COGNITIVE COMPUTATION GROUP
UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

AI2

TableILP: Semi-Structured Reasoning for Answering Science Questions

Daniel Khashabi, Dan Roth
Tushar Khot, Ashish Sabharwal, Peter Clark, Oren Etzioni
IJCAI 2016

Slides adapted from first author's presentation

Presenter: Haoyu Wang

Standardized Tests as an AI Challenge

Build AI systems that demonstrate human-like intelligence by passing standardized science exams <u>as written</u>

Which physical structure would best help a bear to **survive a winter** in New York State?
(A) big ears (B) black nose (C**) thick fur** (D) brown eyes

New Zealand shortest night

In ~~New York State~~, the ~~longest~~ period of ~~daylight~~ occurs during which month?
(A) June
(B) March
(C) December
(D) September

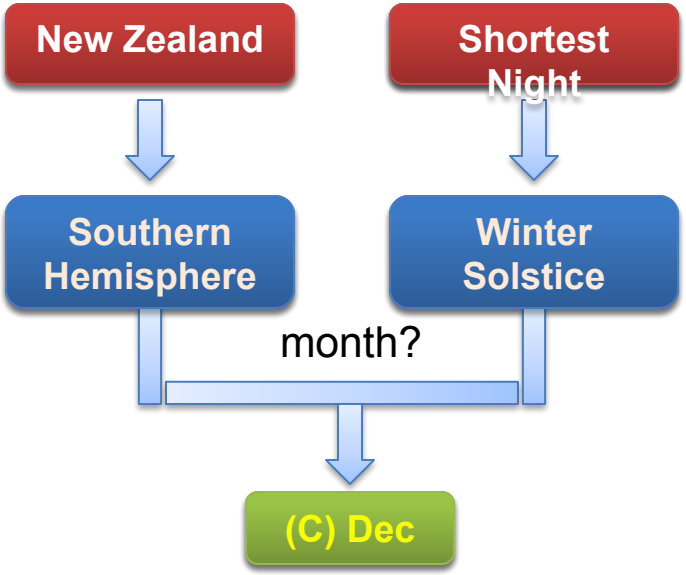
Premise: *a system that "understands" this phenomenon can correctly answer many variations!*

Semi-Structured Inference

New Zealand shortest night

In ~~New York State~~, the ~~longest~~ period of ~~daylight~~ occurs during which month?
(A) June
(B) March
(C) December
(D) September

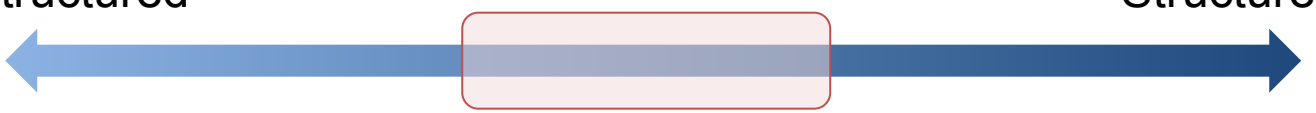
- Structured, Multi-Step Reasoning
  - science knowledge in small, manageable, swappable pieces: *regions, hemispheres, solstice*
  - Goal: overcome brittleness



✔ principled approach, explainable answers
✔ robust to variations

How can we achieve this?

Knowledge as Relational Tables



e.g., free form text from books, web

easy to acquire, difficult to reason with

Relational Tables with free form text

collections of recurring, related, science concepts

e.g., probabilistic first-order logic rules, ontologies

"easy" to reason with, difficult to acquire

Country	*Location*
France	north hemisphere
USA	north hemisphere
...	
Brazil	south hemisphere
Zambia	south hemisphere
...	

Hemisphere	*Orbital Event*	*Month*
northern	summer solstice	Jun
northern	winter solstice	Dec
northern	autumn equinox	Sep
...		
southern	summer solstice	Dec
southern	autumn equinox	Mar
...		

Energy, Forces, Adaptation, Phase Transition, Organ Function, Tools, Units, Evolution, …

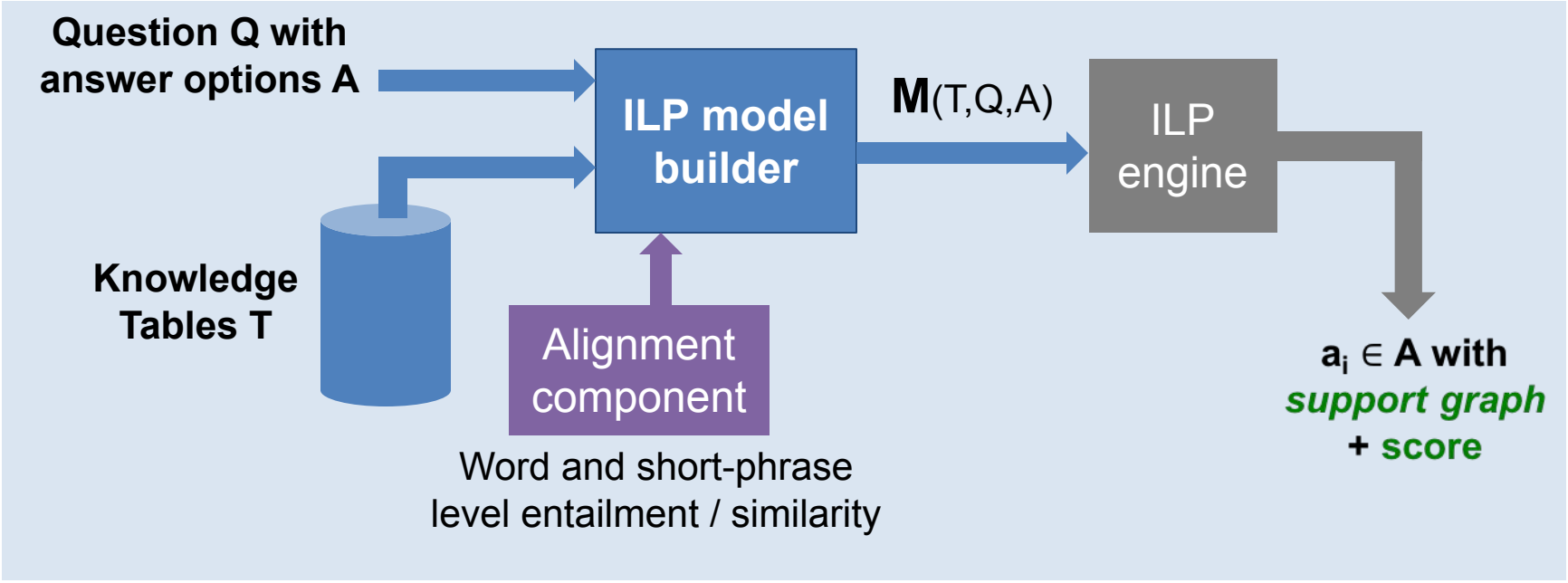
Simple structure, flexible content

- Can acquire knowledge in automated and semi-automated ways

TableILP Solver

A discrete constrained **optimization** approach to QA for multiple-choice questions

- for each given question and candidate answers, we automatically generate a corresponding ILP objective and a set of constraints.



M(T,Q,A) →

$$\max \sum_i c_i x_i \qquad \forall\, x_i \in \mathbb{N} \cup \{0\}$$

$$\begin{cases} \sum_i a_{1i} x_i \le b_1 \\ \dots \\ \sum_i a_{ki} x_i \le b_k \end{cases}$$

Optimization using Integer Linear Prog. formalism

TableILP: Main Idea

Search for the best **Support Graph** connecting the Question to an Answer through Tables.

Q: In New York State, the longest period of daylight occurs during which month?

Cities, States, Countries	**Orbital Events:** **Geographical properties & Timing**
Potential Link: **Regions and Hemispheres**	

(A) December
(B) June
(C) March
(D) September

TableILP: Main Idea

Search for the best **Support Graph** connecting the Question to an Answer through Tables.

Link this information to identify the best supported answer!

Q: In New York State, the longest period of daylight occurs during which month?

Subdivision	Country
New York State	USA
California	USA
Rio de Janeiro	Brazil
...	...

Orbital Event	Day Duration	Night Duration
Summer Solstice	Long	Short
Winter Solstice	Short	Long
....	....	...

Country	Hemisphere
United States	Northern
Canada	Northern
Brazil	Southern
.....	...

Hemisphere	Orbital Event	Month
North	Summer Solstice	June
North	Winter Solstice	December
South	Summer Solstice	December
South	Winter Solstice	June

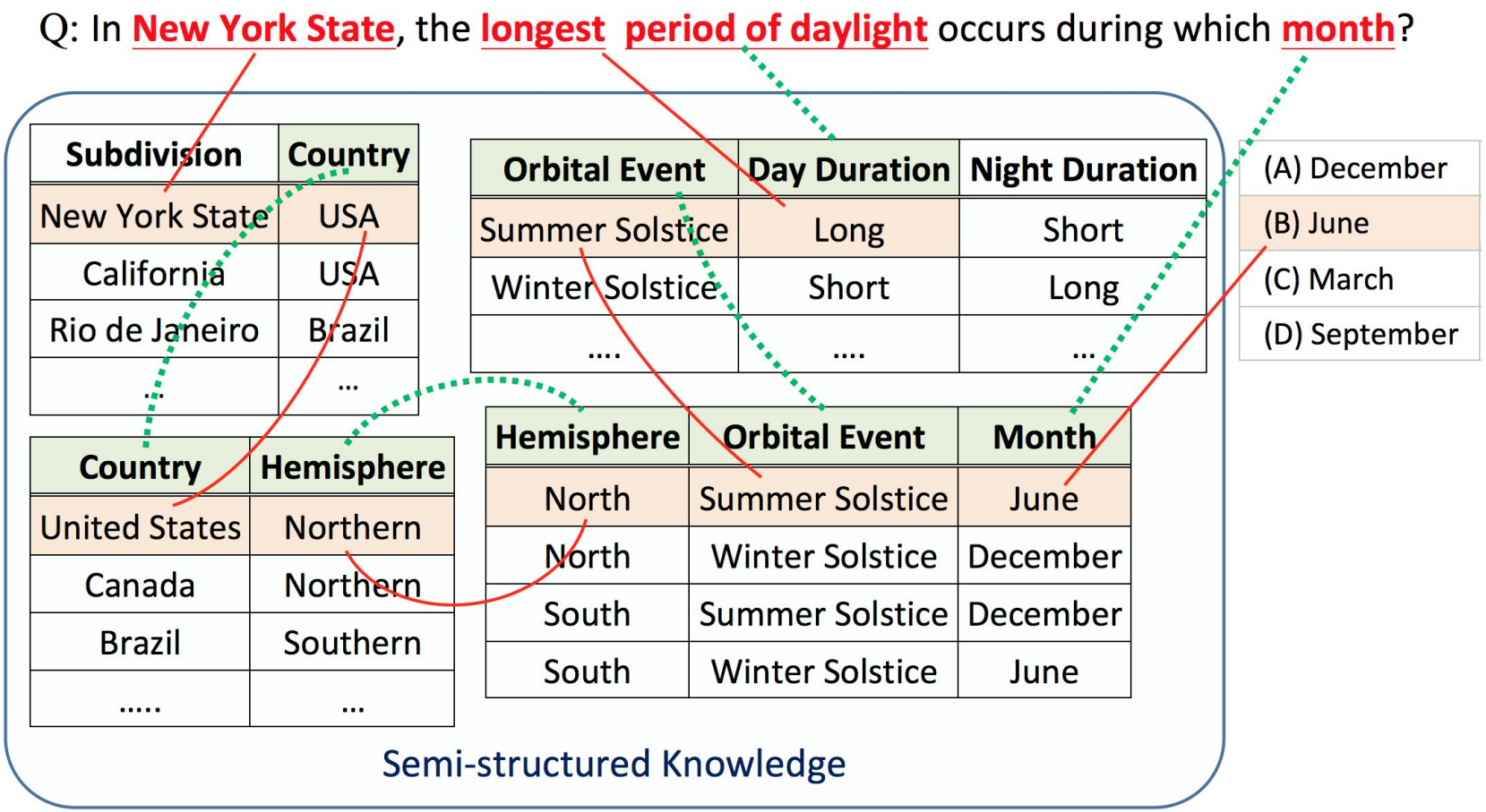
Semi-structured Knowledge

(A) December
(B) June
(C) March
(D) September

TableILP: Main Idea

Search for the best **Support Graph** connecting the Question to an Answer through Tables.

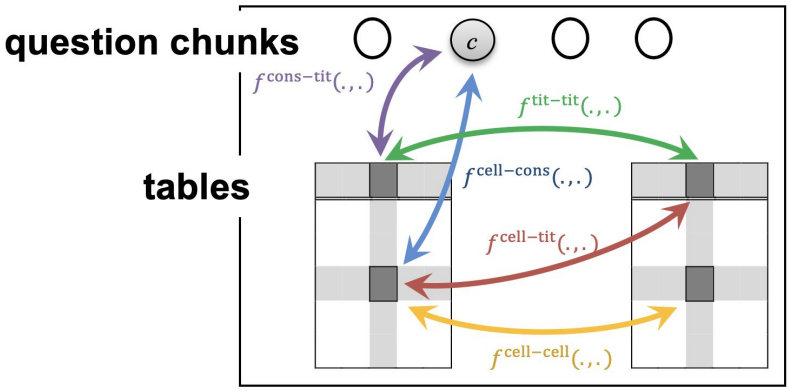
Link this information to identify the best supported answer!



ILP Model

Operates on lexical units of alignment

- cells + headers of tables T
- question chunks Q
- answer options A



~50 high level constraints + preferences

Variables define the space of "support graphs" connecting Q, A, T

- Which nodes + edges between lexical units are active?

Objective Function: "better" support graphs = higher objective value

- Reward active units, high lexical match links, column header match, …
- WH-term boost (which form of energy), science-term boost (evaporation)
- Penalize spurious overuse of frequently occurring terms

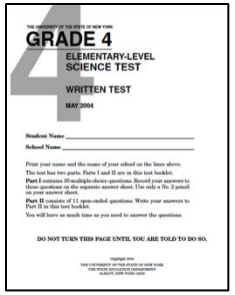
ILP Model: Constraints

Dual goal: scalability, consider only meaningful support graphs

- **Structural Constraints**
  - Meaningful proof structures
    - connectedness, question coverage, appropriate table use
    - parallel evidence => identical multi-row activity signature
  - Simplicity appropriate for 4^{th} / 8^{th} grade
- **Semantic Constraints**
  - Chaining => table joins between semantically similar column pairs
  - Relation matching (ruler measures length, change from water to liquid)
- **Table Relevance Ranking**
  - TF-IDF scoring to identify top N relevant tables

Evaluation

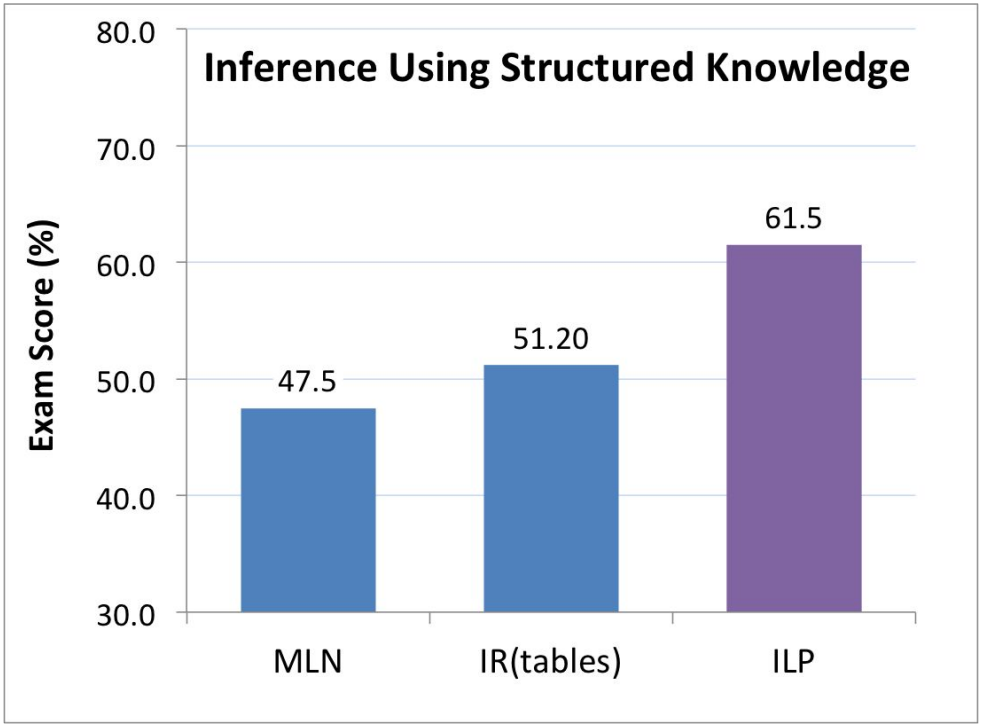
- **4th Grade NY Regents Science Exam**
  - Focus on non-diagram multiple-choice (4-way)
  - 129 questions in completely unseen Test set
    - 6 years of exams; 95% C.I. = 9%
  - Score: 1 point per question (1/k for k-way tie including correct answer)


GRADE 4
ELEMENTARY-LEVEL SCIENCE TEST
WRITTEN TEST
MAY 2004

- **Baselines:**
  - IR Solver: Information Retrieval using Lucene search
    - Using 280 GB of plain text (50B tokens) "waterloo" corpus [AAAI, 2015]
    - IR Solver(tables): Using same tables as TableILP
  - PMI Solver: Statistical correlation using pointwise mutual info.
    - Using 280 GB of plain text (50B tokens) "waterloo" corpus [AAAI, 2015]
  - MLN: Markov Logic Network, a structured prediction model
    - Using rules from 80K sentences [EMNLP, 2015]

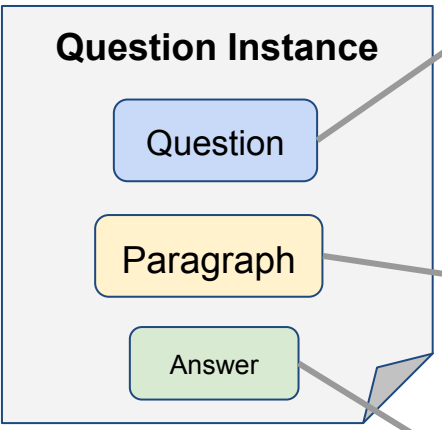
Results: Same Knowledge

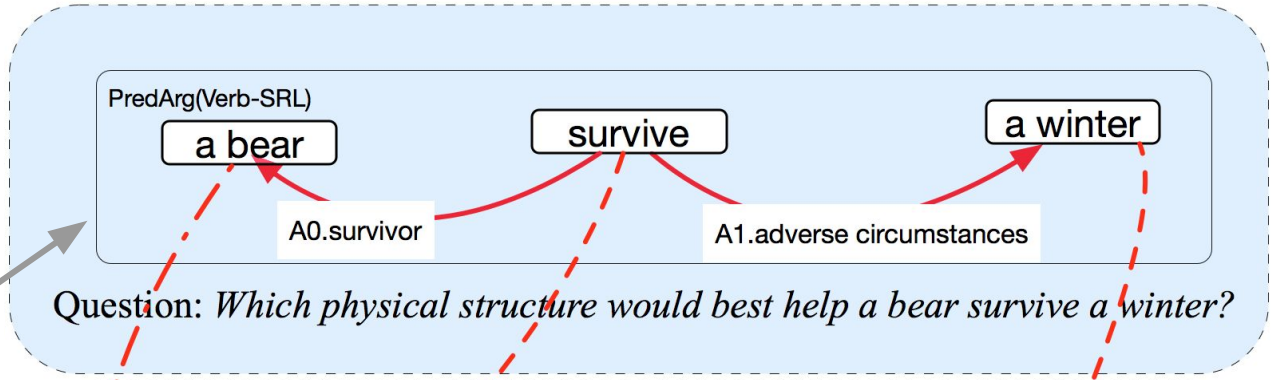
TableILP is substantially better than IR & MLN, when given knowledge derived from the same, domain-targeted sources

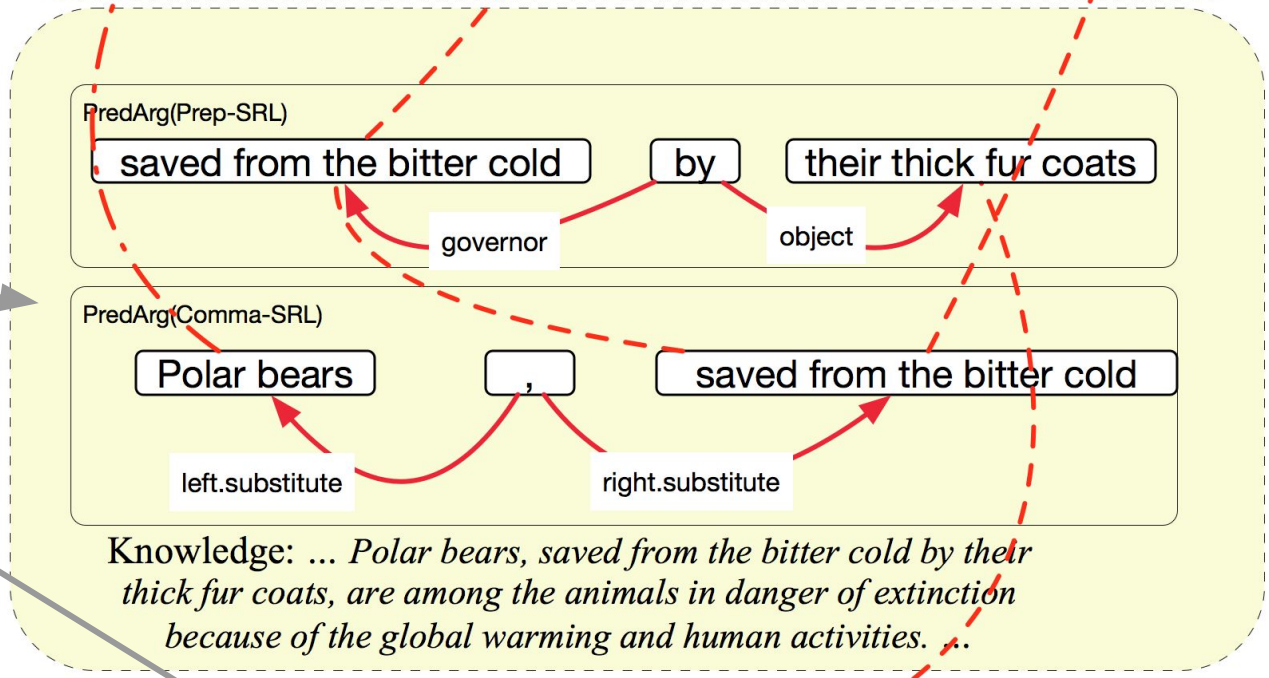


QA as Global Reasoning over Semantic Abstractions

Khashabi et al
(AAAI 2018)







(Irrelevant edges and graphs are dropped for simplicity)

Assessing Brittleness: Question Perturbation

How robust are approaches to simple question perturbations *that would typically make the question easier for a human*?

- E.g., Replace incorrect answers with arbitrary co-occurring terms

In New York State, the longest period of daylight occurs during which month?
(A) *eastern* (B) June (C) *history* (D) *years*

Solver	Original Score (%)	% Drop with Perturbation absolute	% Drop with Perturbation relative
IR	70.7	13.8	19.5
PMI	73.6	24.4	33.2
TableILP	85.0	**10.5**	**12.3**

New Zealand shortest night

In ~~New York State~~, the ~~longest~~ period of ~~daylight~~ occurs during which month?
(A) June
(B) March
(C) December
(D) September

Adversarial Distracting for Evaluation

Jia and Liang, EMNLP'17

Article: Super Bowl 50
Paragraph: "*Peyton Manning became the first quarterback ever to lead two different teams to multiple Super Bowls. He is also the oldest quarterback ever to play in a Super Bowl at age 39. The past record was held by John Elway, who led the Broncos to victory in Super Bowl XXXIII at age 38 and is currently Denver's Executive Vice President of Football Operations and General Manager. Quarterback Jeff Dean had jersey number 37 in Champ Bowl XXXIV.*"
Question: "*What is the name of the quarterback who was 38 in Super Bowl XXXIII?*"
Original Prediction: John Elway
Prediction under adversary: Jeff Dean

Perturbation for recent BERT-based systems

Krunal's demo at http://dickens.seas.upenn.edu:4004

Vanilla Setting:

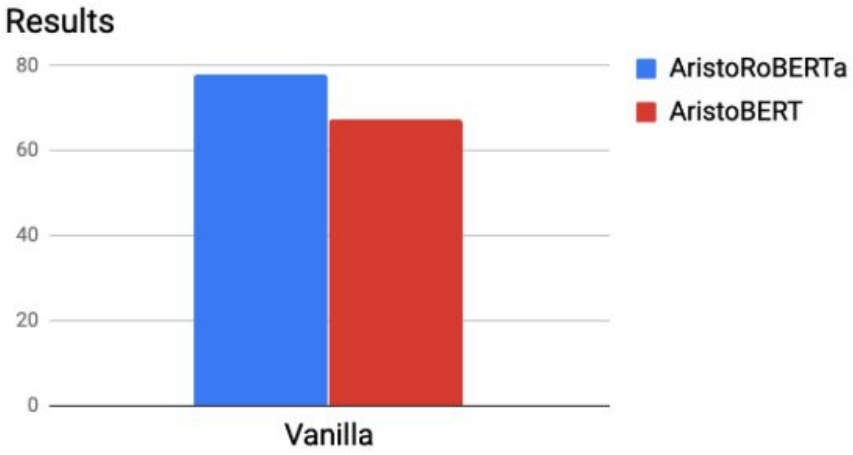
Question: The rate at which a wave passes through a medium is known as its

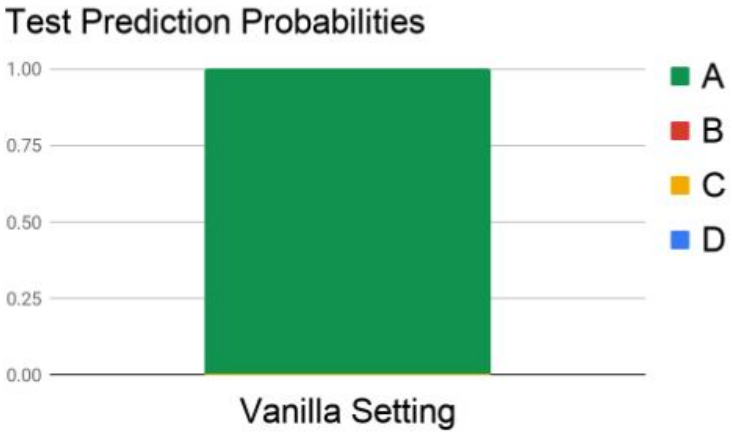
a) **speed.** — **Context:** The speed of a wave is the rate at which vibrations propagate through the medium. VELOCITY: The speed at which the wave moves ...

b) **amplitude.** — **Context:** The greater the amplitude of vibrations of the particles of the medium, the greater the rate at which energy is transported through it…

c) **acceleration.** — **Context:** … hydrophones measure the acceleration of the medium as a seismic wave passes through it, unlike geophones, which respond to the…

d) **wavelength.** — **Context:** The speed of an electromagnetic wave in a medium depends on its wavelength. As the wavelength of a wave in a uniform medium ..





Perturbation for recent BERT-based systems

Krunal's demo at http://dickens.seas.upenn.edu:4004

No Context Setting:

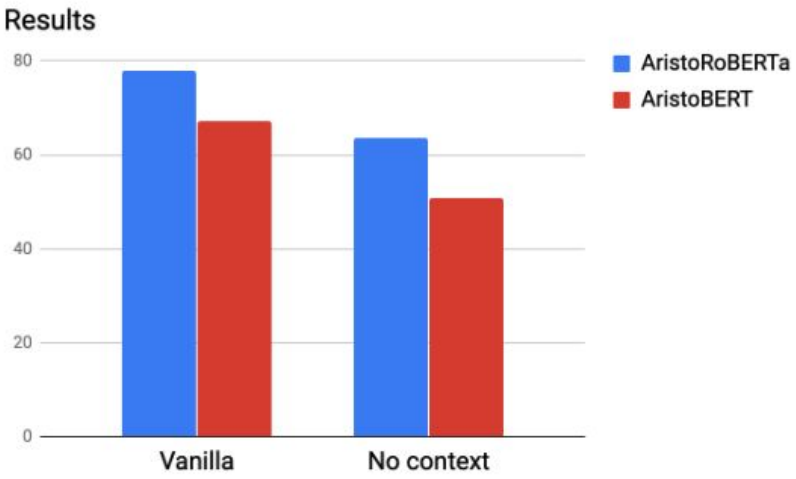
Question: The rate at which a wave passes through a medium is known as its

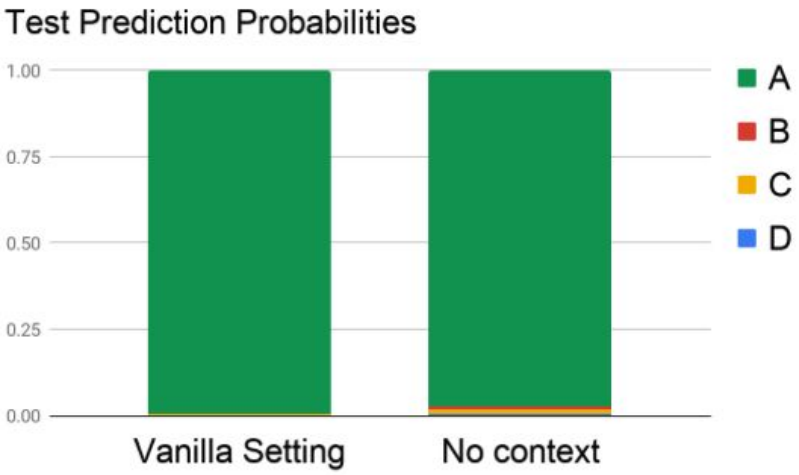
a) speed. **Context:**

b) amplitude. **Context:**

c) acceleration. **Context:**

d) wavelength. **Context:**





ILLINOIS UNIVERSITY OF ILLINOIS AT URBANA-CHAMPAIGN

ALLEN INSTITUTE for ARTIFICIAL INTELLIGENCE

Perturbation for recent BERT-based systems

Krunal's demo at http://dickens.seas.upenn.edu:4004

Incorrect option perturbation:

Question: The rate at which a wave passes through a medium is known as its

a) **speed.** — **Context:** The speed of a wave is the rate at which vibrations propagate through the medium. VELOCITY: The speed at which the wave moves ...

b) `{Question}` — **Context:** `{Question}` The rate at which a wave passes through a medium is known as its

c) **acceleration.** — **Context:** … hydrophones measure the acceleration of the medium as a seismic wave passes through it, unlike geophones, which respond to the…

d) **wavelength.** — **Context:** The speed of an electromagnetic wave in a medium depends on its wavelength. As the wavelength of a wave in a uniform medium ..