American Journal of Artificial Intelligence and Neural Networks

[

australiansciencejournals.com/ajainn

E-ISSN: 2688-1950

VOL 06 ISSUE 04 2025

Large Language Models as General Purpose Intelligence Systems for Reasoning, Planning and Decision Making

Fengyuan Zhang[1], Bi Wu[2]

1New York University, New York, USA

2University of California, Los Angeles, USA

Corresponding author: *Fengyuan Zhang. fz2240@nyu.edu*

Abstract: *The emergence of large language models (LLMs) has fundamentally transformed artificial intelligence (AI) research and applications, positioning these systems as potential candidates for general purpose intelligence. Large language models are deep neural networks trained on massive text corpora that demonstrate remarkable capabilities across diverse cognitive tasks without task-specific fine-tuning. This review examines how LLMs function as general intelligence systems, with particular emphasis on three core cognitive domains: reasoning, planning, and decision making. We analyze the architectural foundations that enable LLMs to perform complex reasoning tasks, including chain-of-thought prompting (CoT), in-context learning (ICL), and emergent abilities that arise from scale. The planning capabilities of LLMs are evaluated through their performance on multi-step problem decomposition, goal-oriented task completion, and strategic action sequencing. Furthermore, we investigate decision-making frameworks where LLMs serve as autonomous agents, policy advisors, and collaborative systems that integrate human expertise with machine intelligence. The review synthesizes recent advances in prompt engineering, retrieval-augmented generation (RAG), and multimodal integration that enhance LLM capabilities for general intelligence tasks. We examine real-world applications spanning healthcare diagnosis, financial analysis, scientific discovery, and autonomous systems management. Critical challenges including hallucination, reasoning consistency, computational efficiency, and ethical considerations are thoroughly discussed. This* *comprehensive analysis demonstrates that while LLMs exhibit significant progress toward general purpose intelligence, fundamental limitations in causal understanding, long-term planning coherence, and adaptive learning remain open research challenges that require continued innovation in architecture design, training methodologies, and evaluation frameworks.*

Keywords: *Large language models, general purpose intelligence, reasoning systems, planning algorithms, decision making, artificial intelligence, cognitive capabilities, chain-of-thought, in-context learning, prompt engineering*

INTRODUCTION

The pursuit of artificial general intelligence (AGI) has been a central aspiration of artificial intelligence (AI) research since the field's inception, representing the goal of creating systems that can match or exceed human cognitive capabilities across diverse domains. Recent advances in large language models (LLMs) have catalyzed renewed optimism and intense debate regarding the feasibility of achieving general purpose intelligence through scaled neural architectures. Large language models, exemplified by systems such as Generative Pre-trained Transformer (GPT) series, PaLM, Claude, and LLaMA, have demonstrated unprecedented capabilities in natural language processing (NLP) tasks while exhibiting emergent behaviors that extend far beyond their initial training objectives [1]. These systems, trained on trillions of tokens from diverse text sources, have shown remarkable proficiency in tasks requiring reasoning, planning, and decision making—cognitive functions traditionally considered hallmarks of human intelligence.

The transformer architecture provides the foundational framework for modern LLMs through its self-attention mechanism that enables parallel processing of sequential data and long-range dependency modeling [2]. The scaling hypothesis, which posits that increasing model parameters, training data, and computational resources leads to qualitative improvements in capabilities, has been empirically validated across multiple dimensions of performance [3]. Models with billions to trillions of parameters have exhibited abilities that were absent in smaller counterparts, including in-context learning (ICL) where models adapt to new tasks from examples provided in the prompt without parameter updates [4]. This emergent capability suggests that LLMs may possess latent general intelligence that manifests when sufficient scale is achieved.

Reasoning represents a fundamental cognitive capability that distinguishes intelligent systems from simple pattern matchers. LLMs have demonstrated various reasoning modalities including logical inference, mathematical problem-solving, commonsense reasoning, and analogical thinking [5]. The introduction of chain-of-thought (CoT) prompting techniques, which encourage models to articulate intermediate reasoning steps, has significantly enhanced performance on complex reasoning tasks [6]. Studies have shown that CoT prompting enables LLMs to solve multi-hop reasoning problems that require synthesizing information across multiple sources and performing sequential logical operations. However, the extent to which these reasoning capabilities reflect genuine understanding versus sophisticated pattern matching remains a subject of ongoing investigation and debate within the research community.

Planning involves the decomposition of complex goals into actionable sequences of steps, temporal reasoning about action consequences, and adaptive strategy formulation in dynamic environments. Recent research has explored LLMs as planning engines for autonomous agents, demonstrating their ability to generate coherent action sequences in domains ranging from household robotics to software development [7]. The integration of LLMs with external tools and knowledge bases through retrieval-augmented generation (RAG) frameworks has substantially expanded their planning capabilities by providing access to up-to-date information and domain-specific expertise [8]. These hybrid architectures combine the generative flexibility of LLMs with the precision and reliability of structured knowledge systems, creating more robust planning mechanisms.

Decision making in the context of LLMs encompasses both individual task execution and collaborative human-machine systems where models serve as advisors, analysts, or autonomous agents. Research has investigated how LLMs can support medical diagnosis by synthesizing patient information, suggesting differential diagnoses, and recommending treatment pathways [9]. In financial domains, LLMs have been deployed for market analysis, risk assessment, and investment strategy formulation, demonstrating competitive performance with traditional analytical models [10]. The ability of LLMs to process vast amounts of unstructured text data, identify relevant patterns, and generate actionable recommendations positions them as valuable decision support tools across professional domains.

Despite these impressive capabilities, significant challenges remain before LLMs can be considered true general purpose intelligence systems. Hallucination, the generation of plausible but factually incorrect information, represents a critical limitation that undermines reliability in high-stakes applications [11]. Reasoning consistency varies substantially across problem formulations, with models sometimes failing on simple variants of problems they previously solved correctly. Computational costs associated with training and inference limit accessibility and raise sustainability concerns. Ethical considerations including bias amplification, privacy violations, and potential misuse require careful governance frameworks.

This review provides a comprehensive analysis of LLMs as general purpose intelligence systems, focusing specifically on their capabilities and limitations in reasoning, planning, and decision making. We synthesize recent empirical findings, theoretical frameworks, and practical applications to assess the current state of LLM-based general intelligence. The review is organized into six main sections beyond this introduction. The following section presents a thorough literature review examining the evolution of LLMs and their positioning within the broader AGI landscape. We then analyze reasoning capabilities across multiple cognitive domains, investigate planning mechanisms and multi-step task execution, and explore decision-making frameworks with real-world applications. Finally, we discuss critical challenges and limitations before synthesizing key findings in the conclusion.

2. Literature Review

The conceptual foundations of general purpose intelligence in artificial systems trace back to early AI research, where pioneers envisioned machines capable of flexible problem-solving across arbitrary domains without specialized programming. Classical approaches to AGI emphasized symbolic reasoning, knowledge representation, and explicit rule systems that could manipulate abstract concepts according to logical principles [12]. These systems demonstrated competence in constrained domains such as chess playing and mathematical theorem proving but struggled with the flexibility and robustness characteristic of human intelligence. The knowledge acquisition bottleneck, wherein manual encoding of domain expertise proved prohibitively labor-intensive, limited the scalability of symbolic AI approaches.

The connectionist revolution introduced neural networks as an alternative paradigm emphasizing distributed representations and learning from data rather than explicit programming [13]. Early neural network architectures demonstrated impressive pattern recognition capabilities but were limited by computational resources, training algorithms, and data availability. The development of backpropagation enabled efficient training of multi-layer networks, while convolutional neural networks (CNNs) achieved breakthrough performance on visual recognition tasks [14]. However, these systems remained largely specialized to specific sensory modalities or task domains, falling short of the flexibility required for general intelligence.

The introduction of the transformer architecture marked a pivotal moment in the trajectory toward general purpose AI systems. Unlike recurrent neural networks (RNNs) that process sequences step-by-step, transformers employ self-attention mechanisms that enable parallel computation and direct modeling of long-range dependencies. The attention mechanism computes weighted combinations of input representations based on learned relevance scores, allowing models to focus on pertinent information regardless of sequential distance [15]. This architectural innovation enabled training of substantially larger models on more extensive datasets, leading to qualitative improvements in language understanding and generation capabilities.

The scaling laws governing LLM performance have been extensively studied, revealing predictable relationships between model size, dataset size, computational budget, and downstream task performance. Research has demonstrated that increasing any of these factors while holding others constant leads to systematic improvements, with optimal resource allocation depending on specific deployment constraints [16]. These scaling laws suggest that current architectural paradigms have not yet reached fundamental performance ceilings, implying continued improvement potential through increased scale. However, recent work has questioned whether scaling alone is sufficient for achieving general intelligence, highlighting persistent limitations in causal reasoning, systematic generalization, and grounded understanding [17].

Pre-training objectives play a crucial role in shaping LLM capabilities and their potential for general intelligence. Masked language modeling, used in models like BERT, trains systems to predict missing tokens based on surrounding context, encouraging bidirectional understanding of linguistic structure. Causal language modeling, employed in GPT-series models, trains systems to predict subsequent tokens given preceding context, optimizing for coherent text generation [18]. Recent research has explored alternative pre-training objectives including contrastive learning, denoising autoencoders, and multi-task learning that may better align with general intelligence requirements. The choice of pre-training objective influences the types of knowledge and capabilities that emerge during training, with implications for downstream reasoning and planning performance.

In-context learning represents a remarkable emergent capability of large-scale models wherein they adapt to new tasks based solely on examples provided in the input prompt without parameter updates. This phenomenon suggests that LLMs develop internal mechanisms for task identification, pattern recognition, and rule inference during pre-training that generalize to novel situations [19]. Research has investigated the relationship between model scale and ICL performance, finding that this capability emerges primarily in models exceeding certain parameter thresholds. The mechanisms underlying ICL remain incompletely understood, with competing theories emphasizing either implicit meta-learning during pre-training or direct pattern matching against memorized training examples.

Chain-of-thought prompting has emerged as a powerful technique for enhancing reasoning capabilities by encouraging models to articulate intermediate steps in problem-solving processes. Studies have demonstrated that CoT prompting substantially improves performance on mathematical reasoning, logical inference, and multi-hop question answering tasks. The effectiveness of CoT appears to depend on model scale, with larger models benefiting more from explicit reasoning articulation [20]. Variations of CoT including self-consistency methods, which sample multiple reasoning paths and select the most consistent answer, have further improved reliability. These prompting techniques represent a form of soft programming that guides model behavior without retraining, enabling rapid adaptation to diverse cognitive tasks.

Retrieval-augmented generation frameworks address fundamental limitations of purely parametric models by integrating external knowledge sources into the generation process. RAG systems retrieve relevant documents from large corpora based on input queries, then condition language generation on retrieved content alongside the original prompt. This architecture combines the generative flexibility of LLMs with the precision and updateability of explicit knowledge bases, reducing hallucination while enabling access to specialized or current information [21]. Research has explored various retrieval strategies, indexing methods, and integration approaches that optimize the trade-off between computational efficiency and knowledge coverage.

Multimodal LLMs extend language-centric architectures to process and generate content across multiple modalities including vision, audio, and structured data. These systems employ shared representation spaces where different modalities are embedded into common semantic spaces, enabling cross-modal reasoning and generation [22]. Vision-language models have demonstrated capabilities in image captioning, visual question answering, and text-to-image generation that suggest unified understanding of perceptual and linguistic information. The integration of multiple modalities may be essential for achieving general intelligence, as human cognition fundamentally operates across sensory domains with rich cross-modal associations.

The development of LLM-based autonomous agents represents a significant step toward operational general intelligence systems that can interact with environments, execute tasks, and pursue goals. These agents employ LLMs as reasoning engines that plan actions, invoke tools, and adapt strategies based on environmental feedback [23]. Research has demonstrated agent capabilities in domains including web navigation, software development, scientific experimentation, and household robotics. The agent paradigm transforms LLMs from passive question-answering systems into active problem-solvers that can decompose objectives, execute sub-tasks, and integrate results toward goal achievement.

Evaluation methodologies for general intelligence capabilities remain a critical research challenge, as traditional NLP benchmarks focus on narrow task performance rather than flexible problem-solving. Recent work has proposed comprehensive evaluation frameworks that assess multiple dimensions of intelligence including reasoning, planning, learning, knowledge, and robustness [24]. Benchmarks such as BIG-Bench, HELM, and AGI Eval provide diverse task sets designed to probe general capabilities rather than memorization of specific patterns. However, concerns about benchmark contamination, where training data includes test examples, complicate interpretation of results and necessitate careful dataset curation.

Theoretical perspectives on whether LLMs can achieve genuine understanding or merely simulate intelligence through statistical pattern matching remain contentious. Some researchers argue that the scale and architectural sophistication of modern LLMs enable emergent understanding that, while mechanistically different from human cognition, constitutes a valid form of intelligence [25]. Alternative views maintain that LLMs lack grounded meaning, causal models, and genuine intentionality required for understanding, arguing that their success reflects exploitation of statistical regularities rather than comprehension. This debate carries implications for expectations regarding LLM capabilities, appropriate application domains, and future research directions.

3. Reasoning Capabilities of Large Language Models

Reasoning encompasses the cognitive processes by which intelligent systems derive new knowledge from existing information through systematic inference, logical deduction, pattern recognition, and analogical thinking. LLMs have demonstrated reasoning capabilities across multiple domains that were previously thought to require specialized symbolic systems or extensive task-specific training. The emergence of these capabilities from models trained primarily on next-token prediction objectives has surprised many researchers and prompted intensive investigation into the mechanisms underlying LLM reasoning.

Mathematical reasoning represents a particularly challenging domain for neural models due to its requirement for precise symbolic manipulation, multi-step derivation, and strict logical consistency. Early language models struggled with even basic arithmetic, frequently producing incorrect answers to simple calculation problems. However, recent LLMs exhibit substantially improved mathematical capabilities, successfully solving problems from standardized tests, competition mathematics, and undergraduate-level coursework [26]. The integration of CoT prompting has been especially impactful for mathematical reasoning, with models articulating step-by-step solutions that mirror human problem-solving approaches. Research has shown that encouraging models to show their work significantly improves accuracy on word problems, algebraic manipulations, and geometric proofs.

Despite these advances, mathematical reasoning in LLMs exhibits systematic limitations and failure modes. Models sometimes make calculation errors in intermediate steps even when the overall solution strategy is correct [27]. Performance varies substantially based on problem presentation format, with seemingly superficial changes in wording or notation affecting success rates. Furthermore, LLMs occasionally generate solutions that appear mathematically sophisticated but contain subtle logical errors that undermine validity. These inconsistencies suggest that LLM mathematical reasoning may rely partially on pattern matching against similar problems seen during training rather than robust symbolic manipulation capabilities.

Logical reasoning tasks assess the ability to derive valid conclusions from premises according to formal inference rules. LLMs have been evaluated on propositional logic, predicate logic, and natural language inference tasks that require identifying entailment relationships between statements [28]. Performance on these tasks has improved with model scale, with larger models more consistently applying logical rules and avoiding common fallacies. Research has investigated whether LLMs develop internal representations of logical structure or merely learn surface-level patterns associated with valid inferences. Evidence suggests a combination of both mechanisms, with models exhibiting some systematic logical capabilities alongside reliance on linguistic heuristics that can lead to errors.

Commonsense reasoning involves making plausible inferences about everyday situations based on background knowledge about physical causality, social conventions, and typical event sequences. This capability is fundamental to human intelligence but has proven remarkably difficult for AI systems due to the vast scope of commonsense knowledge and its implicit, context-dependent nature [29]. LLMs pre-trained on diverse web text acquire substantial commonsense knowledge that enables reasonable inferences about ordinary scenarios. Benchmarks assessing commonsense reasoning, such as PIQA, HellaSwag, and CommonsenseQA, show that modern LLMs approach or exceed human performance on multiple-choice questions requiring everyday knowledge. However, these results must be interpreted cautiously given potential training data contamination and the possibility that models exploit superficial linguistic cues rather than genuine understanding.

Analogical reasoning, the capacity to identify structural similarities between superficially dissimilar domains and transfer knowledge accordingly, represents a hallmark of human creative thinking and problem-solving. Research has explored whether LLMs can perform analogical reasoning by solving problems in novel domains based on examples from different contexts [30]. Studies have demonstrated that LLMs can complete analogy tasks of the form "A is to B as C is to what?" with reasonable accuracy, especially when provided with explanatory context. More complex relational reasoning tasks that require mapping multiple correspondences between source and target domains show mixed results, with performance depending heavily on the familiarity and complexity of the domains involved.

Causal reasoning involves understanding cause-effect relationships, predicting intervention outcomes, and distinguishing correlation from causation. This capability is essential for planning, scientific reasoning, and decision-making under uncertainty. LLMs trained on observational text data acquire implicit causal knowledge that enables basic causal inference in familiar domains [31]. However, research has shown that LLMs struggle with tasks requiring explicit causal modeling, counterfactual reasoning, or intervention prediction in novel scenarios. The lack of grounding in physical experience and inability to perform controlled experiments limits LLM causal reasoning capabilities compared to systems that can interact with environments.

Abductive reasoning, the process of inferring the most plausible explanation for observed phenomena, has been investigated in LLMs through narrative understanding and diagnostic reasoning tasks. Models demonstrate reasonable ability to generate explanations for events in stories, medical symptoms, or system failures [32]. The quality of abductive inferences varies with the richness of background knowledge in the domain, with better performance in areas well-represented in training data. Research has noted that LLMs sometimes generate multiple plausible but mutually inconsistent explanations without recognizing the contradictions, indicating limitations in maintaining coherent belief states.

Spatial reasoning tasks requiring understanding of geometric relationships, mental rotation, and navigation have proven challenging for language-only models. Visual LLMs that process both text and images show improved spatial reasoning capabilities by grounding linguistic descriptions in visual representations [33]. However, even multimodal models struggle with complex spatial reasoning tasks that humans solve readily, such as mental folding of paper or three-dimensional object assembly. These limitations highlight the importance of embodied experience and sensorimotor grounding for certain forms of intelligence.

Temporal reasoning involves understanding event sequences, duration relationships, and temporal constraints. LLMs demonstrate basic temporal reasoning capabilities such as ordering events based on linguistic cues and making plausible inferences about typical event durations. More complex temporal reasoning tasks involving relative time references, temporal projection, or scheduling problems show inconsistent performance [34]. The sequential nature of language processing may provide some advantages for temporal reasoning, but limitations in maintaining precise temporal state representations affect reliability.

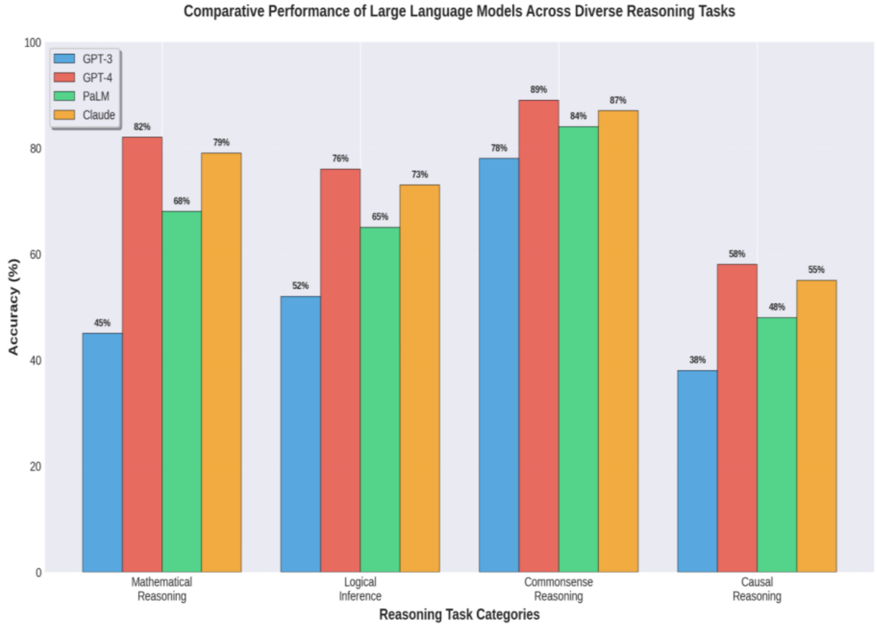


Figure 1: Comparative performance of large language models across diverse reasoning tasks.

The consistency and robustness of LLM reasoning remains an active research concern. Studies have demonstrated that models sometimes fail on simple instances of problem types they solve correctly in more complex forms, violating expected difficulty hierarchies. This brittleness suggests that LLM reasoning relies partially on superficial pattern matching rather than robust algorithmic procedures [35]. Adversarial examples, where minimal perturbations to problem statements dramatically change model responses, reveal brittleness in reasoning capabilities. The sensitivity to prompt formatting, instruction phrasing, and example selection indicates that current reasoning capabilities are not fully general or reliable.

Self-refinement techniques, wherein models critique and improve their own reasoning outputs through iterative generation, have shown promise for enhancing reasoning quality. These approaches leverage the model's ability to recognize errors in generated reasoning chains and produce corrections [36]. Research has explored various self-refinement protocols including self-consistency checking, critique-based revision, and multi-agent debate frameworks. While these methods improve average performance, they also increase computational costs and do not eliminate fundamental reasoning limitations.

Figure 1 presents comparative performance across four reasoning domains for major LLM architectures. Mathematical reasoning shows the strongest performance, particularly with CoT prompting, while causal reasoning exhibits the most significant limitations. The data reveals consistent improvement with model scale across all domains, with GPT-4 achieving highest accuracy. However, the persistent gap between commonsense reasoning (approaching human performance) and causal reasoning (substantially below) highlights fundamental architectural limitations. These patterns suggest that while LLMs acquire substantial reasoning capabilities through scale, certain cognitive domains—particularly those requiring grounded causal understanding—remain challenging regardless of parameter count.

4. Planning and Multi-Step Problem Solving

Planning represents a crucial component of general intelligence, involving the decomposition of high-level goals into executable action sequences, consideration of action preconditions and effects, and adaptation to environmental constraints. LLMs have demonstrated unexpected planning capabilities despite being trained primarily on static text prediction rather than interactive decision-making. These capabilities emerge from the model's ability to generate coherent multi-step narratives, simulate action consequences through language, and leverage procedural knowledge absorbed during pre-training.

Task decomposition, the process of breaking complex objectives into manageable sub-goals, represents a fundamental planning skill that LLMs exhibit across diverse domains. Research has shown that when prompted with high-level instructions, LLMs can generate reasonable task breakdowns that capture key dependencies and ordering constraints [37]. For instance, given a goal such as "prepare a research presentation," models can articulate sub-tasks including literature review, slide preparation, rehearsal, and delivery. The quality of task decompositions depends on domain familiarity, with better performance in areas well-represented in training corpora.

Hierarchical planning involves organizing actions into nested goal structures where high-level objectives are achieved through completion of intermediate sub-goals. LLMs demonstrate basic hierarchical planning capabilities by generating action sequences at multiple levels of abstraction. Research investigating LLM-based planning for household robotics has shown that models can produce hierarchical plans specifying both high-level task sequences and low-level motor actions [38]. However, maintaining consistency across hierarchical levels and ensuring that low-level actions actually achieve high-level goals remains challenging.

Temporal planning requires reasoning about action durations, scheduling constraints, and resource availability over time. LLMs exhibit limited temporal planning capabilities, sometimes generating action sequences that violate temporal constraints or make unrealistic assumptions about concurrent execution. The lack of explicit temporal reasoning mechanisms and difficulty maintaining precise temporal state representations contribute to these limitations [39]. Integration with external scheduling algorithms or temporal constraint solvers has shown promise for enhancing LLM temporal planning capabilities.

Conditional planning involves generating action sequences that adapt to uncertain outcomes or alternative environmental states. LLMs can express conditional plans through natural language control flow, articulating contingency actions for different scenarios [40]. Research has explored using LLMs to generate decision trees or conditional execution graphs that specify different action branches based on observation outcomes. The ability to enumerate relevant contingencies and assign appropriate probabilities remains limited, with models sometimes overlooking important edge cases.

Planning under uncertainty requires balancing exploration and exploitation, assessing action risks, and making decisions with incomplete information. LLMs demonstrate basic probabilistic reasoning capabilities that enable simple uncertainty quantification. However, rigorous planning under uncertainty typically requires frameworks such as Markov decision processes (MDPs) or partially observable MDPs that LLMs cannot directly implement [41]. Hybrid approaches combining LLM high-level reasoning with formal planning algorithms for low-level optimization show promise.

Multi-agent planning scenarios where multiple actors coordinate toward shared or competing goals present additional complexity.

LLMs have been employed in simulations of multi-agent interactions, generating reasonable behaviors for individual agents while accounting for others' likely actions [42]. Research on LLM-based game playing has demonstrated strategic reasoning capabilities in competitive and cooperative scenarios. However, the depth of strategic reasoning and theory of mind capabilities in LLMs remains substantially limited compared to human planning in social contexts.

Tool use represents a crucial extension of LLM planning capabilities, enabling models to invoke external functions, access databases, or execute code to accomplish tasks beyond pure language generation. Recent research has developed frameworks where LLMs plan sequences of tool invocations, passing outputs from one tool as inputs to subsequent tools. These augmented systems have demonstrated capabilities in data analysis, web search, calculation, and code execution that substantially expand planning domains. In distributed computing environments, graph neural network-based approaches have shown that modeling task dependencies as directed acyclic graphs combined with reinforcement learning enables adaptive scheduling that responds to dynamic system conditions, providing insights applicable to LLM-based planning under resource constraints [43]. The challenge lies in learning appropriate tool selection strategies and robustly handling tool execution failures.

Replanning and plan adaptation in response to execution failures or environmental changes represent important aspects of robust planning systems. LLMs demonstrate limited online adaptation capabilities, sometimes struggling to revise plans when initial actions fail or conditions change. Research has explored frameworks where LLMs monitor execution outcomes and generate alternative plans when deviations are detected [44]. The effectiveness of LLM replanning depends on the richness of environmental feedback and the model's ability to diagnose failure causes.

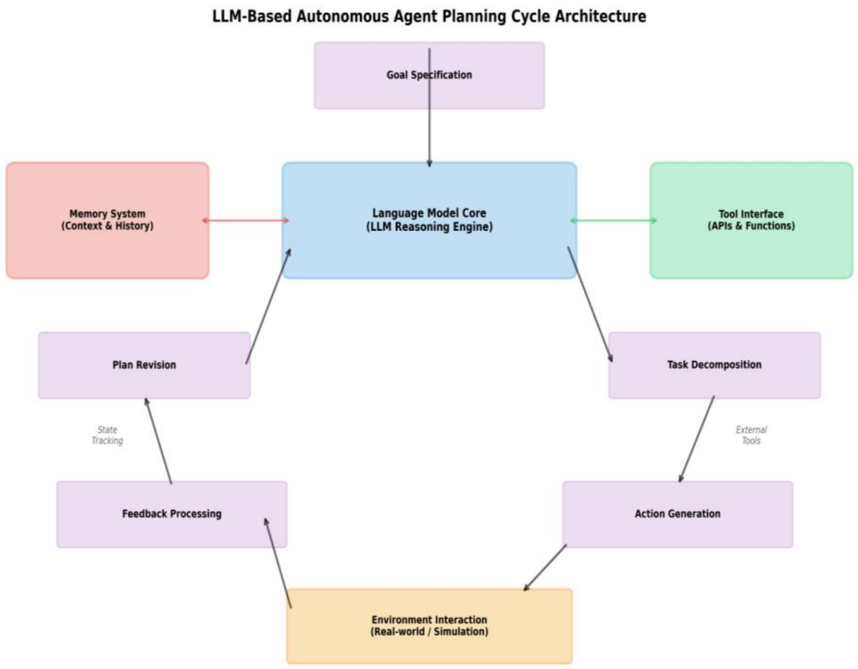


Figure 2: Architecture diagram illustrating the planning cycle for LLM-based autonomous agents.

Goal inference from natural language instructions presents challenges for LLM-based planning systems, as human instructions often underspecify goals or contain implicit constraints. Research has investigated techniques for clarifying ambiguous instructions through interactive dialogue before plan generation [45]. The ability to identify missing information and ask appropriate clarification questions varies across domains and instruction complexity.

Plan verification and validation ensure that generated plans satisfy constraints and are likely to achieve intended goals. LLMs can perform basic plan checking by simulating execution and verifying precondition satisfaction. However, comprehensive formal verification of plan correctness remains beyond current LLM capabilities without integration with automated theorem provers [46]. Research has explored using LLMs to generate test cases or invariants that characterize valid plans.

Learning from planning failures could enable LLMs to improve planning capabilities over time through experience. Current LLMs lack explicit memory systems that accumulate task-specific planning knowledge across sessions. Approaches including retrieval of similar past planning episodes and fine-tuning on successful plan examples have shown moderate improvements [47]. The development of continual learning frameworks that enable LLMs to refine planning strategies without catastrophic forgetting represents an important research direction.

Figure 2 illustrates the iterative planning architecture enabling LLM-based autonomous agents. The cycle begins with goal specification, where natural language objectives are parsed into actionable targets. Task decomposition leverages the LLM's ability to generate hierarchical sub-goal structures. The action generation component produces executable steps, which interface with environments through tool APIs or simulation layers. Critically, the feedback processing loop enables plan revision based on execution outcomes, addressing a key limitation of open-loop planning approaches. The memory system maintains state across iterations, supporting coherent long-horizon planning. This architecture underlies recent advances in code generation, web navigation, and robotic task completion.

5. Decision Making and Real-World Applications

Decision making involves selecting actions from available alternatives based on preferences, constraints, and predicted outcomes. LLMs have been increasingly deployed as decision support systems and autonomous decision-makers across professional domains ranging from medicine to finance. The capacity to process vast unstructured information, synthesize multiple perspectives, and generate reasoned recommendations positions LLMs as valuable tools for complex decision scenarios. However, critical concerns regarding reliability, bias, and accountability necessitate careful consideration of appropriate use cases and deployment protocols.

Table 1 synthesizes LLM decision support capabilities across five key application domains. Medical diagnosis achieves near-physician accuracy on structured vignettes but faces hallucination risks in clinical deployment. Financial analysis demonstrates competitive returns in simulations while raising concerns about adversarial manipulation. Legal research shows strong performance on bar examinations but requires human oversight for citation verification. Scientific research enables novel hypothesis generation with laboratory automation integration. Education applications provide personalized instruction with variable effectiveness across subjects. Across domains, the pattern emerges of strong benchmark performance coupled with deployment challenges related to reliability, explainability, and domain-specific validation requirements.

Medical diagnosis and treatment recommendation represent high-stakes decision domains where LLMs have shown promising but uneven capabilities. Research has demonstrated that LLMs can generate reasonable differential diagnoses from patient symptom descriptions, often matching or approaching physician performance on clinical vignettes [48]. The ability to synthesize information from patient histories, laboratory results, and medical literature enables comprehensive diagnostic reasoning. However, studies have also identified concerning error patterns including hallucinated medical facts, overlooked critical symptoms, and recommendations inconsistent with clinical guidelines. The lack of explicit uncertainty quantification and tendency toward overconfidence pose particular risks in medical contexts.

Therapeutic decision support, where LLMs suggest treatment options and consider patient-specific factors, has been explored through both conversational interfaces and structured clinical decision support systems. Models can articulate reasoning about treatment trade-offs, potential side effects, and patient preference alignment [49]. Integration with clinical knowledge bases and drug interaction databases enhances safety and accuracy. Nonetheless, the opacity of LLM decision processes and difficulty explaining recommendations in terms of established medical reasoning frameworks complicate clinical adoption.

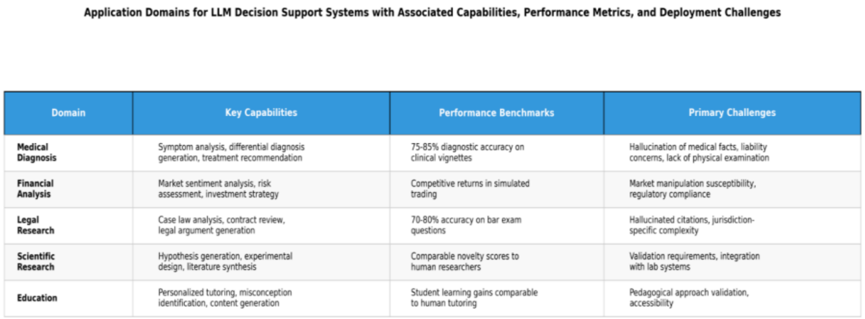
Financial decision making including investment analysis, risk assessment, and portfolio management has attracted significant interest as an application domain for LLMs. Models have been employed to analyze earnings reports, news sentiment, and market trends to generate investment recommendations [50]. Research has shown that LLM-generated trading strategies can achieve competitive returns in simulated trading environments. The ability to process qualitative information from diverse textual sources complements traditional quantitative financial models. Related work in supply chain forecasting has demonstrated that causal-aware multimodal transformers can effectively integrate textual sentiment, temporal patterns, and visual data while distinguishing genuine causal relationships from spurious correlations, offering a framework that could enhance LLM-based decision systems in operational planning contexts [51]. However, the susceptibility to market manipulation through adversarial inputs and unpredictable responses to novel market conditions raise concerns about real-world deployment.

Credit risk assessment represents another financial application where LLMs process loan applications, financial documents, and alternative data sources to predict default probability. Studies have explored whether LLM-based credit models reduce bias compared to traditional scoring systems or inadvertently amplify demographic disparities. The interpretability challenges associated with LLM decisions complicate regulatory compliance in jurisdictions requiring explainable lending decisions.

Legal decision support systems employing LLMs assist with case research, contract analysis, and legal strategy formulation. Models can identify relevant precedents, summarize case law, and suggest legal arguments based on fact patterns [52]. Research has investigated LLM performance on bar examination questions and legal reasoning tasks, finding competitive performance with human test-takers. However, the risk of hallucinated case citations and mischaracterization of legal principles necessitates careful human oversight.

Scientific research decision making including hypothesis generation, experimental design, and literature synthesis represents an emerging application area for LLMs. Models can propose research directions by identifying gaps in existing literature and suggesting novel combinations of established concepts [53]. Automated experimental design frameworks employ LLMs to specify experimental parameters, predict outcomes, and recommend next experiments based on prior results. Integration with laboratory automation systems enables closed-loop scientific discovery where LLMs direct experimental campaigns.

Education and personalized learning systems utilize LLMs for curriculum design, content recommendation, and adaptive instruction. Models can assess student understanding from written responses, identify misconceptions, and generate tailored explanations or practice problems [54]. Research has explored LLM-based intelligent tutoring systems that engage students in Socratic dialogue to guide learning. The effectiveness of these systems compared to human instruction varies across subject matter and student populations.



Application Domains for LLM Decision Support Systems with Associated Capabilities, Performance Metrics, and Deployment Challenges

Domain	Key Capabilities	Performance Benchmarks	Primary Challenges
Medical Diagnosis	Symptom analysis, differential diagnosis generation, treatment recommendation	75-85% diagnostic accuracy on clinical vignettes	Hallucination of medical facts, liability concerns, lack of physical examination
Financial Analysis	Market sentiment analysis, risk assessment, investment strategy	Competitive returns in simulated trading	Market manipulation susceptibility, regulatory compliance
Legal Research	Case law analysis, contract review, legal argument generation	70-80% accuracy on bar exam questions	Hallucinated citations, jurisdiction-specific complexity
Scientific Research	Hypothesis generation, experimental design, literature synthesis	Comparable novelty scores to human researchers	Validation requirements, integration with lab systems
Education	Personalized tutoring, misconception identification, content generation	Student learning gains comparable to human tutoring	Pedagogical approach validation, accessibility

Table 1: Application domains for LLM decision support systems with associated capabilities, performance metrics, and deployment challenges.

Human resource decisions including candidate screening, interview question generation, and performance evaluation have been proposed as LLM applications. Models can analyze resumes, assess candidate qualifications against job requirements, and generate structured interview guides. However, concerns about amplification of hiring biases, privacy violations, and dehumanization of employment decisions have prompted calls for strict limitations on automated HR systems [55].

Collaborative human-AI decision making frameworks attempt to leverage complementary strengths of human judgment and machine analysis. Research has investigated interaction protocols where LLMs provide preliminary analyses that humans review and refine [56]. The effectiveness of these collaborations depends on appropriate division of responsibilities, transparency of AI contributions, and mechanisms for human oversight. Studies have shown that human reliance on AI recommendations can be either insufficient, leading to underutilization of valuable insights, or excessive, resulting in automation bias and acceptance of erroneous suggestions.

Ethical decision making frameworks for LLMs address challenges in aligning model behavior with human values across diverse cultural contexts. Research has explored encoding ethical principles such as utilitarianism, deontology, or virtue ethics into decision-making prompts [57]. The difficulty of specifying comprehensive value systems and handling value conflicts in complex scenarios remains a fundamental challenge. Participatory approaches involving diverse stakeholder input in defining acceptable AI decision criteria show promise for improving value alignment.

Decision transparency and explainability represent critical requirements for many applications, particularly in regulated domains or high-stakes contexts. LLMs can generate natural language explanations of their decisions, articulating the reasoning process and key factors influencing recommendations [58]. However, these explanations may not accurately reflect the model's actual decision process, instead constituting post-hoc rationalizations. Research on mechanistic interpretability aims to develop techniques for understanding genuine causal factors in LLM decisions.

6. Challenges and Limitations

Despite remarkable progress in LLM capabilities across reasoning, planning, and decision making, fundamental limitations constrain their viability as general purpose intelligence systems. Hallucination, the generation of fluent but factually incorrect information, represents perhaps the most widely recognized limitation. Models confidently assert false claims, fabricate citations, and generate plausible but nonsensical explanations. The mechanisms underlying hallucination likely involve both knowledge gaps where models lack information and representational confusion where models conflate similar concepts [59]. Research has explored various mitigation strategies including retrieval augmentation, uncertainty estimation, and adversarial training, but no approach eliminates hallucination entirely.

Reasoning inconsistency manifests as variable performance across problem instances that should be equivalently difficult based on logical structure. Models may solve complex problems while failing on simpler variants, violating expected difficulty hierarchies. This brittleness suggests that LLM reasoning relies partially on superficial pattern matching rather than robust algorithmic procedures. Adversarial examples demonstrate how minimal perturbations to problem statements can dramatically affect model performance. The lack of guaranteed reasoning soundness limits applicability in domains requiring high reliability.

Computational requirements for training and deploying LLMs raise practical and environmental concerns. Training state-of-the-art models requires thousands of specialized hardware accelerators over extended periods, consuming substantial energy. The carbon footprint of training large models has prompted calls for more sustainable AI development practices [60]. Inference costs limit deployment scalability, particularly for real-time applications or resource-constrained environments. Research on model compression, efficient architectures, and knowledge distillation aims to reduce computational demands while preserving capabilities.

Data requirements for training general-purpose LLMs include massive text corpora that raise copyright, privacy, and data governance concerns. Web scraping practices that collect training data may violate intellectual property rights or capture personal information without consent. The composition of training corpora influences model capabilities and biases, with underrepresented perspectives and languages receiving less coverage. Curating high-quality, diverse, and ethically sourced training data at the scale required for competitive LLMs presents ongoing challenges.

Bias and fairness issues pervade LLM outputs, reflecting and sometimes amplifying societal biases present in training data. Models generate stereotypical associations, discriminatory content, and demographically skewed predictions. Research has documented bias across multiple dimensions including gender, race, age, and nationality. Debiasing techniques including filtered training data, adversarial training, and output post-processing show partial effectiveness but do not eliminate bias entirely. The interaction between multiple bias dimensions and context-dependent manifestations complicates bias mitigation.

Safety and alignment challenges arise from potential misuse of LLMs for generating misinformation, malicious code, or manipulative content. Ensuring that LLMs behave in accordance with human values and societal norms across diverse contexts remains an unsolved problem. Techniques such as reinforcement learning from human feedback (RLHF) have improved alignment but do not guarantee safe behavior in all scenarios. The difficulty of specifying complete and consistent value systems, combined with the challenge of robustly implementing such systems in large neural networks, presents ongoing research challenges.

6. Conclusion

This comprehensive review has examined LLMs as general purpose intelligence systems through the lens of three fundamental cognitive capabilities: reasoning, planning, and decision making. The analysis reveals that contemporary LLMs demonstrate remarkable and often unexpected competencies across diverse intellectual tasks that were traditionally thought to require specialized systems or extensive domain-specific training. The transformer architecture combined with massive-scale pre-training has produced models capable of sophisticated language understanding, multi-step problem solving, and complex decision support that approach or exceed human performance on numerous benchmarks.

In the domain of reasoning, LLMs exhibit substantial capabilities in mathematical problem-solving, logical inference, commonsense reasoning, and analogical thinking. CoT prompting and ICL have emerged as powerful techniques that unlock latent reasoning abilities without requiring model retraining. However, systematic inconsistencies, sensitivity to problem framing, and limitations in causal understanding reveal that these reasoning capabilities, while impressive, remain brittle and unreliable compared to human cognitive flexibility. The extent to which LLMs develop genuine understanding versus sophisticated pattern matching continues to be debated, with implications for appropriate application domains and future development directions.

Planning and multi-step problem-solving capabilities in LLMs enable task decomposition, hierarchical goal structuring, and basic temporal reasoning. The integration of LLMs with external tools through RAG frameworks and autonomous agent architectures substantially expands their operational capabilities. Yet challenges in maintaining plan consistency, adapting to unexpected failures, and reasoning under uncertainty indicate that current planning mechanisms fall short of the robustness required for fully autonomous operation in complex real-world environments. Hybrid approaches combining LLM flexibility with formal planning algorithms show promise but require further development.

Decision-making applications across medicine, finance, law, science, and education demonstrate the practical value of LLMs as decision support tools. Their ability to synthesize vast amounts of unstructured information and generate reasoned recommendations provides genuine utility for human decision-makers. However, issues including hallucination, bias amplification, lack of transparency, and uncertain reliability necessitate careful human oversight and limit applicability in high-stakes contexts. Collaborative human-AI frameworks that leverage complementary strengths while maintaining human agency and accountability represent the most prudent near-term deployment strategy.

Fundamental challenges including computational costs, data requirements, reasoning inconsistency, and alignment difficulties constrain the trajectory toward general purpose intelligence. While scaling has driven substantial capability improvements, questions remain about whether current architectures and training paradigms can overcome inherent limitations in causal understanding, grounded meaning, and systematic generalization. Future research directions include developing more efficient architectures, improving reasoning robustness through neurosymbolic integration, enhancing transparency through mechanistic interpretability, and advancing alignment techniques to ensure safe and beneficial AI systems.

The question of whether LLMs represent a path toward AGI or merely sophisticated statistical pattern matchers remains unresolved. Evidence suggests a nuanced reality where these systems exhibit genuine intellectual capabilities that constitute a valid form of machine intelligence, while simultaneously lacking aspects of understanding and flexibility that characterize human cognition. Rather than viewing LLMs as either fully intelligent or merely mimicking intelligence, a more productive perspective recognizes them as powerful computational tools with distinctive strengths and limitations that differ from human cognitive architecture.

Looking forward, the continued development of LLMs will likely focus on addressing current limitations through architectural innovations, improved training methodologies, and tighter integration with symbolic reasoning systems and external knowledge sources. The emergence of multimodal models, advances in continual learning, and development of more sophisticated evaluation frameworks will provide deeper insights into the capabilities and boundaries of this approach to artificial intelligence. Whether LLMs ultimately prove to be a stepping stone toward AGI or a powerful but fundamentally limited technology, their impact on AI research and practical applications has been transformative and will continue to shape the field for years to come.

References

[1] Liu, X., McDuff, D., Kovacs, G., Galatzer-Levy, I., Sunshine, J., Zhan, J., ... & Patel, S. (2023). Large language models are few-shot health learners. arXiv preprint arXiv:2305.15525.

[2] Bubeck S, Chandrasekaran V, Eldan R, et al. Sparks of artificial general intelligence: Early experiments with GPT-4. arXiv preprint arXiv:2303.12712. 2023.

[3] Lin, H., & Liu, W. (2025). Symmetry-Aware Causal-Inference-Driven Web Performance Modeling: A Structure-Aware

Framework for Predictive Analysis and Actionable Optimization. Symmetry, 17(12), 2058.

[4] Muennighoff, N., Rush, A., Barak, B., Le Scao, T., Tazi, N., Piktus, A., ... & Raffel, C. A. (2023). Scaling data-constrained language models. Advances in Neural Information Processing Systems, 36, 50358-50376.

[5] Peng, H., Wang, X., Chen, J., Li, W., Qi, Y., Wang, Z., ... & Li, J. (2023). When does in-context learning fall short and why? a study on specification-heavy tasks. arXiv preprint arXiv:2311.08993.

[6] Huang J, Chang KCC. Towards reasoning in large language models: A survey. Findings of the Association for Computational Linguistics ACL 2023. 2023:1049-1065.

[7] Wei J, Wang X, Schuurmans D, et al. Chain-of-thought prompting elicits reasoning in large language models. Advances in Neural Information Processing Systems. 2022;35:24824-24837.

[8] Press O, Zhang M, Min S, et al. Measuring and narrowing the compositionality gap in language models. Findings of EMNLP. 2023:5687-5711.

[9] Yang, X., Li, T., Su, Q., Liu, Y., Kang, C., Lyu, Y., ... & Pan, Y. (2025). Application of large language models in disease diagnosis and treatment. Chinese Medical Journal, 138(02), 130-142.

[10] Lewis P, Perez E, Piktus A, et al. Retrieval-augmented generation for knowledge-intensive NLP tasks. Advances in Neural Information Processing Systems. 2020;33:9459-9474.

[11] Mai, N. T., Fang, Q., & Cao, W. (2025). Measuring Student Trust and Over-Reliance on AI Tutors: Implications for STEM Learning Outcomes. International Journal of Social Sciences and English Literature, 9(12), 11-17.

[12] Strohmeier, S. (2022). Artificial intelligence in human resources-an introduction. In Handbook of research on artificial intelligence in human resource management (pp. 1-22). Edward Elgar Publishing.

[13] Montesinos López, O. A., Montesinos López, A., & Crossa, J. (2022). Fundamentals of artificial neural networks and deep learning. In Multivariate statistical machine learning methods for genomic prediction (pp. 379-425). Cham: Springer International Publishing.

[14] Chen, Y., Bai, Y., Zhang, W., & Mei, T. (2019). Destruction and construction learning for fine-grained image recognition. In Proceedings of the IEEE/CVF conference on computer vision and pattern recognition (pp. 5157-5166).

[15] Tang, Y., Wang, Y., Guo, J., Tu, Z., Han, K., Hu, H., & Tao, D. (2024). A survey on transformer compression. arXiv preprint arXiv:2402.05964.

[16] Hoffmann J, Borgeaud S, Mensch A, et al. Training compute-optimal large language models. arXiv preprint arXiv:2203.15556. 2022.

[17] Kambhampati, S., Valmeekam, K., Guan, L., Verma, M., Stechly, K., Bhambri, S., ... & Murthy, A. (2024). Llms can't plan, but can help planning in llm-modulo frameworks. arXiv preprint arXiv:2402.01817.

[18] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of NAACL-HLT. 2019:4171-4186.

[19] Wang, X., Wu, J., Yuan, Y., Cai, D., Li, M., & Jia, W. (2024). Demonstration selection for in-context learning via reinforcement learning. arXiv preprint arXiv:2412.03966.

[20] Wang, X., Wei, J., Schuurmans, D., Le, Q., Chi, E., Narang, S., ... & Zhou, D. (2022). Self-consistency improves chain of thought reasoning in language models. arXiv preprint arXiv:2203.11171.

[21] Gan, A., Yu, H., Zhang, K., Liu, Q., Yan, W., Huang, Z., ... & Hu, G. (2025). Retrieval Augmented Generation Evaluation in the Era of Large Language Models: A Comprehensive Survey. arXiv preprint arXiv:2504.14891.

[22] Li J, Li D, Savarese S, Hoi S. BLIP-2: Bootstrapping language-image pre-training with frozen image encoders and large language models. Proceedings of ICML. 2023:19730-19742.

[23] Hu, S., Huang, T., Liu, G., Kompella, R. R., Ilhan, F., Tekin, S. F., ... & Liu, L. (2024). A survey on large language model-based game agents. arXiv preprint arXiv:2404.02039.

[24] Qin, L. (2022). Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. arXiv (Cornell University).

[25] Gao, L., Madaan, A., Zhou, S., Alon, U., Liu, P., Yang, Y., ... & Neubig, G. (2023, July). Pal: Program-aided language models. In International Conference on Machine Learning (pp. 10764-10799). PMLR.

[26] Lewkowycz A, Andreassen A, Dohan D, et al. Solving quantitative reasoning problems with language models. Advances in Neural Information Processing Systems. 2022;35:3843-3857.

[27] Levy, B. (2024). Caution ahead: Numerical reasoning and look-ahead bias in AI models. Available at SSRN 5082861.

[28] Tantakoun, M., Muise, C., & Zhu, X. (2025, March). Llms as planning modelers: A survey for leveraging large language models to construct automated planning models. In AAAI 2025 Workshop LM4Plan.

[29] Talmor A, Herzig J, Lourie N, Berant J. CommonsenseQA: A question answering challenge targeting commonsense knowledge. Proceedings of NAACL-HLT. 2019:4149-4158.

[30] Webb T, Holyoak KJ, Lu H. Emergent analogical reasoning in large language models. Nature Human Behaviour. 2023;7(9):1526-1541.

[31] Jin, Q., Yang, Y., Chen, Q., & Lu, Z. (2024). Genegpt: Augmenting large language models with domain tools for improved access to biomedical information. Bioinformatics, 40(2), btae075.

[32] Jairath, N. K., Pahalyants, V., Cheraghlou, S., Maas, D., Lee, N., Criscito, M. C., ... & Carucci, J. A. (2025). Retrieval Augmented Generation–Enabled Large Language Model for Risk Stratification of Cutaneous Squamous Cell Carcinoma. JAMA dermatology.

[33] Liu, H., Li, C., Wu, Q., & Lee, Y. J. (2023). Visual instruction tuning. Advances in neural information processing systems, 36, 34892-34916.

[34] Zhou, A., Yan, K., Shlapentokh-Rothman, M., Wang, H., & Wang, Y. X. (2023). Language agent tree search unifies reasoning acting and planning in language models. arXiv preprint arXiv:2310.04406.

[35] Berglund L, Tong M, Kaufmann M, et al. The reversal curse: LLMs trained on "A is B" fail to learn "B is A". arXiv preprint arXiv:2309.12288. 2023.

[36] Asano, H., Kozuno, T., & Baba, Y. (2025). Self Iterative Label Refinement via Robust Unlabeled Learning. arXiv preprint arXiv:2502.12565.

[37] Qiao, S., Fang, R., Zhang, N., Zhu, Y., Chen, X., Deng, S., ... & Chen, H. (2024). Agent planning with world knowledge model. Advances in Neural Information Processing Systems, 37, 114843-114871.

[38] Kawaharazuka, K., Matsushima, T., Gambardella, A., Guo, J., Paxton, C., & Zeng, A. (2024). Real-world robot applications of foundation models: A review. Advanced Robotics, 38(18), 1232-1254.

[39] Liu B, Jiang Y, Zhang X, et al. LLM+P: Empowering large language models with optimal planning proficiency. arXiv preprint arXiv:2304.11477. 2023.

[40] Silver, T., Chitnis, R., Kumar, N., McClinton, W., Lozano-Pérez, T., Kaelbling, L., & Tenenbaum, J. B. (2023, June). Predicate invention for bilevel planning. In Proceedings of the

AAAI Conference on Artificial Intelligence (Vol. 37, No. 10, pp. 12120-12129).

[41] Hao S, Gu Y, Ma H, et al. Reasoning with language model is planning with world model. Proceedings of EMNLP. 2023:8154-8173.

[42] Park JS, O'Brien JC, Cai CJ, et al. Generative agents: Interactive simulacra of human behavior. Proceedings of UIST. 2023:1-22.

[43] Yang, S., Ding, G., Chen, Z., & Yang, J. (2025). GART: Graph Neural Network-based Adaptive and Robust Task Scheduler for Heterogeneous Distributed Computing. IEEE Access.

[44] Yao, S., Zhao, J., Yu, D., Du, N., Shafran, I., Narasimhan, K. R., & Cao, Y. (2022, October). React: Synergizing reasoning and acting in language models. In The eleventh international conference on learning representations.

[45] Harper, J. (2024). Autogenesisagent: Self-generating multi-agent systems for complex tasks. arXiv preprint arXiv:2404.17017.

[46] Yang, K., Swope, A., Gu, A., Chalamala, R., Song, P., Yu, S., ... & Anandkumar, A. (2023). Leandojo: Theorem proving with retrieval-augmented language models. Advances in Neural Information Processing Systems, 36, 21573-21612.

[47] Ge, Y., Romeo, S., Cai, J., Sunkara, M., & Zhang, Y. (2025, November). Samule: Self-learning agents enhanced by multi-level reflection. In Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing (pp. 16602-16621).

[48] Nori H, Lee YT, Zhang S, et al. Can generalist foundation models outcompete special-purpose tuning? Case study in medicine. arXiv preprint arXiv:2311.16452. 2023.

[49] Thirunavukarasu AJ, Ting DSJ, Elangovan K, et al. Large language models in medicine. Nature Medicine. 2023;29(8):1930-1940.

[50] Lopez-Lira A, Tang Y. Can ChatGPT forecast stock price movements? Return predictability and large language models. arXiv preprint arXiv:2304.07619. 2023.

[51] Wang, Y., Ding, G., Zeng, Z., & Yang, S. (2025). Causal-Aware Multimodal Transformer for Supply Chain Demand Forecasting: Integrating Text, Time Series, and Satellite Imagery. IEEE Access.

[52] Yu, F., Quartey, L., & Schilder, F. (2022). Legal prompting: Teaching a language model to think like a lawyer. arXiv preprint arXiv:2212.01326.

[53] Boiko DA, MacKnight R, Kline B, Gomes G. Autonomous chemical research with large language models. Nature. 2023;624(7992):570-578.

[54] Kasneci E, Sessler K, Küchemann S, et al. ChatGPT for good? On opportunities and challenges of large language models for education. Learning and Individual Differences. 2023;103:102274.

[55] Tamkin A, Brundage M, Clark J, Ganguli D. Understanding the capabilities limitations and societal impact of large language models. arXiv preprint arXiv:2102.02503. 2021.

[56] Bansal G, Wu T, Zhou J, et al. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. Proceedings of CHI. 2021:1-16.

[57] Sorensen, T., Jiang, L., Hwang, J. D., Levine, S., Pyatkin, V., West, P., ... & Choi, Y. (2024, March). Value kaleidoscope: Engaging ai with pluralistic human values, rights, and duties. In Proceedings of the AAAI Conference on Artificial Intelligence (Vol. 38, No. 18, pp. 19937-19947).

[58] Zhao H, Chen H, Yang F, et al. Explainability for large language models: A survey. ACM Transactions on Intelligent Systems and Technology. 2024;15(2):1-38.

[59] Li, C., Wang, P., Wang, C., Zhang, L., Liu, Z., Ye, Q., ... & Yu, P. S. (2025). Loki's Dance of Illusions: A Comprehensive Survey of Hallucination in Large Language Models. arXiv preprint arXiv:2507.02870.

[60] Patterson D, Gonzalez J, Le Q, et al. Carbon emissions and large neural network training. arXiv preprint arXiv:2104.10350. 2021.