From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project

Peter Clark
November 2019
Science Questions: A Grand Challenge...
Over the last decade, I began to think about a "Digital Aristotle", an easy-to-use, all-encompassing knowledge storehouse...to advance the field of AI.
Over the last decade, I began to think about a "Digital Aristotle", an easy-to-use, all-encompassing knowledge storehouse... to advance the field of AI.

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.
(D) The particles increase in volume.
Elementary Science Tests as a Grand Challenge

Machines that can:
- Answer a wide variety of questions
- Answer complex questions
- Commonsense and world knowledge
Elementary Science Tests as a Grand Challenge

Machines that can:

- Answer a wide variety of questions
- Answer complex questions
- Commonsense and world knowledge

**AND** a task that is:

- Clearly Measurable
- Graduated
- Not gameable
- Ambitious but Realistic
- Motivating!
Elementary Science Tests as a Grand Challenge

Machines that can:

- Answer a wide variety of questions
- Answer complex questions
- Commonsense and world knowledge

**AND** a task that is:

- Clearly Measurable
- Graduated
- Not gameable
- Ambitious but Realistic
- Motivating!
Some Example Questions

Which object is the best conductor of electricity?
(A) a wax crayon  (B) a plastic spoon
(C) a rubber eraser  (D) an iron nail
Which object is the best conductor of electricity?
(A) a wax crayon  (B) a plastic spoon
(C) a rubber eraser  (D) an iron nail
Some Example Questions

Which object is the best conductor of electricity?
(A) a wax crayon  (B) a plastic spoon  
(C) a rubber eraser  (D) an iron nail

City administrators can encourage energy conservation by
(1) lowering parking fees  
(2) building larger parking lots  
(3) decreasing the cost of gasoline  
(4) lowering the cost of bus and subway fares
Question Categories Not Covered

- Diagrams

<table>
<thead>
<tr>
<th>Date</th>
<th>Sunrise</th>
<th>Sunset</th>
</tr>
</thead>
<tbody>
<tr>
<td>February 6</td>
<td>7:00 a.m.</td>
<td>5:20 p.m.</td>
</tr>
<tr>
<td>February 16</td>
<td>6:50 a.m.</td>
<td>5:30 p.m.</td>
</tr>
<tr>
<td>February 22</td>
<td>6:40 a.m.</td>
<td>5:40 p.m.</td>
</tr>
</tbody>
</table>

Source: www.sunsetsunset.com
Question Categories Not Covered

- Diagrams

- Direct Answer Questions
Progression on NY Regents 8th Grade (NDMC)

(hidden test set, questions as written, NDMC, 5 years/119 qns)
The Allen AI Science Challenge

Is your model smarter than an 8th grader?

The Allen Institute for Artificial Intelligence (AI2) is working to improve humanity through fundamental advances in artificial intelligence. One critical but challenging problem in AI is to demonstrate the ability to consistently understand and correctly answer general questions about the world.

The Aristo project at AI2 is focused on building such a system. One way Aristo "learns" is by extracting facts from various sources and processing them into a structured knowledge base. When taking an exam, questions are parsed and processed along with the knowledge base to provide a "best answer".
THE BEST AI STILL FLUNKS 8TH GRADE SCIENCE

(hidden test set, questions as written, NDMC, 5 years/119 qns)
Progression on NY Regents 8th Grade (NDMC)

- 2014: 36.4%
- 2015: 58.1%
- 2016: 63.1%
- 2017: 72.2%
- 2018: 73.1%

- Baseline retrieval methods
- Tables
- Statistics + retrieval + reasoning (AAAI’16)
- (random)

(hidden test set, questions as written, NDMC, 5 years/119 qns)
Progression on NY Regents 8th Grade (NDMC)

- 2014: 36.4%
- 2015: 58.1%
- 2016: 63.1%
- 2017: 72.2%
- 2018: 73.1%
- 2019: 90.7%

- Language models + specialist solvers
- Statistics + retrieval + reasoning (AAAI’16)
- Baseline retrieval methods
- Tables

(random)

(hidden test set, questions as written, NDMC, 5 years/119 qns)
Progression on NY Regents 8th Grade (NDMC)

Separate test on 3 latest exams (2017-2019): 93.3%

(hidden test set, questions as written, NDMC, 5 years/119 qns)
Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?
Aristo: an over-simplified overview

- An ensemble architecture

Diagram:
- Text (Web, Science) flows to Structured Representation.
- Structured Representation feeds into Retrieval & Statistics and Inference Solvers.
- Language Models combine with Inference Solvers.
- Combined output goes to Combiner (Ensemble) for Answer.
Retrieval and Statistical Solvers

- Information Retrieval Solver
  - question + answer option that best matches a corpus sentence
Information Retrieval Solver

- question + answer option that best matches a corpus sentence
- Aristo Corpus:
  - Web crawl from Univ Waterloo (330GB)
  - (science parts of) Wikipedia
  - Science textbooks
Retrieval and Statistical Solvers

- **Information Retrieval Solver**
  - question + answer option that best matches a corpus sentence
  - Aristo Corpus:
    - Web crawl from Univ Waterloo (330GB)
    - (science parts of) Wikipedia
    - Science textbooks

- **PMI**
  - between question and answer words
Retrieval and Statistical Solvers

- **Information Retrieval Solver**
  - question + answer option that best matches a corpus sentence
  - Aristo Corpus:
    - Web crawl from Univ Waterloo (330GB)
    - (science parts of) Wikipedia
    - Science textbooks

- **PMI**
  - between question and answer words

- **ACME**
  - link question to answer via terms in a **termbank**
  - heavy use of vector spaces
Retrieval and Statistical Solvers

- **Information Retrieval Solver**
  - question + answer option that best matches a corpus sentence
  - Aristo Corpus:
    - Web crawl from Univ Waterloo (330GB)
    - (science parts of) Wikipedia
    - Science textbooks

- **PMI**
  - between question and answer words

- **ACME**
  - link question to answer via terms in a termbank
  - heavy use of vector spaces

Infections may be caused by (1) mutations (2) microorganisms (3) toxic substances (4) climate change
Retrieval and Statistical Solvers

- **Information Retrieval Solver**
  - question + answer option that best matches a corpus sentence
  - Aristo Corpus:
    - Web crawl from Univ Waterloo (330GB)
    - (science parts of) Wikipedia
    - Science textbooks

- **PMI**
  - between question and answer words

- **ACME**
  - link question to answer via terms in a **termbank**
  - heavy use of vector spaces

Infections may be caused by:
- (1) mutations
- (2) microorganisms
- (3) toxic substances
- (4) climate change

**Products contaminated with microorganisms may cause infection.**
Aristo: an over-simplified overview

- An ensemble architecture

[Diagram showing the flow of information from Text to Web, Science, Structured Representation, Retrieval & Statistics, Inference Solvers, Language Models, Combiner (Ensemble), and Answer]
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

Planets and Moon reflect light back from the Sun

Moon orbits planets

Moon is a large natural satellite

Primary objects in solar system are planets and moons
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

Stormy weather negatively affects a coastline by (A) causing erosion (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses
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

Stormy weather negatively affects a coastline by (A) causing erosion (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses
Tuple Inference

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

Primary objects in solar system are planets and moons

Stormy weather negatively affects a coastline by (A) causing erosion (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses
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

Primary objects in solar system are planets and moons.

Stormy weather negatively affects a coastline by (A) causing erosion (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses.
Stormy weather negatively affects a coastline by (A) causing erosion (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses.
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

Primary objects in solar system are planets and moons

Stormy weather negatively affects a coastline by (A) causing erosion (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses
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

Primary objects in solar system are planets and moons.

Stormy weather negatively affects a coastline by (A) causing erosion (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses
Stormy weather negatively affects a coastline by (A) causing erosion (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses.
Why does a toy car roll farther on a wood floor than on a thick carpet? (A) The floor has less resistance. (B) The floor has more traction (C) …
Why does a toy car roll farther on a wood floor than on a thick carpet? (A) **The floor has less resistance.** (B) The floor has more traction (C) ...
Why does a toy car roll farther on a wood floor than on a thick carpet? (A) The floor has less resistance. (B) The floor has more traction (C) …

- smoothness up
- friction down
- heat down

- speed up
- distance up
Why does a toy car roll farther on a wood floor than on a thick carpet? (A) The floor has less resistance. (B) The floor has more traction. (C) ...
Qualitative Knowledge

Why does a toy car roll farther on a wood floor than on a thick carpet? (A) The floor has less resistance. (B) The floor has more traction (C) ...
Why does a toy car roll farther on a wood floor than on a thick carpet? (A) The floor has less resistance. (B) The floor has more traction (C) ...
Where was Facebook launched? (A) Cambridge (B) Silicon Valley

H₀: Facebook was launched in Cambridge.

P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in Silicon Valley.
P3: Harvard University is at Cambridge, Massachusetts.
P4: Harvard is only a few miles from Boston.

Relevance Model

P1: 0.4
P2: 0.1
P3: 0.4
P4: 0.1

Sentence-wise

Facebook was launched at Harvard University. Facebook headquarters was set up in Silicon Valley. Harvard University is at Cambridge, Massachusetts. Harvard is only a few miles from Boston.

Multi-Level Aggregator

Entails?
Where was Facebook launched? (A) Cambridge (B) Silicon Valley

**H₀:** Facebook was launched in Cambridge.

- **P1:** Facebook was launched at Harvard University. **Relevance:** P1: 0.4
- **P2:** Facebook headquarters was set up in Silicon Valley. **Relevance:** P2: 0.1
- **P3:** Harvard University is at Cambridge, Massachusetts. **Relevance:** P3: 0.4
- **P4:** Harvard is only a few miles from Boston. **Relevance:** P4: 0.1

Facebook was launched at Harvard University. Facebook headquarters was set up in Silicon Valley. Harvard University is at Cambridge, Massachusetts. Harvard is only a few miles from Boston. **Entails?**

*Multi-Level Aggregator*
Where was Facebook launched? (A) Cambridge (B) Silicon Valley

H₀: Facebook was launched in Cambridge.

P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in Silicon Valley.
P3: Harvard University is at Cambridge, Massachusetts.
P4: Harvard is only a few miles from Boston.

Multi-Level Aggregator

Entails?

Facebook was launched at Harvard University. Facebook headquarters was set up in Silicon Valley. Harvard University is at Cambridge, Massachusetts. Harvard is only a few miles from Boston.
Where was Facebook launched? (A) Cambridge (B) Silicon Valley

\[ H_0: \text{Facebook was launched in Cambridge.} \]

- P1: Facebook was launched at Harvard University. \( R(P1) = 0.4 \)
- P2: Facebook headquarters was set up in Silicon Valley. \( R(P2) = 0.1 \)
- P3: Harvard University is at Cambridge, Massachusetts. \( R(P3) = 0.4 \)
- P4: Harvard is only a few miles from Boston. \( R(P4) = 0.1 \)

\[
\text{Facebook was launched at Harvard University. Facebook headquarters was set up in Silicon Valley. Harvard University is at Cambridge, Massachusetts. Harvard is only a few miles from Boston.}
\]

\[
\text{Entails?}
\]
Where was Facebook launched? (A) Cambridge (B) Silicon Valley

H₀: Facebook was launched in Cambridge.

P1: Facebook was launched at Harvard University.
P2: Facebook headquarters was set up in Silicon Valley.
P3: Harvard University is at Cambridge, Massachusetts.
P4: Harvard is only a few miles from Boston.

Facebook was launched at Harvard University. Facebook headquarters was set up in Silicon Valley. Harvard University is at Cambridge, Massachusetts. Harvard is only a few miles from Boston.

Which substance is a compound? (A) sodium (B) chlorine (C) table salt (D) salt water

Answer: (C) table salt

Score  Premise
0.653  Table salt, for instance, is a compound of sodium and chlorine.
0.543  An example of an inorganic substance is table salt.
Aristo: an over-simplified overview

- An ensemble architecture
What part of a plant needs sunlight to do its job? (A) leaf...
What part of a plant needs sunlight to do its job? (A) leaf...
What part of a plant needs sunlight to do its job? (A) leaf...
What part of a plant needs sunlight to do its job? (A) leaf...
Similar Progress on 4th Grade NDMC
<table>
<thead>
<tr>
<th>Test Set</th>
<th>Num Q</th>
<th>IR</th>
<th>PMI</th>
<th>ACME</th>
<th>TupInf</th>
<th>Mulite</th>
<th>AristoBERT</th>
<th>AristoRoBERTa</th>
<th>ARISTO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regents 4th</td>
<td>109</td>
<td>64.45</td>
<td>66.28</td>
<td>67.89</td>
<td>63.53</td>
<td>69.72</td>
<td>86.24</td>
<td>88.07</td>
<td>89.91</td>
</tr>
<tr>
<td>Regents 8th</td>
<td>119</td>
<td>66.60</td>
<td>69.12</td>
<td>67.65</td>
<td>61.41</td>
<td>68.91</td>
<td>86.55</td>
<td>88.24</td>
<td>91.60</td>
</tr>
<tr>
<td>Regents 12th</td>
<td>632</td>
<td>41.22</td>
<td>46.95</td>
<td>41.57</td>
<td>35.35</td>
<td>56.01</td>
<td>75.47</td>
<td>82.28</td>
<td>83.54</td>
</tr>
<tr>
<td>ARC-Easy</td>
<td>2376</td>
<td>74.48</td>
<td>77.76</td>
<td>66.60</td>
<td>57.73</td>
<td>64.69</td>
<td>81.78</td>
<td>82.88</td>
<td>86.99</td>
</tr>
<tr>
<td>ARC-Challenge</td>
<td>1172</td>
<td>n/a†</td>
<td>n/a†</td>
<td>20.44</td>
<td>23.73</td>
<td>37.36</td>
<td>57.59</td>
<td><strong>64.59</strong></td>
<td>64.33</td>
</tr>
</tbody>
</table>
### Individual Solver Performances

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Num Q</th>
<th>IR</th>
<th>PMI</th>
<th>ACME</th>
<th>TupInf</th>
<th>Multee</th>
<th>AristoBERT</th>
<th>AristoRoBERTa</th>
<th>ARISTO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regents 4th</td>
<td>109</td>
<td>64.45</td>
<td>66.28</td>
<td>67.89</td>
<td>65.53</td>
<td>69.72</td>
<td>86.24</td>
<td>88.07</td>
<td>89.91</td>
</tr>
<tr>
<td>Regents 8th</td>
<td>119</td>
<td>66.60</td>
<td>69.12</td>
<td>67.65</td>
<td>61.41</td>
<td>68.91</td>
<td>86.55</td>
<td>88.24</td>
<td>91.60</td>
</tr>
<tr>
<td>Regents 12th</td>
<td>632</td>
<td>41.22</td>
<td>46.95</td>
<td>41.57</td>
<td>35.35</td>
<td>56.01</td>
<td>75.47</td>
<td>82.28</td>
<td>83.54</td>
</tr>
<tr>
<td>ARC-Easy</td>
<td>2376</td>
<td>74.48</td>
<td>77.76</td>
<td>66.60</td>
<td>57.73</td>
<td>64.69</td>
<td>81.78</td>
<td>82.88</td>
<td>86.99</td>
</tr>
<tr>
<td>ARC-Challenge</td>
<td>1172</td>
<td>n/a</td>
<td>n/a</td>
<td>20.44</td>
<td>23.73</td>
<td>37.36</td>
<td>57.59</td>
<td><strong>64.59</strong></td>
<td>64.33</td>
</tr>
</tbody>
</table>

Most of the heavy lifting....
## Individual Solver Performances

<table>
<thead>
<tr>
<th>Test Set</th>
<th>Num Q</th>
<th>IR</th>
<th>PMI</th>
<th>ACME</th>
<th>TupInf</th>
<th>Multee</th>
<th>AristoBERT</th>
<th>AristoRoBERTa</th>
<th>ARISTO</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regents 4th</td>
<td>109</td>
<td>64.45</td>
<td>66.28</td>
<td>67.89</td>
<td>63.53</td>
<td>69.72</td>
<td>86.24</td>
<td>88.07</td>
<td>89.91</td>
</tr>
<tr>
<td>Regents 8th</td>
<td>119</td>
<td>66.60</td>
<td>69.12</td>
<td>67.65</td>
<td>61.41</td>
<td>68.91</td>
<td>86.55</td>
<td>88.24</td>
<td>91.60</td>
</tr>
<tr>
<td>Regents 12th</td>
<td>632</td>
<td>41.22</td>
<td>46.95</td>
<td>41.57</td>
<td>35.35</td>
<td>56.01</td>
<td>75.47</td>
<td>82.28</td>
<td>83.54</td>
</tr>
<tr>
<td>ARC-Easy</td>
<td>2376</td>
<td>74.48</td>
<td>77.76</td>
<td>66.60</td>
<td>57.73</td>
<td>64.69</td>
<td>81.78</td>
<td>82.88</td>
<td>86.99</td>
</tr>
<tr>
<td>ARC-Challenge</td>
<td>1172</td>
<td>n/a</td>
<td>n/a</td>
<td>20.44</td>
<td>23.73</td>
<td>37.36</td>
<td>57.59</td>
<td>64.59</td>
<td>64.33</td>
</tr>
</tbody>
</table>

Most of the heavy lifting....
Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?
1. Checking for annotation artifacts

(A) friction
(B) light
(C) force
(D) weather
1. Checking for annotation artifacts

(A) friction
(B) light
(C) force
(D) weather

<table>
<thead>
<tr>
<th>Test dataset</th>
<th>“Answer only” score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regents 4th</td>
<td>38.53</td>
</tr>
<tr>
<td>Regents 8th</td>
<td>37.82</td>
</tr>
<tr>
<td>Regents 12th</td>
<td>47.94</td>
</tr>
<tr>
<td>ARC-Easy</td>
<td>36.17</td>
</tr>
<tr>
<td>ARC-Challenge</td>
<td>35.92</td>
</tr>
<tr>
<td>All</td>
<td>37.11</td>
</tr>
</tbody>
</table>
2. Is it fooled by “obviously wrong” answers?

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
(D) weather [selected, correct]
2. Is it fooled by “obviously wrong” answers?

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
(D) weather [selected, correct]

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
(D) weather
(E) joule
(F) gradient
(G) trench
(H) add heat
2. Is it fooled by “obviously wrong” answers?

The condition of the air outdoors at a certain time of day is known as:
(A) friction
(B) light
(C) force
(D) weather [selected, correct]

The condition of the air outdoors at a certain time of day is known as:
(A) friction  (E) joule
(B) light  (F) gradient [selected]
(C) force  (G) trench
(D) weather [correct]  (H) add heat
2. Is it fooled by “obviously wrong” answers?

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
**D) weather [selected, correct]**

Retrain

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
(D) weather [correct]

The condition of the air outdoors at a certain time of day is known as
(A) friction
(B) light
(C) force
**D) weather [correct, selected]**
2. Is it fooled by “obviously wrong” answers?

The condition of the air outdoors at a certain time of day is known as
(A) friction (E) joule
(B) light (F) gradient [selected]
(C) force (G) trench
(D) weather [correct, selected]

The condition of the air outdoors at a certain time of day is known as
(A) friction (E) joule
(B) light (F) gradient [selected]
(C) force (G) trench
(D) weather [correct, selected]
City administrators can encourage energy conservation by
(1) lowering parking fees
(2) building larger parking lots
(3) decreasing the cost of gasoline
(4) lowering the cost of bus and subway fares
2. Is it fooled by “obviously wrong” answers?

The condition of the air outdoors at a certain time of day is known as:
(A) friction  (E) joule
(B) light  (F) gradient [selected]
(C) force  (G) trench
(D) weather [correct, selected]

<table>
<thead>
<tr>
<th>Test dataset</th>
<th>4-way MC</th>
<th>Adversarial 8-way MC</th>
<th>% drop (relative)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Regents 4th</td>
<td>87.1</td>
<td>76.1</td>
<td>12.6</td>
</tr>
<tr>
<td>Regents 8th</td>
<td>78.9</td>
<td>76.4</td>
<td>3.1</td>
</tr>
<tr>
<td>Regents 12th</td>
<td>75.3</td>
<td>58.0</td>
<td>22.9</td>
</tr>
<tr>
<td>ARC-Easy</td>
<td>74.1</td>
<td>65.7</td>
<td>11.3</td>
</tr>
<tr>
<td>ARC-Challenge</td>
<td>55.5</td>
<td>47.4</td>
<td>14.0</td>
</tr>
<tr>
<td>ALL</td>
<td>69.1</td>
<td>59.5</td>
<td>13.8</td>
</tr>
</tbody>
</table>

Drop of (only) ≈ 10 points

Retrain
City administrators can encourage energy conservation by:
1. lowering parking fees
2. building larger parking lots
3. decreasing the cost of gasoline
4. lowering the cost of bus and subway fares
City administrators can encourage energy conservation by:
1. lowering parking fees
2. building larger parking lots
3. decreasing the cost of gasoline
4. lowering the cost of bus and subway fares
3. More than Pattern Matching?

City administrators can encourage energy conservation by
(1) lowering parking fees
(2) building larger parking lots
(3) decreasing the cost of gasoline
(4) lowering the cost of bus and subway fares

Which of the following organs does a squirrel **not** have
(A) a brain
(B) gills
(C) a heart
(D) lungs
3. More than Pattern Matching?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Conjunction</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Polarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>World tracking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

2019 Report Card for Aristo
3. More than Pattern Matching?

2019 Report Card for Aristo

Subject | Grade | Teacher Comments
--- | --- | ---
Negation | | 
Conjunction | | 
Polarity | | 
World tracking | | 
Factivity | | 
Counting | | 

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

Which of the following is not tall? (A) Alan (B) Bob (C) Charlie (D) David [correct]
Synthetic Conjunction Test

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
Synthetic Conjunction Test

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

- 1 conjunct: 98%
- 2 conjuncts: 95%
- 3 conjuncts: 94.5%
- 4 conjuncts: 80%
Synthetic Conjunction Test

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

1 conjunct: 98%  
2 conjuncts: 95%  
3 conjuncts: 94.5%  
4 conjuncts: 80%  
+ 1 negation 76.5%  
+ 1 negation 76%  
+ 1 negation 75%
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

1 conjunct: 98%
2 conjuncts: 95%
3 conjuncts: 94.5%
4 conjuncts: 80%


Which of the following is old and red and light and big and not short? (A) Alan (B) Bob (C) Charlie (D) David
### 3. More than Pattern Matching?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td></td>
</tr>
<tr>
<td><strong>Conjunction</strong></td>
<td>B+</td>
<td>Nice work!</td>
</tr>
<tr>
<td>Polarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>World tracking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Aristo**

94%

80% - 98%
3. More than Pattern Matching?

2019 Report Card for Aristo

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td>Nice work!</td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td></td>
</tr>
<tr>
<td>Polarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>World tracking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Context: For a given medium, sound has a slower speed at lower temperatures.

Question: If Jim turns the thermostat down in his room while listening to music, what will happen to the speed of the sound waves in the room? (A) they will speed up (B) they will slow down [correct]
3. More than Pattern Matching?

2019 Report Card for Aristo

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td>Nice work!</td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td></td>
</tr>
<tr>
<td>Polarity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>World tracking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

94%  
80% - 98%

Context: For a given medium, sound has a slower speed at lower temperatures.

Question: If Jim turns the thermostat **down** in his room while listening to music, what will happen to the speed of the sound waves in the room?  
(A) they will speed up  (B) they will slow down  [correct]
### 3. More than Pattern Matching?

**2019 Report Card for Aristo**

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td>Nice work!</td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td></td>
</tr>
<tr>
<td><strong>Polarity</strong></td>
<td>D+</td>
<td>Could ace this with more study!</td>
</tr>
<tr>
<td>World tracking</td>
<td></td>
<td>94%</td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td>80% - 98%</td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td>67.1%</td>
</tr>
</tbody>
</table>

**Context:**
For a given medium, sound has a slower speed at lower temperatures.

**Question:**
If Jim turns the thermostat down in his room while listening to music, what will happen to the speed of the sound waves in the room?  
(A) they will speed up (B) they will slow down **[correct]**
### 3. More than Pattern Matching?

#### 2019 Report Card for Aristo

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td>Nice work!</td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td></td>
</tr>
<tr>
<td>Polarity</td>
<td>D+</td>
<td>Could ace this with more study!</td>
</tr>
<tr>
<td>World tracking</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- **94%**
- **80% - 98%**
- **67.1%**
3. More than Pattern Matching?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Aristo Teacher Comments</th>
<th>%</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td>Nice work!</td>
<td>94%</td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td></td>
<td>80% -98%</td>
</tr>
<tr>
<td>Polarity</td>
<td>D+</td>
<td>Could ace this with more study!</td>
<td>67.1%</td>
</tr>
<tr>
<td>World tracking</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Context: If someone travels for longer, they will travel further.

Question: John and Rita are going for a run. Rita gets tired and takes a break on the park bench. After twenty minutes in the park, who has run farther? (A) John [correct] (B) Rita
3. More than Pattern Matching?

2019 Report Card for

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Aristo Teacher Comments</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td>Nice work!</td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td>Could ace this with more study!</td>
</tr>
<tr>
<td>Polarity</td>
<td>D+</td>
<td></td>
</tr>
<tr>
<td>World tracking</td>
<td>C</td>
<td></td>
</tr>
<tr>
<td>Factivity</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

If someone **regretted** that a particular thing happened then:

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

(A) that thing might or might not have happened.
(B) that thing didn’t happen.
(C) that thing happened [correct]
3. More than Pattern Matching?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Teacher Comments</th>
<th>Aristo</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td>Nice work!</td>
<td>94%</td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td></td>
<td>80% - 98%</td>
</tr>
<tr>
<td>Polarity</td>
<td>D+</td>
<td>Could ace this with more study!</td>
<td>67.1%</td>
</tr>
<tr>
<td>World tracking</td>
<td>C</td>
<td></td>
<td>72.5%</td>
</tr>
<tr>
<td>Factivity</td>
<td>D</td>
<td></td>
<td>66.5%</td>
</tr>
<tr>
<td>Counting</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Source: Allen Institute for Artificial Intelligence
### 3. More than Pattern Matching?

<table>
<thead>
<tr>
<th>Subject</th>
<th>Grade</th>
<th>Aristo</th>
<th>Teacher Comments</th>
<th>Score</th>
</tr>
</thead>
<tbody>
<tr>
<td>Negation</td>
<td>A</td>
<td></td>
<td>Nice work!</td>
<td>94%</td>
</tr>
<tr>
<td>Conjunction</td>
<td>B+</td>
<td></td>
<td></td>
<td>80% - 98%</td>
</tr>
<tr>
<td>Polarity</td>
<td>D+</td>
<td></td>
<td>Could ace this with more study!</td>
<td>67.1%</td>
</tr>
<tr>
<td>World tracking</td>
<td>C</td>
<td></td>
<td></td>
<td>72.5%</td>
</tr>
<tr>
<td>Factivity</td>
<td>D</td>
<td></td>
<td></td>
<td>66.5%</td>
</tr>
<tr>
<td><strong>Counting</strong></td>
<td><strong>D</strong></td>
<td><strong>Red</strong></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Daniel picked up the football. Daniel dropped the football. Daniel got the milk.

How many objects is Daniel holding? (A) zero (B) one (C) two (D) three
Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?
What part of a plant needs sunlight to do its job? (A) leaf..
What part of a plant needs sunlight to do its job? (A) leaf...
4. Where is Aristo failing?

- Case study on 30 failures:
  - Good support for correct answer: 17 (56.7%)
  - 8 (26.7%)
  - 4 (13.3%)
4. Where is Aristo failing?

- Case study on 30 failures:

  - 17 (56.7%) Good support for correct answer
  - 8 (26.7%) Good support for incorrect answer
  - 4 (13.3%) Good support for incorrect answer
4. Where is Aristo failing?

- Case study on 30 failures:
  - No good support: 17 (56.7%)
  - Good support for correct answer: 8 (26.7%)
  - Good support for incorrect answer: 4 (13.3%)
4. Where is Aristo failing?

- Case study on 30 failures:
  - Reading Comprehension (IR won’t help): 8 (26.7%)
  - Good support for correct answer: 4 (13.3%)
  - Good support for incorrect answer: 0
  - No good support: 17 (56.7%)
1. Good support for the correct answer (13%)

Which is the best unit to measure distances between Earth and other solar systems in the universe? (A) miles (B) kilometers (C) light years (D) astronomical units
Which is the best unit to measure distances between Earth and other solar systems in the universe? (A) miles (B) kilometers (C) light years (D) astronomical units.

In general, distances in the solar system are measured in astronomical units.
Which is the best unit to measure distances between Earth and other solar systems in the universe? (A) miles (B) kilometers (C) light years (D) astronomical units

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

Distances between Earth and the stars are often measured in terms of light-years.
Which of these objects will most likely float in water? (A) glass marble (B) steel ball (C) hard rubber ball (D) table tennis ball
2. Good support for the incorrect answer (3%)  

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

- I remember it had like a rubber ball in it, which would maybe float up…
- We played soccer with a giant rubber ball that floated like a balloon.
- Rubber toys floated on the water.
3. No good support for the correct answer (57%)

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. (D) The particles increase in volume.

• No good single supporting sentence
3. No good support for the correct answer (57%)

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. (D) The particles increase in volume.

- No good single supporting sentence

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 (D) their method of catching food

- Need question decomposition
3. No good support for the correct answer (57%)

Which characteristic applies to animals in only one of these taxonomic groups: reptiles, mammals, birds, amphibians, or fishes? (A) have hair (B) lay eggs (C) have webbed feet (D) breathe with gills

- Boolean reasoning
3. No good support for the correct answer (57%)

Which characteristic applies to animals in only one of these taxonomic groups: reptiles, mammals, birds, amphibians, or fishes? (A) have hair (B) lay eggs (C) have webbed feet (D) breathe with gills

- Boolean reasoning

Which geologic structure will most likely take the longest time to form? (A) a fault (B) a sinkhole (C) a river meander (D) a mountain range

- Cross-option comparative
4. Reading Comprehension (27%)

- Story (experimental method)

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 (D) amount of growth
4. Reading Comprehension (27%)

- Story (experimental method)

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 (D) amount of growth

- Meta/sentiment

Which statement is an opinion? (A) Many plants are green. (B) Many plants are beautiful. (C) Plants require sunlight. (D) Plants can grow in different places.
About how long does it take for the Moon to complete one revolution around Earth? (A) 7 days (B) 30 days (C) 90 days (D) 365 days

- Because it takes the moon about **27.3 days** to complete one orbit around the Earth, the moon moves a little bit further around the Earth each day.
- It takes **27.3 days** for the moon to complete one revolution around the earth.
- The moon completes one revolution of the Earth in about **29.5 days**.
- The Moon completes one revolution around the Earth in **27.32166 days**.
Outline

- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?
What virus structure is similar in function to a cell membrane?
(A) protein shell (B) internal protein...
1. Question Decomposition

What virus structure is similar in function to a cell membrane?
(A) protein shell (B) internal protein...

Structure-function of membrane proteins. Membrane protein structure and function; Structure and function of membrane proteins; Shin's research interests involve the structure and function of cell membrane proteins, including influenza hemagglutinin protein and an HIV virus spike protein that are responsible for cellular-viral membrane fusion. Biological structure analysis by electron crystallography to characterize cell-membrane proteins and viruses; Structure-Function Analysis of the Influenza Virus Ion Channel Influenza virus protein M 2 is a small (97-residue) integral membrane protein that spans the cell membrane once and is minimally a disulfide-linked homotetramer. Biophysical chemists study protein structure and the functional structure of cell membranes. A huge unsolved question of cell membrane structure and function is the structure of membrane proteins. Viruses Proteins and Cell Membranes. Cell Membranes | top | Composition and Structure | Membrane proteins | Membrane functions |
What virus structure is similar in function to a cell membrane? (A) protein shell (B) internal protein...

- What is the function of a cell membrane?
  - Surrounds and protects, gives structure, regulates material, ...
- What part of the virus surrounds and protects it?
  - Protein shell, protein layer, ...

- GapQA *(EMNLP’19)*
- New dataset coming
2. Multihop Reasoning

Which conducts electricity? (A) suit of armor (B) cotton candy
2. Multihop Reasoning

Which conducts electricity? (A) suit of armor (B) cotton candy

Retrieval 1:
The reciprocal of the electrical resistivity is the electrical conductivity. Electrical conductivity is the capacity of metal to conduct an electric current. Electrical Conductivity Water without minerals will not conduct electricity.
2. Multihop Reasoning

Which conducts electricity? (A) suit of armor (B) cotton candy

Retrieval 1:
The reciprocal of the electrical resistivity is the electrical conductivity.
Electrical conductivity is the capacity of metal to conduct an electric current.
Electrical Conductivity Water without minerals will not conduct electricity.

Retrieval 2:
It was not suited to be a center for extensive metal-working.
A suit of armor is a historical type of personal body armour made from metal.
Resisting arrest is a criminal charge, but civil suits can be filed.
2. Multihop Reasoning

Which conducts electricity? (A) suit of armor (B) cotton candy

Retrieval 1:
The reciprocal of the electrical resistivity is the electrical conductivity. Electrical conductivity is the capacity of metal to conduct an electric current. Electrical Conductivity Water without minerals will not conduct electricity.

Retrieval 2:
It was not suited to be a center for extensive metal-working. A suit of armour is a historical type of personal body armour made from metal. Resisting arrest is a criminal charge, but civil suits can be filed.

Form Chains:
“suit of armor...made from metal” AND “...metal conduct electrical current”
=> “suit of armor conducts electricity”
2. Multihop Reasoning

Which conducts electricity? (A) suit of armor (B) cotton candy

Retrieval 1:
The reciprocal of the electrical resistivity is the electrical conductivity.
Electrical conductivity is the capacity of metal to conduct an electric current.
Electrical Conductivity Water without minerals will not conduct electricity.

Retrieval 2:
It was not suited to be a center for extensive metal-working.
A suit of armour is a historical type of personal body armour made from metal.
Resisting arrest is a criminal charge, but civil suits can be filed.

Form Chains:
“suit of armor...made from metal” AND “...metal conduct electrical current”
=> “suit of armor conducts electricity”

“Resisting arrest...suits can be filed” AND “reciprocal of resistivity is conductivity”
=> “suit of armor conducts electricity”
2. Multihop Reasoning

Which conducts electricity? (A) suit of armor (B) cotton candy

Retrieval 1:
The reciprocal of the electrical resistivity is the electrical conductivity.
Electrical conductivity is the capacity of metal to conduct an electric current.
Electrical Conductivity Water without minerals will not conduct electricity.

Retrieval 2:
It was not suited to be a center for extensive metal-working.
A suit of armour is a historical type of personal body armour made from metal.
Resisting arrest is a criminal charge, but civil suits can be filed.

Form Chains:
“suit of armor...made from metal” AND “...metal conduct electrical current”
=> “suit of armor conducts electricity”

“Resisting arrest...suits can be filed” AND “reciprocal of resistivity is conductivity”
=> “suit of armor conducts electricity”

Train system to recognize good chains
3. Modeling World States

Photosynthesis

Roots absorb water from the soil.
The water flows to the leaf.
Light and CO2 enter leaf.
Light, water, CO2 form sugar.
Photosynthesis

Roots absorb water from the soil.

The water flows to the leaf.

Light and CO2 enter leaf.

Light, water, CO2 form sugar.
Photosynthesis

Roots absorb water from the soil.

**The water flows to the leaf.**

Light and CO2 enter leaf.

Light, water, CO2 form sugar.
Photosynthesis

Roots absorb water from the soil.

The water flows to the leaf.

Light and CO2 enter leaf.

Light, water, CO2 form sugar.

Where is the sugar created?
Photosynthesis

Roots absorb water from the soil.

The water flows to the leaf.

Light and CO₂ enter leaf.

Light, water, CO₂ form sugar.

Where is the sugar created? **Light, water, CO₂** [BiDAF]
Can you pick up a penny with a magnet?
4. Explanation and Instruction

Can you pick up a penny with a magnet?

Yes
Can you pick up a penny with a magnet?

Yes

Why?

Because
- pennies are made of metal
- metals are magnetic
4. Explanation and Instruction

Can you pick up a penny with a magnet?

Yes

Why?

Because
- pennies are made of metal
- metals are magnetic

Actually:
Not all metals are magnetic.
Copper is not magnetic.

Try again!
4. Explanation and Instruction

Can you pick up a penny with a magnet?

Yes

Why?

Because
  - pennies are made of metal
  - metals are magnetic

Actually:
Not all metals are magnetic.
Copper is not magnetic.

Try again!

No – because:
  - pennies are made of copper
  - copper is not magnetic
Summary

ARISTO

- Surprising success!
- Reflects rapid progress of NLP
- >> “just pattern matching”
Summary

- Surprising success!
- Reflects rapid progress of NLP
- >> “just pattern matching”
- BUT:
  - falls short with some compositional questions
  - many other AI aspects missing
Summary

ARISTO

- Surprising success!
- Reflects rapid progress of NLP
- >> “just pattern matching”
- BUT:
  - falls short with some compositional questions
  - many other AI aspects missing

What do we need going forward?
- Reintroduce structured reasoning but with language-like representations
Surprising success!
Reflects rapid progress of NLP
>> “just pattern matching”
BUT:
- falls short with some compositional questions
- many other AI aspects missing

What do we need going forward?
- Reintroduce structured reasoning *but* with language-like representations

Thank you!