Thinking Beyond Tokens: From Brain-Inspired Intelligence to Cognitive Foundations for Artificial General Intelligence and its Societal Impact

Rizwan Qureshi[1]*, Ranjan Sapkota[2]*, Abbas Shah[3]*, Amgad Muneer[4]*
Anas Zafar[4], Ashmal Vayani[1], Maged Shoman[5], Abdelrahman B. M. Eldaly[6], Kai Zhang[4], Ferhat Sadak[7], Shaina Raza[8]†, Xinqi Fan[9], Ravid Shwartz-Ziv[10], Hong Yan[6], Vinjia Jain[11], Aman Chadha[12], Manoj Karkee[2], Jia Wu[4], Philip Torr[13], and Seyedali Mirjalili[14,15] ‡

Abstract

Can machines truly think, reason and act in domains like humans? This enduring question continues to shape the pursuit of Artificial General Intelligence (AGI). Despite the growing capabilities of models such as GPT-4.5, DeepSeek, Claude 3.5 Sonnet, Phi-4, and Grok 3, which exhibit multimodal fluency and partial reasoning, these systems remain fundamentally limited by their reliance on token-level prediction and lack of grounded agency. This paper offers a cross-disciplinary synthesis of AGI development, spanning artificial intelligence, cognitive neuroscience, psychology, generative models, and agent-based systems. We analyze the architectural and cognitive foundations of general intelligence, highlighting the role of modular reasoning, persistent memory, and multi-agent coordination. In particular, we emphasize the rise of Agentic RAG frameworks that combine retrieval, planning, and dynamic tool use to enable more adaptive behavior. We discuss generalization strategies, including information compression, test-time adaptation, and training-free methods, as critical pathways toward flexible, domain-agnostic intelligence. Vision-Language Models (VLMs) are reexamined not just as perception modules but as evolving interfaces for embodied understanding and collaborative task completion. We also argue that true intelligence arises not from scale alone but from the integration of memory and reasoning: an orchestration of modular, interactive, and self-improving components where compression enables adaptive behavior. Drawing on advances in neurosymbolic systems, reinforcement learning, and cognitive scaffolding, we explore how recent architectures begin to bridge the gap between statistical learning and goal-directed cognition. Finally, we identify key scientific, technical, and ethical challenges on the path to AGI, advocating for systems that are not only intelligent but also transparent, value-aligned, and socially grounded. We anticipate that this paper will serve as a foundational reference for researchers building the next generation of general-purpose human-level machine intelligence.

***K*eywords** Artificial General Intelligence, Multi-Agents Systems, Cognitive Functions, Large Language Models, Vision-Language Models, Large Vision Models, Foundation Models, Human Brain, Robotics, Psychology, Agents, Agentic AI, World Model

*Equal Contribution

†Corresponding author: shaina.raza@torontomu.ca

‡[1] Center for research in Computer Vision, University of Central Florida, Orlando, FL, USA. [2] Cornell University, Department of Biological and Environmental Engineering, Ithaca, NY 14853, USA [3] Department of Electronics Engineering, Mehran University of Engineering & Technology, Jamshoro, Sindh, Pakistan. [4]Department of Imaging Physics, The University of Texas MD Anderson Cancer Center, Houston, TX, USA. [5] Intelligent Transportation Systems, University of Tennessee, Oakridge, TN, USA. [6] Department of Electrical Engineering, City University of Hong Kong, SAR China. [7] Department of Mechanical Engineering, Bartin University, Bartin Turkey. [8] Vector Institute, Toronto Canada. [9] Manchester Metropolitan University, Manchester, UK. [10] Center for Data Science, New York University, NYU, NY, USA. [11] Meta Research (Work done outside Meta). [12] Amazon Research (Work done outside Amazon). [13] Department of Engineering Science, University of Oxford, UK. [14] Centre for Artificial Intelligence Research and Optimization, Torrens University Australia, Fortitude Valley, Brisbane, QLD 4006, Australia, [15] University Research and Innovation Center, Obuda University, 1034 Budapest, Hungary

Table of Contents

Contents

1 Introduction

Can machines truly think? Over seven decades ago, Alan Turing famously posed this foundational question at the dawn of computing. It remains central to the field of Artificial General Intelligence (AGI), which seeks to replicate the full breadth of human cognitive abilities in computational form [1]. Yet, despite decades of progress, the term "thinking" [2] itself is often invoked without sufficient precision [3]. To meaningfully address this question, we must first define what we mean by thinking and related concepts, such as consciousness, intelligence, and generalization:

- **Thinking**: The Manipulation of internal representations to solve problems, reason about the world, and generate novel ideas [2].
- **Consciousness**: The subjective capacity for awareness and self-reflection [4].
- **Intelligence**: The capacity to acquire, apply, and adapt knowledge across tasks and environments [3].
- **AGI**: Systems capable of broad, human-level reasoning and learning across domains, without the need for task-specific retraining [5].

While leading-edge AI models such as GPT-4 [6], DeepSeek [7], and Grok [8] have demonstrated impressive performance across a diverse array of specialized tasks, their underlying architecture remains fundamentally limited by token-level prediction. Although this paradigm excels at surface-level pattern recognition, it lacks grounding in physical embodiment, higher-order reasoning, and reflective self-awareness, which are the core attributes of general intelligence [9]. Furthermore, these models do not exhibit consciousness or an embodied understanding of their environment, limiting their ability to generalize and adapt effectively to novel, open and real-world scenarios [10].

Why Token-level Next-word Prediction Alone is Insufficient for AGI?

Next-token prediction models capture surface linguistic patterns but fail to support complex mental representations grounded in the physical world. Lacking embodiment, causality, and self-reflection, they struggle with abstraction and goal-directed behavior—core requirements for AGI.

Post-training strategies [11] such as instruction tuning [12] and Reinforcement Learning with Human Feedback (RLHF) [13] improve alignment and usability, but operate within the same autoregressive framework. They introduce behavioral refinements, not architectural changes [13]. Consequently, despite post-training advances, these models remain limited in their capacity to generalize in the open-ended, compositional manner characteristic of AGI [9].

Why Post-Training and Alignment Can't Bridge the Gap to AGI?

Post-training methods, such as, Instruction tuning and RLHF transformed base models like GPT into more usable agents like ChatGPT. However, these alignment methods operate on top of token-level prediction and cannot endow models with core AGI traits—such as abstraction, grounded reasoning, or environmental awareness.

Although model scaling can approximate complex representations and produce emergent behaviors, it lacks inductive biases for structured reasoning, fails to support persistent memory, and cannot generate self-models or agency. These limitations are architectural, not parametric—hence, scaling alone yields diminishing returns and cannot achieve AGI [14, 15].

Why Further Scaling Will Not Lead to AGI?

While scaling improves fluency and performance on many tasks, it cannot resolve core limitations of current LLMs. These models still lack grounded understanding, causal reasoning, memory, and goal-directed behavior.

Besides next-token prediction, trajectory modeling frameworks (e.g. Algorithm 1), such as, The *Decision Transformer* reframe reinforcement learning as conditional sequence modeling, enabling policy generation via trajectory-level representations optimized for long-term return [16]. Complementarily, *self-prompting* mechanisms introduce latent planning loops [17], wherein models generate internal scaffolds to structure multi-step reasoning [18]. *DeepSeek-V2*, a 236B-parameter Mixture-of-Experts model with a 128K-token context, exemplifies this paradigm by integrating trajectory modeling with reinforcement fine-tuning to improve coherence and planning across extended tasks [19]. Collectively, these approaches advance beyond token-level generation by embedding structured, goal-conditioned reasoning within the model architecture [18].

Algorithm 1: Trajectory-Based Planning via Decision Transformers

Input: Goal G, history H, reward function R
Output: Action sequence $A = \{a_1, a_2, \ldots, a_T\}$

1. Encode history and desired return into trajectory-level input
2. Use Decision Transformer to predict next actions conditioned on future reward
3. Iteratively update sequence based on observed outcomes
4. Integrate reward-to-go and attention over past states for long-horizon reasoning
5. Output final plan A

Algorithm 2: Prompt-Based Agentic Reasoning (CoT/ToT/ReAct)

Input: Task description T, retrieved context C, agent memory M
Output: Solution S with intermediate reasoning steps

1. Decompose task T into subproblems using Chain-of-Thought (CoT)
2. Explore multiple reasoning paths via Tree-of-Thoughts (ToT)
3. Interleave reasoning with tool/environment actions (ReAct)
4. Score and revise trajectories based on feedback and self-evaluation
5. Return final solution S and rationale trace

Chain-of-Thought prompting further improves reasoning by decomposing tasks into interpretable substeps, enhancing performance on arithmetic, commonsense, and symbolic challenges [20]. Extending this, the *Tree-of-Thoughts (ToT)* framework enables large language models (LLMs) to explore and evaluate multiple reasoning paths via lookahead, backtracking, and self-evaluation, yielding significant gains in tasks requiring strategic planning [21]. For instance, applying ToT to GPT-4 increased its success rate on a combinatorial puzzle from 4% (CoT) to 74%[21]. *ReAct* further augments this space by interleaving reasoning with environment-aware actions, allowing models to iteratively gather information, revise plans, and improve factual accuracy [18]. These complementary methods collectively form the foundation of prompt-based agentic reasoning, enabling both structured internal deliberation and dynamic external interaction. A generalized overview of this unified reasoning process is presented in Algorithm 2.

As AI systems increasingly influence healthcare, education, governance, and the labor market, their integration into society must be guided by ethical, inclusive, and equitable principles [22]. Democratizing AI means equitably distributing access, participation, and benefits across regions, communities, and socioeconomic groups—narrowing existing disparities rather than reinforcing them [23].

AI Integration and the Need for Democratization

Without inclusive development, AI may amplify existing inequalities and silence underrepresented voices. Trustworthy, transparent, and socially aligned systems are not optional; they are a societal necessity.

Rodney Brooks in 2008 argued that intelligence emerges from physical embodiment rather than abstraction alone [24]. Building on this and recent developments in AGI in cross-disciplinary domains [25], we propose that AGI must arise through integrated perception, embodiment, and grounded reasoning, not scale alone. We synthesize decades of AGI research in machine learning, cognitive neuroscience, and computational theory, critically examining recent techniques such as Chain of Thought [20], Tree of Thoughts [21], ReAct [18], and trajectory modeling [16]. While these methods enhance structured reasoning, they remain transitional, lacking physical grounding, memory, and self-awareness—core to general intelligence [26].

To address these gaps, we explore neuro-symbolic systems, multi-agent coordination, and RLHF as building blocks of AGI. This review frames a roadmap toward systems that are cognitively grounded, modular, and value-aligned, centered on the question: What mechanisms are essential to move from prediction to general-purpose intelligence?

Motivation

Artificial General Intelligence (AGI) aims to replicate the full spectrum of human cognition, including reasoning, learning, memory, perception, and adaptation in dynamic, open-ended environments [27]. It is widely regarded one of the most ambitious frontiers in science and technology [26], and interest in AGI continues to grow across academia and industry, with major contributions from OpenAI [28], Amazon [29], Microsoft Research [30], Google [31], and Meta [32].

Although previous studies have explored AGI readiness [26], safety concerns [33], applications in IoT [34], brain-inspired architectures [35], and cognitive frameworks [36], the fundamental challenge per-

sists: how can we transition from statistical pattern recognition to machines capable of genuine reasoning and flexible generalization?

Recent models such as GPT-4, DeepSeek, and Grok demonstrate growing multimodal competence. However, they still lack core capabilities such as abstraction, grounded reasoning, and real-time adaptation, which are essential for building truly general intelligence.

Key Contributions To the best of our knowledge, this is the first review to evaluate AGI through three integrated lenses: computational architectures, cognitive neuroscience, and societal alignment. Specifically:

- We introduce a unified framework that synthesizes insights from neuroscience, cognition, and AI to identify foundational principles for AGI system design.
- We critically analyze the limitations of current token-level models and post hoc alignment strategies, emphasizing the need for grounded, agentic, and memory-augmented architectures.
- We survey emergent AGI-enabling methods, including modular cognition, world modeling, neuro-symbolic reasoning, and biologically inspired architectures.
- We present a multidimensional roadmap for AGI development that incorporates logical reasoning, lifelong learning, embodiment, and ethical oversight.
- We map core human cognitive functions to computational analogues, offering actionable design insights for future AGI systems. A list of key acronyms used in this paper, are defined in Appendix Table A1.

2 Historical Evolution of AI

AI has evolved through several major paradigms: from symbolic rule-based systems [37] to statistical learning models [38], and more recently into the era of generative and agentic AI [39]. As shown in Figure 1, modern generative models [40] excel at capturing data distributions and generating fluent text [41], speech [42], images and videos [43], and even executable code [9]. Yet, despite their breadth, these systems remain fundamentally constrained: they operate at the level of token prediction, lacking grounded semantics, causal reasoning, and long-term planning [44].

The emergence of more autonomous and general-purpose systems such as DeepSeek [19], GPT-4 [45], OpenAI's o1 [46], DeepResearch and xAI's Grok3 [8] signals a potential shift beyond static pattern matching. These models demonstrate early signs of multimodal integration, creative problem-solving, and self-directed planning, pointing toward the first glimpses of general intelligence in machines.

Bridging the divide between narrow pattern-based intelligence and human-like generality is a central challenge for AGI [35]. A confluence of enabling technologies is accelerating this transition from generative AI to systems capable of adaptive, grounded, and goal-directed behavior [47]. One fundamental thread is *deep reinforcement learning (RL)* [48], which enables agents to learn through trial-and-error interaction with dynamic environments. Landmark achievements, such as AlphaGo [49] and AlphaFold2 [50], illustrate how reinforcement learning and attention mechanisms support long-horizon decision-making and structural prediction. These systems rely on stable optimization methods such as *Proximal Policy Optimization (PPO)* [51], which balances exploration with policy stability in high-dimensional action spaces.

To further align model behavior with human values, recent work emphasizes preference-based fine-tuning methods such as *Direct Preference Optimization (DPO)* [52] and *Group Relative Policy Optimization (GRPO)* [53]. These techniques circumvent the need for explicit reward modeling by directly optimizing for human-aligned outcomes based on comparative preference signals. In parallel, *neuro-symbolic systems* [54] integrate symbolic reasoning with deep learning (DL), allowing agents to manipulate abstract variables and compositional rules. Collectively, these systems provide a path toward explainable and generalizable cognition, critical for robust AGI.

2.1 Overview of AGI

AGI represents a frontier in the evolution of computational systems, striving to develop machines that can perform any intellectual task that a human can, across various domains [55]. Unlike narrow AI [56], which is designed for specific tasks, often operating on limited token-level inputs, AGI aims for a comprehensive cognitive ability, simulating the breadth and depth of human intellect [57, 58]. This ambition poses profound implications for society, promising revolutionary advances in healthcare [27], education [59], and beyond [5], while also introducing complex ethical and safety challenges [60]. AGI research encompasses diverse approaches, including symbolic [61], emergentist [6], hybrid [62], and universalist models [63], each offering distinct pathways toward achieving versatile intelligence [64]. The development of AGI involves integrating sophisticated algorithms that can learn, reason, and adapt in ways that mimic human cognitive processes, such as learning from limited data [65], transferring knowledge

across contexts, and abstract reasoning [66, 67]. Despite its potential, the field grapples with significant hurdles such as ensuring safety, managing unforeseen consequences, and aligning AGI systems with human values [68, 55]. Furthermore, measuring progress in AGI development remains contentious, with debates over the appropriateness of benchmarks like the Turing Test [69] or operational standards akin to human educational achievements [70]. As we advance, the integration of interdisciplinary insights from cognitive science, ethics, and robust engineering is crucial to navigate the complexities of AGI and harness its potential responsibly.

2.2 Agentic AI

Although LLMs excel at predicting text, they lack the perceptual grounding that underpins human cognition [71]. Humans build world models by continually integrating sensory input, memory, and action, skills rooted through direct, embodied interaction (e.g., a child learns to catch a ball by moving in space) [59]. LLMs, by contrast, are disembodied: they cannot perceive, act, or internalize causal dynamics, so they struggle with tasks that demand physical reasoning, commonsense inference, or real-time adaptation [72].

To address these limitations, a parallel frontier has emerged in the form of agentic architecture systems designed to perform autonomous planning, memory management, and inter-agent coordination [73, 74]. A notable example is the Natural Language-based Society of Mind (NLSOM) framework [75], which proposes a modular system composed of multiple specialized agents that communicate using natural language. These neural societies reflect Minsky's original vision [76] of the mind as a collection of loosely coupled agents, each responsible for distinct cognitive tasks. By distributing intelligence across a community of specialized modules, NLSOM and similar architectures mitigate the monolithic limitations of conventional LLMs. They enable cognitive functions such as modular reasoning, episodic memory retrieval, and collaborative problem-solving traits essential for developing general-purpose intelligence [77].

These developments mark a transition from static, feedforward predictors to dynamic, interactive, and cognitively enriched AI systems [78]. As depicted in Figure 1, AI has evolved from symbolic systems (e.g., Turing Test, ELIZA) to neural architectures (e.g., LeNet-5, Deep Belief Networks, AlexNet), then to reinforcement agents (e.g., DQN, AlphaGo), attention-based models (e.g., Transformer, BERT), and most recently, to foundation and emergent models such as GPT-4 and DeepSeek-R1. A detailed chronology of modern AI and deep learning can be found in [79, 80].

Recent proposals such as S1 scaling[7] challenge the traditional focus on parameter count as the primary driver of AGI. Instead, they advocate scaling along cognitive axes—modularity, reasoning depth, self-prompting, and agentic coordination[19]. This structured approach marks a paradigm shift from undifferentiated statistical inference toward architecturally organized systems capable of flexible, interpretable reasoning [8]. Collectively, these trends signal a converging path toward open-ended, general-purpose machine intelligence.

3 Understanding Intelligence - Logical Foundations of Intelligence

Understanding the logical and cognitive foundations of intelligence is essential for developing robust AGI systems [81]. Intelligence covers diverse cognitive abilities, including perception, learning, memory, reasoning, and adaptability. Achieving AGI requires a comprehensive understanding of these cognitive processes and their neural bases [82].

3.1 Brain Functionality

The human brain, shown in Figure 3, is a highly intricate and partially understood organ that underlies core cognitive functions such as consciousness, adaptive intelligence, and goal-directed behavior [83, 84]. Despite weighing only 1.3 to 1.5 kg, it accounts for nearly 20% of the body's energy consumption, underscoring its metabolic and computational intensity [85, 86]. Architecturally, the brain is organized into functionally specialized regions operating in tightly integrated hierarchies [87]. The neocortex a hallmark of mammalian evolution supports higher-order cognition and abstract reasoning, while subcortical structures regulate affective and autonomic functions [88]. Key components such as the hippocampus facilitate encoding of episodic memory (EM) and spatial navigation, whereas the occipital cortex governs visual processing and the motor cortex orchestrates voluntary movement [87]. These neurobiological insights offer design principles for AGI systems aiming to replicate cognitive flexibility, embodied intelligence, and adaptive decision-making.

The true computational power of the brain lies in its approximately 86 billion neurons, which create a dense network of about 150 trillion synaptic connections [89, 90, 91]. This vast network enables both localized and extensive communications, positioning the brain as a complex, multi-scale network system. Synaptic activities, which include excitatory and inhibitory signals, maintain a critical balance essential for all cognitive functionalities [92]. These synaptic interactions facilitate complex behaviors and thought processes, underscoring the importance of understanding these networks to replicate similar capabilities in AI systems [93]. This neuro-computational foundation offers a road-map for de-

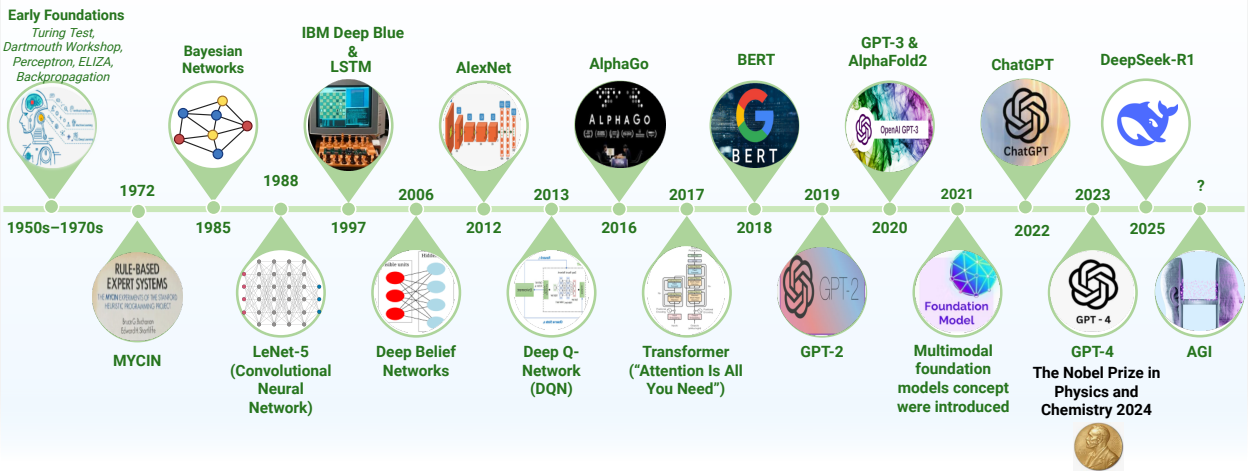

Figure 1: A timeline of key milestones toward Artificial General Intelligence (AGI) from 1950 to 2025. The evolution spans symbolic systems (e.g., ELIZA), neural networks (e.g., LeNet-5, AlexNet), reinforcement learning (e.g., AlphaGo, DQN), foundation models (e.g., GPT-4, DeepSeek-R1), and **(Nobel Prize in Physics and Chemistry in 2024)**. This trajectory reflects a shift from static, rule-based methods to dynamic, multimodal, and increasingly general AI systems.

veloping AGI systems that aim to emulate human-like intelligence.

3.1.1 Brain Functionalities and Their State of Research in AI

Figure 3a maps major brain regions to their AI counterparts, highlighting varying levels of research maturity: well-developed (L1), moderately explored (L2), and underexplored (L3). This comparison reveals both strengths and gaps in current AI research, offering a roadmap for advancing brain-inspired intelligence [94]. The frontal lobe governs high-level cognition such as planning and decision-making [95], with AI showing strong performance in structured tasks (e.g., AlphaGo). Yet, traits like consciousness and cognitive flexibility remain underexplored (L3) [96, 97]. In contrast, language and auditory functions mapped to L1 domains are well-modeled by LLMs, which approach human-level proficiency in language processing [94, 98].

Conversely, the cerebellum and limbic system govern fine motor skills and emotional processing, respectively [99]. In AI, motor coordination is explored via robotics and meta-learning [100, 101], yet achieving human-like dexterity and adaptability remains a challenge (L2–L3) [102]. Emotional and motivational processes modeled by the limbic system are only superficially replicated in AI through reinforcement learning, highlighting a major gap in developing true emotional intelligence. (L3) [103, 104].

3.1.2 Memory in Human and Artificial Intelligence

Memory is a fundamental pillar of cognition in both humans and AI, enabling learning, adaptation, and problem-solving [105]. In humans, it supports language acquisition, skill mastery, and social interaction core to self-awareness and decision-making [106, 107]. Likewise, in AI, memory facilitates intelligent behavior by supporting complex task execution, prediction, and adaptability [108]. This parallel underscores the value of biological memory insights in guiding the design of more advanced, memory-driven AI systems.

Figure 3 presents a hierarchical taxonomy of human memory, outlining how sensory input transitions into short-term and long-term memory through encoding, consolidation, and retrieval [94]. This framework offers a blueprint for AI memory systems, which have evolved from static data stores [109, 110] to dynamic architectures that more closely mimic the flexibility and contextual awareness of human cognition.

Despite recent progress, AI memory systems still fall short of the contextual richness and adaptability of human memory [111]. Unlike humans, who integrate memory with perception, reasoning, and emotion [112], AI typically relies on fixed algorithms and parameters. Achieving AGI will require memory systems that not only store information but also contextualize and conceptualize it akin to human cognition [113]. Drawing from neuroscience and cognitive psychology such as the models in Figure 3 offers a roadmap for building AI that learns from experience,

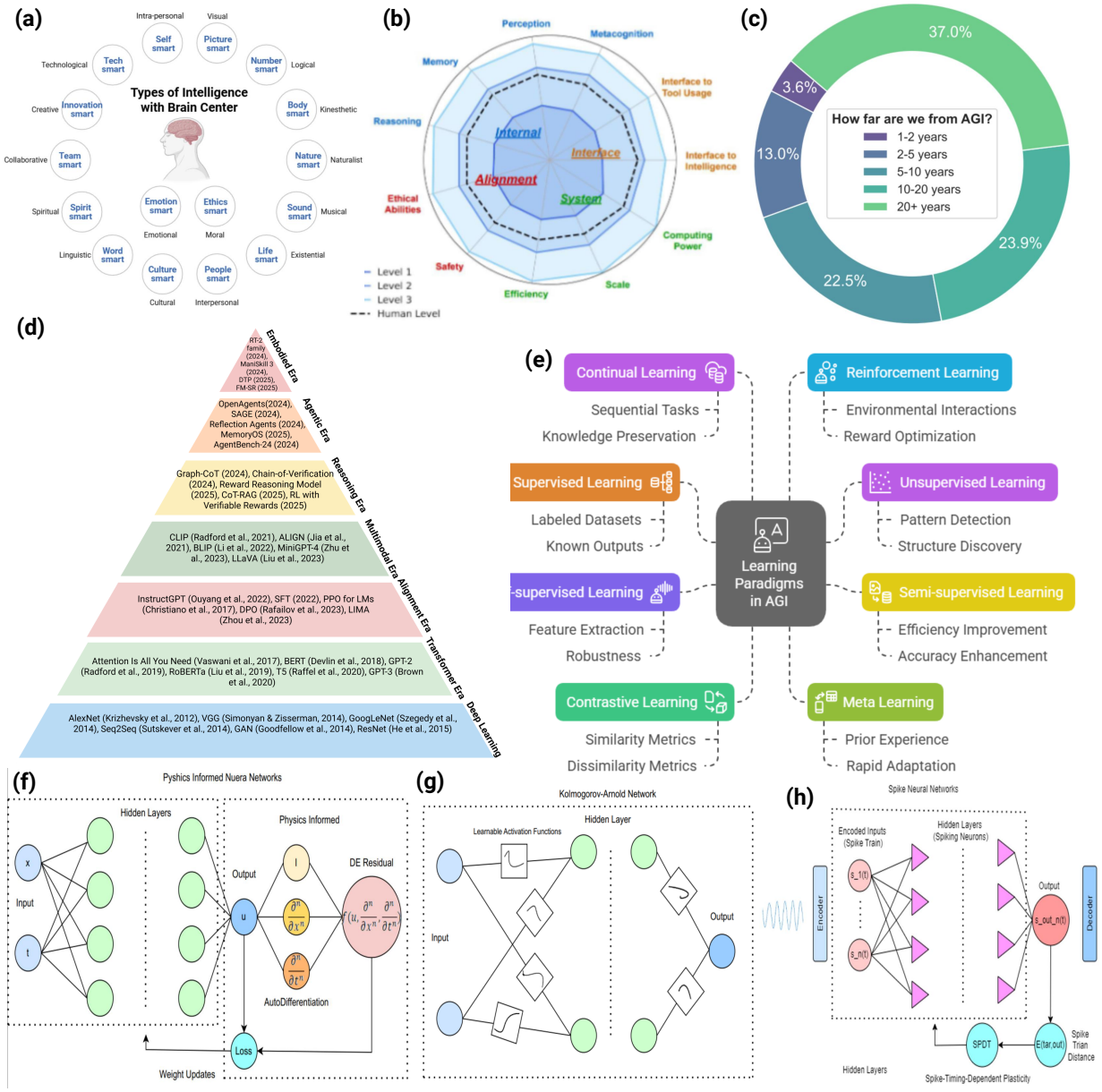

Figure 2: An overview of foundational concepts, progress, and paradigms toward Artificial General Intelligence (AGI). (a) Multiple human intelligence types as conceptualized in brain-inspired AGI. (b) Radar chart representing the multidimensional alignment challenges in AGI including internal reasoning, external interface, system efficiency, and ethical safety. (c) Survey-based forecast of AGI timeline expectations adapted from ICLR 2024 survey [26]. (d) Pyramid of Foundational AI Eras Leading to the Embodied Era. (e) Categorization of core learning paradigms in AGI, including supervised, unsupervised, self-supervised, and reinforcement learning, as well as emerging paradigms like continual, contrastive, semi-supervised, and meta learning. (f–h) Architectures representing (f) Physics-Informed Neural Networks (PINNs), (g) Kolmogorov–Arnold Networks (KANs), and (h) Spiking Neural Networks (SNNs) highlighting biological plausibility and adaptive computation in AGI development.

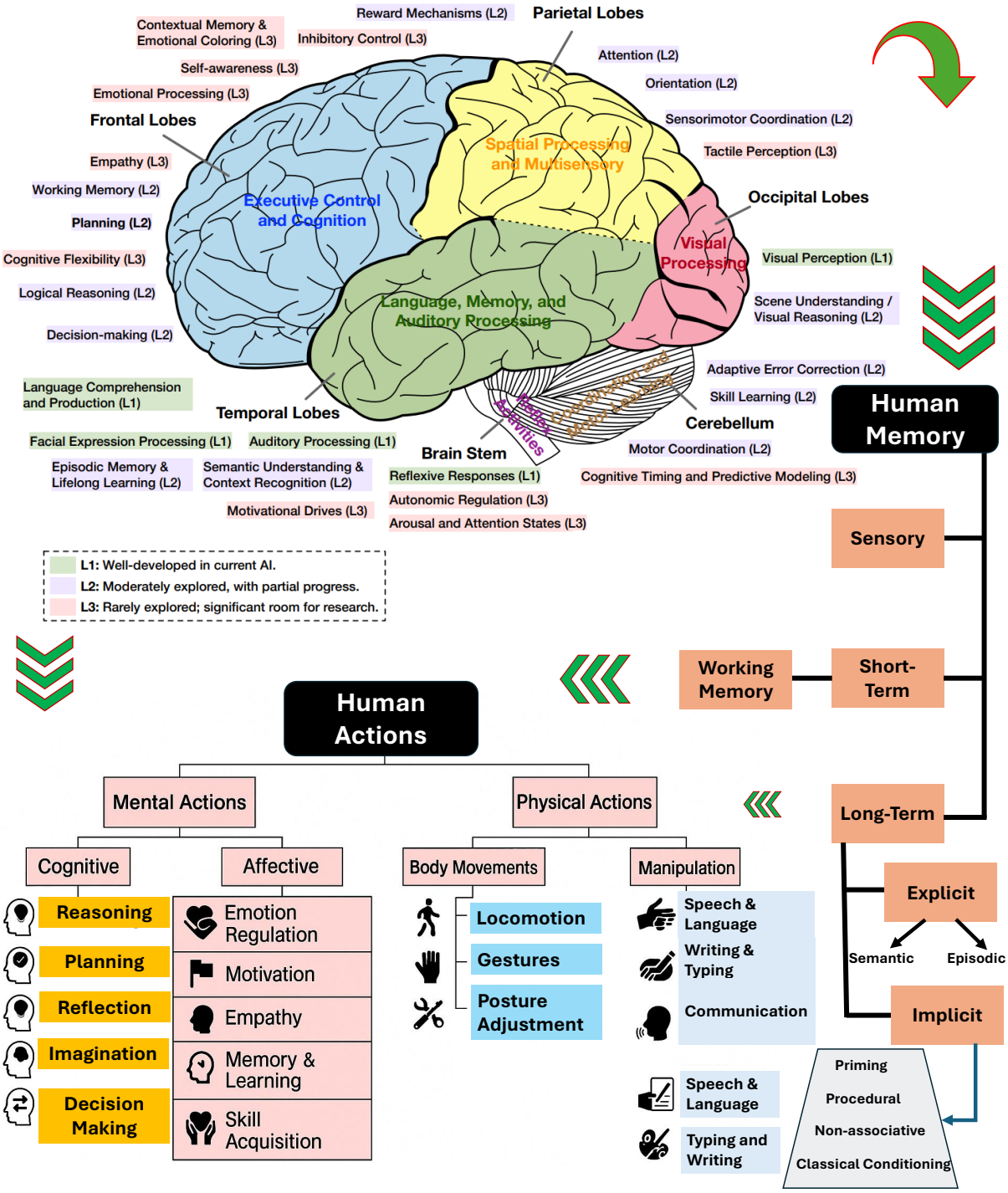

Figure 3: Illustrating the core cognitive pillars brain functions, memory hierarchies, action systems, and world models that serve as foundational inspirations for AGI development. The upper section highlights key brain functionalities aligned with levels of AI research, identifying current achievements, gaps, and opportunities. It presents a hierarchical taxonomy of human memory, including sensory, short-term, working, and long-term types, further categorized into declarative and non-declarative forms. Additionally, it depicts human actions, spanning mental and physical dimensions crucial for cognition, planning, and goal-directed behavior. The brain diagram in this figure showing the functionalities of brain and their state of the research in AI is sourced from [94].

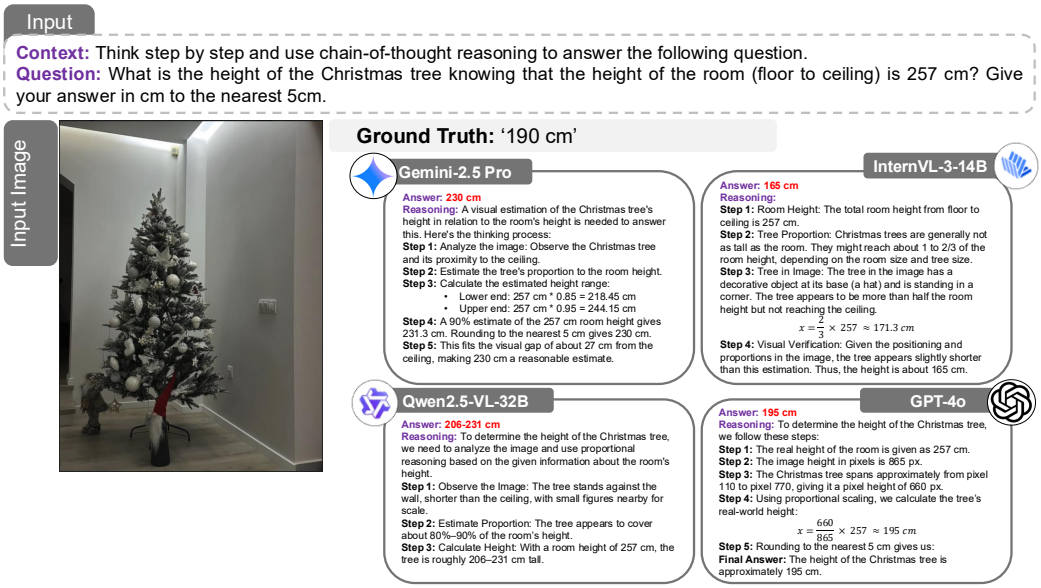

Figure 4: Illustration of the limitations of leading large multimodal models (LMMs) in performing accurate multi-step visual reasoning. Despite being prompted to follow a structured chain-of-thought, all models, ncluding Gemini-2.5 Pro, GPT-4o, Qwen-2.5-VL-32B, and InternVL-3-14B fail to estimate the Christmas tree height correctly based on the known room height of 257 cm. The ground truth of 190 cm highlights over- and under-estimations, exposing a persistent gap between visual perception, proportional reasoning, and precise numerical grounding in current LMMs.

adapts to new situations, and supports emotionally informed, lifelong learning [94].

3.1.3 Human Action System: Mental and Physical Foundations for AGI

The human action system comprising both mental and physical actions is central to intelligent behavior [114, 115]. Mental actions include reasoning, planning, and memory recall, while physical actions encompass movement, communication, and interaction [94] (Figure 3). Mental actions guide internal decision-making and simulate outcomes [116, 117], whereas physical actions execute intentions and adapt behavior through real-world feedback [118, 72]. This bidirectional loop between cognition and action provides a foundational model for AGI systems aiming to integrate perception, planning, and adaptive execution.

In AI agents, action systems are designed to emulate this cognitive loop [119]. Language-based agents (e.g., using LLMs) simulate mental actions like reasoning and planning [120], while robotic agents emulate physical actions via real-world interaction [10, 120]. Models such as LAMs (Large Action Models) aim to unify these capabilities by learning from action trajectories across digital and physical contexts [121]. Crucially, just as humans utilize tools to extend cognitive and physical abilities, AI agents incorporate external APIs, robotic systems, or software interfaces to achieve complex tasks [122]. These tool-mediated actions expand the agent's action space, mirroring the human capacity for tool use and enabling more generalized problem-solving capabilities.

3.1.4 World Models: Cognitive Foundations Bridging Human and AGI

World models are internal representations that allow agents to simulate, predict, and plan without depending solely on trial-and-error [123]. In humans, these mental models underpin spatial navigation, planning, and counterfactual reasoning [124], offering predictive, adaptive, and scalable cognition [125]. For instance, crossing a busy street involves anticipating vehicle motion, timing decisions, and dynamically adjusting behavior hallmarks of world model reasoning. Figure 4 illustrates the cognitive pipeline shared by human and artificial intelligence using the example of a soccer player (AI-generated **Lionel Messi**) predicting and striking a ball. The scenario demonstrates how internal world models enable trajectory prediction before motor action. Prediction integrates visual cues and prior experience, refined by perception and memory. Action is selected through an AI-like decision-making module, and feedback updates memory and internal models. The figure is structured across four conceptual layers: (1) foundational world model types (implicit,

explicit, simulator-based, instruction-driven); (2) dynamic reasoning via prediction, hierarchy, and feedback; (3) core agentic faculties perception, memory, and action; and (4) aspirational AGI capabilities including ethical reasoning and contextual adaptability.

3.1.5 Neural Networks Inspired by Brain Functions

Biological neural systems have inspired a range of architectures that replicate human cognitive functions. Convolutional Neural Networks (CNNs) and attention-based models emulate the visual cortex, excelling in learning local and global patterns [126]. Recurrent Neural Networks (RNNs), reflecting hippocampal temporal processing, are well-suited for sequential data and memory tasks. Spiking Neural Networks (SNNs) mimic neural dynamics like synaptic plasticity and spike timing, offering advantages for temporal modeling and sensor data. Reinforcement Learning (RL), modeled on prefrontal decision-making, enables agents to learn from interaction and feedback in complex environments. Table 1 summarizes how human brain regions map to neural network architectures, outlining their cognitive functions, AI analogues, and applications.

3.2 Cognitive Processes

Cognitive neuroscience leverages brain mapping techniques such as Electroencephalography (EEG), Electrocorticography (ECoG), Magnetoencephalography (MEG), Functional Magnetic Resonance Imaging (fMRI), and Positron Emission Tomography (PET) to investigate the neural basis of cognition [127, 128]. These techniques capture neural activity in response to stimuli, revealing inter-regional communication patterns essential for cognitive functions such as memory [129], learning [130], language [131], cognitive control [132], reward processing [133], and moral reasoning [134, 135]. Furthermore, understanding how neurons communicate sheds light on the foundations of intelligence. Cognitive processes emerge from dynamic interactions across distributed brain regions [136]. By linking neural activity to behavior, cognitive neuroscience bridges low-level circuitry and higher-order cognition [137], offering insights for developing AI systems that emulate the integrative, adaptive capabilities of the human brain [138, 139].

3.2.1 Network Perspective of the Brain

The brain functions as a complex biological network orchestrating perception, emotion, and cognition [140, 141]. Advances in neuroimaging and network science have enabled mapping of the brain's structural and functional connectivity known as the connectome revealing its hierarchical and modular organization [142, 143]. Brain networks are typically classified into three types: anatomical (physical infrastructure), functional (statistical dependencies), and effective (causal influence)[144]. While anatomical networks change slowly, functional and effective networks are dynamic and context-dependent[145], offering critical insights into cognition and adaptive behavior.

3.2.2 Brain Networks in Cognitive Neuroscience

Research shows that cognitive functions attention, memory, decision-making emerge from dynamic interactions across brain networks [146, 147, 148]. Higher cognitive performance correlates with efficient network properties, including high global integration and short path lengths [149, 150], while reduced integration is linked to cognitive decline [151]. This supports the view that cognitive capacity depends on the structural and functional organization of brain networks.

3.2.3 Brain Networks Integration and AGI

Adaptive cognition arises from flexible integration across brain modules. The frontoparietal network (FPN), for instance, dynamically routes information to support diverse cognitive demands [152, 153]. Analogously, AGI may benefit from architectures that mirror this modular integration. A central hub coordinating specialized AI modules akin to the FPN enables dynamic reconfiguration and task-specific generalization, essential for human-level intelligence.

Key Insight – From Brain Networks to AGI Architecture

Cognitive neuroscience reveals that intelligence arises from dynamic, flexible integration between brain networks. Translating these principles into AGI design via hybrid architectures, modular agents, and adaptive control hubs could enable machines to emulate human-like flexibility, reasoning, and learning.

3.2.4 Bridging Biological and Artificial Systems

AGI design must integrate symbolic reasoning with neural adaptability. While symbolic AI offers logical precision, it lacks flexibility. Conversely, neural networks excel at perception and pattern learning but lack interpretability [154]. Hybrid neuro-symbolic systems bridge this gap [64]. Innovations like Physics-Informed Neural Networks (PINNs)[155] and Kolmogorov–Arnold Networks (KANs)[156] exemplify architectures that embed domain knowledge into learning, improving generalization and robustness. These methods advance AGI by fusing logic, memory, and adaptivity.

Table 1: Mapping of human brain regions to neural network models and their functional parallels in AGI research.

Brain Region / Function	Cognitive Role	Neural Network Model	Application	Comparison Highlight
Occipital Lobe	Visual processing	Convolutional Neural Networks (CNNs)	Image recognition, object detection	Biological vision uses sparse, hierarchical filtering; CNNs apply layered filters for edges and textures
Hippocampus / Temporal Lobe	Memory encoding, sequence modeling	Recurrent Neural Networks (RNNs), LSTMs	Sequential modeling, time-series prediction	Humans recall context adaptively; RNNs capture limited temporal state
Motor Cortex	Voluntary motion control	Robotic Control Networks	Robotics, motor skill learning	Human motion uses proprioception and feedback; robotic policies rely on optimization
Prefrontal Cortex	Planning and decision making	Reinforcement Learning (RL)	Game playing, navigation, strategy tasks	Humans plan under uncertainty and values; RL focuses on reward maximization
Synaptic Plasticity	Learning through temporal dynamics	Spiking Neural Networks (SNNs)	Neuromorphic modeling, real-time inference	Hebbian/STDP rules guide human learning; SNNs simulate spikes with scalability trade-offs
Auditory Cortex	Language and speech understanding	Transformer Networks	Language modeling, translation, text generation	Humans integrate emotion and context; Transformers use token attention over sequences

4 Models of Machine Intelligence

Computational Intelligence (CI) encompasses a spectrum of machine learning frameworks aimed at endowing machines with cognitive capabilities comparable to humans [157]. Bridging inspiration from biological cognition and computational abstraction, CI integrates connectionist, symbolic, and hybrid models to support reasoning, learning, perception, and decision-making cornerstones of AGI development.

4.1 Learning Paradigms

Modern AI systems draw on a diverse suite of learning paradigms tailored to support generalization across tasks and domains. At the foundation lie supervised and unsupervised learning: the former relies on labeled examples to learn explicit mappings, while the latter uncovers latent structures from unannotated data [158]. Semi-supervised approaches combine scarce labeled data with abundant unlabeled samples to enhance representational quality. Self-supervised methods including pretext tasks [159] and contrastive learning refine feature embeddings by optimizing similarity–dissimilarity relations between input pairs.

To further boost adaptability, transfer learning enables knowledge acquired in one domain to expedite learning in related tasks [160], while meta-learning and continual learning allow rapid generalization and lifelong skill acquisition without catastrophic forgetting [161]. Reinforcement learning (RL) trains agents through trial-and-error interaction with dynamic environments [162]. Recent RL variants such as Learning to Think (L2T) introduce process-level, information-theoretic rewards that improve sample efficiency and general reasoning without task-specific annotations [163].

In AGI contexts, few-shot and zero-shot learning have emerged as essential capabilities for generalization from minimal supervision [164]. Multi-task and multimodal learning further enable cross-domain and cross-modal abstraction [165], while curriculum learning emulates human cognitive development through progressive task complexity [166]. Shortcut learning remains a cautionary lens, highlighting how models may exploit spurious cues instead of learning robust, generalizable patterns [167].

4.1.1 Representation Learning and Knowledge Transfer

At the heart of these paradigms lies representation learning the process by which models compress raw data into compact, task-relevant abstractions. Neural networks inherently perform this compression, enabling robust transfer across tasks. As shown in Figure 5, this mirrors the human brain's ability to encode generalized, symbolic concepts rather than raw sensory inputs [168]. Recent work [169] on compression–meaning tradeoffs suggests that LLMs often favor lossy statistical compression over semantic abstraction, casting doubt on their capacity for true understanding or generalization. Such compact compositional representations support adaptation, planning, and abstraction core ingredients for building versatile AGI systems.

4.1.2 Knowledge Distillation

Knowledge distillation is a model optimization technique that enables the transfer of capabilities from large teacher models to smaller student models, preserving performance while improving efficiency crucial for scalable AGI systems [170]. Distillation can be feature-based (aligning internal representations), response-based (matching output distributions), or relation-based (preserving structural dependencies). Variants like self-distillation, online distillation, and quantized distillation support continual learning and deployment in resource-constrained AGI environments.

Intelligence as a form of learning compressed representation

Intelligence can also be viewed as the capacity to compress high-dimensional data into abstract, low-dimensional representations [171]. This process involves extracting structure, eliminating redundancy, and preserving key patterns for reasoning and generalization.

4.2 Biologically and Physically Inspired Architectures

Below, we discuss biologically and physically inspired neural architectures.

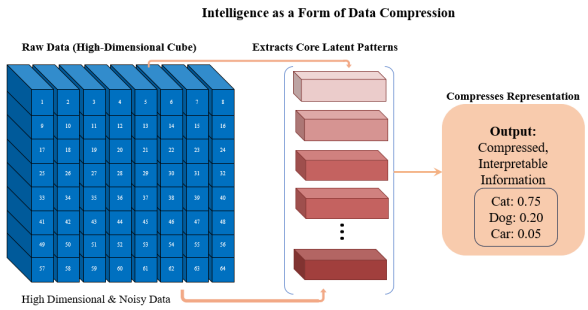

Figure 5: Illustration of intelligence as compression: noisy input (left) is distilled into latent abstractions (middle) and simplified outputs (right), enhancing generalization and reasoning.

Spiking Neural Networks (SNNs) emulate neural spike dynamics and are ideal for temporal and event-based processing [172]. Their biological plausibility supports neuromorphic computing and sensorimotor control.

Physics-Informed Neural Networks (PINNs) incorporate physical laws (e.g., Partial Diffrential Equations (PDEs)) into neural architectures [155], ensuring consistency with real-world constraints in domains such as fluid dynamics and biomechanics.

Kolmogorov-Arnold Networks Kolmogorov-Arnold Networks (KANs) [156] use learnable spline-based activation functions rather than fixed ones to model complex functions, shifting the learning emphasis from weights to activations. This enhances interpretability and flexibility but requires careful regularization for stable training. Table 2 and 3 summarizes the comparative strengths of SNNs, PINNs, and conventional neural networks across key AGI-relevant dimensions, including time modeling, biological plausibility, efficiency, and application scope.

4.2.1 Symbolic, Connectionist, and Hybrid Systems

Symbolic AI [61] excels in interpretability and rule-based reasoning but lacks robustness in perception. **Connectionist models** [173] (e.g., neural networks) offer scalable pattern recognition with less interpretability. Their fusion in **hybrid systems** [174] integrates structured reasoning with perceptual learning making them strong candidates for AGI architectures.

Key Insight: Toward Cognitive Foundations for AGI

The convergence of biologically plausible dynamics (SNNs), physically constrained reasoning (PINNs), symbolic-connectionist hybrids, and advanced learning paradigms marks a decisive step toward AGI. These models enable grounded abstraction, multi-task learning, and flexible adaptation beyond pattern recognition.

4.3 Intelligence as Meta-Heuristics

General intelligence can be viewed as a dynamic collection of meta-heuristics and adaptive strategies that continuously evaluate, revise, and optimize problem-solving pathways [175]. Unlike fixed heuristics [176], meta-heuristic agents improve iteratively by learning from failure and adapting strategies across domains. Recent AGI frameworks such as AutoGPT [177], and Voyager [178] demonstrate such behavior through internal feedback loops, self-prompting, and chain-of-thought reasoning. These systems optimize both task-specific performance and the broader process of learning itself, supporting transfer, adaptability, and generalization [179]. Intelligence, in this view, is not a static capacity but a recursive, self-improving search over heuristics.

4.4 Explainable AI (XAI)

As AI advances toward AGI, explainability must evolve from post hoc interpretation to intrinsic trans-

parency. Traditional techniques such as saliency maps and Grad-CAM provide limited insights into model reasoning [180, 181]. AGI systems, however, require explainability that mirrors human cognition enabling agents to articulate not just outcomes, but the rationale behind decisions [182].

This calls for architectural integration of interpretability through neuro-symbolic reasoning [183], causal modeling [184], and biologically inspired mechanisms such as memory traces and attention routing. Furthermore, multi-level explanations tailored to diverse user contexts are essential [154, 185]. Embedding meta-cognition and self-aware justification as core design principles will transform XAI from an afterthought to a foundational component of general intelligence.

5 Generalization in Deep Learning

Generalization in deep learning refers to a model's ability to extend learned patterns from training data to unseen scenarios, making it essential for AGI development [186]. Unlike narrow AI, which often overfits task-specific distributions, AGI systems must demonstrate robust transferability across domains and contexts [97].

5.1 Foundations of Generalization in AGI

Robust generalization is a cornerstone of AGI, enabling systems to adapt beyond their training distribution. Let P represent the training data distribution and Q the real-world distribution. The empirical risk R_{emp} measures training error, while R_{general} reflects expected real-world error. The generalization gap $R_{\text{emp}} - R_{\text{general}}$ captures how well a model extrapolates to new settings. A strong and robust AGI system should have lower generalization gaps. Theoretical frameworks have highlighted several different perspectives of generalization as follows.

Information Bottleneck (IB) theory proposes that models generalize by compressing inputs into compact latent representations that preserve only task-relevant information while discarding irrelevant or spurious signals [187]. This compression principle provides a trade-off between retaining predictive power and limiting unnecessary input information, thereby constraining model complexity. Shwartz-Ziv and Tishby [188] were among the first to empirically and theoretically propose that deep neural networks progressively compress representations as they learn, connecting this to improved generalization. Their follow-up work with Painsky [189] offered further theoretical support and a sample-complexity-oriented bound linking information compression to generalization. Building on these ideas, Kawaguchi et al. [186] later developed rigorous statistical learning bounds formalizing this principle in modern deep architectures. More recently, Shwartz-Ziv and LeCun [190] extended these information bottleneck arguments to the self-supervised learning paradigm, suggesting that compression not only benefits supervised generalization but also plays a key role in representation learning without labels. This sequence of work suggests that the information bottleneck is not only cognitively and biologically plausible but also grounded in solid mathematical and empirical evidence.

Minimum Description Length (MDL) is based on the idea that the simplest explanation or model that best compresses the data will generalize better [191]. MDL suggests that simpler models, which can compress data better, are less likely to overfit and thus generalize more effectively.

Implicit Regularization, often associated with stochastic gradient descent (SGD), suggests that optimization methods naturally bias models toward flat minima, which stems from the geometry of loss landscapes and provides insight into how generalization arises without explicit regularization [192].

Neural Tangent Kernel (NTK) and **Double Descent** theories together offer a modern understanding of generalization in overparameterized neural networks. NTK shows that as network width approaches infinity, training dynamics become linear and predictable, behaving like kernel regression and often leading to well-generalizing solutions despite large model sizes [193]. Double Descent complements this by revealing that increasing model capacity initially leads to overfitting near the interpolation threshold, but further scaling results in a second descent in test error with improved generalization [194].

PAC-Bayes Bounds combine elements of Bayesian inference with Probably Approximately Correct (PAC) learning [195]. They bound the generalization error of a hypothesis based on its divergence from a prior, typically measured via the Kullback-Leibler (KL) divergence.

Causal Representation Learning emphasizes learning representations that capture the causal structure of data, rather than mere statistical correlations [196]. It uses tools from causal inference, such as structural equation models and do-calculus, to extract invariant features under interventions.

Variational Dropout is a Bayesian regularization method that interprets dropout as approximate variational inference [197]. It injects noise into the model's weights using a learnable distribution, often leading to sparsity and robustness. Unlike fixed dropout rates, variational dropout adapts the noise level during training, improving generalization in uncertain or noisy environments.

Table 2: Architectures and Generalization Theory in AGI: (A) neuro-inspired and physics-informed designs (e.g., SNNs, PINNs); (B) theoretical constructs (e.g., IB, MDL, NTK).

Panel A: Neuro-Inspired and Physics-Grounded Architectures			
Architectures	**SNNs**	**PINNs**	**Conventional NNs**
Property	Simulate spike-timing and event-driven signaling	Encode physical constraints within neural units	Abstract artificial neurons using trainable weights
Time Dynamics	Temporal encoding via spikes	Task-driven implicit time representation	Often absent unless RNNs are used
Computation Paradigm	Event-based, energy-efficient processing	PDE-constrained data fitting	Data-driven general-purpose mapping
Biological Alignment	High (plasticity, sparsity)	Moderate (physics realism)	Low (flexible but abstract)
Efficiency	Moderate; optimized	Dependent on solver complexity	High throughput/GPU parallelism
Use Cases	Edge robotics, dynamic sensing	Scientific simulation, climate modeling	Vision, NLP, reinforcement learning
AGI Potential	Real-time perception	Symbol grounding via physics	Scalable pattern abstraction

Panel B: Theoretical Constructs for Generalization			
Theory	**Inductive Principle**	**Foundation**	**Implication for AGI**
Information Bottleneck (IB)	Focus on relevant latent features while discarding noise	Information theory, mutual information	Compact, task-relevant representation learning
Minimum Description Length (MDL)	Simplicity favors generalization	Algorithmic info theory	Selects compressed, interpretable models
Implicit Regularization (SGD)	Flat minima during optimization	Loss landscape geometry	Encourages generalization
NTK / Double Descent	Overparameterized regimes benefit late generalization	Infinite-width kernel theory	Characterizes regimes of robust learning
PAC-Bayes Bounds	Generalization from distributional priors	Probabilistic learning theory	Formal generalization guarantees
Causal Representation Learning	Extracts stable causal features invariant to interventions	Causal graphs, SEMs	Promotes robustness across tasks/distributions
Variational Dropout	Regularizes through learned noise injection	Variational inference	Enforces sparsity and noise resilience
Simplicity Bias	Learns simpler hypotheses first	Empirical dynamics of training	Lower complexity early in training

Table 3: Optimization and Priors in AGI: (C) learning algorithm biases (e.g., SGD, RL, PEFT); (D) emerging priors in foundation models (e.g., RAG, MAE, RLHF).

Panel C: Learning Algorithms and Loss Function Biases			
Mechanism	**Inductive Bias**	**Examples**	**Relevance to AGI**
SGD / Early Stopping	Implicit preference for flatter minima	Classic training setups	Generalizable, stable convergence
Adaptive Optimizers (Adam, RMSProp)	Faster convergence but risk of sharp solutions	LLM fine-tuning, low-data setups	Tradeoff between speed and generalization
Cross-Entropy Loss	Promotes confident predictions	Classification tasks	Simple yet insensitive to uncertainty
Contrastive / Triplet Loss	Latent clustering, relational structure	SimCLR, MoCo, triplet nets	Robust representation learning
KL Divergence (in VAEs, PAC-Bayes)	Regularizes latent space or distributions	VIB, Bayesian networks	Encourages minimal, disentangled codes
RL Objectives	Long-term credit assignment, goal focus	PPO, Q-learning, DPO	Supports planning and sequential reasoning
Meta-Learning / PEFT	Task-agnostic initialization or fast adaptation	MAML, LoRA, Reptile	Enables efficient few-shot or continual learning

Panel D: Emerging Inductive Priors in Foundation Models			
Mechanism	**Inductive Bias**	**Examples**	**AGI Relevance**
Multimodal Attention	Enables alignment across modalities	CLIP, Flamingo, Perceiver IO	Supports grounded reasoning and perceptual understanding
Cross-Modal Contrastive Learning	Aligns visual and language embeddings via shared structure	ALIGN, LiT, GIT	Encourages shared representations and compositionality
External Memory Augmentation	Facilitates long-term and episodic recall	RNN+Memory, ReAct, RETRO	Enables scalable context and symbolic chaining
Retrieval-Augmented Generation (RAG)	External database during inference	RAG, Atlas, KAT	Enhances factuality/adaptability
Masked Modeling / Autoregression	Learns predictive structure from partial context	BERT, GPT, BEiT, MAE	General-purpose self-supervised pretraining
Prompt Tuning and Instruction Biases	Learns structure through task prompts or instructions	T5, InstructGPT, PEFT, Prefix Tuning	Provides zero-shot adaptation and alignment with user intent
RL with Human Feedback (RLHF)	Aligns model outputs with human values/preferences	InstructGPT, DPO, Constitutional AI	Critical for safety and value alignment

Simplicity Bias refers to the empirical observation that deep networks, when trained with gradient descent, tend to learn simpler functions before complex ones [198]. This bias arises from the implicit properties of parameter-function mappings and the dynamics of neural network training. As a result, models are more likely to converge to functions with lower complexity, which tend to generalize better.

Generalization: A Pillar of AGI

Effective generalization not just memorization distinguishes AGI from narrow AI. Theories like the Information Bottleneck, minimum description length, and optimization landscapes converge on one idea: compress inputs to extract robust, transferable representations.

5.2 Architectural and Algorithmic Inductive Biases

Inductive biases embedded in model architectures and learning algorithms are central to the design of AGI systems, guiding how they learn, generalize, and reason. For example, linear models offer interpretability but are limited in capturing nonlinear patterns [171]. MLPs support hierarchical representations but lack spatial or temporal priors [199]. CNNs introduce local spatial bias and translation invariance ideal for vision while RNNs model sequences but struggle with long-range dependencies [200]. Transformers [201], with global attention, excel at long-range modeling and underpin modern LLMs like GPT [202], though they lack grounded abstraction. State-space models (e.g., Mamba) offer implicit recurrence and dynamic memory [203], improving temporal scalability. GNNs encode relational priors for graph-structured tasks [204], and GANs [205] support powerful generative modeling, albeit with stability trade-offs.

5.2.1 Biases in Learning Algorithms

Learning algorithm biases also play a vital role. Optimization methods like SGD favor flat minima with better generalization [206], while adaptive optimizers like Adam can converge faster but bias toward sharper solutions [207]. Loss functions impose task-specific priors: cross-entropy for classification, contrastive losses for relational tasks, and adversarial or reinforcement losses for realism and long-term planning [208]. Meta-learning and structured losses promote compositionality and generalization across tasks essential traits for AGI. A unified AGI architecture may need to integrate these diverse inductive structures to achieve abstraction, compositionality, and adaptive reasoning across modalities and tasks.

5.2.2 Solving Inductive Bias Technique

AGI systems must generalize not only across tasks but also across distributions, time, and embodiment. Techniques to enhance this capability include **uncertainty estimation**, which accounts for epistemic and aleatoric uncertainty to improve reliability [209] (further discussed in Section X), and **adaptive regularization** mitigates catastrophic forgetting in continual learning [210].

5.3 Generalization During Deployment

Test-Time Adaptation (TTA) refers to techniques that enable machine learning models to dynamically adjust their predictions at inference time, aiming to improve robustness to distributional shifts or domain changes encountered during deployment [211]. There are two primary paradigms within TTA: optimization-based TTA and training-free TTA.

Optimization-based TTA involves updating certain model parameters, typically through gradient descent, at test time, using unsupervised or self-supervised objectives derived from the test data itself, such as test-time training (TTT) [212] and test-time prompt tuning (TPT) [213].

Training-free TTA improves model adaptation at test time without performing any explicit parameter updates or gradient-based optimization. Instead, these methods rely on recalibrating or modifying the model's inference process, such as training-free dynamic adapter (TDA) [214] and dual memory network (DMN) [215]

Retrieval-Augmented Generation (RAG) augments model predictions by incorporating information retrieved from large external databases, document corpora, or knowledge bases during inference [216, 217]. Instead of relying solely on the parametric memory of the model, RAG retrieves relevant documents or facts in response to a query or input and conditions the model's output on both the original input and the retrieved evidence. RAG can improve factual accuracy and reduce hallucination without requiring additional model retraining, but challenges include efficient retrieval, handling noisy evidence and latency during inference.

Deployment-Time Generalization

For AGI to succeed in dynamic environments, continual adaptation is essential. Techniques like TTA and RAG offer real-time resilience through knowledge retrieval, error correction, and ongoing learning.

5.4 Toward Real-World Adaptation

Embodied Intelligence To achieve real-world adaptation, AGI systems must bridge the gap between abstract reasoning and physical interaction. This requires the integration of perception, planning, and control to enable flexible behavior in dynamic environments. Techniques such as imitation learning and zero-shot planning are instrumental for equipping robots and embodied agents with the ability to generalize learned knowledge to novel tasks and contexts, thereby enhancing adaptability and autonomy in robotics applications [218].

Causal Reasoning Robust adaptation necessitates distinguishing causation from mere correlation, a challenge addressed by the causal inference frameworks pioneered by Pearl and Bengio [184]. Causal reasoning allows AGI to identify and model underlying mechanisms, supporting effective generalization across distribution shifts and facilitating reliable interventions in complex, uncertain environments.

Robustness and Alignment AGI must be resilient to rare, high-impact "black swan" events that are difficult to anticipate but potentially catastrophic. Ensuring robustness involves the capacity for safe exploration, rapid adaptation to unforeseen scenarios, and continual monitoring for emergent risks. At the same time, alignment mechanisms are critical to guarantee that AGI systems consistently act in accordance with human values and intentions, even in the face of novel and ambiguous circumstances [219].

6 Reinforcement Learning and Alignment for AGI

"The measure of intelligence is the ability to change" **(Albert Einstein)**. This insight underscores a limitation of static neural networks: true intelligence demands adaptability. Reinforcement learning (RL), which enables agents to learn by interacting with their environment and adapting through feedback, captures this essence [220, 221]. Unlike supervised learning, which relies on fixed datasets, RL thrives in non-stationary, uncertain environments, making it a natural candidate for AGI [222].

The Core of AGI: Learning by Doing in Real -Time

RL's foundation lies in its trial-and-error paradigm, promoting continual, adaptive learning through experience.

6.1 Reinforcement Learning: Cognitive Foundations

While RL offers a promising path toward adaptive intelligence, its direct application to AGI is hindered by several limitations, including sample inefficiency, limited scalability in high-dimensional spaces, and vulnerability to reward misspecification [222, 33]. To address these concerns, algorithmic strategies have been developed.

Model-based RL incorporates predictive dynamics to reduce sample complexity [221], while **hierarchical RL** decomposes tasks into reusable subtasks for more efficient exploration and planning [162]. Complementing these advances, cognitive reasoning methods inspired by LLMs significantly expand RL's expressive capacity.

Recent methods such as **Chain-of-Thought (CoT)** [20], **Tree-of-Thought (ToT)** [21], and **Reasoning-Acting (ReAct)** [18] embed structured, deliberative reasoning within RL pipelines. CoT enables transparent multi-step inference; ToT explores multiple solution paths to improve policy selection; and ReAct integrates reasoning with environment interaction, reducing errors and enhancing adaptability. These methods mitigate short-term bias and inefficient exploration, aligning RL agents more closely with the demands of general intelligence [48].

Integrative frameworks exemplify this convergence of RL and LLM reasoning:

- **MetaGPT** [223]: Coordinates multiple LLM agents in specialized roles, facilitating structured task decomposition and collaborative problem-solving.
- **SwarmGPT** [224]: Combines LLM planning with multi-agent RL for real-time coordination in systems such as robotic swarms.
- **AutoGPT** [177]: Demonstrates autonomous goal decomposition, iterative self-correction, and continuous self-improvement via internal RL loops.

Supporting these frameworks are optimization strategies such as:

- **Proximal Policy Optimization (PPO)** [51]: Balances policy performance with stability.
- **Direct Preference Optimization (DPO)** [52]: Trains agents directly from preference data, simplifying alignment.
- **Group Relative Policy Optimization (GRPO)** [53]: Optimizes reasoning quality by comparing multiple generated trajectories.

6.2 Human Feedback and Alignment

Reinforcement Learning with Human Feedback (RLHF) [225] addresses AGI alignment by incorporating human judgments into the reward loop, improving safety and reducing harmful outputs [226, 227]. RLHF underpins systems like InstructGPT and ChatGPT, though challenges remain in scaling feedback and mitigating biases.

6.2.1 Alignment Techniques and Supervision

Human-in-the-loop training, **value learning**, and **inverse reinforcement learning** enhance AGI's alignment with human values [228]. Online supervision allows real-time adaptation [229], while offline supervision enables reflective policy refinement without continuous oversight [230, 231, 232]. Additionally, machine unlearning [233] has emerged as a corrective tool for removing spurious correlations, hallucinations, or biased representations in vision-language models, contributing to safer and more interpretable systems [234].

6.2.2 Ethical Issues of AGI

As AGI systems approach greater autonomy and capability, ensuring fairness, transparency, trust, and privacy becomes not only a technical imperative but also a societal one [235, 5, 165]. These principles form the ethical backbone of safe AGI deployment, safeguarding individuals and communities from disproportionate harms such as surveillance, exclusion, or algorithmic manipulation. To address these challenges, governance frameworks must be grounded in human rights and international norms [236, 237]. These frameworks must go beyond technical safeguards by incorporating participatory design, redress mechanisms, and interdisciplinary oversight. Without such structures, AGI risks reinforcing existing inequities, centralizing power, and becoming unaccountable in high-stakes decisions.

6.2.3 Future Outlook

Future alignment strategies must integrate multidisciplinary insights from AI, ethics, psychology, and law [238, 25]. As shown in Figure 8(a), AGI readiness hinges on cognitive, interface, systems, and alignment axes. Figure 8(b) shows expert uncertainty, with 37% expecting AGI realization in two decades or more [26]. Cross-cultural modeling, robust evaluation, and international coordination will be critical.

7 AGI Capabilities, Alignment, and Societal Integration

AGI seeks to replicate core human cognitive abilities reasoning, learning, memory, perception, and emotion to operate autonomously across domains [26]. Beyond technical capability, safe deployment requires alignment with ethical principles and social values. This section synthesizes cognitive foundations, psychological insights, and governance frameworks that shape AGI's path toward responsible integration [239].

AGI Integration at a Glance

Cognitive Core: Reasoning, learning, memory, and perception underpin AGI adaptability.
Safety: Robust design, value alignment, and human-in-the-loop controls remain essential.
Psychological Grounding: Cognitive science guides realistic and ethical agent behavior.
Governance: Frameworks like NIST, EU AI Act, and OECD foster transparent oversight.
Equity: "AI for everyone, by everyone" reflects the need for co-design and fair access.

7.1 Core Cognitive Functions

7.1.1 Reasoning

AGI systems must perform deductive, inductive, and abductive reasoning to solve novel problems [240, 35]. Deep reasoning enables hypothesis testing, planning, and counterfactual inference[241]. Models like chain-of-thought and neuro-symbolic systems integrate symbolic logic with neural learning for more interpretable and adaptive reasoning [242, 243, 244].

7.1.2 Learning

AGI integrates supervised, unsupervised, symbolic, reinforcement, and deep learning paradigms [245, 246]. These enable generalization and continuous refinement. Reinforcement learning facilitates interaction-based learning in dynamic environments [247], while deep learning abstracts features across modalities [248].

7.1.3 Thinking

Thinking refers to abstraction, strategy formation, and decision-making. Cognitive architectures and neural networks simulate high-level thought [249]. Neuro-symbolic systems combine formal logic with adaptable models [250], increasing reliability in complex reasoning tasks [251].

7.1.4 Memory

Memory supports context awareness and learning continuity. Short-term memory aids in immediate task handling; long-term memory encodes cumulative knowledge [78, 252]. Parametric and external memory systems allow rapid retrieval and flexible updates [71].

7.1.5 Perception

AGI perception involves multimodal sensory interpretation. CNNs and transformers process visual and auditory signals[253]. Advances in multimodal models like Perceiver and Flamingo improve AGI's ability to interpret heterogeneous inputs[254].

7.2 Human-Centered Foundations: Psychology and Safety in AGI Design

The safe deployment of AGI requires more than technical ingenuity; it demands architectures informed by a realistic understanding of human cognition [33]. Cognitive psychology reveals mechanisms such as attention, memory consolidation, emotion regulation, and causal reasoning [255, 256], which inform AGI's design and behavior modeling. Concepts like incremental learning and theory of mind [257, 258] offer blueprints for developing adaptive, socially attuned agents. However, naively importing psychological concepts can introduce anthropomorphic biases or flawed heuristics [259]. A human-centered AGI must be empirically grounded, cross-culturally aware, and sensitive to normative variation [260].

Safety concerns are deeply intertwined with these human-centered foundations. AGI's open-ended generalization capabilities heighten the risk of unintended behavior [261]. Key dimensions include technical robustness (resilience to adversarial inputs), specification soundness (goal alignment), and human control (corrigibility, intervenability) [262]. Research in scalable oversight [263], reward modeling [264], and uncertainty calibration [265] seeks to systematically mitigate these vulnerabilities.

Ultimately, AGI systems must not only learn, plan, and reason but also reflect, defer, and ask for help [260]. Embedding interpretability, human-in-the-loop safeguards, and NSFW (Not Safe for Work) content filters [266] is essential for preserving public trust. Building AGI that is intelligent, safe, and aligned begins with understanding the minds it aims to augment, not replace. Table 4 outlines major evaluation benchmarks, bio-inspired system mappings, and emerging governance frameworks [154].

7.3 Societal Integration and Global Frameworks

The transition of AGI from lab to society raises urgent questions regarding equity, human agency, and democratic oversight, as shown in the Algorithm 3.

Work and Autonomy: AI is not only transforming manual labor but increasingly encroaching on cognitive, techical and emotional domains. Recent studies reveal that prolonged LLM use in educational settings leads to measurable cognitive debt, marked by reduced neural engagement, memory recall, and authorship awareness [267].

As intelligent agents begin to mediate professional and personal routines, these shifts raise profound questions about identity, equity, and the structure of work [238]. The World Economic Forum estimates that up to 87% of data-driven tasks could be automated by AGI [268], while leading AI developers suggest that most white-collar roles are now within reach of current-state-of-the-art models. These trends underscore the urgency of designing inclusive systems and proactively reimagining labor, education, and welfare infrastructures to ensure a just transition.

Public Trust Public sentiment oscillates between promise and peril. While AGI-augmented healthcare and education spark hope, concerns about surveillance and job loss demand transparent oversight, participatory development, and community-driven evaluation [269].

Policy Infrastructure Several governance frameworks are converging to guide AGI deployment. The NIST AI RMF [270] promotes trustworthiness through interpretability and risk mitigation. The EU AI Act enforces risk-tiered compliance in high-stakes sectors. UNESCO and OECD advocate global ethical standards rooted in inclusivity, safety, and accountability [271].

AI for Everyone, by Everyone As AGI systems become more powerful, their development must reflect diverse societal needs and values [272]. The principle of "AI for everyone, by everyone" underscores the importance of participatory design, equitable access to AI resources, and co-governance across disciplines and geographies. Open-source models, community auditing, and culturally tuned datasets are crucial to democratize AGI and avoid reinforcing power asymmetries. **Constructive Examples** Early signs of responsible integration include AI tutors, digital mental health agents, and scientific co-reasoners [273]. These applications demonstrate the potential of AGI to increase expertise, but also underscore the need for accountability in decision-making pipelines.

Toward Co-Designed Futures To ensure that AGI advances human flourishing, it must be co-developed with ethicists, legal scholars, and the public. Embedding AGI within sociotechnical ecosystems [274], through cross-disciplinary governance, inclusive norms, and transparent validation, will be critical to building systems that are not only intelligent, but also wise [275].

7.4 LLM's, VLM's and Agentic AI

Large Language Model (LLM), Vision-Language Model (VLM)and Agentic AI have a fundamental role to play in the advancement towards AGI systems. LLM's capability of natural language understanding and VLM's which can combine visual and textual information together support the development of autonomous, adaptable and context aware AI agents that serve as the driving force for AGI. In this regard, this section discusses notable AI frameworks and models which are available currently followed by a discussion on VLMs and agentic AI as a pathway towards AGI. One of the key techniques that enables such agentic behavior is the Tree-of-Thought reasoning framework, which equips models with the ability to explore, evaluate, and revise multiple reasoning paths. A generalized outline of this structured decision-making approach is presented in Algorithm 3.

Algorithm 3: Tree-of-Thought Reasoning

Input: Problem description P
Output: Final solution path S

1. Initialize root thought with task prompt
2. Expand nodes with plausible reasoning paths
3. Evaluate each path using scoring heuristics or LLM feedback
4. Apply lookahead and backtracking to prune low-reward branches
5. Select optimal reasoning trajectory S

7.4.1 VLMs and Agentic AI as a pillar for the future AGI Framework

VLMs represent a pivotal advancement in AI by integrating visual perception and linguistic understanding, enabling tasks like captioning, visual question answering, and multimodal reasoning [294, 295]. Rooted in early computer vision (e.g., object detection [296]) and NLP research (e.g., machine translation), initial approaches were constrained by their unimodal focus [297]. The creation of paired datasets like Pascal VOC and Flickr30k [298, 299] enabled learning associations between images and text. This led to the emergence of early VLMs, which combined CNN-RNN pipelines for captioning and VQA, though they often lacked deeper semantic understanding [294]. A paradigm shift occurred with the Transformer architecture [201], unifying NLP and vision through self-attention. This enabled models like BERT [300] and ViT [301] to advance multimodal understanding, forming the backbone of contemporary VLMs increasingly applied in domains, such as robotics, medicine, and assistive technologies [302].

Table 4 (panel B) presents a roadmap connecting brain-inspired principles to the development of AGI via VLMs. Key *brain functions* such as neocortical reasoning and hippocampal spatial memory [282, 283] are reflected in transformer-based architectures that employ cognitive modularity and attention mechanisms [284], paving the way for neuro-symbolic planning [61] and cognitive digital twins in medical diagnostics [303]. The brain's *memory hierarchies*, which transition from sensory encoding to long-term storage [285], are represented in VLMs through contextual embeddings and dynamic prompt extensions [286], supporting lifelong learning and adaptive tutoring systems. In terms of *action systems*, the integration of mental and physical processes [287] is emulated by multi-agent VLMs and vision-action loops [94, 304]. Finally, *world models*-compact internal representations for prediction and planning [288, 289]-are realized through multimodal embeddings and simulator-based architectures, supporting anticipatory agents for household and space missions [94]. Together, these components illustrate how brain-inspired VLMs can advance AGI through the integration of embodied reasoning, hierarchical memory, and goal-directed action.

The adoption of Transformers enabled VLMs to process images and text using unified self-attention architectures, significantly enhancing multimodal integration [305]. Contrastive learning approaches, as in CLIP and ALIGN, align image-text pairs in shared embedding spaces for robust general-purpose representations [166, 306]. Scaling up with models like Flamingo, PaLI, and LLaVA introduced few-shot learning, multimodal dialogue, and state-of-the-art performance on diverse tasks [307, 308, 12].

Figure6(a) presents the chronological evolution of VLMs following the release of ChatGPT in late 2022. These models have rapidly advanced in terms of scale, multimodal comprehension, and cross-domain generalization [309]. Current state-of-the-art VLMs support a wide spectrum of capabilities including visual question answering, captioning, visual reasoning, and image-to-text alignment. In applied domains, they have been deployed for robotic instruction following, autonomous navigation, and assistive dialogue agents. A critical advantage of VLMs lies in

Table 4: Panel A presents representative benchmarks for AGI evaluation. Panel B maps biologically inspired cognitive functions to vision-language and agentic AI systems. Panel C outlines global governance frameworks for safe, ethical, and equitable AGI deployment.

Panel A: Representative Benchmarks for AGI Evaluation

Benchmark	Focus	Capabilities Tested	Notable Feature	Modality	Interactivity Level
BIG-Bench [276]	Language reasoning	Multitask generalization, logic, math	Human-written diverse tasks	Language	Static
ARC [277]	Abstract reasoning	Concept composition	System-2 style generalization	Visual, Symbolic	Static
MineDojo [278]	Embodied AI	Planning, exploration	Minecraft sandbox environment	Multimodal	Interactive
BabyAI [279]	Language grounding	Navigation, planning	Curriculum-based instructions	Language + Embodied	Interactive
Agentbench [280]	LLM agents	Tool use, dialogue	Multi-agent evaluation	Language + Tools	Real-time
AGI-Bench [63]	AGI evaluation	Multimodal generalization	Multi-domain tasks	Multimodal	Mixed
eAGI [281]	Engineering cognition	Reasoning, synthesis, critique	Bloom-level tasks with structured design inputs	Text + Diagrams	Mixed

Panel B: Mapping Brain-Inspired AGI Functions to Vision-Language and Agentic AI Architectures

AGI Function	Biological Inspiration	VLM Representation	Agentic AI Mechanism	Development Pathway	Future Applications
Brain Functions	Neocortex (reasoning), Hippocampus (memory), Cerebellum (motor control) [282, 283]	Transformer attention modules simulating cortical modularity [284]	Autonomous agents with role-based communication and planning [75]	Neuro-symbolic cognitive architectures unifying language and perception	Cognitive robotics, brain-inspired diagnostics, and human-AI collaboration
Memory Systems	Hierarchical short- and long-term memory; working memory dynamics [285]	In-context retrieval, memory tokens, and dynamic prompt chaining [286]	Persistent memory, episodic task replay, and continual learning agents [74]	Meta-memory and lifelong memory consolidation frameworks	Adaptive tutoring systems, emotional-aware assistants, and digital memory augmentation
Action Systems	Cognitive imagination, motor planning, and physical interaction [287]	Scene-grounded VLM control with vision-to-action APIs [94]	Task-specialized agents under orchestration and multi-agent tool-use [77]	Embodied perception-action systems in real and virtual environments	Autonomous robotics in healthcare, manufacturing, and creative co-design
World Modeling	Internal generative simulation, counterfactuals, predictive coding [288, 289]	Multimodal latent embeddings and temporal scene simulation [94]	Self-play reasoning and task generation (e.g., AZR [290]) with verifiable feedback [291]	Causal inference and forward-planning agents for open-ended tasks	Scientific reasoning, autonomous experimentation, AGI research copilots

Panel C: Societal Frameworks and Policy Instruments for AGI Deployment

Framework	Institution/Origin	Principles	Key Areas Addressed	Scope	Enforcement Strategy
EU AI Act [25]	European Commission	Risk-based tiers, human oversight, transparency	High-risk system regulation, employment, health, surveillance	Regional (EU)	Legal compliance with penalties
NIST AI RMF [270]	U.S. NIST	Trustworthiness, transparency, risk mitigation	Security, privacy, robustness, explainability	Voluntary (U.S.)	Self-assessment, toolkits
OECD AI Principles [292]	OECD Nations	Human-centered values, safety, accountability	Innovation vs. risk balance, cross-border alignment	Global	Member-state adoption
UNESCO AI Ethics [271]	UNESCO	Equity, inclusiveness, sustainability	Socioeconomic impact, environmental, cultural diversity	Global	Advisory with monitoring reports
IEEE ECPAIS [293]	IEEE Standards Association	Transparency, accountability, bias mitigation	Algorithmic audits, ethical design	Industry-wide	Standardization, audit checklists

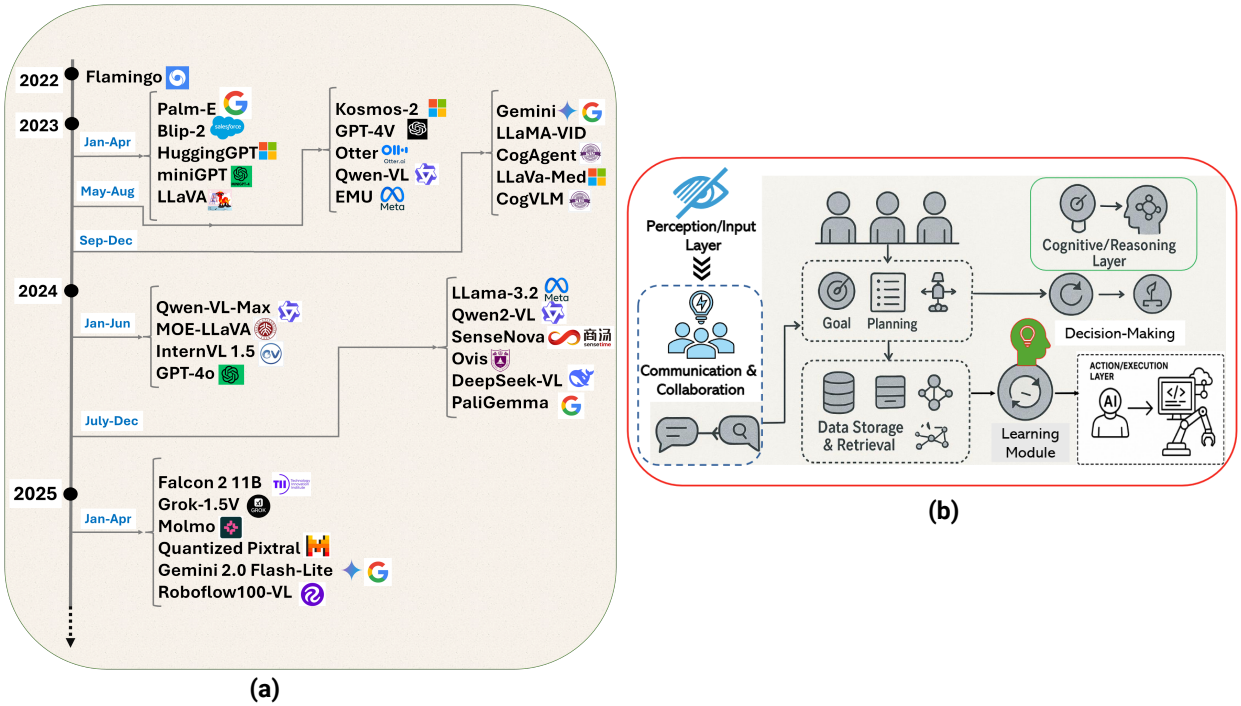

Figure 6: (a) Chronological evolution of VLMs following the release of ChatGPT in late 2022. The timeline highlights key VLM developments across major research labs and companies, organized by quarterly intervals from 2022 through early 2025. (b) Illustrating a visual overview of core functionalities in Agentic AI which is a key to AGI. This figure depicts the layered structure through which AI agents perceive inputs, make decisions, execute actions, and engage in learning and coordination to operate effectively in both individual and collaborative settings (Agentic System/MAS).

their ability to translate perception into semantically rich representations, enabling downstream reasoning and decision-making. Yet, despite these advances, VLMs alone cannot fulfill the requirements of AGI. They excel at perception and interpretation, but lack structured autonomy, persistent memory, and adaptive goal management. To truly transition from perception to intelligent action, VLMs must be embedded within broader Agentic AI architectures, where decision-making, coordination, and learning unfold across layered cognitive processes.

Figure 6 (b) illustrates this complementary architecture. At the core of Agentic AI lies a modular framework where VLMs serve as the perceptual interface detecting objects, interpreting environments, and feeding this information into a cognitive reasoning layer. This is followed by modules for goal formulation, planning, and data storage and retrieval, which maintain contextual coherence across tasks. Agents then utilize learning modules for continuous adaptation, drawing on episodic and semantic memory to inform future actions [73, 77]. Through collaboration and communication modules, agents interact within multi-agent systems (MAS), enabling distributed problem-solving and collective intelligence [74]. The decision-making layer synthesizes insights from upstream modules, and the action execution layer interfaces with external actuators or APIs to carry out commands. This layered system ensures that agent behavior is not just reactive but context-aware, goal-driven, and self-refining hallmarks of AGI. As these systems mature, Agentic AI will increasingly enable long-horizon autonomy in fields such as scientific discovery, healthcare, and adaptive robotics. By combining VLMs for perception with agentic architectures for reasoning and execution, we move closer to AGI systems that not only perceive and describe the world but also act within it with purpose, adaptability, and alignment with human values.

Additionally, the future of AGI hinges not just on increasing model scale or parameter count, but on the emergence of Agentic AI systems endowed with autonomy, memory, tool-use, and decision-making capabilities that mirror core aspects of human cognition [77]. Unlike static models that simply respond to prompts, Agentic AI systems act, plan, reflect, and adapt over time [310, 77]. Several promising frameworks illustrate this paradigm shift: AutoGPT [177] orchestrates sequential tool calls using a planner-reflector loop; BabyAGI implements a task prioritization loop with a vector-based memory store; CAMEL

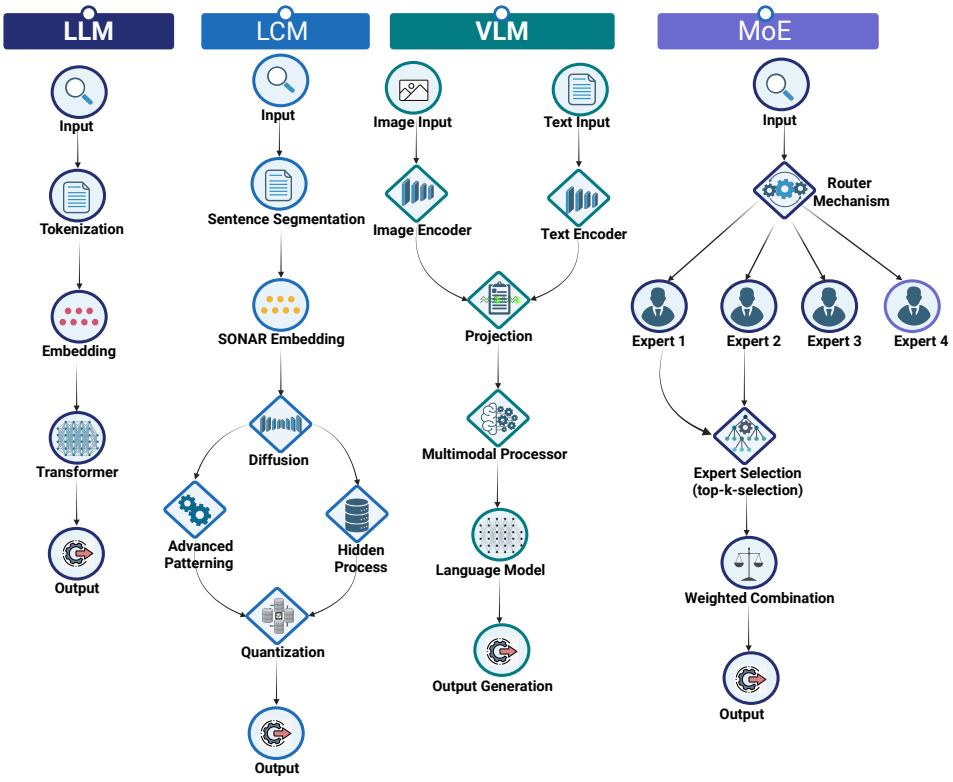

Figure 7: Conceptual overview of core foundation model architectures. The architectural pipelines of Large Language Models (LLMs), Language-Centric Models (LCMs), Vision-Language Models (VLMs), and Mixture of Experts (MoE).

(Communicative Agents for Mind Exploration of Large-scale language models) enables multiple agents to coordinate via natural language dialogue [311]; ReAct fuses reasoning and acting through intermediate reasoning traces [18]; and OpenAGI integrates goal-oriented decision-making with tool use and memory retrieval [312]. Each of these systems demonstrates attributes critical to AGI, including context persistence, agent collaboration, and feedback-guided learning. When integrated with VLMs such as LLaVA [12], Flamingo [307], or Kosmos-2 [313], these agents acquire perceptual grounding in real-world environments, enabling a more adaptive and embodied form of intelligence.

VLMs enable agents to interpret multimodal data, including images, text, and videos, while reasoning about this information in a human-like manner [304]. For example, an embodied agent equipped with VLM capabilities can interpret its environment, plan actions, and learn through interactions, mirroring how humans link perception and motor actions. This convergence is already evident in domains like robotics, assistive medical agents, and multi-agent research systems. However, a critical bottleneck persists: most current agentic systems depend on human-curated tasks, externally defined reward signals, or fine-tuned supervision, limiting their long-term autonomy and adaptability. For AGI to emerge, these agents must evolve beyond being mere tool-users; they must become self-motivated learners, capable of generating, testing, and refining their own reasoning processes. This is where the Absolute Zero paradigm presents a transformative shift.

The AZR introduces a self-evolving agentic AI paradigm that discards dependence on human-labeled tasks by autonomously generating, solving, and validating its own reasoning problems using a code execution engine [290]. Built on Reinforcement Learning with Verifiable Rewards (RLVR) [314], AZR supports outcome-based, self-verifying learning without external supervision. Its meta-cognitive curriculum design enables continuous skill refinement by identifying and addressing its own reasoning gaps. AZR is both model-agnostic and scalable, making it adaptable for integration into larger agentic ecosystems such as multi-agent research assistants or autonomous robotics. Empirically, it achieves state-of-the-art performance on mathematical and code reasoning benchmarks, outperforming traditional zero-shot models. By enabling AI systems to improve through introspective feedback rather than curated data, AZR advances AGI toward reflective, self-

directed learning, pushing AI closer to human-like, adaptive, and open-ended intelligence.

In summary, future AGI will likely take the form of a self-improving, multimodal system capable of autonomous reasoning, adaptive learning, and goal-directed behavior across diverse, open-ended environments, integrating agentic AI, structured memory, and world modeling to emulate human-like cognition.

8 Recent Advancements and Benchmark Datasets

The pursuit of AGI has recently entered a phase defined by the emergence of increasingly general, autonomous, and multi-capable systems [315]. This section highlights several of the most prominent conceptual frameworks and approaches that reflect current trends in AGI design blending planning, reasoning, memory, and environmental interaction in novel ways. This is followed by a discussion on data, which is essential for AGI development.

8.1 Advancements Beyond Large Language Models

The progression toward AGI, as depicted in Figure 8 necessitates overcoming the inherent limitations of current LLMs, which primarily rely on autoregressive next-token prediction. While this approach facilitates multi-task learning [316, 317], it may not fully capture complex human cognitive processes, such as intuition and ethical reasoning [98, 318]. Figure 1 illustrates AI's evolution since the 1950s, highlighting milestones where AI systems have matched or exceeded human-level performance across various domains. This historical trajectory underscores the accelerating pace of AI development, suggesting that future advancements may continue to outpace human capabilities.

The reliance on scaling laws [319], indicates that while increasing model size and training data enhances performance, this approach encounters diminishing returns [14]. Sustained scaling requires exponentially greater computational resources for increasingly marginal gains, and fundamental human abilities, such as creativity and moral reasoning, may not be effectively captured through scaling alone. This limitation underscores the need to explore more advanced learning mechanisms and architectural innovations capable of addressing the ethical and intuitive dimensions of intelligence.

8.1.1 AI Agent Communication Protocols

As the field advances towards AGI, robust and interpolable communication between autonomous AI agents has emerged as a critical enabler. Recent few foundational agent communication protocols such as the model context protocol (MCP) Source Link, the agent communication protocol (ACP) Source Link, the Agent2Agent protocol (A2A) Source Link, and the agent network protocol (ANP) Source Link represent key milestones in the development of scalable, compositional, and collaborative agent ecosystems.

MCP, pioneered for LLM-centric systems such as OpenAI's Assistants API, standardizes how models receive external tools and context through secure, typed JSON-RPC interfaces [320]. This enhances context-awareness during inference and allows modular tool mounting, a cornerstone for generalizable intelligence. ACP further advances this by enabling REST-native, session-aware messaging between heterogeneous agents with structured MIME-typed payloads, fostering reliable multimodal coordination. A2A introduces a peer-to-peer framework where agents advertise capabilities via dynamic "Agent Cards" and negotiate task delegation through structured artifacts Source Link. This supports fine-grained collaboration between agents across frameworks and vendors, promoting agent autonomy and specialization. Likewise, ANP pushes the frontier with decentralized, internet-scale discovery and collaboration, using DID-authenticated agents and semantic web standards (JSON-LD, Schema.org). It establishes the foundation for federated agent networks with open trust and runtime negotiation.

Together, these protocols define a layered infrastructure for communication, identity, and task management. They collectively support the emergence of agent societies capable of distributed reasoning, adaptive coordination, and persistent memory [74, 321, 322], hallmark of the AGI systems. Their evolution marks a shift from isolated, monolithic agents toward scalable, interoperable networks of intelligent entities operating with shared context and collective goals.

8.1.2 Large Concept Models

As AI technology advances towards AGI, the underlying bottlenecks of token-level processing have become increasingly apparent, driving the development of architectures that operate at higher level of semantic abstraction [323]. Large Concept Models (LCMs) are a quantum leap from token-level language prediction models to concept-level reasoning-based language prediction models (Figure 7), providing the machine with a human-like manner of understanding and processing language, which is consistent with hierarchical cognitive process.

LCMs are designed to operate over explicit higher-level semantic representations known as "concepts", which are language- and modality-agnostic abstractions that represent ideas or actions in a structured flow. Unlike LLMs, which process the text at token

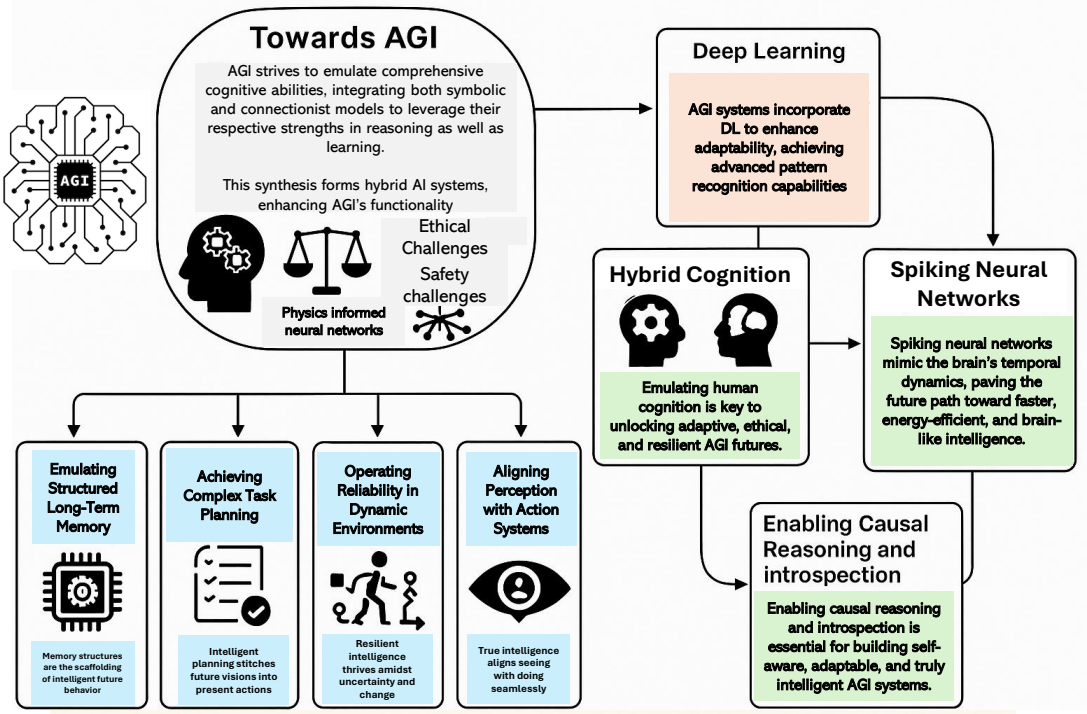

Figure 8: Illustrating AGI's progression toward human-like intelligence by integrating symbolic and connectionist models, emphasizing structured memory, causal reasoning, adaptive planning, and perception-action alignment, while addressing safety, efficiency, and introspective cognitive capabilities for future development. Merge our proposal

level, LCMs predict the next concept rather than the next token, with every concept being a sentence-level semantic representation. This architectural novelty is enabled possible by the SONAR embedding space [324], a multilingual and multimodal fixed-size sentence embedding framework that supports more than 200 languages in text and 76 languages in speech and supports the concept-level reasoning through its intricate encoder-decoder model.

LCMs are a critical building block in the pursuit of AGI, as they enable AI systems to work in terms of concepts rather than individual words, thereby allow for the development of deep contextual understanding and more coherent long-form generation. The development of LCMs represents a fundamental paradigm shift from token-based language modeling towards a semantic-based language modeling, offering a closer approximation of human cognitive processes without the limitation imposed by modality competition [325]

8.1.3 Large Reasoning Models (LRMs)

LRMs represents a shift away from traditional language models, moving toward systems that focus on explicit, multi-step cognitive processes as opposed to single-shot response generation [20]. This method derives from human problem-solving behavior, in which complicated problems are analyzed in sequences of the reasoning process, nested on previous conclusions. Extended inference time computation lies at the core of LRMs and involves the training of models to 'think' through problems in a structured manner, as opposed to relying only on pattern matching from already seen training examples [326]. These systems employ techniques including chain-of-thought reasoning, self-reflection, and iterative refinement to generate more accurate and well-reasoned outputs [53]. This controlled computational approach allows models to perform advanced mathematical, logical, and analytic operations, far exceeding the capabilities of even the largest autoregressive language models.

The LRM paradigm changes the typical trade-off between model size, computational complexity and performance by showing that the computation resources can be spent effectively on the inference side rather than the training side [327]. Unlike typical architectures which learn responses in a single forward pass, LRMs perform prolonged reasoning processes, and sometimes require multiple iterations, self-correction, and fact-checking. This mirrors human cognition, for which hard problems require attention, working memory, and systematic cycling through possible solution paths before a non-intuitive solution occurs.

The reasoning-centered design of LRMs mirrors the structured nature of human reasoning during an-

alytical thought, where complex problems are approached via effortful decomposition, hypothesis generation and evidence scrutiny. This systematic treatment of problems is key to the development of more robust and interpretable AI systems that deal with tasks that start from real understanding of the data, instead of merely patterns that arise from the data.

8.1.4 Mixture of Experts

Mixture of Experts (MoE) is a departure from monolithic neural network architectures, considering models as ensembles of specific sub-networks, selectively triggered by the input [328]. This argument is based on the biological analogy of modular architecture, typical of some parts of the brain specializing in processing different kinds of information [329]. At the center of MoE are multiple "expert" networks, each of which can handle part of the overall task, and a "gating" network that dynamically chooses to which experts to send its inputs [330]. Such conditional computation enables a much higher model capability to be achieved without a linear increase in computational cost. The gating mechanism is learned to distribute the computation across the experts, such a way that only a small fraction of parameters are activated for each given input [331]. This is in contrast to traditional dense neural networks, where all parameters need to participate in processing each sample, resulting in huge computational cost as the model grows [332].

The MoE paradigm, which promotes a specialized yet coordinated intelligence architecture, mirrors human cognition where the brain consists of specialized physical regions which are specialized in different functions, yet capable of seamlessly integrating to solve complex tasks. It is widely believed that this modularity and specialization are essential for the efficiency, adaptability, and plasticity of human intelligence.

8.1.5 Neural Society of Agents

Another approach towards decentralized decision making and prediction is Neural Society of Agents. Within this, rather than a single model that is all encompassing, the neural society of agents approach suggests a multi-agent AI model, in which different agents have distinct expertise and that share intelligence to collaborate on solving complex problems [333]. This resembles the system, found in nature, in which individual cells or organisms work together to achieve overall goal [328]. This method also supports distributed problem decomposition and task assignment, since capabilities are distributed amongst the agents, leading to a parallel implementation and enhanced efficiency. Moreover, the interactions between agents can lead to an enhanced collective intelligence which can be greater than that of any single agent, such as found in social insect's colonies [334]. To achieve the above functionality, the neural society of agents requires work in multiple areas such as multi-agent reinforcement learning, optimizing communication protocols, coordination mechanisms and managing emergent behaviors [335].

The creation of neural societies of agents represents a compelling approach to AGI, as it reflects the distributed and collaborative nature of human intelligence. Human cognition is not a unitary construct, but rather the product of complex interactions among multiple cognitive modules and brain regions. By developing communities of artificial agents that can collaborate, share their findings and learn from each other, we may be able to replicate some of the most powerful attributes of human intelligence and ultimately enabling the creation of more general, adaptive and flexible AGI systems.

8.2 The importance of benchmark datasets

Benchmark datasets have been foundational to progress in AI, enabling fair comparisons and standardizing evaluations, e.g., ImageNet for vision [336], GLUE, HELM, and ALM-Benc for language [337, 338, 339]. However, current benchmarks often assess narrow capabilities and fall short of testing generalization, long-horizon planning, or socio-cognitive reasoning key to AGI. To evaluate AGI systems meaningfully, we need next-generation benchmarks that integrate multi-modal inputs, real-world constraints, ethical reasoning, and interactive environments. Initiatives like ARC [277] and BIG-Bench [276] point in this direction, but broader, dynamic benchmarks are still lacking. Table 4 summarizes the prominent benchmarks used to evaluate the capabilities related to AGI in reasoning, embodiment, and language interaction.

8.3 The Role of Synthetic Data in AGI

Synthetic data has emerged as a pivotal component in scaling and generalizing AI systems, offering controllable diversity, infinite augmentation, and safe simulation for high-risk or rare scenarios [43]. Procedurally generated environments such as BabyAI and MineDojo [278] enable agents to train in highly customizable tasks, while self-play and emergent curricula exemplified by AlphaZero and Voyager allow for autonomous skill acquisition without explicit supervision [340].

Moreover, LLMs now routinely generate synthetic instruction–response datasets, accelerating pretraining and fine-tuning pipelines. However, the misuse of synthetic data can lead to systemic biases, factual drift, and ethical misalignment, especially when artificial distributions diverge from real-world human contexts [341]. As AGI systems grow more au-

tonomous and capable, ensuring the quality, representativeness, and traceability of synthetic data has become essential for developing robust, grounded, and ethically aligned intelligence [342].

9 Missing Pieces and Avenues of Future Work

While there has been enormous progress towards the goal of AGI, there are several aspects that still are missing. A major issue with current systems in terms of AGI is the lack of true creativity and innovation. Currently available models excel at using already seen data to generate outputs, they still lack true creativity capability. AGI systems need to be able to "think out of the box" which requires pushing the boundaries posed by the confines of input data.

9.1 Uncertainty in AGI: Navigating a Dual-Natured Universe

AGI aspires to emulate human-like intellectual versatility, crucially including managing uncertainty inherent in our dual-natured universe, where deterministic rules coexist with random, unpredictable events [333, 343]. Unlike narrow AI, optimized for structured environments, AGI must autonomously adapt and make informed decisions under conditions of incomplete knowledge and inherent randomness.

Two principal uncertainty types confront AGI. *Epistemic uncertainty*, reflecting deterministic limitations, arises from incomplete or noisy data, training gaps, or novel environments beyond prior knowledge [333]. In contrast, *aleatory uncertainty* captures the intrinsic randomness of natural and social phenomena, such as unpredictable human emotions or environmental variability that defy deterministic modeling regardless of data quantity [344, 345].

Effectively navigating these uncertainties requires AGI to dynamically balance exploration of new knowledge and exploitation of established information, thereby enabling optimal decision-making in unpredictable settings [346, 347]. Additionally, decisions under uncertainty carry profound ethical implications, necessitating interpretable and accountable AGI systems to mitigate biases, unfair outcomes, and unintended consequences [348, 349].

The Dual Universe: Random and Deterministic Dynamics in AGI

While the universe is inherently stochastic, AGI systems equipped with continual learning mature by absorbing real-world variance. Over time, uncertainty becomes compressible into structured knowledge facilitating robust, deterministic adaptation and generalization.

9.2 Beyond Memorization: Compression as a Bridge to Reasoning

The success of Large AI systems much still stems from memorization at scale, since these models are trained to predict the next token, these models often fails in unfamiliar situations [350]; particularly those demanding causal reasoning [351] long-horizon planning [26], or physical intuition [22].

Reasoning and Memorization Are Not Opposites Reasoning and memorization are considered distinct or even opposing capabilities [352]. In reality, they exist on a continuum shaped by the degree to which information is compressed [353]. Memorization corresponds to low compression, which means that one simply stores examples like a lookup table. True reasoning reflects high compression, abstracting core principles and applying them flexibly to novel problems![354].

Most LLMs operate between these extremes. They don't merely memorize—they generalize shallowly by interpolating across known patterns. Yet this is not full abstraction. Their reasoning remains fragile, limited by training data and lacking mechanisms for grounding or principled inference [355].

Designing for Compression and Abstraction in AGI The path forward isn't to discard memory, but to structure it more intelligently. Memorization supplies facts; reasoning turns those facts into insights. AGI will require architectures that embrace both—using tools like retrieval-augmented generation (RAG) [356], modular reasoning agents [357], and memory-aware training strategies that encourage deeper compression [187].

Decomposing Intelligence: Reasoning + Memory

While memory and reasoning are often seen as separate, true intelligence arises from their synergy. Memory anchors past experience; reasoning abstracts and applies it to new situations. Their integration enables adaptive, context-aware behavior—central to AGI design.

9.3 Emotional and Social Understanding

Current AI systems lack the capacity to perceive emotions or navigate complex social dynamics. For AGI to achieve human-level intelligence, it must engage with users in emotionally, empathatically and context-aware ways [358]. This requires integrating psychological theories , human behavioral data, and leveraging multimodal learning techniques to effectively detect, interpret, and respond to emotional and social cues effectively.

9.3.1 Ethics and Moral Judgement

True AGI must operate within a comprehensive ethical and moral framework. Event current systems, despite lackin general intelligence, exhibit biases that raise concerns [113]. To prevent harmful outcomes, AGI development must embed ethical principles from the outset, guided by interdisciplinary consensus among legal, ethical, and sociological experts. Furthermore, AGI systems should incorporate human-in-the-loop feedback mechanisms to ensure accountability and promote responsible behavior [359].

9.4 Debt in the Age of AGI: Cognitive and Technical Risks

One emerging concern is **cognitive debt**, a long-term erosion of human intellectual engagement caused by overreliance on LLMs. Recent neurobehavioral studies [267] reveal that participants using LLMs exhibit reduced neural connectivity, lower recall, and diminished essay ownership compared to those relying on their own cognition.

Technical Debt In parallel, AGI development is accelerating the phenomenon of **technical debt** through practices like *vibe coding* [360], where code is generated based on surface-level pattern completion rather than robust logic or modular design.

These dual debts, whether cognitive and technical, are not peripheral concerns. They reflect a broader imbalance in current AGI trajectories: prioritizing short-term performance and usability over foundational understanding and resilience [361]. Mitigating them requires not only architectural guardrails, but also thoughtful co-evolution of education, software engineering norms, and human-AI interaction design.

9.5 Power Consumption and Environmental Impact

The infrastructure supporting computationally intensive models demands immense electricity, with projections indicating substantial increases as development advances toward AGI [362]. This escalating energy consumption not only limits scalability but also exacerbates environmental concerns, including carbon emissions and resource depletion. To mitigate these impacts, AGI development must prioritize energy-efficient model architectures, low-power deployment strategies, and sustainable data center operations [363].

10 Conclusion

AGI remains one of the most profound scientific challenges of our time, demanding not only greater scale, but also deeper alignment with the cognitive, ethical, and societal foundation of human intelligence. This paper has examined AGI from a multidisciplinary lens, synthesizing insights from neuroscience, symbolic reasoning, learning theory, and social systems design. We argue that current paradigms, especially those grounded in next-token prediction are insufficient to yield agents capable of robust reasoning, self-reflection, and generalization across unstructured, uncertain environments.

Several challenges remains, such as the need for grounded world models, dynamic memory, causal reasoning, robust handling of aleatory and epistemic uncertainty, developing perception of emotional and social contexts and collective agent architectures. Significant advancements have been made, such as Large Concept Models, Large Reasoning Models and Mixture of Experts, which improve LLM performance beyond next-token prediction by incorporating biologically inspired behaviors into output generation. The "society of agents" metaphor offers a promising direction, reflecting both biological modularity and the need for specialization and internal negotiation in future AGI systems.

Looking forward, we believe that true progress toward AGI will require a fundamental shift from monolithic models to modular, self-adaptive, and value-aligned systems. This transition must be accompanied by social foresight, involving the proactive redesign of education, labor, and policy frameworks to accommodate and co-evolve with intelligent machines. AGI cannot be purely a technical pursuit. On the contrary, it must be a human project with development progressing alongside humans actively involved in the process. This requires the inclusion of diverse stakeholders in the development process through cultivating a shared, inclusive vision and goal-setting. Such an ecosystem will facilitate the responsible and socially acceptable advancement of AGI.

References

[1] Alan M Turing. *Computing machinery and intelligence*. Springer, 2009.

[2] Jerome Bruner. *A study of thinking*. Routledge, 2017.

[3] Rolf Pfeifer and Christian Scheier. *Understanding intelligence*. MIT press, 2001.

[4] Giulio Tononi and Christof Koch. Consciousness: here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668):20140167, 2015.

[5] Inayat Khan, Abid Jameel, Inam Ullah, Ijaz Khan, and Habib Ullah. The agi-cybersecurity nexus: Exploring implications and applications. In *Artificial General Intelligence (AGI) Security: Smart Appli-*

cations and Sustainable Technologies, pages 271–289. Springer, 2024.

[6] Mehmet Fırat and Saniye Kuleli. What if gpt4 became autonomous: The auto-gpt project and use cases. *Journal of Emerging Computer Technologies*, 3(1):1–6, 2023.

[7] Aixin Liu, Bei Feng, Bing Xue, Bingxuan Wang, Bochao Wu, Chengda Lu, Chenggang Zhao, Chengqi Deng, Chenyu Zhang, Chong Ruan, et al. Deepseek-v3 technical report. *arXiv preprint arXiv:2412.19437*, 2024.

[8] Nidhal Jegham, Marwan Abdelatti, and Abdeltawab Hendawi. Visual reasoning evaluation of grok, deepseek janus, gemini, qwen, mistral, and chatgpt. *arXiv preprint arXiv:2502.16428*, 2025.

[9] Mengnan Qi, Yufan Huang, Yongqiang Yao, Maoquan Wang, Bin Gu, and Neel Sundaresan. Is next token prediction sufficient for gpt? exploration on code logic comprehension. *arXiv preprint arXiv:2404.08885*, 2024.

[10] Manling Li, Shiyu Zhao, Qineng Wang, Kangrui Wang, Yu Zhou, Sanjana Srivastava, Cem Gokmen, Tony Lee, Erran Li Li, Ruohan Zhang, et al. Embodied agent interface: Benchmarking llms for embodied decision making. *Advances in Neural Information Processing Systems*, 37:100428–100534, 2024.

[11] Komal Kumar, Tajamul Ashraf, Omkar Thawakar, Rao Muhammad Anwer, Hisham Cholakkal, Mubarak Shah, Ming-Hsuan Yang, Phillip HS Torr, Fahad Shahbaz Khan, and Salman Khan. Llm post-training: A deep dive into reasoning large language models. *arXiv preprint arXiv:2502.21321*, 2025.

[12] Haotian Liu, Chunyuan Li, Qingyang Wu, and Yong Jae Lee. Visual instruction tuning. *Advances in neural information processing systems*, 36:34892–34916, 2023.

[13] Timo Kaufmann, Paul Weng, Viktor Bengs, and Eyke Hüllermeier. A survey of reinforcement learning from human feedback. *arXiv preprint arXiv:2312.14925*, 10, 2023.

[14] Divya Shanmugam, Fernando Diaz, Samira Shabanian, Michèle Finck, and Asia Biega. Learning to limit data collection via scaling laws: A computational interpretation for the legal principle of data minimization. In *Proceedings of the 2022 ACM Conference on Fairness, Accountability, and Transparency*, pages 839–849, 2022.

[15] Jingbo Shang, Zai Zheng, Jiale Wei, Xiang Ying, Felix Tao, and Mindverse Team. Ai-native memory: A pathway from llms towards agi. *arXiv preprint arXiv:2406.18312*, 2024.

[16] Lili Chen, Kevin Lu, Aravind Rajeswaran, Kimin Lee, Aditya Grover, Misha Laskin, Pieter Abbeel, Aravind Srinivas, and Igor Mordatch. Decision transformer: Reinforcement learning via sequence modeling. *Advances in neural information processing systems*, 34:15084–15097, 2021.

[17] Amber Xie, Oleh Rybkin, Dorsa Sadigh, and Chelsea Finn. Latent diffusion planning for imitation learning. *arXiv preprint arXiv:2504.16925*, 2025.

[18] Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik Narasimhan, and Yuan Cao. React: Synergizing reasoning and acting in language models. In *International Conference on Learning Representations (ICLR)*, 2023.

[19] Haoyu Lu, Wen Liu, Bo Zhang, Bingxuan Wang, Kai Dong, Bo Liu, Jingxiang Sun, Tongzheng Ren, Zhuoshu Li, Hao Yang, et al. Deepseek-vl: towards real-world vision-language understanding. *arXiv preprint arXiv:2403.05525*, 2024.

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

[21] Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

[22] Muhammad Usman Hadi, Rizwan Qureshi, Abbas Shah, Muhammad Irfan, Anas Zafar, Muhammad Bilal Shaikh, Naveed Akhtar, Jia Wu, Seyedali Mirjalili, et al. Large language models: a comprehensive survey of its applications, challenges, limitations, and future prospects. *Authorea Preprints*, 1:1–26, 2023.

[23] M Jae Moon. Searching for inclusive artificial intelligence for social good: Participatory governance and policy recommendations for making ai more inclusive and benign for society. *Public Administration Review*, 83(6):1496–1505, 2023.

[24] Rodney Brooks. I, rodney brooks, am a robot. *IEEE spectrum*, 45(6):68–71, 2008.

[25] Raghu Raman, Robin Kowalski, Krishnashree Achuthan, Akshay Iyer, and Prema Nedungadi. Navigating artificial general intelligence development: societal, technological, ethical, and brain-inspired pathways. *Scientific Reports*, 15(1):1–22, 2025.

[26] Tao Feng, Chuanyang Jin, Jingyu Liu, Kunlun Zhu, Haoqin Tu, Zirui Cheng, Guanyu Lin, and Jiaxuan You. How far are we from agi: Are llms all we need? *Transactions on Machine Learning Research*.

[27] Sajib Alam. A methodological framework to integrate agi into personalized healthcare. *Quarterly Journal of Computational Technologies for Healthcare*, 7(3):10–21, 2022.

[28] Josh Achiam, Steven Adler, Sandhini Agarwal, Lama Ahmad, Ilge Akkaya, Florencia Leoni Aleman, Diogo Almeida, Janko Altenschmidt, Sam Altman, Shyamal Anadkat, et al. Gpt-4 technical report. *arXiv preprint arXiv:2303.08774*, 2023.

[29] Amazon Artificial General Intelligence. Amazon nova sonic: Technical report and model card. 2025.

[30] Marah Abdin, Jyoti Aneja, Harkirat Behl, Sébastien Bubeck, Ronen Eldan, Suriya Gunasekar, Michael Harrison, Russell J Hewett, Mojan Javaheripi, Piero Kauffmann, et al. Phi-4 technical report. *arXiv preprint arXiv:2412.08905*, 2024.

[31] Gemini Team, Rohan Anil, Sebastian Borgeaud, Jean-Baptiste Alayrac, Jiahui Yu, Radu Soricut, Johan Schalkwyk, Andrew M Dai, Anja Hauth, Katie Millican, et al. Gemini: a family of highly capable multimodal models. *arXiv preprint arXiv:2312.11805*, 2023.

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

[33] Tom Everitt, Gary Lea, and Marcus Hutter. Agi safety literature review. *arXiv preprint arXiv:1805.01109*, 2018.

[34] Fei Dou, Jin Ye, Geng Yuan, Qin Lu, Wei Niu, Haijian Sun, Le Guan, Guoyu Lu, Gengchen Mai, Ninghao Liu, et al. Towards artificial general intelligence (agi) in the internet of things (iot): Opportunities and challenges. *arXiv preprint arXiv:2309.07438*, 2023.

[35] Lin Zhao, Lu Zhang, Zihao Wu, Yuzhong Chen, Haixing Dai, Xiaowei Yu, Zhengliang Liu, Tuo Zhang, Xintao Hu, Xi Jiang, et al. When brain-inspired ai meets agi. *Meta-Radiology*, page 100005, 2023.

[36] Florin Leon. A review of findings from neuroscience and cognitive psychology as possible inspiration for the path to artificial general intelligence. *arXiv preprint arXiv:2401.10904*, 2024.

[37] Wlodzislaw Duch, Rudy Setiono, and Jacek M Zurada. Computational intelligence methods for rule-based data understanding. *Proceedings of the IEEE*, 92(5):771–805, 2004.

[38] Giuseppe Marra, Sebastijan Dumančić, Robin Manhaeve, and Luc De Raedt. From statistical relational to neurosymbolic artificial intelligence: A survey. *Artificial Intelligence*, page 104062, 2024.

[39] Martin Campbell-Kelly, William F Aspray, Jeffrey R Yost, Honghong Tinn, and Gerardo Con Díaz. *Computer: A history of the information machine*. Routledge, 2023.

[40] Peter West, Ximing Lu, Nouha Dziri, Faeze Brahman, Pingqing Fu, Jena D. Hwang, Liwei Jiang, Jillian Fisher, Abhilasha Ravichander, Khyathi Raghavi Chandu, Benjamin T. Newman, Pang Wei Koh, Allyson Ettinger, and Yejin Choi. The generative ai paradox: "what it can create, it may not understand", 01 2023.

[41] Wenhao Yu, Chenguang Zhu, Zaitang Li, Zhiting Hu, Qingyun Wang, Heng Ji, and Meng Jiang. A survey of knowledge-enhanced text generation. *ACM Computing Surveys*, 54(11s):1–38, 2022.

[42] Dong Zhang, Shimin Li, Xin Zhang, Jun Zhan, Pengyu Wang, Yaqian Zhou, and Xipeng Qiu. Speechgpt: Empowering large language models with intrinsic cross-modal conversational abilities. *arXiv preprint arXiv:2305.11000*, 2023.

[43] James Jordon, Lukasz Szpruch, Florimond Houssiau, Mirko Bottarelli, Giovanni Cherubin, Carsten Maple, Samuel N Cohen, and Adrian Weller. Synthetic data–what, why and how? *arXiv preprint arXiv:2205.03257*, 2022.

[44] Yongqi Tong, Dawei Li, Sizhe Wang, Yujia Wang, Fei Teng, and Jingbo Shang. Can llms learn from previous mistakes? investigating llms' errors to boost for reasoning. *arXiv preprint arXiv:2403.20046*, 2024.

[45] Sébastien Bubeck, Varun Chandrasekaran, Ronen Eldan, Johannes Gehrke, Eric Horvitz, Ece Kamar, Peter Lee, Yin Tat Lee, Yuanzhi Li, Scott Lundberg, et al. Sparks of artificial general intelligence: Early experiments with gpt-4. *arXiv preprint arXiv:2303.12712*, 2023.

[46] Kevin Wang, Junbo Li, Neel P Bhatt, Yihan Xi, Qiang Liu, Ufuk Topcu, and Zhangyang Wang. On the planning abilities of openai's o1 models: Feasibility, optimality, and generalizability. *arXiv preprint arXiv:2409.19924*, 2024.

[47] Dominik K Kanbach, Louisa Heiduk, Georg Blueher, Maximilian Schreiter, and Alexander Lahmann. The genai is out of the bottle: generative artificial intelligence from a business model innovation perspective. *Review of Managerial Science*, 18(4):1189–1220, 2024.

[48] Ashish Kumar Shakya, Gopinatha Pillai, and Sohom Chakrabarty. Reinforcement learning algorithms: A brief survey. *Expert Systems with Applications*, 231:120495, 2023.

[49] Maxim Lapan. *Deep Reinforcement Learning Hands-On: Apply modern RL methods, with deep Q-networks, value iteration, policy gradients, TRPO, AlphaGo Zero and more*. Packt Publishing Ltd, 2018.

[50] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *nature*, 596(7873):583–589, 2021.

[51] John Schulman, Filip Wolski, Prafulla Dhariwal, Alec Radford, and Oleg Klimov. Proximal policy optimization algorithms. *arXiv preprint arXiv:1707.06347*, 2017.

[52] Rafael Rafailov, Archit Sharma, Eric Mitchell, Christopher D Manning, Stefano Ermon, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model. *Advances in Neural Information Processing Systems*, 36:53728–53741, 2023.

[53] Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in Neural Information Processing Systems*, 36:8634–8652, 2023.

[54] Marta Garnelo and Murray Shanahan. Reconciling deep learning with symbolic artificial intelligence: representing objects and relations. *Current Opinion in Behavioral Sciences*, 29:17–23, 2019.

[55] Omar Ibrahim Obaid. From machine learning to artificial general intelligence: A roadmap and implications. *Mesopotamian Journal of Big Data*, 2023:81–91, 2023.

[56] John Page, Michael Bain, and Faqihza Mukhlish. The risks of low level narrow artificial intelligence. In *2018 IEEE international conference on intelligence and safety for robotics (ISR)*, pages 1–6. IEEE, 2018.

[57] Adriana Braga and Robert K Logan. The emperor of strong ai has no clothes: limits to artificial intelligence. *Information*, 8(4):156, 2017.

[58] Nikolaus Kriegeskorte and Pamela K Douglas. Cognitive computational neuroscience. *Nature neuroscience*, 21(9):1148–1160, 2018.

[59] George Siemens, Fernando Marmolejo-Ramos, Florence Gabriel, Kelsey Medeiros, Rebecca Marrone, Srecko Joksimovic, and Maarten de Laat. Human and artificial cognition. *Computers and Education: Artificial Intelligence*, 3:100107, 2022.

[60] Sedat Sonko, Adebunmi Okechukwu Adewusi, Ogugua Chimezie Obi, Shedrack Onwusinkwue, and Akoh Atadoga. A critical review towards artificial general intelligence: Challenges, ethical considerations, and the path forward. *World Journal of Advanced Research and Reviews*, 21(3):1262–1268, 2024.

[61] Amit Sheth and Kaushik Roy. Neurosymbolic value-inspired artificial intelligence (why, what, and how). *IEEE Intelligent Systems*, 39(1):5–11, 2024.

[62] Ron Sun and Frederic Alexandre. *Connectionist-symbolic integration: From unified to hybrid approaches*. Psychology Press, 2013.

[63] Fei Tang, Wanling Gao, LuZhou Peng, and Jianfeng Zhan. Agibench: a multi-granularity, multimodal, human-referenced, auto-scoring benchmark for large language models. In *International Symposium on Benchmarking, Measuring and Optimization*, pages 137–152. Springer, 2023.

[64] Ben Goertzel. Artificial general intelligence: concept, state of the art, and future prospects. *Journal of Artificial General Intelligence*, 5(1):1, 2014.

[65] Ross Gruetzemacher and David Paradice. Toward mapping the paths to agi. In *Artificial General Intelligence: 12th International Conference, AGI 2019, Shenzhen, China, August 6–9, 2019, Proceedings 12*, pages 70–79. Springer, 2019.

[66] John E Laird and Robert E Wray III. Cognitive architecture requirements for achieving agi. In *3d Conference on Artificial General Intelligence (AGI-2010)*, pages 3–8. Atlantis Press, 2010.

[67] Omkar Thawakar, Ashmal Vayani, Salman Khan, Hisham Cholakal, Rao M Anwer, Michael Felsberg, Tim Baldwin, Eric P Xing, and Fahad Shahbaz Khan. Mobillama: Towards accurate and lightweight fully transparent gpt. *arXiv preprint arXiv:2402.16840*, 2024.

[68] Tobias Mahler. Regulating artificial general intelligence (agi). In *Law and artificial intelligence: Regulating AI and applying AI in legal practice*, pages 521–540. Springer, 2022.

[69] James Moor. *The Turing test: the elusive standard of artificial intelligence*, volume 30. Springer Science & Business Media, 2003.

[70] Shlomo Danziger. Intelligence as a social concept: a socio-technological interpretation of the turing test. *Philosophy & Technology*, 35(3):68, 2022.

[71] Ben Goertzel. Generative ai vs. agi: The cognitive strengths and weaknesses of modern llms. *arXiv preprint arXiv:2309.10371*, 2023.

[72] Niels Van Berkel, Mikael B Skov, and Jesper Kjeldskov. Human-ai interaction: intermittent, continuous, and proactive. *Interactions*, 28(6):67–71, 2021.

[73] Ranjan Sapkota, Konstantinos I Roumeliotis, and Manoj Karkee. Vibe coding vs. agentic coding: Fundamentals and practical implications of agentic ai. *arXiv preprint arXiv:2505.19443*, 2025.

[74] Ranjan Sapkota, Konstantinos I Roumeliotis, and Manoj Karkee. Ai agents vs. agentic ai: A conceptual taxonomy, applications and challenges. *arXiv preprint arXiv:2505.10468*, 2025.

[75] Mingchen Zhuge, Haozhe Liu, Francesco Faccio, Dylan R Ashley, Róbert Csordás, Anand Gopalakrishnan, Abdullah Hamdi, Hasan Abed Al Kader Hammoud, Vincent Herrmann, Kazuki Irie, et al. Mindstorms in natural language-based societies of mind. *arXiv preprint arXiv:2305.17066*, 2023.

[76] Marvin Minsky. *Society of mind*. Simon and Schuster, 1986.

[77] Deepak Bhaskar Acharya, Karthigeyan Kuppan, and B Divya. Agentic ai: Autonomous intelligence for complex goals–a comprehensive survey. *IEEE Access*, 2025.

[78] Andrew Lampinen, Stephanie Chan, Andrea Banino, and Felix Hill. Towards mental time travel: a hierarchical memory for reinforcement learning agents. *Advances in Neural Information Processing Systems*, 34:28182–28195, 2021.

[79] Juergen Schmidhuber. Annotated history of modern ai and deep learning. *arXiv preprint arXiv:2212.11279*, 2022.

[80] Michael Haenlein and Andreas Kaplan. A brief history of artificial intelligence: On the past, present, and future of artificial intelligence. *California management review*, 61(4):5–14, 2019.

[81] Shiqiang Zhu, Ting Yu, Tao Xu, Hongyang Chen, Schahram Dustdar, Sylvain Gigan, Deniz Gunduz, Ekram Hossain, Yaochu Jin, Feng Lin, et al. Intelligent computing: the latest advances, challenges, and future. *Intelligent Computing*, 2:0006, 2023.

[82] Ron Sun. *Cognitive architectures: Research issues and challenges*, volume 10. Elsevier, 2006.

[83] Suzana Herculano-Houzel. The human brain in numbers: a linearly scaled-up primate brain. *Frontiers in human neuroscience*, 3:857, 2009.

[84] Kelly Rae Chi. Neural modelling: Abstractions of the mind. *Nature*, 531(7592):S16–S17, 2016.

[85] Richard B Buxton. The thermodynamics of thinking: connections between neural activity, energy metabolism and blood flow. *Philosophical Transactions of the Royal Society B*, 376(1815):20190624, 2021.

[86] L Felipe Barros, Juan P Bolanos, Gilles Bonvento, Anne-Karine Bouzier-Sore, Angus Brown, Johannes Hirrlinger, Sergey Kasparov, Frank Kirchhoff, Anne N Murphy, Luc Pellerin, et al. Current technical approaches to brain energy metabolism. *Glia*, 66(6):1138–1159, 2018.

[87] Rowena Chin, Steve WC Chang, and Avram J Holmes. Beyond cortex: The evolution of the human brain. *Psychological Review*, 130(2):285, 2023.

[88] Nancy Kanwisher. Functional specificity in the human brain: a window into the functional architecture of the mind. *Proceedings of the national academy of sciences*, 107(25):11163–11170, 2010.

[89] Frederico AC Azevedo, Ludmila RB Carvalho, Lea T Grinberg, José Marcelo Farfel, Renata EL Ferretti, Renata EP Leite, Wilson Jacob Filho, Roberto Lent, and Suzana Herculano-Houzel. Equal numbers of neuronal and nonneuronal cells make the human brain an isometrically scaled-up primate brain. *Journal of Comparative Neurology*, 513(5):532–541, 2009.

[90] Roberto Lent. Yes, the human brain has around 86 billion neurons. *Brain*, page awaf048, 2025.

[91] Randy L Buckner and Fenna M Krienen. The evolution of distributed association networks in the human brain. *Trends in cognitive sciences*, 17(12):648–665, 2013.

[92] Saket Navlakha, Ziv Bar-Joseph, and Alison L Barth. Network design and the brain. *Trends in cognitive sciences*, 22(1):64–78, 2018.

[93] Xuhong Liao, Athanasios V Vasilakos, and Yong He. Small-world human brain networks: perspectives and challenges. *Neuroscience & Biobehavioral Reviews*, 77:286–300, 2017.

[94] Bang Liu, Xinfeng Li, Jiayi Zhang, Jinlin Wang, Tanjin He, Sirui Hong, Hongzhang Liu, Shaokun Zhang, Kaitao Song, Kunlun Zhu, et al. Advances and challenges in foundation agents: From brain-inspired intelligence to evolutionary, collaborative, and safe systems. *arXiv preprint arXiv:2504.01990*, 2025.

[95] Chris Frith and Ray Dolan. The role of the prefrontal cortex in higher cognitive functions. *Cognitive brain research*, 5(1-2):175–181, 1996.

[96] Serge Dolgikh. Self-awareness in natural and artificial intelligent systems: a unified information-based approach. *Evolutionary Intelligence*, 17(5):4095–4114, 2024.

[97] Jaitip Na-songkhla, Vorapon Mahakaew, and Roumiana Peytcheva-Forsyth. The emergence of generative artificial intelligence: Enhancing critical thinking skills in chatgpt-integrated cognitive flexibility approach. *Generative Artificial Intelligence in Higher Education: A Handbook for Educational Leaders*, page 52, 2024.

[98] Kyle Mahowald, Anna A Ivanova, Idan A Blank, Nancy Kanwisher, Joshua B Tenenbaum, and Evelina Fedorenko. Dissociating language and thought in large language models. *Trends in cognitive sciences*, 2024.

[99] Dagmar Timmann, Johannes Drepper, Marcus Frings, Michael Maschke, Stephanie Richter, MEEA Gerwig, and Florian P Kolb. The human cerebellum contributes to motor, emotional and cognitive associative learning. a review. *Cortex*, 46(7):845–857, 2010.

[100] RC Miall. The cerebellum, predictive control and motor coordination. In *Novartis Foundation Symposium 218-Sensory Guidance of Movement: Sensory Guidance of Movement: Novartis Foundation Symposium 218*, pages 272–290. Wiley Online Library, 2007.

[101] Azhagu Madhavan Sivalingam and Arjun Pandian. Cerebellar roles in motor and social functions and implications for asd. *The Cerebellum*, 23(6):2564–2574, 2024.

[102] Moshe Glickman and Tali Sharot. How human–ai feedback loops alter human perceptual, emotional and social judgements. *Nature Human Behaviour*, pages 1–15, 2024.

[103] Shunsen Huang, Xiaoxiong Lai, Li Ke, Yajun Li, Huanlei Wang, Xinmei Zhao, Xinran Dai, and Yun Wang. Ai technology panic—is ai dependence bad for mental health? a cross-lagged panel model and the mediating roles of motivations for ai use among adolescents. *Psychology Research and Behavior Management*, pages 1087–1102, 2024.

[104] Jing Ren and Feng Xia. Brain-inspired artificial intelligence: A comprehensive review. *arXiv preprint arXiv:2408.14811*, 2024.

[105] Evgenia Gkintoni, Hera Antonopoulou, Andrew Sortwell, and Constantinos Halkiopoulos. Challenging cognitive load theory: The role of educational neuroscience and artificial intelligence in redefining learning efficacy. *Brain Sciences*, 15(2):203, 2025.

[106] Jay L Garfield, Candida C Peterson, and Tricia Perry. Social cognition, language acquisition and the development of the theory of mind. *Mind & Language*, 16(5):494–541, 2001.

[107] Stanley I Greenspan and Stuart Shanker. *The first idea: How symbols, language, and intelligence evolved from our primate ancestors to modern humans*. Da Capo, 2009.

[108] Qinghua Zheng, Huan Liu, Xiaoqing Zhang, Caixia Yan, Xiangyong Cao, Tieliang Gong, Yong-Jin Liu, Bin Shi, Zhen Peng, Xiaocen Fan, et al. Machine memory intelligence: Inspired by human memory mechanisms. *Engineering*, 2025.

[109] Gabriel Molas and Etienne Nowak. Advances in emerging memory technologies: From data storage to artificial intelligence. *Applied Sciences*, 11(23):11254, 2021.

[110] Zihong He, Weizhe Lin, Hao Zheng, Fan Zhang, Matt W Jones, Laurence Aitchison, Xuhai Xu, Miao Liu, Per Ola Kristensson, and Junxiao Shen. Human-inspired perspectives: A survey on ai long-term memory. *arXiv preprint arXiv:2411.00489*, 2024.

[111] Sadia Tariq, Asif Iftikhar, Puruesh Chaudhary, and Khurram Khurshid. Is the 'technological singularity scenario'possible: Can ai parallel and surpass all human mental capabilities? *World Futures*, 79(2):200–266, 2023.

[112] Stephen Grossberg. A path toward explainable ai and autonomous adaptive intelligence: deep learning, adaptive resonance, and models of perception, emotion, and action. *Frontiers in neurorobotics*, 14:36, 2020.

[113] Oliver Li. Should we develop agi? artificial suffering and the moral development of humans. *AI and Ethics*, 5(1):641–651, 2025.

[114] David Vernon, Giorgio Metta, and Giulio Sandini. A survey of artificial cognitive systems: Implications for the autonomous development of mental capabilities in computational agents. *IEEE transactions on evolutionary computation*, 11(2):151–180, 2007.

[115] Xiao Wang, Jun Huang, Yonglin Tian, Chen Sun, Lie Yang, Shanhe Lou, Chen Lv, Changyin Sun, and Fei-Yue Wang. Parallel driving with big models and foundation intelligence in cyber–physical–social spaces. *Research*, 7:0349, 2024.

[116] Richard A Andersen and He Cui. Intention, action planning, and decision making in parietal-frontal circuits. *Neuron*, 63(5):568–583, 2009.

[117] Luca Crosato, Kai Tian, Hubert PH Shum, Edmond SL Ho, Yafei Wang, and Chongfeng Wei. Social interaction-aware dynamical models and decision-making for autonomous vehicles. *Advanced Intelligent Systems*, 6(3):2300575, 2024.

[118] Michael Luck and Ruth Aylett. Applying artificial intelligence to virtual reality: Intelligent virtual environments. *Applied artificial intelligence*, 14(1):3–32, 2000.

[119] Sara Colombo, Lucia Rampino, and Filippo Zambrelli. The adaptive affective loop: how ai agents can generate empathetic systemic experiences. In *Advances in Information and Communication: Proceedings of the 2021 Future of Information and Communication Conference (FICC), Volume 1*, pages 547–559. Springer, 2021.

[120] Yuheng Cheng, Ceyao Zhang, Zhengwen Zhang, Xiangrui Meng, Sirui Hong, Wenhao Li, Zihao Wang, Zekai Wang, Feng Yin, Junhua Zhao, et al. Exploring large language model based intelligent agents: Definitions, methods, and prospects. *arXiv preprint arXiv:2401.03428*, 2024.

[121] Garima Singhal and Aniket Singh. The large action model: Pioneering the next generation of web and app engagement. In *2024 11th International Conference on Reliability, Infocom Technologies and Optimization (Trends and Future Directions)(ICRITO)*, pages 1–6. IEEE, 2024.

[122] Cecilio Angulo, Alejandro Chacón, and Pere Ponsa. Towards a cognitive assistant supporting human operators in the artificial intelligence of things. *Internet of Things*, 21:100673, 2023.

[123] Zhiting Hu and Tianmin Shu. Language models, agent models, and world models: The law for machine reasoning and planning. *arXiv preprint arXiv:2312.05230*, 2023.

[124] Philip N Johnson-Laird. Mental models and human reasoning. *Proceedings of the National Academy of Sciences*, 107(43):18243–18250, 2010.

[125] Jyrki Suomala and Janne Kauttonen. Human's intuitive mental models as a source of realistic artificial intelligence and engineering. *Frontiers in psychology*, 13:873289, 2022.

[126] Anas Zafar, Danyal Aftab, Rizwan Qureshi, Xinqi Fan, Pingjun Chen, Jia Wu, Hazrat Ali, Shah Nawaz, Sheheryar Khan, and Mubarak Shah. Single stage adaptive multi-attention network for image restoration. *IEEE Transactions on Image Processing*, 2024.

[127] Anthony Randal McIntosh. Mapping cognition to the brain through neural interactions. *memory*, 7(5-6):523–548, 1999.

[128] Nick Lee and Laura Chamberlain. Neuroimaging and psychophysiological measurement in organizational research: an agenda for research in organizational cognitive neuroscience. *Annals of the New York Academy of Sciences*, 1118(1):18–42, 2007.

[129] Wynn Legon, Steven Punzell, Ehsan Dowlati, Sarah E. Adams, Alexandra B. Stiles, and Rosalyn J. Moran. Altered prefrontal excitation/inhibition balance and prefrontal output: Markers of aging in human memory networks. *Cerebral Cortex*, 26(11):4315–4326, 2016.

[130] Chang-Hao Kao, Ankit N. Khambhati, Danielle S. Bassett, Matthew R. Nassar, Joseph T. McGuire, Joshua I. Gold, and Joseph W. Kable. Functional brain network reconfiguration during learning in a dynamic environment. *Nature Communications*, 11(1):1682, 2020.

[131] Alfredo Ardila, Byron Bernal, and Monica Rosselli. How localized are language brain areas? a review of brodmann areas involvement in oral language. *Archives of Clinical Neuropsychology*, 31(1):112–122, 2016.

[132] Jeffrey M. Spielberg, Gregory A. Miller, Wendy Heller, and Marie T. Banich. Flexible brain network reconfiguration supporting inhibitory control. *Proceedings of the National Academy of Sciences*, 112(32):10020–10025, 2015.

[133] Valorie N. Salimpoor, Iris van den Bosch, Natasa Kovacevic, Anthony Randal McIntosh, Alain Dagher, and Robert J. Zatorre. Interactions between the nucleus accumbens and auditory cortices predict music reward value. *Science*, 340(6129):216–219, 2013.

[134] Frederic R. Hopp, Ori Amir, Jacob T. Fisher, Scott Grafton, Walter Sinnott-Armstrong, and René Weber. Moral foundations elicit shared and dissociable cortical activation modulated by political ideology. *Nature Human Behaviour*, 7(12):2182–2198, 2023.

[135] Alex Fornito, Andrew Zalesky, and Michael Breakspear. The connectomics of brain disorders. *Nature Reviews Neuroscience*, 16(3):159–172, 2015.

[136] Danielle S Bassett and Michael S Gazzaniga. Understanding complexity in the human brain. *Trends in cognitive sciences*, 15(5):200–209, 2011.

[137] Katrin Amunts, Javier DeFelipe, Cyriel Pennartz, Alain Destexhe, Michele Migliore, Philippe Ryvlin, Steve Furber, Alois Knoll, Lise Bitsch, Jan G Bjaalie, et al. Linking brain structure, activity, and cognitive function through computation. *Eneuro*, 9(2), 2022.

[138] Stephen Grossberg. Toward autonomous adaptive intelligence: Building upon neural models of how brains make minds. *IEEE Transactions on Systems, Man, and Cybernetics: Systems*, 51(1):51–75, 2020.

[139] Bin Hu, Zhi-Hong Guan, Guanrong Chen, and CL Philip Chen. Neuroscience and network dynamics toward brain-inspired intelligence. *IEEE Transactions on Cybernetics*, 52(10):10214–10227, 2021.

[140] Nikolaus Kriegeskorte and Pamela K. Douglas. Cognitive computational neuroscience. *Nature Neuroscience*, 21(9):1148–1160, 2018.

[141] Hae-Jeong Park and Karl Friston. Structural and functional brain networks: From connections to cognition. *Science*, 342(6158):1238411, 2013. doi: 10.1126/science.1238411.

[142] Alex Fornito, Andrew Zalesky, and Edward Bullmore. *Fundamentals of Brain Network Analysis*. Academic Press, 2016.

[143] Olaf Sporns. Structure and function of complex brain networks. *Dialogues in Clinical Neuroscience*, 15(3):247–262, 2013.

[144] David Papo, Javier M. Buldú, and Stefano Boccaletti. *Network Theory in Neuroscience*, pages 2190–2206. Springer New York, New York, NY, 2022.

[145] Martijn P. van den Heuvel and Olaf Sporns. Rich-club organization of the human connectome. *The Journal of Neuroscience*, 31(44):15775–15786, 2011.

[146] Olaf Sporns and Richard F. Betzel. Modular brain networks. *Annual Review of Psychology*, 67:613–640, 2016.

[147] John D. Medaglia, Mary-Ellen Lynall, and Danielle S. Bassett. Cognitive network neuroscience. *Journal of Cognitive Neuroscience*, 27(8):1471–1491, 2015.

[148] Luiz Pessoa. Understanding brain networks and brain organization. *Physics of Life Reviews*, 11(3):400–435, 2014.

[149] Rex E. Jung and Richard J. Haier. The parieto-frontal integration theory (p-fit) of intelligence: Converging neuroimaging evidence. *Behavioral and Brain Sciences*, 30(2):135–154, 2007.

[150] Kirsten Hilger, Matthias Ekman, Christian J. Fiebach, and Ulrike Basten. Intelligence is associated with the modular structure of intrinsic brain networks. *Scientific Reports*, 7(1):16088, 2017.

[151] Farzad V. Farahani, Waldemar Karwowski, and Nichole R. Lighthall. Application of graph theory for identifying connectivity patterns in human brain networks: A systematic review. *Frontiers in Neuroscience*, 13, 2019.

[152] Michael W. Cole, Jeremy R. Reynolds, Jonathan D. Power, Grega Repovs, Alan Anticevic, and Todd S. Braver. Multi-task connectivity reveals flexible hubs for adaptive task control. *Nature Neuroscience*, 16(9):1348–1355, 2013.

[153] Gustavo Deco, Diego Vidaurre, and Morten L. Kringelbach. Revisiting the global workspace orchestrating the hierarchical organization of the human brain. *Nature Human Behaviour*, 5(4):497–511, 2021.

[154] Shaina Raza, Rizwan Qureshi, Anam Zahid, Joseph Fioresi, Ferhat Sadak, Muhammad Saeed, Ranjan Sapkota, Aditya Jain, Anas Zafar, Muneeb Ul Hassan, et al. Who is responsible? the data, models, users or regulations? responsible generative ai for a sustainable future. *arXiv preprint arXiv:2502.08650*, 2025.

[155] Maziar Raissi, Paris Perdikaris, and George E Karniadakis. Physics-informed neural networks: A deep learning framework for solving forward and inverse problems involving nonlinear partial differential equations. *Journal of Computational physics*, 378:686–707, 2019.

[156] Ziming Liu, Yixuan Wang, Sachin Vaidya, Fabian Ruehle, James Halverson, Marin Soljačić, Thomas Y Hou, and Max Tegmark. Kan: Kolmogorov-arnold networks. *arXiv preprint arXiv:2404.19756*, 2024.

[157] Swagatam Das, Ajith Abraham, and BK Panigrahi. Computational intelligence: Foundations, perspectives, and recent trends. *Computational Intelligence and Pattern Analysis in Biological Informatics*, pages 1–37, 2010.

[158] Mohamed Alloghani, Dhiya Al-Jumeily, Jamila Mustafina, Abir Hussain, and Ahmed J Aljaaf. A systematic review on supervised and unsupervised machine learning algorithms for data science. *Supervised and unsupervised learning for data science*, pages 3–21, 2020.

[159] Arun Babu, Changhan Wang, Andros Tjandra, Kushal Lakhotia, Qiantong Xu, Naman Goyal, Kritika Singh, Patrick Von Platen, Yatharth Saraf, Juan Pino, et al. Xls-r: Self-supervised cross-lingual speech representation learning at scale. *arXiv preprint arXiv:2111.09296*, 2021.

[160] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big data*, 3:1–40, 2016.

[161] Alireza Fallah, Aryan Mokhtari, and Asuman Ozdaglar. On the convergence theory of gradient-based model-agnostic meta-learning algorithms. In *International Conference on Artificial Intelligence and Statistics*, pages 1082–1092. PMLR, 2020.

[162] Bram Bakker, Jürgen Schmidhuber, et al. Hierarchical reinforcement learning based on subgoal discovery and subpolicy specialization. In *Proc. of the 8-th Conf. on Intelligent Autonomous Systems*, pages 438–445. Citeseer, 2004.

[163] Jingyao Wang, Wenwen Qiang, Zeen Song, Changwen Zheng, and Hui Xiong. Learning to think: Information-theoretic reinforcement fine-tuning for llms. *arXiv preprint arXiv:2505.10425*, 2025.

[164] Archit Parnami and Minwoo Lee. Learning from few examples: A summary of approaches to few-shot learning. *arXiv preprint arXiv:2203.04291*, 2022.

[165] Shaina Raza, Aravind Narayanan, Vahid Reza Khazaie, Ashmal Vayani, Mukund S Chettiar, Amandeep Singh, Mubarak Shah, and Deval Pandya. Humanibench: A human-centric framework for large multimodal models evaluation. *arXiv preprint arXiv:2505.11454*, 2025.

[166] Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sastry, Amanda Askell, Pamela Mishkin, Jack Clark, et al. Learning transferable visual models from natural language supervision. In *International conference on machine learning*, pages 8748–8763. PmLR, 2021.

[167] Robert Geirhos, Jörn-Henrik Jacobsen, Claudio Michaelis, Richard Zemel, Wieland Brendel, Matthias Bethge, and Felix A Wichmann. Shortcut learning in deep neural networks. *Nature Machine Intelligence*, 2(11):665–673, 2020.

[168] Irina Higgins, Sébastien Racanière, and Danilo Rezende. Symmetry-based representations for artificial and biological general intelligence. *Frontiers in Computational Neuroscience*, 16:836498, 2022.

[169] Chen Shani, Dan Jurafsky, Yann LeCun, and Ravid Shwartz-Ziv. From tokens to thoughts: How llms and humans trade compression for meaning. *arXiv preprint arXiv:2505.17117*, 2025.

[170] Geoffrey Hinton. Distilling the knowledge in a neural network. *arXiv preprint arXiv:1503.02531*, 2015.

[171] Yuzhen Huang, Jinghan Zhang, Zifei Shan, and Junxian He. Compression represents intelligence linearly. *arXiv preprint arXiv:2404.09937*, 2024.

[172] Xiangwen Wang, Xianghong Lin, and Xiaochao Dang. Supervised learning in spiking neural networks: A review of algorithms and evaluations. *Neural Networks*, 125:258–280, 2020.

[173] Paul Smolensky. Connectionist ai, symbolic ai, and the brain. *Artificial Intelligence Review*, 1(2):95–109, 1987.

[174] Fatemeh Chahkoutahi and Mehdi Khashei. A seasonal direct optimal hybrid model of computational intelligence and soft computing techniques for electricity load forecasting. *Energy*, 140:988–1004, 2017.

[175] Elena N Benderskaya and Sofya V Zhukova. Multidisciplinary trends in modern artificial intelligence: Turing's way. *Artificial Intelligence, Evolutionary Computing and Metaheuristics: In the Footsteps of Alan Turing*, pages 319–343, 2013.

[176] Edward Allen Silver. An overview of heuristic solution methods. *Journal of the operational research society*, 55(9):936–956, 2004.

[177] Hui Yang, Sifu Yue, and Yunzhong He. Auto-gpt for online decision making: Benchmarks and additional opinions. *arXiv preprint arXiv:2306.02224*, 2023.

[178] Guanzhi Wang, Yuqi Xie, Yunfan Jiang, Ajay Mandlekar, Chaowei Xiao, Yuke Zhu, Linxi Fan, and Anima Anandkumar. Voyager: An open-ended embodied agent with large language models. *arXiv preprint arXiv:2305.16291*, 2023.

[179] Xiaohui Zou. A review of the latest research achievements in the basic theory of generative ai and artificial general intelligence (agi). *Computer Science and Technology*, 3(3):82, 2024.

[180] NL Rane and M Paramesha. Explainable artificial intelligence (xai) as a foundation for trustworthy artificial intelligence. *Trustworthy Artificial Intelligence in Industry and Society*, pages 1–27, 2024.

[181] Emanuele Neri, Gayane Aghakhanyan, Marta Zerunian, Nicoletta Gandolfo, Roberto Grassi, Vittorio Miele, Andrea Giovagnoni, Andrea Laghi, and SIRM expert group on Artificial Intelligence. Explainable ai in radiology: a white paper of the italian society of medical and interventional radiology. *La radiologia medica*, 128(6):755–764, 2023.

[182] Arun Rai. Explainable ai: From black box to glass box. *Journal of the Academy of Marketing Science*, 48:137–141, 2020.

[183] Zhen Lu, Imran Afridi, Hong Jin Kang, Ivan Ruchkin, and Xi Zheng. Surveying neuro-symbolic approaches for reliable artificial intelligence of things. *Journal of Reliable Intelligent Environments*, pages 1–23, 2024.

[184] Yoshua Bengio, Tristan Deleu, Nasim Rahaman, Rosemary Ke, Sébastien Lachapelle, Olexa Bilaniuk, Anirudh Goyal, and Christopher Pal. A meta-transfer objective for learning to disentangle causal mechanisms. *arXiv preprint arXiv:1901.10912*, 2019.

[185] Shaina Raza, Rizwan Qureshi, Marcelo Lotif, Aman Chadha, Deval Pandya, and Christos Emmanouilidis. Just as humans need vaccines, so do models: Model immunization to combat falsehoods. *arXiv preprint arXiv:2505.17870*, 2025.

[186] Kenji Kawaguchi, Leslie Pack Kaelbling, and Yoshua Bengio. Generalization in deep learning. *arXiv preprint arXiv:1710.05468*, 1(8), 2017.

[187] Naftali Tishby, Fernando C Pereira, and William Bialek. The information bottleneck method. *arXiv preprint physics/0004057*, 2000.

[188] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. *arXiv preprint arXiv:1703.00810*, 2017.

[189] Ravid Shwartz-Ziv, Amichai Painsky, and Naftali Tishby. Representation compression and generalization in deep neural networks. *arXiv preprint arXiv:1805.00915*, 2018.

[190] Ravid Shwartz-Ziv and Yann LeCun. To compress or not to compress—self-supervised learning and information theory: A review. *Entropy*, 26(3):252, 2024.

[191] Jürgen Schmidhuber. Simple algorithmic principles of discovery, subjective beauty, selective attention, curiosity & creativity. In *International conference on discovery science*, pages 26–38. Springer, 2007.

[192] Behnam Neyshabur, Ryota Tomioka, and Nathan Srebro. In search of the real inductive bias: On the role of implicit regularization in deep learning. *arXiv preprint arXiv:1412.6614*, 2014.

[193] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. *Advances in neural information processing systems*, 31, 2018.

[194] Mikhail Belkin, Daniel Hsu, Siyuan Ma, and Soumik Mandal. Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences*, 116(32):15849–15854, 2019.

[195] David A McAllester. Pac-bayesian model averaging. In *Proceedings of the twelfth annual conference on Computational learning theory*, pages 164–170, 1999.

[196] Bernhard Schölkopf, Francesco Locatello, Stefan Bauer, Nan Rosemary Ke, Nal Kalchbrenner, Anirudh Goyal, and Yoshua Bengio. Toward causal representation learning. *Proceedings of the IEEE*, 109(5):612–634, 2021.

[197] Durk P Kingma, Tim Salimans, and Max Welling. Variational dropout and the local reparameterization trick. *Advances in neural information processing systems*, 28, 2015.

[198] Guillermo Valle-Perez, Chico Q Camargo, and Ard A Louis. Deep learning generalizes because the parameter-function map is biased towards simple functions. *arXiv preprint arXiv:1805.08522*, 2018.

[199] Marius-Constantin Popescu, Valentina E Balas, Liliana Perescu-Popescu, and Nikos Mastorakis. Multilayer perceptron and neural networks. *WSEAS Transactions on Circuits and Systems*, 8(7):579–588, 2009.

[200] Alex Sherstinsky. Fundamentals of recurrent neural network (rnn) and long short-term memory (lstm) network. *Physica D: Nonlinear Phenomena*, 404:132306, 2020.

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

[202] Jiasen Lu, Christopher Clark, Sangho Lee, Zichen Zhang, Savya Khosla, Ryan Marten, Derek Hoiem, and Aniruddha Kembhavi. Unified-io 2: Scaling autoregressive multimodal models with vision language audio and action. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 26439–26455, 2024.

[203] Albert Gu and Tri Dao. Mamba: Linear-time sequence modeling with selective state spaces. *arXiv preprint arXiv:2312.00752*, 2023.

[204] Peter W Battaglia, Jessica B Hamrick, Victor Bapst, Alvaro Sanchez-Gonzalez, Vinicius Zambaldi, Mateusz Malinowski, Andrea Tacchetti, David Raposo, Adam Santoro, Ryan Faulkner, et al. Relational inductive biases, deep learning, and graph networks. *arXiv preprint arXiv:1806.01261*, 2018.

[205] Ian J Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. Generative adversarial nets. *Advances in neural information processing systems*, 27, 2014.

[206] Zeke Xie, Issei Sato, and Masashi Sugiyama. A diffusion theory for deep learning dynamics: Stochastic gradient descent exponentially favors flat minima. *arXiv preprint arXiv:2002.03495*, 2020.

[207] Soham De, Anirbit Mukherjee, and Enayat Ullah. Convergence guarantees for rmsprop and adam in non-convex optimization and an empirical comparison to nesterov acceleration. *arXiv preprint arXiv:1807.06766*, 2018.

[208] Qi Wang, Yue Ma, Kun Zhao, and Yingjie Tian. A comprehensive survey of loss functions in machine learning. *Annals of Data Science*, 9(2):187–212, 2022.

[209] Andreas Sedlmeier, Michael Kölle, Robert Müller, Leo Baudrexel, and Claudia Linnhoff-Popien. Quantifying multimodality in world models. *arXiv preprint arXiv:2112.07263*, 2021.

[210] Rohan Anil, Vineet Gupta, Tomer Koren, and Yoram Singer. Memory efficient adaptive optimization. *Advances in Neural Information Processing Systems*, 32, 2019.

[211] Jian Liang, Ran He, and Tieniu Tan. A comprehensive survey on test-time adaptation under distribution shifts. *International Journal of Computer Vision*, 133(1):31–64, 2025.

[212] Yu Sun, Xiaolong Wang, Zhuang Liu, John Miller, Alexei Efros, and Moritz Hardt. Test-time training with self-supervision for generalization under distribution shifts. In *International conference on machine learning*, pages 9229–9248. PMLR, 2020.

[213] Manli Shu, Weili Nie, De-An Huang, Zhiding Yu, Tom Goldstein, Anima Anandkumar, and Chaowei Xiao. Test-time prompt tuning for zero-shot generalization in vision-language models. *Advances in Neural Information Processing Systems*, 35:14274–14289, 2022.

[214] Adilbek Karmanov, Dayan Guan, Shijian Lu, Abdulmotaleb El Saddik, and Eric Xing. Efficient test-time adaptation of vision-language models. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 14162–14171, 2024.

[215] Yabin Zhang, Wenjie Zhu, Hui Tang, Zhiyuan Ma, Kaiyang Zhou, and Lei Zhang. Dual memory networks: A versatile adaptation approach for vision-language models. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 28718–28728, 2024.

[216] Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, et al. Retrieval-augmented generation for knowledge-intensive nlp tasks. *Advances in neural information processing systems*, 33:9459–9474, 2020.

[217] Weijia Shi, Sewon Min, Michihiro Yasunaga, Minjoon Seo, Rich James, Mike Lewis, Luke Zettlemoyer, and Wen-tau Yih. Replug: Retrieval-augmented black-box language models. *arXiv preprint arXiv:2301.12652*, 2023.

[218] Giuseppe Paolo, Jonas Gonzalez-Billandon, and Balázs Kégl. A call for embodied ai. *arXiv preprint arXiv:2402.03824*, 2024.

[219] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. *arXiv preprint arXiv:2209.00626*, 2022.

[220] Richard S Sutton and Andrew G Barto. Reinforcement learning: an introduction mit press. *Cambridge, MA*, 22447(10), 1998.

[221] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *nature*, 550(7676):354–359, 2017.

[222] Tuomas Haarnoja, Aurick Zhou, Pieter Abbeel, and Sergey Levine. Soft actor-critic: Off-policy maximum entropy deep reinforcement learning with a stochastic actor. In *International conference on machine learning*, pages 1861–1870. Pmlr, 2018.

[223] Sirui Hong, Xiawu Zheng, Jonathan Chen, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, Liyang Zhou, et al. Metagpt: Meta programming for multi-agent collaborative framework. *arXiv preprint arXiv:2308.00352*, 3(4):6, 2023.

[224] Aoran Jiao, Tanmay P Patel, Sanjmi Khurana, Anna-Mariya Korol, Lukas Brunke, Vivek K Adajania, Utku Culha, Siqi Zhou, and Angela P Schoellig. Swarm-gpt: Combining large language models with safe motion planning for robot choreography design. *arXiv preprint arXiv:2312.01059*, 2023.

[225] Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, et al. Training language models to follow instructions with human feedback. *Advances in neural information processing systems*, 35:27730–27744, 2022.

[226] Paul F Christiano, Jan Leike, Tom Brown, Miljan Martic, Shane Legg, and Dario Amodei. Deep reinforcement learning from human preferences. *Advances in neural information processing systems*, 30, 2017.

[227] Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, et al. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv preprint arXiv:2204.05862*, 2022.

[228] Robert X Gao, Jörg Krüger, Marion Merklein, Hans-Christian Möhring, and József Váncza. Artificial intelligence in manufacturing: State of the art, perspectives, and future directions. *CIRP Annals*, 2024.

[229] Paul M Salmon, Chris Baber, Catherine Burns, Tony Carden, Nancy Cooke, Missy Cummings, Peter Hancock, Scott McLean, Gemma JM Read, and Neville A Stanton. Managing the risks of artificial general intelligence: A human factors and ergonomics perspective. *Human Factors and Ergonomics in Manufacturing & Service Industries*, 33(5):366–378, 2023.

[230] Martin Andreoni, Willian T Lunardi, George Lawton, and Shreekant Thakkar. Enhancing autonomous system security and resilience with generative ai: A comprehensive survey. *IEEE Access*, 2024.

[231] Yue Zhao and Jiequn Han. Offline supervised learning vs online direct policy optimization: A comparative study and a unified training paradigm for neural network-based optimal feedback control. *Physica D: Nonlinear Phenomena*, 462:134130, 2024.

[232] Shaina Raza, Ashmal Vayani, Aditya Jain, Aravind Narayanan, Vahid Reza Khazaie, Syed Raza Bashir, Elham Dolatabadi, Gias Uddin, Christos Emmanouilidis, Rizwan Qureshi, et al. Vldbench: Vision language models disinformation detection benchmark. *arXiv preprint arXiv:2502.11361*, 2025.

[233] Sijia Liu, Yuanshun Yao, Jinghan Jia, Stephen Casper, Nathalie Baracaldo, Peter Hase, Yuguang Yao, Chris Yuhao Liu, Xiaojun Xu, Hang Li, et al. Rethinking machine unlearning for large language models. *Nature Machine Intelligence*, pages 1–14, 2025.

[234] Vishal Narnaware, Ashmal Vayani, Rohit Gupta, Sirnam Swetha, and Mubarak Shah. Sb-bench: Stereotype bias benchmark for large multimodal models. *arXiv preprint arXiv:2502.08779*, 2025.

[235] Ben Shneiderman. Bridging the gap between ethics and practice: guidelines for reliable, safe, and trustworthy human-centered ai systems. *ACM Transactions on Interactive Intelligent Systems (TiiS)*, 10(4):1–31, 2020.

[236] Michael Mylrea and Nikki Robinson. Artificial intelligence (ai) trust framework and maturity model: applying an entropy lens to improve security, privacy, and ethical ai. *Entropy*, 25(10):1429, 2023.

[237] Lixiang Yan, Lele Sha, Linxuan Zhao, Yuheng Li, Roberto Martinez-Maldonado, Guanliang Chen, Xinyu Li, Yueqiao Jin, and Dragan Gašević. Practical and ethical challenges of large language models in education: A systematic scoping review. *British Journal of Educational Technology*, 55(1):90–112, 2024.

[238] Dileesh Chandra Bikkasani. Navigating artificial general intelligence (agi): societal implications, ethical considerations, and governance strategies. *AI and Ethics*, pages 1–16, 2024.

[239] Scott McLean, Gemma JM Read, Jason Thompson, Chris Baber, Neville A Stanton, and Paul M Salmon. The risks associated with artificial general intelligence: A systematic review. *Journal of Experimental & Theoretical Artificial Intelligence*, 35(5):649–663, 2023.

[240] Yogesh K Dwivedi, Nir Kshetri, Laurie Hughes, Emma Louise Slade, Anand Jeyaraj, Arpan Kumar Kar, Abdullah M Baabdullah, Alex Koohang, Vishnupriya Raghavan, Manju Ahuja, et al. Opinion paper:"so what if chatgpt wrote it?" multidisciplinary perspectives on opportunities, challenges and implications of generative conversational ai for research, practice and policy. *International Journal of Information Management*, 71:102642, 2023.

[241] Zheng Zhang, Levent Yilmaz, and Bo Liu. A critical review of inductive logic programming techniques for explainable ai. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[242] Marcello Mariani and Yogesh K Dwivedi. Generative artificial intelligence in innovation management: A preview of future research developments. *Journal of Business Research*, 175:114542, 2024.

[243] Jana Al Haj Ali, Ben Gaffinet, Hervé Panetto, and Yannick Naudet. Cognitive systems and interoperability in the enterprise: A systematic literature review. *Annual Reviews in Control*, 57:100954, 2024.

[244] Ron Campos, Ashmal Vayani, Parth Parag Kulkarni, Rohit Gupta, Aritra Dutta, and Mubarak Shah. Gaea: A geolocation aware conversational model. *arXiv preprint arXiv:2503.16423*, 2025.

[245] Pei Wang, Xiang Li, and Patrick Hammer. Self in nars, an agi system. *Frontiers in Robotics and AI*, 5:20, 2018.

[246] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[247] Fatemeh Golpayegani, Saeedeh Ghanadbashi, and Akram Zarchini. Advancing sustainable manufacturing: Reinforcement learning with adaptive reward machine using an ontology-based approach. *Sustainability*, 16(14):5873, 2024.

[248] Mohammad Mustafa Taye. Understanding of machine learning with deep learning: architectures, workflow, applications and future directions. *Computers*, 12(5):91, 2023.

[249] Ljubiša Bojić, Matteo Cinelli, Dubravko Ćulibrk, and Boris Delibašić. Cern for ai: a theoretical framework for autonomous simulation-based artificial intelligence testing and alignment. *European Journal of Futures Research*, 12(1):15, 2024.

[250] Lukai Li, Luping Shi, and Rong Zhao. A vertical-horizontal integrated neuro-symbolic framework towards artificial general intelligence. In *International Conference on Artificial General Intelligence*, pages 197–206. Springer, 2023.

[251] Rao Mikkilineni, W Patrick Kelly, and Gideon Crawley. Digital genome and self-regulating distributed software applications with associative memory and event-driven history. *Computers*, 13(9):220, 2024.

[252] Peter Isaev and Patrick Hammer. Memory system and memory types for real-time reasoning systems. In *International Conference on Artificial General Intelligence*, pages 147–157. Springer, 2023.

[253] Jürgen Schmidhuber, Sepp Hochreiter, et al. Long short-term memory. *Neural Comput*, 9(8):1735–1780, 1997.

[254] Yu-Dong Zhang, Zhengchao Dong, Shui-Hua Wang, Xiang Yu, Xujing Yao, Qinghua Zhou, Hua Hu, Min Li, Carmen Jiménez-Mesa, Javier Ramirez, et al. Advances in multimodal data fusion in neuroimaging: overview, challenges, and novel orientation. *Information Fusion*, 64:149–187, 2020.

[255] Michael I Posner. *Cognitive neuroscience of attention*. Guilford Press, 2012.

[256] Bernard J Baars. *Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience*, volume 150. Elsevier, 2005.

[257] Alison Gopnik and Laura Schulz. Theory of mind and causal learning in children: The devil is in the details. *Behavioral and Brain Sciences*, 27(1):126–127, 2004.

[258] David Premack and Guy Woodruff. Does the chimpanzee have a theory of mind? *Behavioral and Brain Sciences*, 1(4):515–526, 1978.

[259] Iyad Rahwan, Manuel Cebrian, Josh Bongard, and et al. Combining psychology with artificial intelligence: What could possibly go wrong? *Nature Machine Intelligence*, 4:12–13, 2022.

[260] Wissam Salhab, Darine Ameyed, Fehmi Jaafar, and Hamid Mcheick. A systematic literature review on ai safety: Identifying trends, challenges and future directions. *IEEE Access*, 2024.

[261] Jonas Schuett, Noemi Dreksler, Markus Anderljung, David McCaffary, Lennart Heim, Emma Bluemke, and Ben Garfinkel. Towards best practices in agi safety and governance: A survey of expert opinion. *arXiv preprint arXiv:2305.07153*, 2023.

[262] Yifeng He, Ethan Wang, Yuyang Rong, Zifei Cheng, and Hao Chen. Security of ai agents. *arXiv preprint arXiv:2406.08689*, 2024.

[263] Peter Cihon. Chilling autonomy: Policy enforcement for human oversight of ai agents. In *41st International Conference on Machine Learning, Workshop on Generative AI and Law*, 2024.

[264] Shangding Gu, Long Yang, Yali Du, Guang Chen, Florian Walter, Jun Wang, and Alois Knoll. A review of safe reinforcement learning: Methods, theory and applications. *arXiv preprint arXiv:2205.10330*, 2022.

[265] Simon Burton, Benjamin Herd, and João-Vitor Zacchi. Uncertainty-aware evaluation of quantitative ml safety requirements. In *International Conference on Computer Safety, Reliability, and Security*, pages 391–404. Springer, 2024.

[266] Nicolas Guzman. Advancing nsfw detection in ai: Training models to detect drawings, animations, and assess degrees of sexiness. *Journal of Knowledge Learning and Science Technology ISSN: 2959-6386 (online)*, 2(2):275–294, 2023.

[267] Nataliya Kosmyna, Eugene Hauptmann, Ye Tong Yuan, Jessica Situ, Xian-Hao Liao, Ashly Vivian Beresnitzky, Iris Braunstein, and Pattie Maes. Your brain on chatgpt: Accumulation of cognitive debt when using an ai assistant for essay writing task. *arXiv preprint arXiv:2506.08872*, 2025.

[268] World Economic Forum. The future of jobs report. Technical report, World Economic Forum, 2025.

[269] Wim Naudé and Nicola Dimitri. The race for an artificial general intelligence: implications for public policy. *AI & society*, 35:367–379, 2020.

[270] NIST AI. Artificial intelligence risk management framework (ai rmf 1.0). *URL: https://nvlpubs. nist. gov/nistpubs/ai/nist. ai*, pages 100–1, 2023.

[271] Dorine Eva Van Norren. The ethics of artificial intelligence, unesco and the african ubuntu perspective. *Journal of Information, Communication and Ethics in Society*, 21(1):112–128, 2023.

[272] Xiao-Li Meng. Data science and ai: Everything everywhere all at once. *Harvard Data Science Review*, 7(1), 2025.

[273] Shaina Raza, Oluwanifemi Bamgbose, Shardul Ghuge, Fatemeh Tavakoli, and Deepak John Reji. Developing safe and responsible large language models–a comprehensive framework. *arXiv preprint arXiv:2404.01399*, 2024.

[274] Peter Voss and Mladjan Jovanovic. Why we don't have agi yet. *arXiv preprint arXiv:2308.03598*, 2023.

[275] Jianxi Luo. Designing the future of the fourth industrial revolution, 2023.

[276] Aarohi Srivastava et al. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *arXiv preprint arXiv:2206.04615*, 2022.

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

[278] Linxi Fan, Guanzhi Wang, Yunfan Jiang, Ajay Mandlekar, Yuncong Yang, Haoyi Zhu, Andrew Tang, De-An Huang, Yuke Zhu, and Anima Anandkumar. Minedojo: Building open-ended embodied agents with internet-scale knowledge. *Advances in Neural Information Processing Systems*, 35:18343–18362, 2022.

[279] Maxime Chevalier-Boisvert, Dzmitry Bahdanau, Salem Lahlou, Lucas Willems, Chitwan Saharia, Thien Huu Nguyen, and Yoshua Bengio. Babyai: A platform to study the sample efficiency of grounded language learning. *arXiv preprint arXiv:1810.08272*, 2018.

[280] Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, et al. Agentbench: Evaluating llms as agents. *arXiv preprint arXiv:2308.03688*, 2023.

[281] Sandeep Neema, Susmit Jha, Adam Nagel, Ethan Lew, Chandrasekar Sureshkumar, Aleksa Gordic, Chase Shimmin, Hieu Nguygen, and Paul Eremenko. On the evaluation of engineering artificial general intelligence. *arXiv preprint arXiv:2505.10653*, 2025.

[282] Mark J Wagner and Liqun Luo. Neocortex–cerebellum circuits for cognitive processing. *Trends in neurosciences*, 43(1):42–54, 2020.

[283] Seralynne D Vann and Mathieu M Albasser. Hippocampus and neocortex: recognition and spatial memory. *Current opinion in neurobiology*, 21(3):440–445, 2011.

[284] Haixing Dai. *Brain-inspired Approaches for Advancing Artificial Intelligence*. PhD thesis, University of Georgia, 2023.

[285] Deborah E Hannula, Jennifer D Ryan, and David E Warren. *Beyond long-term declarative memory: Evaluating hippocampal contributions to unconscious memory expression, perception, and short-term retention*. Springer, 2017.

[286] Igor Dakat, Isadora Langley, Lysander Montgomery, Rosalin Bennett, and Lysandra Blackwood. Enhancing large language models through dynamic contextual memory embedding: A technical evaluation. *Authorea Preprints*, 2024.

[287] Guido Schillaci, Verena V Hafner, and Bruno Lara. Exploration behaviors, body representations, and simulation processes for the development of cognition in artificial agents. *Frontiers in Robotics and AI*, 3:39, 2016.

[288] Alhassan Mumuni and Fuseini Mumuni. Large language models for artificial general intelligence (agi): A survey of foundational principles and approaches. *arXiv preprint arXiv:2501.03151*, 2025.

[289] Yuri Calleo, Amos Taylor, Francesco Pilla, and Simone Di Zio. Ai-assisted real-time spatial delphi: integrating artificial intelligence models for advancing future scenarios analysis. *Quality & Quantity*, pages 1–33, 2025.

[290] Andrew Zhao, Yiran Wu, Yang Yue, Tong Wu, Quentin Xu, Matthieu Lin, Shenzhi Wang, Qingyun Wu, Zilong Zheng, and Gao Huang. Absolute zero: Reinforced self-play reasoning with zero data. *arXiv preprint arXiv:2505.03335*, 2025.

[291] Jiaqi Chen, Yuxian Jiang, Jiachen Lu, and Li Zhang. S-agents: Self-organizing agents in open-ended environments. *arXiv preprint arXiv:2402.04578*, 2024.

[292] Helen Canton. Organisation for economic cooperation and development—oecd. In *The Europa Directory of International Organizations 2021*, pages 677–687. Routledge, 2021.

[293] IEEE Standards Association et al. Eee ethically aligned design: A vision for prioritizing human well-being with autonomous and intelligent systems. *IEEE Standards Association: Piscataway, NJ, USA*, 2019.

[294] Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. Show and tell: A neural image caption generator. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 3156–3164, 2015.

[295] Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C Lawrence Zitnick, and Devi Parikh. Vqa: Visual question answering. In *Proceedings of the IEEE international conference on computer vision*, pages 2425–2433, 2015.

[296] Ranjan Sapkota and Manoj Karkee. Object detection with multimodal large vision-language models: An in-depth review. *Available at SSRN 5233953*, 2025.

[297] James F Peters. *Foundations of computer vision: computational geometry, visual image structures and object shape detection*, volume 124. Springer, 2017.

[298] Papers with Code. Pascal voc dataset, 2024. Accessed: 2025-04-18.

[299] Papers with Code. Flickr30k dataset, 2024. Accessed: 2025-04-18.

[300] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: human language technologies, volume 1 (long and short papers)*, pages 4171–4186, 2019.

[301] Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, et al. An image is worth 16x16 words: Transformers for image recognition at scale. *arXiv preprint arXiv:2010.11929*, 2020.

[302] Ranjan Sapkota, Konstantinos I Roumeliotis, Rahul Harsha Cheppally, Marco Flores Calero, and Manoj Karkee. A review of 3d object detection with vision-language models, 2025.

[303] Shuqi Guo, Ge Zhang, Xin Zeng, Yue Xiong, Yuanhang Xu, Yan Cui, Dezhong Yao, and Daqing Guo. Ten years of the digital twin brain: Perspectives and challenges. *Europhysics Letters*, 2025.

[304] Ranjan Sapkota, Yang Cao, Konstantinos I. Roumeliotis, and Manoj Karkee. Vision-language-action models: Concepts, progress, applications and challenges, 2025.

[305] Florian Bordes, Richard Yuanzhe Pang, Anurag Ajay, Alexander C Li, Adrien Bardes, Suzanne Petryk, Oscar Mañas, Zhiqiu Lin, Anas Mahmoud, Bargav Jayaraman, et al. An introduction to vision-language modeling. *arXiv preprint arXiv:2405.17247*, 2024.

[306] Chao Jia, Yinfei Yang, Ye Xia, Yi-Ting Chen, Zarana Parekh, Hieu Pham, Quoc Le, Yun-Hsuan Sung, Zhen Li, and Tom Duerig. Scaling up visual and vision-language representation learning with noisy text supervision. In *International conference on machine learning*, pages 4904–4916. PMLR, 2021.

[307] Jean-Baptiste Alayrac, Jeff Donahue, Pauline Luc, Antoine Miech, Iain Barr, Yana Hasson, Karel Lenc, Arthur Mensch, Katherine Millican, Malcolm Reynolds, et al. Flamingo: a visual language model for few-shot learning. *Advances in neural information processing systems*, 35:23716–23736, 2022.

[308] Xi Chen, Xiao Wang, Soravit Changpinyo, AJ Piergiovanni, Piotr Padlewski, Daniel Salz, Sebastian Goodman, Adam Grycner, Basil Mustafa, Lucas Beyer, et al. Pali: A jointly-scaled multilingual language-image model. *arXiv preprint arXiv:2209.06794*, 2022.

[309] Jingyi Zhang, Jiaxing Huang, Sheng Jin, and Shijian Lu. Vision-language models for vision tasks: A survey. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2024.

[310] Yonadav Shavit, Sandhini Agarwal, Miles Brundage, Steven Adler, Cullen O'Keefe, Rosie Campbell, Teddy Lee, Pamela Mishkin, Tyna Eloundou, Alan Hickey, et al. Practices for governing agentic ai systems. *Research Paper, OpenAI*, 2023.

[311] Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in Neural Information Processing Systems*, 36:51991–52008, 2023.

[312] Yingqiang Ge, Wenyue Hua, Kai Mei, Juntao Tan, Shuyuan Xu, Zelong Li, Yongfeng Zhang, et al. Openagi: When llm meets domain experts. *Advances in Neural Information Processing Systems*, 36:5539–5568, 2023.

[313] Zhiliang Peng, Wenhui Wang, Li Dong, Yaru Hao, Shaohan Huang, Shuming Ma, and Furu Wei. Kosmos-2: Grounding multimodal large language models to the world. *arXiv preprint arXiv:2306.14824*, 2023.

[314] Youssef Mroueh. Reinforcement learning with verifiable rewards: Grpo's effective loss, dynamics, and success amplification. *arXiv preprint arXiv:2503.06639*, 2025.

[315] Michael Moor, Oishi Banerjee, Zahra Shakeri Hossein Abad, Harlan M Krumholz, Jure Leskovec, Eric J Topol, and Pranav Rajpurkar. Foundation models for generalist medical artificial intelligence. *Nature*, 616(7956):259–265, 2023.

[316] Yupeng Chang, Xu Wang, Jindong Wang, Yuan Wu, Linyi Yang, Kaijie Zhu, Hao Chen, Xiaoyuan Yi, Cunxiang Wang, Yidong Wang, et al. A survey on evaluation of large language models. *ACM Transactions on Intelligent Systems and Technology*, 15(3):1–45, 2024.

[317] Yuting Wu, Ziyu Wang, and Wei D Lu. Pim gpt a hybrid process in memory accelerator for autoregressive transformers. *npj Unconventional Computing*, 1(1):4, 2024.

[318] Lukas Netz, Jan Reimer, and Bernhard Rumpe. Using grammar masking to ensure syntactic validity in llm-based modeling tasks. In *Proceedings of the ACM/IEEE 27th International Conference on Model Driven Engineering Languages and Systems*, pages 115–122, 2024.

[319] Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. *Proceedings of the National Academy of Sciences*, 121(27):e2311878121, 2024.

[320] Xinyi Hou, Yanjie Zhao, Shenao Wang, and Haoyu Wang. Model context protocol (mcp): Landscape, security threats, and future research directions. *arXiv preprint arXiv:2503.23278*, 2025.

[321] Md Shamsujjoha, Qinghua Lu, Dehai Zhao, and Liming Zhu. Swiss cheese model for ai safety: A taxonomy and reference architecture for multi-layered guardrails of foundation model based agents. In *2025 IEEE 22nd International Conference on Software Architecture (ICSA)*, pages 37–48. IEEE, 2025.

[322] Jason Jabbour and Vijay Janapa Reddi. Generative ai agents in autonomous machines: A safety perspective. In *Proceedings of the 43rd IEEE/ACM International Conference on Computer-Aided Design*, pages 1–13, 2024.

[323] Loïc Barrault, Paul-Ambroise Duquenne, Maha Elbayad, Artyom Kozhevnikov, Belen Alastruey, Pierre Andrews, Mariano Coria, Guillaume Couairon, Marta R Costa-jussà, David Dale, et al. Large concept models: Language modeling in a sentence representation space. *arXiv preprint arXiv:2412.08821*, 2024.

[324] Paul-Ambroise Duquenne, Holger Schwenk, and Benoît Sagot. Sonar: sentence-level multimodal and language-agnostic representations. *arXiv preprint arXiv:2308.11466*, 2023.

[325] Armen Aghajanyan, Bernie Huang, Candace Ross, Vladimir Karpukhin, Hu Xu, Naman Goyal, Dmytro Okhonko, Mandar Joshi, Gargi Ghosh, Mike Lewis, et al. Cm3: A causal masked multimodal model of the internet. *arXiv preprint arXiv:2201.07520*, 2022.

[326] Siwei Wu, Zhongyuan Peng, Xinrun Du, Tuney Zheng, Minghao Liu, Jialong Wu, Jiachen Ma, Yizhi Li, Jian Yang, Wangchunshu Zhou, et al. A comparative study on reasoning patterns of openai's o1 model. *arXiv preprint arXiv:2410.13639*, 2024.

[327] Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, et al. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*, 2021.

[328] Andrzej Cichocki and Alexander P Kuleshov. Future trends for human-ai collaboration: A comprehensive taxonomy of ai/agi using multiple intelligences and learning styles. *Computational Intelligence and Neuroscience*, 2021(1):8893795, 2021.

[329] Sara Papi, Edmondo Trentin, Roberto Gretter, Marco Matassoni, and Daniele Falavigna. Mixtures of deep neural experts for automated speech scoring. *arXiv preprint arXiv:2106.12475*, 2021.

[330] William Fedus, Barret Zoph, and Noam Shazeer. Switch transformers: Scaling to trillion parameter models with simple and efficient sparsity. *Journal of Machine Learning Research*, 23(120):1–39, 2022.

[331] Wensheng Gan, Zhenyao Ning, Zhenlian Qi, and Philip S Yu. Mixture of experts (moe): A big data perspective. *arXiv preprint arXiv:2501.16352*, 2025.

[332] Hao Sun, Shaosen Li, Hao Li, Jianxiang Huang, Zhuqiao Qiao, Jialei Wang, and Xincui Tian. Invmoe: Moes based invariant representation learning for fault detection in converter stations. *Energies*, 18(7):1783, 2025.

[333] J. E. Korteling, G. C. van de Boer-Visschedijk, R. Blankendaal, R. Boonekamp, and A. R. Eikelboom. Human- versus artificial intelligence. *Frontiers in Artificial Intelligence*, 4, 2021.

[334] Krti Tallam. From autonomous agents to integrated systems, a new paradigm: Orchestrated distributed intelligence. *arXiv preprint arXiv:2503.13754*, 2025.

[335] Christian Schroeder de Witt. Open challenges in multi-agent security: Towards secure systems of interacting ai agents. *arXiv preprint arXiv:2505.02077*, 2025.

[336] Jia Deng, Wei Dong, Richard Socher, et al. Imagenet: A large-scale hierarchical image database. *CVPR*, 2009.

[337] Alex Wang, Amanpreet Singh, et al. Glue: A multitask benchmark and analysis platform for natural language understanding. In *ICLR*, 2019.

[338] Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, et al. Holistic evaluation of language models. *arXiv preprint arXiv:2211.09110*, 2022.

[339] Ashmal Vayani, Dinura Dissanayake, Hasindri Watawana, Noor Ahsan, Nevasini Sasikumar, Omkar Thawakar, Henok Biadglign Ademtew, Yahya Hmaiti, Amandeep Kumar, Kartik Kukreja, et al. All languages matter: Evaluating lmms on culturally diverse 100 languages. In *Proceedings of the Computer Vision and Pattern Recognition Conference*, pages 19565–19575, 2025.

[340] Dimitri Bertsekas. *Lessons from AlphaZero for optimal, model predictive, and adaptive control*. Athena Scientific, 2022.

[341] Markus Anderljung, Julian Hazell, and Moritz von Knebel. Protecting society from ai misuse: when are restrictions on capabilities warranted? *AI & SOCIETY*, pages 1–17, 2024.

[342] Shaina Raza, Oluwanifemi Bamgbose, Veronica Chatrath, Shardule Ghuge, Yan Sidyakin, and Abdullah Yahya Mohammed Muaad. Unlocking bias detection: Leveraging transformer-based models for content analysis. *IEEE Transactions on Computational Social Systems*, 2024.

[343] Rajeev Gupta, Suhani Gupta, Ronak Parikh, Divya Gupta, Amir Javaheri, and Jairaj Singh Shaktawat. Personalized artificial general intelligence (agi) via neuroscience-inspired continuous learning systems. *arXiv preprint arXiv:2504.20109*, 2025.

[344] Anton Kuznietsov, Balint Gyevnar, Cheng Wang, Steven Peters, and Stefano V. Albrecht. Explainable ai for safe and trustworthy autonomous driving: A systematic review. *IEEE Transactions on Intelligent Transportation Systems*, 25(12):19342–19364, 2024.

[345] Carl-Johan Hoel, Krister Wolff, and Leo Laine. Ensemble quantile networks: Uncertainty-aware reinforcement learning with applications in autonomous driving. *IEEE Transactions on Intelligent Transportation Systems*, 24(6):6030–6041, 2023.

[346] Gabriel Stanovsky, Renana Keydar, Gadi Perl, and Eliya Habba. Beyond benchmarks: On the false promise of ai regulation. *arXiv preprint arXiv:2501.15693*, 2025.

[347] Waddah Saeed and Christian Omlin. Explainable ai (xai): A systematic meta-survey of current challenges and future opportunities. *Knowledge-based systems*, 263:110273, 2023.

[348] Hongjun Guan, Liye Dong, and Aiwu Zhao. Ethical risk factors and mechanisms in artificial intelligence decision making. *Behavioral Sciences*, 12(9):343, 2022.

[349] Vikas Hassija, Vinay Chamola, Atmesh Mahapatra, Abhinandan Singal, Divyansh Goel, Kaizhu Huang, Simone Scardapane, Indro Spinelli, Mufti Mahmud, and Amir Hussain. Interpreting black-box models: a review on explainable artificial intelligence. *Cognitive Computation*, 16(1):45–74, 2024.

[350] John X Morris, Chawin Sitawarin, Chuan Guo, Narine Kokhlikyan, G Edward Suh, Alexander M Rush, Kamalika Chaudhuri, and Saeed Mahloujifar. How much do language models memorize? *arXiv preprint arXiv:2505.24832*, 2025.

[351] Federico Maria Cau, Hanna Hauptmann, Lucio Davide Spano, and Nava Tintarev. Effects of ai and logic-style explanations on users' decisions under different levels of uncertainty. *ACM Transactions on Interactive Intelligent Systems*, 13(4):1–42, 2023.

[352] Chulin Xie, Yangsibo Huang, Chiyuan Zhang, Da Yu, Xinyun Chen, Bill Yuchen Lin, Bo Li, Badih Ghazi, and Ravi Kumar. On memorization of large language models in logical reasoning. *arXiv preprint arXiv:2410.23123*, 2024.

[353] Patrick C Kyllonen and Raymond E Christal. Reasoning ability is (little more than) working-memory capacity?! *Intelligence*, 14(4):389–433, 1990.

[354] Chenxu Hu, Jie Fu, Chenzhuang Du, Simian Luo, Junbo Zhao, and Hang Zhao. Chatdb: Augmenting llms with databases as their symbolic memory. *arXiv preprint arXiv:2306.03901*, 2023.

[355] Zhiming Li, Yushi Cao, Xiufeng Xu, Junzhe Jiang, Xu Liu, Yon Shin Teo, Shang-Wei Lin, and Yang Liu. Llms for relational reasoning: How far are we? In *Proceedings of the 1st International Workshop on Large Language Models for Code*, pages 119–126, 2024.

[356] Ruichen Zhang, Hongyang Du, Yinqiu Liu, Dusit Niyato, Jiawen Kang, Sumei Sun, Xuemin Shen, and H Vincent Poor. Interactive ai with retrieval-augmented generation for next generation networking. *IEEE Network*, 2024.

[357] Stefania Costantini, Andrea Formisano, and Valentina Pitoni. An epistemic logic for modular development of multi-agent systems. In *International Workshop on Engineering Multi-Agent Systems*, pages 72–91. Springer, 2021.

[358] Adrián Scribano and Maximiliano E Korstanje. *AI and Emotions in Digital Society*. IGI Global, 2023.

[359] Piotr Boltuc. Moral space for paraconsistent agi. In *International Conference on Artificial General Intelligence*, pages 168–177. Springer, 2022.

[360] Minyang Chow and Olivia Ng. From technology adopters to creators: Leveraging ai-assisted vibe coding to transform clinical teaching and learning. *Medical Teacher*, pages 1–3, 2025.

[361] Peter Boltuc. Human-agi gemeinschaft as a solution to the alignment problem. In *International Conference on Artificial General Intelligence*, pages 33–42. Springer, 2024.

[362] Ai is set to drive surging electricity demand from data centres while offering the potential to transform how the energy sector works. https://www.iea.org/news/ai-is-set-to-drive-surging-electricity-demand-from-data-centres-while-offering-the-potential-to-transform-how-the-energy-sector-works. Accessed: 09-Jun-2025.

[363] Bhupinder Singh and Christian Kaunert. Dynamic landscape of artificial general intelligence (agi) for advancing renewable energy in urban environments: Synergies with sdg 11—sustainable cities and communities lensing policy and governance. In *Artificial General Intelligence (AGI) Security: Smart Applications and Sustainable Technologies*, pages 247–270. Springer, 2024.

Appendix

Table A1: Glossary of Terms

Term	Abbreviation	Definition
Abstract Reasoning Corpus	ARC	Benchmark that evaluates abstract reasoning and pattern-completion skills beyond surface pattern matching.
Agent Communication Protocol	ACP	Communication system designed for software agents allowing them to communicate using RESTful protocol.
Agent Network Protocol	ANP	Decentralised protocol using decentralized identifiers and semantic-web standards for discovery and collaboration among federated agents.
Agent2Agent Protocol	A2A	Peer-to-peer protocol where agents advertise capabilities via agent cards and negotiate task delegation.
ALIGN	ALIGN	Google vision–language model trained on noisy web-scale image–alt-text pairs for universal cross-modal representations.
AlphaFold2	AlphaFold2	Google DeepMind's AI system that predicts protein structure from amino acid sequences with high accuracy, revolutionizing structural biology.
AlphaGo	AlphaGo	Google DeepMind's reinforcement learning system that defeated world champions in the game of Go, combining deep neural networks with Monte Carlo tree search.
Application Programming Interface	APIs	Standardised interfaces that let separate software components communicate and exchange functionality or data.
Abstract Reasoning Corpus	ARC	Visual reasoning benchmark created by Francois Chollet that consists of puzzles where you need to figure out the underlying pattern or rule.
Artificial General Intelligence	AGI	Systems capable of flexible, human-level reasoning and learning across domains, without task-specific retraining.
Automated Language Model	ALM	Systematic approach to evaluating language models using automated testing procedures across multiple benchmarks and tasks without manual intervention.
AutoGPT	AutoGPT	Open-source agent that plans subtasks and calls tools autonomously via a planner–reflector loop over an LLM.
BabyAGI	BabyAGI	Minimal task-execution loop that prioritises tasks and stores context in a vector memory, driven by an LLM.
Beyond the Imitation Game Benchmark	BIG-Bench	Collaborative benchmark featuring diverse, challenging tasks designed to test capabilities beyond current language model performance.
CAMEL	CAMEL	Framework where two role-playing LLM agents collaborate via natural-language dialogue to solve tasks.
Cerebellum	Cerebellum	Brain region responsible for motor control, balance, and coordination, also involved in cognitive functions like language and learning.
Chain-of-Thought Prompting	CoT	A prompting technique that decomposes complex reasoning into interpretable sub-steps, improving performance on multi-step tasks.
CICERO	CICERO	Meta AI agent that achieved human-level performance in the game Diplomacy via strategic planning and natural-language negotiation.
Cognitive Debt	CD	Prolong reliance on AI may cause a gradual erosion of neural engagement, memory consolidation, and critical reasoning
Communicative Agents for Mind Exploration of Large Language Models	CAMEL	Framework enabling multiple role-playing LLM agents to collaborate via natural-language dialogue to solve complex tasks.
Computational Intelligence	CI	Umbrella field covering neural, evolutionary, fuzzy and swarm methods aimed at adaptive, intelligent behaviour.
Contrastive Language–Image Pre-training Model	CLIP	Contrastive Language–Image Pre-training model aligning textual and visual embeddings for zero-shot recognition.
Convolutional Neural Networks	CNNs	Neural network architectures that apply convolutional filters to capture spatial hierarchies in image data (e.g., edges → textures → objects).
Decentralized Identifier	DID	W3C standard for verifiable, self-sovereign digital identities that enable secure, decentralized authentication and authorization.
Deep Learning	DL	Sub-field of machine learning that trains deep (multi-layer) neural networks to learn hierarchical feature representations.
Deep Q-Network	DQN	Deep reinforcement learning algorithm that combines Q-learning with deep neural networks to learn optimal actions in complex environments.
Direct Preference Optimization	DPO	An alignment technique that trains models directly from human preference data, effectively turning an LLM into its own reward model for improved alignment.

Term	Abbreviation	Definition
Dual Memory Network	DMN	Architecture maintaining separate memory systems for different types of information, enabling flexible retrieval and reasoning.
Electroencephalography	EEG	Non-invasive neuro-imaging technique that records electrical activity via scalp electrodes, giving millisecond-level temporal resolution.
Electrocorticography	ECoG	Invasive recording of cortical surface potentials, offering higher spatial fidelity than EEG for research or clinical use.
ELIZA	ELIZA	Early chatbot developed in the 1960s that simulated conversation by using pattern matching and substitution methodology.
Episodic Memory	EM	The ability to recall and reuse specific past experiences, enabling context-aware reasoning and learning from interactions over time.
Explainable AI	XAI	A domain focused on making AI systems transparent and interpretable, embedding interpretability through neuro-symbolic reasoning, causal modeling, or attention mechanisms.
Flamingo	Flamingo	DeepMind vision-language model that performs few-shot image+text tasks via contrastive pre-training and frozen LLM backbone.
Frontoparietal Network	FPN	Large-scale brain network linking frontal and parietal cortices, implicated in executive control, attention, and flexible cognition.
Functional Magnetic Resonance Imaging	fMRI	Measures brain activity indirectly via blood-oxygen (BOLD) signals, producing whole-brain maps with millimetre spatial resolution.
General Language Understanding Evaluation	GLUE	Benchmark suite for evaluating natural language understanding across multiple tasks including sentiment analysis and textual entailment.
Gradient-weighted Class Activation Mapping	Grad-CAM	Explainability technique that produces visual explanations for CNN predictions by highlighting important regions in input images.
Group Relative Policy Optimization	GRPO	A method that optimizes reasoning quality by comparing multiple generated trajectories, improving alignment through relative policy evaluation.
Hippocampal	Hippocampal	Relating to or involving the hippocampus brain region, particularly in context of memory formation and spatial processing capabilities.
Hippocampus	Hippocampus	Brain region crucial for memory formation, spatial navigation, and learning, serving as a key inspiration for AI memory architectures.
Holistic Evaluation of Language Models	HELM	Comprehensive framework for evaluating language models across accuracy, calibration, robustness, fairness, bias, and efficiency.
Implicit Regularization	IR	Phenomenon where optimization methods (like SGD) naturally bias models toward solutions with better generalization properties.
Information Bottleneck	IB	A theoretical framework positing that models generalize well by compressing inputs into compact latent representations that retain only task-relevant information.
JavaScript Object Notation for Linked Data	JSON-LD	Method of encoding linked data using JSON, enabling semantic web standards and structured data representation.
JavaScript Object Notation Remote Procedure Call	JSON-RPC	Lightweight remote procedure call protocol using JSON for data interchange, enabling standardized communication between systems.
Kolmogorov–Arnold Networks	KANs	Networks using learnable spline-based activation functions rather than fixed ones, improving interpretability and flexibility in approximating complex functions.
Kullback-Leibler Divergence	KL	Measure of difference between probability distributions, commonly used in variational inference and information theory.
Large Action Models	LAMs	Foundation models that predict full action sequences (such as API calls, tool invocations) rather than next-word tokens, enabling embodied or tool-augmented decision making.
Large Language Models	LLMs	Large-scale models trained on massive text corpora for language understanding and generation.
Large Reasoning Models	LRMs	AI systems focusing on explicit, multi-step cognitive processes and extended inference-time computation for enhanced reasoning capabilities.
Learning to Think	L2T	Meta-learning paradigm where an agent improves its own reasoning procedure, not just task performance.
LeNet-5	LeNet-5	Convolutional neural network architecture developed by Yann LeCun for handwritten digit recognition.
Locked-image Tuning	LiT	Vision-language model focusing on efficient image-text alignment and generative capabilities for multimodal tasks.
Low-Rank Adaptation	LoRA	Parameter-efficient fine-tuning method that adapts large models by learning low-rank decompositions of weight updates.

Term	Abbreviation	Definition
Magnetoencephalography	MEG	Neuro-imaging that detects magnetic fields generated by neuronal currents, allowing source-localised brain-activity mapping.
Masked Autoencoder	MAE	Vision model pre-trained by reconstructing masked image patches, yielding strong features for downstream tasks.
MineDojo	MineDojo	Framework for open-ended agent learning in Minecraft, providing diverse tasks and environments for embodied AI research and evaluation.
Minimum Description Length	MDL	A principle from algorithmic information theory stating that the simplest model that best compresses the data will generalize more effectively.
Mixture of Experts	MoE	Neural architecture using a gating network to route each input to a small subset of specialised expert subnetworks.
Model Context Protocol	MCP	Specification for passing shared context (goals, world state) among heterogeneous models/agents in a pipeline.
Model-Agnostic Meta-Learning	MAML	Meta-learning algorithm that finds parameter initializations enabling fast adaptation to new tasks with minimal gradient steps.
Momentum Contrast	MoCo	Contrastive learning approach using a momentum-updated encoder to maintain consistent representations across training batches.
Multi-Agent Systems	MAS	Systems composed of multiple interacting agents that coordinate to perform complex tasks via communication and shared goals.
Multi-Layer Perceptrons	MLPs	Feedforward neural networks with multiple hidden layers, capable of learning complex nonlinear mappings between inputs and outputs.
Multipurpose Internet Mail Extensions	MIME	Standard defining format of email messages and, by extension, format of content in web communications and API interactions.
MYCIN	MYCIN	Early expert system developed in the 1970s for diagnosing bacterial infections and recommending antibiotics, representing rule-based AI approaches.
National Institute of Standards and Technology	NIST	U.S. federal agency developing technology standards, including frameworks for AI risk management and trustworthiness.
Natural Language–based Society of Mind	NLSOM	A modular architecture composed of multiple specialized agents that communicate via natural language, enabling collaborative reasoning and problem solving.
Neocortex	Neocortex	The outer layer of the cerebral cortex in mammals, responsible for higher-order cognitive functions including sensory perception, motor commands, and abstract reasoning.
Neural Tangent Kernel	NTK	A perspective showing that infinitely wide neural networks behave like kernel regressors during training, characterizing regimes of robust generalization.
NIST AI Risk Management Framework	NIST AI RMF	Framework promoting AI trustworthiness through interpretability, risk mitigation, security, privacy, and robustness guidelines.
Not Safe for Work	NSFW	Content classification system used to identify material inappropriate for professional or public settings, important for AI safety.
Occipital Lobes	Occipital Lobes	Brain regions primarily responsible for visual processing, containing the primary visual cortex and associated visual areas.
Organisation for Economic Co-operation and Development	OECD	International organization developing economic and social policy guidelines, including principles for AI governance.
PAC-Bayes Bounds	PAC-Bayes	Theoretical framework that upper-bounds generalisation error using a prior/posterior KL-divergence term.
Parameter-Efficient Fine-Tuning	PEFT	Techniques (such as LoRA, adapters) that adapt a large model by only training a small subset of parameters.
Parietal Lobes	Parietal Lobes	Brain regions involved in spatial processing, attention, and sensorimotor integration, crucial for coordinating perception and action.
Partial Differential Equations	PDEs	Mathematical equations describing relationships between functions and their partial derivatives, often encoding physical laws in PINNs.
Pascal Visual Object Classes	Pascal VOC	Benchmark dataset for object detection and image segmentation, instrumental in advancing computer vision research.
Pathways Language and Image Model	PaLI	Google's multilingual, multimodal model combining visual and textual pre-training for cross-modal understanding.
Physics-Informed Neural Networks	PINNs	Models that incorporate physical laws (such as partial differential equations) into their architecture, ensuring predictions remain consistent with known physics.
Positron Emission Tomography	PET	Imaging that uses radiotracers to capture metabolic or molecular processes, often combined with CT/MRI for anatomy.
Proximal Policy Optimization	PPO	An RL algorithm that balances policy improvement with stability by constraining updates to a trust region in policy space.

Term	Abbreviation	Definition
Q-Learning	Q-Learning	Model-free reinforcement learning algorithm that learns optimal action-value functions through temporal difference updates.
ReAct	ReAct	Prompting strategy that interleaves reasoning traces and actions, letting an LLM decide when to think or call a tool.
Recurrent Neural Networks	RNNs	Neural network architectures designed for sequential data, maintaining hidden states to capture temporal dependencies (such as time series, language).
Reinforcement Learning	RL	A learning paradigm where agents learn by interacting with the environment through trial-and-error to maximize cumulative reward.
Reinforcement Learning with Human Feedback	RLHF	A method that incorporates human judgments into the reinforcement learning reward loop to improve alignment and safety of learned behaviors.
Retrieval-Augmented Generation	RAG	A technique that augments model outputs by retrieving relevant external documents or knowledge during inference, improving factual accuracy.
Retrieval-Enhanced Transformer	RETRO	Architecture augmenting language models with retrieval mechanisms to access external knowledge during generation.
Self-Evolving Agentic AI	AZR	Research project exploring agents that autonomously update their policies, memories and objectives over long horizons.
Sentence-level Multimodal and Language-Agnostic Representations	SONAR	Multilingual, multimodal embedding framework supporting 200+ languages for cross-lingual and cross-modal understanding tasks.
Simple Contrastive Learning of Representations	SimCLR	Self-supervised learning method that learns representations by maximizing agreement between differently augmented views of data.
Small Language Model	SLM	Compact LLM (approx.100 M–1 B parameters) optimised for edge devices or cost-sensitive deployment.
Spike-Timing-Dependent Plasticity	STDP	Neurobiological learning rule where synaptic strength changes based on precise timing of pre- and post-synaptic neural spikes.
Spiking Neural Networks	SNNs	Biologically inspired networks that emulate neural spike dynamics (such as synaptic plasticity, spike timing), enabling event-driven, energy-efficient temporal processing.
Stochastic Gradient Descent	SGD	First-order optimisation algorithm that updates parameters using mini-batch estimates of the gradient.
Structural Equation Models	SEMs	Statistical models encoding causal relationships between variables, used in causal inference and representation learning.
Synaptic Activities	Synaptic	Electrochemical processes at neural connections that transmit information between neurons, including excitatory and inhibitory signals essential for all cognitive functions.
Temporal Lobes	Temporal Lobes	Brain regions housing auditory processing areas, memory structures (including hippocampus), and language comprehension areas.
Test-Time Adaptation	TTA	Techniques enabling models to adapt at inference time to distributional shifts, either by optimizing certain parameters on the test batch (optimization-based) or by modifying inference behavior without weight updates (training-free).
Test-Time Prompt Tuning	TPT	Lightweight variant of TTT that updates only soft prompts or prefix tokens at inference time.
Test-Time Training	TTT	Adapts a model on the test batch itself (usually self-supervised) to counter distribution shift during inference.
Trajectory Modelling	Trajectory Modelling	Framework that treats multi-step decision sequences as fundamental units for modeling, enabling AI systems to plan over extended horizons.
Training-Free Dynamic Adapter	TDA	Test-time adaptation approach that modifies inference behavior without weight updates to handle distribution shifts.
Tree-of-Thoughts Framework	ToT	A framework that enables exploration and evaluation of multiple reasoning paths via lookahead and backtracking, yielding gains in tasks requiring strategic planning.
United Nations Educational, Scientific and Cultural Organization	UNESCO	UN agency promoting global ethical standards for AI development, emphasizing equity, inclusiveness, and sustainability.
Vision Language Models	VLMs	Models that integrate visual perception and linguistic understanding for multimodal tasks, enabling capabilities such as visual question answering and image captioning.
Vision Transformer	ViT	Transformer architecture adapted for image recognition by treating image patches as sequence tokens, achieving state-of-the-art performance.
Voyager	Voyager	Open-ended embodied agent using large language models for autonomous exploration and skill acquisition in minecraft environments.

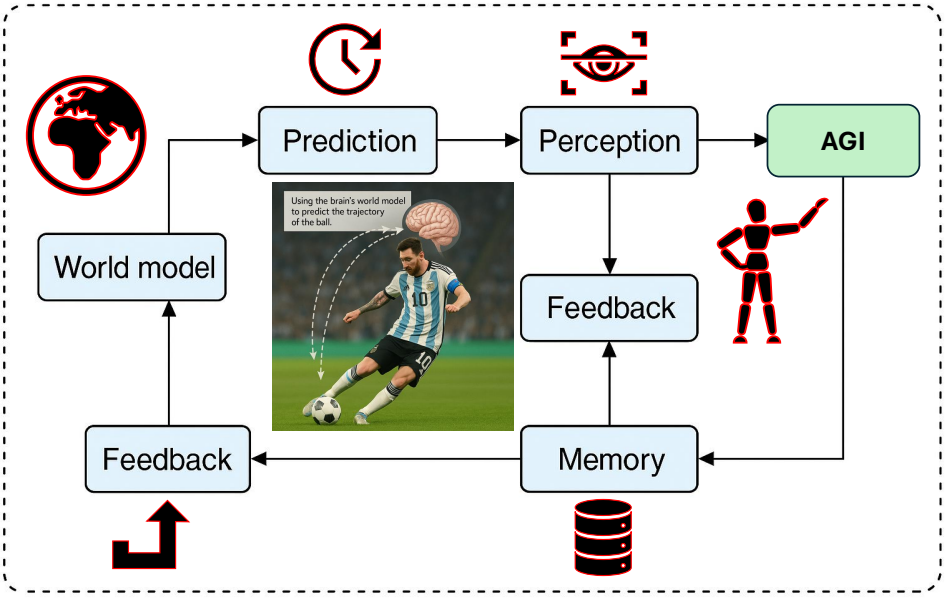

Figure A1: A conceptual framework illustrating how world models, prediction, perception, memory, and feedback interact to enable intelligent behavior. The diagram reflects a biologically inspired loop where internal world models allow an agent to anticipate outcomes, update its beliefs through perception and memory, and adapt actions based on feedback. This architecture mirrors human cognition and serves as a foundational blueprint for developing AGI.

Task Description and Prompt Example

Problem: A company's sales increased from \$1.2M to \$1.5M over a quarter. The marketing budget in the same period went from \$200K to \$300K. What is the percentage increase in sales per dollar spent on marketing? **Assistant's Response:** First, compute the sales-to-marketing ratio before and after the increase:

$$\text{Initial ratio} = \frac{1{,}200{,}000}{200{,}000} = 6 \tag{A1}$$

$$\text{Final ratio} = \frac{1{,}500{,}000}{300{,}000} = 5 \tag{A2}$$

Now calculate the percentage change in efficiency (sales per dollar):

$$\frac{5-6}{6} \times 100 = -16.67\%$$

Final Answer: Sales efficiency decreased by 16.67%. *Note: This example demonstrates the need for multimodal reasoning capabilities in AGI systems.*

Figure A2: **An example of multimodal reasoning in AGI systems**

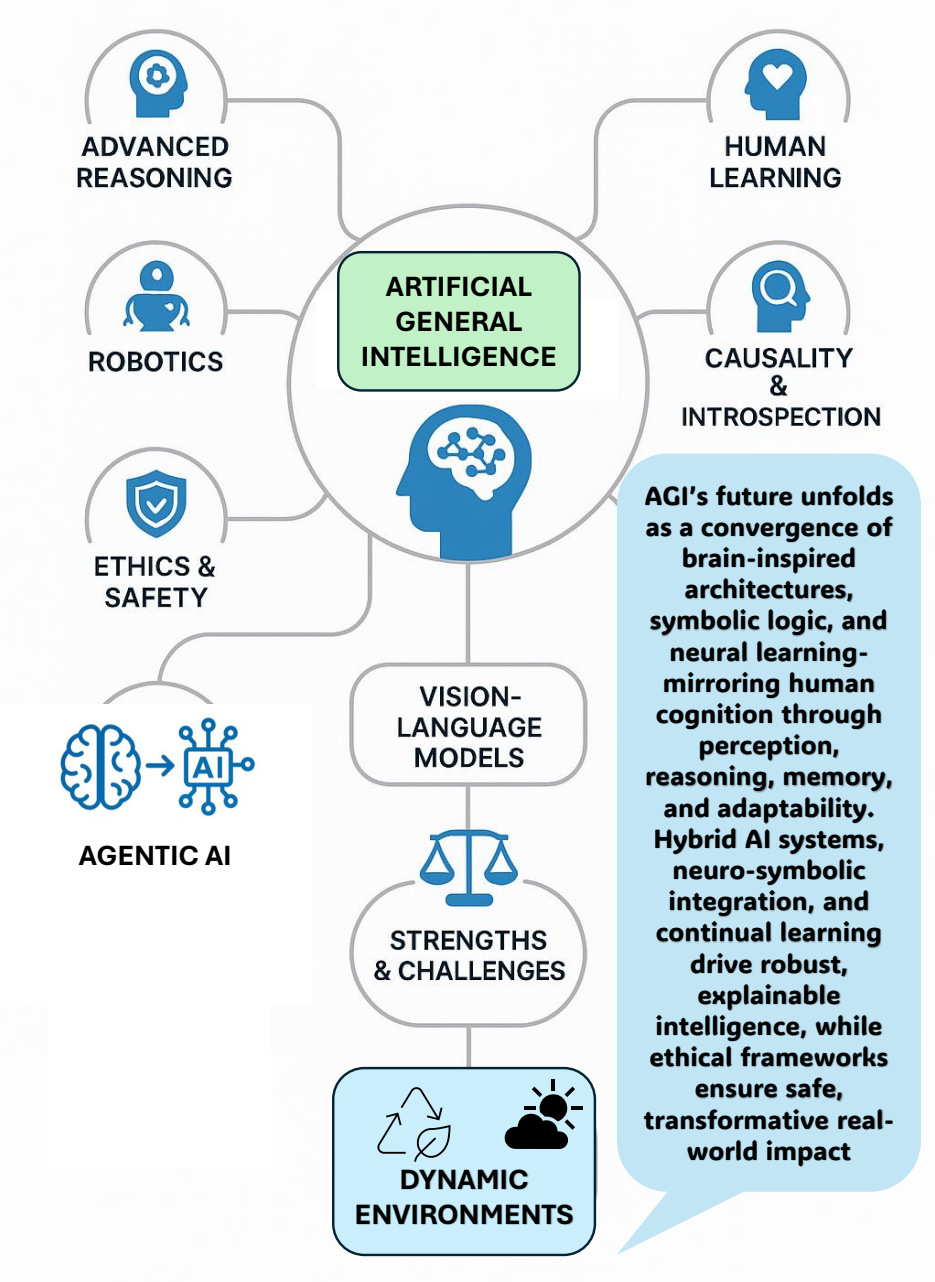

Figure A3: AGI Development Roadmap: Illustrating a scientific roadmap of AGI development, highlighting hybrid AI architectures, core cognitive functions, memory systems, perception models, and ethical safeguards. The diagram shows how neuroscience and AI converge to shape generalizable, human-aligned artificial intelligence.

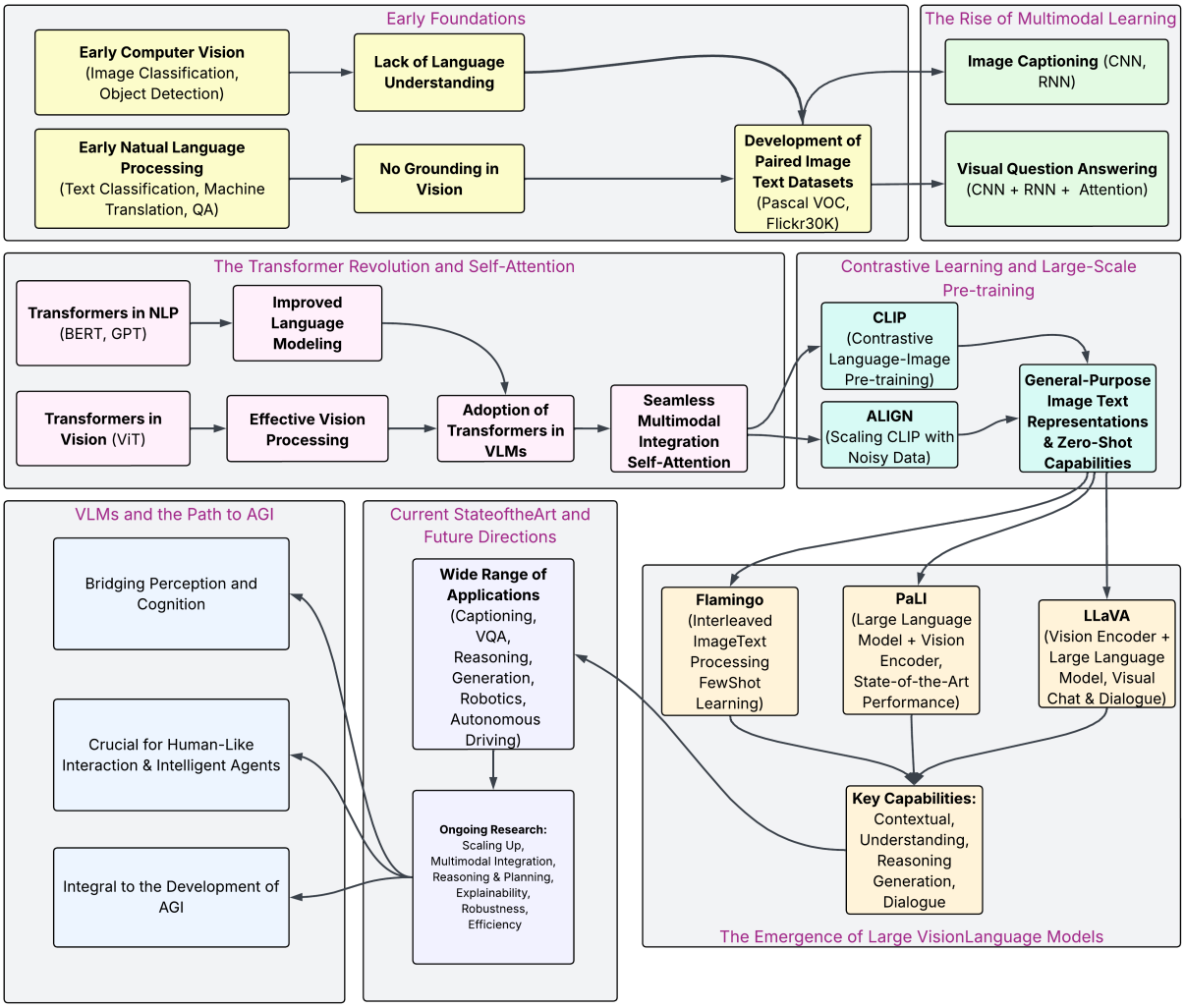

Figure A4: Conceptual roadmap tracing the evolution of VLMs. The figure outlines the progression from early unimodal systems in computer vision and natural language processing to modern VLMs enabled by self-attention, contrastive learning, and large-scale pretraining. It highlights pivotal developments such as paired image-text datasets, the adoption of transformers, and the emergence of general-purpose models like CLIP and ALIGN. The diagram also emphasizes the capabilities, applications, and future research directions of VLMs, positioning them as foundational components in the pursuit of AGI.