Analyzing Elementary School Olympiad Math Tasks as a Benchmark for AGI

Alexey Potapov(✉), Oleg Scherbakov, Vitaly Bogdanov, Vita Potapova, Anatoly Belikov, Sergey Rodionov, and Artem Yashenko

SingularityNET Foundation, Amsterdam, The Netherlands
{alexey,olegshcherbakov,vitaly,abelikov,sergey, yashenko}@singularitynet.io

Abstract. Many benchmarks and challenges for AI and AGI exist, which help to reveal both short- and long-term topics and directions of research. We analyze elementary school Olympiad math tasks as a possible benchmark for AGI that can occupy a certain free niche capturing some limitations of the existing neural and symbolic systems better than other existing both language understanding and mathematical tests. A detailed comparison and analysis of implications of AGI is provided.

Keywords: AGI · AI evaluation · Math tasks · Language understanding

1 Introduction

Having some metric to estimate progress in a certain domain is considered as a necessity in contemporary AI practice. At the same time, there is no generally accepted standard AGI benchmark, although theoretical metrics of AGI exist (e.g. [1]) as well as different empirical tests and challenges have been proposed (e.g.[1]). However, each of them either requires a real AGI to pass it, or, in contrary, can be (partially) solved by narrow AI techniques, or at least favors a certain approach to AGI or a type of proto-AGI systems (for example, reinforcement learning models will be favored by certain environments, while such challenges as General Game Playing discourage the use of learning at all). It is a not uncommon opinion that comparing different proto-AGI or measuring progress towards AGI in an unbiased way is very hard [2].

The paper [3] overviewed thirty computer models addressing intelligent test problems, and came to the conclusion that these models have different purposes and applications, and have a limited connection between each other. But still, AGI benchmarks are far from worthless by themselves, and possess a considerable methodological importance, because they help to understand limitations of the existing methods and reveal possible directions of further research. Although the effort to encourage future computer models taking intelligence test problems to link with and build upon previous research

[1] https://www.general-ai-challenge.org/.

B. Goertzel et al. (Eds.): AGI 2020, LNAI 12177, pp. 279–289, 2020.
https://doi.org/10.1007/978-3-030-52152-3_29

made in [3] is really useful, we see some objective reasons in the diversity of the existing intelligent tests.

Indeed, although standard benchmarks exist for many domains in narrow AI, these benchmarks also fail to specify an ultimate goal even within rather particular tasks, and optimizing some metric is not an end in itself but only an intermediate goal, which we managed to formulate based on our current understanding of the task, which can be imprecise or even misleading. It frequently appears that the state-of-the-art methods are steadily improving their scores on some benchmark, but are doing this in the way we just "don't like", and then the benchmarks themselves start to being criticized and improved upon. For example, the visual question answering (VQA) datasets were criticized [4] for lacking compositional questions, allowing confidently answering questions without looking at images, etc., which were fixed in other benchmarks (e.g. [5]), which, in turn, had other drawbacks and limitations and were further improved upon. However, these drawbacks were not so obvious from the beginning, and a perfect benchmark would be difficult to create even for such restricted task as VQA. This should be even truer for AGI.

In this paper, we do not pretend to create an ultimate AGI metric, but discuss yet another possible AGI-ish benchmark, which, however, has some advantages and can have a certain utility as discussed below. The basic idea is to compose a dataset using elementary school mathematical Olympiad tasks. A similar proposal to use mathematical puzzles as a challenging competition for AI [6] has been made, but without referring to Olympiad tasks as a source for arranging a concrete dataset and without relevance to AGI. In the following sections, we compare this idea with some related benchmarks highlighting differences and consequences for AGI, which are worth discussing even before creating the benchmark itself.

2 Related Works and Discussion

Natural Language Understanding

Language is frequently considered as one of the main differences between human and animal intelligence. An extreme form of focusing on language is expressed in "equation": "language – sound = thinking". The seminal Turing test was essentially a natural language understanding (NLU) test, while the main point of Searle's Chinese room argument was to show that computers (physical symbol systems) are incapable of language understanding in principle. Nowadays, many benchmarks in narrow AI exist for question answering, dialogs, text generation and other language processing tasks.

Modern deep neural network (DNN) models may show nearly human or even superhuman scores on some benchmarks. However, the way they do this (in comparison with more traditional symbolic systems) is the source of ongoing debates. Is it really possible to map arbitrary sentences to a large, but fixed vector space of their meanings? Do DNN models really understand sentences, or mostly memorize huge text corpora and recall them? Is it possible to understand texts in natural language without even attempting to represent real-world situations, described in them?

Some tests and challenges exist, which try showing the lack of understanding in the existing models. One example is the Winograd Schema Challenge (WSC) [7]. Questions in WSC follow the same pattern and contain an ambiguous pronoun to be associated with nouns using knowledge and commonsense reasoning. WSC is reasonably difficult: best models demonstrate ~70% accuracy that is not too low, though, to deprive of hope for solving this challenge by incrementally improving and tweaking the existing models. Also, it may appear that the challenge can be solved using purely linguistic knowledge and simple ontological relations. Another drawback of WSC is that it contains only 150 schemas, which apparently cannot be used for training and extracting necessary knowledge from the dataset itself (although its recent analogue, WinoGrande[2], contains 44k problems).

The standard General Language Understanding Evaluation (GLUE) benchmark [8] includes WSC along with other 8 NLU tasks including sentiment analysis, semantic similarity of sentences, and others. Each of these tasks highlights one or another aspect of language understanding, and all together they cannot be called narrow. However, they all are still too focused on the language domain itself. For example, the Corpus of Linguistic Acceptability (CoLA) requires distinguishing between (grammatically) acceptable and inacceptable sentences, e.g. "*John tried to be a good boy*" and "*Who does John visit Sally because he likes?*" correspondingly.

Consider the following question from WSC as an example: "Joan made sure to thank Susan for all the help she had [given/received]. Who had [given/received] help?". Apparently, in order to answer it, a model does need to "know" that it is usually a person, who receives help, who thanks a person who helps. However, it doesn't really need to understand what it means to help. What it really needs is just an ontological relation – not its real-world grounding.

Other NLU tasks can require using some factual encyclopedic knowledge, but without its real understanding. Some tests involve scientific knowledge also. For example, the Aristo project [9, 10] dataset includes such questions as "Which object in our solar system reflects light and is a satellite that orbits around one planet? (A) Moon (B) Earth (C) Mercury (D) Sun", which requires not only language processing and basic reasoning abilities, but also commonsense and scientific knowledge representation and manipulation. Such tests have their own utility, but they don't require an understanding of what it means to orbit around a planet or to reflect light. What is necessary is just a set of relations or facts "The Moon orbits around the Earth", "The Earth is a planet", etc. Indeed, these are so-called open book questions for understanding of qualitative relationships.

Let us consider a few examples of elementary school math tasks for comparison:

- A group of girls stands in a circle. Emily is the fifth on the left from Mary and the sixth on the right from Emily. How many girls are in the group?
- Nicole takes a sheet of paper and cuts it into 9 pieces. She then takes one of these pieces and cuts it into 9 smaller pieces. She then takes another piece and cuts it into 9 smaller pieces and finally cuts one of the smaller pieces into 9 tiny pieces. How many pieces of paper has the original sheet been cut into?
- How many different cubes are there with two faces colored green and four faces colored yellow?

[2] https://leaderboard.allenai.org/winogrande/.

Imagine how these tasks can be solved by an AI system, e.g. a DNN model. It should be noted that the tasks are quite unique. There can be a few more tasks involving standing in circles or cutting sheets of paper, but they will be formulated in a different way and require inferring other consequences. At the same time, quite a large number of different tasks exist, and these tasks are not wiredrawn, but "real-world" in sense that they are really given to human children. Apparently, our AI system cannot just memorize the training dataset and recall similar tasks. These tasks don't require the extensive use of factual encyclopedic knowledge (which can be memorized), but suppose a deeper understanding of what a circle is or what cutting is that goes beyond pair-wise relations between symbolic atoms and requires at least some modeling of corresponding "physical" situations. It will not be enough to map the sentences into some semantic vector space. The system will most likely require having an internal model of girls standing in a circle and explicitly reason over it.

We believe such tasks are more indicative of what "understanding" is and their formulations cover quite a wide spectrum of aspects of natural language also (but of course not all of them, e.g. sentiment analysis is not covered). We don't say that other NLU tests are worse, but we claim that the mentioned math tasks require dealing explicitly with an additional important aspect of natural language understanding, which is rarely highlighted in other NLU tasks (which, however, better cover some other aspects). Besides NLU, these math tasks require some form of reasoning, which is also important for AGI benchmarking.

Recently, SuperGLUE [11] benchmark was proposed with a new set of natural understanding tasks. Although these tasks are more difficult, they are also purely textual and do not heavily require symbol (textual entities) grounding.

Visual (and Physical) Reasoning

The lack of necessity of grounding linguistic entities in the real world in purely textual NLU tasks is not a novel observation and has been addressed in multimodal benchmarks, which most often rely on visual input. Interestingly, many school mathematical tasks involve images, and can be considered as questions about images, which make them similar to VQA tasks.

As mentioned above, the earlier VQA datasets were criticized for that relatively high scores on them could be achieved with the use of superficial correlations between textual tokens in questions without both reasoning and clear grounding of words in images. Some of consequent datasets (e.g. [4]) introduced different biases in training and test subsets to prevent using superficial correlations. More interesting is that considerable efforts have been made to stimulate the focus on reasoning in VQA. In particular, CLEVR is a synthetic dataset with simple scenes, but complex questions about objects and their spatial relations. Later, similar complex questions were generated using Visual Genome for real-world scenes [12].

Images in math problems are mostly composed of abstract shapes or simple objects and are closer to CLEVR in this respect, but they are not generated by a simple formal process. They are much closer to real-world VQA than CLEVR in terms of "open-endedness". Although they don't require recognizing a great variety of real objects (which is of course an important, but a sort of vision-domain-specific property), they require a deeper image understanding than traditional VQA datasets. Consider Fig. 1.

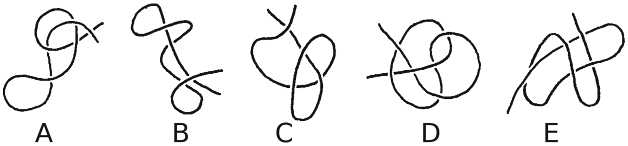

Fig. 1. Which ropes will be tightened into knots if they are pulled by the ends?

It can be seen that while VQA tasks require just extracting bounding boxes of discrete objects and discrete relations between them, math tasks require analyzing images in finer details. Also, while a DNN might be able to learn from thousands of examples some features enough to answer the question about the ropes, it will not generalize to other such tasks and learn from few examples.

Apparently, school math problems require much more complex and open-ended reasoning in comparison to synthetic questions of low diversity, which are really compositional but hardly require reasoning. Indeed, they can be answered by a direct seq2seq mapping of textual questions to imperative programs.

We don't claim that math questions with images form a perfect VQA dataset, but such a dataset can be quite indicative in terms of structural image understanding and visual reasoning (showing how far the state-of-the-art VQA models from real visual reasoning even over such simplistic images).

It should be noted that there are types of tasks, which use images and (optionally) textual questions as input, although they are not considered as VQA tasks. One example of such tasks is Physical Bongard Problems (e.g., [13]), which requires categorizing simple synthetic scenes based on their physical properties (e.g. stable/unstable configuration, small objects fall down, etc.).

Physical Bongard Problems are conceptually similar to the math tasks under discussion in that answers to them don't directly follow from images, but require some internal representation of depicted situations, over which reasoning is carried out. Of course, there are many differences in details, and these two sets of problems don't intersect, but complement each other. Physical Bongard Problems also don't contain textual questions and are devoted to a relatively restricted subdomain of naive physics (concretely, dynamics and object interaction). Both these properties are good for some purposes, but make Physical Bongard Problems hackable by narrow methods (especially taking into account that not too many problems exist).

It should be mentioned that physical problems were also considered in the context of cognitive psychology, in particular as a test case for analogical reasoning and transfer learning (e.g. [14]). However, the possibility to solve the particular tests being used in such studied by hand-crafted or narrow methods wasn't analyzed.

Elementary school mathematical Olympiad tasks don't require extensive physical knowledge or detailed simulation. Instead, they highlight the necessity to represent scenes or situations in a way that allows reasoning over them.

Mathematical Tests

We have compared school math tasks with NLU and VQA tasks showing their utility in AGI testing, but one may wonder if there are other existing benchmarks based on math tasks. Indeed, the ability of mathematicians to decompose, abstract and solve real world problems was the golden standard of thinking and intelligent processing during evolution of AI research agenda especially at the early stages of AI field establishment. To solve even simple math puzzles humans use analytical abilities such as logical and spatial-temporal reasoning as well as intuition, understanding and common sense. To find out if AI systems have capabilities of handling non-trivial math and reasoning problems several challenges have already been proposed. IMO (International Mathematical Olympiad) Grand Challenge[3] is probably one of the most well known. This challenge calls for building an AI system that can win a gold medal in the IMO competition among humans.

It may appear that the IMO Grand Challenge already brings our proposal to its ultimate form. However, there is an essential difference between them. IMO tasks are purely mathematical and are provided to AI in a formalized representation.

In contrast, texts of elementary school math tasks don't define formal constructions for conducting inference, but describe real-world situations, which require constructing some models that formalize these situations with higher or lower precision. For example, if we consider two objects moving towards each other, we can sum up their velocities only as an approximation (in contrast to the relativity theory, we suppose existence of some global time and no speed limit). Thus, formalization is achieved not by a direct text2math mapping, but through simulation (imagination) of the situation (in this context, it is interesting to note the discussion on the nature of mathematical knowledge and its relation to AI [15]).

Even after reconstructing the situation, the task can remain underformalized. In fact, complete formalization and inference over it can be cumbersome even in pure math tasks. Indeed, consider the task "prove that at least one of two numbers is divisible by 3 if their product is divisible by 3" – a fully formalized solution may be surprisingly long, especially if it doesn't rely on lemmas about simple factoring. At the same time, we can imagine that the product of two number is composed of 3 and the rest part, which is divided into two pieces belonging to different initial numbers, and "3" should "go" into one of them making it divisible by 3. It is convincing, although not really formal. Answers to less formal tasks can be obtained by "physical" simulation or via knowledge-based reasoning. For example, for the task of cutting a sheet of paper, we can imagine how this sheet is cut, although we need to suppose some commonsense-based invariance during this simulation, i.e. to figure out if it matters or not where it is cut, in what order, etc. Alternatively, we can just know that cutting a sheet of paper destroys it, and thus cutting one piece into 9 pieces increases the total number of pieces by 8. But even if start with this simplistic "formalization of cutting" and represent the process as an algorithm that takes a list with one element as input and iteratively removes one random element from the list and inserts 9 new elements into it, a complete formal proof that the length of the list produced by this algorithm is independent on random choices will be not that short.

Consider the task "Bella colors all the small squares that lie on the two longest diagonals of a square grid. She colors 2021 small squares. What is the size of the square grid?". When we write down equation 2size – 1 = 2021 relying on the fact that the number of squares in the longest diagonals is the same as the size of the board, and they

[3] https://github.com/IMO-grand-challenge/.

have one common square, the answer is obvious. But it's semi-formal. To be completely formal, it should contain definitions of boards, diagonals, etc. as mathematical objects. These definitions can be cumbersome. Of course, we can rely on formerly proved lemmas about diagonals, etc. (we can imagine Agda or Coq-style definition of boards, diagonals, coloring and so on as dependent types), but still the mapping to this formalization is not that straightforward.

Even when some tasks rely on physics to a nearly zero extent, and suppose a more direct translation into, say, algebraic representation, they are first translated into some representation of a "real-world" situation. In fact, the skill of using algebraic representation is not natural and should be specially developed prior to solving math tasks per se (and actually, it was discovered by humankind just a few centuries ago) as was pointed out by George Pólya long time ago. Only higher-grade tasks become purely mathematical, when pupils have developed an internal representation of this abstract domain separately (or on top of) perceptual world representation and simulation.

Consider the task: "Bill lacks 8 cents to buy the apple, while Mary lacks 1 cent to buy the apple. How much does the apple cost if Bill and Mary cannot buy it even if they put their money together?" It is very simple mathematically, and it supposes quite a straightforward complete formalization, but still, humans (both children and adults) rarely solve it via this complete formalization. Rather, they arrive at the solution semi-formally. First of all, we'd be surprised: how is it possible that they have not enough money together if Mary lacks just 1 cent? Eureka! Bill has no money at all. We don't bother with writing down the following system: $a + 8 = x$, $b + 1 = x$, $a + b < x$, $a \geq 0$, $b \geq 0$. Besides the fact that the last two inequalities require some background knowledge and commonsense assumptions, this is not really how we solve this task.

One can claim that the abstract world of IMO-type math problems is no less important than the world of clocks, buses, sheets of paper and so on, and the ability to solve IMO tasks is more indicative from the AGI-ish point of view. However, all real-world tasks (related not necessarily to everyday environment, but to any object or system of scientific study) differ from IMO tasks in that they involve very complex objects, many properties of which are not necessary, while some other important properties are missing and should be filled in with default or commonsense values. Isolating the problem (even already given in natural language) from the rest of Universe and representing it in a solvable way is absent in IMO Grand Challenge, and it can be more difficult than solving a formalized task.

The main difficulty of applying symbolic systems to real-world tasks consists in translating input data into representations, over which these systems can reason. At the same time, end-to-end trainable deep learning models have rather weak reasoning capabilities (and fail to learn to reason as well).

Indeed, recently an attention of the research community has been shifted to estimation of the ability of DNN models to solve math-alike problems. Neural models successfully handle many of the general text problems, but parsing and answering math questions is a very special task which is at least at the first glance cannot be directly generalized from standard pretrained model. However some of the researchers are trying to experimentally evaluate such generalization properties of DNN models at least in restricted problem-set conditions.

In the paper [16] researchers introduce the Mathematics Dataset consisting of many different types of mathematics problems that cover topics in algebra, arithmetic, basic combinatory and probability theory. There are two types of tests: interpolation and extrapolation tests. Interpolation tests assume that all types of questions were presented during the training but test set questions have to be presented at most 2% of the total test set size. Extrapolation tests estimate generalization capabilities of the trained models to work with tasks, which differ from training ones by larger numbers, more numbers involved in equation, more compositions, and (if it was a probability question) larger samplers. The authors have also examined several popular general purpose models. All of the models were modern neural architectures for solving sequence-to-sequence problems: recurrent neural architectures, and attentional/transformer architecture. The authors also claim that they tried to use advanced neural models with external memory, like Differentiable Neural Computers [17], which could be potentially well suited for solving mathematical questions. But it is reported that there is no significant outcome from the usage of these models. The researchers also have shown some interesting flaws in models performance on very simple tasks of adding series of "ones", where "one" appears n times for $n > 5$. It is especially interesting because the models could correctly predict results for longer sequences of far bigger numbers. The major takeaway from this study is that the modern DNN models do not generalize well to the specific problem domain like math questions even in well-controlled environments consisting of formally defined tasks though in natural language.

Interestingly enough, more recent Tensor Product Transformer model [18] has shown some promising results on the Mathematics Dataset. The dataset includes tasks like "*What is the first derivative of 13 * a ** 2 – 627434 * a + 11914106?*" or even such complex tasks as "*Let r(g) be the second derivative of 2 * g ** 3/3 – 21 * g ** 2/2 + 10 * g. Let z be r(7). Factor – z * s + 6 – 9 * s ** 2 + 0 * s + 6 * s ** 2*", which is mathematically involved, but doesn't require reasoning over or formalizing real-world situations and corresponds to a closed domain.

Elementary school mathematical Olympiad tasks are difficult simultaneously for neural and symbolic systems, while most of the other benchmarks favor either symbolic or neural approaches (or at least seem to favor). Apparently, passing IMO Grand Challenge requires much more sophisticated symbolic reasoning, which is not covered by elementary school math tasks, but passing the former will also not make the goal of solving the latter any closer. So, these are really different benchmarks.

Of course, there are also other tests, which don't suppose formalized math tasks as input. For example, GEOS [19] and ARIS [20] projects are closely related to Aristo, but GEOS is focused on answering geometry questions with supporting diagram information, while ARIS suggests dealing with elementary arithmetic problems. A typical example of the GEOS problem is the following (Fig. 2)

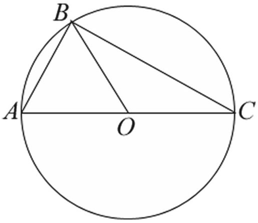

Fig. 2. In the figure, triangle ABC is inscribed in the circle with center O and diameter AC. If AB = AO, what is the degree measure of angle ABO?

The figures as well as the textual descriptions in GEOS are much more restricted, and their formal representation in terms of such predicates as Equals (AB, AO), IsTriangle (ABC), IsCenterOf (O, circle), Is AC, diameter) can be extracted (see the end-to-end geometry project solver[4]) rendering GEOS not too useful for testing AGI systems.

Here is one task from ARIS problem set: “Last week Tom had $74. He washed cars over the weekend and now has $86. How much money did he make washing cars?” It can be seen that the questions are concerned with very basic arithmetic, but the main challenge is to extract necessary information from the plain text description.

Another interesting initiative is the SemEval [21] project that provides a benchmark for testing AI abilities to pass high school Scholastic Achievement Tests (SAT). The dataset consists of 2200 training, 500 development, and 1000 test questions which were derived from Math SAT study guides. The question can have or have not some supplementary reference information presented in the form of a diagram.

Both ARIS and SemEval are similar to the elementary school mathematical Olympiad tasks in that the problem of understanding the task is more difficult than the problem of solving its formalized version. However, the ARIS and SemEval contain much more standard tasks of not too many types, which formalization is typically more straightforward, and which require much more restricted representations and simpler reasoning or problem solving capabilities. There are also other challenges and systems, which try to solve even more restricted forms of math problems and puzzles. One such system is LOGICIA [22], which is trying to deal with logic grid puzzles.

Some modifications to these benchmarks exist. For example, [https://www.aclweb.org/anthology/S17-1029.pdf] proposes to enrich the training set samples with detailed demonstrative solutions in natural language, but they also focus on SAT style geometry problems [23].

These are creativity, diversity, and originality of Olympiad tasks, which make them especially interesting from the AGI testing perspective in comparison to more restricted mathematical tests, which are good for advancing state-of-the-art models locally. Even if the training set is large enough, the process of solving tasks from the test set will not be routine. To see this, it is enough to try applying geometry SAT task solvers to the tasks like in Fig. 1). Consider also the following task as an example: using 6 matchsticks is it possible to create 4 equilateral triangles?

[4] http://geometry.allenai.org/.

Apparently, it is not yet another task on symbolic differentiation abundant both in test and training sets. It is unique and its only difficulty (even for humans) is to choose the correct solution space. A default formalization of this problem has no solution on the plane, but is easily solvable in 3D. An AI system that really understands natural language should not just represent coordinates of matchstick ends as points in 3D, but should consider 2D formalization also (what is about non-Euclidian spaces?), and even more, should consider points formed by intersections of matches, and should ask if it is allowed to break matches into pieces.

3 Conclusion

We have discussed (elementary) school mathematical Olympiad tasks as a rich source for (proto-)AGI systems benchmarking. The domain of these tasks is open-ended and diverse. Instead of requiring vast but shallow encyclopedic knowledge about facts and pair-wise relations, they require a more restricted amount of commonsense knowledge grounded in simulation or abstract models of reality. This renders memorization adopted by most DNN models not too useful (that is in agreement with a more general recent criticism of DNNs, e.g., in [24]).

The tasks under discussion require some creative reasoning, which may be non-trivial for elementary school pupils or even adults, but it is much less complex than what the existent automated theorem provers successfully deal with. The main problem here is to understand the task (e.g., but not necessarily, to adequately formalize it within some symbolic system), that is difficult for both neural and symbolic systems.

Thus, elementary school math tasks require a diverse set of cognitive skills are challenging for the existing AI systems, while manageable by young children without special training and extensive domain-specific knowledge.

Programming Olympiad tasks (as well as of other school subjects) can be used for a further extension of this idea. In fact, programming tasks highlight some issues even better. Indeed, it should be quite obvious that seq2seq models translating natural language descriptions to the code will be useless in open-ended domains unless language is somehow grounded in an interpreter (that gives real meaning to text tokens and symbols). However, programming tasks are more involved and don't replace math tasks, but extend them. They deserve a separate study in future work.

References

1. Hernández-Orallo, J., Minaya-Collado, N.: A formal definition of intelligence based on an intensional variant of Kolmogorov complexity. In: Proceedings of the International Symposium of Engineering of Intelligent Systems (EIS 1998), pp. 146–163. ICSC Press (1998)
2. Goertzel, B.: Artificial general intelligence: concept, state of the art, and future prospects. J. Artif. Gen. Intell. **5**(1), 1–48 (2014)
3. Hernández-Orallo, J., Martínez-Plumed, F., Schmid, U., Siebers, M., Dowe, D.L.: Computer models solving intelligence test problems: progress and implications. Artif. Intell. **230**, 74–107 (2016)

4. Agrawal, A., et al.: Don't just assume; look and answer: overcoming priors for visual question answering. In: Proceedings of IEEE Conference on CVPR, pp. 4971–4980 (2018)
5. Johnson, J., et al.: CLEVR: a diagnostic dataset for compositional language and elementary visual reasoning. arXiv preprint arXiv:1612.06890 (2016)
6. Chesani, F., Mello, P., Milano, M.: Solving mathematical puzzles: a challenging competition for AI. AI Mag. **38**(3), 83–94 (2017)
7. Ackerman, E.: Can winograd schemas replace turing test for defining human-level AI? IEEE Spectrum (2014)
8. Wang, A., et al.: GLUE: a multi-task benchmark and analysis platform for natural language understanding. arXiv preprint arXiv:1804.07461 (2018)
9. Clark, P.: Elementary school science and math tests as a driver for AI: take the Aristo challenge! In: Twenty-Seventh IAAI Conference (2015)
10. Clark, P., et al.: From 'F' to 'A' on the N.Y. regents science exams: an overview of the aristo project. arXiv preprint arXiv:1909.01958 (2019)
11. Wang, A.: SuperGLUE: a stickier benchmark for general-purpose language understanding systems. arXiv preprint arXiv:1905.00537 (2019)
12. Hudson, D.A., Manning, Ch.D.: GQA: a new dataset for real-world visual reasoning and compositional question answering. arXiv preprint arXiv:1902.09506 (2019)
13. Weitnauer, E., Ritter, H.: Physical bongard problems. In: Iliadis, L., Maglogiannis, I., Papadopoulos, H. (eds.) AIAI 2012. IAICT, vol. 381, pp. 157–163. Springer, Heidelberg (2012). https://doi.org/10.1007/978-3-642-33409-2_17
14. Klenk, M., Forbus, K.: Analogical model formulation for transfer learning in AP physics. Artif. Intell. **173**(18), 1615–1638 (2009)
15. Sloman, A.: Kantian philosophy of mathematics and young robots. In: Autexier, S., Campbell, J., Rubio, J., Sorge, V., Suzuki, M., Wiedijk, F. (eds.) CICM 2008. LNCS (LNAI), vol. 5144, pp. 558–573. Springer, Heidelberg (2008). https://doi.org/10.1007/978-3-540-85110-3_45
16. Saxton, D., Grefenstette, E., Hill, F., Kohli, P.: Analysing mathematical reasoning abilities of neural models. In: International Conference on Learning Representations (2019). https://openreview.net/forum?id=H1gR5iR5FX
17. Graves, A., et al.: Hybrid computing using a neural network with dynamic external memory. Nature **538**(7626), 471–476 (2016)
18. Schlag, I., et al.: Enhancing the transformer with explicit relational encoding for math problem solving. arXiv preprint arXiv:1910.06611 (2019)
19. Seo, M., et al.: Solving geometry problems: combining text and diagram interpretation. In: Proceedings Conference on Empirical Methods in Natural Language Processing, pp. 1466–1476 (2015)
20. Hosseini, M., Hajishirzi, H., Etzioni, O., Kushman, N.: Learning to solve arithmetic word problems with verb categorization. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 523–533 (2014)
21. Hopkins, M., et al.: SemEval 2019 task 10: math question answering. In: Proceedings of the 13th International Workshop on Semantic Evaluation (SemEval-2019), pp. 893–899 (2019)
22. Mitra, A., Baral, C.: Learning to automatically solve logic grid puzzles. In: Proceedings Conference on Empirical Methods in Natural Language Processing, pp. 1023–1033 (2015)
23. Sachan, M., Xing, E.: Learning to solve geometry problems from natural language demonstrations in textbooks. In: Proceedings of the 6th Joint Conference on Lexical and Computational Semantics, pp. 251–261 (2017)
24. Marcus, G.: The next decade in AI: four steps towards robust artificial intelligence. arXiv:2002.06177 (2020)