Knowledge Graph and Symbolic Approaches for the ARC-AGI-2 Benchmark: Architecture, Performance, and Implications

I. Introduction

A. Context: The Quest for Artificial General Intelligence (AGI) and the Role of Benchmarks

The pursuit of Artificial General Intelligence (AGI), defined as AI systems capable of performing any intellectual task that a human being can, represents a significant frontier in artificial intelligence research.[1] This ambition marks a departure from narrow AI, which excels at specific, predefined tasks, towards systems demonstrating general, adaptive intelligence.[3] Measuring progress towards AGI necessitates robust evaluation tools. Benchmarks play a critical role in this endeavor, serving not only as progress indicators but also as instruments to discern specific capabilities, identify gaps, guide innovation, and inspire research directions.[3] However, traditional approaches often conflate task-specific skill acquisition with genuine intelligence.[7] Measuring skill alone, particularly when influenced heavily by prior knowledge or vast training data, can mask a system's underlying generalization power and adaptability.[8]

A more refined perspective, articulated by François Chollet, defines intelligence not merely by skill but by *skill-acquisition efficiency* over a broad scope of tasks, relative to priors, experience, and generalization difficulty.[8] This definition emphasizes a system's ability to adapt to novel problems unforeseen by its creators.[8] This contrasts sharply with definitions centered on automating economically valuable work, which, while a useful goal, is considered an incorrect measure of intelligence itself.[8]

B. The Abstraction and Reasoning Corpus (ARC) Family

To operationalize this concept of intelligence, Chollet introduced the Abstraction and Reasoning Corpus (ARC) in 2019.[2] ARC, later termed ARC-AGI, is specifically designed as a benchmark to measure *fluid intelligence* – the capacity to reason, solve novel problems, and adapt to new situations – rather than *crystallized intelligence*, which relies on accumulated knowledge and skills.[7] The initial version, ARC-AGI-1 [7], aimed to distinguish skill acquisition from general intelligence.[7]

Recognizing limitations in ARC-AGI-1, such as its binary pass/fail nature, lack of human difficulty calibration, and susceptibility to brute-force solutions [7], the ARC Prize Foundation launched ARC-AGI-2 in March 2025.[4] This second iteration significantly increases the difficulty for AI systems while maintaining relative ease for humans,

explicitly targeting reasoning, adaptability, and, crucially, efficiency.[4] Furthermore, work has commenced on ARC-AGI-3, envisioned for early 2026, which plans to move beyond static tasks towards interactive environments assessing exploration, data gathering, goal setting, and action efficiency.[7]

C. The Emergence of Knowledge-Graph and Symbolic Approaches

The significant challenges posed by ARC-AGI, particularly ARC-AGI-2, have exposed limitations in prevailing AI paradigms, most notably large language models (LLMs) based on transformer architectures.[3] These models, despite their successes in various domains, struggle profoundly with the abstract reasoning, compositional logic, and contextual rule application demanded by ARC tasks.[3] This performance gap has spurred interest in alternative or complementary approaches, including those leveraging structured knowledge representation and explicit reasoning mechanisms characteristic of knowledge graphs (KGs) and symbolic AI.

The hypothesis is that the explicit structure, relational information, and logical manipulation capabilities offered by KGs and symbolic systems might be better suited to handle the types of abstract, compositional reasoning required by ARC.[23] This aligns with the growing field of Neuro-Symbolic AI (NeSy), which seeks to combine the perceptual strengths of neural networks with the reasoning capabilities of symbolic systems.[23] Such hybrid approaches aim to achieve robustness, interpretability, and better generalization from limited data.[36]

D. Report Objectives and Structure

This report provides an expert-level technical analysis of research concerning knowledge-graph-based and related symbolic/graph-based solvers specifically applied to the ARC-AGI-2 benchmark. Drawing exclusively upon the provided research materials, the analysis focuses on the architecture, performance, efficiency, interpretability, and broader implications of these approaches for AI and AGI research. The subsequent sections will delve into the specifics of the ARC-AGI-2 challenge, analyze the performance gap hindering current AI models, examine the architectures and mechanisms of KG and symbolic solvers, evaluate their performance and efficiency, consider interpretability aspects, and finally, discuss the implications and future directions for this line of research.

II. The ARC-AGI-2 Challenge

A. Design Philosophy: "Easy for Humans, Hard for AI"

The fundamental design principle underpinning the ARC-AGI benchmark series is

"Easy for Humans, Hard for AI".[3] This philosophy intentionally selects tasks that humans, leveraging innate cognitive abilities, find relatively straightforward, yet which pose significant difficulties for contemporary AI systems. The rationale behind this approach is to move beyond benchmarks that test specialized, often superhuman skills or vast memorized knowledge (termed "PhD++ problems").[3] Instead, ARC-AGI aims to illuminate the fundamental capability gaps that prevent current AI from achieving human-like general intelligence, particularly in areas of fluid reasoning, adaptability, and generalization from limited experience.[3] The ultimate goal, as articulated by the ARC Prize Foundation, is to measure progress towards AGI by tracking the closure of this gap between tasks easy for humans and hard for AI; when no such tasks remain, AGI could be considered achieved.[3]

B. Task Structure and Format

ARC tasks manifest as visual puzzles presented on grids.[6] Each grid is a rectangular matrix, ranging in size from 1x1 to 30x30 cells.[6] Each cell within the grid can hold one of ten distinct values, typically represented as integers from 0 to 9 and visualized as colors.[6]

A single ARC task comprises two main parts: a set of demonstration pairs and a set of test pairs.[6] The demonstration section typically includes three input-output grid pairs that illustrate a specific, hidden transformation rule or concept.[6] The test section provides one or two input grids for which the solver must generate the corresponding output grid(s).[6] The core challenge for the test-taker (human or AI) is to infer the underlying abstract rule from the few demonstration examples and apply it correctly to the test input(s).[6] Critically, this involves not only filling the output grid cells with the correct colors/values but also predicting the correct dimensions (height and width) of the output grid.[8] Solutions must be exact; any deviation, even a single incorrect cell, results in failure for that attempt.[6]

To ensure a fair comparison between human and artificial intelligence and to focus on core reasoning abilities, ARC tasks are designed to rely only on a minimal set of "Core Knowledge Priors".[8] These priors represent fundamental cognitive concepts believed to be innate or acquired very early in human development, universally shared, and not dependent on specific cultural knowledge or language.[8] The explicitly listed priors include:

- **Objectness**: Understanding that the visual world can be parsed into objects that persist, move cohesively, and interact through contact.[8]
- **Goal-directedness**: Recognizing that some objects can be agents with intentions and goals.[8]

- **Numbers & Counting**: Basic arithmetic concepts like addition, subtraction, comparison, and the ability to count or sort objects.[8]
- **Basic Geometry & Topology**: Concepts of shape, distance, orientation, spatial relationships (like inside/outside), and transformations (like rotation, translation, reflection, scaling, deformation).[8]

By strictly limiting the assumed knowledge to these primitives, ARC forces solvers to demonstrate genuine problem-solving and generalization abilities rather than relying on pre-existing, domain-specific expertise.[8]

C. ARC-AGI-2 Specific Enhancements and Challenges

ARC-AGI-2 represents a deliberate evolution from its predecessor, designed to address ARC-AGI-1's limitations and pose a more significant challenge to AI reasoning systems.[4] While ARC-AGI-1 was criticized for being a binary pass/fail benchmark lacking nuance, potentially susceptible to brute-force methods, and lacking difficulty calibration based on human performance [7], ARC-AGI-2 incorporates several key enhancements.

A major improvement is **difficulty calibration**. The tasks within the ARC-AGI-2 evaluation sets (public, semi-private, and private) have been calibrated using performance data gathered from controlled studies involving over 400 human participants.[4] This ensures that the different evaluation sets are roughly equivalent in difficulty (Independent and Identically Distributed - IID) for both humans and AI, allowing for more reliable comparisons of non-overfit scores across sets.[4] Every task included in the evaluation sets was successfully solved by at least two humans in two attempts or less, matching the rules applied to AI solvers.[4]

Crucially, ARC-AGI-2 tasks are specifically designed to target cognitive abilities where current AI systems, particularly advanced reasoning models, demonstrably struggle.[3] Studies conducted during the benchmark's design identified key areas of weakness, leading to the inclusion of tasks emphasizing:

1. **Symbolic Interpretation**: This requires understanding that symbols (colors/shapes on the grid) possess meaning beyond their mere visual or geometric properties.[3] AI systems might perform surface-level checks like symmetry or transformations but fail to assign semantic significance to the symbols within the task's context.[3]
2. **Compositional Reasoning**: This involves the simultaneous application of multiple rules, or the sequential application of rules where later steps depend on the outcomes of earlier ones, especially when rules interact.[3] Systems often succeed

when only one or a few global rules are present but falter when complex interactions are required.[4]

3. **Contextual Rule Application**: This demands the ability to apply rules differently based on the specific context within the grid, essentially requiring conditional logic or control flow in the reasoning process.[3] AI systems tend to fixate on superficial patterns rather than grasping the underlying principles that determine which rule applies in which situation.[3]

The deliberate inclusion of tasks targeting these specific reasoning bottlenecks makes ARC-AGI-2 not just a measure of general capability but also a diagnostic instrument. By observing where sophisticated AI systems fail on ARC-AGI-2, researchers can pinpoint the precise types of abstract, flexible, and compositional reasoning that need to be developed. The benchmark effectively highlights the shortcomings of approaches relying heavily on pattern matching learned from vast datasets, pushing the field towards architectures capable of deeper, more structured cognitive operations.

D. Evaluation Metrics: Accuracy and Efficiency

Evaluating performance on ARC-AGI-2 involves two key dimensions: accuracy and efficiency. The primary accuracy metric is **Pass@2**.[4] A system is considered to have passed a task if it produces the exactly correct output grid for all test inputs within that task, given a maximum of two attempts per test input.[4] This two-attempt allowance accommodates tasks with potential inherent ambiguity or helps mitigate unintentional errors in the dataset design, while still penalizing random guessing.[4] The overall score is typically the percentage of tasks passed within the evaluation set.[58] Some teams may also track pixel-level correctness as an auxiliary metric, although this is not used for official scoring.[29]

A defining feature of ARC-AGI-2, distinguishing it significantly from ARC-AGI-1, is the explicit emphasis on **efficiency**.[3] Performance is often reported alongside a cost-per-task metric (e.g., in USD) or evaluated under strict computational constraints (e.g., runtime limits on specific hardware).[5] The ARC Prize 2025 competition, for instance, imposes a 12-hour runtime limit on specified CPU/GPU configurations (L4x4 GPUs mentioned) with no internet access allowed.[10]

This focus on efficiency is not merely a practical consideration; it is deeply intertwined with the benchmark's underlying philosophy of intelligence. By measuring efficiency alongside capability, ARC-AGI-2 directly operationalizes Chollet's definition of intelligence as *efficient* skill acquisition.[3] It actively discourages solutions that rely on computationally exorbitant brute-force search or massive scaling, which might

demonstrate skill but not necessarily intelligence in this framework.[3] The goal is to incentivize the development of solutions that mirror human cognitive efficiency – the ability to find solutions with minimal resources.[3] This dual focus makes ARC-AGI-2 a fundamentally more demanding benchmark than its predecessor, evaluating not just *what* solution is found, but *how* it is found.

III. The Performance Gap: Why ARC-AGI-2 is Hard for Current AI

A. Extremely Low Scores for State-of-the-Art Models

The difficulty of ARC-AGI-2 for current artificial intelligence systems is starkly illustrated by their performance scores. While designed to be solvable by humans – with average human performance around 60% and curated human panels achieving 100% solvability within the two-attempt rule [1] – even the most advanced AI models demonstrate remarkably poor results.

Pure LLMs, those without specialized reasoning engines or search mechanisms integrated, consistently score 0% on ARC-AGI-2 tasks.[4] This indicates that the pattern recognition and sequence completion abilities honed by training on vast text corpora are fundamentally insufficient for the abstract, compositional reasoning required. Even sophisticated models incorporating reasoning techniques, such as OpenAI's o-series (e.g., o3, o4-mini), Google's Gemini 2.0 Flash, Anthropic's Claude 3.7, and DeepSeek-R1, achieve only single-digit percentage scores, typically below 4%.[1] For instance, leaderboard data shows o3 (medium) at 3.0%, o3-mini (high) also at 3.0%, ARChitects (the 2024 ARC Prize winner, likely a hybrid system) at 2.5%, and Gemini 2.0 Flash at 1.3%.[4] This dramatic gap between human and AI performance underscores the benchmark's success in isolating capabilities central to general intelligence that current AI lacks.

B. Limitations of LLMs and Scaling

Several fundamental limitations of current AI models, particularly LLMs, contribute to their poor performance on ARC-AGI-2.

Firstly, **memorization is insufficient**.[1] Each ARC task is designed to be unique and novel.[20] Models cannot rely on recalling patterns or solutions encountered during training, as the evaluation tasks demand generalization to entirely new problem structures. LLMs are often characterized as operating via a "memorize, fetch, apply" paradigm, excelling at retrieving and applying learned patterns but failing when faced with true novelty.[5]

Secondly, LLMs primarily demonstrate **crystallized intelligence** (skill based on

accumulated knowledge) rather than the **fluid intelligence** (adaptability, novel problem-solving) that ARC targets.[1] Their success on many benchmarks reflects their vast training data, not necessarily an innate ability to reason abstractly or adapt efficiently.[8]

Thirdly, LLMs exhibit specific **reasoning deficiencies** directly relevant to ARC-AGI-2's challenges. They struggle with symbolic interpretation (assigning meaning beyond visual form), compositional reasoning (handling interacting rules), and contextual rule application (conditional logic).[3] They may fixate on superficial visual patterns or fail to decompose problems into logical steps.[3]

Fourthly, **visual and spatial reasoning weaknesses** are often cited.[6] While ARC grids can be represented numerically or textually, the underlying logic often involves spatial relationships, object manipulation, and geometric transformations that seem difficult for models primarily trained on linear sequences of text.[20] Some argue that the bottleneck lies in the model's inability to form good initial hypotheses based on visual intuition, leading to an overly large search space for subsequent reasoning steps.[70]

Fifthly, even when LLM-based approaches achieve some success, they are often **computationally inefficient**.[5] Models like o3, which employ Chain-of-Thought (CoT) reasoning combined with search or program synthesis, can incur costs of hundreds or even thousands of dollars per task, drastically failing the efficiency requirements of ARC-AGI-2.[4]

Finally, there is a growing consensus that simply **scaling up current LLM architectures is insufficient** to conquer ARC-AGI-2.[3] The benchmark's resilience suggests that fundamental architectural or algorithmic innovations ("new ideas") are necessary to bridge the gap to human-level performance.[3]

The consistent failure of even multi-trillion parameter models like OpenAI's o3 or Anthropic's Claude 3.7 on ARC-AGI-2 provides compelling empirical support for the argument that the dominant transformer-based scaling paradigm may be encountering fundamental limitations. These models excel on tasks that can be solved by leveraging the vast statistical patterns learned during pre-training, effectively demonstrating impressive crystallized intelligence. However, ARC-AGI-2 demands fluid intelligence: the ability to reason abstractly, compose procedures, understand context, and adapt efficiently to complete novelty. The stark performance difference suggests that these capabilities do not spontaneously emerge from scaling alone and may require qualitatively different architectural approaches, potentially incorporating

more explicit structure or reasoning mechanisms.

C. Role of Test-Time Adaptation and Program Synthesis

Significantly, the approaches that have shown *some* measurable progress beyond baseline LLM performance on ARC benchmarks often incorporate mechanisms for **test-time adaptation (TTA)** or **program synthesis**.[7]

TTA involves fine-tuning or otherwise adapting the model *during the evaluation phase*, using the specific demonstration examples provided within each task.[11] This allows the model to specialize its knowledge or search strategy for the immediate problem at hand, moving beyond reliance on its static pre-trained state. Program synthesis involves generating an explicit program (often in a Domain-Specific Language or DSL) that captures the transformation rule observed in the demonstration pairs; this program is then executed on the test input.[5]

OpenAI's o-series models, for example, are described as using CoT reasoning combined with search and synthesis mechanisms.[3] This suggests that these models perform some form of search over potential reasoning paths or programs at test time, guided by the base LLM.[5] The (relative) success of TTA and program synthesis approaches reinforces the idea that solving ARC requires dynamic, task-specific processing during evaluation, rather than just passive application of pre-learned functions.

IV. Knowledge-Graph and Symbolic Solvers for ARC-AGI-2

A. Rationale: Why Structured Approaches?

The documented struggles of connectionist models like LLMs on ARC-AGI-2 motivate the exploration of approaches incorporating more explicit structure and reasoning, such as those based on knowledge graphs (KGs) and symbolic AI. Several rationales underpin this direction:

1. **Addressing LLM Limitations**: Symbolic and graph-based methods inherently offer mechanisms for explicit, step-by-step reasoning, handling compositionality, representing context sensitivity, and ensuring verifiable transformations – capabilities often lacking or unreliable in end-to-end neural models when applied to ARC tasks.[23]
2. **Handling Core Knowledge Priors**: ARC tasks rely on fundamental concepts like objectness, basic geometry, and topology.[8] Symbolic representations and graph structures provide natural frameworks for explicitly encoding and manipulating these priors, potentially leading to more robust reasoning grounded in these

fundamental concepts.[8]

3. **Facilitating Program Synthesis**: A dominant paradigm for tackling ARC involves program synthesis, where solvers search for or generate programs in a Domain-Specific Language (DSL) that encode the task's transformation logic.[6] DSLs are inherently symbolic. Knowledge graphs could potentially enhance this process by providing structured knowledge to guide the search, define constraints, or represent program components and their relationships.
4. **Improving Interpretability**: Compared to the opaque internal workings of large neural networks, reasoning processes based on symbolic manipulations or traversals of graph structures can offer greater transparency and explainability.[22] This is a key goal of neuro-symbolic AI.[36]

B. Architectural Approaches Mentioned in Research

The provided research materials describe several architectural paradigms relevant to KG and symbolic solving of ARC tasks:

- **Explicit Knowledge Graph Construction**:
  - One research direction explicitly proposes converting ARC task data (input-output grids, demonstrated transformations) into a formal knowledge graph structure.[38] This involves defining DSLs for properties and transformations, structuring the KG, and then using this graph to extract "core knowledge" relevant to the task. An abductive symbolic solver then utilizes this extracted knowledge to synthesize a solution, aiming to limit the search space and provide logically grounded intermediate steps.[38]
  - A Kaggle notebook submission describes a system using a lightweight, in-memory "Agentic KG".[91] This KG acts dynamically during solving, updating usage and success counts for different transformation "agents" in real-time, effectively implementing a form of learning from experience within the constraints of the competition environment. Detailed transformation logs are kept separately for offline analysis.[91] *(Note: Details beyond this description were unavailable [91])*.
  - A more general framework, Reasoning on Graphs (RoG), synergizes LLMs with existing KGs.[90] It uses the LLM to generate potential relation paths (plans) grounded in the KG, retrieves valid reasoning paths based on these plans, and then uses the LLM again to perform reasoning along these retrieved paths. While not specifically designed for ARC, its principles of using KGs to structure and ground LLM reasoning could be adapted.
- **Neuro-Symbolic Architectures**:
  - This paradigm explicitly aims to combine neural networks (for perception,

pattern recognition, intuition) with symbolic systems (for logic, reasoning, structure).[23] This hybridization is often motivated by the desire to capture the strengths of both approaches, potentially mirroring the dual-process theories of human cognition (System 1/System 2).[36]
    - Specific implementations for ARC include:
        - **NSA (Neuro-Symbolic ARC)**: This system uses a transformer model, pre-trained on synthetic data and fine-tuned at test time (TTA), to propose promising transformation primitives or search directions within a symbolic DSL. A combinatorial search engine then uses these proposals to find the actual solution program more efficiently.[23]
        - **DreamCoder Adaptation**: The DreamCoder framework, which learns a library of reusable code components through program synthesis, has been adapted for ARC.[21] This involves designing a suitable DSL (e.g., PeARL - Perceptual Abstraction and Reasoning Language) and using a neural "recognition model" to guide the search for programs within this language.[21]
        - **General Hybrid Proposals**: Other suggestions include using deep learning as a perception front-end to parse inputs into discrete symbolic representations for a reasoning engine, adding symbolic modules to DL models, or using DL models to guide or prune the search space of discrete program synthesis algorithms.[20]
- **Graph-Based Representations (Non-Explicit KG)**:
    - Some approaches use graph structures to represent ARC tasks without necessarily building a formal, queryable knowledge graph.
        - **ARGA (Abstract Reasoning with Graph Abstractions)**: This framework converts ARC grid images into object-centric graph representations.[22] It then employs a search algorithm to find a program within a graph-based DSL that operates on these abstract representations.[22]
        - **Graph Neural Networks (GNNs)**: While specific GNN-based ARC *solvers* are not detailed extensively, GNNs are mentioned as relevant components for graph reasoning.[90] Their ability to learn representations from graph structures could potentially be integrated into larger ARC solving systems, perhaps for feature extraction or guiding symbolic search. One submission mentions integrating GNNs into a modular system for relational understanding.[92]
- **Program Synthesis with Symbolic DSLs**:
    - A significant body of work on ARC, particularly early successes and many current systems, relies heavily on program synthesis using hand-crafted Domain-Specific Languages (DSLs).[6] These DSLs encapsulate symbolic

operations relevant to ARC tasks (e.g., geometric transformations, object manipulations). The core challenge becomes efficiently searching the vast space of possible programs constructible from the DSL primitives.[19]
    - A recent trend involves using LLMs to *guide* this synthesis process.[11] The LLM might generate candidate programs, suggest promising search directions, or even help debug generated code, combining neural pattern matching with symbolic program construction.

The clear trend emerging from these varied approaches is a move towards **hybrid architectures**. Purely connectionist models like LLMs falter due to reasoning and generalization limitations on ARC-AGI-2.[4] Purely symbolic approaches, like early brute-force DSL search, while more interpretable, often struggle with the complexity, novelty, and efficiency demands of the benchmark.[6] The most promising avenues appear to lie in combining these paradigms: using neural networks for their strengths in perception, pattern recognition, or providing heuristic guidance, while leveraging symbolic structures (DSLs, graphs, KGs) for their ability to handle explicit reasoning, compositionality, constraints, and potentially offer better few-shot generalization and interpretability.[20] This convergence reflects the broader interest and potential seen in the field of neuro-symbolic AI.

C. Transformation Agents and Symbolic Operations

The concept of modular "transformation agents" is explicitly mentioned in the description of the Agentic KG solver [91], suggesting components dedicated to specific types of grid manipulations. While specific agent examples like ColorRemapAgent or TilerAgent are not detailed in the available snippets for that particular solver, the underlying principle aligns with the structure of DSL-based approaches.

DSLs, by their nature, define a set of symbolic operations or transformations that can be applied to the ARC grids or their representations. The research materials allude to various categories of such operations:

- **Geometric/Topological Transformations**: Common operations include mirroring (diagonal or axis-aligned), rotation, translation, scaling, deformation, repetition of patterns, and combining shapes.[8] Specific function names like dmirror (diagonal_mirror) appear in discussions of DSL refinement.[84]
- **Object-Based Manipulations**: Many tasks involve identifying discrete objects within the grid and applying rules based on them. Operations might include moving objects, changing their color, counting them, finding the smallest bounding box (subgrid), or applying transformations conditioned on object properties.[8] The Core Knowledge prior of "Objectness" is central here.[8]

- **Grid-Level Operations**: These include resizing the grid, copying the input grid to the output (often as a starting point for modification), resetting the grid (filling with a default color like 0), flood-filling connected areas of the same color, and applying masks (e.g., selecting rectangular regions, borders, checkerboard patterns, or custom bitmap shapes).[6]
- **Color/Value Operations**: Getting the color of a cell (get_color), changing colors based on rules, or applying color mappings are fundamental.[84]
- **Higher-Order/Functional Operations**: Some DSLs incorporate functional programming concepts, like fork (combine_two_function_results) mentioned in [84], allowing for more complex program structures.

These symbolic operations form the building blocks that program synthesis or KG-based reasoning systems use to construct solutions to ARC tasks. The effectiveness of any such system is heavily dependent on the choice and definition of these fundamental primitives within its DSL or knowledge representation.

V. Performance, Efficiency, and Scalability

A. Performance of KG/Symbolic/Graph-Based Solvers

Evaluating the performance of knowledge-graph, symbolic, and graph-based solvers specifically on the ARC-AGI-2 benchmark is challenging due to limited reported data in the provided materials. Most available performance metrics pertain to ARC-AGI-1 or are for general LLMs/hybrid systems on ARC-AGI-2, providing context rather than direct results for KG solvers.

- **Direct KG Solvers**: No specific ARC-AGI-2 accuracy scores are available for the explicitly KG-based solvers described (e.g., the Agentic KG [91] or the symbolic solver using KG-extracted knowledge [38]). A hybrid system incorporating a Graph Neural Network (GNN) was reported to outperform traditional neural models by 15-20% on *difficult ARC tasks*, but the specific ARC version (1 or 2) and task subset were not specified.[92]
- **Neuro-Symbolic Solvers**:
  - The NSA (Neuro-Symbolic ARC) system achieved a score of 75/400 (18.75%) on the ARC-AGI-1 Evaluation Set when using Test-Time Adaptation (TTA).[50] This score significantly surpassed comparison baselines reported in the same study, including the graph-based ARGA method.[50] While this is ARC-AGI-1 data, it demonstrates the potential of neuro-symbolic approaches to outperform simpler symbolic or purely neural methods on ARC tasks. Extrapolating this advantage to the more difficult ARC-AGI-2 remains speculative.

    - Adaptations of the DreamCoder framework using the PeARL DSL reportedly solved three times more tasks than a previous implementation [51], and ensembles combining DreamCoder with LLMs showed improvement over the state-of-the-art at the time.[21] However, absolute ARC-AGI-2 scores are not provided.
- **Graph-Based Solvers (ARGA)**: The ARGA system scored 9/400 (2.25%) on the ARC-AGI-1 Evaluation set, with a variant (ARGAe) reaching 22/400 (5.5%).[50] This indicates some capability but falls short of the performance achieved by the NSA neuro-symbolic system on the same dataset.[50]
- **Program Synthesis (DSL-based)**: Historically, DSL-based program synthesis approaches achieved scores around 20-33% on the ARC-AGI-1 private evaluation set.[6] More recently, LLM-guided program synthesis pushed scores to 42-43% on public/semi-private ARC-AGI-1 leaderboards.[20] The top performers in the ARC Prize 2024 competition, using hybrid induction (program synthesis) and transduction (direct prediction) methods, reached impressive scores of 53.5-55.5% on the ARC-AGI-1 private evaluation set.[6]
- **ARC-AGI-2 Context**: For comparison, the current state-of-the-art scores reported on ARC-AGI-2 are very low: the ARChitects team (2024 winner, likely using a hybrid approach developed for ARC-1) scored 2.5% [4], and OpenAI's o3 (medium) model scored 3.0%.[4] These low scores highlight the significantly increased difficulty of ARC-AGI-2.

B. Efficiency Analysis

Efficiency is a primary evaluation criterion for ARC-AGI-2, reflecting the benchmark's focus on intelligence as efficient skill acquisition.[3] Symbolic and KG-based approaches are often motivated by the potential for greater efficiency compared to computationally intensive deep learning or exhaustive search methods.[37]

- The Agentic KG approach was explicitly designed with efficiency in mind, using an in-memory graph to stay within competition memory limits.[91]
- Neuro-symbolic systems like NSA aim to improve efficiency by using the neural component to prune the vast search space faced by the symbolic combinatorial search.[45]
- The ARGA graph-based method was noted for its efficiency on certain types of tasks.[22]

However, symbolic reasoning itself, especially complex search or inference over large knowledge structures, can be computationally demanding.[19] The ARC-AGI-2 leaderboard provides crucial context on the current performance-efficiency

landscape:

Table 1: Comparative Performance and Efficiency on ARC-AGI-2 (Selected Systems)

AI System	System Type (Primary Approach)	ARC-AGI-2 Score (%)	Cost/Task ($)	Notes/Limitations
o3 (medium)	CoT + Synthesis (LLM-based)	3.0%	$2.53	High capability, moderate cost
o3-mini (high)	CoT (LLM-based)	3.0%	$0.55	Similar capability to o3 medium, lower cost
ARChitects (2024)	Custom Hybrid (Likely NeSy/PS)	2.5%	$0.20	2024 Winner (ARC-1), efficient for its score
o4-mini (Medium)	CoT (LLM-based)	2.4%	$0.23	Good efficiency, slightly lower score
DeepSeek R1	CoT (LLM-based)	1.3%	$0.08	Low score, very efficient
Gemini 2.0 Flash	Base LLM	1.3%	$0.004	Low score, extremely efficient
Human Panel	**Human**	**100.0%**	**~$17.00**	**Benchmark reference (cost based on studies)**
Avg. Human	**Human**	**~60.0%**	**~$17.00**	**Benchmark reference (cost based on studies)**

Data sourced from.[4] System types are inferred based on descriptions in snippets. Costs are approximate and may vary.

This data reveals a significant trade-off. The most cost-effective systems currently achieving any score on ARC-AGI-2 are base LLMs or simpler CoT models (Gemini Flash, DeepSeek R1), but their accuracy is minimal (1.3%). Achieving the current state-of-the-art scores (2.5-3.0%) requires more complex and costly systems like o3 or the ARChitects' hybrid approach. Notably, the ARChitects' solution demonstrates significantly better cost-efficiency ($0.20/task) compared to o3 medium ($2.53/task) for a slightly lower score, suggesting hybrid approaches might offer a better balance. However, even these costs are substantial compared to the extremely low cost of Gemini Flash. This highlights a persistent challenge: improving reasoning capability on ARC-AGI-2 currently comes at a significant efficiency cost. Future KG and symbolic solvers must aim to improve accuracy substantially while maintaining or exceeding the efficiency levels demonstrated by systems like ARChitects, thereby breaking the current capability-efficiency trade-off. Simply being symbolic does not automatically guarantee efficiency at competitive performance levels.

C. Scalability Challenges

Scaling KG, symbolic, and hybrid approaches to effectively tackle the full range of ARC-AGI-2 tasks presents several challenges:

- **Combinatorial Explosion**: The primary hurdle for methods involving search (e.g., program synthesis, graph traversal) is the potentially enormous search space. As task complexity increases, the number of possible programs or reasoning paths can grow exponentially, making exhaustive or even guided search computationally infeasible within practical time limits.[19]
- **DSL Limitations**: Program synthesis approaches are fundamentally limited by the expressiveness and completeness of their underlying DSL.[79] Crafting a DSL that is general enough to cover the diverse logic of ARC tasks, yet constrained enough to allow efficient search, is a difficult design problem. The DSL must capture the necessary core knowledge primitives and allow for their flexible composition.
- **Neuro-Symbolic Integration Complexity**: Effectively merging neural and symbolic components is non-trivial.[48] Ensuring that the neural component provides useful guidance without overriding correct symbolic reasoning, or that the symbolic component can effectively utilize potentially noisy neural outputs, requires careful architectural design and training strategies. Furthermore, issues like "reasoning shortcuts" can arise, where the hybrid system finds a solution using incorrect or unintended internal logic, compromising reliability.[39] Hardware

acceleration for the symbolic parts of these systems is also less developed compared to neural network accelerators.[48]

- **Knowledge Representation**: For KG-based approaches, designing an appropriate ontology or schema to represent ARC tasks, their components (objects, properties, transformations), and the underlying core knowledge is crucial but challenging given the abstract and varied nature of the tasks.[38]

VI. Interpretability Considerations

A. Potential Advantages of Symbolic/Graph Structures

A frequently cited motivation for exploring symbolic, graph-based, and neuro-symbolic approaches is their potential for enhanced interpretability compared to end-to-end deep learning models.[36]

- **Traceable Reasoning**: Systems that generate explicit symbolic programs (via DSLs) or follow reasoning paths on a knowledge graph can theoretically offer a step-by-step trace of how a solution was derived.[22] This contrasts with the "black box" nature of deep neural networks, where understanding the internal decision-making process is often difficult. The programs generated by the ARGA system, for example, were noted as being "easy to understand".[22] The KG approach proposed in [38] explicitly aims to ground solutions in extracted core knowledge, enhancing logical transparency.[38]
- **Explicit Structure**: Knowledge graphs and other graph representations make relationships between entities and concepts explicit, which can aid human understanding of the problem structure and the solver's approach.[22]
- **Neuro-Symbolic Goals**: Improving explainability is often a core objective driving neuro-symbolic research, alongside performance and generalization.[36]

B. Challenges and Nuances

Despite the theoretical advantages, achieving meaningful interpretability with these approaches, especially on complex tasks like those in ARC-AGI-2, faces significant hurdles:

- **System Complexity**: As symbolic systems (large DSLs, intricate KGs) or hybrid models grow in complexity to handle diverse tasks, their reasoning traces can become convoluted and difficult for humans to follow and verify.
- **Opacity in Hybrids**: In neuro-symbolic systems, the neural components responsible for perception or guiding the search remain largely opaque.[39] This limits the overall transparency of the system, as crucial parts of the process are not easily interpretable. The interaction points between neural and symbolic

modules are critical and can be sources of unexpected behavior.

- **Reasoning Shortcuts**: A critical issue identified in neuro-symbolic research is the phenomenon of "reasoning shortcuts".[39] Models might learn to produce the correct final output while relying on incorrect intermediate reasoning steps or misinterpreting symbolic concepts (e.g., confusing pedestrians with red lights because both imply "stop" in a traffic scenario [39]). This occurs when the provided knowledge or task structure allows for spurious correlations to lead to the right answer. Such shortcuts undermine the trustworthiness and the perceived interpretability of the system, as the apparent reasoning path does not reflect the true mechanism. Specialized benchmarks and evaluation techniques (like rsbench [39]) are being developed specifically to detect and mitigate these issues.

Therefore, while symbolic and graph structures offer a more transparent *framework* compared to purely neural networks, this structural transparency does not automatically guarantee trustworthy or meaningful interpretability for complex reasoning tasks like ARC-AGI-2. The complexity of the tasks, the potential opacity of neural components in hybrid systems, and the risk of reasoning shortcuts necessitate dedicated methods to validate the *faithfulness* and *semantic correctness* of the reasoning process itself. Simply observing a symbolic trace is insufficient; verifying that the trace accurately reflects a sound and intended logical progression remains an active and important research challenge.

VII. Implications and Future Research Directions

A. ARC-AGI as a Research Driver

The ARC-AGI benchmark series, particularly ARC-AGI-2, serves as a significant catalyst and directional guide for AGI research.[3] It acts as a "North Star," pushing the community to focus on fundamental aspects of intelligence like fluid reasoning, few-shot learning, generalization to novelty, and efficient adaptation during testing.[3] The ARC Prize competitions, with substantial funding and an emphasis on open-source contributions, further amplify this effect, incentivizing researchers and labs to tackle the benchmark and share novel solutions.[3]

While success on ARC is viewed as a potential milestone towards AGI, potentially enabling new programming paradigms like programming-by-example [8], it is explicitly positioned by its creators not as a definitive test for AGI, but rather as a valuable research tool designed to focus attention on key unsolved problems in AI reasoning and adaptation.[5]

B. Significance of KG/Symbolic Approaches for AGI Research

Within this context, research into knowledge graph, symbolic, and neuro-symbolic solvers for ARC holds particular significance. These approaches directly confront the reasoning, generalization, and adaptability limitations observed in current dominant AI paradigms like scaled LLMs.[23] By forcing the development of systems that can handle explicit structure, compositional rules, and context-dependent logic, this research pushes towards more robust, flexible, and potentially more human-like AI reasoning.

Furthermore, developing effective architectures that integrate structured knowledge (KGs) or symbolic reasoning (DSLs, logic) with perceptual capabilities (neural networks) for ARC could yield blueprints applicable to a wide range of AI challenges beyond the benchmark itself.[23] Success in this area could inform the design of more trustworthy, explainable, and data-efficient AI systems for complex real-world applications requiring reasoning and adaptability.

C. Limitations and Open Questions

Despite their potential, significant challenges and open questions remain for KG and symbolic approaches in the context of ARC-AGI-2:

- **Scalability and Efficiency**: As previously discussed, ensuring these methods are computationally tractable and meet the stringent efficiency requirements of ARC-AGI-2 is a primary obstacle.[19]
- **Knowledge Representation**: Designing KGs, ontologies, or DSLs that are sufficiently expressive to capture the diverse and abstract logic of ARC tasks, yet structured enough to support efficient reasoning, remains a difficult problem.[38] How can these representations be learned automatically or adapted dynamically rather than relying solely on manual crafting?
- **Neuro-Symbolic Integration**: Finding the optimal ways to combine neural and symbolic components – leveraging the strengths of each without introducing new failure modes or bottlenecks – is a core research challenge.[20]
- **Benchmark Validity and Scope**: Questions remain about whether ARC, even ARC-AGI-2, fully captures all necessary aspects of general intelligence.[54] While designed to test fluid reasoning, it may not adequately cover other cognitive dimensions. ARC-AGI-1 had known flaws [7], and future research may uncover limitations in ARC-AGI-2 as AI capabilities advance.

D. Future Research Directions

The challenges posed by ARC-AGI-2 point towards several key future research directions for KG, symbolic, and neuro-symbolic approaches:

- **Advanced Neuro-Symbolic Architectures**: Continued development of novel

NeSy architectures specifically designed for the types of abstract, compositional, and contextual reasoning demanded by ARC.[41] This includes exploring different ways to integrate perception, memory, attention, and symbolic manipulation.
- **Automated Knowledge Representation**: Research into methods for automatically constructing or learning relevant knowledge graphs, ontologies, or DSL primitives directly from ARC tasks or related data, reducing reliance on manual engineering.[38]
- **KG-Guided Program Synthesis**: Exploring how explicit knowledge graphs can be used to constrain, guide, or verify the process of program synthesis, potentially making it more efficient and reliable than purely neural-guided or brute-force approaches.[20]
- **Meta-Learning and Adaptability**: Investigating meta-learning techniques to enable faster adaptation and generalization from the few examples provided in each ARC task, potentially by learning reusable reasoning strategies or adapting symbolic representations on the fly.[23]
- **Preparing for ARC-AGI-3**: The planned shift towards interactive environments in ARC-AGI-3 necessitates research into agents capable of active exploration, planning, goal management, and efficient action selection in dynamic settings.[7] KG and symbolic methods, with their strengths in planning, state representation, and logical inference, appear particularly relevant for this next stage. Collaboration and input from the research community are actively solicited for the design of ARC-AGI-3.[7]

The deliberate progression of the ARC benchmark series – from ARC-AGI-1's focus on basic fluid intelligence, to ARC-AGI-2's emphasis on complex reasoning and efficiency, and towards ARC-AGI-3's focus on interactive agency – maps out a clear trajectory designed by the benchmark creators. This trajectory pushes AI research systematically away from static pattern recognition towards more dynamic, adaptive, efficient, and ultimately, more agentic forms of intelligence. Knowledge graphs and symbolic methods, which excel at representing structure, planning sequences of actions, and performing logical inference, seem naturally aligned with the increasing demands of this evolving benchmark landscape. Their relevance and potential importance are therefore likely to grow as the field progresses towards tackling ARC-AGI-3 and the broader challenges of AGI.

VIII. Conclusion

A. Summary of Findings

The ARC-AGI-2 benchmark stands as a formidable challenge to current AI systems,

designed explicitly to measure fluid intelligence, abstract reasoning, and efficient adaptation to novelty – areas where prevailing models like LLMs demonstrate significant weaknesses, often scoring in the low single digits or zero. This performance gap underscores the limitations of current scaling paradigms and motivates the exploration of alternative approaches, particularly those incorporating structured knowledge and explicit reasoning mechanisms found in knowledge graphs (KGs), symbolic AI, and neuro-symbolic hybrids. Research in this area explores various architectures, including explicit KG construction for task representation and knowledge extraction, neuro-symbolic systems combining neural perception/guidance with symbolic search/reasoning (e.g., NSA, DreamCoder adaptations), graph-based representations focusing on object-centric abstractions (e.g., ARGA), and sophisticated program synthesis techniques often guided by neural models or operating within symbolic DSLs.

B. Current Status and Potential

Based on the reviewed materials, dedicated KG-based solvers for ARC-AGI-2 are still in the proposal or early development stages, with limited specific performance data available for this benchmark version. However, related symbolic and neuro-symbolic approaches have shown promise on ARC-AGI-1, often outperforming baseline methods, suggesting potential viability. The trend clearly favors hybrid systems that attempt to leverage the strengths of both neural and symbolic paradigms. While these approaches hold theoretical advantages in terms of handling ARC's specific reasoning challenges, potential for better generalization from few examples, and enhanced interpretability, they face substantial practical hurdles. Key challenges include managing the computational complexity and ensuring the efficiency required by ARC-AGI-2, designing adequate and scalable knowledge representations (KGs or DSLs), effectively integrating neural and symbolic components, and ensuring the trustworthiness and semantic correctness of the reasoning process beyond mere structural transparency.

C. Broader Significance

The research focused on tackling ARC-AGI-2, including efforts involving KGs and symbolic methods, is highly significant for the broader field of AI. By pushing beyond the limitations of current models, this work directly addresses core challenges related to achieving more robust, adaptable, and general intelligence. ARC-AGI-2 and the associated ARC Prize initiative effectively steer research towards novel architectures and algorithms capable of abstraction, compositionality, contextual reasoning, and efficiency – key ingredients potentially needed for AGI. Knowledge graph and symbolic approaches represent a critical research thrust in this direction, offering

pathways to integrate structured knowledge and explicit reasoning into AI systems. While significant obstacles remain, continued progress in developing and evaluating these structured approaches against demanding benchmarks like ARC-AGI-2 will be crucial in understanding the architectural requirements for future intelligent systems and potentially accelerating the journey towards artificial general intelligence. The emphasis on new ideas and efficient, open-source solutions remains paramount.

Works cited

1. We're getting close now....ARC-AGI v2 is getting solved at rapid pace, high score already at 12.4% (humans score 60%, o3 (medium) scores 3%) : r/singularity - Reddit, accessed May 4, 2025, https://www.reddit.com/r/singularity/comments/1k902kd/were_getting_close_now_arcagi_v2_is_getting_solved/
2. Is AGI Here? A Deep Dive into OpenAI's o3 Model and ARC-AGI Benchmarks | Dan Sasser, accessed May 4, 2025, https://dansasser.me/posts/is-agi-here-a-deep-dive-into-openais-o-3-model-and-arc-agi-benchmarks/
3. ARC Prize launches its toughest AI benchmark yet: ARC-AGI-2 - AI News, accessed May 4, 2025, https://www.artificialintelligence-news.com/news/arc-prize-launches-toughest-ai-benchmark-yet-arc-agi-2/
4. Announcing ARC-AGI-2 and ARC Prize 2025, accessed May 4, 2025, https://arcprize.org/blog/announcing-arc-agi-2-and-arc-prize-2025
5. OpenAI o3 Breakthrough High Score on ARC-AGI-Pub, accessed May 4, 2025, https://arcprize.org/blog/oai-o3-pub-breakthrough
6. Solving ARC-AGI Challenge With AI Agents - The Hudson Weekly, accessed May 4, 2025, https://hudsonweekly.com/solving-arc-agi-challenge-with-ai-agents/
7. ARC-AGI-2 Overview (in-depth presentation) : r/singularity - Reddit, accessed May 4, 2025, https://www.reddit.com/r/singularity/comments/1jwb0s5/arcagi2_overview_indepth_presentation/
8. What is ARC-AGI? - ARC Prize, accessed May 4, 2025, https://arcprize.org/arc
9. What is ARC-AGI? - ARC Prize, accessed May 4, 2025, https://arcprize.org/arc-agi
10. ARC-AGI 2025: A research review - lewish.io, accessed May 4, 2025, https://lewish.io/posts/arc-agi-2025-research-review
11. ARC Prize 2024: Technical Report, accessed May 4, 2025, https://arcprize.org/media/arc-prize-2024-technical-report.pdf
12. arcprize/ARC-AGI-2 - GitHub, accessed May 4, 2025, https://github.com/arcprize/ARC-AGI-2
13. What is the ARC AGI Benchmark and its significance in evaluating LLM capabilities in 2025, accessed May 4, 2025, https://www.adaline.ai/blog/what-is-the-arc-agi-benchmark-and-its-significance-in-evaluating-llm-capabilities-in-2025

14. About ARC - Lab42, accessed May 4, 2025, https://lab42.global/arc/
15. Using LLMs to Solve the ARC-AGI Challenge | SMU Guildhall, accessed May 4, 2025, https://www.smu.edu/guildhall/academics/research/using-llms-to-solve-the-arc-agi-challenge
16. The Abstraction and Reasoning Challenge (ARC), accessed May 4, 2025, https://pgpbpadilla.github.io/chollet-arc-challenge
17. Abstraction and Reasoning Challenge - Kaggle, accessed May 4, 2025, https://www.kaggle.com/competitions/abstraction-and-reasoning-challenge
18. [D] François Chollet Announces New ARC Prize Challenge – Is It the Ultimate Test for AI Generalization? : r/MachineLearning - Reddit, accessed May 4, 2025, https://www.reddit.com/r/MachineLearning/comments/1de2b16/d_fran%C3%A7ois_chollet_announces_new_arc_prize/
19. Understanding and Benchmarking Artificial Intelligence: OpenAI's o3 Is Not AGI - arXiv, accessed May 4, 2025, https://arxiv.org/pdf/2501.07458
20. How to Beat ARC-AGI by Combining Deep Learning and Program Synthesis, accessed May 4, 2025, https://arcprize.org/blog/beat-arc-agi-deep-learning-and-program-synthesis
21. [2402.03507] Neural networks for abstraction and reasoning: Towards broad generalization in machines - arXiv, accessed May 4, 2025, https://arxiv.org/abs/2402.03507
22. Graphs, Constraints, and Search for the Abstraction and Reasoning Corpus - arXiv, accessed May 4, 2025, https://arxiv.org/abs/2210.09880
23. ARC Prize 2025 | Kaggle, accessed May 4, 2025, https://www.kaggle.com/competitions/arc-prize-2025/discussion/573220
24. #84: Could Program Synthesis Unlock AGI? - Hugging Face, accessed May 4, 2025, https://huggingface.co/blog/Kseniase/fod84
25. #84: Could Program Synthesis Unlock AGI? - Turing Post, accessed May 4, 2025, https://www.turingpost.com/p/fod84
26. fchollet/ARC-AGI: The Abstraction and Reasoning Corpus - GitHub, accessed May 4, 2025, https://github.com/fchollet/ARC-AGI
27. ARC Prize, accessed May 4, 2025, https://arcprize.org/
28. Is Your AI Smart Enough? Test It with ARC AGI v2! - Labellerr, accessed May 4, 2025, https://www.labellerr.com/blog/arc-agi-v2/
29. Official Guide - ARC Prize, accessed May 4, 2025, https://arcprize.org/guide
30. ARC-AGI-2 Overview With Francois Chollet - YouTube, accessed May 4, 2025, https://www.youtube.com/watch?v=TWHezX43I-4
31. ARC Prize Foundation - a North Star for AGI, accessed May 4, 2025, https://arcprize.org/blog/arc-prize-2025
32. François Chollet on why LLMs won't scale to AGI - Effective Altruism Forum, accessed May 4, 2025, https://forum.effectivealtruism.org/posts/MGpJpN3mELxwyfv8t/francois-chollet-on-why-llms-won-t-scale-to-agi
33. The ARC AGI 2 Benchmark is destroying LLMs - YouTube, accessed May 4, 2025, https://m.youtube.com/shorts/fTe3Z7oYKTU

34. LLMs Hit a New Low on ARC-AGI-2 Benchmark, Pure LLMs Score 0%, accessed May 4, 2025, https://analyticsindiamag.com/ai-news-updates/llms-hit-a-new-low-on-arc-agi-2-benchmark-pure-llms-score-0/
35. LLMs Can't Reason - The Reversal Curse, The Alice In Wonderland Test, And The ARC - AGI Challenge - CustomGPT.ai, accessed May 4, 2025, https://customgpt.ai/llm-reasoning-vs-memorization/
36. ARC, Neuro-Symbolic AI, Intermediate Language | Road to AGI | Recap 01, accessed May 4, 2025, https://dev.to/nucleoid/roadtoagi-recap-01-arc-neuro-symbolic-ai-intermediate-language-40cd
37. Is the $0.42 per task budget realistic for developers to reach 85% accuracy on the ARC-AGI-2 test in the Arc Prize 2025 contest? - Quora, accessed May 4, 2025, https://www.quora.com/Is-the-0-42-per-task-budget-realistic-for-developers-to-reach-85-accuracy-on-the-ARC-AGI-2-test-in-the-Arc-Prize-2025-contest
38. Abductive Symbolic Solver on Abstraction and Reasoning Corpus - arXiv, accessed May 4, 2025, https://arxiv.org/html/2411.18158v1
39. A Neuro-Symbolic Benchmark Suite for Concept Quality and Reasoning Shortcuts, accessed May 4, 2025, https://proceedings.neurips.cc/paper_files/paper/2024/file/d1d11bf8299334d354949ba8738e8301-Paper-Datasets_and_Benchmarks_Track.pdf
40. [2501.05435] Neuro-Symbolic AI in 2024: A Systematic Review - arXiv, accessed May 4, 2025, https://arxiv.org/abs/2501.05435
41. System 2 Reasoning via Generality and Adaptation - OpenReview, accessed May 4, 2025, https://openreview.net/attachment?id=F6EHMaugaq&name=pdf
42. System-2 Reasoning via Generality and Adaptation - OpenReview, accessed May 4, 2025, https://openreview.net/pdf?id=brqBUZpj3K
43. Neural networks for abstraction and reasoning - PMC - PubMed Central, accessed May 4, 2025, https://pmc.ncbi.nlm.nih.gov/articles/PMC11561310/
44. How to Think About Benchmarking Neurosymbolic AI? - CEUR-WS.org, accessed May 4, 2025, https://ceur-ws.org/Vol-3432/paper22.pdf
45. [2501.04424] NSA: Neuro-symbolic ARC Challenge - arXiv, accessed May 4, 2025, https://arxiv.org/abs/2501.04424
46. LAMDASZ-ML/Awesome-Neuro-Symbolic-Learning-with-LLM - GitHub, accessed May 4, 2025, https://github.com/LAMDASZ-ML/Awesome-Neuro-Symbolic-Learning-with-LLM
47. A Neuro-Symbolic Benchmark Suite for Concept Quality and Reasoning Shortcuts - arXiv, accessed May 4, 2025, https://arxiv.org/html/2406.10368v2
48. Towards Efficient Neuro-Symbolic AI: From Workload Characterization to Hardware Architecture - ResearchGate, accessed May 4, 2025, https://www.researchgate.net/publication/384245563_Towards_Efficient_Neuro-Symbolic_AI_From_Workload_Characterization_to_Hardware_Architecture
49. How to Think About Benchmarking Neurosymbolic AI?, accessed May 4, 2025, https://www.cs.ox.ac.uk/isg/conferences/tmp-proceedings/NeSy2023/paper22.pd

f
50. NSA: Neuro-symbolic ARC Challenge - arXiv, accessed May 4, 2025, https://arxiv.org/html/2501.04424v1
51. Neural networks for abstraction and reasoning: Towards broad generalization in machines, accessed May 4, 2025, https://arxiv.org/html/2402.03507v1
52. The KANDY Benchmark: Incremental Neuro-Symbolic Learning and Reasoning with Kandinsky Patterns - arXiv, accessed May 4, 2025, https://arxiv.org/html/2402.17431v1
53. arc-prize-2024-solution-by-the-architects - Kaggle, accessed May 4, 2025, https://www.kaggle.com/code/dfranzen/arc-prize-2024-solution-by-the-architects/input
54. Exploring ARC-AGI: The Test That Measures True AI Adaptability - Unite.AI, accessed May 4, 2025, https://www.unite.ai/exploring-arc-agi-the-test-that-measures-true-ai-adaptability/
55. ARC-AGI Without Pretraining | iliao2345 - Isaac Liao, accessed May 4, 2025, https://iliao2345.github.io/blog_posts/arc_agi_without_pretraining/arc_agi_without_pretraining.html
56. ARC Prize 2024: Technical Report - arXiv, accessed May 4, 2025, https://arxiv.org/html/2412.04604v1
57. ARC Prize 2024: Technical Report - arXiv, accessed May 4, 2025, https://arxiv.org/html/2412.04604v2
58. ARC Prize 2025 | Kaggle, accessed May 4, 2025, https://www.kaggle.com/competitions/arc-prize-2025
59. Leaderboard - ARC Prize, accessed May 4, 2025, https://arcprize.org/leaderboard
60. Unveiling OpenAI o3: From benchmarks to real world | Our Insights | Plante Moran, accessed May 4, 2025, https://www.plantemoran.com/explore-our-thinking/insight/2025/01/unveiling-openai-o3-from-benchmarks-to-real-world
61. OpenAI o3 Released: Benchmarks and Comparison to o1 - Helicone, accessed May 4, 2025, https://www.helicone.ai/blog/openai-o3
62. Analyzing o3 and o4-mini with ARC-AGI, accessed May 4, 2025, https://arcprize.org/blog/analyzing-o3-with-arc-agi
63. OpenAI's O3: Features, O1 Comparison, Benchmarks & More | DataCamp, accessed May 4, 2025, https://www.datacamp.com/blog/o3-openai
64. 2025 Competition Details - ARC Prize, accessed May 4, 2025, https://arcprize.org/competition
65. ARC-AGI-2 Leaderboard : r/singularity - Reddit, accessed May 4, 2025, https://www.reddit.com/r/singularity/comments/1jj2kez/arcagi2_leaderboard/
66. OpenAI o3 performance on ARC-AGI - Reddit, accessed May 4, 2025, https://www.reddit.com/r/OpenAI/comments/1hiptxb/openai_o3_performance_on_arcagi/
67. Why ARC-AGI is not Proof that Models are incapable of Reasoning : r/OpenAI - Reddit, accessed May 4, 2025, https://www.reddit.com/r/OpenAI/comments/1g8a1pw/why_arcagi_is_not_proof_t

hat_models_are_incapable/
68. ARC-AGI is a genuine AGI test but o3 cheated - LessWrong, accessed May 4, 2025, https://www.lesswrong.com/posts/KHCyituifsHFbZoAC/arc-agi-is-a-genuine-agi-test-but-o3-cheated
69. o3 scores <5% on ARC-AGI-2 (but the test looks ... harder?) - Reddit, accessed May 4, 2025, https://www.reddit.com/r/singularity/comments/1jj2vbz/o3_scores_5_on_arcagi2_but_the_test_looks_harder/
70. The real bottleneck of ARC-AGI : r/ArtificialInteligence - Reddit, accessed May 4, 2025, https://www.reddit.com/r/ArtificialInteligence/comments/1jluzxw/the_real_bottleneck_of_arcagi/
71. Arc Prize 2024 Winners and Technical Report | Hacker News, accessed May 4, 2025, https://news.ycombinator.com/item?id=42343215
72. ARC Prize 2024 Winners & Technical Report Published, accessed May 4, 2025, https://arcprize.org/blog/arc-prize-2024-winners-technical-report
73. Test-Time Training for Transductive Transformer Models in ARC-AGI Challenge - OpenReview, accessed May 4, 2025, https://openreview.net/pdf?id=TtGONY7UKy
74. The LLM ARChitect: Solving ARC-AGI Is A Matter of Perspective - Dr. Amit Puri, accessed May 4, 2025, https://cached.amitpuri.com/assets/pdf/The%20LLM%20ARChitect%20Solving%20ARC-AGI%20Is%20A%20Matter%20of%20Perspective.pdf
75. [2412.04604] ARC Prize 2024: Technical Report - arXiv, accessed May 4, 2025, https://arxiv.org/abs/2412.04604
76. ARC Prize 2024, accessed May 4, 2025, https://arcprize.org/2024-results
77. ARC Prize 2024: Technical Report - ResearchGate, accessed May 4, 2025, https://www.researchgate.net/publication/386555197_ARC_Prize_2024_Technical_Report
78. Reflection System for the Abstraction and Reasoning Corpus - OpenReview, accessed May 4, 2025, https://openreview.net/forum?id=GU4CjUNNb5
79. arXiv:2411.08706v1 [cs.LG] 13 Nov 2024 - OpenReview, accessed May 4, 2025, https://openreview.net/pdf/15a431dc87a7907669635a54e0f00465dd0809a5.pdf
80. Towards Efficient Neurally-Guided Program Induction for ARC-AGI - arXiv, accessed May 4, 2025, https://arxiv.org/html/2411.17708v1
81. ConceptSearch: Towards Efficient Program Search Using LLMs for Abstraction and Reasoning Corpus (ARC) - arXiv, accessed May 4, 2025, https://arxiv.org/html/2412.07322v2
82. arXiv:2411.17708v1 [cs.AI] 13 Nov 2024, accessed May 4, 2025, https://arxiv.org/pdf/2411.17708?
83. System-2 Reasoning via Generality and Adaptation - arXiv, accessed May 4, 2025, https://arxiv.org/html/2410.07866v1
84. Capturing Sparks of Abstraction for the ARC Challenge - arXiv, accessed May 4, 2025, https://arxiv.org/html/2411.11206v1
85. Understanding and Benchmarking Artificial Intelligence: OpenAI's o3 Is Not AGI -

arXiv, accessed May 4, 2025, https://arxiv.org/html/2501.07458v1
86. ConceptSearch: Towards Efficient Program Search Using LLMs for Abstraction and Reasoning Corpus (ARC), accessed May 4, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/34259/36414
87. Solving ARC-AGI Challenge with AI Agents - Cases, accessed May 4, 2025, https://cases.media/article/solving-arc-agi-challenge-with-ai-agents
88. Generalized Planning for the Abstraction and Reasoning Corpus | Proceedings of the AAAI Conference on Artificial Intelligence, accessed May 4, 2025, https://ojs.aaai.org/index.php/AAAI/article/view/29996
89. OpenAI's Latest Model Shows AGI Is Inevitable. Now What? - Lawfare, accessed May 4, 2025, https://www.lawfaremedia.org/article/openai's-latest-model-shows-agi-is-inevitable.-now-what
90. Reasoning on Graphs: Faithful and Interpretable Large Language Model Reasoning - arXiv, accessed May 4, 2025, https://arxiv.org/html/2310.01061v2
91. safeAI ARC 2 2025 - Kaggle, accessed May 4, 2025, https://www.kaggle.com/code/bliztafree/safeai-arc-2-2025/notebook
92. ARC Prize 2025 Paper Submission | PDF | Artificial Intelligence - Scribd, accessed May 4, 2025, https://www.scribd.com/document/853475966/ARC-Prize-2025-Paper-Submission
93. [2407.05816] Graph Reasoning Networks - arXiv, accessed May 4, 2025, http://arxiv.org/abs/2407.05816
94. ARC: A Generalist Graph Anomaly Detector with In-Context Learning - arXiv, accessed May 4, 2025, https://arxiv.org/html/2405.16771v1
95. volotat/ARC-Game: The Abstraction and Reasoning Corpus made into a web game - GitHub, accessed May 4, 2025, https://github.com/volotat/ARC-Game
96. Resources - ARC Prize, accessed May 4, 2025, https://arcprize.org/resources