Abstraction and Reasoning Abilities in Artificial Intelligence Applied to Solving the ARC Prize: A Systematic Literature Review

Zakhar Zinkevich
Computer Science Department, Berlin School of Economics and Law, Berlin, Germany

zak.von.zak@gmail.com

Abstract: In recent years, the development of AI-based systems has seen a drastic increase in popularity and investment. To assess and measure specific capabilities of AI-based systems, different benchmarks have been established. AI-driven approaches tend to outperform humans on most of these benchmarks, but no AI-based system was able to surpass average human performance on the Abstraction and Reasoning Corpus (ARC) benchmark. This paper presents an extensive PRISMA-guided literature review that assesses and classifies techniques and technologies utilized by solution approaches for the ARC benchmark. 538 manuscripts are screened, resulting in an inclusion of 65 publications in the final systematic literature review. As a result, a knowledge graph consisting of review protocols of manuscripts is created, that provides further insight into classification of solution approaches. Furthermore, an estimate of possible synergies and ensemble combinations between different approaches is provided by analyzing the task-level performance of solution approaches. The estimation is conducted based on the heat-maps created using the Szymkiewicz-Simpson coefficient and the Gain coefficient.

Keywords: Abstraction and reasoning corpus, ARC, ARC Prize, ARC-AGI, PRISMA, Systematic literature review

1. Introduction

1.1 Rationale (What is ARC?)

The Abstraction and Reasoning Corpus (ARC) was introduced by François Chollet in 2019 (Chollet, 2019) and is a general Artificial Intelligence (AI) benchmark aimed at measuring the abstraction and reasoning abilities of AI-based systems. Its purpose is to enable the comparison of those systems with each other as well as with humans. Chollet (2019) claims in his paper, that training AI-based systems on unlimited training data or providing unlimited priors does not make a system intelligent, but rather enables it to develop some task-specific shortcuts that allow it to excel at an above-human level of performance on a selected set of tasks. In contrast, the goal of the ARC benchmark is to measure the skill-acquisition efficiency on novel tasks, thus implicitly measuring the generalization abilities of an AI-system.

The Abstraction and Reasoning Corpus is a dataset consisting of subsets of problems, which require generalization and abstraction abilities to solve them. They are represented as numerical two-dimensional grids, which may have a different number of cells even inside of a single subset. These grids contain unique symbols representing colours. Each problem in a subset consists of an input grid and an output grid (see Figure 1). A test-taker (human or machine) always has access to training examples (i.e., both input grid and output grid are available) and to the test problem (i.e., only the input grid is provided, and the output grid needs to be constructed by the test-taker). The test-taker has a maximum of three attempts to solve the test problem. The feedback to each of the solution suggestions is binary: correct or incorrect. The goal of the test-taker is to find the generalization and abstraction rules needed to solve the problem. All the problem subsets are said to be different from each other, which makes it difficult to “buy” the solution by providing extensive training data for a specific task.

ARC was introduced in 2019, but still stays unsolved up to this day. The lack of progress on such an important benchmark inspired the establishment of the ARC Prize in 2024 (Chollet et al., 2025). It promised a $1,000,000+ prize for someone, who can achieve the accuracy above 85% on the ARC dataset. Unfortunately, none of the submissions has been able to make it so far. Nevertheless, the analysis of these submissions and the AI approaches empowering them are essential for the further development in this area.

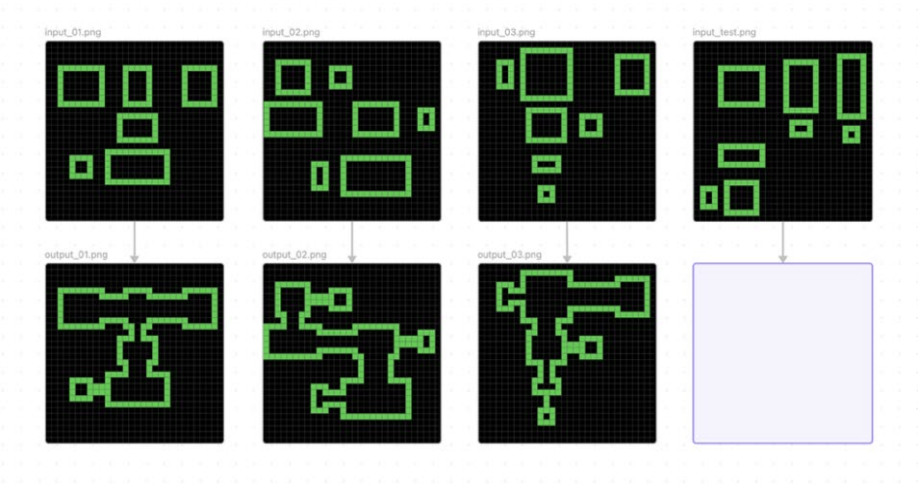

Figure 1: An example of an ARC task. The top row represents input grids and the button row depicts the output grids

1.2 Objectives

The objectives of this systematic literature review have been compiled using the PICO framework that was introduced by Richardson et. al (1995). Using this framework, the systematic review objectives can be posed as follows:

- **Population:** This systematic literature review aims to assess generalization and abstraction abilities in AI-based systems. Therefore, a variety of AI-based approaches applied to solving the ARC Prize 2024 was selected as the population.
- **Intervention:** An intervention of this systematic literature review is the ARC benchmark. All the submissions to the ARC Prize 2024 were tested using this benchmark to measure the generalization and abstraction abilities.
- **Comparator:** One of the goals of the ARC benchmark is to facilitate a fair comparison between humans and AI-based systems. Consequently, human performance has been chosen as the comparator for this systematic review. By comparing the performance of a given AI-based solution approach to the average human performance reported by LeGris et. al (2024), it becomes possible to assess whether the goal of achieving human-level performance has been accomplished.
- **Outcome:** The main goal of this systematic review is to gather information about publicly accessible AI-based solution approaches to the ARC Prize 2024. Furthermore, a classification of the techniques and technologies used by those approaches is performed. Finally, the performance of these approaches is compared at the single-task-level, and theoretical future ensemble combinations are described based on that data.

2. Methodology

2.1 Eligibility Criteria

To examine most of the available solution approaches to the ARC benchmark that fulfil the objectives described in Section 1.2 a specific set of eligibility criteria was considered. This set consists of inclusion as well as exclusion criteria. The following list presents the inclusion criteria for this study:

- A given manuscript must be no older than *January 2024*. The reason why exactly this date range was chosen is that the ARC Prize was introduced in 2024. Furthermore, this criterium is aimed to cut off non-relevant older publications and allow a more precise analysis of the newest solution approaches.
- The language of the publication must be either *English* or *German*. The information in these languages could be analysed by the author directly eliminating the need of translation, that could potentially be biased.
- A title or an abstract of a given manuscript must mention at least one of the following terms: *ARC, ARC-AGI, ARC Prize, ARC Prize 2024, AGI, Abstraction, Reasoning*.
- Importantly, manuscripts *without a peer review* (e.g. preprints from ArXiv) were considered eligible for this systematic literature review (see Section 2.5 for the bias assessment).

During the full-text-based selection process, an additional set of exclusion criteria was considered, as all the manuscripts at that point met the inclusion criteria described above. A manuscript is considered ineligible if:

- none of the inclusion criteria was satisfied or
- the AI-based system described in the manuscript was not applied to solving the ARC benchmark. This means, some of the AI-based systems were excluded from this systematic review even though they were specifically designed to perform abstraction and reasoning. The motivation for this exclusion criterium is based on the "comparator" point described in the PICO objectives formulation in Section 1.2.

2.2 Information Sources

This section provides a detailed overview of all information sources used for this systematic literature review. Table 1 represents all crucial characteristics of each of the information sources including the type of the information source (e.g., database, register, website, etc.), the date of the last access, and the type of that access (e.g., direct interaction, Application Programming Interface (API) call, etc.).

Table 1: Information sources

Name	Type	Last Access Date	Type of Access
ArXiv	Open-access archive	April 27, 2025	API call
Google Scholar	Academic search engine	May 3, 2025	Direct interaction
ACM Digital Library	Scientific database	May 3, 2025	Direct interaction
Semantic Scholar	Research tool	May 3, 2025	Direct interaction
Others	Project supervisor	March 3, 2025	Direct interaction

It must be clarified that results retrieved from some of the information sources might overlap, as search engines, for instance, might deliver manuscripts that are stored elsewhere.

2.3 Search Strategy

This section describes a way each of the information sources was accessed: what queries were used to retrieve any relevant information and the way these queries were executed. To simplify a further consolidation and selection process, a new data formatting scheme was introduced. The metadata about each of the manuscripts was structurally saved in a table with a following header: Title, Authors, Abstract, Link, and Publish Date. Queries used to retrieve needed metadata from each of the information sources are provided in Table 2.

Table 2: Queries used for metadata retrieval

Information Source	Query for metadata retrieval
ArXiv	(ti:"ARC" OR ti:"ARC-AGI" OR ti:"ARC-Prize" OR ti:"ARC Prize 2024" OR ti:"Abstraction" OR ti:"Reasoning" OR abs:"ARC" OR abs:"ARC-AGI" OR abs:"ARC-Prize" OR abs:"ARC Prize 2024" OR abs:"Abstraction" OR abs:"Reasoning") AND (cat:cs.AI OR cat:cs.LG OR cat:cs.NE OR cat:cs.CV OR cat:cs.MA)
Google Scholar	ARC, ARC-AGI, ARC Prize, ARC Prize 2024, Abstraction, Reasoning
ACM Digital Library	[All: arc, arc-agi, arc prize, arc prize 2024, abstraction, reasoning] AND [E-Publication Date: Past year] AND [E-Publication Date: (01/01/2025 TO 05/31/2025)]
Semantic Scholar	ARC, ARC-AGI, ARC Prize, ARC Prize 2024, Abstraction, Reasoning
Other	No specific query was used. Metadata was extracted manually.

2.4 Selection Process

This section describes the process of selecting articles for the final review. To make this process more manageable and streamlined, it was split into different stages:

- **Title Filtering:** Every title collected in the metadata table of a specific information source was examined, i.e., whether it refers to the topic of this systematic review and whether it fulfills the inclusion criteria or not.
- **Abstract Filtering:** For every record not filtered out in a previous steep, an assessment of whether an abstract of this record fulfills the inclusion criteria was performed.

- **Duplicates Filtering:** After the first two filtering steps were performed on the information source level, all the eligible records were consolidated into a common table. Consequently, all duplicates inside this common table were reconciliated leaving only one unique record á manuscript.
- **Metadata Clean Up:** An additional cleaning and reconciliation step was carried out to ensure metadata consistency. Specifically, it was explicitly examined whether metadata of all records fits the inclusion criteria (e.g. it was checked, if there were any manuscripts older than 2024 that would have had to be eliminated).
- **Full-Text Access:** After all pervious steps were carried out, a full text of a manuscript was accessed for each of the records in the metadata table. The content of every article from the metadata table was examined, i.e., whether it fits the exclusion criteria. In that case, the corresponding article was excluded from any further analysis.

2.5 Bias Assessment

The main goal of this section is to provide a full and straightforward overview of possible bias resulting from the design of this systematic literature review. Well known risk bias assessment frameworks such as AMSTAR (Shea et al., 2017) or AHRQ RRB (Agency for Healthcare Research and Quality, 2019) could not be applied to this systematic review directly, as they are primarily designed for clinical trials. Table 3 was, nevertheless, inspired by those bias assessment frameworks.

Table 3: Bias assessment

ID	Title	Explanation	Modesty
BA1	One-person study	This systematic literature review was performed by one person, thus implying potential bias when searching, selecting, filtering and analyzing manuscripts.	High
BA2	Manual work	Several actions, such as metadata extraction and filtering, were performed manually, thus increasing the chance of errors and towards specific research results.	Moderate
BA3	Use of unpublished or not peer-reviewed manuscripts	Several manuscripts have not yet been peer-reviewed or published at the time of the access (mainly articles from ArXiv).	Low
BA4	Reading list provided by the supervisor	The reading list provided by the supervising professor can be potentially biased.	Low

In conclusion, it can be said that most of the records from Table 3 have a relatively low modesty and do not render the results of this systematic review ineligible of any further research. Furthermore, the selection process has been designed with a goal to mitigate the BA2 (see step 4, Section 2.4). Moreover, all the data generated during the collection, selection and analysis processes is supplied together with this paper which further mitigates the bias BA1.

3. Results

3.1 Study Selection

This section provides an overview of the collection and selection results from Chapter 2. Results are reported in compliance to the PRISMA guidelines (Page et al., 2021) and are visualized in Figure 2.

A total of 538 manuscripts across 5 different information sources were identified. After duplicates filtering, a total of 494 manuscripts was passed down to the screening phase, where 404 publications were marked as ineligible according to the selection process explained in Section 2.4. Therefore, the full text of 88 manuscripts was revised to assess their eligibility. During this process, further 36 records were filtered out. Finally, by combining these manuscripts with literature from other sources, a final number of 62 studies were included in the final systematic review.

Figure 2: PRISMA 2020 flow diagram for new systematic reviews which included searches of databases, registers and other sources

3.2 Study Classification

During the full-text screening process, each manuscript was read and analyzed. To structure this process, a custom review protocol was developed. The template of this protocol includes the following elements: *Title*, *Metadata (e.g. date of the protocol creation), Tags (see Table 4), Author(s), Summary*, *Publish Date,* and *Link*.

Table 4: Tag descriptions used in the systematic review

Tag	Explanation
#DSL	The solution approach uses a Domain Specific Language (DSL).
#program_synthesis	The solution approach depicted in this paper uses program synthesis.
#llm	An LLM is utilized in the described solution approach.
#novel_model_architecture	This paper uses a custom or a novel model architecture to solve ARC tasks.
#transformers	This publication uses a variation of a transformers architecture to solve ARC tasks.
#data_generation	This publication utilizes a unique approach for data generation.
#test-time-training	The solution approach uses a variation of Test-Time Training (TTT).
#fine-tuning	The solution approach employs a version of Fine Tuning (FT) to solve ARC tasks.
#induction	This solution approach can be categorized as induction.
#transduction	This solution approach can be categorized as transduction.
#abduction	The solution approach utilizes abduction to solve ARC tasks.
#arc_tools	New tools for ARC tasks solutions are introduced in the paper.
#reinforcement_learning	A version of a Reinforcement Learning (RL) algorithm is used to solve ARC tasks.
#search_algorithm	A novel search algorithm is introduced in the paper.
#inference_mehods	Some inference methods such as Monte Carlo Tree Search (MCTS) or voting techniques were used.
#derived_arc_dataset	A novel dataset was derived from the ARC-AGI in this publication.
#human_performance	Human performance was analyzed in this paper.
#meta_learning	A version of meta-learning was applied to solve ARC tasks.
#positional_encodings	Positional Encodings (PEs) play a vital role in this solution approach.

Tag	Explanation
#paywall	This article is behind a paywall and the full text could not be accessed.
#arc_irrelevant	This article is irrelevant to the systematic review. This means, this manuscript was excluded based on the exclusion criteria.

Table 4 shows all the tags used in the categorization and classification of the solution approaches. During the full-text screen procedure, similarities between different articles were documented using tags. Every time a new tag was introduced, it was first added to Table 4. Subsequently, all the review protocols analyzed before the introduction of a new tag were inspected one more time to assign a new tag to them, if it was needed.

Figure 3 shows a distribution of tags across all manuscripts assessed during the full-text assessment procedure. Notably, one review protocol can have multiple tags. The most relevant solution approaches to the ARC Prize 2024 from Figure 3 are discussed in Section 4.1.

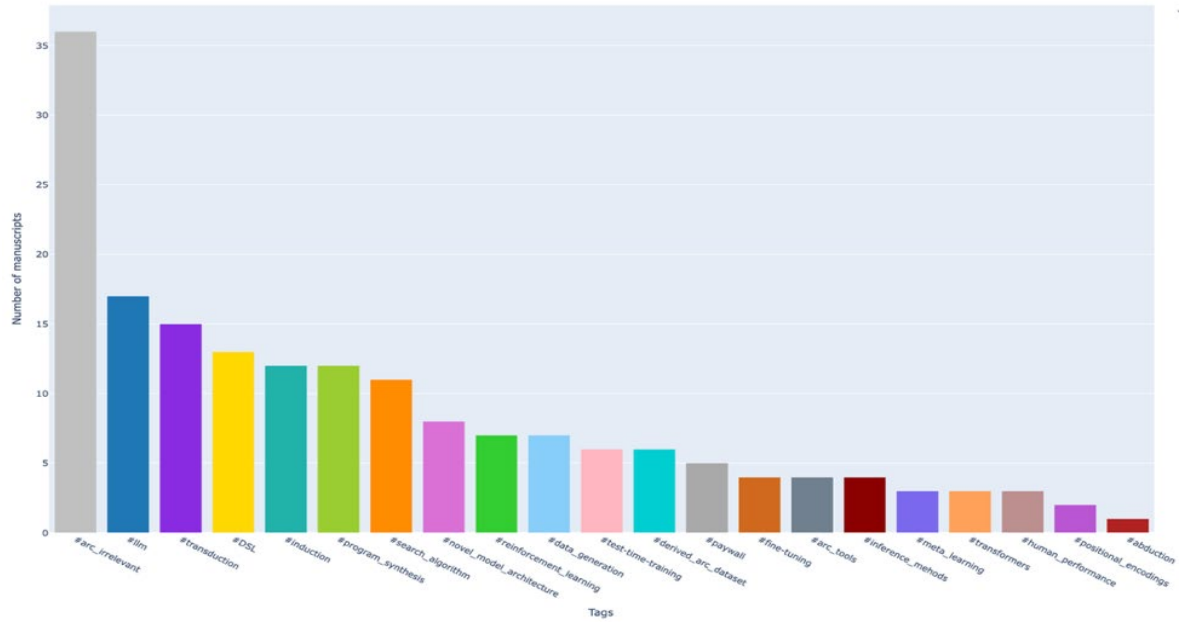

Figure 3: Tags distribution

Furthermore, a knowledge graph consisting of all review protocols was created using the Markdown language editor Obsidian (https://obsidian.md). A set of review protocols $P = \{p_1, p_2, \dots\}$ and a set of tags $T = \{t_1, t_2, \dots\}$ build a set of vertices of the knowledge graph as follows:

$$V = P \cup T$$

A set of edges $E = \{\{x, y\} \mid x, y \in V \wedge x \neq y\}$ contains all pairs of vertices that have any connection to each other, e.g. a review protocol is tagged with one or more tags. The resulting knowledge graph is, therefore, a connected and undirected graph:

$$G = (V, E)$$

Additionally, one further Obsidian feature was utilized for this systematic review: when hovering over a vertex $v \in V$, all the first-order neighbors of this vertex $W = \{w \in V, | \{w, v\} \in E \wedge v \neq w\}$ are automatically highlighted (see Figure 4).

3.3 Performance Data Extraction

All manuscripts reviewed during the full-text access stage were scanned for performance data of a solution approach described in each paper. As mentioned in Section 1.2, the goal was to compare different solutions based on their per-task performance. To do that, a set of 400 training task from the ARC benchmark was chosen, as it has been hypothesized that most of the solution approaches would be tested on this set or on a subset of it. During the review process, (Drori et al., 2025) was identified to have tested the largest number of different solution approaches and models on this set of tasks.

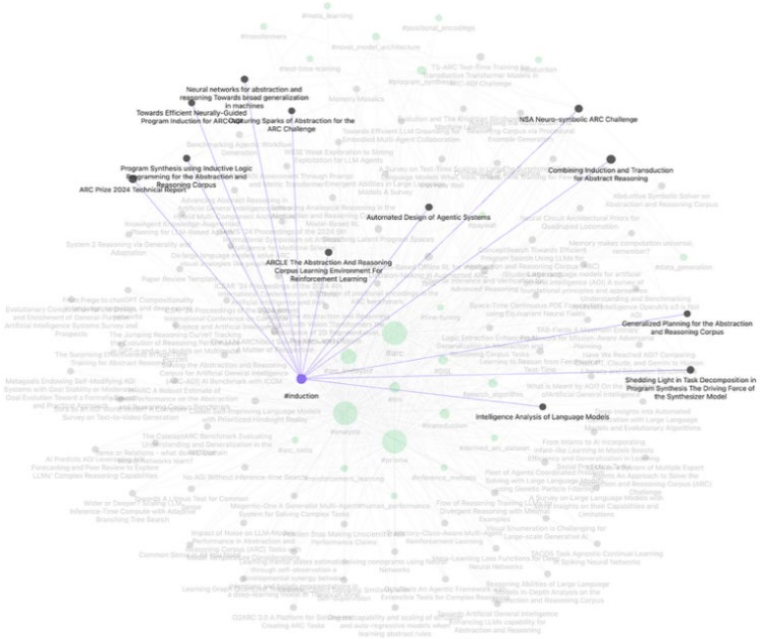

Figure 4: Highlighting of the knowledge graph

3.3.1 Szymkiewicz-Simpson coefficient

This coefficient represents the proportion of tasks solved by the stronger model (say A) that were solved by the weaker model (say B):

$$Overlap(A,B) = \frac{|A \cap B|}{\min(|A|,|B|)}$$

The Szymkiewicz-Simpson coefficient can be considered a performance-normalized similarity metric, as it makes it possible to determine whether two approaches solve similar problems, meaning, a stronger approach is just a more powerful version of a weaker one, or whether they solve distinct types of tasks (see Figure 5). It is regarded to be performance-normalized, because the coefficient does not depend on a performance of a particular model, but rather on the coefficient between the models (Bober-Irizar and Banerjee, 2024).

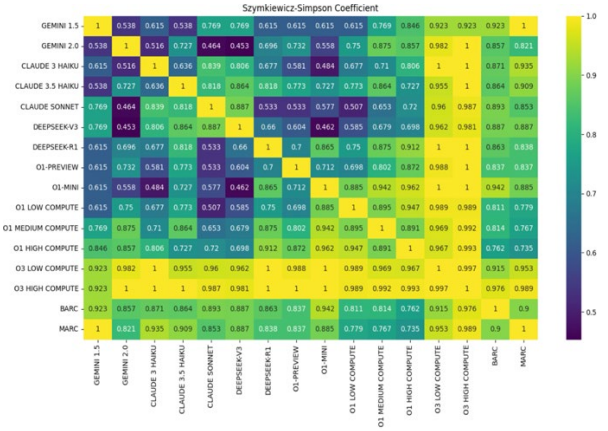

Figure 5: The Szymkiewicz-Simpson coefficient of the performance data from (Drori et al., 2025)

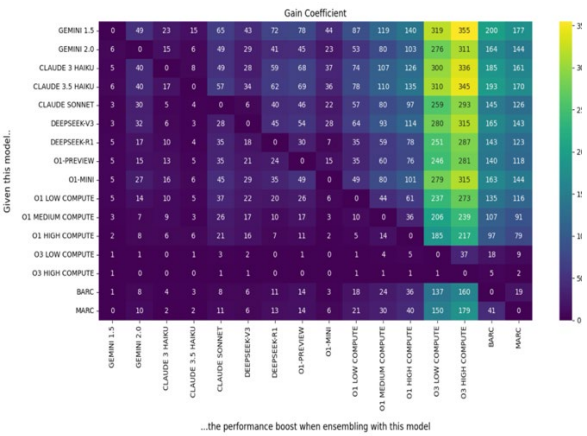

Figure 6: The Gain coefficient of the performance data from (Drori et al., 2025).

3.3.2 Gain coefficient

This is an asymmetric coefficient that shows the number of tasks that can be theoretically solved by combining different solution approaches:

$$Gain(A,B) = |A \cap B| - |A|$$

The gain coefficient provides insights to the question, which solution approach performs better and how much better this performance is in comparison to any other model (see Figure 6). This coefficient can also be seen as a measure of unique tasks that can be solved only by a combination of two solution approaches (Bober-Irizar and Banerjee, 2024).

4. Discussion

4.1 Interpretation of Results of the Selection and Classification Processes

Based on a collection of review protocols, tag frequency statistics, and knowledge graph representation, following the statements can be made:

- A large number of solution approaches to the ARC Prize 2024 utilize a Domain Specific Language (DSL) in combination with an LLM. Most frequently, they consist of a set of primitives that are used to locate and transform objects in a grid. Notably, Ouellette (2024) claims that a larger set of initial primitives of a given DSL theoretically leads to a larger search space that has to be explored by a model. Ouellette (2024) was, nevertheless, able to empirically show that the solution time of his approach scales sub-linearly. Further solution approaches that utilize a DSL are: (Andrews, 2024; Batorski et al., 2025; Bober-Irizar and Banerjee, 2024; Butt et al., 2024; Chollet et al., 2025; Hodel, 2024; Lei et al., 2024; Lim et al., 2024; Rocha et al., 2024; Singhal and Shro, 2025; Zenkner et al., 2025).
- Using the knowledge graph from Figure 4 it was empirically determined that 8/13 of all solution approaches that utilize a DSL were also tagged as approaches employing induction. Furthermore, a similar analysis shows, that exactly 10 in 13 of DSL-based solution approaches rely on program synthesis. Based on this empirical analysis it can be concluded that many ARC solution approaches that rely on a DSL often utilize it during program synthesis and thus can be categorized as inductive approaches.
- Notably, the number of manuscripts tagged with **#transaction** is bigger than those with **#induction**. Interestingly, Li et al. (2024) were able to statically show that induction- and abduction-based models solve different types of tasks. Thus, it can be hypothesized that combining induction and transduction approaches can potentially result in a better overall performance.
- Furthermore, 5 out of 6 TTT-based approaches were also classified as transduction approaches. This can be explained by the fact that TTT is a performance-enhancement technique utilized during the inference time by transductive models. According to François Chollet (2025), TTT-based models show a strong performance on abstraction and reasoning tasks.
- Moreover, several solution approaches utilize combinations of agents or LLMs to generate multiple answers during inference time. This poses a problem of selecting the best possible solution from different candidates. To solve these problems, 4 manuscripts employ different inference techniques (e.g. Chain-of-Thought (CoT) (Wei et al., 2023) or techniques as intra-transformation or global voting (Akyürek et al., 2025)). On the other hand, several search strategies are also employed to solve a similar problem not only for picking the final answer, but also for exploring and exploiting intermediate states during inference time (e.g. Abstraction Branching Monte Carlo Tree Search (Misaki et al., 2025)).

4.2 Interpretation of the Performance Data Extraction Process

4.2.1 Interpretation of the Szymkiewicz-Simpson coefficient heatmap

The data shown in Figure 5 is derived from the per-task performance of different solution approaches including several LLMs (e.g. Gemeni 2.0 and Claude Sonnet) and some specifically engineered approaches: BARC (Li et al., 2024), MARC (Akyürek et al., 2025). Using this data representation, it can be concluded that solution approaches with a low Szymkiewicz-Simpson coefficient, for instance pairs of Gemini 2.0 and Claude Sonnet or Deepseek-V3 and Gemini 2.0, could theoretically deliver a better performance when used together, as they tend to solve different types of tasks.

On the contrary, model pairs like Deepseek-R1 and o1-high-compute or o1- medium-compute and Gemini 2.0 have a Szymkiewicz-Simpson coefficient of 1, meaning that theoretically no performance improvement can be expected when these pairs of models are used together. It shall be noted that o3-high and o3-low models are able to solve almost 90% of the ARC tasks, thus implicitly inducing a near-one Szymkiewicz-Simpson coefficient when compared with any other solution approach. Therefore, the Szymkiewicz-Simpson coefficient cannot be applied to these models to make any assumptions.

4.2.2 Interpretation of the Gain coefficient heatmap

Figure 6 can be used to estimate several additional tasks that can be theoretically solved by a combination of two models. Due to the asymmetric nature of the heatmap it can be clearly seen that the two extra engineered approaches, BARC and MARC, tend to boost the performance of all other solution approaches that utilize LLMs. A maximal performance improvement can be theoretically observed by combining Gemini 1.5 and BARC.

Notably, as in the case of the Szymkiewicz-Simpson coefficient, an extremely good performance of o3 family of models on the ARC benchmark can also lead to a misleading interpretation of gain coefficient heatmap results. Specifically, combining Gemini 1.5 with o3-high-compute models results in an improvement of 355 additionally solved tasks. This happens not because these models solve complementary types of tasks, but rather because o3-high-compute solves almost all the tasks.

5. Conclusion and Future Work

A PRISMA-guided publications search, selection and evaluation of abstraction and reasoning abilities in AI-based systems applied to solving the ARC Prize was conducted. Five information sources were accessed and 538 potentially relevant manuscripts were identified. After the selection and full-text screening phases, 62 studies were included in the final systematic review. During the full-text assessment step, all manuscripts were tagged based on the technologies and techniques they utilize. The tagging enabled the creation of a knowledge graph that was further used for the classification of solution approaches and insights extraction.

Based on the tagging performed in this systematic literature review it can be concluded that most of the solution approaches to the ARC Prize utilize a DSL in combination with program synthesis, thus making these approaches inductive in nature. On the contrary, transductive solution approaches mostly employ inference techniques as CoT and search techniques as MCTS. Furthermore, the per-task performance data of 16 solution approaches was visualized using the Szymkiewicz-Simpson and the Gain coefficients.

This study holds significant potential for further development. The suggested structure of review protocols is flexible enough to accommodate other possible extensions. An evaluation of further solution approaches could be considered in the future, so that the coefficient heatmaps can provide more insightful information for possible combinations of multiple approaches. Moreover, a way to compare more than two approaches at a time would be an insightful topic to explore.

Ethics declaration: The research provided in this paper did not require any ethical clearance.

AI declaration: ChatGPT (gpt-4o) was used to choose colors for the tag distribution diagram (see Figure 3).

References

Ademola, S.S., Yusuf, A.O., Olalekan, A.A. et al. (2024) 'The Rise of Large Language Models (LLMs) in Academic Research: Opportunities and Challenges', Journal of Digitovation and information system, 4(2), pp. 144–159. https://doi.org/10.54433/JDIIS.2024100043.

Agency for Healthcare Research and Quality (2019) Methods Guide – Chapter: Assessing the Risk of Bias in Systematic Reviews of Health Care Interventions. https://effectivehealthcare.ahrq.gov/products/methods-bias-update/methods.

Akyürek, E., Damani, M., Zweiger, A. et al. (2025) The Surprising Effectiveness of Test-Time Training for Few-Shot Learning. http://arxiv.org/abs/2411.07279.

Andrews, M. (2024) Capturing Sparks of Abstraction for the ARC Challenge. http://arxiv.org/abs/2411.11206.

Batorski, P., Brinkmann, J. and Swoboda, P. (2025) NSA: Neuro-symbolic ARC Challenge. http://arxiv.org/abs/2501.04424.

Bober-Irizar, M. and Banerjee, S. (2024) Neural networks for abstraction and reasoning: Towards broad generalization in machines. http://arxiv.org/abs/2402.03507.

Butt, N., Manczak, B., Wiggers, A. et al. (2024) CodeIt: Self-Improving Language Models with Prioritized Hindsight Replay. http://arxiv.org/abs/2402.04858.

Chollet, F. (2019) On the Measure of Intelligence. https://philpapers.org/rec/CHOOTM.

Chollet, F., Knoop, M., Kamradt, G. et al. (2025) ARC Prize 2024: Technical Report. http://arxiv.org/abs/2412.04604.

Drori, I., Longhitano, G., Mao, M. et al. (2025) Diverse Inference and Verification for Advanced Reasoning. http://arxiv.org/abs/2502.09955.

Hodel, M. (2024) Addressing the Abstraction and Reasoning Corpus via Procedural Example Generation. http://arxiv.org/abs/2404.07353.

LeGris, S., Vong, W.K., Lake, B.M. et al. (2024) H-ARC: A Robust Estimate of Human Performance on the Abstraction and Reasoning Corpus Benchmark. http://arxiv.org/abs/2409.01374.

Lei, C., Lipovetzky, N. and Ehinger, K.A. (2024) Generalized Planning for the Abstraction and Reasoning Corpus. http://arxiv.org/abs/2401.07426.

Li, W.-D., Hu, K., Larsen, C. et al. (2024) Combining Induction and Transduction for Abstract Reasoning. http://arxiv.org/abs/2411.02272.

Lim, M., Lee, S., Abitew, L.W. et al. (2024) Abductive Symbolic Solver on Abstraction and Reasoning Corpus. http://arxiv.org/abs/2411.18158.

Misaki, K., Misaki, K., Imajuku, Y. et al. (2025) Wider or Deeper? Scaling LLM Inference-Time Compute with Adaptive Branching Tree Search. http://arxiv.org/abs/2503.04412.

Ouellette, S. (2024) Towards Efficient Neurally-Guided Program Induction for ARC-AGI. http://arxiv.org/abs/2411.17708.

Page, M.J., McKenzie, J.E., Bossuyt, P.M. et al. (2021) 'The PRISMA 2020 statement: an updated guideline for reporting systematic reviews', BMJ, 372. https://doi.org/10.1136/bmj.n71.

Rajpurkar, P., Zhang, J., Lopyrev, K. et al. (2016) SQuAD: 100,000+ Questions for Machine Comprehension of Text. http://arxiv.org/abs/1606.05250.

Richardson, W.S., WilsonM.C., Nishikawa, J. et al. (1995) 'The well-built clinical question: a key to evidence-based decisions', ACP Journal Club, 123(3), pp. A12–A13.

Rocha, F.M., Dutra, I. and Costa, V.S. (2024) Program Synthesis using Inductive Logic Programming for the Abstraction and Reasoning Corpus. http://arxiv.org/abs/2405.06399.

Shea, B.J., Reeves, B.C., Wells, G. et al. (2017) 'AMSTAR 2: a critical appraisal tool for systematic reviews that include randomised or non-randomised studies of healthcare interventions, or both', BMJ, 358. https://doi.org/10.1136/bmj.j4008.

Singhal, K. and Shroff, G. (2025) 'ConceptSearch: Towards Efficient Program Search Using LLMs for Abstraction and Reasoning Corpus (ARC) (Student Abstract)', Proceedings of the AAAI Conference on Artificial Intelligence, 39(28), pp. 29493–29494. https://doi.org/10.1609/aaai.v39i28.35300.

Statista (2024) Artificial intelligence (AI) market size worldwide from 2020 to 2030 (in billion U.S. dollars) [Graph]. https://www-statista-com.ezproxy.hwr-berlin.de/forecasts/1474143/global-ai-market-size.

Wang, A., Wang, X., Schuurmans, D. et al. (2019) GLUE: A Multi-Task Benchmark and Analysis Platform for Natural Language Understanding. http://arxiv.org/abs/1804.07461

Wei, J. et al. (2023) Chain-of-Thought Prompting Elicits Reasoning in Large Language Models. http://arxiv.org/abs/2201.11903.

Zenkner, J., Sesterhenn, T. and Bartelt, C. (2025) Shedding Light in Task Decomposition in Program Synthesis: The Driving Force of the Synthesizer Model. http://arxiv.org/abs/2503.08738.