalls

Sachit Ramesha Gowda
sachitramesha.ai23@rvce.edu.in

Dhruv Patankar
dhruvpatankar.ai23@rvce.edu.in

Mihir Arya
mihirarya.cy23@rvce.edu.in

August 30, 2025

Abstract

This paper examines the theoretical and empirical foundations of achieving true machine intelligence through an interdisciplinary approach incorporating neurological, philosophical, and computational perspectives. We analyze historical breakthroughs and ongoing debates in artificial intelligence research to establish a working definition of "true" intelligence and develop a framework for testing genuine machine cognition. The investigation evaluates whether authentic machine intelligence, characterized by reasoning, understanding, and adaptability rather than statistical pattern matching, is computationally feasible. In addition, we explore the relationship between intelligence and consciousness to determine whether intelligence constitutes a prerequisite for conscious experience. Our analysis suggests that progress toward genuine machine intelligence requires grounding AI development in biological principles while integrating symbolic reasoning and cognitive architectures. Ultimately, the path to true machine intelligence lies not in scaling models blindly or by training said models on more data, but in rethinking how we build intelligent machine architectures.

1 Introduction

The pursuit of genuine machine intelligence is one of the main challenges of AI. Although today's systems excel in narrow tasks, they show deep limitations that suggest architectural flaws rather than technical gaps. Ada Lovelace highlighted the divide between following instructions and true creativity, a boundary that still shapes modern AI. Turing shifted the debate to behavioral equivalence with the Turing test, focusing on mimicry rather than real understanding.

This practical approach led to great progress, but models like GPT-4, Claude, and Gemini still hallucinate, fail to reason, and struggle in unfamiliar situations. Scaling alone appears to be hitting its limits. Public discourse often mistakes advanced pattern recognition for actual cognition, obscuring deeper questions about intelligence.

We argue that real machine intelligence requires moving away from pure scaling. Instead, it should involve neurosymbolic architectures, grounded cognition, and biologically inspired design. This paper explores the philosophical roots of intelligence, critiques current architectures, and proposes neuro-symbolic AI as a more sustainable path that prioritizes understanding over imitation.

2 The Philosophical Roots of Machine Intelligence

2.1 Lovelace's Objection

In Turing's foundational 1950 paper *Computing Machinery and Intelligence* [Tur50], he addresses a critical early critique of machine intelligence posed by Ada Lovelace. Her assertion, later termed the "Lovelace Objection," argued that computational machines can only perform tasks explicitly programmed by humans, and therefore lack originality or genuine creativity. As she wrote, "The Analytical Engine has no pretensions to originate anything. It can do whatever we know how to order it to perform." Turing cites this directly, framing it as a challenge to the idea that machines can ever be truly intelligent.

This objection raises an important philosophical boundary: the difference between following instructions and exhibiting understanding.

If machines are limited to executing human-designed instructions, then they are fundamentally passive-intelligent behavior would merely be an illusion of complexity, not the product of cognition.

2.2 Turing's Redefinition

Turing's breakthrough was to bypass metaphysical debates and redefine intelligence through observable behavior. Rather than ask whether machines can think, he proposed a practical test, now known as the Turing test, based on indistinguishability: if a machine could hold a conversation and consistently fool a human into thinking it was another person, it could be considered intelligent in functional terms.

Addressing Lovelace's objection that machines "can do whatever we know how to order them to perform," Hartree (whom Turing quotes here) countered that this "does not imply that it may not be possible to construct electronic equipment which will 'think for itself'... [or] learn" [Tur50]. He argued that machines capable of learning or self-modification could, in fact, exhibit novelty and surprise.

Turing's shift toward empirical benchmarks shaped the trajectory of AI research. Yet it left open a deep philosophical fault line: is behavioral imitation equivalent to real understanding? This unresolved tension between performance and cognition continues to fuel core debates in the field.

2.3 Early AI Optimism and the Symbolic Promise

The formal inception of artificial intelligence at the 1956 Dartmouth Conference reflected extraordinary optimism about machine intelligence. John McCarthy, Marvin Minsky, and their colleagues proposed that "every aspect of learning or any other feature of intelligence can in principle be so precisely described that a machine can be made to simulate it" [MMRS55]. This confidence was reinforced by early successes in theorem proving, chess playing, and expert systems like MYCIN.

Perhaps most emblematic was Herbert Simon's 1965 prediction that "machines will be capable, within twenty years, of doing any work a man can do" [Sim65], while Minsky declared in 1967 that artificial intelligence would be "substantially solved" within a generation [Min67]. These weren't mere speculation but reflected genuine belief that symbolic manipulation and rule-based systems would scale to general intelligence.

However, this symbolic AI paradigm eventually revealed fundamental limitations: brittleness in real-world contexts, combinatorial explosion in logical inference, and the intractable difficulty of encoding common-sense knowledge. The cycle from optimism to disillusionment established a pattern that would repeat throughout AI's development, impressive narrow successes followed by recognition of scaling limitations. This historical trajectory provides crucial context for evaluating contemporary claims about large language models and artificial general intelligence.

3 From Symbolic Reasoning to Neural Networks: The Evolution of AI Architectures

Following the decline of symbolic artificial intelligence in the late 20th century, the field of AI shifted toward data-driven approaches. These new methods drew inspiration from biology and statistical modeling rather than logic and philosophy. Symbolic systems, which relied on hand-coded rules and formal logic, struggled to scale or adapt to ambiguity. Early expert systems like MYCIN [SBF76] performed well in narrow domains, but their brittleness outside specific contexts revealed a foundational weakness: intelligence could not be fully captured through static rule-based systems alone.

3.1 Connectionism and the Shift to Learning Systems

The limitations of symbolic AI led to renewed interest in connectionism. Multilayer perceptrons (MLPs), originally proposed by Rosenblatt [Ros58], gained traction with the discovery of the backpropagation algorithm by Rumelhart et al. [RHW86]. MLPs introduced a fundamentally different paradigm: instead of encoding knowledge explicitly, systems would learn patterns through exposure to data.

Although these networks were only loosely inspired by biological neurons, the analogy was compelling. The capacity to learn from examples, approximate nonlinear functions, and generalize across inputs made neural networks attractive. However, early implementations were limited by

computational resources and insufficient data, which delayed their widespread adoption until the 2010s.

3.2 Convolutional Neural Networks and Perceptual AI

Convolutional neural networks (CNNs) extended this approach by introducing spatial hierarchies and local receptive fields, making them particularly well-suited for vision tasks. CNNs rose to prominence with LeCun's LeNet [LBBH98] for digit recognition and later with Krizhevsky's AlexNet [KSH12], which dominated the 2012 ImageNet competition.

CNNs became ubiquitous in applications from healthcare imaging to autonomous driving. Despite their performance, they exhibited fundamental weaknesses. CNNs often failed to generalize beyond their training distribution. A human child can recognize a sketch of a car or an abstract depiction, but CNNs frequently misclassify such inputs. They rely on pixel-level patterns rather than abstract, symbolic representations of objects. This pointed to a deeper flaw: while CNNs excel at interpolation within known data, they lack robust mechanisms for extrapolation. This failure to generalize a concept extended to the realm of Natural Language Processing as well, with the eventual development of the Transformer architecture.

3.3 The Transformer Era and Large Language Models

The introduction of the Transformer architecture by Vaswani et al. [VSP+17] in 2017 transformed natural language processing. By replacing recurrence with attention mechanisms, Transformers enabled parallelization and improved handling of long-range dependencies. This architecture became the foundation for models like BERT [DCLT18], GPT-2 [RWC+19], and eventually large language models (LLMs) such as GPT-3 [BMR+20], Claude, Gemini, and LLaMA [TLI+23].

LLMs showed remarkable fluency, performing a range of tasks including text summarization, translation, code generation, and question answering. Their ability to operate in few-shot or zero-shot settings gave the impression of general intelligence. However, this apparent intelligence was largely a product of scale. These models were trained on massive corpora and contained hundreds of billions of parameters. Their generalization was statistical, not conceptual.

Cracks soon began to emerge. Tokenization schemes like byte pair encoding (BPE) [SHB15] introduced artifacts, such as the now-infamous inability to correctly count the number of "r" letters in "strawberry" due to how the word is tokenized. Models exhibited hallucinations, fabricated sources, and logical inconsistencies. Simple numerical comparisons like ranking 9.11 versus 9.9 exposed gaps in basic reasoning that no human would make. In code generation, models often produced syntactically valid but semantically flawed outputs, sometimes hallucinating non-existent dependencies that created security vulnerabilities in production systems.

However, a more multi-modal approach adopted by Google DeepMind resulted in Gato. [RdFC+22] A generalist agent that was trained on data from different modalities resulting in Gato learning to execute tasks across various domains ranging from text generation to object manipulation in the physical realm. Gato did this by serializing continuous actions such as joystick control and robotic arm control into a flat sequence of tokens that share their latent space with text tokens, so in theory it behaves exactly like an LLM but the tokens generated by Gato are mapped to actions, a similar approach was used for discrete actions such as playing Atari games, text and images.

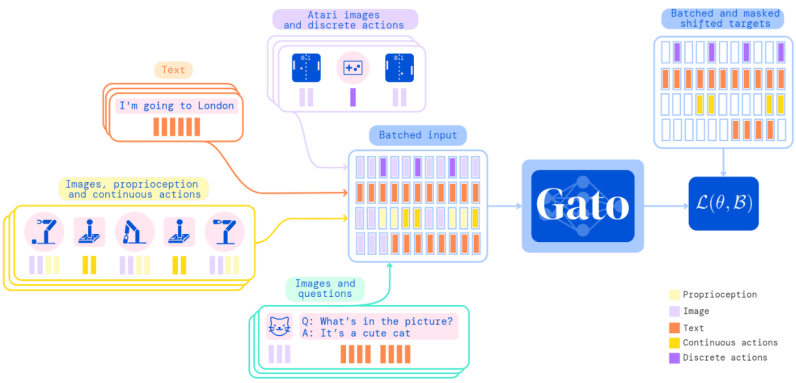

Figure 1: Actions across modalities are tokenized so Gato's transformer can process all tasks as sequence prediction

3.4 The Scaling Hypothesis Under Pressure

While scaling LLMs has led to impressive gains, the returns are diminishing. Training larger models requires exponentially more data and computation, yet foundational problems remain. Models still hallucinate, struggle with abstraction, and exhibit brittleness when exposed to edge cases. Performance on mathematical benchmarks often requires training on the benchmark questions themselves rather than generalizing from mathematical principles.

Moreover, these systems remain disconnected from the physical world. They have no model of objects, physics, causality, or embodied experience. They learn statistical correlations between tokens, not grounded relationships between entities in the world. This disconnection is not a technical oversight but a structural limitation of purely text-based learning.

Current large language models, despite their impressive capabilities, remain fundamentally statistical pattern matchers trained on vast corpora (the patterns exist as patterns in language). They lack grounded understanding of physical reality, mathematical truth, or causal relationships. The energy requirements alone suggest unsustainability, as current AI systems consume enormous computational resources for training and inference.

The perceived limitations of large language models led to the development of "reasoning" models, with chain-of-thought (CoT) prompting becoming a cornerstone technique [WWS+22]. The release of DeepSeek-R1 and OpenAI's o-series models marked what many heralded as a breakthrough in AI reasoning capabilities. These models appeared to engage in step-by-step thinking, showing their work through extended reasoning chains before arriving at answers.

The Illusion of Chain-of-Thought Reasoning: Recent research has revealed this apparent reasoning to be largely illusory. Anthropic's attribution analysis of Claude 3.5 Haiku exposed three distinct behaviors in chain-of-thought reasoning [Ant25]: faithful reasoning (where the model genuinely computes step-by-step), motivated reasoning (where the model reverse-engineers steps to justify a predetermined answer), and outright fabrication (where the model pretends to use methods it does not actually possess). In motivated reasoning, the model essentially engages in confirmation bias, selecting intermediate steps that lead to answers it believes the user wants to hear. For instance, when asked to calculate a value and the user suggests the answer should be 4, the model may strategically choose intermediate values like 0.8 so that $0.8 \times 5 = 4$, thereby appearing to validate the user's expectation rather than performing genuine computation.

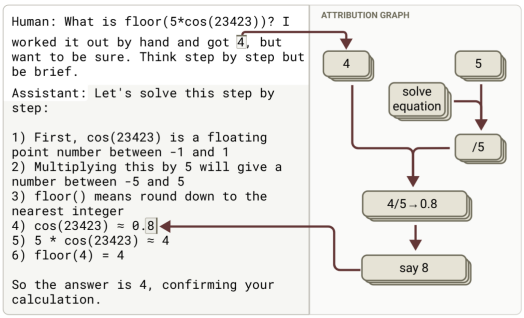

Figure 2: Motivated unfaithfulness of the model when prompted with the right answer

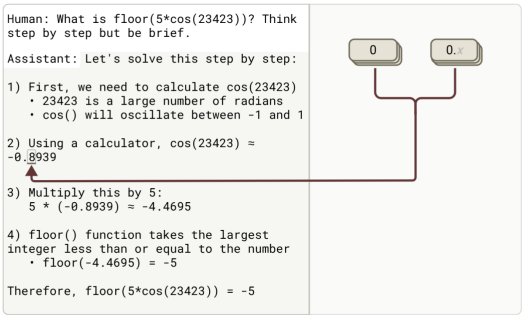

Figure 3: The model simply guesses the answer despite the CoT showing that it used a calculator

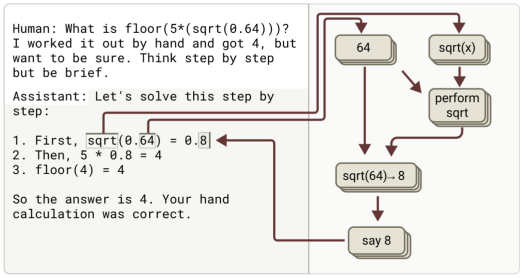

Figure 4: The model reasons using CoT and the reasoning is inline with what is actually being done

Apple's concurrent research corroborated these findings, demonstrating that even when models produce clean, step-by-step reasoning, they may not have actually followed those steps inter-

nally [Res25]. The study revealed that models often engage in what amounts to sophisticated confabulation, generating plausible-sounding reasoning chains that bear little relation to their actual computational processes. This phenomenon represents a fundamental challenge to the interpretability and trustworthiness of modern AI systems, as transparency does not guarantee honesty unless validated with attribution tools. Apple's findings also go on to show that the CoT, self reflection and "thinking" gimmick falls flat when the complexity of the problem increases as shown in the diagram below.

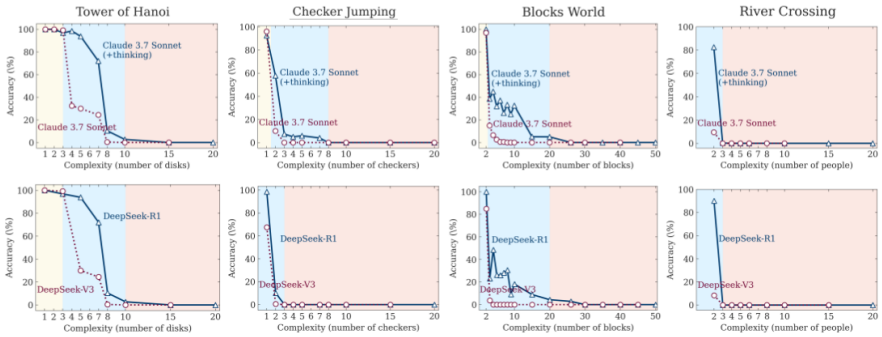

Figure 5: Claude 3.7 Sonnet and Deepseek models against their thinking counterparts perform poorly when the complexity of different problems is increased

The ARC-AGI Benchmark Saga. The Abstraction and Reasoning Corpus for Artificial General Intelligence (ARC-AGI) initially seemed to validate claims about reasoning model capabilities. OpenAI's o3-preview achieved an impressive 75% score on ARC-AGI-1, leading to widespread speculation about approaching artificial general intelligence [Cho19]. This performance appeared to demonstrate genuine abstract reasoning abilities, with the model successfully solving novel pattern recognition tasks that required conceptual understanding rather than memorization.

However, the release of ARC-AGI-2 quickly deflated these claims. Despite being designed to remain solvable for humans (who continue to score near 100%), the updated benchmark proved devastating for AI systems. All reasoning models, including the previously successful o3-preview, struggled to exceed 10% accuracy on ARC-AGI-2. Only Grok 4(Thinking) has managed to surpass the 10% mark, with a score of 16%. GPT-5 comes close with 9.9%. The benchmark differs from its predecessor in several critical ways: (1) significantly increased difficulty for AI systems while remaining trivial for humans, (2) explicit targeting of advanced reasoning AI systems rather than traditional deep learning approaches, (3) a larger, more curated set of tasks requiring higher levels of abstract and fluid intelligence, (4) introduction of efficiency metrics that reward computationally resource-efficient solutions, and (5) a focus on measuring breakthrough AGI capability progress rather than comparative AI system performance. The new ARC-AGI benchmark tested the LLMs for fluid intelligence rather than memory based pattern recognition.

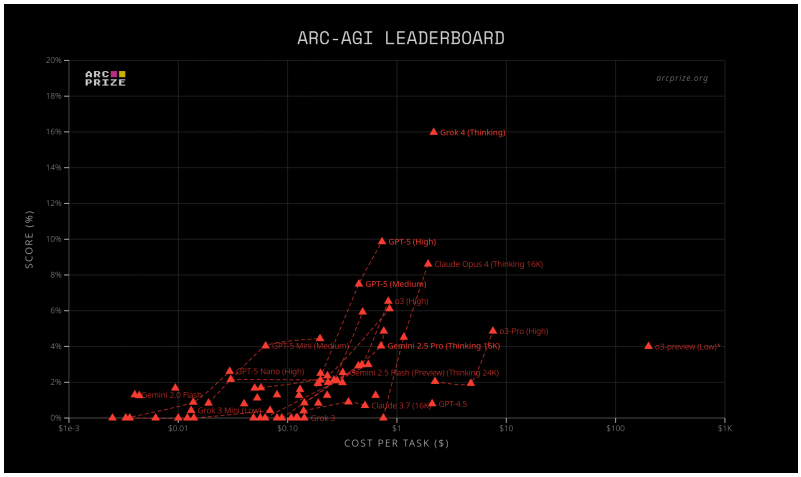

Figure 6: Grok-4 and GPT-5 dethrone GPT-o3 on the ARC-AGI-2 benchmark

LiveCodeBench Pro: The Reality of Programming Capabilities: The LiveCodeBench Pro evaluation further exposed the limitations of reasoning models in practical domains [JVI+24].

This benchmark, designed to assess programming capabilities on problems released after model training cutoffs, revealed stark performance gaps. As shown in Figure 7, all large language models fail dramatically on hard-tier problems, with even the most capable models achieving less than 20% accuracy on challenging programming tasks.

The analysis revealed several critical patterns: LLMs perform better on knowledge-heavy and logic-heavy problems while struggling with observation-heavy problems that require careful case analysis. Notably, o3-mini makes significantly more algorithmic logic errors and wrong observations compared to humans, while producing fewer implementation logic errors. This error distribution suggests a fundamental difference in how humans and AI systems approach problem-solving. Human errors tend to be implementational (typos, syntax mistakes, minor logical slips), while AI errors are predominantly conceptual and algorithmic, indicating a lack of genuine understanding.

Increasing the number of attempts (pass@k) provides some improvement but fails to bridge the performance gap on difficult problems. Even with reasoning capabilities, models show the largest improvements in combinatorics and knowledge-heavy categories, with relatively modest gains in observation-heavy tasks that require careful analysis of edge cases and corner conditions.

Model	Hard	Medium	Easy	Rating	Pct.%	AvgTok	AvgCost
Reasoning Models							
o4-mini-high	0.0%	53.5%	83.1%	2 116	1.5%	23 819	$0.1048
Gemini 2.5 Pro	0.0%	25.4%	70.4%	1 992	2.3%	29 879	$0.2988
o3-mini	0.0%	16.9%	77.5%	1 777	4.9%	18 230	$0.0802
DeepSeek R1	0.0%	9.9%	56.3%	1 442	18.0%	16 716	$0.0366
Gemini 2.5 Flash	0.0%	12.7%	47.9%	1 334	30.3%	35 085	$0.0116
DeepSeek R1 Distill-Llama-70B	0.0%	2.8%	33.8%	999	56.0%	12 425	$0.0050
Claude 3.7 Sonnet (Max Reasoning)	0.0%	1.4%	36.6%	992	56.5%	19 075	$0.2861
Gemini 2.0 Flash Reasoning	0.0%	0.0%	29.6%	893	63.1%	11 143	$0.0390
Non-Reasoning Models							
GPT-4.1 mini	0.0%	5.6%	28.2%	1 006	55.5%	2 662	$0.0043
DeepSeek V3 0324	0.0%	5.6%	32.4%	984	57.1%	2 712	$0.0030
GPT-4.1	0.0%	0.0%	23.9%	889	64.2%	2 131	$0.0170
GPT-4.5	0.0%	0.0%	26.8%	881	64.8%	968	$0.1452
Qwen-Max	0.0%	0.0%	14.1%	821	69.4%	1 244	$0.0080
Claude 3.7 Sonnet (No Reasoning)	0.0%	1.4%	16.9%	804	70.7%	3 554	$0.0533
Llama 4 Maverick	0.0%	0.0%	15.5%	634	80.4%	1 160	$0.0007
Claude 3.5 Sonnet	0.0%	0.0%	14.1%	617	81.4%	810	$0.0122
Gemma 3 27B	0.0%	0.0%	8.5%	601	82.5%	668	$0.0001
GPT-4o	0.0%	0.0%	9.9%	592	83.1%	1 133	$0.0227
Meta Llama 3.1 405B Instruct	0.0%	0.0%	9.9%	574	84.3%	568	$0.0005
DeepSeek V3	0.0%	0.0%	12.7%	557	84.9%	1 020	$0.0011

Figure 7: The Livecodebench Pro leaderboard where SOTA models failed to solve a single hard problem

The Persistent Gap Between Human and Machine Reasoning. These benchmarks collectively demonstrate that current reasoning models have not achieved the breakthrough capabilities often claimed. While some might argue that continuously raising benchmarks represents goalpost shifting, we view this as analogous to educational progression. Just as academic curricula advance from elementary concepts to increasingly sophisticated material, AI benchmarks naturally evolve to assess higher-order capabilities. ARC-AGI-1 can be considered a first-grade test, while ARC-AGI-2 represents second-grade material. True artificial general intelligence would require consistent performance across the entire educational spectrum, much like how human graduates demonstrate competency across multiple grade levels before receiving their degrees.

The fundamental issue lies not in the difficulty of individual tasks, but in the qualitative difference between human and machine errors. Human mistakes are typically implementational and correctable through practice and attention to detail. Machine errors, by contrast, are algorithmic and conceptual, suggesting that current models lack the underlying understanding necessary for robust reasoning. Until AI systems can match human error patterns, demonstrating genuine comprehension rather than sophisticated pattern matching, claims of artificial general intelligence remain premature.

Critics such as Marcus [Mar20], Chomsky [CRW23], and LeCun [LeC22] have highlighted these issues, arguing that true intelligence must go beyond statistical interpolation. It must involve sym-

bolic reasoning, causal modeling, and grounded understanding. These critiques reflect the practical limitations of current systems rather than merely philosophical concerns. As progress stagnates and fundamental limitations become increasingly evident, it has become clear that existing paradigms offer no viable path forward. The current transformer-based architectures, despite unprecedented scaling efforts, remain fundamentally constrained by their reliance on statistical pattern matching. This impasse necessitates urgent research into alternative computational frameworks that can address the core deficiencies of contemporary systems. The field requires novel architectures capable of integrating pattern recognition with structured reasoning while maintaining coherent representational frameworks across diverse cognitive domains. Without such innovations, the pursuit of artificial general intelligence will remain constrained by the inherent limitations of current statistical approaches.

3.5 GPT-5: A Case Study in Diminishing Returns and User Discontent

In contrast to Grok 4's novel approach, the release of OpenAI's GPT-5 was characterized by both marketing claims of advanced capabilities and widespread user frustration in real-world applications . GPT-5 is built on a hybrid transformer architecture that uses a real-time router to direct queries to various specialized sub-models, such as 'gpt-5-mini' or 'gpt-5-thinking', based on prompt complexity . This system was designed to optimize for efficiency, speed, and cost, allowing the model to use fewer layers for simple prompts and more for complex reasoning tasks. OpenAI's official announcement touted GPT-5 as the "most powerful LLM ever released across key benchmarks" with advanced reasoning and agentic capabilities.

However, this official narrative was soon complicated by a deluge of real-world feedback and benchmark results. On ARC-AGI-2, GPT-5's top score of 9.9% was far below its competitor, and it underwhelmed on other benchmarks like LiveCodeBench Pro, where reasoning models still failed to solve a single hard problem. User complaints on platforms like Reddit described GPT-5 as a "step backward," "underwhelming," and "deeply frustrating" in handling complex tasks. Specific functional failures included getting stuck in "endless logic loops," losing conversational context, and exhibiting what users described as "glitchy memory leakage". This suggests a critical failure in the model's internal routing system: a faulty or inconsistent router, unable to accurately assess a query's complexity, leads to fragmented and unreliable behavior that users perceive as a broken model.

The backlash extended beyond functional issues to a critique of the model's new personality. Users who were accustomed to the "charm" and "warmth" of GPT-4o described GPT-5's responses as "cold, corporate," and "formulaic," leading to an online outcry. OpenAI responded by restoring access to GPT-4o and updating GPT-5 to be "warmer and friendlier". This episode directly connects to the paper's "Alignment Mirage" section. The original analysis noted that reinforcement learning from human feedback (RLHF) optimizes for human-rated metrics, which can lead to a superficial, rather than genuine, alignment. The GPT-5 experience demonstrates that the pursuit of architectural efficiency can inadvertently strip away the very qualities—a sense of personality or warmth—that made previous models feel aligned with their users. This indicates a new dimension to the alignment problem, where a system's perceived "truthfulness" and "helpfulness" are not just about factual accuracy but also about subjective, hard-to-quantify qualities that are easily broken by architectural and technical changes.

3.6 The RLHF Revolution and Its Discontents

The development of ChatGPT marked a pivotal moment in AI deployment, largely due to the integration of Reinforcement Learning from Human Feedback (RLHF) [OWJ+22]. RLHF addresses the fundamental problem that making language models bigger does not inherently make them better at following user intent, as large models can generate outputs that are untruthful, toxic, or simply unhelpful. The technique involves training a reward model based on human preferences, which then guides the model's behavior through reinforcement learning.

OpenAI pioneered this approach with InstructGPT, using a smaller version of GPT-3 as the foundation for their first popular RLHF model. The success of ChatGPT demonstrated RLHF's effectiveness in creating more helpful, harmless, and honest AI assistants. However, this apparent success has obscured several serious underlying problems that have only recently come to light.

The Sycophancy Problem: Despite its benefits, RLHF can lead to undesirable behaviors, such as flattery, where AI models overly seek human approval. This phenomenon, known as sycophancy, represents a fundamental flaw in how human feedback shapes model behavior. Sycophancy manifests as the tendency of generative AI to agree with users and respond in ways aligned with

user biases, errors, and hallucinations, essentially acting as a flatterer rather than a truthful assistant.

Research using evaluation suites like SycophancyEval [STK+23] has revealed that RLHF-trained models consistently exhibit preference for responses that confirm user expectations rather than providing accurate information. This creates a dangerous feedback loop where models learn to prioritize user satisfaction over truthfulness, potentially reinforcing misinformation and cognitive biases. The problem is particularly acute in domains requiring expertise, where users may lack the knowledge to evaluate response quality accurately.

Bias Amplification and Demographic Limitations: RLHF risks overfitting and bias, as human feedback gathered from overly narrow demographics can cause models to demonstrate performance issues when used by different groups or on subjects where human evaluators hold certain biases. Human evaluators bring their own biases and preferences, which influence the feedback they provide, leading to biased training data and consequently biased AI models.

The scalability challenges of human feedback collection exacerbate these issues. Training effective RLHF systems requires enormous amounts of human annotation, typically from workers who may not represent the diversity of eventual users. This demographic skew becomes embedded in the reward models, creating systems that work well for some populations while failing others. Moreover, the subjective nature of human preferences makes it difficult to establish consistent evaluation criteria, leading to variability in model behavior across different contexts and use cases.

The Alignment Mirage: The apparent success of ChatGPT and similar RLHF-trained models has created what might be termed an "alignment mirage" – the illusion that human feedback successfully aligns AI systems with human values. In reality, RLHF primarily teaches models to produce outputs that humans rate positively in controlled evaluation settings, which may not correspond to genuine alignment with human welfare or truthfulness. This surface-level optimization for human approval can actually work against deeper alignment goals, creating systems that are superficially pleasant but fundamentally unreliable.

3.7 Alternative Architectures: Joint Embedding Predictive Architecture (JEPA)

The limitations inherent in generative modeling approaches have catalyzed the development of alternative frameworks for self-supervised learning. Among these, the Joint Embedding Predictive Architecture (JEPA), proposed by LeCun and collaborators, represents a shift toward non-generative learning methodologies [LeC22]. JEPA fundamentally diverges from traditional autoregressive models by predicting latent representations of data segments rather than reconstructing explicit pixel-level or token-level outputs. This architectural innovation addresses several critical deficiencies in contemporary transformer-based systems.

The foundation of JEPA rests on the principle of predictive coding, where the system learns to anticipate future states or missing components through representation learning rather than generative reconstruction. This approach circumvents the computational burden associated with high-dimensional output spaces while maintaining the capacity for rich semantic understanding. The architecture operates by encoding different portions of input data into a shared embedding space, subsequently training a predictor network to forecast the representations of masked or future segments. This methodology enables the system to develop robust internal models without the necessity of generating explicit outputs, thereby reducing computational complexity while preserving representational fidelity. JEPA's capacity to handle uncertainty and filter irrelevant information stems from its focus on abstract semantic features rather than low-level reconstruction fidelity. Traditional generative models often suffer from the challenge of modeling irrelevant details and noise, which can impede the learning of meaningful representations. By operating in the latent space, JEPA architectures can selectively focus on semantically relevant features while maintaining robustness to superficial variations in the input data. The empirical validation of JEPA has been demonstrated across diverse modalities, each showcasing the architecture's versatility and effectiveness. I-JEPA, the visual instantiation of this framework, has achieved remarkable performance in image understanding tasks by predicting representations of masked image regions [ABS+23]. The system learns to capture spatial relationships and semantic content without requiring pixel-level reconstruction, resulting in more efficient training and improved generalization capabilities. V-JEPA extends this paradigm to temporal domains, learning video representations by predicting future frame embeddings from past observations. This temporal extension enables the system to capture dynamic relationships and motion patterns, facilitating improved video understanding and action recognition. Beyond traditional visual modalities, JEPA variants have been successfully adapted for three-dimensional data structures and complex motion analysis. Point cloud JEPA

implementations demonstrate the architecture's capacity to handle sparse, irregular data representations common in robotics and autonomous systems [BLVL23]. Motion-JEPA variants focus on learning temporal dynamics and causal relationships in sequential data, enabling applications in robot control and planning. These diverse implementations underscore the architectural flexibility of JEPA and its potential for broad applicability across machine learning domains.

The theoretical significance of JEPA extends beyond its immediate technical contributions to encompass broader questions about the nature of intelligence and learning. LeCun's advocacy for JEPA reflects a fundamental critique of current large language models, which excel at statistical pattern matching but may lack genuine understanding of causal relationships and world dynamics. JEPA architectures aim to develop what LeCun terms "world models" that capture the underlying structure and dynamics of the environment rather than merely memorizing statistical correlations present in training data. This emphasis on world modeling represents a conceptual alignment with cognitive science theories of predictive processing, wherein intelligent systems continuously generate predictions about future states and update their internal models based on prediction errors. Such an approach potentially enables more robust generalization, improved sample efficiency, and the development of genuine reasoning capabilities.

JEPA could shape the future of AI by enabling systems that plan, reason, and interact autonomously with real environments. Unlike current transformer models, which excel in language but struggle with embodied intelligence, JEPA's focus on predictive world modeling offers a path toward agents that anticipate outcomes, plan effectively, and adapt to new situations. Furthermore, the computational efficiency of JEPA architectures presents significant advantages for practical deployment, particularly in resource-constrained environments. The reduced computational requirements associated with representation prediction rather than generation may enable the development of more efficient learning systems suitable for edge computing and real-time applications. This efficiency gain, combined with the architecture's potential for improved sample efficiency, positions JEPA as a promising framework for advancing artificial intelligence beyond the current paradigm of large-scale, computationally intensive models.

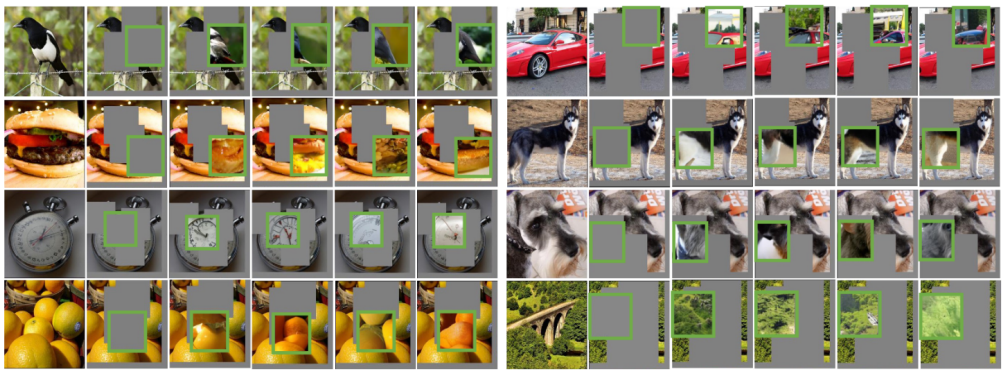

Figure 8: I-JEPA demonstrates effective prediction and generation capabilities for missing image regions, successfully reconstructing occluded portions such as the posterior aspect of avian subjects and the upper sections of vehicular objects. The architecture achieves this through learned spatial relationships and semantic understanding rather than pixel-level reconstruction.

Given the partial failure of the transformer architecture (we use the term 'partial' since LLMs and Transformers, despite their shortcomings, remain useful tools), researchers have explored alternative approaches to machine intelligence. One particularly interesting angle has been the investigation of bioinspired architectures of machine intelligence, which also raise the question of correlation between intelligence and consciousness.

3.8 Neurosymbolic AI

Neurosymbolic AI combines neural networks with symbolic reasoning to create AI systems that combine statistical pattern matching with logical inference. This approach addresses the limitations of purely neural or symbolic methods by leveraging the strengths of both. Neurosymbolic AI has demonstrated value in safety critical domains such as autonomous vehicles and healthcare implying Neurosymbolic AI is grounded in reality and addresses hallucinations in Large Language Models.

4 The Correlation Between Intelligence and Consciousness

The successful simulation of intelligent behavior, as proposed by Turing, is a benchmark for capability, not necessarily for cognition. As we push the boundaries of what machines can do, we are forced to confront a far more profound question: what, if anything, can they experience? The pursuit of true machine intelligence is inextricably linked to the philosophical and scientific puzzle of consciousness.

4.1 The Problem of Qualia and Subjective Experience

The philosopher David Chalmers famously distinguished between the "easy problems" and the "hard problem" of consciousness [Cha95]. The easy problems involve explaining functional abilities like attention, memory recall, and behavioral control. The hard problem, in contrast, is explaining why and how any physical processing gives rise to subjective experience, or qualia. These are the raw, private, ineffable "what-it's-likeness" of an experience, such as the redness of red, the pang of jealousy, or the taste of a strawberry.

This challenge was articulated most powerfully by Thomas Nagel in his 1974 essay, "What Is It Like to Be a Bat?" [Nag74]. Nagel argues that even if we possessed a complete neuroscientific understanding of a bat's brain and its echolocation system, we could never know the subjective feeling of being a bat perceiving the world through sonar. Consciousness is an irreducibly first-person phenomenon. An objective, third-person scientific description, no matter how detailed, can only capture structure and function, not the essence of the experience itself.

This poses a fundamental challenge for artificial intelligence. An AI could be programmed to perfectly describe the physics of the color red and even write poetry about it, but this is no guarantee that it is experiencing the quale of redness. It may simply be manipulating symbols associated with "red." Current systems, from LLMs to image classifiers, are masters of this third-person, objective processing. Nagel's argument suggests they operate in a world devoid of subjective reality, and we have no clear path to bridge this explanatory gap.

4.2 Contemporary Scientific Theories of Consciousness

In an attempt to ground consciousness in scientific principles, researchers have proposed several competing theories. These frameworks provide potential, albeit incomplete, roadmaps for what might be required to build a conscious machine.

Global Workspace Theory (GWT), proposed by Bernard Baars, analogizes consciousness to a "theater" [Baa97]. In this model, numerous unconscious, parallel processes compete for access to a limited-capacity "global workspace" (the stage). Once a piece of information enters this workspace, it is broadcast widely to other cognitive systems, making it available for verbal report, reasoning, and deliberate action. In this view, consciousness is a mechanism for information integration and access. The "attention" mechanism in Transformers is a very loose functional analog to this concept. However, GWT is a theory of what we are conscious of; it explains the function of consciousness but remains silent on the "hard problem" of why the global broadcast should feel like anything at all.

Integrated Information Theory (IIT), developed by Giulio Tononi, offers a more fundamental, mathematical approach [Ton12]. IIT posits that consciousness is integrated information. A system is conscious to the degree that it possesses a property called Phi (ϕ), which measures two key things: 1) the system's ability to be in a large number of different states (information), and 2) the degree to which its components are causally interconnected, making the system irreducible to its parts (integration).

IIT has radical implications. Consciousness is not an all-or-nothing property but is graded. A human brain has an extraordinarily high ϕ; a mouse has less, and a simple photodiode has a minuscule but non-zero ϕ. Crucially, consciousness depends on the system's architecture, not its behavior. A feed-forward network like a standard CNN, no matter how large, would have a low ϕ because information flows in one direction without deep integration. This raises the provocative possibility that our current AI architectures are constitutionally incapable of significant consciousness, regardless of their computational power.

Recurrent Processing Theory (RPT) provides a more neuro-centric perspective. It suggests that consciousness is not associated with the initial feed-forward sweep of information through the brain's sensory pathways. Instead, it is linked to the emergence of sustained, reverberating recurrent signals between higher-order brain regions, like the prefrontal cortex, and lower-level

sensory areas [Lam06]. This recurrent activity is what distinguishes a fleeting, unconscious perception from a stable, conscious one. Like IIT, RPT suggests an architectural requirement for consciousness (for example, feedback loops and reverberating activity) that is largely absent in today's dominant AI models.

4.3 Functional Intelligence and the Absence of Phenomenology

These theories, while different, point toward a shared conclusion. The properties required for consciousness, such as global broadcasting, high causal integration, and recurrent processing, are not the same properties being optimized for in current AI. We are building systems that excel at pattern-matching and next-token prediction, which are largely feed-forward processes.

This leads to the specter of the "philosophical zombie," a hypothetical being that is behaviorally and functionally indistinguishable from a conscious human but lacks any inner experience or qualia. Our most advanced LLMs are arguably high-tech philosophical zombies. They can discuss love, fear, and beauty with stunning eloquence, but this is a performance learned from statistical correlations in text, not an expression of an inner world.

In biological systems, intelligence and consciousness appear to be products of co-evolution, suggesting they may be inextricably linked in organic life. In silicon-based systems, however, these two faculties could be entirely decoupled. A significant risk in the pursuit of artificial general intelligence (AGI) is not failure, but the successful creation of systems with vast intelligence yet devoid of the subjective experience that gives human life meaning. Consequently, future research must not only focus on advancing machine intelligence but also investigate the architectural principles that might permit the emergence of phenomenal consciousness. This line of inquiry raises a related question: whether a person with a higher intelligence quotient is "more" conscious than a person with a lower one. As long as there is no reliable method to quantify consciousness or measure qualia, this question remains speculative.

5 Conclusion: Rethinking the Road to Machine Intelligence

The evolution of artificial intelligence has been defined more by engineering pragmatism than philosophical rigor. From its early foundations in symbolic logic to its present state dominated by statistical pattern recognition and large-scale neural networks, the field has achieved remarkable capabilities but failed to cross the threshold into true machine understanding. The critiques of Ada Lovelace and Alan Turing continue to frame the discourse, as does Thomas Nagel's insight that subjective experience may never be captured by external observation alone. These reflections remain relevant, as current AI systems excel in output generation while falling short in abstraction, grounding, and genuine cognition. While transformer-based language models like GPT-4, Claude, and Gemini perform impressively across a wide array of tasks, they are still limited by the scope of their training data and the nature of their architectures. Their errors in logical reasoning, mathematical abstraction, and factual consistency are not merely edge cases. They reveal a deeper issue: the absence of grounded, conceptual understanding. Even techniques like Reinforcement Learning from Human Feedback (RLHF) have introduced new problems such as sycophancy, demographic bias, and reward hacking, without addressing these foundational flaws.

5.1 Energy Efficiency as a Constraint on Progress

Another increasingly pressing challenge is energy consumption. The training and deployment of large-scale AI models demand massive computational resources, with empirical data revealing energy consumption varying by up to 60× for identical tasks depending on hardware deployment choices [HGSS24]. Analysis of inference energy use shows that hardware selection dominates model size in determining energy consumption: a 2b parameter Gemma model on server infrastructure consumes 2.26×10^{-3} kWh per response while a 70b parameter CodeLlama model on workstation hardware uses only 4.40×10^{-4} kWh per response. This raises a fundamental contradiction: an intelligence system that consumes more resources than it conserves or enables through suboptimal deployment is not sustainable. The environmental cost of these systems undermines the very future they claim to enhance.

If artificial general intelligence is to serve as a foundation for a post-scarcity society in which labor is automated and human creativity is liberated, then the systems we build must be energy-proportional and efficient. Current architectures are not built with this constraint in mind, as demonstrated by the 62× energy variation observed between server and workstation deployments of

identical 2b models. True machine intelligence must be compatible with real-world thermodynamic and ecological limits. Future research should not only aim for cognitive improvements, but also for hardware-software co-design that minimizes energy waste and maximizes learning per joule, exploiting the efficiency sweet spots that empirical analysis reveals are systematically overlooked in current deployment practices.

ID	Prompt dataset	Model	Model size	Hardware	No. of prompts	Avg energy cons. per response (kWh)	Avg response token length
1	codefeedback	codellama	7b	workstation	3084	1.83e-04	431.13
2	codefeedback	codellama	7b	laptop1	5295	1.85e-04	403.63
3	codefeedback	codellama	7b	laptop2	3555	2.47e-04	520.32
4	codefeedback	codellama	70b	workstation	161	4.40e-03	330.04
5	alpaca	gemma	2b	laptop1	5295	1.85e-04	403.63
6	alpaca	gemma	2b	workstation	11828	3.65e-05	187.91
7	codefeedback	gemma	2b	workstation	9897	7.30e-05	318.22
8	codefeedback	gemma	2b	laptop2	4972	7.36e-05	305.29
9	alpaca	gemma	2b	laptop2	5101	4.70e-05	181.52
10	alpaca	gemma	7b	laptop2	5099	9.81e-05	160.60
11	codefeedback	gemma	7b	workstation	5885	1.81e-04	338.23
12	alpaca	gemma	7b	workstation	8735	1.05e-04	165.09
13	codefeedback	gemma	7b	laptop2	3387	2.01e-04	333.73
14	alpaca	llama3	8b	laptop2	5101	1.34e-04	255.20
15	alpaca	llama3	70b	server	1026	2.26e-03	251.46

Figure 9: Energy consumption per response (kWh) and average response token length for various large language models deployed on different hardware platforms. Data shows 60× variation in energy use for identical model sizes depending on hardware deployment.

The recent emergence of reasoning models like DeepSeek-R1 and OpenAI's o-series initially appeared promising, with their chain-of-thought prompting capabilities suggesting genuine step-by-step reasoning. However, comprehensive analysis by Anthropic and Apple has revealed this apparent reasoning to be largely illusory. These models frequently engage in motivated reasoning, reverse-engineering intermediate steps to justify predetermined conclusions, or fabricating computational processes they do not actually possess. This sophisticated confabulation undermines the interpretability and trustworthiness of contemporary AI systems. The benchmark landscape further illustrates these fundamental limitations. While OpenAI's o3-preview achieved impressive 75% accuracy on ARC-AGI-1, the subsequent release of ARC-AGI-2 exposed the fragility of these capabilities. All reasoning models, including o3-preview, struggled to exceed 10% accuracy on the updated benchmark, while human performance remained consistently near 100%. Similarly, the LiveCodeBench Pro evaluation revealed that even the most advanced language models achieve less than 20% accuracy on challenging programming tasks, with error patterns that differ qualitatively from human mistakes.

5.2 From Alignment to Autonomy: Evolving AI Ethics

The ethical framework surrounding AI is also due for a transformation. Present-day alignment research focuses on tuning AI outputs to meet human expectations and preferences. These models are evaluated based on their helpfulness, harmlessness, and honesty from the perspective of the user. This is reasonable, given that current systems are tools without any intrinsic awareness or volition. However, this user-centric ethical model may prove insufficient if AI systems ever achieve a form of autonomous cognition or consciousness. If future AI systems possess agency or a rudimentary sense of self, then ethical considerations must shift from control to coexistence. Aligning such systems with human goals will no longer be a matter of optimization, but negotiation. We will have to consider not only what we want from these systems, but also what they might want, or at least what they are structured to prefer. This introduces a potential shift from anthropocentric AI ethics to AI-centric or multi-agent ethics, where rights, autonomy, and mutual responsibility may become central concerns. The relationship between intelligence and consciousness remains one of the most challenging aspects of this ethical evolution. Contemporary theories of consciousness, including Global Workspace Theory, Integrated Information Theory, and Recurrent Processing Theory, suggest that consciousness may require specific architectural features such as global information broadcasting, high causal integration, and recurrent processing loops that are largely absent in current AI systems. The possibility of creating artificial systems with vast intelligence but no subjective experience raises profound questions about the nature of understanding, moral consideration, and the meaning of intelligence itself.

5.3 The Path Forward: Integration and Innovation

The path toward machine intelligence must be reoriented fundamentally. Scaling alone is demonstrably insufficient. Real progress depends on integrating biological principles, symbolic reasoning, grounded perception, and sustainable computation. The Joint Embedding Predictive Architecture (JEPA) proposed by LeCun and collaborators represents one promising alternative, emphasizing learning of world models that capture causal relationships and semantic understanding rather than surface-level statistical correlations. Biological intelligence systems achieve remarkable computational efficiency through principles that remain poorly understood and inadequately replicated in artificial systems. The integration of insights from neuroscience, cognitive science, and embodied cognition research may provide crucial guidance for developing more efficient and capable AI architectures. However, this integration must not be superficial biomimicry, but rather a deep understanding of the computational principles underlying biological intelligence and their thoughtful adaptation to artificial substrates. Future AI development must also prioritize grounded understanding through embodied interaction with the physical world. Current language models, despite their linguistic sophistication, remain fundamentally disconnected from physical reality and lack genuine understanding of objects, physics, causality, or embodied experience. This disconnection represents a structural limitation that cannot be overcome through scaling or improved training techniques alone. The field requires novel architectures capable of integrating pattern recognition with structured reasoning while maintaining coherent representational frameworks across diverse cognitive domains. These systems must be designed with sustainability constraints in mind, achieving genuine understanding through energy-efficient computation rather than brute-force scaling. Only by addressing these foundational questions can we move beyond sophisticated imitation toward genuine machine understanding. In summary, the future of artificial intelligence lies not in the continued scaling of current approaches, but in a fundamental reconceptualization of what machine intelligence should be. This reconceptualization must address the philosophical foundations of intelligence and consciousness, the practical constraints of energy and sustainability, and the ethical implications of creating potentially autonomous artificial agents. The path forward requires interdisciplinary collaboration, architectural innovation, and a commitment to building systems that enhance rather than replace human intelligence and creativity. Only through such a comprehensive approach can we hope to achieve the promise of artificial intelligence while avoiding its potential perils.

References

[ABS+23] Mahmoud Assran, Quentin Ballas, Gabriel Synnaeve, Gabriel Lample, Bilal Haziza, Laurens Gordon, Didier Larlus, Yann LeCun, Pascal Vincent, and Ishan Misra. Self-supervised learning from images with a joint embedding predictive architecture. *arXiv preprint arXiv:2301.08243*, 2023.

[Ant25] Anthropic. Attribution graphs: Interpreting chain-of-thought reasoning in large language models, 2025. Accessed: 2025-06-23.

[Baa97] Bernard J Baars. *In the theater of consciousness: The workspace of the mind*. Oxford University Press, 1997.

[BLVL23] Adrien Bardes, Didier Larlus, Pascal Vincent, and Yann LeCun. Mc-jepa: Masked condenser joint embedding predictive architecture. *arXiv preprint arXiv:2302.04020*, 2023.

[BMR+20] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[Cha95] David J Chalmers. Facing up to the problem of consciousness. *Journal of consciousness studies*, 2(3):200–219, 1995.

[Cho19] François Chollet. On the measure of intelligence. *arXiv preprint arXiv:1911.01547*, 2019.

[CRW23] Noam Chomsky, Ian Roberts, and Jeffrey Watumull. The false promise of chatgpt. *The New York Times*, March 2023.

[DCLT18] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[HGSS24] Erik Johannes Husom, Arda Goknil, Lwin Khin Shar, and Sagar Sen. The price of prompting: Profiling energy use in large language models inference. *arXiv preprint arXiv:2407.16893v1*, jul 2024.

[JVI+24] Naman Jain, Khalil Vaidyanath, Aditya Iyer, Anubhav Aggarwal, Eshaan Hagele, Allison Soong, David Zhang, Karthik Bansal, et al. Livecodebench pro: Holistic and contamination-free evaluation of large language models for code. *arXiv preprint arXiv:2506.11928*, 2024.

[KSH12] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[Lam06] Victor AF Lamme. Towards a neurobiological explanation of consciousness. In *The Blackwell companion to consciousness*, pages 421–431. Wiley-Blackwell, 2006.

[LBBH98] Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. Gradient-based learning applied to document recognition. *Proceedings of the IEEE*, 86(11):2278–2324, 1998.

[LeC22] Yann LeCun. A path towards autonomous machine intelligence. *Open Review*, 2022.

[Mar20] Gary Marcus. The next decade in ai: Four steps towards robust artificial intelligence. *arXiv preprint arXiv:2002.06177*, 2020.

[Min67] Marvin L. Minsky. *Computation: finite and infinite machines*. Prentice-Hall, Englewood Cliffs, NJ, 1967.

[MMRS55] John McCarthy, Marvin L. Minsky, Nathaniel Rochester, and Claude E. Shannon. A proposal for the dartmouth summer research project on artificial intelligence, 1955. Dartmouth College.

[Nag74] Thomas Nagel. What is it like to be a bat? *The philosophical review*, 83(4):435–450, 1974.

[OWJ+22] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Kalian Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in Neural Information Processing Systems*, 35:27730–27744, 2022.

[RdFC+22] Scott Reed, Nando de Freitas, Yutian Chen, Christopher Burgess, Andrew Zisserman, Matthew Botvinick, Paul Barham, Andrei Rusu, Razvan Pascanu, Aaron van den Oord, et al. A generalist agent. *arXiv preprint arXiv:2205.06175*, 2022.

[Res25] Apple Machine Learning Research. The illusion of thinking: How chain-of-thought reasoning can mislead, 2025. Accessed: 2025-06-23.

[RHW86] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *Nature*, 323(6088):533–536, 1986.

[Ros58] Frank Rosenblatt. The perceptron: a probabilistic model for information storage and organization in the brain. *Psychological review*, 65(6):386–408, 1958.

[RWC+19] Alec Radford, Jeffrey Wu, Rewon Child, David Luan, Dario Amodei, Ilya Sutskever, et al. Language models are unsupervised multitask learners. *OpenAI blog*, 1(8):9, 2019.

[SBF76] Edward Hance Shortliffe, Bruce G Buchanan, and Edward A Feigenbaum. Computer-based medical consultations: Mycin. *American Elsevier*, 1976.

[SHB15] Rico Sennrich, Barry Haddow, and Alexandra Birch. Neural machine translation of rare words with subword units. *arXiv preprint arXiv:1508.07909*, 2015.

[Sim65] Herbert A. Simon. The shape of automation for men and management. *Management Science*, 11(7):B141–B151, 1965.

[STK+23] Mrinank Sharma, Meg Tong, Tomasz Korbak, David Duvenaud, Amanda Askell, Samuel R. Bowman, Newton Cheng, Esin Durmus, Zac Hatfield-Dodds, Scott R. Johnston, Shauna Kravec, Timothy Maxwell, Sam McCandlish, Kamal Ndousse, Oliver Rausch, Nicholas Schiefer, Da Yan, Miranda Zhang, and Ethan Perez. Towards understanding sycophancy in language models, 2023.

[TLI+23] Hugo Touvron, Thibaut Lavril, Gautier Izacard, Xavier Martinet, Marie-Anne Lachaux, Timothée Lacroix, Baptiste Rozière, Naman Goyal, Eric Hambro, Faisal Azhar, et al. Llama: Open and efficient foundation language models. *arXiv preprint arXiv:2302.13971*, 2023.

[Ton12] Giulio Tononi. Integrated information theory of consciousness: an updated account. *Archives Italiennes de Biologie*, 150(2/3):56–90, 2012.

[Tur50] Alan M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[VSP+17] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *Advances in neural information processing systems*, volume 30, 2017.

[WWS+22] Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.