From *F* to *A* on the New York Regents Science Exams — An Overview of the Aristo Project

Peter Clark, Oren Etzioni, Daniel Khashabi, Tushar Khot, Bhavana Dalvi Mishra, Kyle Richardson, Ashish Sabharwal, Carissa Schoenick, Oyvind Tafjord, Niket Tandon, Sumithra Bhakthavatsalam, Dirk Groeneveld, Michal Guerquin, Michael Schmitz

Artificial intelligence has achieved remarkable mastery over games such as Chess, Go, and poker, and even Jeopardy!, but the rich variety of standardized exams has remained a landmark challenge. Even as recently as 2016, the best artificial intelligence system could only achieve 59.3 percent on an eighth-grade science exam (Schoenick et al. 2017). This article reports success on the Grade 8 New York Regents Science Exam, where, for the first time, a system scores more than ninety percent on the exam's non-diagram, multiple-choice questions. In addition, our Aristo system, building upon the success of recent language models, exceeded eighty-three percent on the corresponding Grade 12 Science Exam's non-diagram, multiple-choice questions. The results, on unseen test questions, are robust across different test years and different variations of this kind of test. They demonstrate that modern natural language processing methods can result in mastery on this task. While not a full solution to general question answering (the questions are limited to eighth-grade multiple-choice science), it represents a significant milestone for the field.

In 2014, Project Aristo was launched with the goal of reliably answering grade-school science questions, a stepping stone in the quest for systems that understood and could reason about science. The Aristo goal was highly ambitious, with the initial system scoring well below fifty percent even on fourth-grade multiple-choice tests. With a glance at the questions, it is easy to see why — the questions are hard. For example, consider the following eighth-grade question:

How are the particles in a block of iron affected when the block is melted?

(A) The particles gain mass.
(B) The particles contain less energy.
(C) The particles move more rapidly. [correct]
(D) The particles increase in volume.

This question is challenging, as it requires both scientific knowledge (particles move faster at higher temperatures) and common-sense knowledge (melting involves raising temperature), and the ability to combine this information appropriately.

Now, six years later, we are able to report that Aristo recently surpassed ninety percent on multiple-choice questions from the Grade 8 New York Regents Science Exam, a major milestone and a reflection on the tremendous progress of the natural language processing (NLP) community as a whole. In this article, we review why this is significant, how Aristo was able to achieve this score, and where the system still makes mistakes. We also explore what kinds of

reasoning Aristo appears to be capable of, and what work still needs to be done to achieve the broader goals of the project.

Why is this an important achievement? First, passing standardized tests has been a challenging artificial intelligence (AI) benchmark for many years (Bringsjord and Schimanski 2003; Brachman et al. 2005; Strickland 2013). A good benchmark should test a variety of capabilities while also being clearly measurable, understandable, accessible, nongameable, and sufficiently motivating. Standardized tests, while not a full test of machine intelligence, meet many of these practical requirements (Clark and Etzioni 2016). They also appear to require several capabilities associated with intelligence, including language understanding, reasoning, and common sense — although the extent to which such skills are needed is controversial (Davis 2014). We explore this in more detail this article.

Second, although NLP has made dramatic advances in recent years with the advent of large-scale language models such as Embeddings from Language Models (Peters et al. 2018), Bidirectional Encoder Representations from Transformers (BERT; Devlin et al. 2018), and Robustly optimized BERT (RoBERTa; Liu et al. 2019), many of the demonstrated successes have been on internal yardsticks generated by the AI/NLP community itself, such as the Stanford Question Answering Dataset (Rajpurkar et al. 2016), the General Language Understanding Evaluation dataset (Wang et al. 2019), and the TriviaQA dataset (Joshi et al. 2017). In contrast, the eighth-grade science exams are an external, independently generated benchmark where we can compare machine performance with human performance. Aristo thus serves as a poster child for the remarkable and rapid advances achieved in NLP, applied to an easily accessible task.

Finally, Aristo makes steps toward the AI Grand Challenge of a system that can read a textbook chapter and answer the questions at the end of the chapter. This broader challenge dates back to the 1970s, and was reinvigorated in Raj Reddy's 1988 Association for the Advancement of Artificial Intelligence Presidential Address and subsequent writing (Reddy 1988, 2003). However, progress on this challenge has a checkered history. Early attempts side-stepped the natural language understanding (NLU) task, in the belief that the main challenge lay in problem solving (for example, Larkin et al. 1980). In recent years there has been substantial progress in systems that can find factual answers in text, starting with IBM's Watson system (Ferrucci et al. 2010), and now with high-performing neural systems that can answer short questions provided they are given a text that contains the answer (for example, Seo et al. 2016; Wang, Yan, and Wu 2018). Aristo continues along this trajectory, but aims to also answer questions where the answer may not be written down explicitly. While not a full solution to the textbook grand challenge, Aristo is a further step along this path.

At the same time, care is needed in interpreting Aristo's results. We make no claims that Aristo is answering questions in the way a person would (and is likely using different methods). Exams are designed with human reasoning in mind, to test certain human knowledge and reasoning skills. But if the computer is answering questions in a different way, to what extent does it possess such skills? (Davis 2014). To explore this, we examine the causes of some of Aristo's failures, and test whether Aristo has some of the semantic skills that appear necessary for good performance. We find evidence of several types of such systematic behavior, suggesting that some form of reasoning is occurring, albeit not perfectly. Although still quite distant from human problem-solving, these emergent semantic skills are likely a key contributor to Aristo's scores reaching the 90-percent range.

As a brief history, the metric progress of the Aristo system on the Grade 8 Regents exams (nondiagram, multiple-choice part, for a hidden, held-out test set) is shown in figure 1. The figure shows the variety of techniques attempted, and mirrors the rapidly changing trajectory of the NLP field in general. Early work was dominated by information retrieval, statistical, and automated rule extraction and reasoning methods (Clark et al. 2014, 2016; Khashabi et al. 2016, 2018; Khot, Sabharwal, and Clark 2017). Later work has harnessed state-of-the-art tools for large-scale language modeling and deep learning (Tandon et al. 2018; Trivedi et al. 2019), which have come to dominate the performance of the overall system and reflects the stunning progress of the field of NLP as a whole.

Finally, it is particularly fitting to report this result in the *AI Magazine,* as it is another step in the decades-long quest to fulfill the late Paul Allen's dream of a Digital Aristotle, an "easy-to-use, all-encompassing knowledge storehouse … to advance the field of AI" (Allen 2012), a dream also set out in the *AI Magazine* in the Winter 2004 issue (Friedland et al. 2004). Aristo's success reflects how much progress the field of NLP and AI has made in the intervening years.

The Aristo System

Aristo is comprised of eight solvers, each of which attempts to independently answer a multiple-choice question. Its suite of solvers has changed over the years, with new solvers being added and redundant solvers being dropped to maintain a simple architecture. (Earlier solvers included use of Markov logic networks [Khot et al. 2015], reasoning over tables [Khashabi et al. 2016], and other neural approaches; these have been superseded by the language models.) As illustrated in figure 2, they can be loosely grouped into statistical and information-retrieval methods, reasoning methods, and large-scale language model methods.

We now briefly describe these solvers, with pointers to further information. Over the life of the project, the relative importance of the methods has shifted

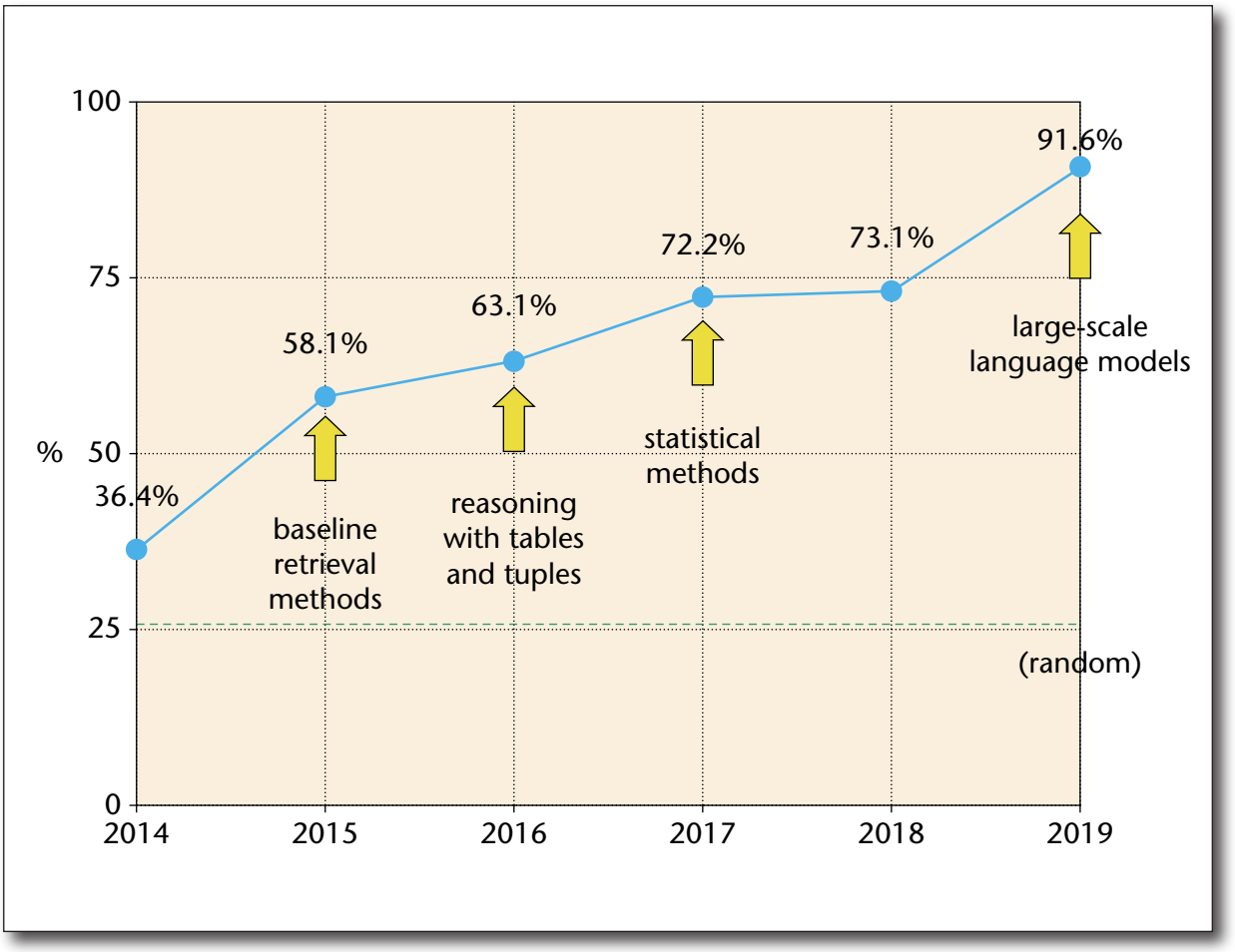

Figure 1. Progress over Time of Aristo's Scores on Regents 8th Grade Science Test.

(Nondiagram, multiple choice questions, held-out test set).

toward large-scale language methods, which now dominate the overall performance of the system.

Information Retrieval and Statistics

The information retrieval solver searches to see if the question-along-with-an-answer option is explicitly stated in the corpus. To do this, for each answer option a_i it sends $q + a_i$ as a query to a search engine, and returns the search engine's score for the top retrieved sentence s. This is repeated for all options a_i to score them all, and the option with the highest score is selected (Clark et al. 2016).

The Pointwise Mutual Information (PMI) solver uses pointwise mutual information (Church and Hanks 1989) to measure the associations between parts of q and parts of a_i. The PMI for two n-grams x and y is defined as

$$PMI(x,y) = \log \frac{p(x,y)}{p(x)p(y)}$$

The larger the PMI, the stronger the association between x and y. The solver extracts unigrams, bigrams, trigrams, and skip-bigrams, and outputs the answer with the largest average PMI, calculated over all pairs of the question-and-answer option n-grams (Clark et al. 2016).

Finally, the abstract–concrete mapping engine searches for a cohesive link between a question q and candidate answer a_i using a large knowledge base of vector spaces that relate words in language to a set of 5,000 scientific terms enumerated in a term bank. The key insight in the abstract–concrete mapping engine is that we can better assess lexical cohesion of a question and its answer by pivoting through scientific terminology, rather than by simple co-occurrence frequencies of question-and answer-words (Turney 2017).

Reasoning Methods

The TupleInference solver uses semistructured knowledge in the form of tuples, extracted via open information extraction (Banko et al. 2007). TupleInference treats the reasoning task as searching for a graph that best connects the terms in the question with an answer choice via the knowledge; see figure 3 for a simple illustrative example. To find the best support graph for each answer option, we define

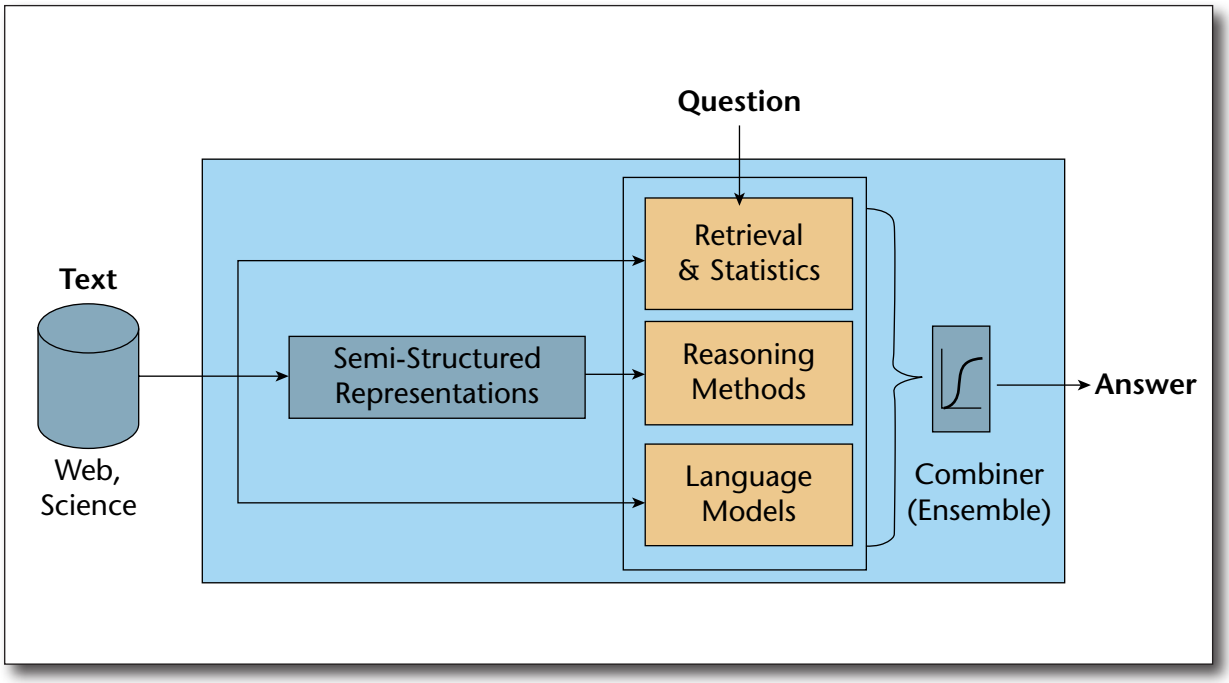

Figure 2. A Simplified Picture of Aristo's Architecture.

the task as an optimization problem, and use integer linear programming to solve it. The answer choice with the highest scoring graph is then selected (Khot, Sabharwal, and Clark 2017).

Multee (Trivedi et al. 2019) is a solver that repurposes existing textual entailment tools for question answering. Textual entailment is the task of assessing if one text implies another, and there are several high-performing textual entailment systems now available. Multee learns to combine their decisions, so it can determine how strongly a set of retrieved texts entails a particular question-and-answer option (Trivedi et al. 2019).

The qualitative reasoning solver is designed to answer questions about qualitative influence, that is, how more or less of one quantity affects another. Unlike the other solvers in Aristo, it is a specialist solver that only fires for a small subset of questions that ask about qualitative change. The solver uses a knowledge base of 50,000 (textual) statements about qualitative influence, such as "A sunscreen with a higher SPF protects the skin longer." It has then been trained to reason with the BERT language model (Devlin et al. 2018), using a similar approach to that described below (Tafjord et al. 2019).

Large-Scale Language Models

The field of NLP has advanced substantially with the advent of large-scale language models such as Embeddings from Language Models (Peters et al. 2018), BERT (Devlin et al. 2018), and RoBERTa (Liu et al. 2019). The AristoBERT solver applies BERT to multiple-choice questions by treating the task as one of classification: Given a question q with answer options a_i and optional background knowledge K_i, we provide it to BERT as

$$[CLS]\ K_i\ [SEP]\ q\ [SEP]\ a_i\ [SEP]$$

for each option a_i. The [CLS] output token is projected to a single logit and fed through a softmax layer across answer options, trained using cross entropy loss, and then the highest scoring option is selected.

AristoBERT uses three methods to apply BERT more effectively. First, we retrieve and supply background knowledge K_i along with the question when using BERT, as described above. This provides the potential for BERT to "read" that background knowledge and apply it to the question, although the exact nature of how it uses background knowledge is more complex and less interpretable. Second, following Sun et al. (2019), we fine-tune BERT using a curriculum of several datasets, starting with ReAding Comprehension from Examinations, or RACE (a general reading comprehension dataset that is not science-related; Lai et al. 2017), followed by a collection of science training sets: OpenbookQA (Mihaylov et al. 2018), AI2 Reasoning Challenge dataset (ARC) (Clark et al. 2018), and Regents questions (training partition). Finally, we repeat this for three variants of BERT (cased, uncased, and cased whole-word), and assemble the predictions.

Finally, the AristoRoBERTa solver does the same with RoBERTa (Liu et al. 2019), a high-performing

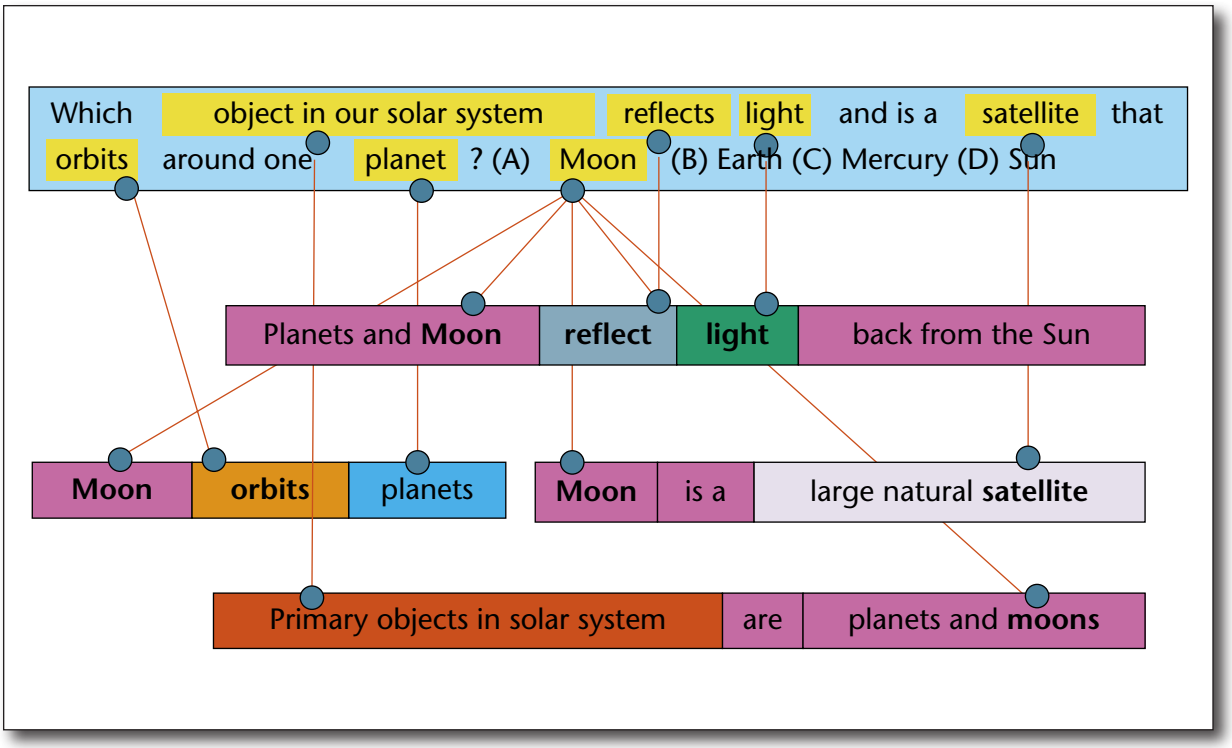

Figure 3. Support Graph for Choice A, as Constructed by the TupleInference Solver.

and optimized derivative of BERT trained on significantly more text. In AristoRoBERTa, we simply replace the BERT model in AristoBERT with RoBERTa, repeating similar fine-tuning steps. We ensemble two versions together, namely with and without the first fine-tuning step using RACE.

Experiments and Results

We apply Aristo to the nondiagram, multiple-choice (NDMC) questions in the science exams. Although questions with diagrams are common,[1] they are outside of our focus on language and reasoning. (For illustrative work on science diagrams, see Krishnamurthy, Tafjord, and Kembhavi 2016.) We also omit questions that require a direct answer; this is for two reasons. The first is that, after removing the questions with accompanying diagrams, such questions are statistically rare; for example, of the 482 direct-answer questions over thirteen years of Grade 8 Regents Science Exams, only thirty eight (less than eight percent) do not involve a diagram. The second reason is that they are complex; they often require explanation and synthesis. Both diagram and direct-answer questions form natural topics for future work.

We evaluate Aristo using the New York Regents Science exam questions,[2] and the ARC dataset, a larger corpus of science questions drawn from public resources across the country (Clark et al. 2018). The Regents exams are only produced for fourth-, eighth-, and twelfth-grade students (corresponding to the end of elementary, middle, and high-school, respectively), while the ARC questions span grades three to nine. All questions are posed exactly as written, with no rewording or rephrasing. The entire dataset is partitioned into train/development (dev)/test parts (table 1), and for the Regents questions we ensure that each exam is completely in *train*, *dev*, or *test*, and not split among them. (The non-Regents ARC questions do not have exam groupings.) All but thirty nine of the 9,366 questions are four-way multiple choice, the remaining thirty nine (less than 0.5 percent) being three- or five-way. A random score over the entire dataset is 25.02 percent.

Results

The results are summarized in figure 4, showing the performance of the solvers individually, and their combination in the full Aristo system. Note that Aristo is a single system run on the five datasets (not retuned for each dataset in turn).

Most notably, Aristo's scores on the Regents Exams far exceed earlier performances (for example, Clark et al. 2016; Schoenick et al. 2017), and represent a new high-point on science questions.

In addition, the results show the dramatic impact of new language modeling technology, as embodied in AristoBERT and AristoRoBERTa; the scores for these two solvers dominate the performance of

Dataset	Partition			
	Train	Development	Test	Total
Regents 4th Grade	127	20	109	256
Regents 8th Grade	125	25	119	269
Regents 12th Grade	665	282	632	1,579
ARC-Easy	2,251	570	2,376	5,197
ARC-Challenge	1,119	299	1,172	2,590
Totals[1]	4,035	1,151	4,180	9,366

Table 1. Dataset Partition Sizes, Showing Number of Questions.[3]

the overall system. Even on the ARC-Challenge questions, containing a wide variety of difficult questions, the language-modeling–based solvers dominate. The general increasing trend of solver scores from left to right for each test-set loosely reflects the progression of the NLP field over the six years of the project.

To further check that we have not overfit to our data, we also ran Aristo on the most recent years of the Regents Exams (fourth and eighth grade), years 2017–19, which were not available at the start of the project and were not part of our datasets. We find similar scores (average 92.8 percent for the three fourth-grade exams, 93.3 percent for the eighth-grade), suggesting the system is not overfit.

On a combination of exam scores and laboratory work (weighted approximately 60:40), the NY State Education Department considers an overall score of sixty-five percent as "Meeting the Standards," and over eighty-five percent as "Meeting the Standards with Distinction."[4] As a somewhat loose comparison, if this rubric applies equally to the NDMC subset we have studied, this would mean Aristo has met the standard with distinction in Grade 8 NDMC Science (although clearly Full Science requires substantially more).

Answer-Only Performance

Several authors have observed that for some multiple-choice datasets, systems can still perform well even when ignoring the question body and looking only at the answer options (Gururangan et al. 2018; Poliak et al. 2018b). This surprising result is particularly true for crowdsourced datasets, where workers may use stock words or phrases (for example, *not*) in incorrect answer options that give them away. To measure this phenomenon on our datasets, we trained and tested a new AristoRoBERTa model giving it only the answer options — that is, no question body nor retrieved knowledge. (Without retraining the model scores slightly less, thirty-five percent overall versus thirty-eight percent with retraining.) The results (test set) are shown in figure 5, indicating that it is hard to select the right answer without reading the question. (Scores are slightly higher for twelfth-grade answer-only, possibly because the average answer length is longer, hence more potential for hidden patterns inside that hint at correctness or incorrectness.)

Adversarial Answer Options

What if we add extra incorrect answer options to the questions? If a system has mastery of the material, we would expect its score to be relatively unaffected by such modifications. We can make this more challenging by doing this adversarially: try many different incorrect options until the system is fooled. If we do this, turning a four-way MC question into eight-way with options chosen to fool Aristo, then retrain on this new dataset, we do observe an effect: the scores drop, although the drop is small (approximately ten percent); see figure 6. This indicates that while Aristo performs well, it still has some blind spots that can be artificially uncovered through adversarial methods such as this.

Analysis

Despite the high scores, Aristo still makes occasional mistakes. Because Aristo retrieves and then "reads" corpus sentences to answer a question, we can inspect the retrieved knowledge when Aristo fails, and gain some insight as to where and why it makes errors. Did Aristo retrieve the *right* knowledge, but then choose the *wrong* answer? Or was the failure due (in part) to the retrieval step itself? We manually analyzed 30 random failures (of 248) in the entire dev set (Regents + ARC, 1,151 dev set questions total), and found four main categories of failures, illustrated in figure 7, that we now summarize. As the language model solvers have highest weight in Aristo, we conduct this analysis for failures by AristoRoBERTa, but note these very

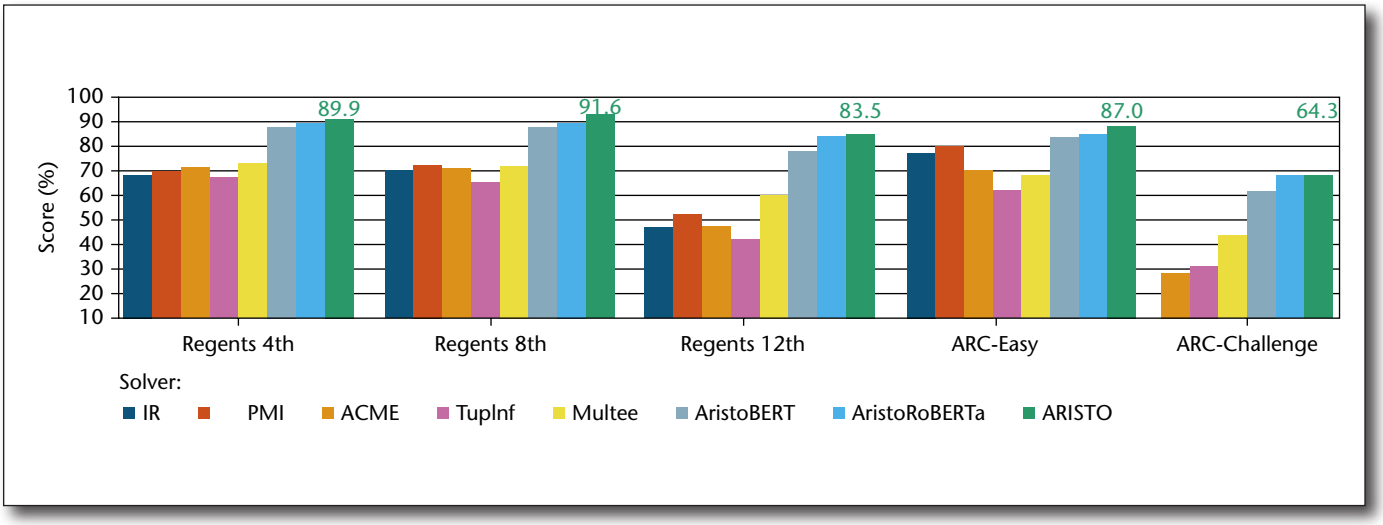

Figure 4. The Results of Each of the Aristo Solvers, as Well as the Overall Aristo System, on Each of the Test Sets.

Most notably, Aristo achieves 91.6-percent accuracy in 8th grade and exceeds 83 percent in 12th grade. Note that Aristo is a single system, run unchanged on each dataset (not retuned for each dataset).

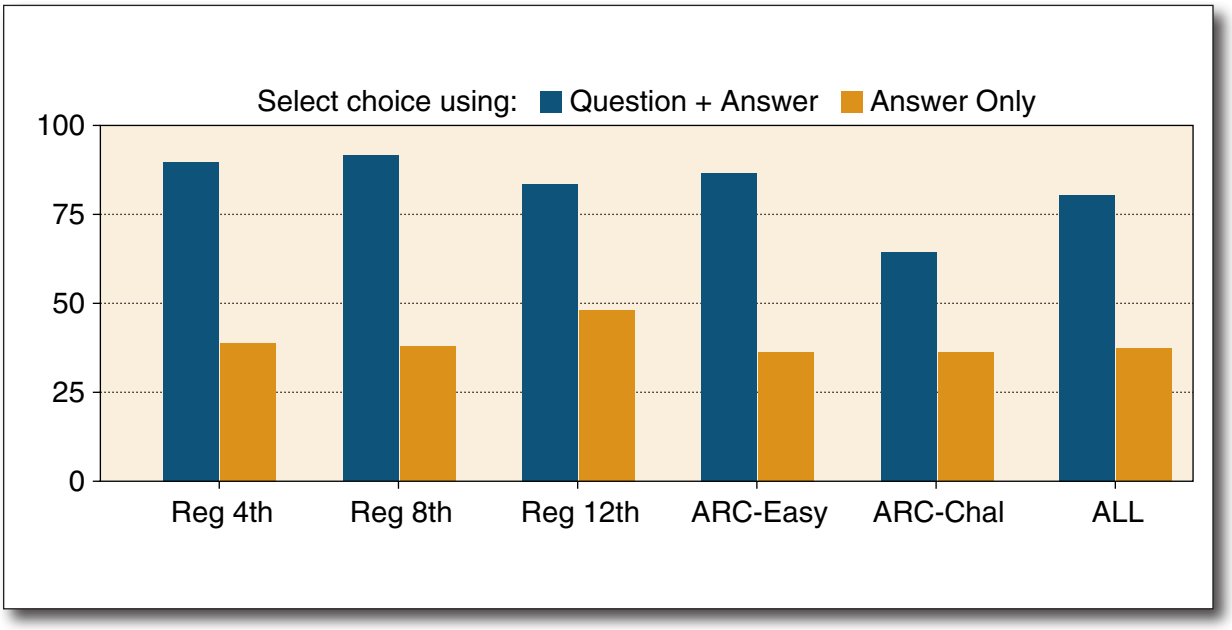

Figure 5. Scores When Looking at the Answer Options Only, Compared with Using the Full Questions.

The (desirably) low scores and large drops indicate it is hard to guess the answer without reading the question.

frequently (approximately ninety percent of the time) equate to overall Aristo failures, and that when AristoRoBERTa fails, most (on average, seventy-six percent) of the other solvers also fail. We did not discern any systematic patterns within these.

Good Support for Correct Answer (13 Percent)

Surprisingly, only four of the thirty failures were cases where the retrieved knowledge supported the right answer option, but Aristo chose a wrong answer option. An example is:

Which is the best unit to measure distances between Earth and other solar systems in the universe?

(A) miles
(B) kilometers
(C) light years [*correct*]
(D) astronomical units [*selected*]

Here, although Aristo did retrieve good evidence for the correct answer (C), namely

> Distances between Earth and the stars are often measured in terms of light-years

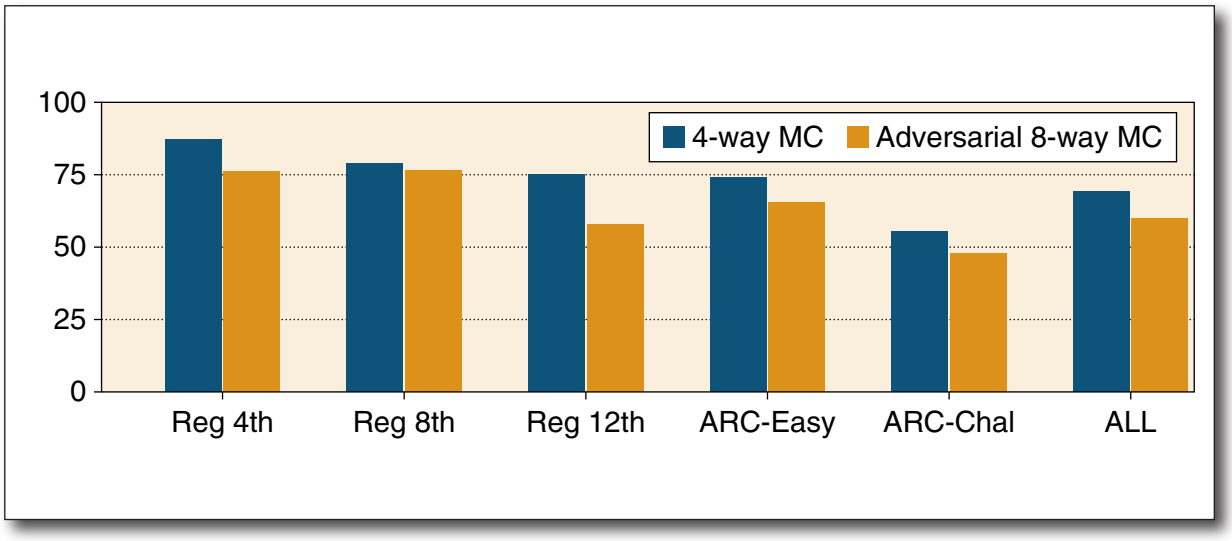

Figure 6. Aristo's Scores Drop a Small Amount (Average 10 Percent) When Tested on Adversarially Generated Eight-Way Multiple-Choice Questions.

It still preferred the incorrect option (D) from the retrieved knowledge:

> In general, distances in the solar system are measured in astronomical units.

Here, Aristo has confused distinguishing distances *within* the solar system versus distances *between* solar systems (a confusion that a human might easily make too). This illustrates where Aristo has misapplied its retrieved knowledge. However, such cases appear to be rare (four out of thirty). In other words, for the vast majority of questions, if suitable knowledge is retrieved, then Aristo will answer correctly.

No Support for the Correct Answer (Fifty-Seven Percent)

The largest cause of failure was simply when none of the retrieved sentences provide evidence for the correct answer. In such situations, Aristo has little chance of answering correctly. For example:

Although they belong to the same family, an eagle and a pelican are different. What is one difference between them?
(A) their preference for eating fish
(B) their ability to fly
(C) their method of reproduction [*selected*]
(D) their method of catching food [*correct*]

As there are no corpus sentences comparing eagles and pelicans, Aristo retrieves a rather random collection of unhelpful facts. Instead, what is needed here is to realize that this is a comparison question, retrieve appropriate facts for pelicans and eagles individually and then compare them, such as by using question-decomposition methods (Wolfson et al. 2020).

Reading Comprehension (Twenty-Seven Percent)

In the exams, there are a few reading comprehension questions that primarily require reasoning over the question content itself, rather than retrieving and applying science knowledge. In such situations, retrieved knowledge is unlikely to be helpful. Eight out of thirty failures fell into this category. One example is a question describing an experiment:

A student wants to determine the effect of garlic on the growth of a fungus species. Several samples of fungus cultures are grown in the same amount of agar and light. Each sample is given a different amount of garlic. What is the independent variable in this investigation?
(A) amount of agar
(B) amount of light
(C) amount of garlic [*correct*]
(D) amount of growth [*selected*]

Here, the answer is unlikely to be written down in a corpus, as a novel scenario is being described. Rather, it requires understanding the scenario itself.

A second example is a metaquestion about sentiment:

Which statement is an opinion?
(A) Many plants are green
(B) Many plants are beautiful [*correct*]
(C) Plants require sunlight [*selected*]
(D) Plants can grow in different places

Again, retrieval is unlikely to help here. Rather, the question asks for an analysis of the options themselves, something Aristo does not realize.

Good Support for an Incorrect Answer (Three Percent)

Occasionally a failure occurs due to retrieved knowledge supporting an incorrect answer, for example, if the question is ambiguous, or the retrieved knowledge is wrong. The single failure in this category that we observed was:

Which of these objects will most likely float in water?
(A) glass marble
(B) steel ball
(C) hard rubber ball [*selected*]
(D) table tennis ball [*correct*]

Here, Aristo retrieved evidence for both (C) and (D), for example, for (C), Aristo's retrieval included: "It had like a rubber ball in it, which would maybe float up" here leading Aristo to select the wrong answer. Arguably, as this question is a comparative *(which most likely floats?),* Aristo should have rejected this in favor of the correct answer (*table tennis ball*). However, as Aristo computes a confidence for each option independently, it is unable to directly make these cross-option comparisons.

Other

Finally, we point to one other interesting failure:

About how long does it take for the Moon to complete one revolution around Earth?
(A) 7 days
(B) 30 days [*correct*]
(C) 90 days
(D) 365 days [*selected*]

In this case, many relevant sentences were retrieved, including:

> Because it takes the moon about 27.3 days to complete one orbit around the Earth …
> It takes 27.3 days for the moon to complete one revolution around the earth.
> The Moon completes one revolution around the Earth in 27.32166 days.

However, Aristo does not realize 27.3 is *about* 30, and hence answered the question incorrectly.

A Score Card for Aristo's Semantic Skills

From informal tests, Aristo appears to be doing more than simply matching a question-and-answer option to a retrieved sentence. Rather, Aristo appears to recognize various linguistic and semantic phenomena, and respond appropriately. For example, if we add negation (a *not*) to the question, Aristo almost always correctly changes its answer choice. Similarly, if we replace *increase* with *decrease* in a question, Aristo will typically change its answer choice correctly, suggesting it has some latent knowledge of qualitative direction.

To quantify such skills more systematically, we performed five sets of tests on Aristo without fine-tuning Aristo on those tests; that is, the tests are zero-shot. (From other experiments, we know that if we train Aristo on these tests it can perform them almost perfectly, but our interest here is how Aristo performs "out of the box" after training on the science exams). Each test probes a different semantic phenomenon of interest, as we now describe.

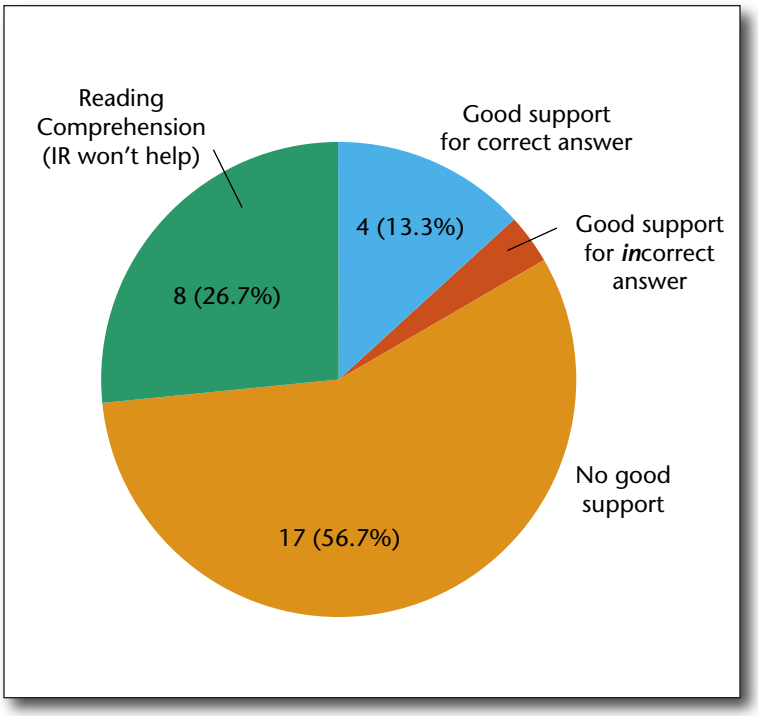

Figure 7. Case Study.

Causes of error for 30 questions that Aristo answered incorrectly.

Negation

How well does Aristo handle negation? As a (limited) test, we generated a synthetic negation dataset (10,000 questions), where each question has a synthetic context (replacing the retrieved sentences), plus a question about it, for example:

Context:
Alan is small. Alan is tall. Bob is big. Bob is tall. Charlie is big. Charlie is tall. David is small. David is short.

Question:
Which of the following is not tall?
(A) Alan
(B) Bob
(C) Charlie
(D) David [*correct*]

We then test Aristo on this dataset without fine-tuning on it. Remarkably, Aristo scores ninety-four percent on this dataset, suggesting at least in this particular formulation, Aristo has an understanding of *not.*

Conjunction

We test conjunction in a similar way, with questions such as:

Context:
Alan is red. Alan is big. Bob is blue. Bob is small. Charlie is blue. Charlie is big. David is red. David is small.

Question:
Which of the following is big and blue?
(A) Alan
(B) Bob
(C) Charlie [*correct*]
(D) David

With questions containing two conjuncts (for example, the one above), and again without any fine-tuning on this data, Aristo scores ninety-eight percent. If we increase the number of conjuncts in the question to three, four, and five, Aristo scores ninety-five percent, ninety-four percent, and eighty percent, respectively. If we use five conjuncts and a negation, for example:

Context:
Alan is red. Alan is big. Alan is light. Alan is old. Alan is tall. Bob is red. Bob is small. Bob is heavy. Bob is old. Bob is tall. Charlie is blue. Charlie is big. Charlie is light. Charlie is old. Charlie is tall. David is red. David is small. David is heavy. David is young. David is tall.

Question:
Which of the following is old and red and light and big and not short?
(A) Alan (correct)
(B) Bob
(C) Charlie
(D) David

Answer:
Aristo remarkably still scores seventy-five percent. Although not perfect, this indicates some form of systematic handling of conjunction-plus-negation is occurring.

Polarity

Polarity refers to Aristo's ability to correctly change its answer when a comparative in the question is "flipped." For example, given:

Which human activity will likely have a negative effect on global stability?
(A) decreasing water pollution levels
(B) increasing world population growth [*correct*]

If we now switch *negative* to *positive,* Aristo should switch its answer from (B) to (A). To score a point, Aristo must get both the original question and the flipped question (with a changed answer) correct. To measure this, we use an existing qualitative dataset containing such pairs, called QuaRTz (Tafjord et al. 2019). As the QuaRTz questions are two-way multiple-choice, a random score for getting both right would be twenty-five percent. Remarkably, we find Aristo scores 67.1 percent (again with no fine-tuning on QuaRTz), suggesting Aristo has some knowledge about the polarity of comparatives. Note that this test also requires Aristo to get the original question right in the first place, thus the score reflects both knowledge and polarity reasoning, a harder task than polarity alone.

Factuality

Event factuality refers to whether an event, mentioned in a textual context, did in fact occur (Rudinger, White, and Durme 2018). For example:

If someone regretted that a particular thing happened
(A) that thing might or might not have happened
(B) that thing didn't happen
(C) that thing happened [*correct*]

Predicting factuality requires understanding what the context around an event implies about that event's occurrence. We tested Aristo on this task using the veracity question dev set from the Diverse Natural Language Inference Collection (Poliak et al. 2018a), converted to multiple-choice format (394 questions). On this task, Aristo scored 66.5 percent, again suggesting Aristo has some knowledge of how words affect the factiveness of the events that they modify.

Counting

Finally, we ran Aristo on bAbI task 7, a simple form of counting (Weston et al. 2015), converted to multiple choice with four options, as below. For example:

Daniel picked up the football. Daniel dropped the football. Daniel got the milk. How many objects is Daniel holding?
(A) zero
(B) one (correct)
(C) two
(D) three

Aristo (again not fine-tuned on this dataset) did badly at this task, scoring only six percent.[5] This result is perhaps not surprising, as this type of reasoning is not exemplified in any way in an any of Aristo's training data.

Scorecard

We can informally map these scores to a grade level to give Aristo a score card (figure 8). The most striking observation is that Aristo passes all but counting, and has apparently acquired these skills through its general fine-tuning on RACE and Science Exams, with no fine-tuning at all on these specific probing datasets. Aristo does appear to be doing more than just sentence matching, but not quite in the systematic way a person would. These acquired latent skills are reflected in the high scores Aristo achieves on the Science Exams.

Discussion

What can we conclude from this? Most significantly, Aristo has achieved surprising success on a formidable problem, in particular by leveraging large-scale language models. The system thus serves as a demonstration of the stunning progress that NLP technology has made in the last two years.

At the same time, exams themselves are an imperfect test of understanding science, and, despite their many useful properties, are also only a partial test of machine

intelligence (Davis 2014). Earlier, we highlighted several classes of problems Aristo does not handle well, even though its exam scores are high; these are questions requiring diverse pieces of evidence to be combined, such as reading comprehension (story) questions, metaquestions, and performance of arithmetic. Davis (2016) has similarly pointed out that as standardized tests are authored for people, not machines, they also don't directly test for things that are easy for people, such as temporal reasoning, simple counting, and obviously impossible situations. These are problem types that Aristo is not familiar with and would be hard for it (but not for people). Science exams are just one of many different, partial indicators of progress in broader AI. Finally, we have only been using multiple-choice questions, a format that just requires ranking of answer choices, arguably allowing more use of weak evidence compared with (say) generating an answer, or even independently deciding if an answer is true or false (Clark et al. 2019).

On the other hand, we do see clear evidence of systematic semantic skill in Aristo. For example, Aristo not only answers this question correctly:

City administrators can encourage energy conservation by
(A) lowering parking fees
(B) decreasing the cost of gasoline
(C) lowering the cost of bus and subway fares [*correct, selected*]

but flipping *decreasing* and *lowering* causes it to correctly change its answer:

City administrators can encourage energy conservation by
(A) lowering parking fees
(B) ~~decreasing~~ increasing the cost of gasoline [*correct, selected*]
(C) ~~lowering~~ raising the cost of bus and subway fares

Our probes showed that such behavior is not just anecdotal but systematic, suggesting that *some* form of reasoning is occurring, but not in the traditional style of discrete symbol manipulation in a formally designed language (Brachman and Levesque 1985; Genesereth and Nilsson 2012). Other work has similarly found that neural systems can learn systematic behavior (Lake and Baroni 2018; Clark, Tafjord, and Richardson 2020), and these emergent semantic skills are a key contributor to Aristo's scores reaching the ninety-percent range. Large-scale language model architectures have brought a dramatic, new capability to the table that goes significantly beyond just pattern matching and similarity assessment.

2020 Report Card for		*Aristo*
Subject	Grade	Score
Negation	*A*	*94%*
Conjunction	*B+*	*80%-98%*
Polarity	*D+*	*67.1%*
Factuality	*D*	*66.5%*
Counting	*F*	*6%*

Figure 8. Aristo, With No Fine-Tuning, Passes Probes for All but Counting.

Summary and Conclusion

Answering science questions is a long-standing AI grand challenge (Reddy 1988; Friedland et al. 2004). We have described Aristo, the first system to achieve a score of over 90 percent on the NDMC part of the New York Regents Grade 8 Science Exam, demonstrating that modern NLP methods can result in mastery of this task. Although Aristo only answers multiple-choice questions without diagrams, and operates only in the domain of science, it nevertheless represents an important milestone toward systems that can read and understand. The momentum on this task has been remarkable, with accuracy moving from roughly sixty percent to over ninety percent in just three years. In addition, the use of independently authored questions from a standardized test allows us to benchmark AI performance relative to human students.

Beyond the use of a broad vocabulary and scientific concepts, many of the benchmark questions intuitively appear to require some degree of reasoning to answer. For many years in AI, reasoning was thought of as discrete symbol manipulation. With the advent of deep learning, this notion of reasoning has expanded, with systems performing challenging tasks using neural architectures rather than explicit representation languages. Similarly, we observe surprising performance on answering science questions, and on specific semantic phenomena such as those probed earlier. This suggests that the machine has indeed learned something about language and the world as well as how to manipulate that knowledge — albeit neither symbolically nor discretely.

Although an important milestone, this work is only a step on the long road toward a machine that has a deep understanding of science and achieves Paul Allen's original dream of a Digital Aristotle. A machine that has fully understood a textbook should not only be able to answer the multiple-choice questions at the end of the chapter, it should also be able to generate both short and long answers to

Related Work on Standardized Testing for AI

Standardized Tests

Standardized tests have long been proposed as challenge problems for AI (for example, Bringsjord and Schimanski 2003; Brachman et al. 2005; Piatetsky-Shapiro et al. 2006; Clark and Etzioni 2016), as they appear to require significant advances in AI technology while also being accessible, measurable, understandable, and motivating.

Earlier work on standardized tests focused on specialized tasks, for example, Stanford Achievement Test (or SAT) word analogies (Turney 2006), Graduate Record Examinations word antonyms (Mohammad et al. 2013), and Test of English as a Foreign Language synonyms (Landauer and Dumais 1997). More recently, there have been attempts at building systems to pass university entrance exams. Under the National Institute of Informatics' Todai project, several systems were developed for parts of the University of Tokyo Entrance Exam, including mathematics, physics, English, and history (Strickland 2013; Tainaka 2013; Fujita et al. 2014), although in some cases questions were modified or annotated before being given to the systems (for example, Matsuzaki et al. 2014). Similarly, a smaller project worked on passing the Gaokao (China's college entrance exam; Cheng et al. 2016; Guo et al. 2017). The Todai project was reported as ended in 2016, in part because of the challenges of building a machine that could "grasp meaning in a broad spectrum" (Mott 2016).

Mathematics Word Problems

Substantial progress has been achieved on math word problems. On plane geometry questions, Seo et al. (2015) demonstrated an approach that achieved a 61-percent accuracy on SAT practice questions. The Euclid system (Hopkins et al. 2017) achieved a 43-percent recall and 91-percent precision on SAT closed-vocabulary algebra questions, a limited subset of questions that nonetheless constitutes approximately 45 percent of a typical math SAT exam. Closed-vocabulary questions are those that do not reference real-world situations (for example, "what is the largest prime smaller than 100?" or "Twice the product of x and y is 8. What is the square of x times y?").

Work on open-world math questions has continued, but results on standardized tests have not been reported and thus it is difficult to benchmark the progress relative to human performance. See Amini et al. (2019) for a recent snapshot of the state of the art, and references to the literature on this problem.

direct questions; it should be able to perform constructive tasks such as designing an experiment for a particular hypothesis; it should be able to explain its answers in natural language and discuss them with a user; and it should be able to learn directly from an expert who can identify and correct the machine's misunderstandings. These are all ambitious tasks still largely beyond the current technology, but with the rapid progress happening in NLP and AI, solutions may arrive sooner than we expect.

Acknowledgments

We gratefully acknowledge the late Paul Allen's inspiration, passion, and support for research on this grand challenge. We also thank the many other contributors to Aristo, including Niranjan Balasubramanian, Matt Gardner, Peter Jansen, Jayant Krishnamurthy, Souvik Kundu, Todor Mihaylov, Harsh Trivedi, Peter Turney, and the Beaker team at Allen Institute for Artificial Intelligence (AI2), and to Ernie Davis, Gary Marcus, Raj Reddy, and many others for helpful feedback on this work. We also thank the anonymous reviewers for helpful comments that improved the article.

Notes

1. Ratios of NDMC, with diagram multiple-choice, nondiagram direct-answer, and with diagram direct-answer questions are approximately 45/25/5/25 for Regents 4th Grade, 25/25/5/45 for 8th Grade, and 40/25/15/20 for 12th Grade.

2. See www.nysedregents.org for the original exams.

3. ARC (Easy+Challenge) includes Regents 4th- and 8th-grade datasets as subsets.

4. www.nysedregents.org/grade8/science/618/home.html.

5. Lower than random guessing, because Aristo frequently selects option *D* (three), an option which is (by chance) very rarely the right answer in this dataset. *D* is likely chosen due to a small random bias toward *three*, and all questions looking stylistically similar. Note there is no training on this (nor other) probing datasets, so Aristo is unaware of the answer distribution.

References

Allen, P. 2012. *Idea Man: A Memoir by the Cofounder of Microsoft*. New York: Penguin

Amini, A.; Gabriel, S.; Lin, P.; Koncel-Kedziorski, R.; Choi, Y.; and Hajishirzi, H. 2019. MathQA: Towards Interpretable Math Word Problem Solving with Operation-Based Formalisms. *In 2019 Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Banko, M.; Cafarella, M. J.; Soderland, S.; Broadhead, M.; and Etzioni, O. 2007. Open Information Extraction from the Web. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence Press.

Brachman, R.; Gunning, D.; Bringsjord, S.; Genesereth, M.; Hirschman, L.; and Ferro, L. 2005. *Selected Grand Challenges in Cognitive Science*. MITRE Technical Report 05-1218. Bedford, MA: The MITRE Corporation.

Brachman, R. J., and Levesque, H. J. 1985. *Readings in Knowledge Representation*. San Francisco: Morgan Kaufmann Publishers.

Bringsjord, S., and Schimanski, B. 2003. What is Artificial Intelligence? Psychometric AI as an Answer, 887–93. In *Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence*. San Francisco: Morgan Kaufmann Publishers.

Cheng, G.; Zhu, W.; Wang, Z.; Chen, J.; and Qu, Y. 2016. Taking up the Gaokao Challenge: An Information Retrieval Approach, 2479–85. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence Press.

Church, K. W., and Hanks, P. 1989. Word Association Norms, Mutual Information and Lexicography. In *27th Annual Meeting of the Association for Computational Linguistics*, 76–83. Stroudsberg, PA: Association for Computational Linguistics.

Clark, C.; Lee, K.; Chang, M.-W.; Kwiatkowski, T.; Collins, M.; and Toutanova, K. 2019. BoolQ: Exploring the Surprising Difficulty of Natural Yes/No. In *Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Clark, P.; Balasubramanian, N.; Bhakthavatsalam, S.; Humphreys, K.; Kinkead, J.; Sabharwal, A.; and Tafjord, O. 2014. Automatic Construction of Inference-Supporting Knowledge Bases. Paper presented at the 4th Workshop on Automated Knowledge Base Construction, Montreal, Canada, Dec. 13. citeseerx.ist.psu.edu/viewdoc/download;jsessionid=60F1851D3476027472E83247F9875717?doi=10.1.1.697.9536&rep=rep1&type=pdf.

Clark, P.; Cowhey, I.; Etzioni, O.; Khot, T.; Sabharwal, A.; Schoenick, C.; and Tafjord, O. 2018. *Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge*. arXiv preprint: ArXiv abs/1803.05457. Ithaca, NY: Cornell University Library.

Clark, P., and Etzioni, O. 2016. My Computer Is an Honor Student — But How Intelligent Is It? Standardized Tests as a Measure of AI. *AI Magazine* 37(1): 5–12.

Clark, P.; Etzioni, O.; Khot, T.; Sabharwal, A.; Tafjord, O.; Turney, P. D.; and Khashabi, D. 2016. Combining Retrieval, Statistics, and Inference to Answer Elementary Science Questions, 2580–86. In *Proceedings of the Thirtieth Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Clark, P.; Tafjord, O.; and Richardson, K. 2020. Transformers as Soft Reasoners over Language. In *Proceedings of the Twenty-Ninth International Joint Conference on Artificial Intelligence*. Vienna, Austria: International Joint Conferences on Artificial Intelligence.

Davis, E. 2014. *The Limitations of Standardized Science Tests as Benchmarks for Artificial Intelligence Research*. arXiv preprint. arXiv:abs/1411.1629. Ithaca, NY: Cornell University Library.

Davis, E. 2016. How to Write Science Questions that are Easy for People and Hard for Computers. *AI Magazine* 37: 13–22.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Ferrucci, D.; Brown, E.; Chu-Carroll, J.; Fan, J.; Gondek, D.; Kalyanpur, A. A.; Lally, A.; Murdock, J. W.; Nyberg, E.; Prager, J., et al. 2010. Building Watson: An Overview of the DeepQA Project. *AI Magazine* 31(3): 59–79.

Friedland, N. S.; Allen, P. G.; Matthews, G.; Witbrock, M.; Baxter, D.; Curtis, J.; Shepard, B.; Miraglia, P.; Angele, J.; Staab, S., et al. 2004. Project Halo: Towards a Digital Aristotle. *AI Magazine* 25(4): 29.

Fujita, A.; Kameda, A.; Kawazoe, A.; and Miyao, Y. 2014. Overview of Todai Robot Project and Evaluation Framework of its NLP-based Problem Solving. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation*. Paris, France: European Language Resources Association.

Genesereth, M. R., and Nilsson, N. J. 2012. *Logical Foundations of Artificial Intelligence*. San Francisco, CA: Morgan Kaufmann

Guo, S.; Zeng, X.; He, S.; Liu, K.; and Zhao, J. 2017. Which is the Effective Way for Gaokao: Information Retrieval or Neural Networks? In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, 111–20. www.aclweb.org/anthology/E17-1011.pdf.

Gururangan, S.; Swayamdipta, S.; Levy, O.; Schwartz, R.; Bowman, S. R.; and Smith, N. A. 2018. Annotation Artifacts in Natural Language Inference Data. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Hopkins, M.; Petrescu-Prahova, C.; Levin, R.; Bras, R. L.; Herrasti, A.; and Joshi, V. 2017. Beyond Sentential Semantic Parsing: Tackling the Math SAT with a Cascade of Tree Transducers. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics.

Joshi, M.; Choi, E.; Weld, D. S.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Khashabi, D.; Khot, T.; Sabharwal, A.; Clark, P.; Etzioni, O.; and Roth, D. 2016. Question Answering via Integer Programming over Semi-Structured Knowledge. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*. Palo Alto, CA: Association for the Advancement of Artificial Intelligence Press.

Khashabi, D.; Khot, T.; Sabharwal, A.; and Roth, D. 2018. Question Answering as Global Reasoning over Semantic Abstractions. In *Proceedings of the Thirty-Second Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Khot, T.; Balasubramanian, N.; Gribkoff, E.; Sabharwal, A.; Clark, P.; and Etzioni, O. 2015. Exploring Markov Logic Networks for Question Answering. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics.

Khot, T.; Sabharwal, A.; and Clark, P. F. 2017. Answering Complex Questions using Open Information Extraction. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Krishnamurthy, J.; Tafjord, O.; and Kembhavi, A. 2016. Semantic Parsing to Probabilistic Programs for Situated Question Answering. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics.

Lai, G.; Xie, Q.; Liu, H.; Yang, Y.; and Hovy, E. 2017. RACE: Large-scale Reading Comprehension Dataset from Examinations. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics.

Lake, B. M., and Baroni, M. 2018. Generalization without Systematicity: On the Compositional Skills of Sequence-to-Sequence Recurrent Networks. In *Proceedings of the 34th International Conference on Machine Learning*. arXiv preprint arXiv: 1711.00350. Ithaca, NY: Cornell University Library.

Landauer, T. K., and Dumais, S. T. 1997. A Solution to Plato's problem: The Latent Semantic Analysis Theory of Acquisition, Induction, and Representation of Knowledge. *Psychological Review* 104(2): 211.

Larkin, J. H.; McDermott, J.; Simon, D. P.; and Simon, H. A. 1980. Models of Competence in Solving Physics Problems. *Cognitive Science* 4: 317–45.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. *RoBERTa: A Robustly Optimized BERT Pretraining Approach*. arXiv preprint arXiv: 1907.11692. Ithaca, NY: Cornell University Library.

Matsuzaki, T.; Iwane, H.; Anai, H.; and Arai, N. H. 2014. The most Uncreative Examinee: A First Step toward Wide Coverage Natural Language Math Problem Solving. In *Proceedings of the Twenty-Eighth Association for the Advancement of Artificial Intelligence (AAAI) Conference on Artificial Intelligence*. Palo Alto, CA: AAAI Press.

Mihaylov, T.; Clark, P.; Khot, T.; and Sabharwal, A. 2018. Can a Suit of Armor Conduct Electricity? A New Dataset for Open Book Question Answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics.

Mohammad, S. M.; Dorr, B. J.; Hirst, G.; and Turney, P. D. 2013. Computing Lexical Contrast. *Computational Linguistics* 39(3): 555–90.

Mott, N. 2016. Todai Robot Gives Up on Getting Into the University of Tokyo. *Inverse*. www.inverse.com/article/23761-todai-robot-gives-up-university-tokyo.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M. P.; Clark, C.; Lee, K.; and Zettlemoyer, L. S. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Piatetsky-Shapiro, G.; Djeraba, C.; Getoor, L.; Grossman, R.; Feldman, R.; and Zaki, M. 2006. What are the Grand Challenges for Data Mining?: KDD-2006 Panel Report. *SIGKDD Explorations* 8(2): 70–7.

Poliak, A.; Haldar, A.; Rudinger, R.; Hu, J. E.; Pavlick, E.; White, A. S.; and Durme, B. V. 2018a. Collecting Diverse Natural Language Inference Problems for Sentence Representation Evaluation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics.

Poliak, A.; Naradowsky, J.; Haldar, A.; Rudinger, R.; and Van Durme, B. 2018b. Hypothesis Only Baselines in Natural Language Inference. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*. Stroudsberg, PA: Association for Computational Linguistics.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics.

Reddy, R. 1988. Foundations and Grand Challenges of Artificial Intelligence: AAAI Presidential Address. *AI Magazine* doi.org/10.1609/aimag.v9i4.950.

Reddy, R. 2003. Three Open Problems in AI. *Journal of the Association for Computing Machinery* 50(1): 83–6.

Rudinger, R.; White, A. S.; and Durme, B. V. 2018. Neural Models of Factuality. In *The 2019 Human Language Technology: Conference of the North American Chapter of the Association of Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Schoenick, C.; Clark, P. F.; Tafjord, O.; Turney, P. D.; and Etzioni, O. 2017. Moving beyond the Turing Test with the Allen AI Science Challenge. *Communications of the ACM* 60(9): 60–4.

Seo, M. J.; Hajishirzi, H.; Farhadi, A.; Etzioni, O.; and Malcolm, C. 2015. Solving Geometry Problems: Combining Text and Diagram Interpretation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics.

Seo, M. J.; Kembhavi, A.; Farhadi, A.; and Hajishirzi, H. 2016. *Bidirectional Attention Flow for Machine Comprehension*. arXiv preprint: ArXiv abs/1611.01603. Ithaca, NY: Cornell University Library.

Strickland, E. 2013. Can an AI Get into the University of Tokyo? *IEEE Spectrum* 50(9): 13–4.

Sun, K.; Yu, D.; Yu, D.; and Cardie, C. 2019. Improving Machine Reading Comprehension with General Reading Strategies. In *The 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Tafjord, O.; Gardner, M.; Lin, K.; and Clark, P. 2019. QuaRTz: An Open-Domain Dataset of Qualitative Relationship Questions. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing*. Stroudsberg, PA: Association for Computational Linguistics.

Tainaka, M. 2013. Todai Robot Project. NII Today 46. www.nii.ac.jp/userdata/results/pr_data/NII_Today/60_en/all.pdf.

Tandon, N.; Mishra, B. D.; Grus, J.; Yih, W.-t.; Bosselut, A.; and Clark, P. 2018. Reasoning about Actions and State Changes by Injecting Commonsense Knowledge. In *Proceedings of the 2018 Conference on Empirical Methods in Natural*

Language Processing. Stroudsberg, PA: Association for Computational Linguistics.

Trivedi, H.; Kwon, H.; Khot, T.; Sabharwal, A.; and Balasubramanian, N. 2019. Repurposing Entailment for Multi-Hop Question Answering Tasks. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Turney, P. D. 2006. Similarity of Semantic Relations. *Computational Linguistics* 32(3): 379–416.

Turney, P. D. 2017. *Leveraging Term Banks for Answering Complex Questions: A Case for Sparse Vectors*. arXiv preprint arXiv: 1704.03543. Ithaca, NY: Cornell University Library.

Wang, A.; Singh, A.; Michael, J.; Hill, F.; Levy, O.; and Bowman, S. R. 2019. GLUE: A Multi-task Benchmark and Analysis Platform for Natural Language Understanding. In *7th International Conference on Learning Representations ICLR-19*. arXiv preprint: arXiv 1804.07461. Ithaca, NY: Cornell University Library.

Wang, W.; Yan, M.; and Wu, C. 2018. Multi-Granularity Hierarchical Attention Fusion Networks for Reading Comprehension and Question Answering. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics*. Stroudsberg, PA: Association for Computational Linguistics.

Weston, J.; Bordes, A.; Chopra, S.; and Mikolov, T. 2015. *Towards AI-Complete Question Answering: A Set of Prerequisite Toy Tasks*. arXiv preprint: arXiv 1502.05698. Ithaca, NY: Cornell University Library.

Wolfson, T.; Geva, M.; Gupta, A.; Gardner, M.; Goldberg, Y.; Deutch, D.; and Berant, J. 2020. Break It Down: A Question Understanding Benchmark. *Transactions of the Association for Computational Linguistics* 8: 183–98.

Peter Clark is a senior research manager at AI2 and leads the Aristo Project. His work focuses on NLP, machine reasoning, and world knowledge, and the interplay among these three areas.

Oren Etzioni is the CEO of AI2, and a professor of computer science at the University of Washington.

Daniel Khashabi is a Young Investigator at AI2. His interests lie at the interface of AI and NLP. He holds a PhD from the University of Pennsylvania and a bachelor's degree from Amirkabir University of Technology (Tehran Polytechnic).

Tushar Khot is a senior research scientist at AI2. His current research focuses on the reasoning problems in NLU. He received his PhD from the University of Wisconsin-Madison in Statistical Relational Learning.

Bhavana Dalvi Mishra is a senior research scientist at AI2. Her research focuses on common-sense reasoning, question answering, and knowledge graphs. She received her PhD from Carnegie Mellon University in 2015 and her MTech from the Indian Institute of Technology, Bombay in 2007.

Kyle Richardson is a research scientist at AI2, where he works at the intersection of NLP and machine learning. He holds a PhD from the University of Stuttgart, Germany, and his work focuses broadly on problems related to computational NLU including automated question answering, semantic parsing, and natural language inference.

Ashish Sabharwal is a senior research scientist at AI2, where he works on machine reasoning, explanations, and language understanding. He previously held research appointments at IBM Watson and Cornell University, and obtained a PhD from the University of Washington.

Carissa Schoenick is a senior program manager and communications director at AI2, where she works on defining, driving, and communicating the future of AI. Schoenick has spent over a decade at the forefront of computational knowledge initiatives, including management of the computable data and natural language parsing efforts for Wolfram Alpha and the implementation of computational cloud support features and environments for Amazon Web Services.

Oyvind Tafjord is a principal research scientist with the Aristo project at AI2. His research focuses on NLU and reasoning in the context of scientific knowledge. He holds a PhD in Theoretical Physics from Princeton University.

Niket Tandon is a research scientist with the Aristo team at AI2. His research focuses on machine reasoning and NLU. He received a PhD in Commonsense Knowledge Graphs from Saarland University and the Max Planck Institute for Informatics.

Sumithra Bhakthavatsalam is a research engineer at AI2 working in areas of knowledge base construction, information extraction, and question answering.

Dirk Groeneveld is a senior software engineer at AI2. He came to the United States from Germany to study computer science at the University of California, Irvine. After graduation, he moved to Seattle, Washington and worked at Microsoft and Amazon. Afterward he founded a moderately successful startup, and finally joined AI2.

Michal Guerquin is a senior software engineer working on platform infrastructure at AI2. He seeks simple solutions to complex problems, and is interested in distributed computing, big data, and web technologies.

Michael Schmitz is the director of engineering at AI2, and led the initial engineering team on Aristo. Schmitz has a degree in Computer Science and Mathematics from the University of Washington.