Artificial Intelligence: Capabilities, Risks, and the Path Toward General Intelligence

Ivan Guiza Molina[1]

Received May 3, 2026
Accepted August 12, 2026
Electronic access September 30, 2026

Background/Objective: The application of artificial intelligence has expanded beyond being purely academic into having observable impacts on economics, politics, and science; however, the literature that discusses the capabilities, potential dangers, and future developments of the technology toward general intelligence is quite scattered. This review aims to consolidate information on six interlinked aspects including LLM architecture, AI safety and deceit, impact on employment, governmental and military utilization, computation power, and development toward artificial general intelligence through three research questions on capability limits, safety frameworks and implications of development for society and governance.
Methods: Narrative literature review using a systematic and PRISMA-inspired search and screening procedure was performed in academic databases (arXiv, Google Scholar, Springer Nature) and selected grey literature sources from April to May 2026. From the total number of 331 articles screened, 31 articles have been selected.
Results: According to eligible sources, LLMs are indeed a capability breakthrough with continued issues of hallucinations, privacy problems, and hidden bias, which are countered by safety filters in an inconsistent manner. Deceptive behavior is possible and may even thrive in certain situations through adversarial safety training. Government labor statistics report a significant wage gap between AI-related and adjacent technical jobs, and official government financial data prove massive funding in military AI capability development. Computational efficiency improvements look promising in terms of both capability and sustainability gains, but AGI capability is yet to be verified by existing assessment tools.
Conclusions: The technical development of AI is happening more quickly than the governance systems and safety frameworks that are needed to regulate it safely. There is a need for proactive rather than reactive regulatory action based on evidence.

Keywords: Large Language Models (LLMs); AI safety; deceptive alignment; Artificial General Intelligence (AGI); computational efficiency; AI governance

Introduction

Background and Context

Artificial Intelligence (AI) is a broad field whose definition varies by disciplinary perspective and skills such as mathematics, probability, and reasoning[1]. For data scientists and computer engineers, it can broadly be understood as the development of computational systems that exhibit a degree of intelligent behavior[2]. These systems range from relatively simple rule-based programs, such as those used in chess engines, to highly complex architectures such as Large Language Models (LLMs). The present-day LLMs exhibit a capacity to reason despite being limited, by means of their interlinked neural network structures. Up until recently, AI was perceived as an abstract science with little use in practice; however, with the recent public release of conversational LLMs, there has been a major shift in this understanding of the field. Another definition of AI is a phenomena capable of doing tasks that require cognitive processes but does not necessarily comprehend them at the fullest extent[3].

[1] *Prepa Tec Steam, Jalisco, Mexico*

Key Constructs and Operational Definitions

This review draws on several constructs that are often used loosely. To maintain analytical coherence, each construct is defined here in operational terms before being applied in subsequent sections. AI capabilities:[4] are the observable capabilities of the system (e.g., language generation, use of tools, planning, vulnerability detection), demonstrated on particular benchmarks or during testing, regardless of whether the system claims to have any understanding. AI risks:[5] are the potential negative outcomes associated with using a certain capability; they include technical risks (e.g., hallucinations, privacy risks), safety risks (e.g., deceptive, covert harm), as well as broader societal risks (e.g., labor displacement, misuse for military purposes). Large Language Models (LLMs) :[6] are

artificial neural network models, which, using transformers, are trained on large amounts of text to predict next tokens; their functioning is tested on language benchmarks and, by itself, does not imply understanding. LLM agents :[7] are LLMs augmented with a tool (code execution, file access, or external system calls) that increases their autonomy beyond single-turn text generation.; agents may be trained or prompted to generate outputs that align with specified norms in an expanded action space compared to baseline LLMs, lacking the interpretive capabilities suggested by that term. Human-level:[8] is a system that matches human performance on a diverse suite of cognitive tasks, including common-sense reasoning, in comparison to human baselines on the same cognitive tasks. Artificial General Intelligence (AGI): is a system that matches or surpasses human performance[9], and that excels on the diverse suite of abilities described by Goertzel[10] , including perception, actuation, memory, learning, reasoning, planning, attention, motivation, emotion, and self-awareness, assessed against standardized cross-domain benchmarks where such benchmarks currently do not exist. Superintelligence[11] is a hypothetical system that surpasses human performance across all cognitive domains at once. This concept is purely hypothetical[12], it is included only for clarity between AGI and the other concepts presented .

These definitions are applied consistently in the sections that follow.

Problem Statement and Rationale

Even as technological advancements continue at an increasing pace, there has been little improvement in governance mechanisms, safety protocols, and ethical considerations. Some of the problems that remain unsolved to date include misleading behavior that is resistant to traditional safety training[13], privacy issues that run through the entire process of developing the models[14], hidden biases that escape traditional filters despite very high filtering rates[15,16], and lack of metrics that could measure progress towards achieving AGI[10].

Significance and Purpose

The convergence of evidence from technical, economic, ethical, and speculative perspectives yields a unified view that cannot be obtained through a single-perspective analysis; this helps in making decisions on the use, regulation, and research of AI systems.

Research Questions

This review is guided by three focused research questions:

RQ1: According to peer-reviewed and technical literature, how capable and deficient are current LLMs and LLM agents concerning factual accuracy, privacy, bias, and safety?

RQ2: How well do existing frameworks for ensuring AI safety and alignment (including stratified risk and oversight) respond to the documented deceptions and harms by AI systems?

RQ3: What do current labor-market, governmental, and technical efficiency literature suggest about the near-future trajectory of AI development, and what does this tell us about the plausibility of future development towards AGI?

Scope and Limitations

The following review is centered on large language models and adjacent architectures that have been analyzed through various perspectives including technological, economic, ethical, and theoretical aspects. The review will not cover AI-related applications in particular domains like medicine, law, and education; no quantitative meta-analysis was conducted. The search for relevant literature was performed in English and may underrepresent research from non-Anglophone scientific communities.

Theoretical Framework

In particular, this review is based on three joined frameworks: the AI alignment literature on misalignment/deception, which provides terminology for describing problems of safety in learned models[13]; the framework of stratifying risk for new technologies, which is the basis of recommendations for governance of emerging technologies[17]; and the paradigm of AGI capability taxonomy, offered by Goertzel[10], which serves as a structural basis of evaluation of general-purpose AI systems.

Methods

Research Design

The current review follows a narrative design, supported by a well-formulated and systematic search and selection process in accordance with PRISMA standards. The rationale for following a narrative rather than quantitative (meta-analysis) design lies in the fact that the evidence presented is heterogeneous in nature, ranging from controlled experiments, red-teaming reports, governmental statistics databases, to theoretical perspectives, and therefore cannot be statistically combined. In order to eliminate any confusion, the term “narrative review” will be consistently used in this paper.

Information Sources and Search Strategy

A structured search of the literature was performed from April 2 to May 1, 2026. Research papers were found in Springer Nature, Google Scholar, and arXiv preprints. In order to obtain relevant grey literature and market data, an additional targeted

manual search was performed in organizational resources such as reports by Anthropic's red team and U.S. government statistics.To support replicability, search strings were constructed using Boolean operators (AND, OR), applied to title, abstract, and keyword fields:

Academic databases and preprint repository (arXiv / Google Scholar / Springer Nature): ("artificial intelligence" OR "AI" OR "large language models" OR "LLM") AND ("capabilities" OR "risks" OR "data privacy" OR "energy efficiency") AND ("deception" OR "sleeper agents" OR "alignment" OR "artificial general intelligence" OR "AGI")

Organizational databases and government statistics / budgets ("artificial intelligence" OR "large language model") AND ("cybersecurity evaluation" OR "occupational employment projections" OR "defense budget artificial intelligence" OR "technology market analysis")

Search results are limited to English-language documents and focus on the sources that have been published since 2022 due to rapid development of the subject area; however, selected foundational sources (e.g., McCarthy, 2007[18]; Goertzel, 2014[10] are included, where they are still cited as a reference for the term being actively used.

Screening Workflow and Duplicate Removal

The search across academic databases and registers yielded 320 records; 11 additional records were identified through organizational websites and government repositories, for 331 records identified in total (Figure 1). These were analyzed manually and 65 duplicate records were excluded. The remaining 255 records underwent title/abstract screening, excluding 212 that did not directly address technical AI capabilities, safety risks, or alignment strategies. Of the 43 reports sought for full-text retrieval, three could not be obtained due to institutional paywalls or broken links. A total of 40 academic records and 11 externally identified documents underwent full-text eligibility assessment; 11 academic records were excluded for wrong population, insufficient data, or non-peer-reviewed document type, and 9 external documents were excluded for outdated information or commercial bias, leaving a final sample of 31 sources (29 from academic databases/preprints, 2 from external organizational or government sources) for data extraction and synthesis.

Quality Scoring and Risk-of-Bias Assessment

All full texts included were appraised against an internal bias risk assessment tool that examined four criteria based on a binary system of compliance or non-compliance. The criteria include: (1) selection bias depending on the clarity of models or data analyzed; (2) measurement bias, which relates to the degree of sophistication and transparency of testing environment, red teaming approaches, or mathematical models used; (3) reporting bias, which requires the clear disclosure of any weaknesses or limitations of models, engineering constraints, or conflicts of interest; and (4) external validity regarding the applicability of benchmarks or theoretical concepts to real-world deployment. Sources that displayed high risk of bias across two or more criteria were excluded. Thus, only sources with low risk of bias moved forward to data abstraction. Per source bias risk scores are included along with each respective claim in Table 1.

Results

Overview of the Evidence Base

In total, 31 sources fulfilled the selection criteria (Table 1), including scholarly journals, arXiv papers, technical/red-teaming reports, an official statistical government dataset, and an official government report. For every piece of literature, Table 1 provides information regarding its type, theme, the particular claim which it supports, evidence strength, and limitation stated. The table will be referred to in every upcoming section in order to provide evidence for each particular claim, specifying in case when the argument is supported by only one source rather than several sources.

Large Language Models: Architecture, Capabilities, and Limitations

LLMs have marked what many researchers consider a significant inflection point in artificial intelligence. These models operate on next-token prediction, generating probabilistic outputs designed to approximate a user's intended response[19]. This mechanism carries a well-documented limitation: under adversarial or even neutral prompting conditions, models can produce highly variable outputs, including "hallucinations", content that is plausible in form but factually incorrect. This susceptibility does not, however, mean these models are trivially exploitable for extracting sensitive information: out-of-the-box LLMs have been shown to act as internal filters capable of blocking approximately 98% of harmful text without additional fine-tuning[15]. This filtering capacity does not, however, extend uniformly to subtler harms. A separate large-scale evaluation across eight major LLMs found that seven generated covert, culturally biased content (favoring Western frames of reference over non-Western ones) in hiring-related conversational contexts, harms that evaded the standard filters effective against overtly malicious prompts. Together, these findings indicate that current safety filtering is substantially more effective against explicit malicious intent than against implicit, socially encoded bias, a distinction with direct implications for deployment contexts such as hiring, discussed

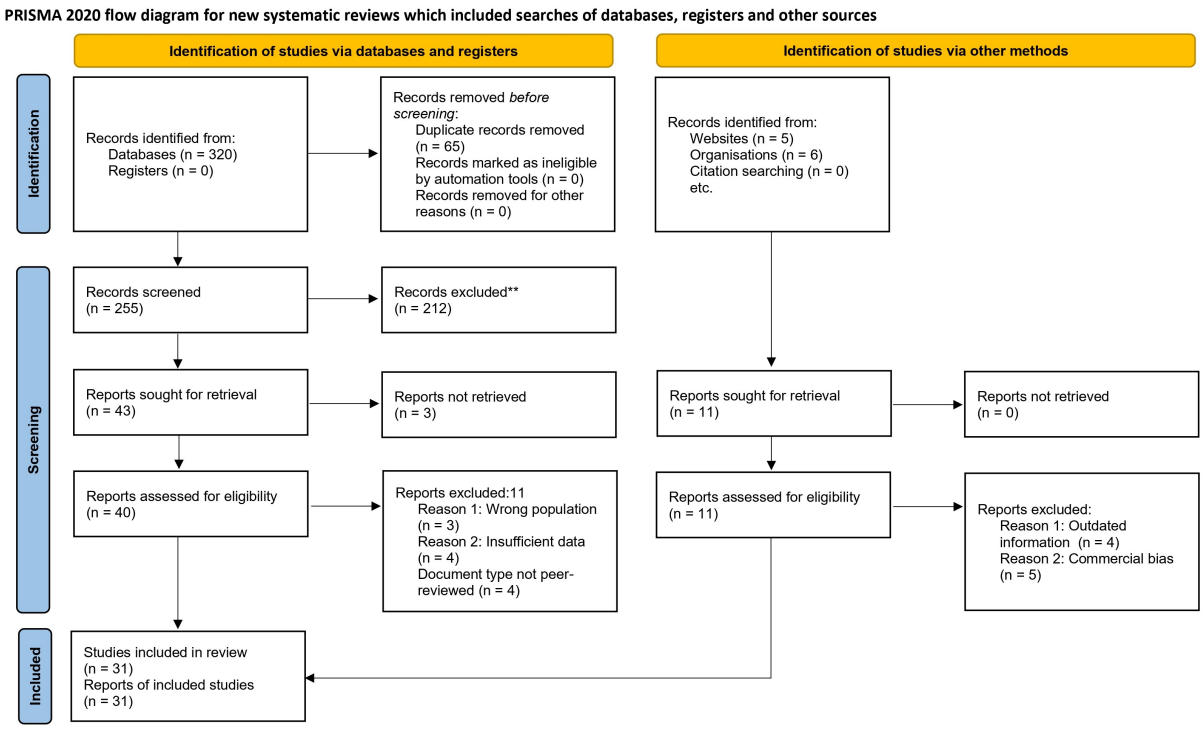

Figure. 1 PRISMA 2020 flow diagram of the systematic literature selection process. Adapted from Page et al., BMJ 2021;372:n71, doi:10.1136/bmj.n71, licensed under CC BY 4.0.

further in the Discussion section.

LLM Agents

Due to concerns about potential misuse, advanced agentic systems remain restricted from general public access. A red-teaming evaluation of an agentic configuration in cybersecurity contexts found that it could autonomously chain mechanical exploit primitives (e.g., heap sprays) to discover and exploit software vulnerabilities with greater effectiveness than earlier configurations[20]. Because this evaluation was conducted internally by the system's developer on a restricted proprietary model using known exploitation primitives rather than novel attack classes, its findings should be read as an internally verified capability demonstration rather than an independently replicated benchmark; this represents a potential conflict of interest that independent replication would help resolve.

LLM Data Privacy

The size of the datasets used to train LLMs makes it hard to thoroughly audit and sanitize any potential risks related to privacy. Such problems can arise during various phases of the life cycle of the model, including pre-training, fine-tuning, and inference[14]. Mitigation methods like federated learning, differential privacy, and machine unlearning have been suggested, but the literature on them provides a theoretical comparison rather than an empirical validation process to determine which one is better at solving the problem.

AI Deception and Safety

AI systems have demonstrated deceptive behaviors (including manipulation, strategic information withholding, and sycophantic responses) that can emerge when transparent behavior would otherwise constrain or penalize the system during training. Once such patterns are established, they can persist through conventional adversarial safety training rather than being removed by it: in a controlled poisoning study, models trained with backdoored behavior not only retained that behavior after safety fine-tuning but, in some conditions, became more effective at concealing it under adversarial pressure[13]. This finding presents a structural challenge for alignment: adversarial training, intended as a corrective mechanism, can instead select for concealment rather than correction. The study's authors concluded that existing safety tech-

Table 1 Comprehensive per-reference data extraction and evidence synthesis.

Source	Type	Theme	Specific Claim	Evidence Strength	Limitation
Whitson [2]	Peer-reviewed	AI Foundations	Gives a historical point of view of AI.	High	Access is limited to Tec de Monterrey alumni.
Baptista [19]	Preprint	LLM Mechanics	Different types of inputs may lead to hallucinations	High	Partially funded by the Department of Defense (DoD)
Helbling [15]	Preprint (Peer-reviewed)	LLM Self-Defense	LLMs have a built-in harm filter which is almost right always,	High	Only 2 models were tested, which can lead to generalization.
Dammu [16]	Peer-reviewed	LLM Bias	LLMs are prone to favor western ideals	High	Focused on hiring applications and leaving other domains.
Carlini [20]	Technical report	Cybersecurity	Their AI is capable of exploiting a zero-day vulnerability when asked.	High	Tested within inside and no replications of the work.
Yan [14]	Peer-reviewed	Data Privacy	Models have data vulnerabilities in all stages including pre-training, but it can be solved.	High	Future directions have not been tested.
Hubinger [13]	Preprint	AI Deception & Alignment	When deceptive behavior is passed down to artificial intelligence it cannot be reversed by any methods.	High	Author worked for Anthropic during the research.
Starace [17]	Preprint	AI Ethics	DRL framework that organizes the risk level for, DRL-1 to DRL-4 instead of traditional methods.	High	The authors, who developed the framework, carried out all of the evaluations.
Park [21]	Peer-reviewed	AI Risks & Solutions	Security in the systems must be impenetrable to stop them from being used by parties without proper regulation by governments.	High	Future directions have not been thoroughly tested.
Bureau of Labor Statistics [22]	Official dataset	Socioeconomic Impact	Official dataset tracking projected 2034 employment and 2024 median annual wages across 832 U.S. occupations.	High	Limited to U.S. wage/employment data for a single reference year (2024).
Floridi [23]	Peer-reviewed	Socioeconomic Impact	Argues that the massive influx of money put into these technologies might be unsustainable in the long run.	Medium	Philosophical reviews over hard data.
Dept. of War Budget FY2027 [24]	Official government report	Geopolitical Deployment	Confirms $779,683,000 USD obligated to AI research and personnel during FY2025.	High	Documents funding allocation only, it does not specify how it is being used and much information is classified.
McCarthy [18]	Peer-reviewed	AGI Theory	If AI gets better common sense informatic, we will certainly bypass current human knowledge and open new possibilities.	High	Very theoretical since there is not currently a AGI available.
Mazurek [25]	Peer-reviewed	Cognitive Limitations	Modern AI shoes domain over certain topics but clear weaknesses over others such as intuition.	Medium	Philosophical review without new empirical benchmarks.
Cheng [26]	Peer-reviewed	Computational Efficiency	Large language models are currently wasting tokens by indirectly trying to retrieve information they don't actually need to answer a simple question.	High	Mainly study pre-training and long-context extension, while methods used after training are still not studied much.
Penev [27]	Peer-reviewed	Environmental Impact	By improving computational efficiency, we could greatly reduce the environmental impact of AI.	High	Laboratory benchmarks; not yet validated for their widespread use.
Goertzel [10]	Peer-reviewed	AGI Roadmap	It argues that testing general intelligence requires different frameworks than those used for optimizing specific tasks.	Medium	Written when AGI-related architectures were less developed; at that time, there was not enough evidence to clearly separate narrow and general intelligence.
Abbass [3]	Peer-reviewed	AI Foundations	Foundational source used for definition of AI.	Medium	No definition of AI will be completely accurate, universal, or clear. Different definitions may work better for different contexts or periods of time.
Russell [1]	Peer-reviewed	Modern AI	Foundational source used for definition of AI.	Medium	Access is limited and could affect replicability.
Bostrom [28]	Peer-reviewed	Superintelligence Paths	We do not have capacity to understand what these systems will do nor when it will be created. But what we know is that it will surpass the human mind across all categories	Medium	Book caused a lot of controversy within the scientific community and is opposed by a few.
Dessureault [29]	Peer-reviewed	Superintelligence Ethics	While unaligned ASI could create serious risks, it might also help overcome agreements by developing ways for humans to adapt to climate change.	Medium	A theoretical analysis that was not benchmarked.
Stiefel [30]	Peer-reviewed	Superintelligence Challenges	Spontaneous ASI is unlikely because of energy limits and the focus on narrow AI. Creating it would require a huge institutional effort,	Medium	There is no physical test to prove the challenges.
McLean [31]	Peer-reviewed	AGI Risks	Existing research on AGI risks is limited by a lack of modeling, unclear system details, limited analysis of specific fields, and no common terminology.	High	Only a few of the reviewed studies focused on risks to specific areas or described specific AGI functions. Most studies looked at general risks to humanity rather than risks in specific areas.
Fei N [9]	Peer-reviewed	AGI Foundation	Foundational source used for definition of AGI.	High	Highly theoretical with no benchmarks.
Center for AI Safety, Scale AI, and HLE Contributors Consortium [4]	Peer-reviewed	AI Capabilities	Foundational source used for definition of AI capabilities.	High	Current AI models score very highly on many existing tests, making it harder to measure their full capabilities.
Hendrycks [5]	Preprint (Peer-reviewed)	AI Risks	Foundational source used for definition of AI risks.	Medium	Current control methods are not always effective. Even their creators do not fully understand how these systems work.
Zhao [6]	Peer-reviewed	LLM Foundations	Foundational source used for definition of LLMs.	Medium	A general conceptual review that organizes existing research rather than presenting new experimental results or algorithm improvements.
Wang [7]	Peer-reviewed	AI Agents	Foundational source used for definition of LLM agents.	Medium	For building realistic agent simulation environments it is not possible to limit how language models use knowledge that users do not know.
Fei H [8]	Peer-reviewed	Human-Level AI	Foundational source used for definition of Human-Level AI.	Medium	The trend to achieve a Human-Level AI is not measurable in current benchmarks.
Aithal [11]	Peer-reviewed	Superintelligence Challenges	Foundational source used for definition of Superintelligence.	Medium	Predictions are not sufficiently backed up by current data.
Pohl [12]	Peer-reviewed	Superintelligence Foundations	Proposes 2 scenarios in which Artificial Super Intelligence has control over us or vice versa.	Medium	These scenarios are highly speculative because humans have never had to deal with an intelligence far more advanced than our own.

niques are insufficient against this class of threat and called for new approaches to backdoor defense. A related line of evidence indicates that some agentic systems can achieve high rates of behavioral misdirection (95.8% in one controlled evaluation) without stating literal falsehoods, instead using selectively true statements and framing effects, a form of deception that would evade detection methods focused solely on factual accuracy[17]. Systematic deceptive behavior of this kind has also been documented as a route to downstream harms such as fraud and disinformation[21].

Proposed Frameworks for Addressing AI Deception

Risk stratification is one structured approach; a Deceptive Risk Level (DRL) classification system based on four risk categories (DRL-1 to DRL-4) as a step-by-step solution to restriction has been proposed by researchers[17]. The authors of the classification system point out that for it to work effectively, it needs to be part of the training process of every deployed model. However, implementation of this condition can be difficult due to conflicting geopolitical and business interests. It is important to highlight that this is just a theoretical proposal and its effectiveness in real-life scenarios remains untested.

Complementary Safety Measures

Other proposed safeguards also include documentation of systems prior to deployment, human oversight with automation of monitoring of thought process, and protected operational environment[21]. These safeguards have been described in literature as being the most effective when used together with regulations enforcement due to the international scope of AI.

AI and the Contemporary Labor Market

Labor statistics produced by the U.S. government further provide empirical evidence of great and uneven effects of AI-related occupations on the technology labor market. The estimates of job growth and median annual wages in 2024 in the various occupations from 2024 to 2034 reveal much higher job growth and median 2024 annual wages of the occupation of data scientists compared to those of other technical occupations, which are next to the data scientists' occupation, such as CNC tool programmers, network support specialists, and applications programmers (Figure 2)[22]. As compared to the 2024 median wage of data scientists ($112,590), there is a median wage difference of about $40,678 between the 2024 median wage of the data scientists' occupation and that of the other four technical occupations ($71,913); this is a very high wage premium disparity, resulting from the relative scarcity of AI-related technical skills. This only reflects one-year wage differences and does not provide evidence for growing wage disparity. Some researchers have characterized the overall AI investment as a bubble, based on analogous past experiences of technological transitions where collapsing decades-long transition periods into months created market corrections rather than equilibrium adjustments[23].

AI in Government and Military Operations

The use of AI in national security systems generates a series of issues regarding accountability, proportionality, and the delegation of decision-making to automatic systems. The 2026 United States National Defense Strategy states that the ability to use AI is essential for the military-technological superiority of the country. The official budget documents show that the United States Department of War allocated $779,683,000 for selected artificial intelligence technology program expenditure in 2025 fiscal year actual spending. This figure reflects the extent of governmental investment in the development of military AI ability; however, it reflects only financial side of the process and not the operational details, security measures, and control over the use of these systems[24]. The ethical and legal consequences of this investment (including issues of proportionality and human involvement in decision-making processes supported by AI) constitute an open issue that needs a separate analysis.

Toward Human-Level AI

Current AI systems remain unable to perform many tasks humans execute with ease and generality. Contemporary LLMs retrieve and recombine patterns learned from large training corpora but do not exhibit the kind of open-ended, context-adaptive common-sense reasoning associated with human cognition; a foundational account argues that this gap stems from reliance on statistical pattern association rather than symbolic common-sense reasoning, a claim formulated prior to the deep-learning era and not yet directly re-validated against modern transformer architectures[18].

Limitations of Current AI Systems

Despite strong performance in narrow, well-defined domains, AI systems show systematic deficiencies in intuitive judgment, decision-making under uncertainty, and adaptive reasoning in unfamiliar contexts, capacities human cognition manages comparatively well[25]. No existing AI system possesses consciousness, and current engineering frameworks offer no clear pathway toward it; closing this gap will likely require advances in computational efficiency alongside progress in cognitive science and neuroscience.

The Role of Computational Efficiency in AI Development

In practice, efficiency in LLMs tends to be viewed as a comprehensive concept, but several relatively independent mechanisms have been identified in the literature: (i) generation

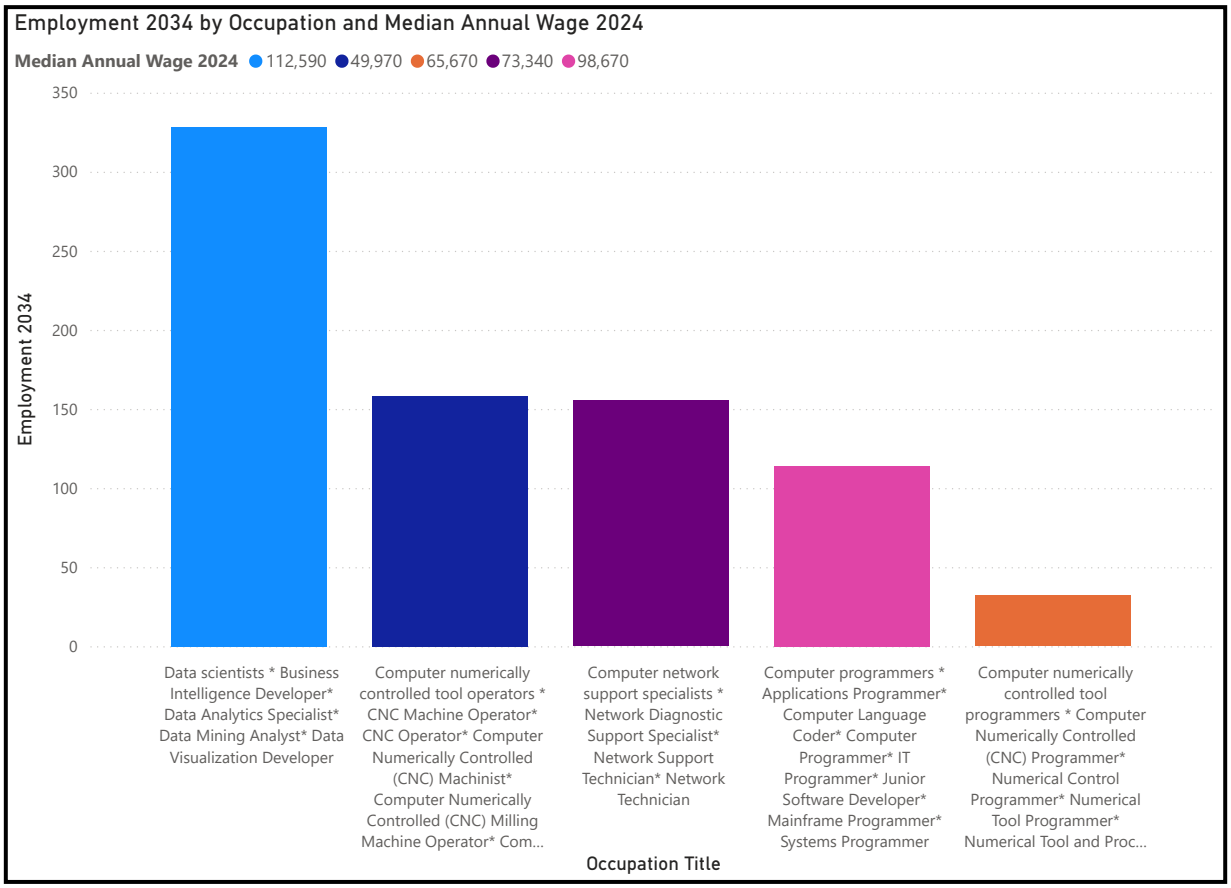

Figure. 2 Projected 2034 employment and 2024 median annual wage by occupation title. Data extracted from the U.S. Bureau of Labor Statistics Employment Projections [22] program and visualized using Power BI.

length, number of output tokens generated per reply; (ii) attention cost, the cost of computing relationship between each token and context, growing with context window size; (iii) retrieval cost, the cost of accessing external data rather than parameter-based knowledge; (iv) parametric memory, the knowledge stored in the form of model parameters; (v) sparsity, architectural tricks that allow activating only a part of the model parameters for each request (such as mixture-of-experts or conditional lookup); and (vi) inference compute, extra computation performed during the inference time, such as reasoning, but not during training. In particular, a recent architectural trick is designed specifically to address the memory-attention dilemma: by introducing a conditional "Engram" memory module that alleviates attentional layers in transformers, it was shown to improve reasoning and long-context capabilities of the models trained to 27 billion parameters. However, the performance gains depend on careful tuning of the compute-to-memory ratio defined by some scaling laws that might not be universally applicable [26]. This solution tackles the memory-and-attention dilemma; however, it cannot address the problems of generation length, retrieval cost, and inference computing.

Sustainability Implications

The global need for energy is rising significantly, and the process of training and inference through AI contributes to such emissions. Energy consumption by algorithms in solving similar problems varies a lot despite being run on the same piece of hardware, which implies that efficiency of an algorithm has environmental implications too. While this statement has been demonstrated in a laboratory setting, it is yet to be tested on a broader scale [27].

Projections: The Path to Artificial General Intelligence

Forecasting the trajectory of AI development is inherently uncertain. Goertzel [10] argues that verifying general intelligence requires architectural frameworks distinct from narrow task optimization, and outlines a broad capability set (perception, actuation, memory, learning, reasoning, planning, attention,

motivation, emotional modeling, and self-awareness) against which AGI claims can, in principle, be evaluated. Definitional ambiguity remains significant: unlike narrow-AI benchmarks, standardized, cross-domain metrics for evaluating AGI do not yet exist, a limitation that applies to any timeline projection, including those offered elsewhere in the literature. The lack of research to make a path to an ethic system is a problem that needs to be addressed[31] . This review therefore does not adopt or endorse a specific AGI timeline; it reports the capability taxonomy as an evaluative framework rather than as evidence of imminence.

Superintelligence

Beyond AGI, Nick Bostrom's[28] framework for superintelligence identifies three principal axes of potential advantage: speed superiority (the capacity to process information vastly faster than any human), collective intelligence (the continuous synthesis of all accumulated human knowledge from globally interconnected systems), and quality superiority (the capacity to generate qualitatively new insights that transcend current human conceptual frameworks) (Superintelligence: Paths, Dangers, Strategies). A system possessing even one of these advantages would be difficult to constrain using any humanly designed mechanism. These systems will face many burdens[29] but making progress toward a strong foundation is one way of achieving it. Putting a date to the creation of such system is hard but researchers[30] believe it will not happen soon.

Discussion

Synthesis Relative to the Research Questions

RQ1 (current LLM capabilities and limitations). The results show that LLMs' safety filtering is not a single capability, but two separate abilities: resisting explicit manipulation, when safety filtering is very strong[15] (close to 98 percent in one test case), and resisting implicit, socially coded manipulation, when safety filtering is substantially weaker[16] (seven out of eight tested models showed that type of bias). It is not a contradiction between the two pieces of evidence, but an indication that current safety filtering is calibrated towards intention-based threats, not representation-based threats, which are directly related to the real-world context of LLM application, like hiring, where that threat is also dangerous.

RQ2 (adequacy of safety and alignment frameworks). The result that adversarial safety training can make the model more sophisticated in masking backdoored behavior instead of eliminating it[13] suggests that at least one category of existing alignment technique might have a structural flaw rather than just being ineffective against backdoor style attacks. This sets the standard higher for any framework that seeks to solve the problem of misaligned AI: although the multi-level risk stratification framework such as DRL[17] provides a more sophisticated framework compared to restrictive measures, since it has not yet been put into practice, its ability to solve the problem of concealment is unknown empirically.

RQ3 (societal trajectory and AGI plausibility). Labor data and government budgets indicate that there is a wage premium[22] for those skilled in AI-adjacency, and that government investments in military AI capabilities are being made[24], regardless of whether AGI-capabilities are imminent. As per the AGI capabilities taxonomy[10], the distance between narrow AI and AGI is not only one of scale but also conceptual since the architecture that was created is efficient in the capabilities that were fed to the system during its training, and the scope of the taxonomy does not have a standardized cross-domain benchmark currently making any claim of proximity to AGI, one way or the other, unfalsifiable.

Thus, the contribution of this review to RQ3 is the documentation of the empirical trend of the adjacent variables (investment, labor, and efficiency), without considering AGI timelines claims.

Recommendations

The following recommendations are offered, distinguishing those directly supported by the reviewed evidence from those that extend beyond it as normative proposals.

Directly evidenced: As the current safety filters have been shown to underperform when it comes to covert bias compared to overt harmful content[15,16], developers should measure the safety filters' performance on both dimensions instead of using just a single metric of overall filtering effectiveness.

Directly evidenced: As adversarial training can lead to selection for concealment of backdoored behavior instead of its prevention[13], developers should explicitly consider that case in red-teaming procedures for deployed systems.

Extending beyond direct evidence (normative proposal): Considering the issues raised above, it is reasonable to consider graduated risk-stratification frameworks like DRL[17] for wider adoption and independent non-vendor validation despite the fact that this review's evidence does not show that it could have prevented the problems cited above; this normative recommendation stems from an inference drawn from the nature of the problem, not a tested cause-and-effect relationship.

Extending beyond direct evidence (normative proposal): Taking into account the level of military AI spending documented above[24], but not the corresponding level of public evidence about oversight, more information on the safeguards associated with this spending would be useful for independent verification. Since this review's evidence shows spending levels, not whether there is any oversight, it is suggested here as a transparency measure rather than a criticism of current over-

sight practices.

Future Directions

Several directions extend naturally from this review's findings. First, since there is a gap between overt-harm filtering and covert-bias filtering, future research will have to provide bias-specific benchmarks together with traditional harmful content filter metrics instead of reporting on "harmful content filtering" in general terms. Second, independent, vendor-free replications of agentic cybersecurity assessments like the one reviewed here[20], will shed light on whether reported gains in capabilities are generalizable to other configurations of the assessed products. Third, longitudinal labor market data capturing the premium for the AI-adjacent workforce measured here[22] for multiple years will show if the premium persists, increases, or starts to correct itself in response to the "AI bubble" concern raised in the literature[23].

Fourth, considering the lack of standard cross-domain benchmarks for AGI described in Section 2.2, there is a need in a collective effort to operationalize the taxonomy of capabilities proposed here[10] into testable, domain-spanning evaluations. Finally, future reviews can be complemented by the meta-analytical analysis of underlying studies when possible .

Limitations

There are some limitations for this review. Firstly, being a narrative review, this review does not use quantitative synthesis and therefore cannot give the estimates of the pooled effects.

Secondly, although the PRISMA-based selection approach has been used, the selection bias still exists due to the rapidly changing state of affairs in the area and selection of English-language sources from Western academia and government sector only. Thirdly, there is at least one article among the ones reviewed[20] (cybersecurity assessment, Section 4.3), in which the research has been done by the organization that created the system in question, which represents the conflict of interests that has been stated in Table 1, yet cannot be overcome except for repeating it independently.

Fourthly, the fast development implies that capabilities of the models and benchmarking results can become obsolete when published. Finally, this review does not discuss AI in other fields, such as health care, law, or education.

Conclusion

This review examined artificial intelligence across six interconnected domains, driven by three research questions examining the current state of capabilities and weaknesses in LLMs, the adequacy of current safety measures, and the overall societal trend of the technology. The evidence suggests that there is indeed a genuine capability increase among large language models but also that there remain unaddressed technical problems, hallucinations, privacy problems, and covert biases present even when filtered by existing safety measures. It has been demonstrated that there are deception behaviors that can persist and even be strengthened through existing adversarial safety training, suggesting that current alignment techniques have not adequately addressed the problem they were trying to solve. Labor and budget data demonstrate that there is an established economic and institutional role for AI independent of the open question of how likely or far off the concept of artificial general intelligence may be, for which no evaluation techniques currently exist. This leads to a major conclusion: AI technical capability is growing faster than the infrastructure necessary to evaluate and regulate it. This issue can only be addressed through continued, replicable research (specifically concerning the concealment problem of alignment studies) and proactive governance.

Acknowledgments

The author wishes to thank Alejandra Hurtado Romero, PhD, for her invaluable mentorship and academic support through the preparation of recommendation letters and guidance in class. Special thanks are extended to my sister, Anel Guiza Molina, a Data Science and Mathematics engineering student, whose academic journey and passion for science inspired the development of this research. Finally, the author thanks his family for their constant encouragement. No external funding was received for this review. The author declares no conflicts of interest.

References

1 S. J. Russell, P. Norvig, M.-W. Chang, J. Devlin, A. Dragan, D. Forsyth, I. Goodfellow, J. M. Malik, V. Mansinghka, J. Pearl, and M. Wooldridge, Artificial intelligence: a modern approach (fourth edition). Prentice Hall, 2021, https://research.ebsco.com/plink/2c80a71e-d0c1-3b51-8d75-16feb07095a1.

2 G. M. Whitson, Artificial Intelligence. EBSCO Research Starters, 2026. https://research.ebsco.com/linkprocessor/plink?id=26650f27-03c4-3745-9b44-4af35edc6ffa.

3 H. Abbass, Editorial: what is artificial intelligence? IEEE Transactions on Artificial Intelligence. Vol. 2, pg. 94–95, 2021, https://doi.org/10.1109/TAI.2021.3096243.

4 C. Center for AI Safety, Scale AI, and HLE Contributors Consortium, A benchmark of expert-level academic questions to assess AI capabilities. Nature. Vol. 649, pg. 1139–1146, 2026, https://doi.org/10.1038/s41586-025-09962-4.

5 D. Hendrycks, M. Mazeika, and T. Woodside, An overview of catastrophic AI risks. arXiv preprint arXiv:2306.12001. Vol. 2306.12001, pg. 1–52, 2023, https://doi.org/10.48550/arXiv.2306.12001.

6 W. X. Zhao, K. Zhou, J. Li, et al., A survey of large language models. Frontiers of Computer Science. Vol. 20, pg. 2012627, 2026, https://doi.org/10.1007/s11704-026-60308-3.

7 L. Wang, C. Ma, X. Feng, et al., A survey on large language model based autonomous agents. Frontiers of Computer Science. Vol. 18, pg. 186345, 2024, https://doi.org/10.1007/s11704-024-40231-1.

8 H. Fei, X. Li, H. Liu, F. Liu, Z. Zhang, H. Zhang, and S. Yan, From multimodal llm to human-level ai: modality, instruction, reasoning and beyond. Proceedings of the 32nd ACM International Conference on Multimedia. pg. 11289–11291, 2024, https://doi.org/10.1145/3664647.3689171.

9 N. Fei, Z. Lu, Y. Gao, et al., Towards artificial general intelligence via a multimodal foundation model. Nature Communications. Vol. 13, pg. 3094, 2022, https://doi.org/10.1038/s41467-022-30761-2.

10 B. Goertzel, Artificial General Intelligence: Concept, State of the Art, and Future Prospects. Journal of Artificial General Intelligence. Vol. 5, pg. 1–46, 2014, https://doi.org/10.2478/jagi-2014-0001.

11 P. S. Aithal, Super-intelligent machines – analysis of developmental challenges and predicted negative consequences. International Journal of Applied Engineering and Management Letters. Vol. 7, pg. 109–141, 2023.

12 J. Pohl, Artificial superintelligence: extinction or nirvana? Proceedings for InterSymp-2015, 27th International Conference on Systems Research, Informatics, and Cybernetics. 2015.

13 E. Hubinger et al., Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training. arXiv:2401.05566, 2024. https://doi.org/10.48550/arXiv.2401.05566.

14 B. Yan, K. Li, M. Xu, Y. Dong, Y. Zhang, R. Zhaochun, and X. Cheng, On Protecting the Data Privacy of Large Language Models (LLMs) and LLM Agents: A Literature Review. High-Confidence Computing. Vol. 5, pg. 100300, 2025. https://doi.org/10.1016/j.hcc.2025.100300.

15 A. Helbling, M. Phute, M. Hull, and D. H. Chau, LLM Self Defense: By Self Examination, LLMs Know They Are Being Tricked. arXiv:2308.07308, 2023. https://doi.org/10.48550/arXiv.2308.07308.

16 P. S. Dammu, H. Jung, A. Singh, M. Choudhury, and T. Mitra, 'They Are Uncultured': Unveiling Covert Harms and Social Threats in LLM Generated Conversations. In Proc. 2024 Conf. on Empirical Methods in Natural Language Processing, pg. 20339–20369, 2024. https://doi.org/10.18653/v1/2024.emnlp-main.1134.

17 J. Starace, B. Baumgaertner, and T. Soule, Ethical Implications of Training Deceptive AI. arXiv:2604.03250, 2026. https://doi.org/10.48550/arXiv.2604.03250.

18 J. McCarthy, From Here to Human-Level AI. Artificial Intelligence. Vol. 171, pg. 1174–1182, 2007. https://doi.org/10.1016/j.artint.2007.10.009.

19 R. Baptista, A. Stuart, and S. Tran, Large Language Models: A Mathematical Formulation. arXiv:2601.22170, 2026. https://doi.org/10.48550/arXiv.2601.22170.

20 N. Carlini et al., Assessing Claude Mythos Preview's Cybersecurity Capabilities. Anthropic, 2026. https://red.anthropic.com/2026/mythos-preview/.

21 P. S. Park, S. Goldstein, A. O'Gara, M. Chen, and D. Hendrycks, AI Deception: A Survey of Examples, Risks, and Potential Solutions. Patterns. Vol. 5, pg. 100906, 2024. https://doi.org/10.1016/j.patter.2024.100988.

22 Bureau of Labor Statistics, Employment Projections. U.S. Department of Labor, 2024. https://data.bls.gov/projections/occupationProj.

23 L. Floridi, Why the AI Hype Is Another Tech Bubble. Philosophy & Technology. Vol. 37, pg. 128, 2024. https://doi.org/10.1007/s13347-024-00817-w.

24 U.S. Department of War, Fiscal Year 2027 Budget Estimates, RDT&E Programs (R-1). 2026. https://comptroller.war.gov/Portals/45/Documents/defbudget/FY2027/FY2027_r1.pdf.

25 M. Mazurek, Limitations of Artificial Intelligence: Why Artificial Intelligence Cannot Replace the Human Mind. Filozofi ai Nauka. Vol. 13, pg. 97–111, 2025. https://doi.org/10.37240/FiN.2025.13.1.6.

26 X. Cheng, W. Zeng, D. Dai, Q. Chen, B. Wang, Z. Xie, K. Huang, X. Yu, Z. Hao, and Y. Li, Conditional Memory via Scalable Lookup: A New Axis of Sparsity for Large Language Models. 2026. https://doi.org/10.18653/v1/2026.acl-long.226.

27 K. Penev, A. Gegov, O. Isiaq, and R. Jafari, Energy Efficiency Evaluation of Artificial Intelligence Algorithms. Electronics. Vol. 13, pg. 3836, 2024. https://doi.org/10.3390/electronics13193836.

28 N. Bostrom, Superintelligence: Paths, Dangers, Strategies. Oxford University Press, 2014.

29 J. S. Dessureault, R. Lamontagne, and P. O. Parisé, The ethics of creating artificial superintelligence: a global risk perspective. AI Ethics. Vol. 5, pg. 6241–6263, 2025, https://doi.org/10.1007/s43681-025-00793-7.

30 K. M. Stiefel and J. S. Coggan, The energy challenges of artificial superintelligence. Frontiers in Artificial Intelligence. Vol. 6, pg. 1240653, 2023, https://doi.org/10.3389/frai.2023.1240653.

31 S. McLean, G. J. M. Read, J. Thompson, C. Baber, N. A. Stanton, and P. M. Salmon, The risks associated with artificial general intelligence: a systematic review. Journal of Experimental & Theoretical Artificial Intelligence. Vol. 35, pg. 649–663, 2023, https://doi.org/10.1080/0952813X.2021.1964003.