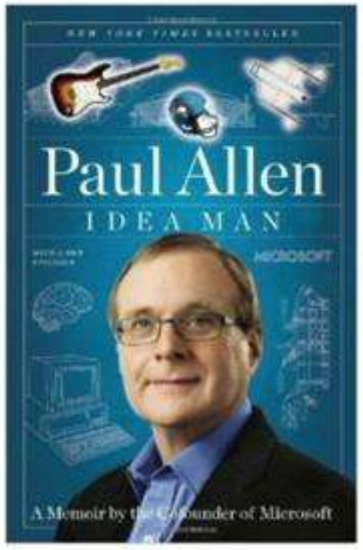
From 'F' to 'A' on the N.Y. Regents Science Exams: An Overview of the Aristo Project

Peter Clark

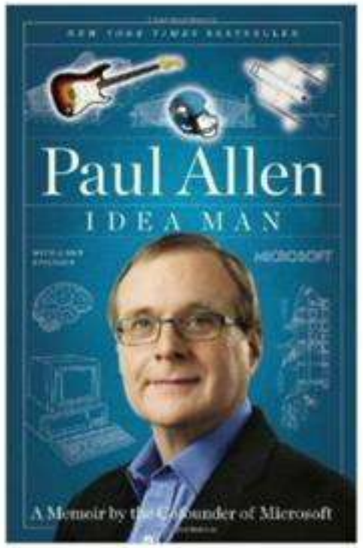
November 2019

Science Questions: A Grand Challenge…



ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

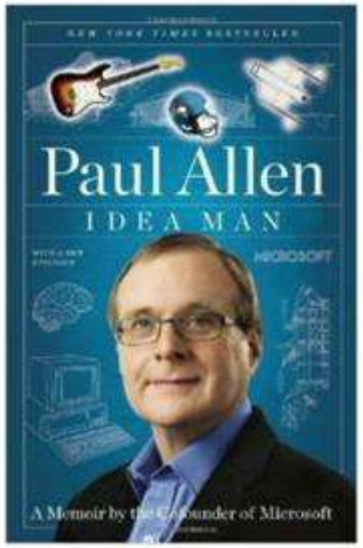
Science Questions: A Grand Challenge…



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.*

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

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.*

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

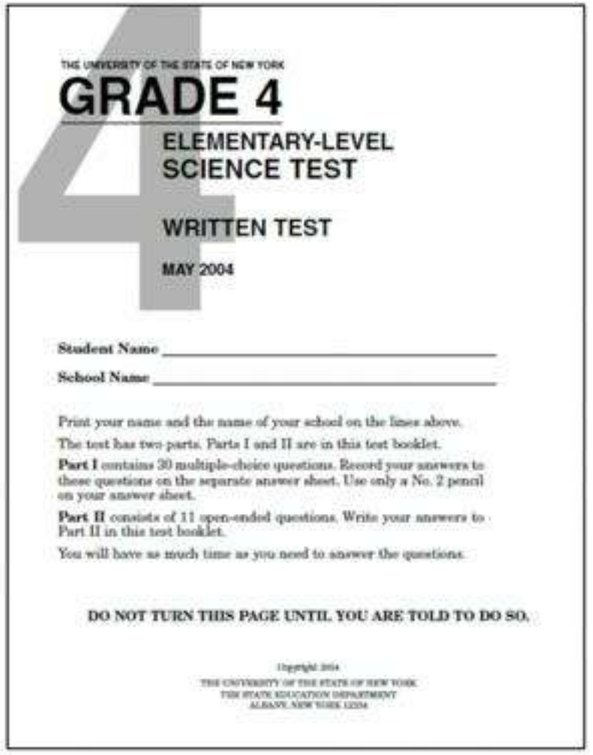
(A) The particles gain mass.
(B) The particles contain less energy.
(C) The particles move more rapidly.
(D) The particles increase in volume.

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Elementary Science Tests as a Grand Challenge

Machines that can:

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



THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 4

ELEMENTARY-LEVEL SCIENCE TEST

WRITTEN TEST

MAY 2004

Student Name _____

School Name _____

Print your name and the name of your school on the lines above.

The test has two parts. Parts I and II are in this test booklet.

Part I contains 30 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.

Part II consists of 11 open-ended questions. Write your answers to Part II in this test booklet.

You will have as much time as you need to answer the questions.

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

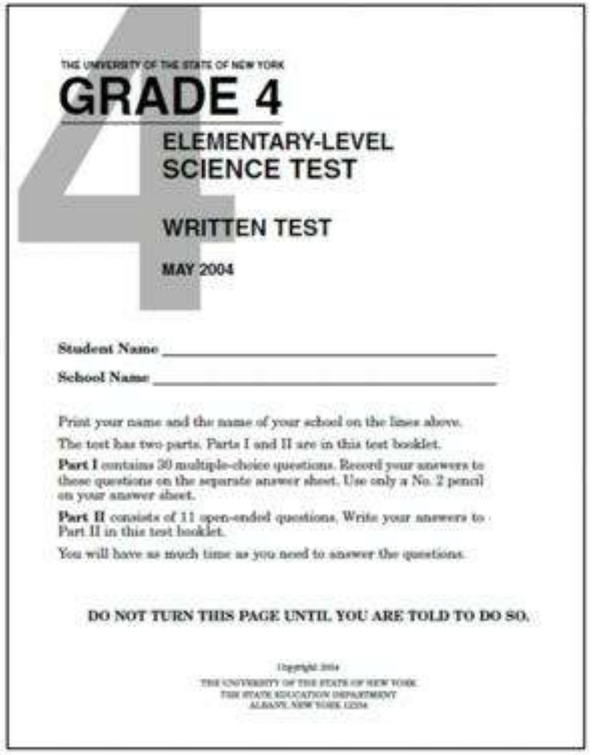
Elementary Science Tests as a Grand Challenge

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!



THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 4

ELEMENTARY-LEVEL SCIENCE TEST

WRITTEN TEST

MAY 2004

Student Name _____

School Name _____

Print your name and the name of your school on the lines above.

The test has two parts. Parts I and II are in this test booklet.

Part I contains 30 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.

Part II consists of 11 open-ended questions. Write your answers to Part II in this test booklet.

You will have as much time as you need to answer the questions.

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.

AI2 ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Elementary Science Tests as a Grand Challenge

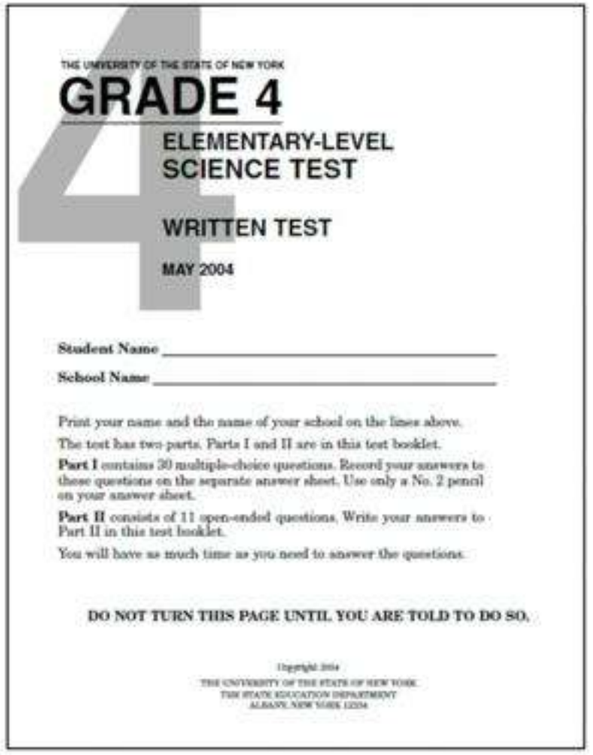
Machines that can:

- 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! ✓

!!!



THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 4

ELEMENTARY-LEVEL SCIENCE TEST

WRITTEN TEST

MAY 2004

Student Name _____

School Name _____

Print your name and the name of your school on the lines above.

The test has two parts. Parts I and II are in this test booklet.

Part I contains 30 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.

Part II consists of 11 open-ended questions. Write your answers to Part II in this test booklet.

You will have as much time as you need to answer the questions.

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.

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

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**

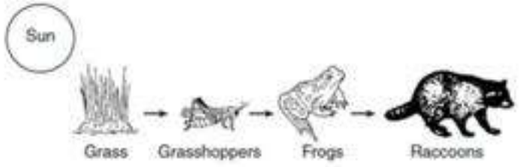
Some Example Questions

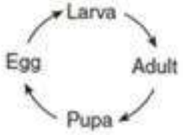
Which object is the best conductor of electricity?
(A) a wax crayon (B) a plastic spoon
(C) a rubber eraser **(D) an iron nail**

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

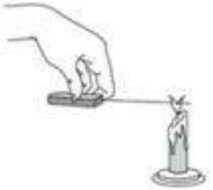
- Diagrams

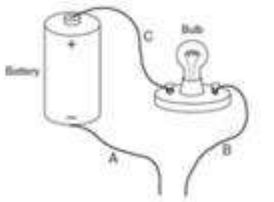


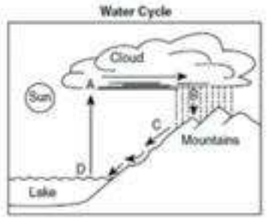


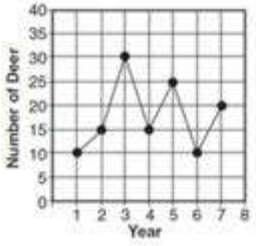
Date	Sunrise	Sunset
February 8	7:00 a.m.	5:20 p.m.
February 15	6:50 a.m.	5:30 p.m.
February 22	6:40 a.m.	5:40 p.m.

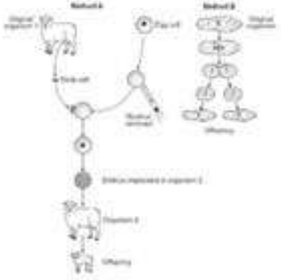
Source: www.sunrisesunset.com

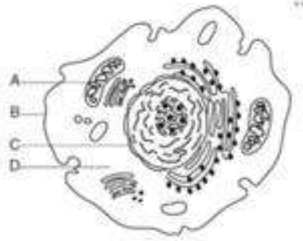


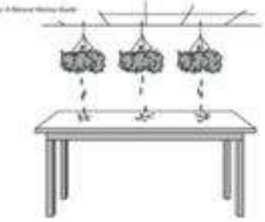


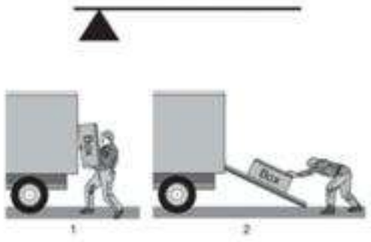






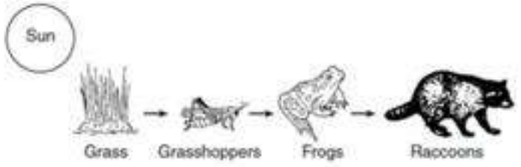


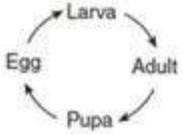




Question Categories Not Covered

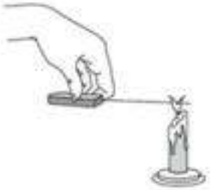
- Diagrams

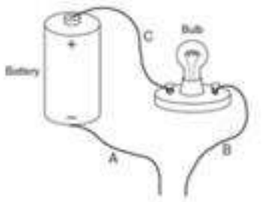


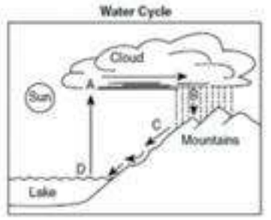


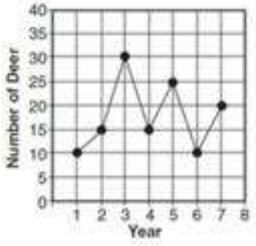
Date	Sunrise	Sunset
February 8	7:00 a.m.	5:20 p.m.
February 15	6:50 a.m.	5:30 p.m.
February 22	6:40 a.m.	5:40 p.m.

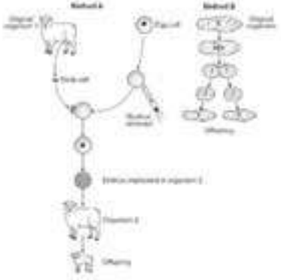
Source: www.sunrisesunset.com

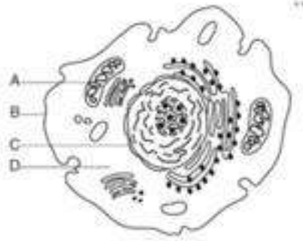


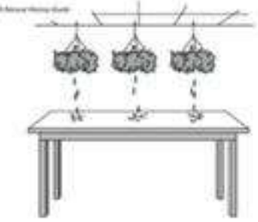


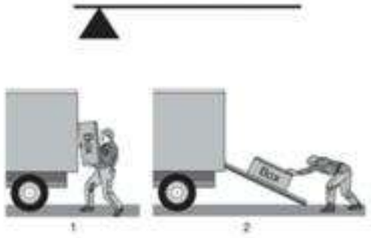






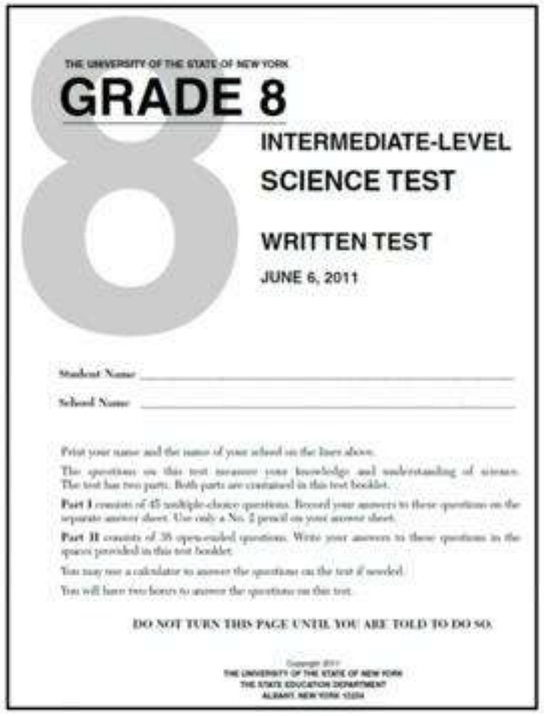






- Direct Answer Questions

Progression on NY Regents 8th Grade (NDMC)



THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 8

INTERMEDIATE-LEVEL SCIENCE TEST

WRITTEN TEST

JUNE 6, 2011

Student Name _____

School Name _____

Print your name and the name of your school on the lines above.

The questions on this test measure your knowledge and understanding of science. The test has two parts. Both parts are contained in this test booklet.

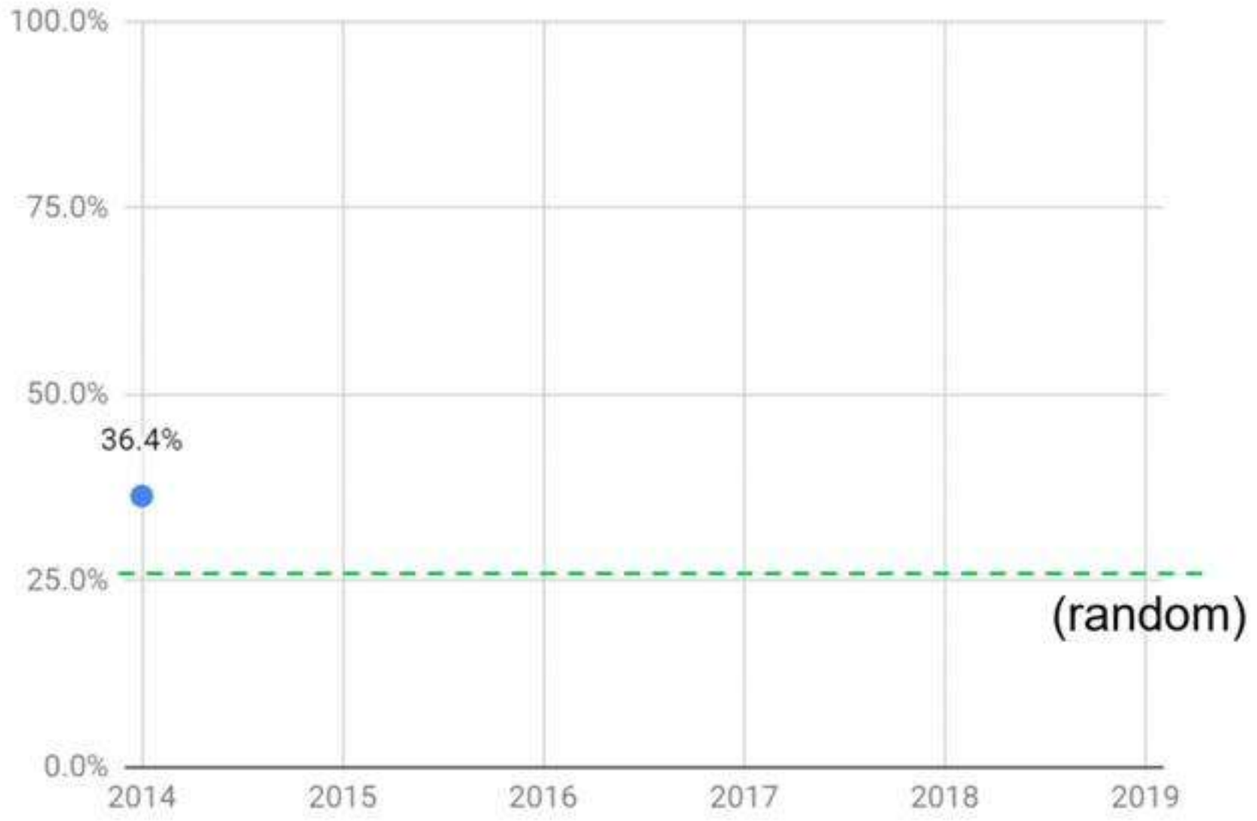
Part I consists of 45 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.

Part II consists of 26 open-ended questions. Write your answers to these questions in the spaces provided in this test booklet.

You may use a calculator to answer the questions on the test if needed.

You will have two hours to answer the questions on this test.

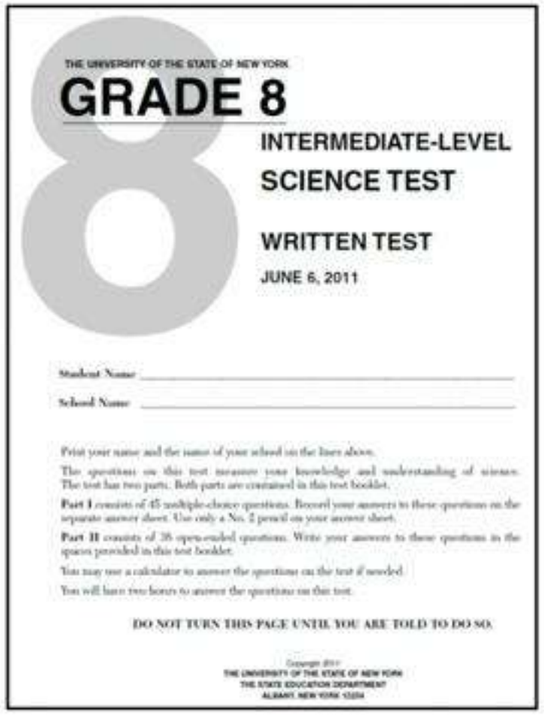
DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.



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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Progression on NY Regents 8th Grade (NDMC)



THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 8

INTERMEDIATE-LEVEL SCIENCE TEST

WRITTEN TEST

JUNE 6, 2011

Student Name _____

School Name _____

Print your name and the name of your school on the lines above.

This questions on this test measure your knowledge and understanding of science. The test has two parts. Both parts are contained in this test booklet.

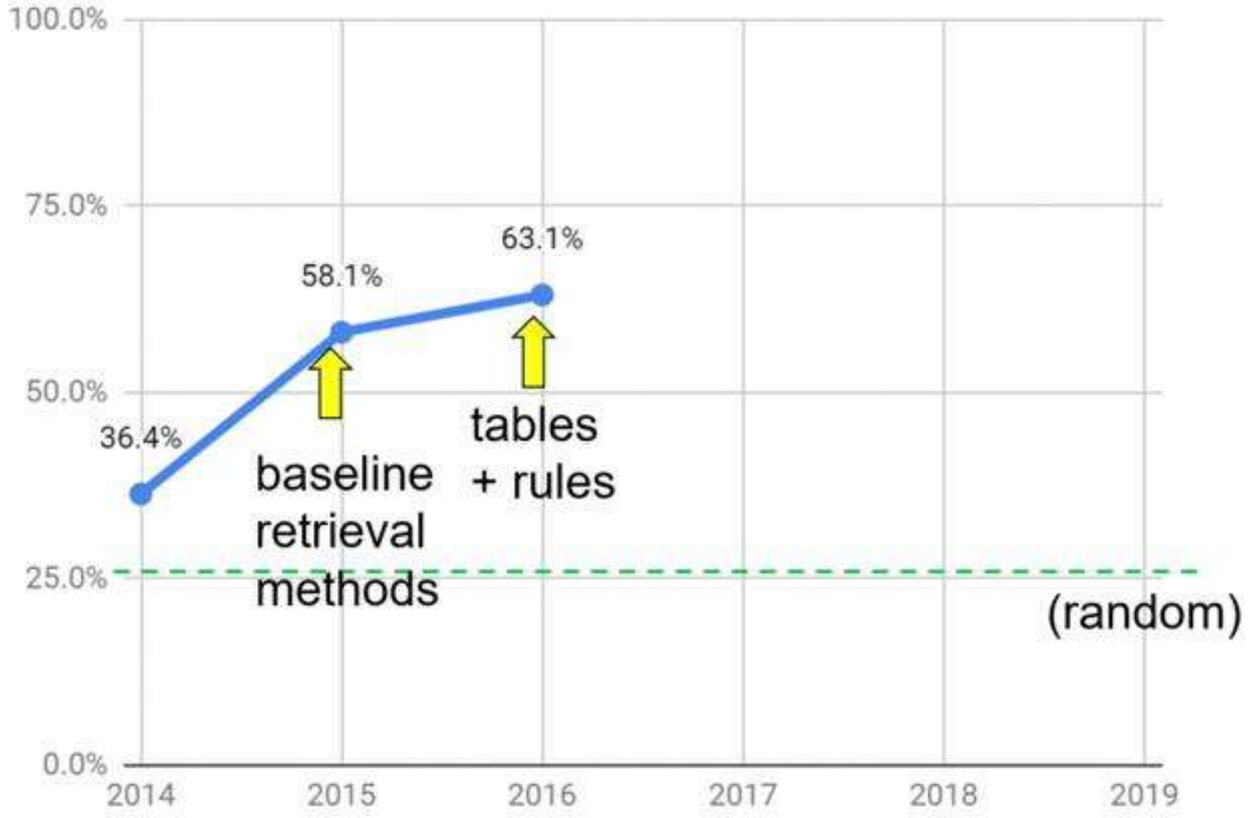
Part I consists of 45 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.

Part II consists of 26 open-ended questions. Write your answers to these questions in the spaces provided in this test booklet.

You may use a calculator to answer the questions on the test if needed.

You will have two hours to answer the questions on this test.

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.



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

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

$80,000 • 119 teams

The Allen AI Science Challenge

Merger and 1st Submission Deadline

Wed 7 Oct 2015 — Sat 13 Feb 2016 (4.0 days to go)

Dashboard

Home
- Data
- Make a submission

Information
- Description
- Evaluation
- Rules
- Prizes
- Timeline

Forum

Leaderboard

Public Leaderboard

1. amsqr
2. Cardal
3. poweredByTalkwalker
4. Generation Gap
5. yamayamada

Competition Details » Get the Data » Make a submission

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

Progression on NY Regents 8th Grade (NDMC)


THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 8

INTERM

SCIEN

WRITTE

JUNE 6, 20

Student Name _____

School Name _____

Print your name and the name of your school on the lines

The questions on this test measure your knowledg

The test has two parts. Both parts are contained in this bo

Part I consists of 45 multiple-choice questions. Record y

separate answer sheet. Use only a No. 2 pencil on your an

Part II consists of 36 open-ended questions. Write you

spaces provided in this test booklet.

You may use a calculator to answer the questions on the t

You will have two hours to answer the questions on this te

DO NOT TURN THIS PAGE UNTIL Y

Copyright 201

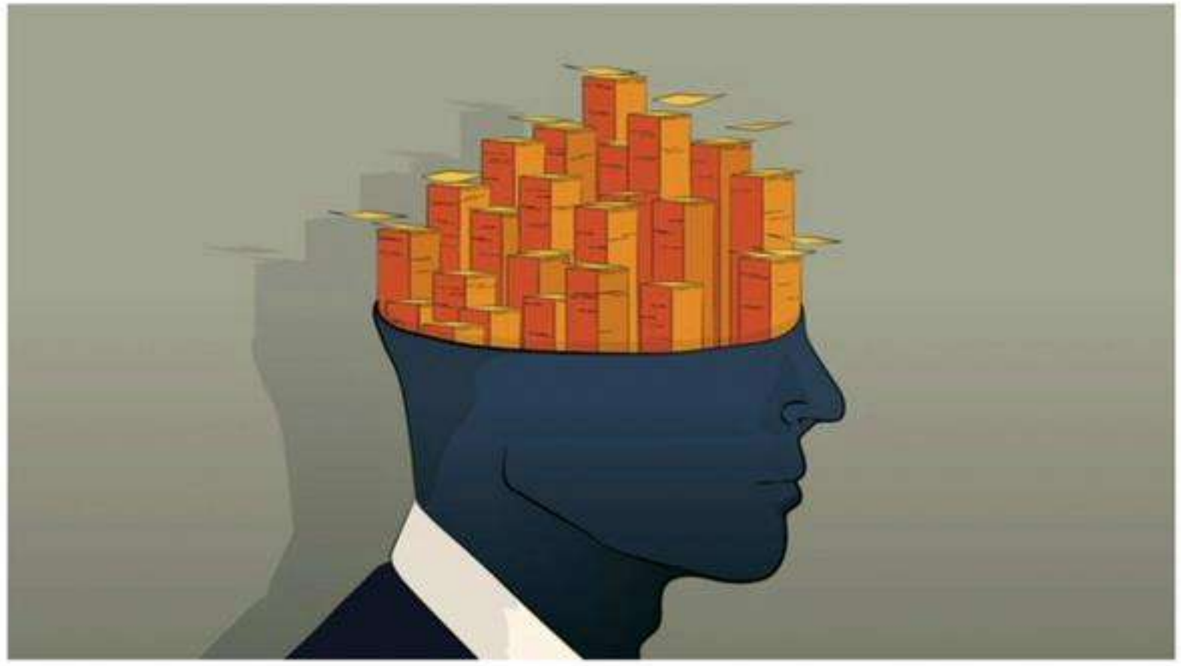
THE UNIVERSITY OF THE STAT

THE STATE EDUCATION DE

ALBANY, NEW YORK

CADE METZ BUSINESS 02.16.16 09:00 AM

THE BEST AI STILL FLUNKS 8TH GRADE SCIENCE

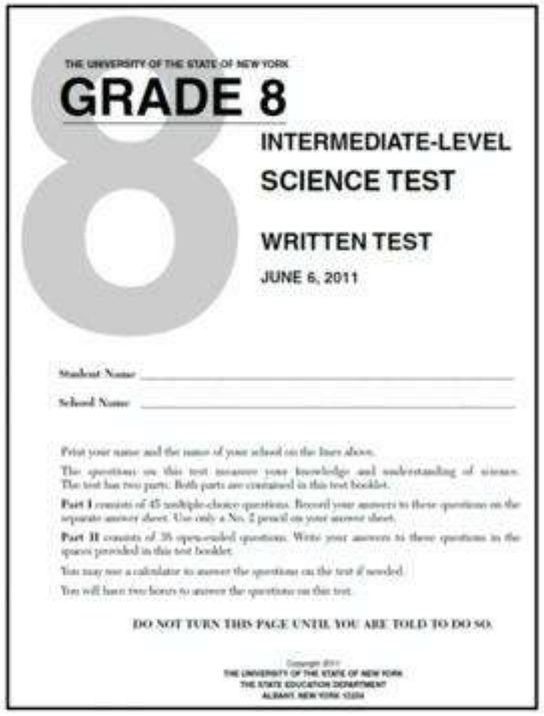


THEN ONE/WIRED

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

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

Progression on NY Regents 8th Grade (NDMC)



THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 8

INTERMEDIATE-LEVEL SCIENCE TEST

WRITTEN TEST

JUNE 6, 2011

Student Name _____

School Name _____

Print your name and the name of your school on the lines above.

The questions on this test measure your knowledge and understanding of science. The test has two parts. Both parts are contained in this test booklet.

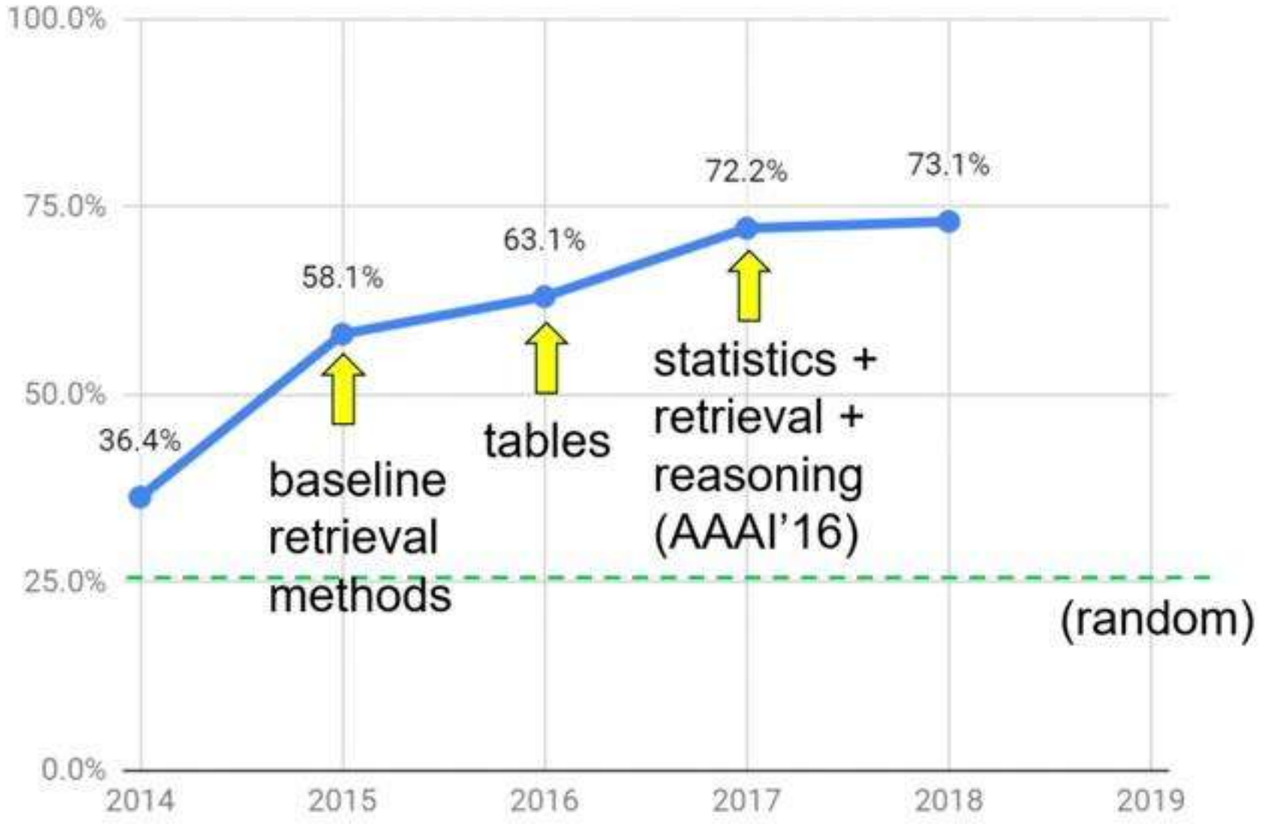
Part I consists of 45 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.

Part II consists of 26 open-ended questions. Write your answers to these questions in the spaces provided in this test booklet.

You may use a calculator to answer the questions on the test if needed.

You will have two hours to answer the questions on this test.

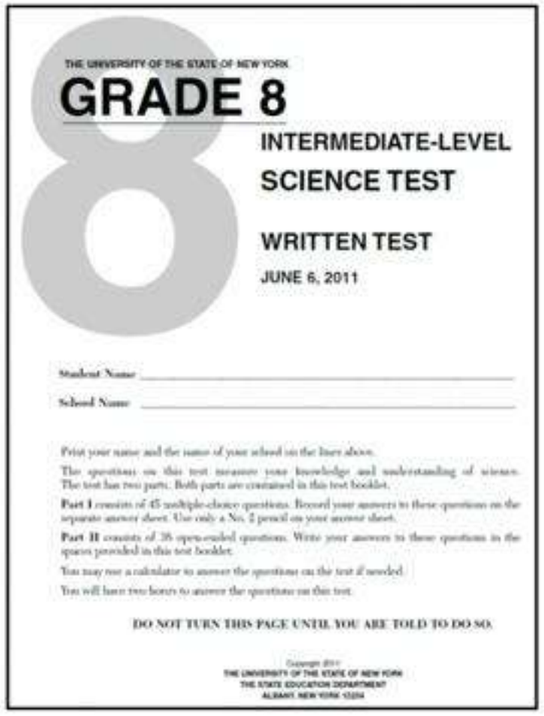
DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.



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

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Progression on NY Regents 8th Grade (NDMC)



THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 8

INTERMEDIATE-LEVEL SCIENCE TEST

WRITTEN TEST

JUNE 6, 2011

Student Name _____

School Name _____

Print your name and the name of your school on the lines above.

The questions on this test measure your knowledge and understanding of science. The test has two parts. Both parts are contained in this test booklet.

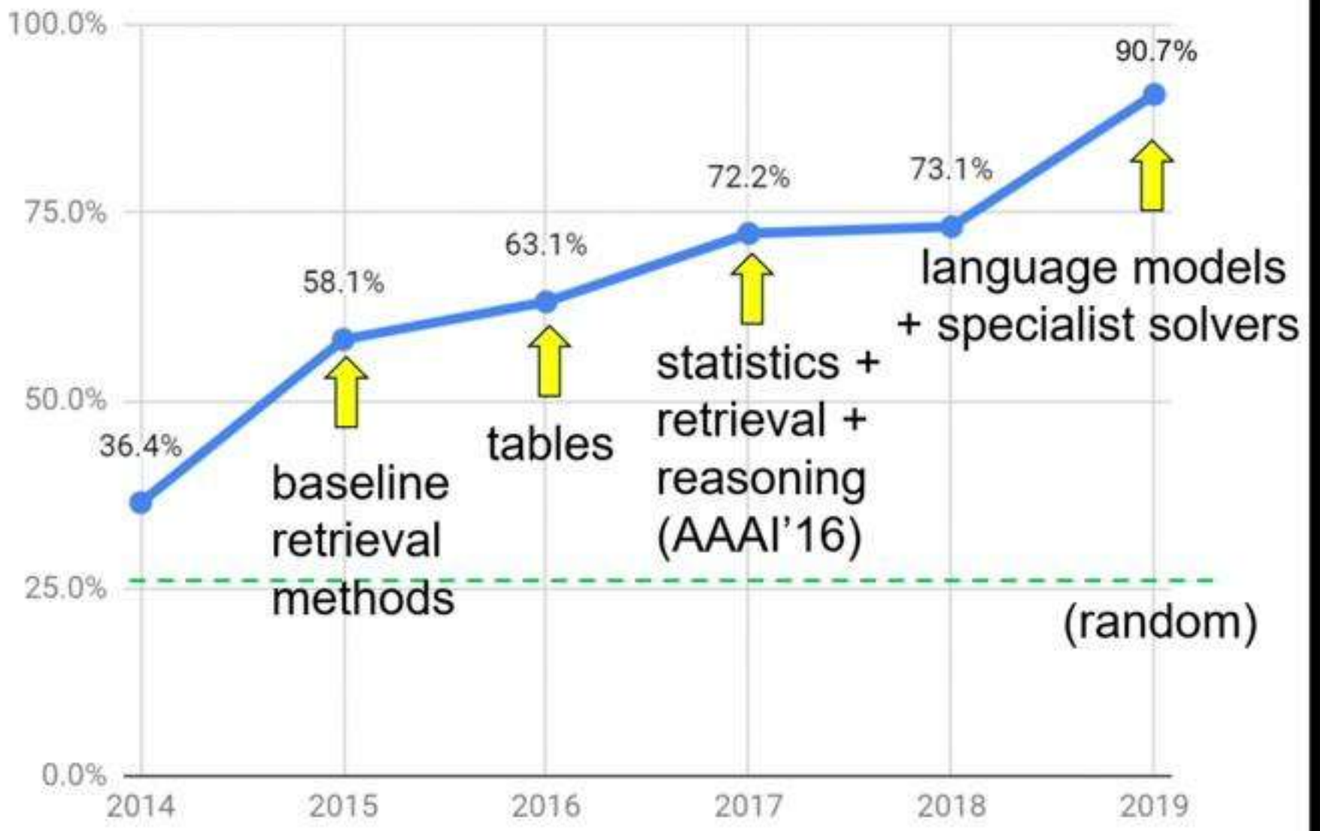
Part I consists of 45 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.

Part II consists of 26 open-ended questions. Write your answers to these questions in the spaces provided in this test booklet.

You may use a calculator to answer the questions on the test if needed.

You will have two hours to answer the questions on this test.

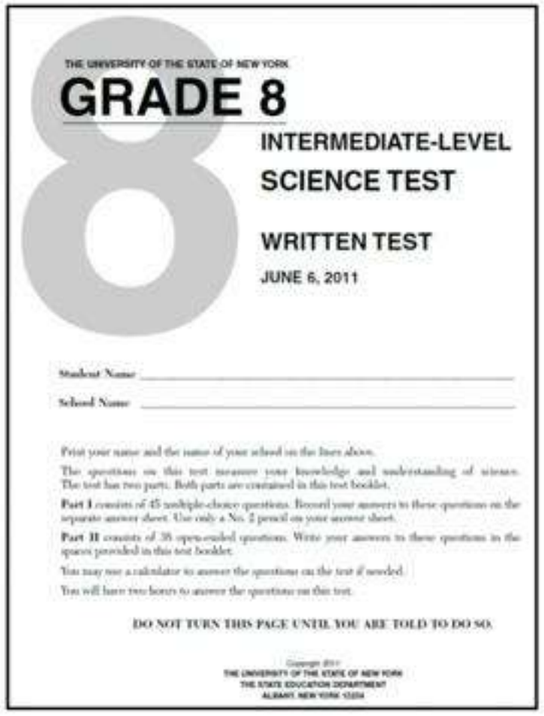
DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.



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

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Progression on NY Regents 8th Grade (NDMC)



THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 8

INTERMEDIATE-LEVEL SCIENCE TEST

WRITTEN TEST

JUNE 6, 2011

Student Name _____

School Name _____

Print your name and the name of your school on the lines above.

The questions on this test measure your knowledge and understanding of science. The test has two parts. Both parts are contained in this test booklet.

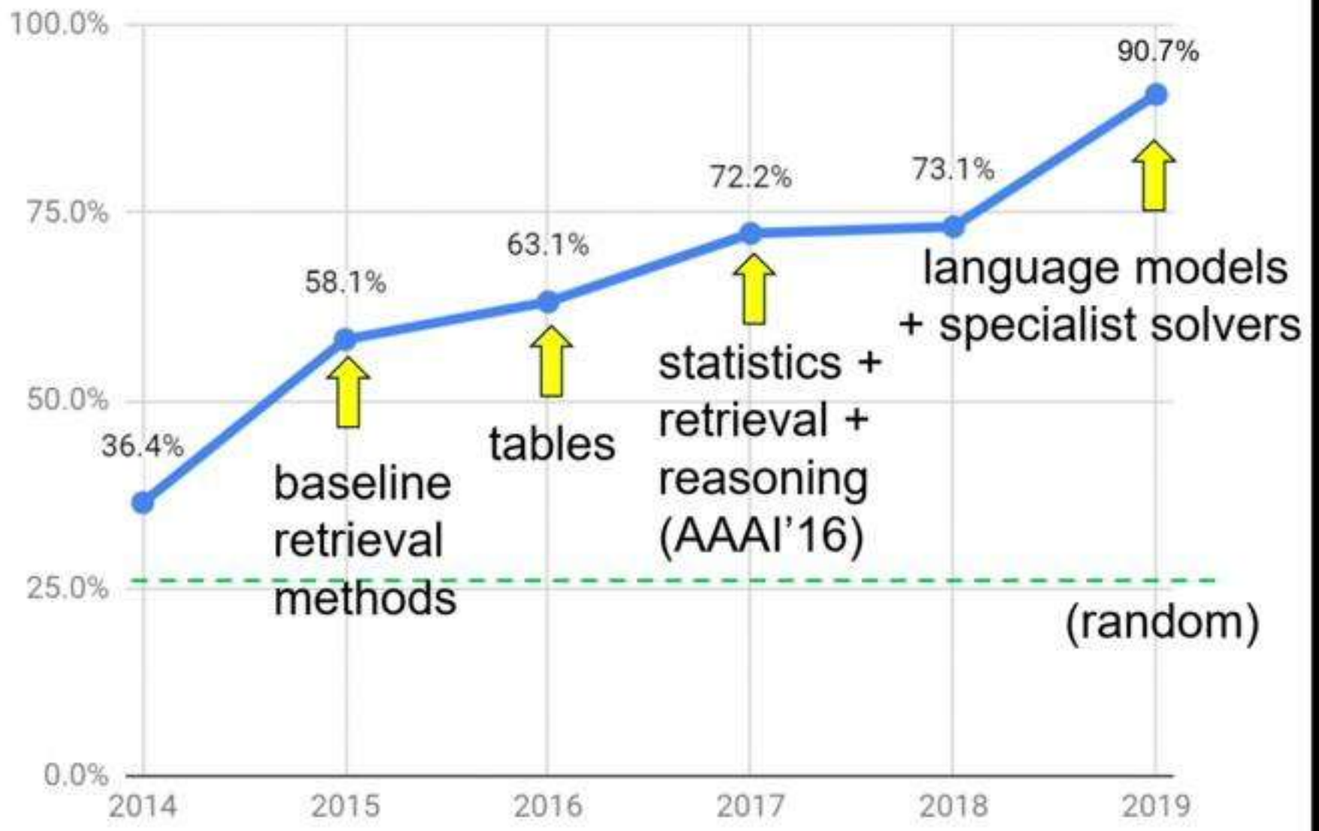
Part I consists of 45 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.

Part II consists of 16 open-ended questions. Write your answers to these questions in the spaces provided in this test booklet.

You may use a calculator to answer the questions on the test if needed.

You will have two hours to answer the questions on this test.

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.

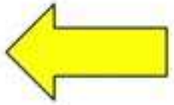


Separate test on 3 latest exams (2017-2019): 93.3%

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

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

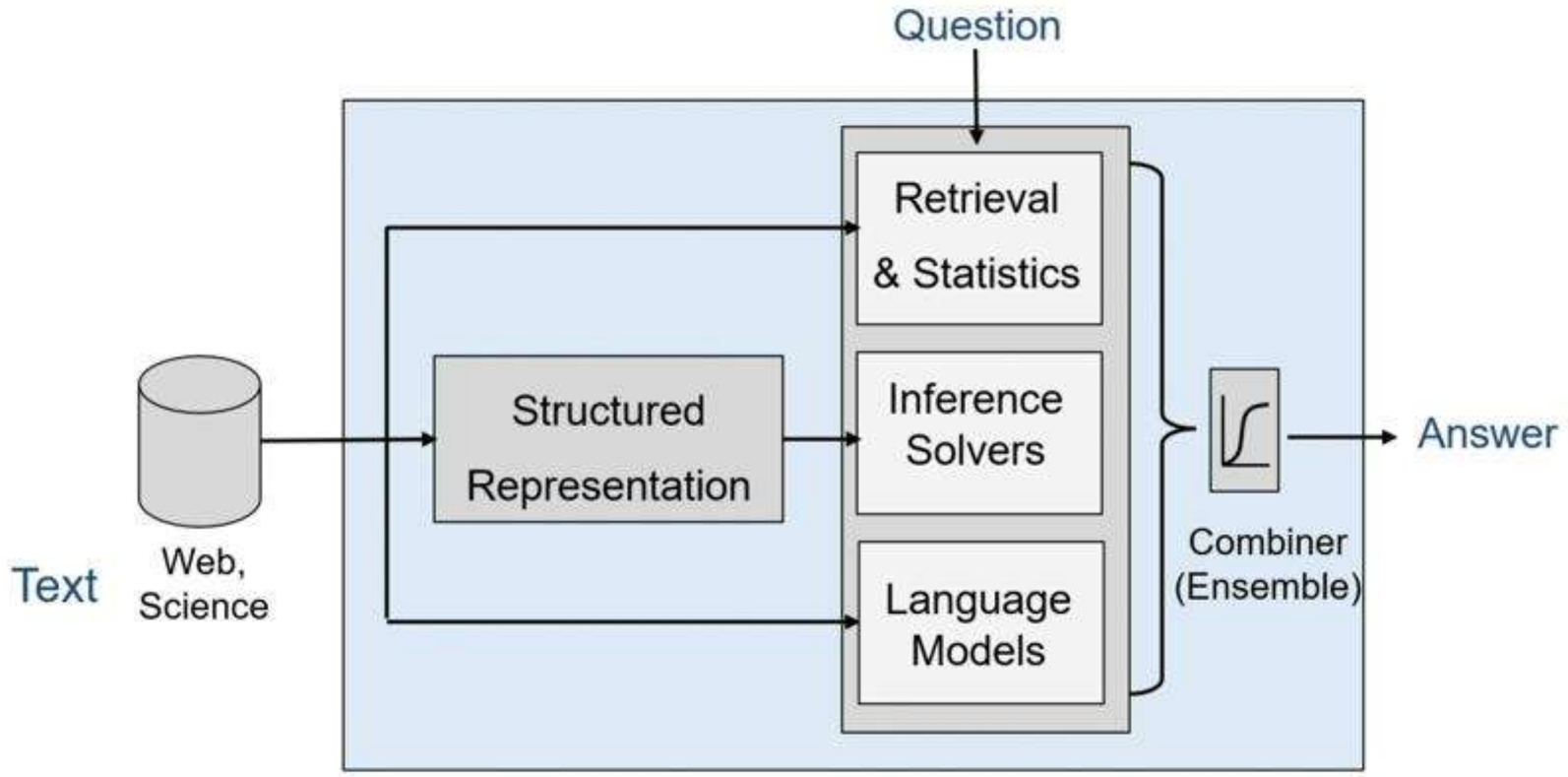
Outline

- 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

ARISTO



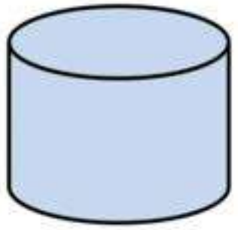
ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Retrieval and Statistical Solvers

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

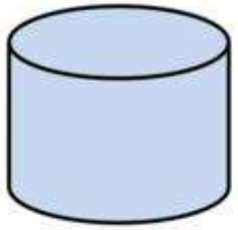
Retrieval and Statistical Solvers

- **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



Retrieval and Statistical Solvers

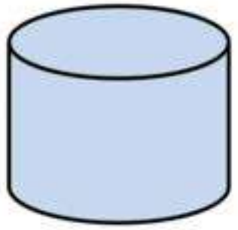
- **Information Retrieval Solver**
  - question + answer option that best matches a corpus sentence
  - Aristo Corpus:
    - Web crawl from Univ Waterloo (330GB)
    - (science parts of) Wikipedia
    - Science textbooks



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

Retrieval and Statistical Solvers

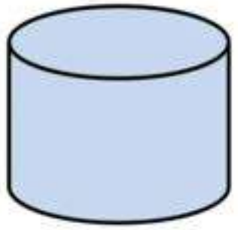
- **Information Retrieval Solver**
  - question + answer option that best matches a corpus sentence
  - Aristo Corpus:
    - Web crawl from Univ Waterloo (330GB)
    - (science parts of) Wikipedia
    - Science textbooks



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

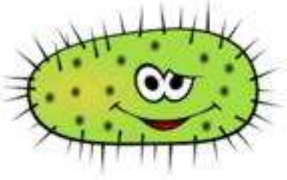
Retrieval and Statistical Solvers

- **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



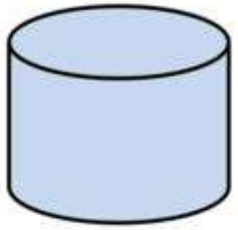
- **PMI**
  - between question and answer words
- **ACME**
  - link question to answer via terms in a ***termbank***
  - heavy use of vector spaces

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



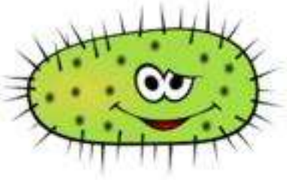
Retrieval and Statistical Solvers

- **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

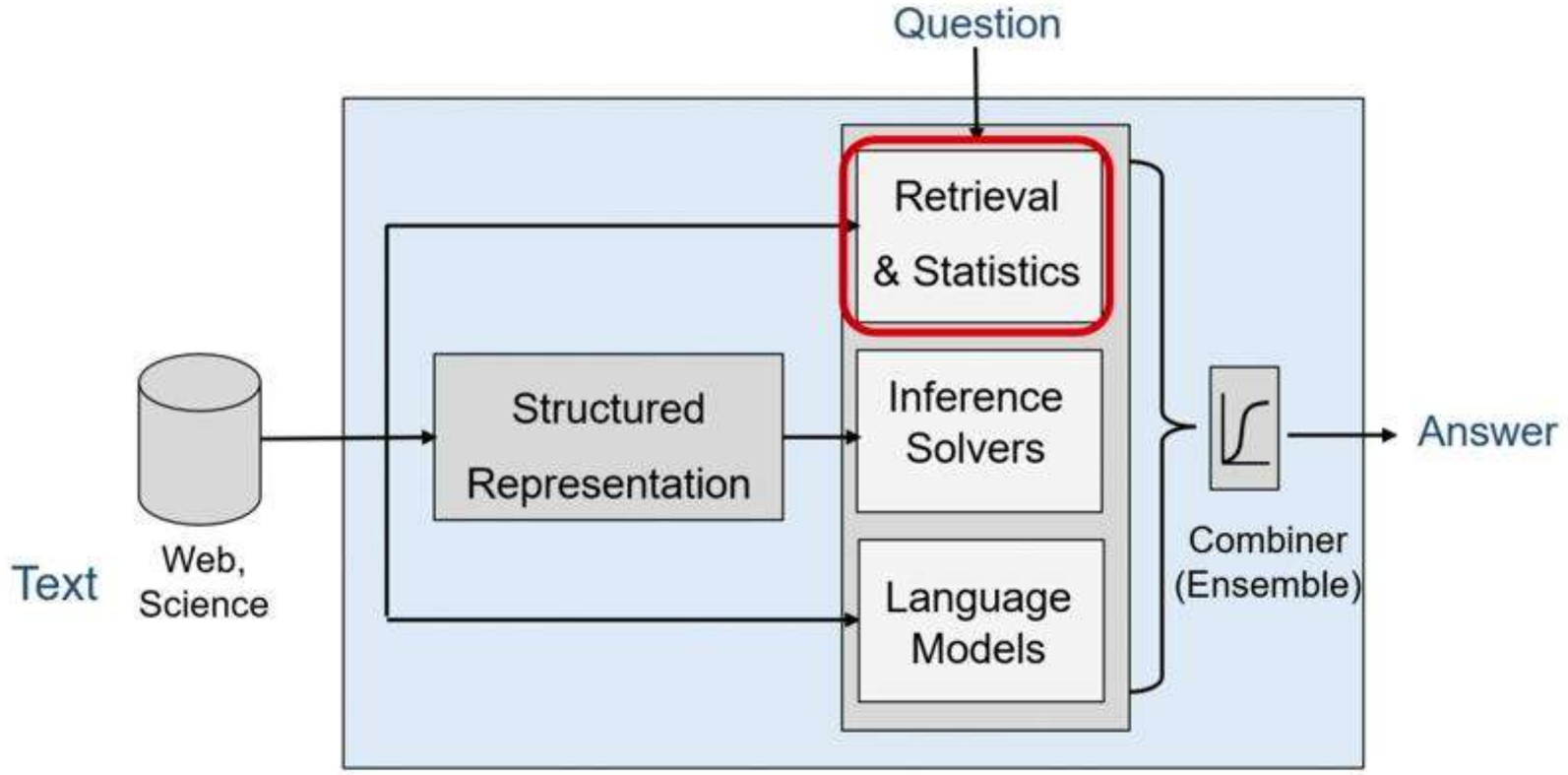


- *Products contaminated with **microorganisms** may cause **infection**.*

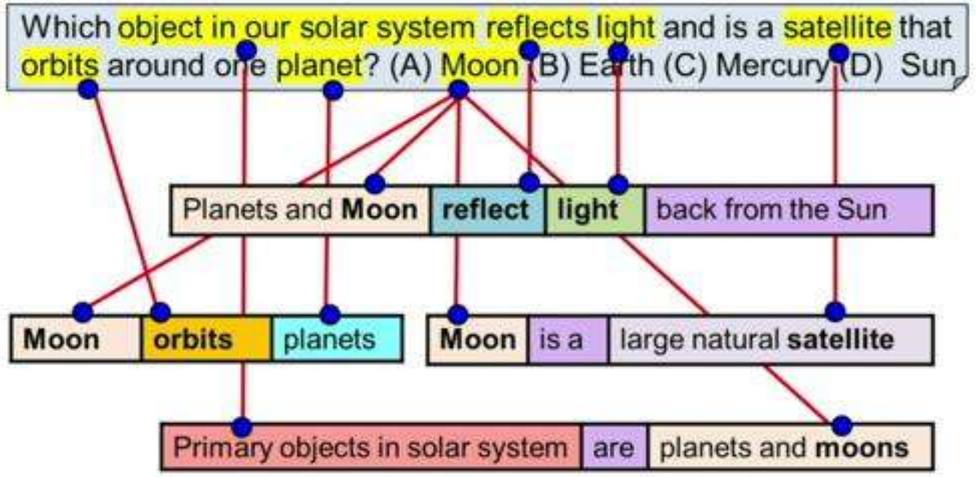
Aristo: an over-simplified overview

- An ensemble architecture

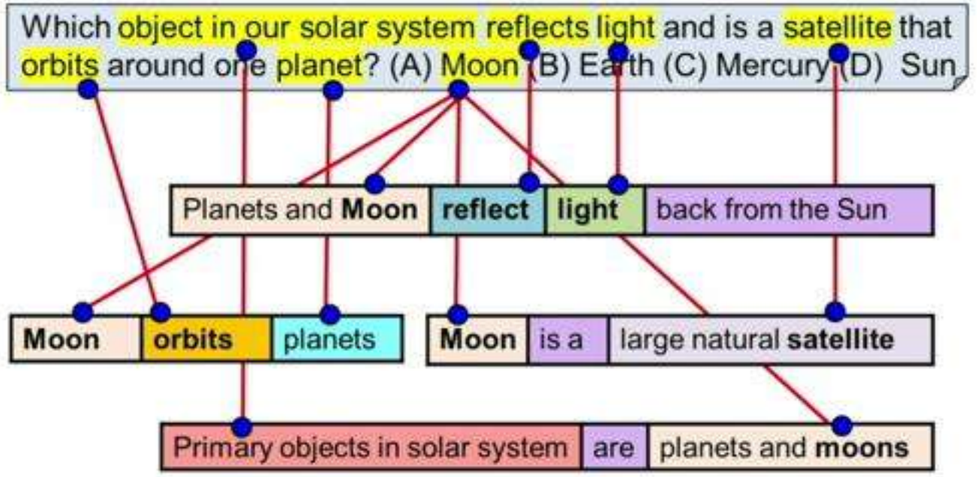
ARISTO



Tuple Inference

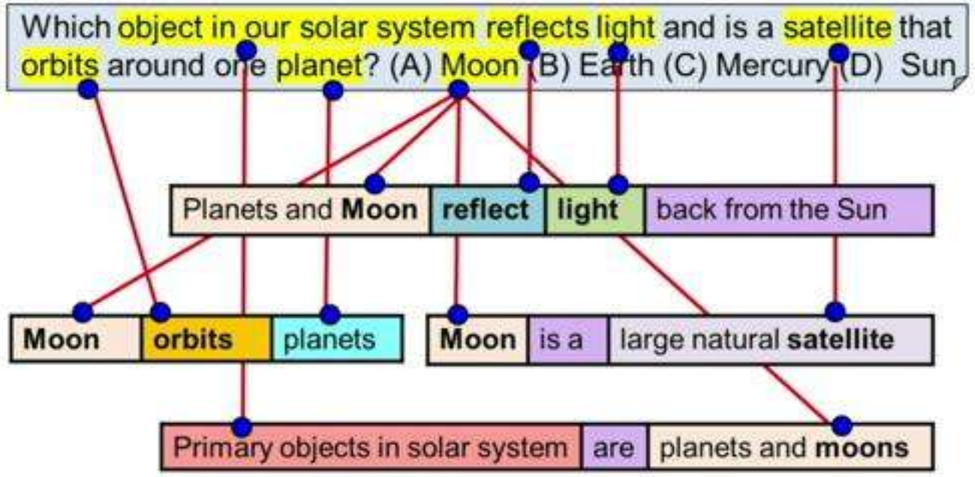


Tuple Inference

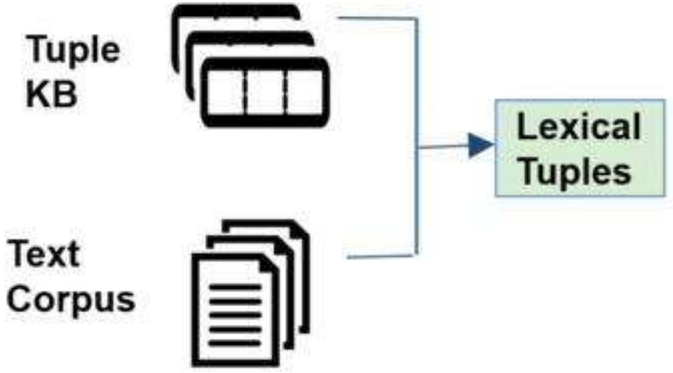


Stormy weather negatively affects a coastline by **(A) causing erosion** (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses

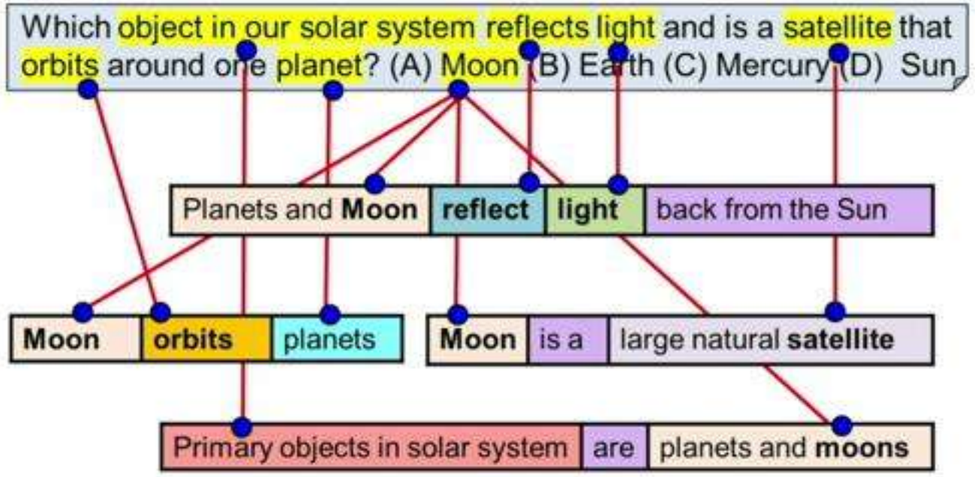
Tuple Inference



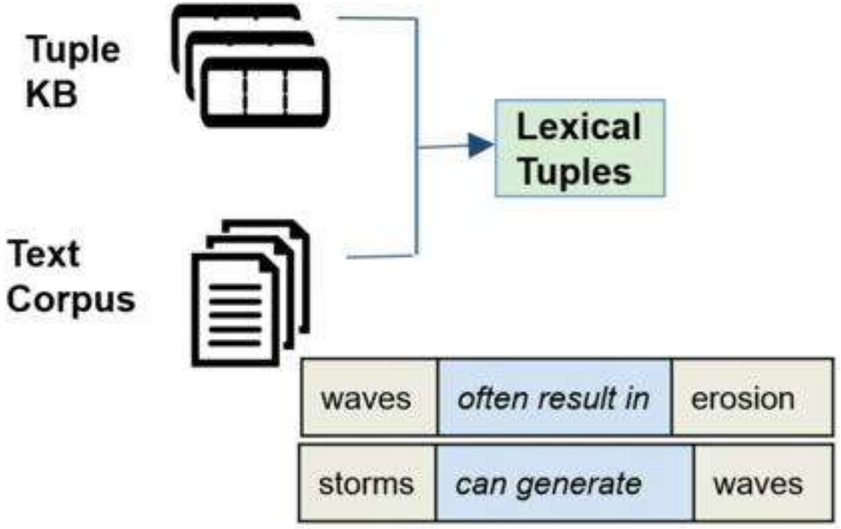
Stormy weather negatively affects a coastline by **(A) causing erosion** (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses



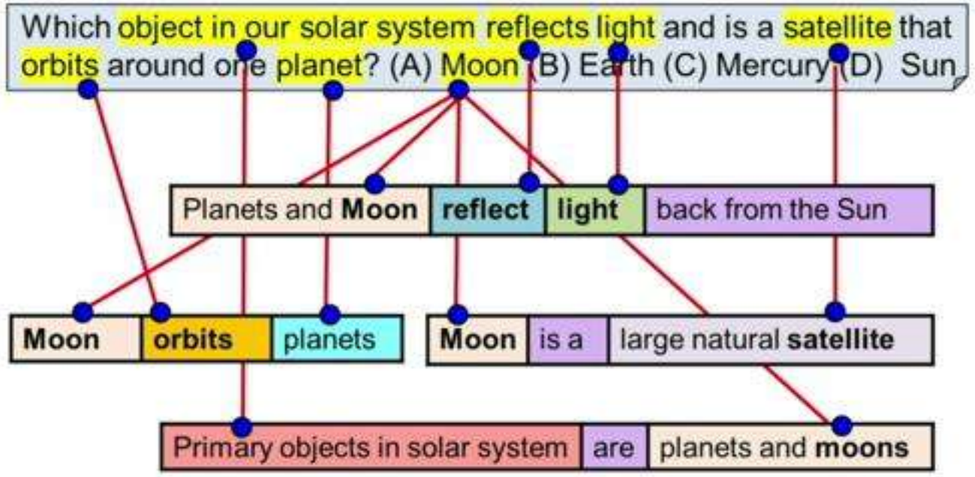
Tuple Inference



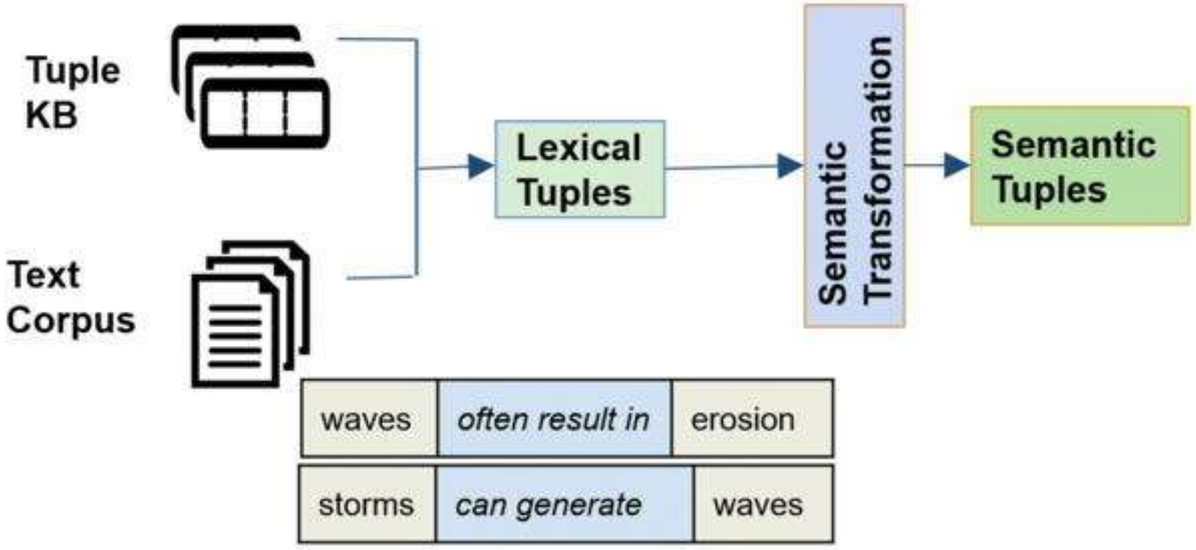
Stormy weather negatively affects a coastline by **(A) causing erosion** (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses



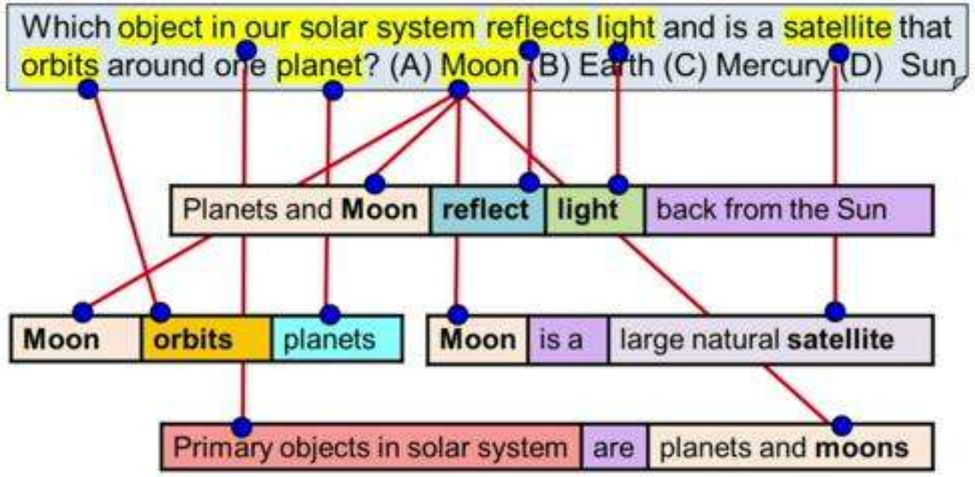
Tuple Inference



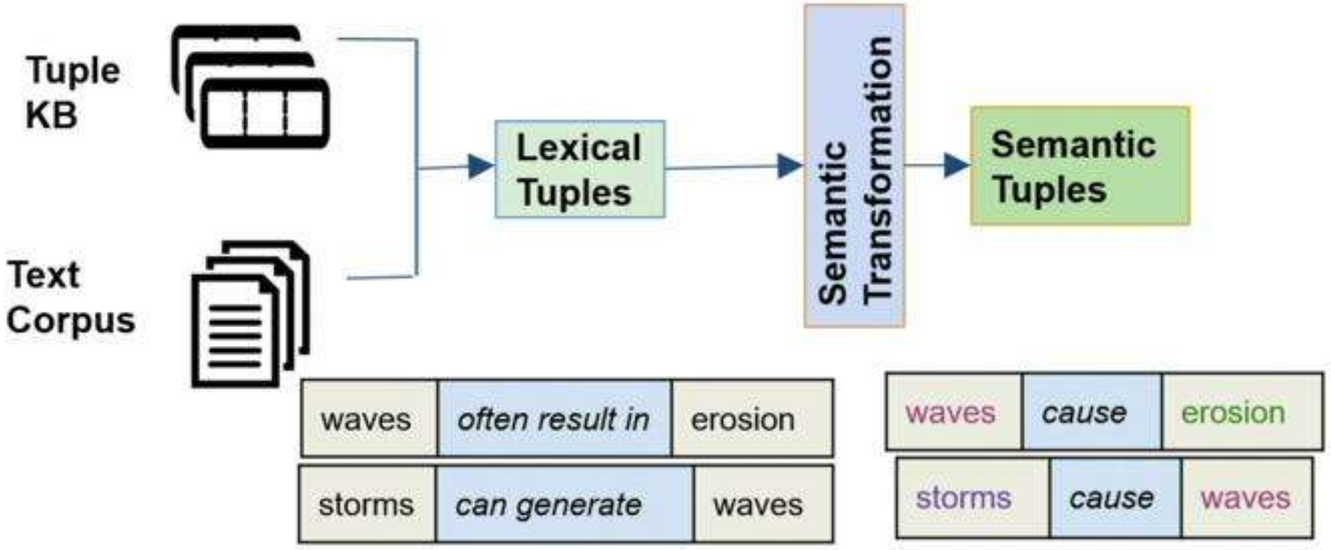
Stormy weather negatively affects a coastline by **(A) causing erosion** (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses



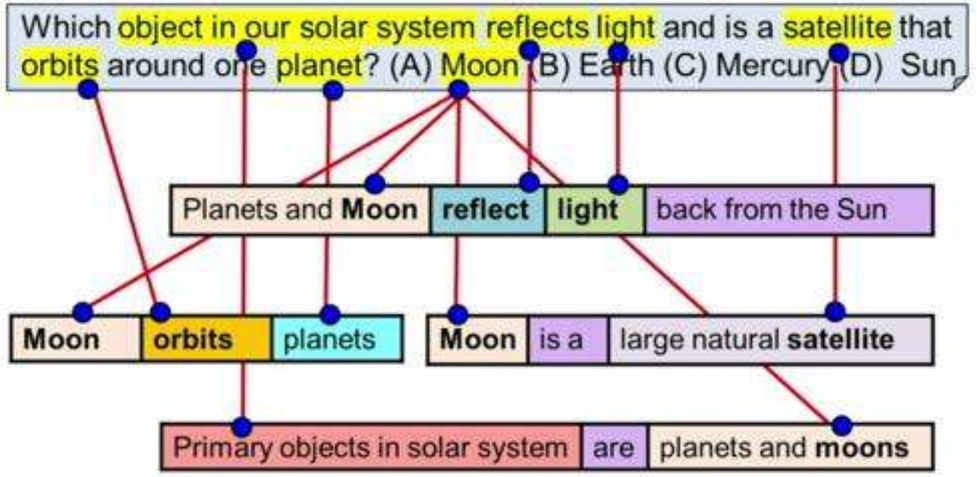
Tuple Inference



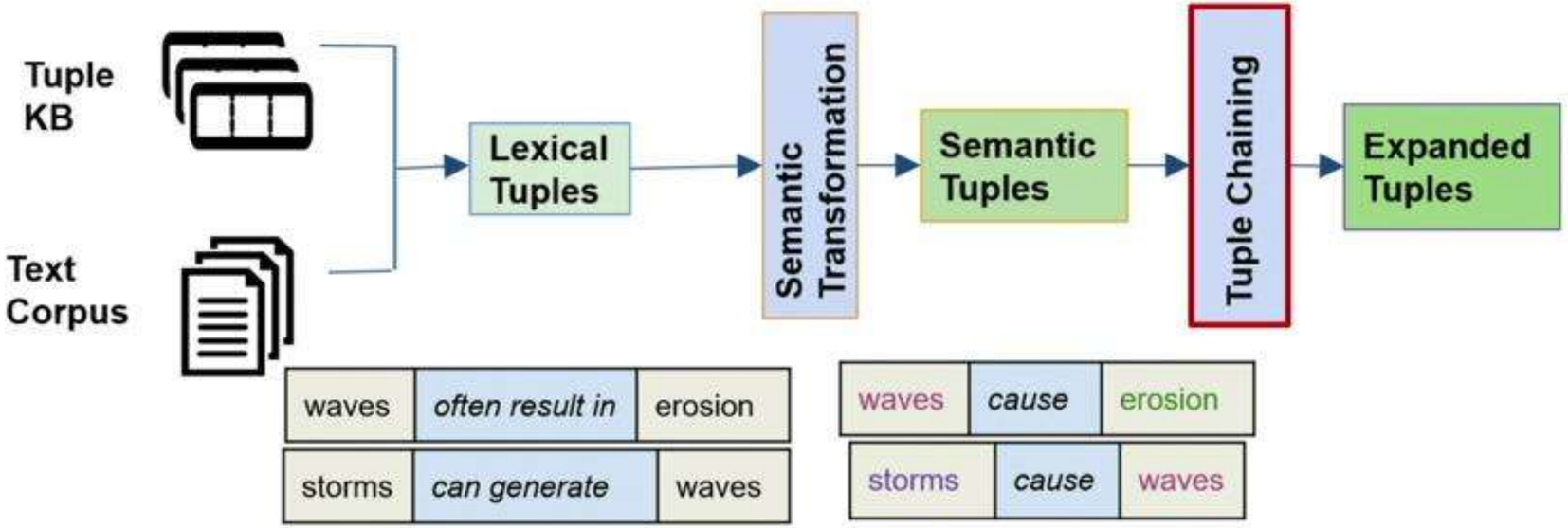
Stormy weather negatively affects a coastline by **(A) causing erosion** (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses



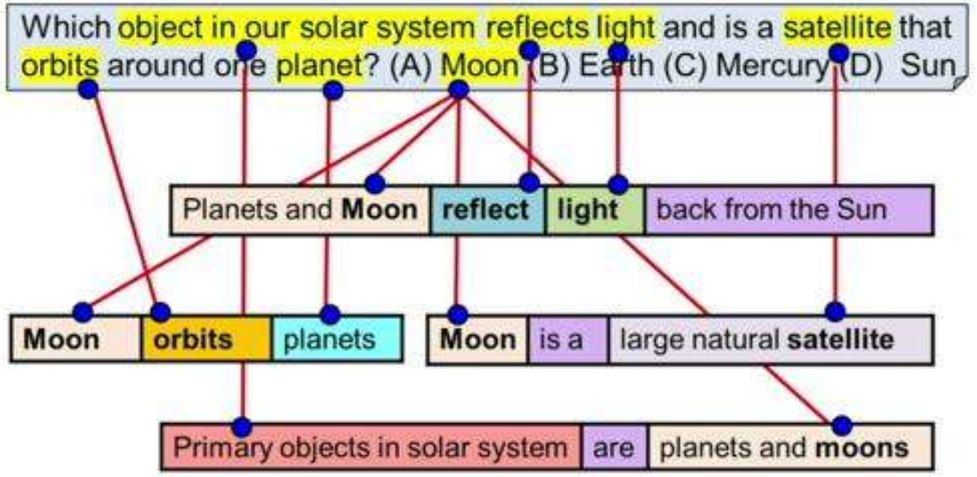
Tuple Inference



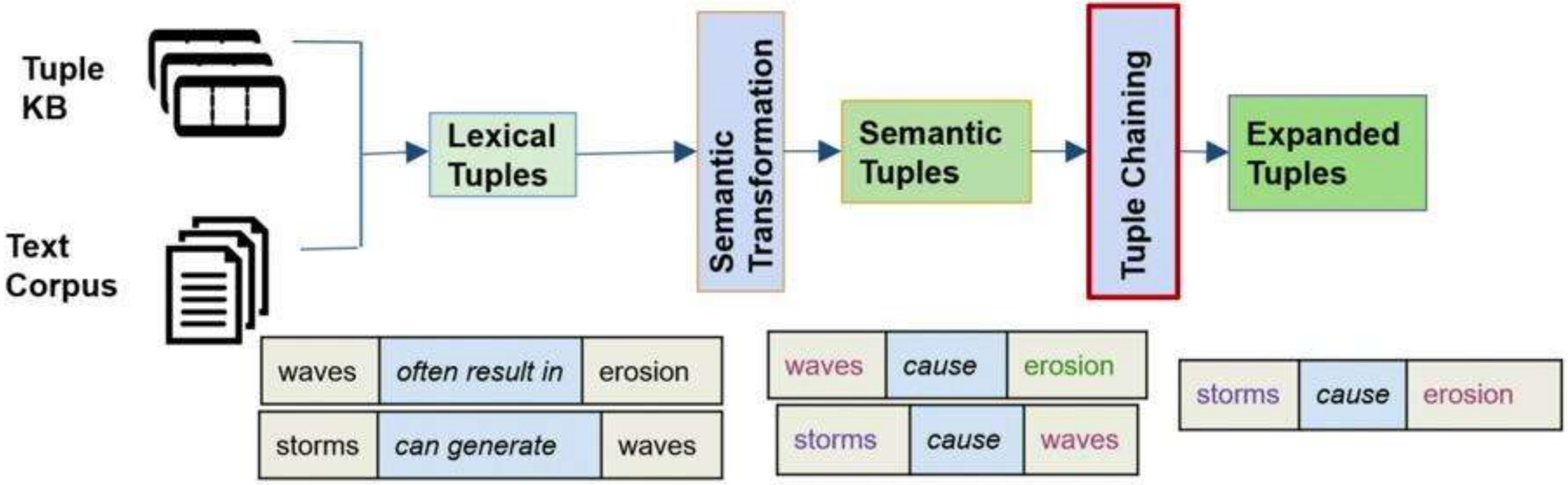
Stormy weather negatively affects a coastline by **(A) causing erosion** (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses



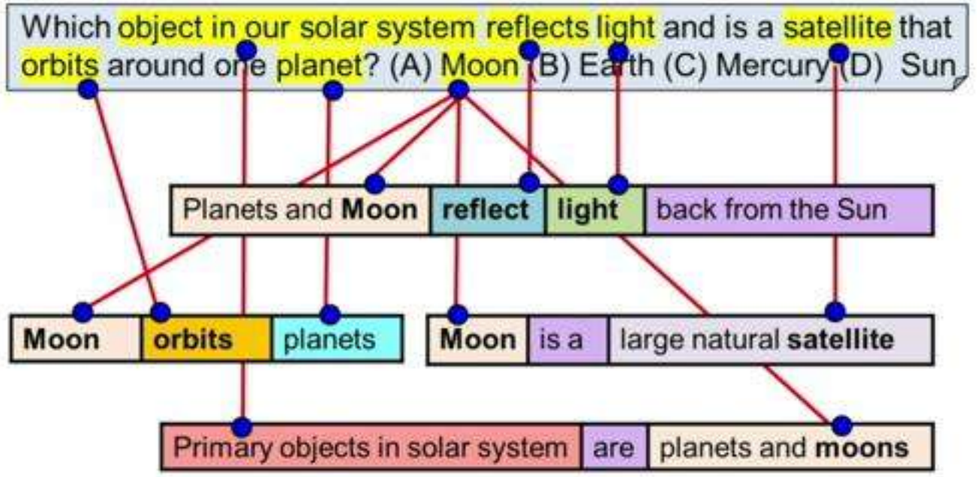
Tuple Inference



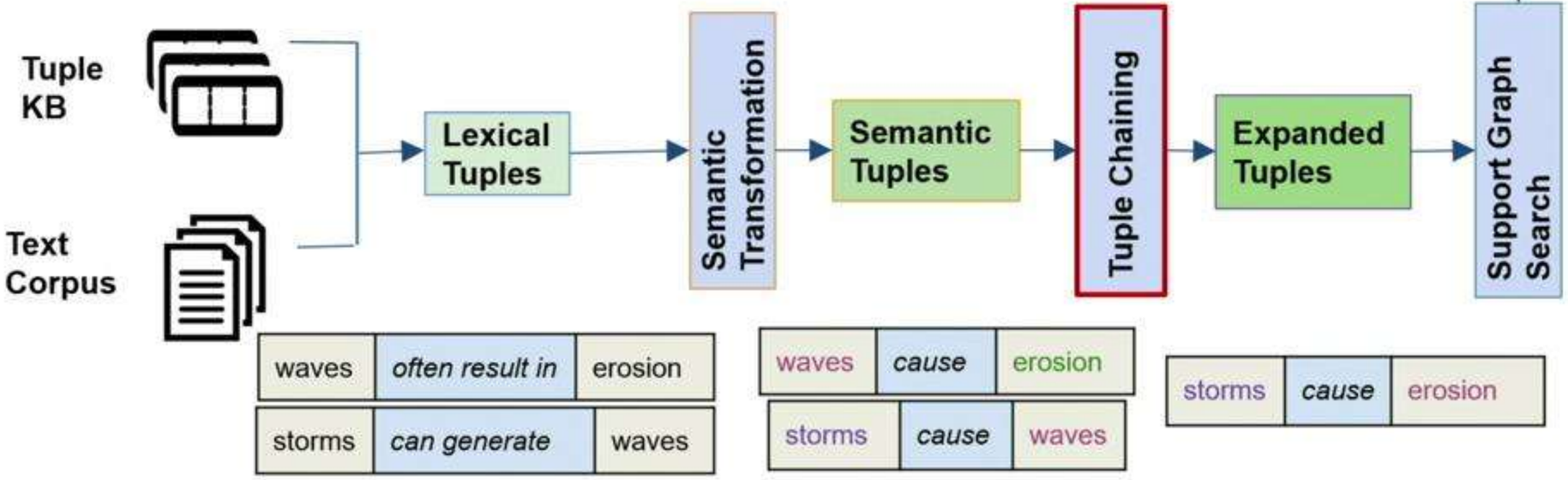
Stormy weather negatively affects a coastline by **(A) causing erosion** (B) causing earthquakes (C) increasing food production (D) increasing the growth of grasses



Tuple Inference



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

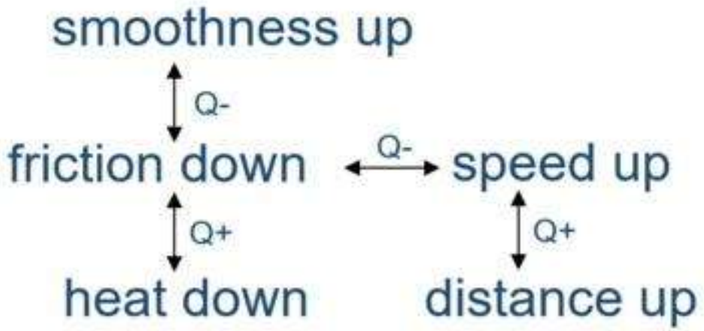


Qualitative Knowledge

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) …

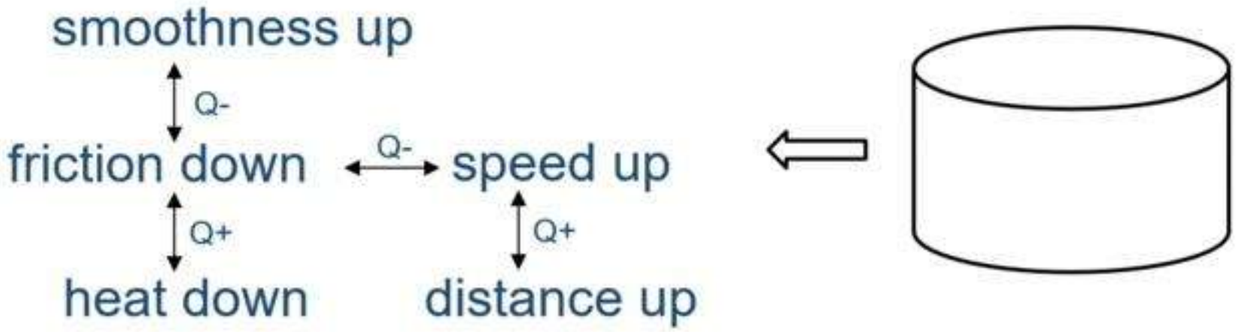
Qualitative Knowledge

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) …



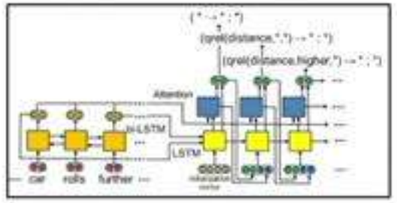
Qualitative Knowledge

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) …



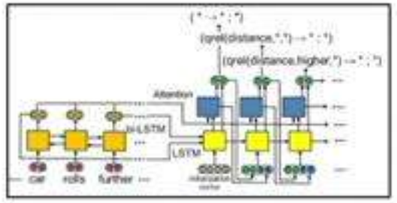
Qualitative Knowledge

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) …



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) …



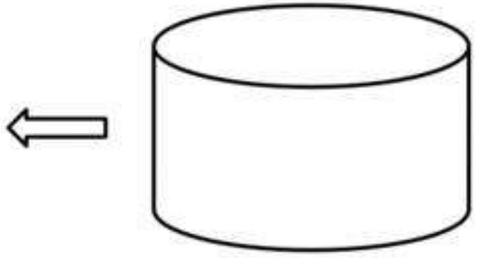
smoothness up

Q-

friction down Q- speed up

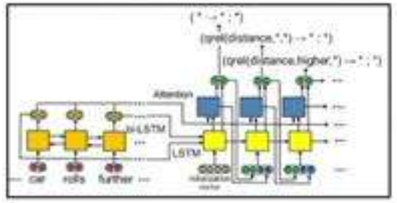
Q+ Q+

heat down **distance up**



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) …



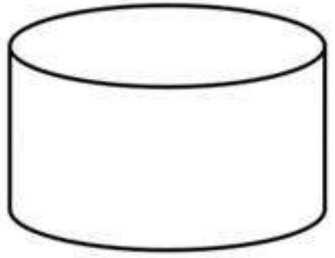
smoothness up

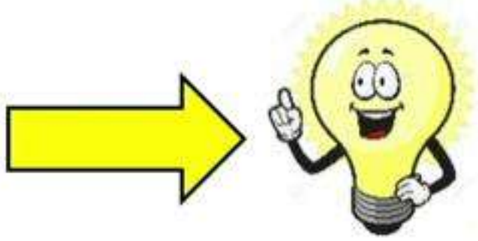
Q-

friction down ↔ speed up

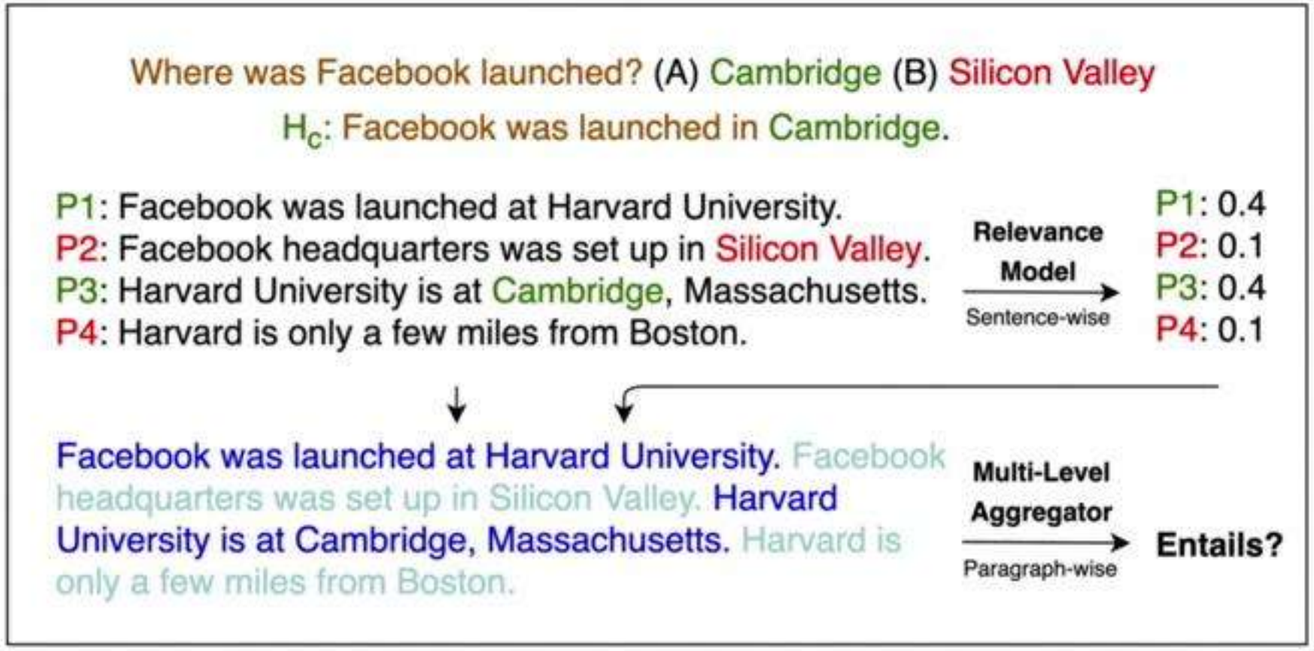
Q+ Q+

heat down **distance up**

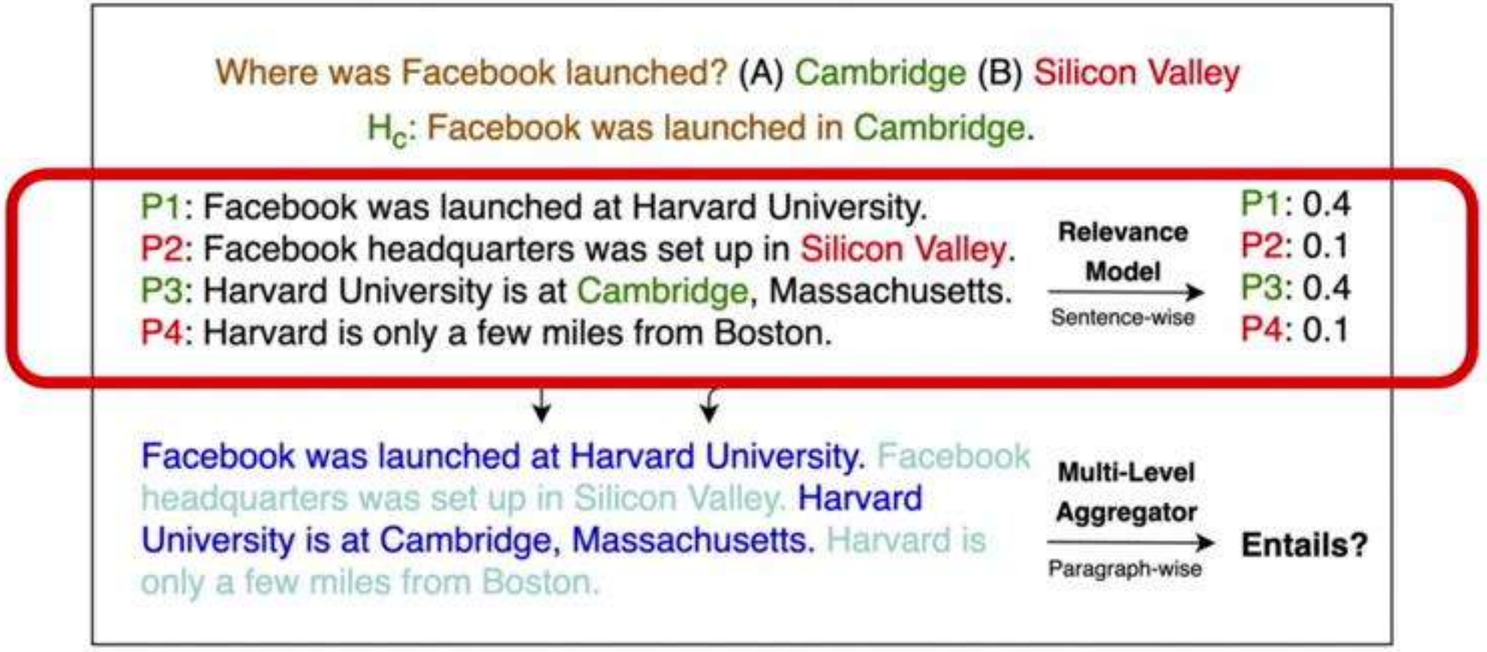




Multee: Repurposing Textual Entailment

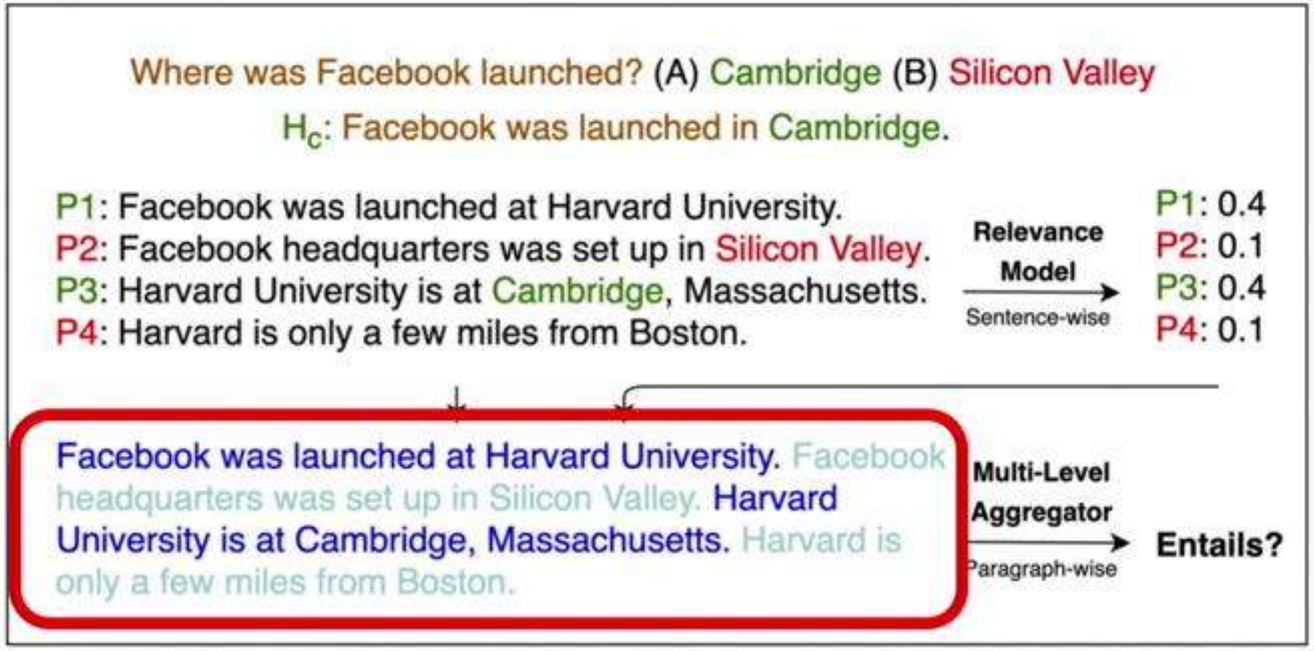


Multee: Repurposing Textual Entailment



ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Multee: Repurposing Textual Entailment



ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Multee: Repurposing Textual Entailment

Multee: Repurposing Textual Entailment

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

Answer: (C) table salt

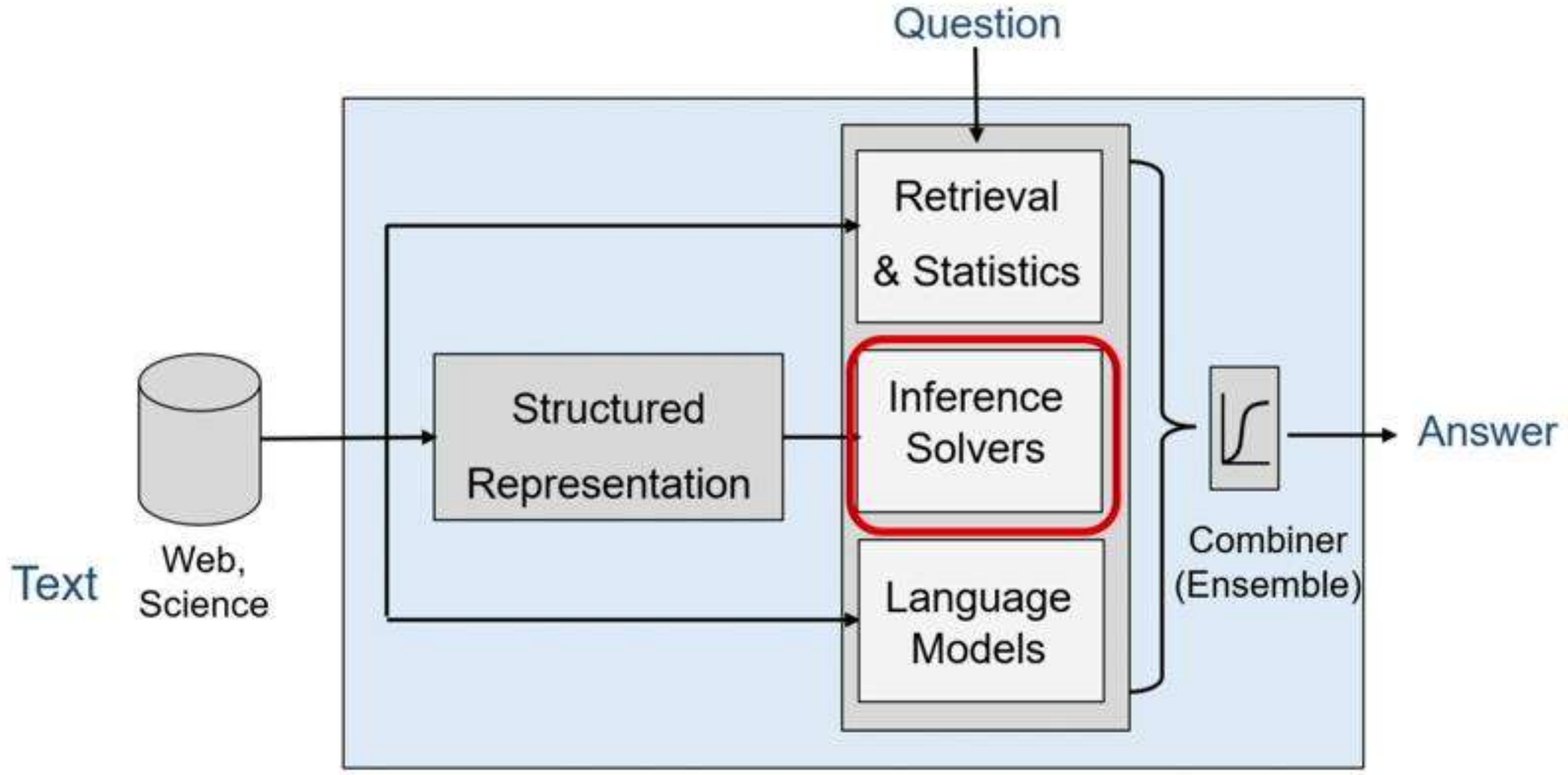
Score	Premise
0.653	Table salt, for instance, is a compound of sodium and chlorine.
0.543	An example of an inorganic substance is table salt.

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Aristo: an over-simplified overview

- An ensemble architecture

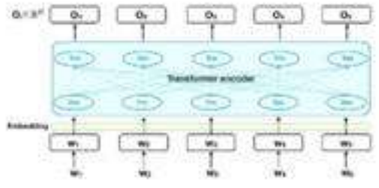
ARISTO



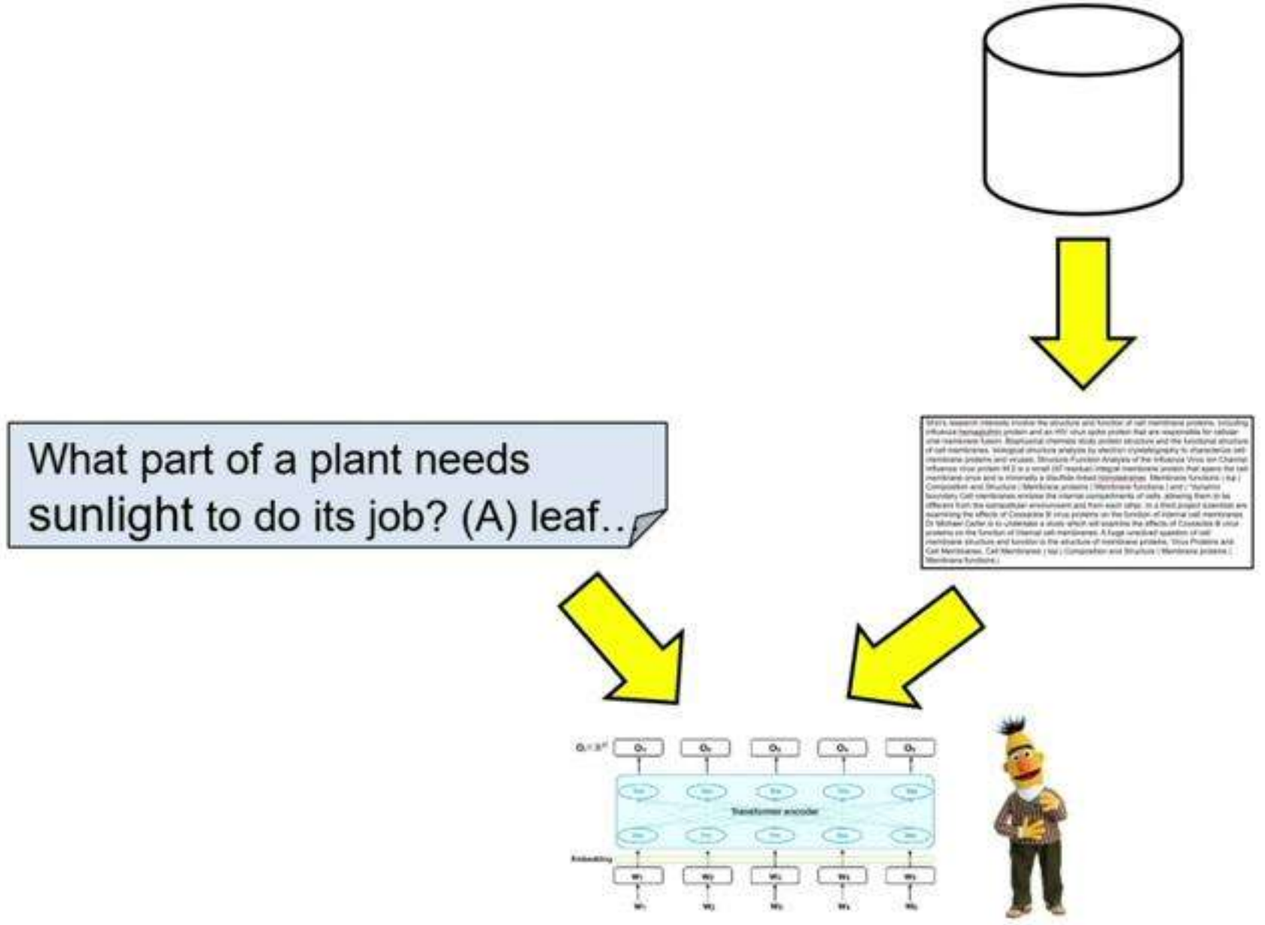
ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

BERT and RoBERTa

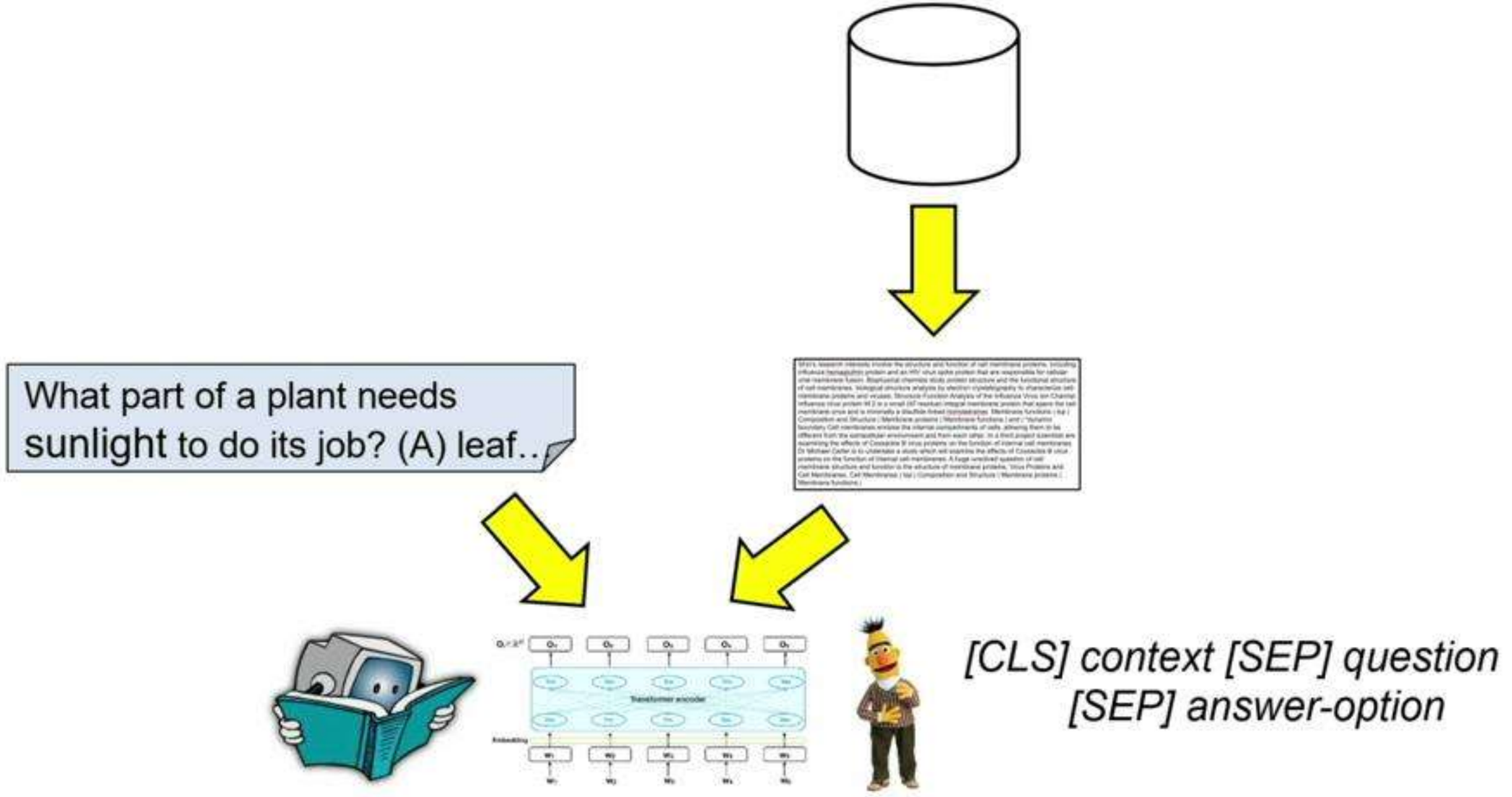
What part of a plant needs sunlight to do its job? (A) leaf..



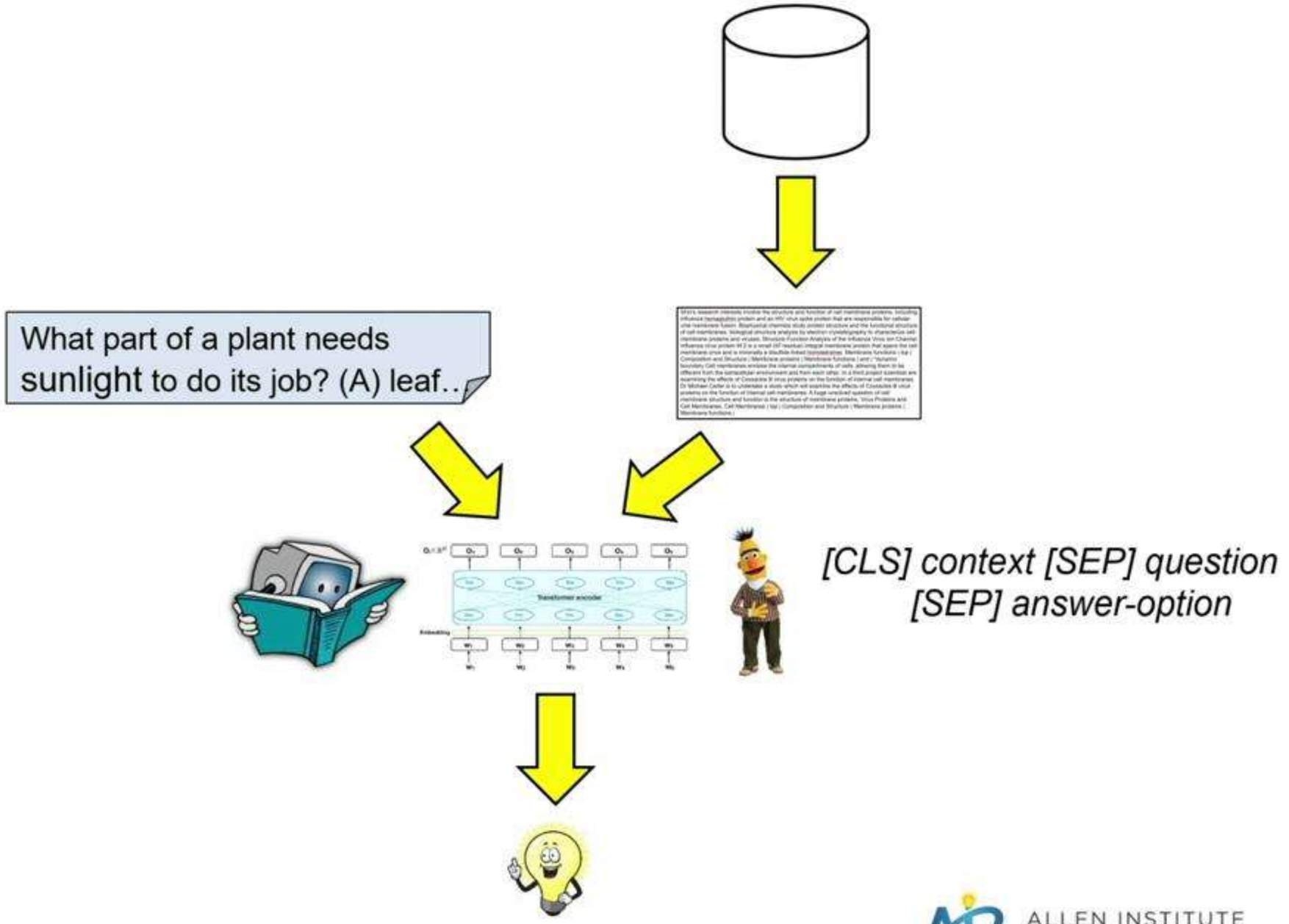
BERT and RoBERTa



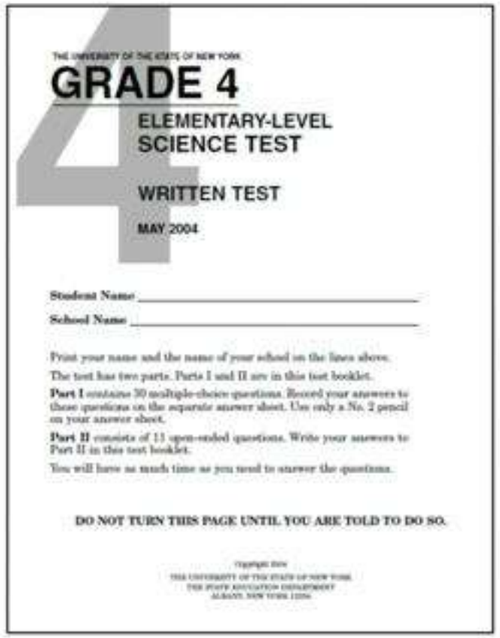
BERT and RoBERTa



BERT and RoBERTa



Similar Progress on 4th Grade NDMC



THE UNIVERSITY OF THE STATE OF NEW YORK

GRADE 4

ELEMENTARY-LEVEL SCIENCE TEST

WRITTEN TEST

MAY 2004

Student Name _____

School Name _____

Print your name and the name of your school on the lines above.

The test has two parts. Parts I and II are in this test booklet.

Part I contains 30 multiple-choice questions. Record your answers to these questions on the separate answer sheet. Use only a No. 2 pencil on your answer sheet.

Part II consists of 11 open-ended questions. Write your answers to Part II in this test booklet.

You will have as much time as you need to answer the questions.

DO NOT TURN THIS PAGE UNTIL YOU ARE TOLD TO DO SO.

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Individual Solver Performances

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

Individual Solver Performances

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

Most of the heavy lifting….

Individual Solver Performances

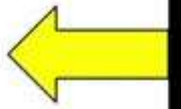
Test Set	Num Q	IR	PMI	ACME	TupInf	Multee	AristoBERT	AristoRoBERTa	ARISTO
Regents 4th	109	64.45	66.28	67.89	63.53	69.72	86.24	88.07	**89.91**
Regents 8th	119	66.60	69.12	67.65	61.41	68.91	86.55	88.24	**91.60**
Regents 12th	632	41.22	46.95	41.57	35.35	56.01	75.47	82.28	**83.54**
ARC-Easy	2376	74.48	77.76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

Most of the heavy lifting….

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

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

[redacted]

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

1. Checking for annotation artifacts

[redacted]

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

Test dataset	"Answer only" score
Regents 4th	38.53
Regents 8th	37.82
Regents 12th	47.94
ARC-Easy	36.17
ARC-Challenge	35.92
All	37.11

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	(E) joule
(B) light	(F) gradient
(C) force	(G) trench
(D) weather	(H) add heat

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

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 **[correct]**
(E) joule
(F) gradient [selected]
(G) trench
(H) add heat

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

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

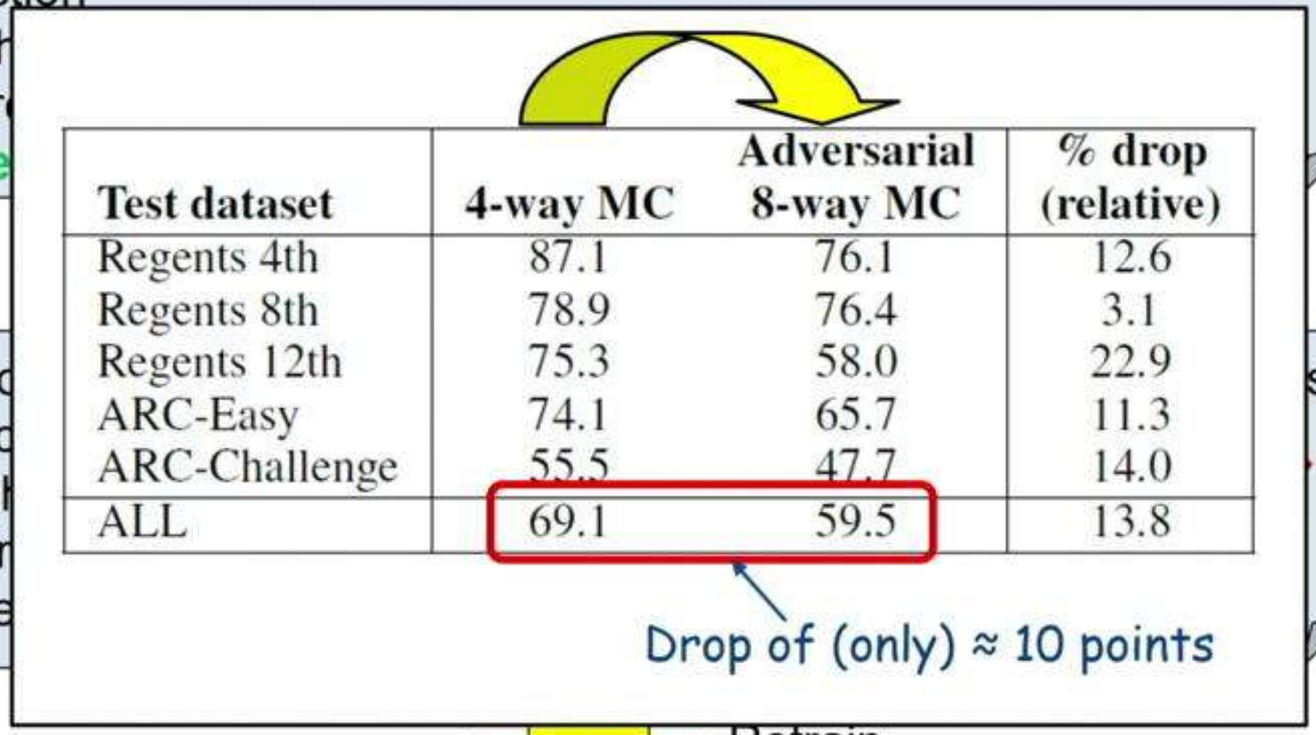
Retrain

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]	

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) ligh
(C) for
(D) we



Test dataset	4-way MC	Adversarial 8-way MC	% drop (relative)
Regents 4th	87.1	76.1	12.6
Regents 8th	78.9	76.4	3.1
Regents 12th	75.3	58.0	22.9
ARC-Easy	74.1	65.7	11.3
ARC-Challenge	55.5	47.7	14.0
ALL	69.1	59.5	13.8

Drop of (only) ≈ 10 points

Retrain

The co
(A) fric
(B) ligh
(C) for
(D) we

s

The condition of the air outdoors at a certain time of day is known as

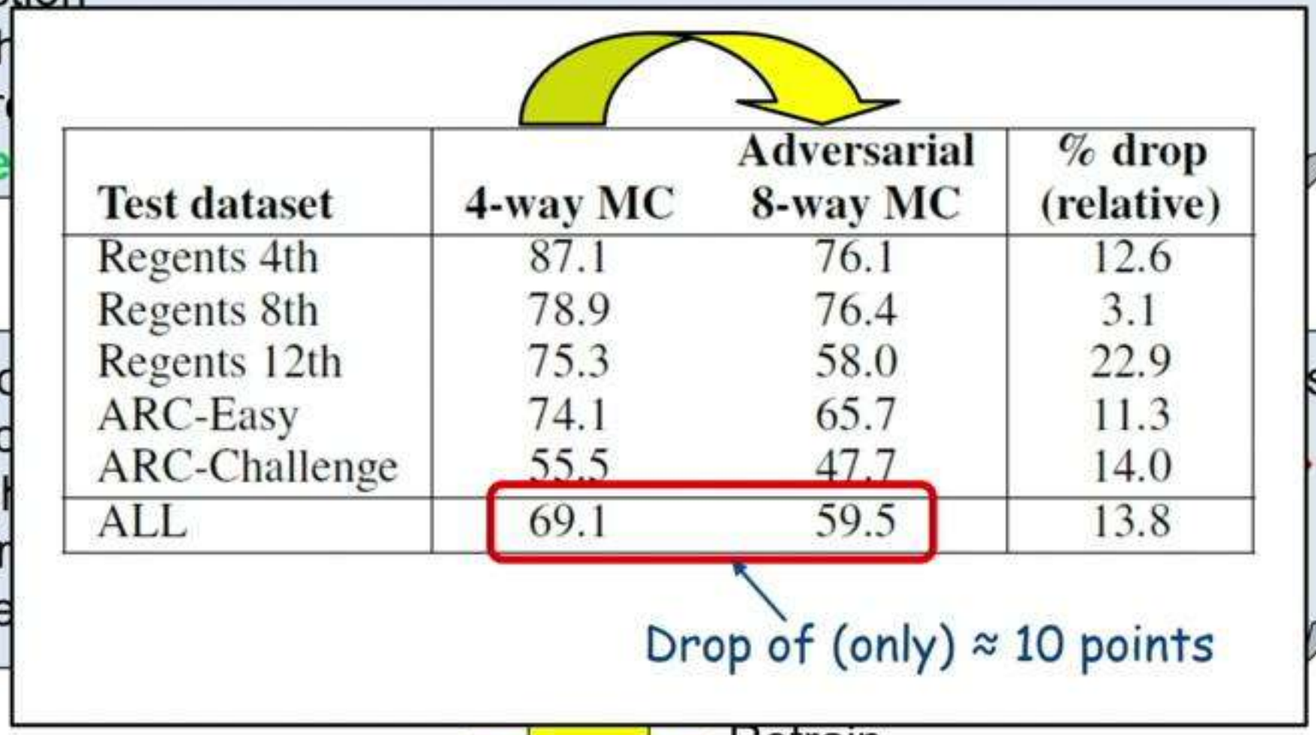
(A) friction	(E) joule
(B) light	(F) gradient [selected]
(C) force	(G) trench
(D) weather **[correct,selected]**	

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

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) ligh
(C) for
(D) we



Test dataset	4-way MC	Adversarial 8-way MC	% drop (relative)
Regents 4th	87.1	76.1	12.6
Regents 8th	78.9	76.4	3.1
Regents 12th	75.3	58.0	22.9
ARC-Easy	74.1	65.7	11.3
ARC-Challenge	55.5	47.7	14.0
ALL	69.1	59.5	13.8

Drop of (only) ≈ 10 points

Retrain

The co
(A) fric
(B) ligh
(C) for
(D) we

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]**
(E) joule
(F) gradient [selected]
(G) trench

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

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

increasing

(4) lowering the cost of bus and subway fares

3. More than Pattern Matching?

increasing
raising

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?

2019 Report Card for *Aristo*

Subject	Grade	Teacher Comments
Negation		
Conjunction		
Polarity		
World tracking		
Factivity		
Counting		

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	88.5%
		76.5%
		76%
		75%

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

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%		88.5%
2 conjuncts:	95%		76.5%
3 conjuncts:	94.5%	+ 1 negation	76%
4 conjuncts:	80%		75%

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.

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?

2019 Report Card for Aristo

Subject	Grade	Teacher Comments
Negation	A	Nice work!
Conjunction	B+	
Polarity		
World tracking		
Factivity		
Counting		

94%

80% -98%

3. More than Pattern Matching?

2019 Report Card for *Aristo*

Subject	Grade	Teacher Comments
Negation	A	Nice work!
Conjunction	B+	
Polarity		
World tracking		
Factivity		
Counting		

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

Subject	Grade	Teacher Comments
Negation	A	Nice work!
Conjunction	B+	
Polarity		
World tracking		
Factivity		
Counting		

94%

80% -98%

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

Question: If Jim turns the thermostat ~~down~~ up 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]***~~ *[correct]*

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

3. More than Pattern Matching?

2019 Report Card for *Aristo*

Subject	Grade	Teacher Comments	
Negation	*A*	*Nice work!*	*94%*
Conjunction	*B+*		*80% -98%*
Polarity	*D+*	*Could ace this with more study!*	*67.1%*
World tracking			
Factivity			
Counting			

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

Question: If Jim turns the thermostat ~~down~~ *up* 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]***~~ *[correct]*

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

3. More than Pattern Matching?

2019 Report Card for Aristo

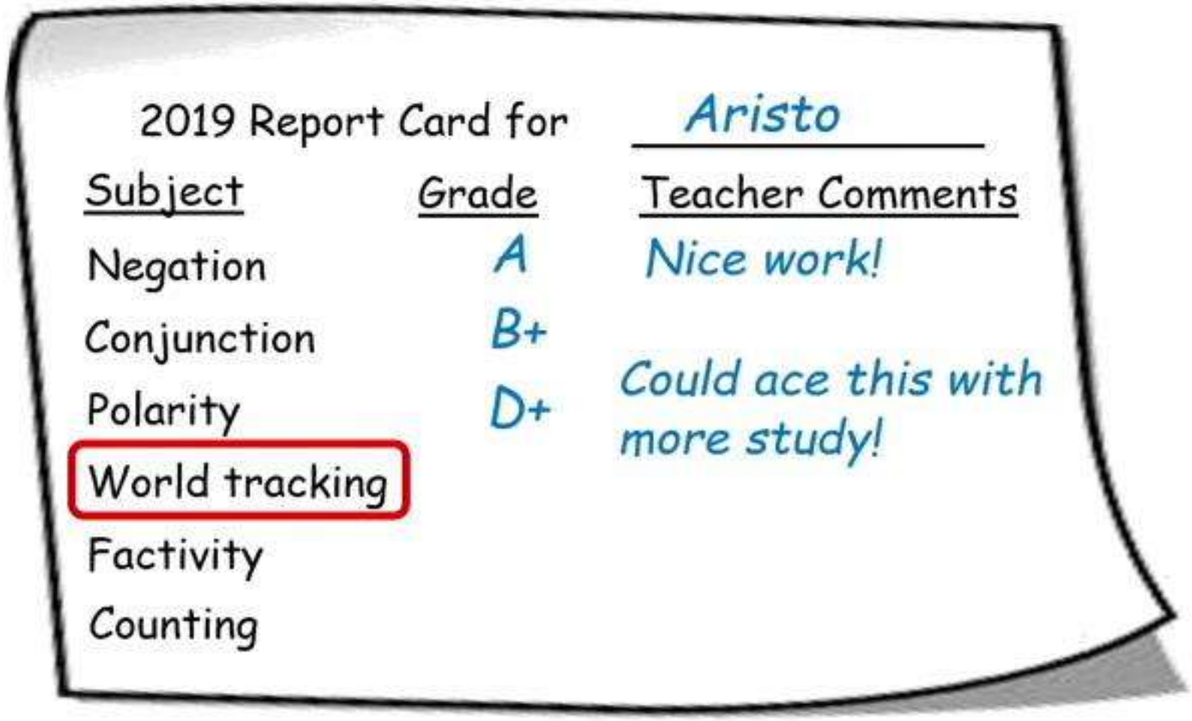
Subject	Grade	Teacher Comments
Negation	A	Nice work!
Conjunction	B+	
Polarity	D+	Could ace this with more study!
World tracking		
Factivity		
Counting		

94%

80% -98%

67.1%

3. More than Pattern Matching?



94%

80% -98%

67.1%

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 *Aristo*

Subject	Grade	Teacher Comments
Negation	*A*	*Nice work!*
Conjunction	*B+*	
Polarity	*D+*	*Could ace this with more study!*
World tracking	*C*	
Factivity		
Counting		

94%

80% -98%

67.1%

72.5%

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]***

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

3. More than Pattern Matching?

2019 Report Card for Aristo

Subject	Grade	Teacher Comments	
Negation	A	Nice work!	94%
Conjunction	B+		80% -98%
Polarity	D+	Could ace this with more study!	67.1%
World tracking	C		72.5%
Factivity	D		66.5%
Counting			

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]***

ALLEN INSTITUTE *for* ARTIFICIAL INTELLIGENCE

3. More than Pattern Matching?

2019 Report Card for *Aristo*

Subject	Grade	Teacher Comments
Negation	A	Nice work!
Conjunction	B+	
Polarity	D+	Could ace this with more study!
World tracking	C	
Factivity	D	
Counting		

94%

80% -98%

67.1%

72.5%

66.5%

3. More than Pattern Matching?

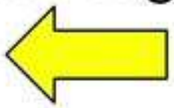
2019 Report Card for *Aristo*

Subject	Grade	Teacher Comments	
Negation	A	Nice work!	94%
Conjunction	B+		80% -98%
Polarity	D+	Could ace this with more study!	67.1%
World tracking	C		72.5%
Factivity	D		66.5%
Counting			

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

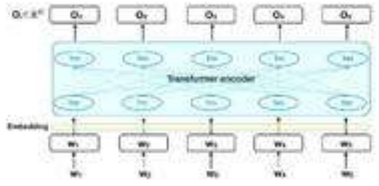
How many objects is Daniel holding? (A) zero **(B) one** (C) two (D) three

Outline

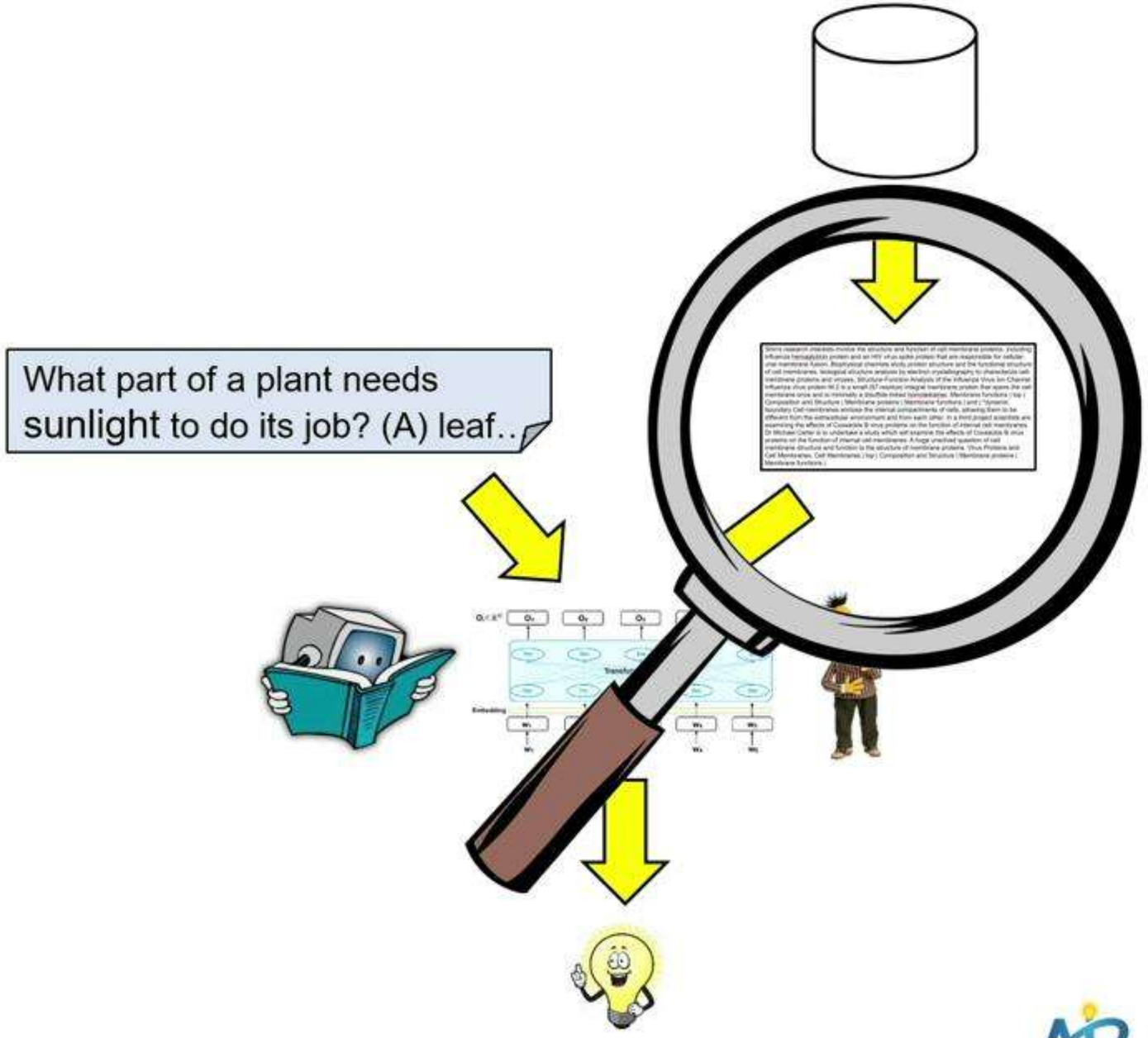
- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail? 
- What are steps forward?

4. Where is Aristo Failing?

What part of a plant needs sunlight to do its job? (A) leaf..



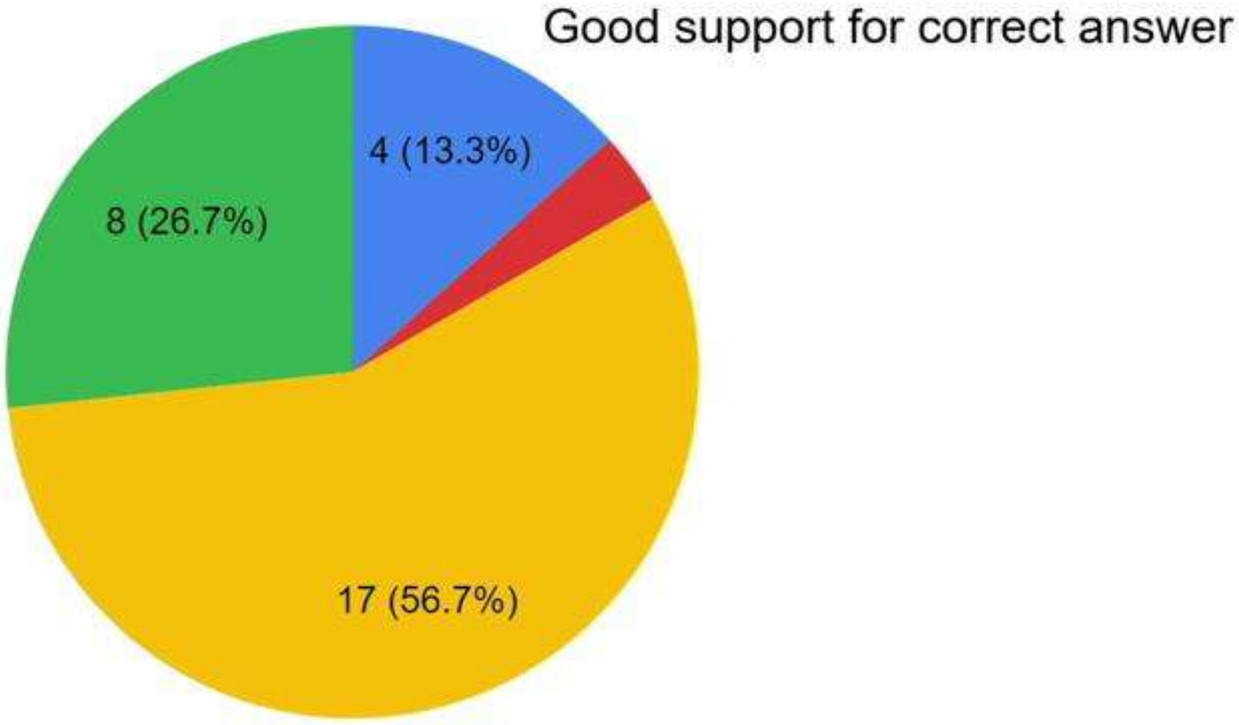
4. Where is Aristo Failing?



ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

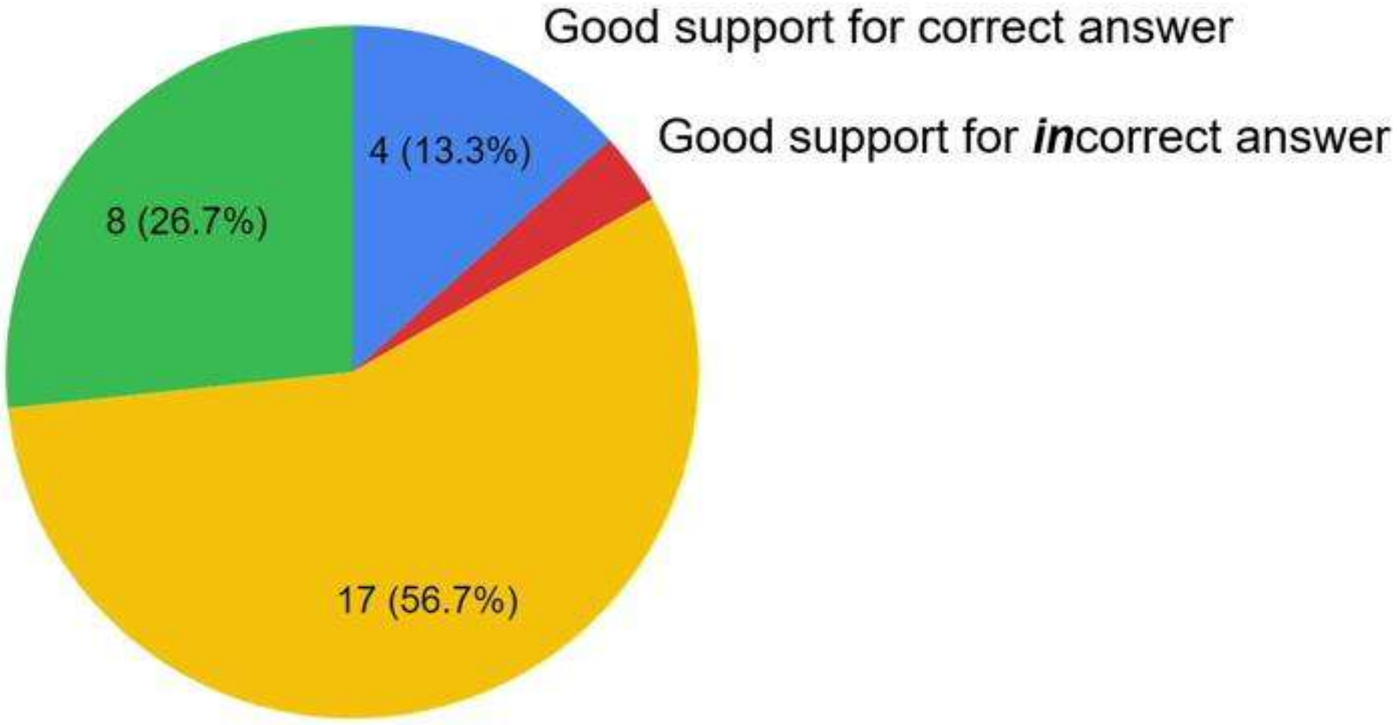
4. Where is Aristo failing?

- Case study on 30 failures:



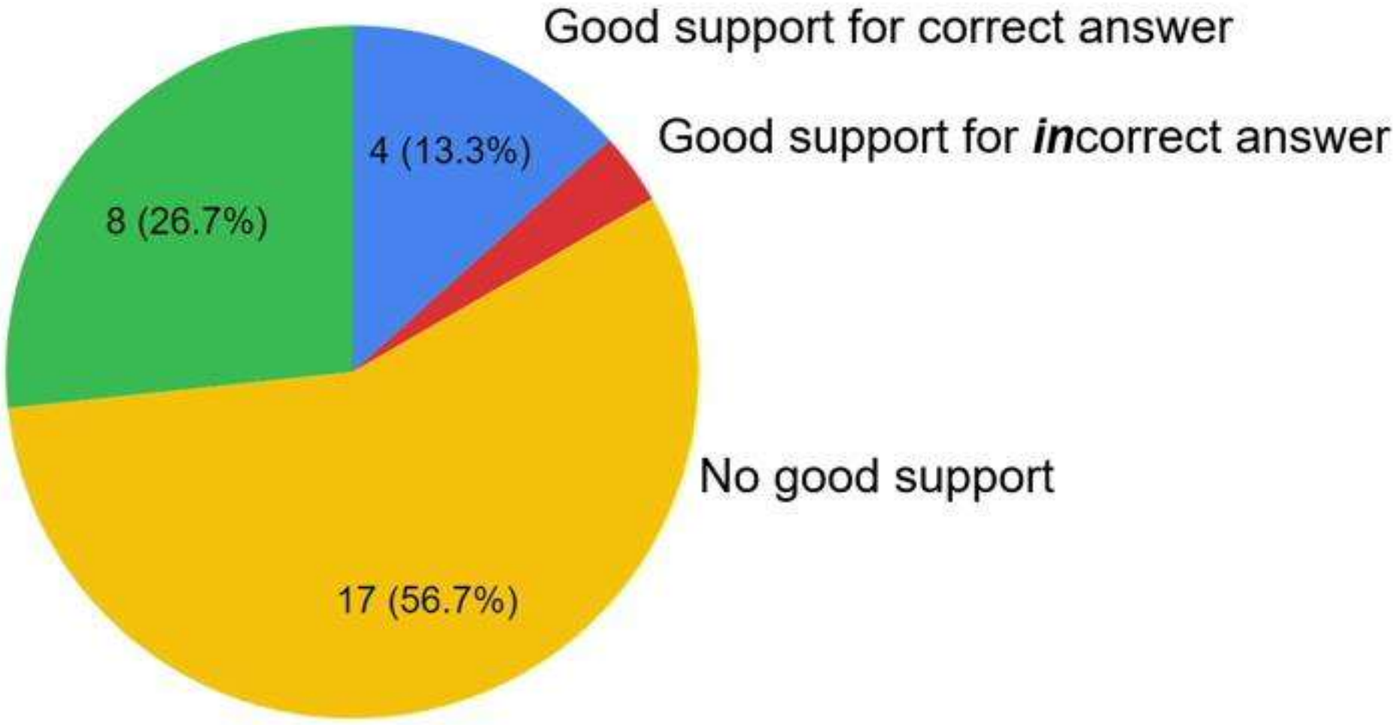
4. Where is Aristo failing?

- Case study on 30 failures:



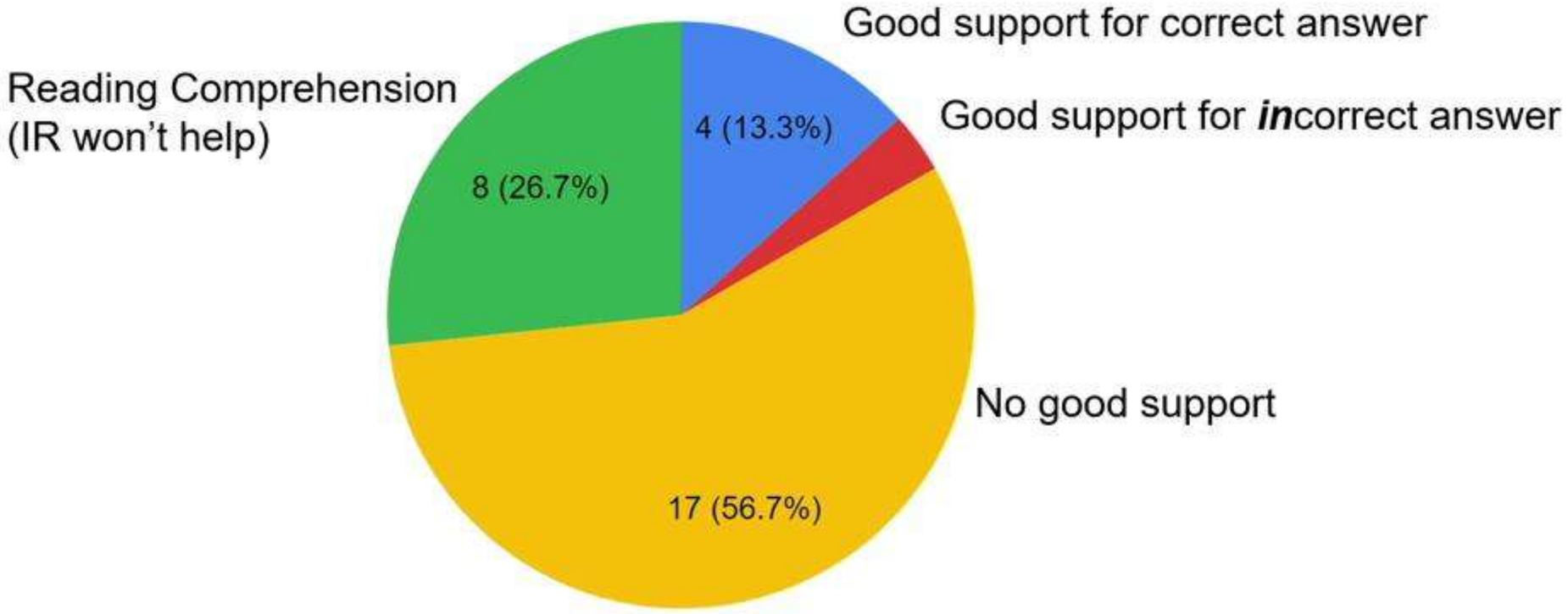
4. Where is Aristo failing?

- Case study on 30 failures:



4. Where is Aristo failing?

- Case study on 30 failures:



ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

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**

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**

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

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**

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.

2. Good support for the ***in***correct 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**

2. Good support for the ***in***correct 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.

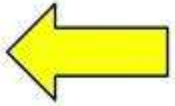
Math Reasoning

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**

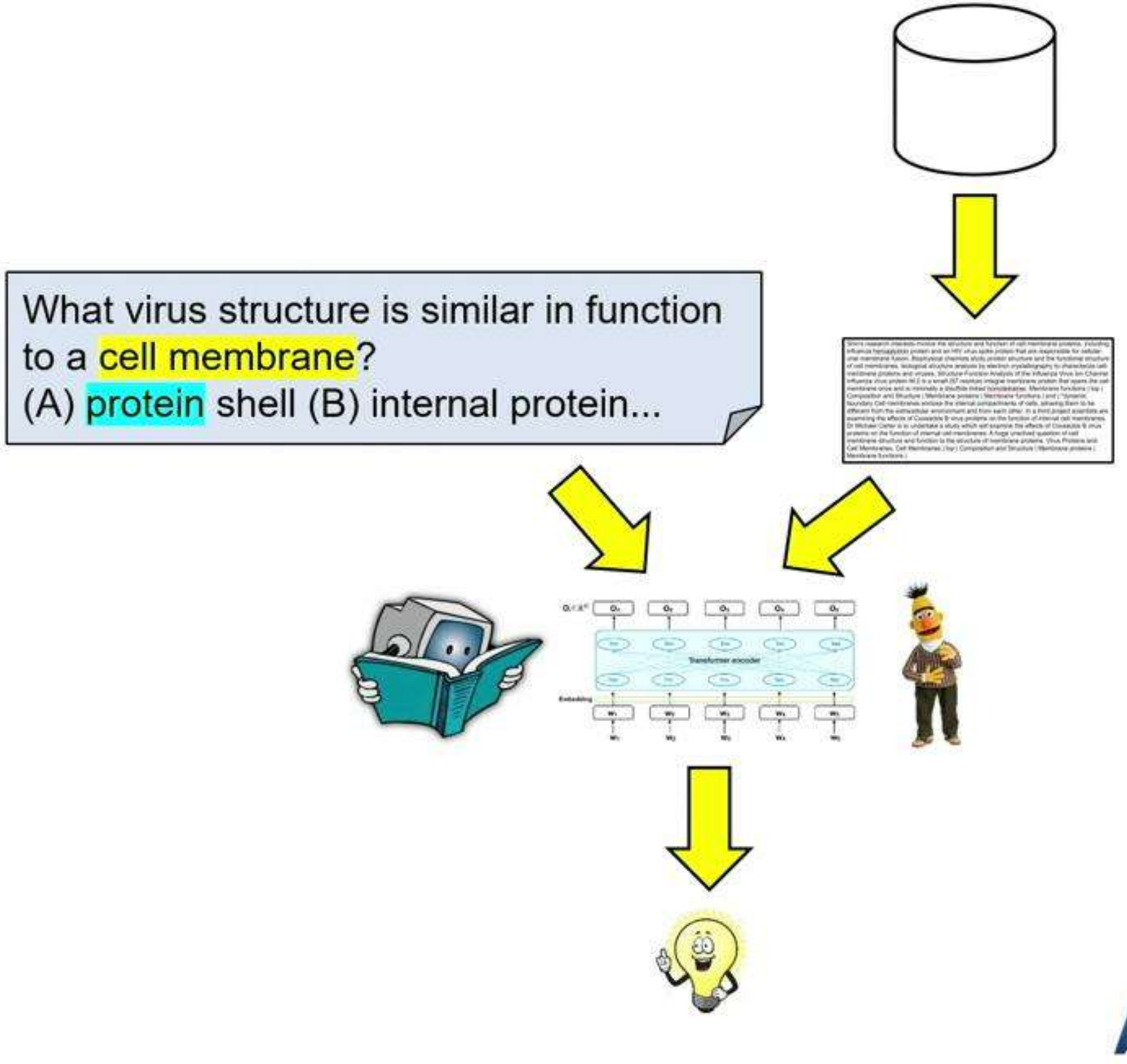
- *Because it takes the moon about **27.3 days** to complete one orbit around the Earth, the moon moves a little bit further around the Earth each day.*
- *It takes **27.3 days** for the moon to complete one revolution around the earth.*
- *The moon completes one revolution of the Earth in about **29.5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

Outline

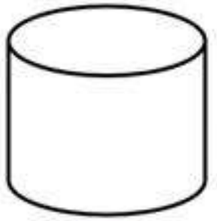
- Introduction
- How does Aristo work?
- What is going on behind the high scores on the exams?
- Where does Aristo fail?
- What are steps forward?



1. Question Decomposition



1. Question Decomposition

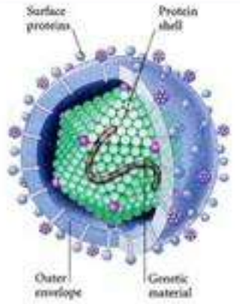


What virus structure is similar in function to a cell membrane?
(A) protein shell (B) internal protein...

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 hemaglutinin 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. Virus Proteins and Cell Membranes. Cell Membranes | top | Composition and Structure | Membrane proteins | Membrane functions |

1. Question Decomposition

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 armour 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

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

3. Modeling World States

Photosynthesis

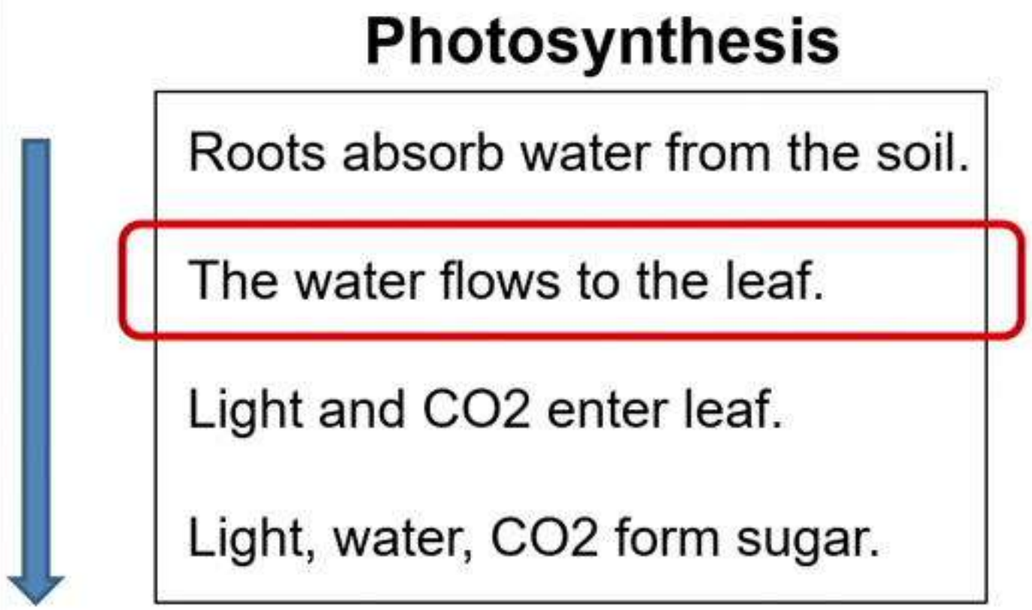
Roots absorb water from the soil.

The water flows to the leaf.

Light and CO_2 enter leaf.

Light, water, CO_2 form sugar.

3. Modeling World States



3. Modeling World States

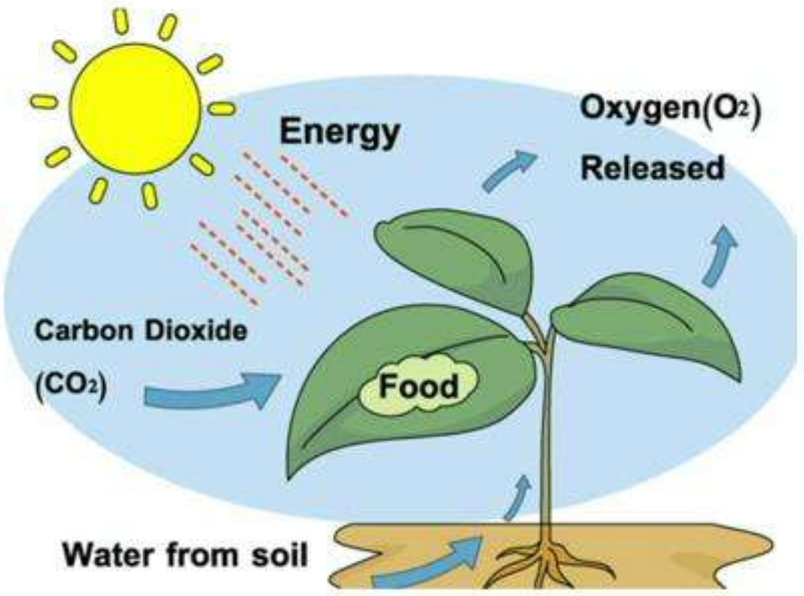
Photosynthesis

Roots absorb water from the soil.

The water flows to the leaf.

Light and CO2 enter leaf.

Light, water, CO2 form sugar.



3. Modeling World States

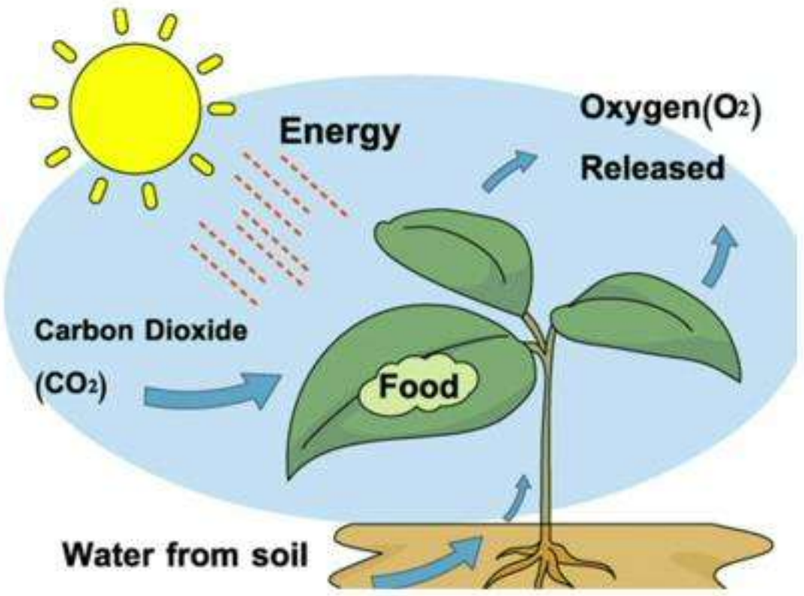
Photosynthesis

Roots absorb water from the soil.

The water flows to the leaf.

Light and CO2 enter leaf.

Light, water, CO2 form sugar.



Where is the sugar created?

3. Modeling World States

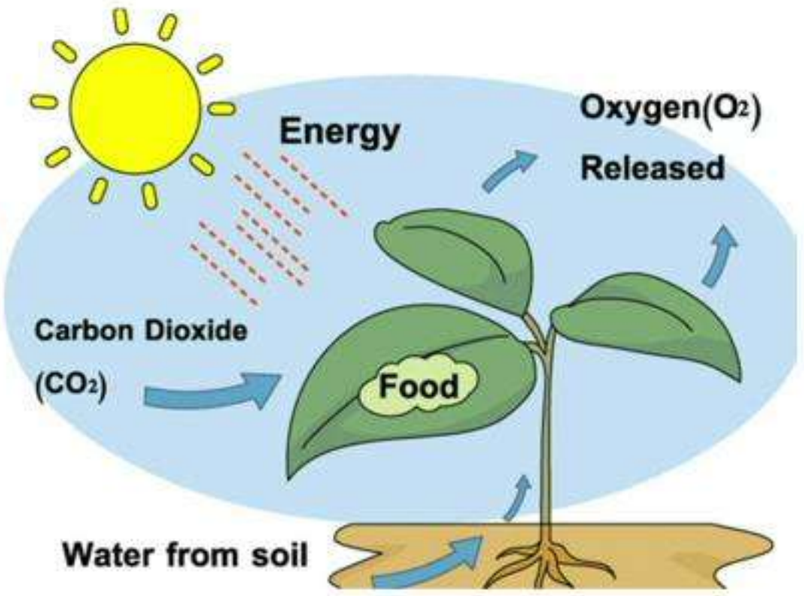
Photosynthesis

Roots absorb water from the soil.

The water flows to the leaf.

Light and CO2 enter leaf.

Light, water, CO2 form sugar.



Where is the sugar created? **Light, water, CO2** [BiDAF]

4. Explanation and Instruction

Can you pick up a penny with a magnet?

4. Explanation and Instruction

Can you pick up a penny with a magnet?

Yes

4. Explanation and Instruction

Can you pick up a penny with a magnet?

Why?

Yes

Because

- *pennies are made of metal*
- *metals are magnetic*

4. Explanation and Instruction

Can you pick up a penny with a magnet?

Why?

Yes

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?

Why?

Yes

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

- 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

- 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

Summary

- 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