Natural Language Processing

Benchmarks/Tasks/Metrics

Neville Walo
natural language processing
natural resources
natural disasters
natural selection
natural history museum
natural language processing definition
Natural Born Killers
Film (1994)
natural mojo
natural numbers
natural gas

Google Suche  Auf gut Glück!
Natural Language Processing

Verarbeitung natürlicher Sprache
What is Natural Language Processing?

- **Big picture**: A set of methods and algorithms for making natural languages accessible to computers

- **Analysis** \((\text{NL} \rightarrow \mathbb{R})\), e.g., topic classification

- **Generation** \((\mathbb{R} \rightarrow \text{NL})\), e.g., chat bots

- **Acquisition** of \(\mathbb{R}\) from knowledge and data, e.g., modeling

- \(\mathbb{R}\): some representation interpretable by a machine
Benchmarks

- GLUE (General Language Understanding Evaluation)
- SuperGLUE (Successor of GLUE)
## SuperGLUE Tasks

<table>
<thead>
<tr>
<th>Name</th>
<th>Identifier</th>
<th>Download</th>
<th>More Info</th>
<th>Metric</th>
</tr>
</thead>
<tbody>
<tr>
<td>Broadcoverage Diagnostics</td>
<td>AX-b</td>
<td></td>
<td></td>
<td>Matthew's Corr</td>
</tr>
<tr>
<td>CommitmentBank</td>
<td>CB</td>
<td></td>
<td></td>
<td>Avg. F1 / Accuracy</td>
</tr>
<tr>
<td>Choice of Plausible Alternatives</td>
<td>COPA</td>
<td></td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>Multi-Sentence Reading Comprehension</td>
<td>MultiRC</td>
<td></td>
<td></td>
<td>F1a / EM</td>
</tr>
<tr>
<td>Recognizing Textual Entailment</td>
<td>RTE</td>
<td></td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>Words in Context</td>
<td>WiC</td>
<td></td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>The Winograd Schema Challenge</td>
<td>WSC</td>
<td></td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>BoolQ</td>
<td>BoolQ</td>
<td></td>
<td></td>
<td>Accuracy</td>
</tr>
<tr>
<td>Reading Comprehension with Commonsense Reasoning</td>
<td>ReCoRD</td>
<td></td>
<td></td>
<td>F1 / Accuracy</td>
</tr>
<tr>
<td>Winogender Schema Diagnostics</td>
<td>AX-g</td>
<td></td>
<td></td>
<td>Gender Parity / Accuracy</td>
</tr>
</tbody>
</table>

[DOWNLOAD ALL DATA]
<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Model</th>
<th>URL</th>
<th>Score</th>
<th>BoolQ</th>
<th>CB</th>
<th>COPA</th>
<th>MultiRC</th>
<th>ReCoRD</th>
<th>RTE</th>
<th>WiC</th>
<th>WSC</th>
<th>AX-g</th>
<th>AX-b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DeBERTa Team - Microsoft</td>
<td>DeBERTa / TuringNLv4</td>
<td></td>
<td>90.3</td>
<td>90.4</td>
<td>95.7/97.6</td>
<td>98.4</td>
<td>88.2/63.7</td>
<td>94.5/94.1</td>
<td>93.2</td>
<td>77.5</td>
<td>95.9</td>
<td>93.3/93.8</td>
<td>66.7</td>
</tr>
<tr>
<td>2</td>
<td>Zinl Wang</td>
<td>T5 + Meena, Single Model (Meena Team - Google Brain)</td>
<td></td>
<td>90.2</td>
<td>91.3</td>
<td>95.8/97.6</td>
<td>97.4</td>
<td>88.3/63.0</td>
<td>94.2/93.5</td>
<td>92.7</td>
<td>77.9</td>
<td>95.9</td>
<td>88.8/89.9</td>
<td>66.5</td>
</tr>
<tr>
<td>3</td>
<td>SuperGLUE Human Baselines</td>
<td>SuperGLUE Human Baselines</td>
<td></td>
<td>89.8</td>
<td>89.0</td>
<td>95.8/98.9</td>
<td>100.0</td>
<td>81.8/51.9</td>
<td>91.7/91.3</td>
<td>93.6</td>
<td>80.0</td>
<td>100.0</td>
<td>99.3/99.7</td>
<td>76.6</td>
</tr>
<tr>
<td>4</td>
<td>T5 Team - Google</td>
<td>T5</td>
<td></td>
<td>89.3</td>
<td>91.2</td>
<td>93.9/96.8</td>
<td>94.8</td>
<td>88.1/63.3</td>
<td>94.1/93.4</td>
<td>92.5</td>
<td>76.9</td>
<td>93.8</td>
<td>92.7/91.9</td>
<td>65.6</td>
</tr>
<tr>
<td>5</td>
<td>Huawei Noah's Ark Lab</td>
<td>NEZHA-Plus</td>
<td></td>
<td>86.7</td>
<td>87.8</td>
<td>94.4/96.0</td>
<td>93.6</td>
<td>84.6/55.1</td>
<td>90.1/89.6</td>
<td>89.1</td>
<td>74.6</td>
<td>93.2</td>
<td>87.1/74.4</td>
<td>58.0</td>
</tr>
</tbody>
</table>
Choice of Plausible Alternatives (COPA)

• Premise: The man broke his toe. What was the CAUSE of this?
  Alternative 1: He got a hole in his sock.
  Alternative 2: He dropped a hammer on his foot.

• Premise: I tipped the bottle. What happened as a RESULT?
  Alternative 1: The liquid in the bottle froze.
  Alternative 2: The liquid in the bottle poured out.

• Premise: I knocked on my neighbor's door. What happened as a RESULT?
  Alternative 1: My neighbor invited me in.
  Alternative 2: My neighbor left his house.
Choice of Plausible Alternatives (COPA)

- Metric: Accuracy

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Model</th>
<th>COPA</th>
<th>MultiRC</th>
<th>ReCoRD</th>
<th>RTE</th>
<th>WIC</th>
<th>WSC</th>
<th>AX-g</th>
<th>AX-b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DeBERTa Team - Microsoft</td>
<td>DeBERTa / TuringNLRv4</td>
<td>98.4</td>
<td>88.2/63.7</td>
<td>94.5/94.1</td>
<td>93.2</td>
<td>77.5</td>
<td>95.9</td>
<td>93.3/93.8</td>
<td>66.7</td>
</tr>
<tr>
<td>2</td>
<td>Zini Wang</td>
<td>T5 + Meena, Single Model (Meena Team - Google Brain)</td>
<td>97.4</td>
<td>88.3/63.0</td>
<td>94.2/93.5</td>
<td>92.7</td>
<td>77.9</td>
<td>95.9</td>
<td>88.8/89.9</td>
<td>66.5</td>
</tr>
<tr>
<td>3</td>
<td>SuperGLUE Human Baselines</td>
<td>SuperGLUE Human Baselines</td>
<td>100.0</td>
<td>81.8/51.9</td>
<td>91.7/91.3</td>
<td>93.6</td>
<td>80.0</td>
<td>100.0</td>
<td>99.3/99.7</td>
<td>76.6</td>
</tr>
<tr>
<td>4</td>
<td>T5 Team - Google</td>
<td>T5</td>
<td>94.8</td>
<td>88.1/63.3</td>
<td>94.1/93.4</td>
<td>92.5</td>
<td>76.9</td>
<td>93.8</td>
<td>92.7/91.9</td>
<td>65.6</td>
</tr>
<tr>
<td>5</td>
<td>Huawei Noah's Ark Lab</td>
<td>NEZHA-Plus</td>
<td>93.6</td>
<td>84.6/55.1</td>
<td>90.1/89.6</td>
<td>89.1</td>
<td>74.6</td>
<td>93.2</td>
<td>87.1/74.4</td>
<td>58.0</td>
</tr>
</tbody>
</table>
Paragraph:
Sent 1: Most young mammals, including humans, like to play.
Sent 2: Play is one way they learn the skills that they will need as adults.
Sent 3: Think about how kittens play.
Sent 4: They pounce on toys and chase each other.
Sent 5: This helps them learn how to be better predators.
Sent 6: Big cats also play.
Sent 7: The lion cubs pictured below are playing.
Sent 8: At the same time, they are also practicing their hunting skills.
Sent 9: The dogs are playing tug-of-war with a toy.
Sent 10: What do you think they are learning by playing together this way?
Sent 11: Human children learn by playing as well.
Sent 12: For example, playing games and sports can help them learn to follow rules.
Sent 13: They also learn to work together.
Sent 14: The young child pictured below is playing in the sand.
Sent 15: She is learning about the world through play.
Sent 16: What do you think she might be learning?

Question: What do human children learn by playing games and sports?

- to follow rules
- They learn to follow rules and work together.
- They learn about the world
- Learn to work together
- skills that they will need as adults
- they learn about how to cheat
- how to hunt
- tug-of-war
- only learns to follow rules
- only learns working together
- hunting skills
Reading Comprehension over Multiple Sentences (MultiRC)

- Metric: F1a / EM (Exact Match)
# Words in Context (WiC)

<table>
<thead>
<tr>
<th>Label</th>
<th>Target</th>
<th>Context-1</th>
<th>Context-2</th>
</tr>
</thead>
<tbody>
<tr>
<td>F</td>
<td>bed</td>
<td>There's a lot of trash on the bed of the river</td>
<td>I keep a glass of water next to my bed when I sleep</td>
</tr>
<tr>
<td>F</td>
<td>land</td>
<td>The pilot managed to land the airplane safely</td>
<td>The enemy landed several of our aircrafts</td>
</tr>
<tr>
<td>F</td>
<td>justify</td>
<td>Justify the margins</td>
<td>The end justifies the means</td>
</tr>
<tr>
<td>T</td>
<td>beat</td>
<td>We beat the competition</td>
<td>Agassi beat Becker in the tennis championship</td>
</tr>
<tr>
<td>T</td>
<td>air</td>
<td>Air pollution</td>
<td>Open a window and let in some air</td>
</tr>
<tr>
<td>T</td>
<td>window</td>
<td>The expanded window will give us time to catch the thieves</td>
<td>You have a two-hour window of clear weather to finish working on the lawn</td>
</tr>
</tbody>
</table>
Words in Context (WiC)

• Metric: Accuracy

<table>
<thead>
<tr>
<th>Rank</th>
<th>Name</th>
<th>Model</th>
<th>URL</th>
<th>Score</th>
<th>BoolQ</th>
<th>CB</th>
<th>COPA</th>
<th>MultiRC</th>
<th>ReCoRD</th>
<th>RTE</th>
<th>WIC</th>
<th>WSC</th>
<th>AX-g</th>
<th>AX-b</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DeBERTa Team - Microsoft</td>
<td>DeBERTa / TuringNLUv4</td>
<td>☑</td>
<td>90.3</td>
<td>90.4</td>
<td>95.7/97.6</td>
<td>98.4</td>
<td>88.2/63.7</td>
<td>94.5/94.1</td>
<td>93.2</td>
<td>77.5</td>
<td>95.9</td>
<td>93.3/93.8</td>
<td>66.7</td>
</tr>
<tr>
<td>2</td>
<td>Zhir Wang</td>
<td>T5 + Meena, Single Model (Meena Team - Google Brain)</td>
<td>☑</td>
<td>90.2</td>
<td>91.3</td>
<td>95.8/97.6</td>
<td>97.4</td>
<td>88.3/63.0</td>
<td>94.2/93.5</td>
<td>92.7</td>
<td>77.9</td>
<td>95.9</td>
<td>88.8/89.9</td>
<td>66.5</td>
</tr>
<tr>
<td>3</td>
<td>SuperGLUE Human Baselines</td>
<td>SuperGLUE Human Baselines</td>
<td>☑</td>
<td>89.8</td>
<td>89.0</td>
<td>95.8/98.9</td>
<td>100.0</td>
<td>81.8/51.9</td>
<td>91.7/91.3</td>
<td>93.6</td>
<td>80.0</td>
<td>100.0</td>
<td>99.3/99.7</td>
<td>76.6</td>
</tr>
<tr>
<td>4</td>
<td>T5 Team - Google</td>
<td>T5</td>
<td>☑</td>
<td>89.3</td>
<td>91.2</td>
<td>93.9/96.8</td>
<td>94.8</td>
<td>88.1/63.3</td>
<td>94.1/93.4</td>
<td>92.5</td>
<td>76.9</td>
<td>93.8</td>
<td>92.7/91.9</td>
<td>65.6</td>
</tr>
<tr>
<td>5</td>
<td>Huawei Noah's Ark Lab</td>
<td>NEZHA-Plus</td>
<td>☑</td>
<td>86.7</td>
<td>87.8</td>
<td>94.4/96.0</td>
<td>93.6</td>
<td>84.6/55.1</td>
<td>90.1/89.6</td>
<td>89.1</td>
<td>74.6</td>
<td>93.2</td>
<td>87.1/74.4</td>
<td>58.0</td>
</tr>
</tbody>
</table>
Machine Translation

• Task: (NL -> NL)
• Metric: ?

<table>
<thead>
<tr>
<th>German</th>
<th>English</th>
</tr>
</thead>
<tbody>
<tr>
<td>Auf dem Tisch ist eine Katze.</td>
<td>There is a cat on the table.</td>
</tr>
<tr>
<td></td>
<td>On the table is a cat.</td>
</tr>
<tr>
<td></td>
<td>There's a cat on the table.</td>
</tr>
<tr>
<td></td>
<td>A cat is on a table.</td>
</tr>
<tr>
<td></td>
<td>On the table is a kitty cat.</td>
</tr>
<tr>
<td></td>
<td>On the table is a bird.</td>
</tr>
<tr>
<td></td>
<td>On the desk there is a cat.</td>
</tr>
<tr>
<td></td>
<td>Cat on table.</td>
</tr>
<tr>
<td></td>
<td>There are cats on the table.</td>
</tr>
<tr>
<td></td>
<td>There was a cat on the table.</td>
</tr>
</tbody>
</table>

16
Machine Translation

• Task: (NL -> NL)

• Metric:
  • BLEU (Bilingual Evaluation Understudy)
  • ROUGE (Recall-Oriented Understudy for Gisting Evaluation)
  • METEOR (Metric for Evaluation of Translation with Explicit ORdering)
N-Gram

• An N-gram is a contiguous sequence of N items from a given sample of text or speech.
• The items can be phonemes, syllables, letters, words or base pairs according to the application.

<table>
<thead>
<tr>
<th>Name</th>
<th>N</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Unigram</td>
<td>1</td>
<td>A</td>
</tr>
<tr>
<td>Bigram</td>
<td>2</td>
<td>AB</td>
</tr>
<tr>
<td>Trigram</td>
<td>3</td>
<td>UNO</td>
</tr>
<tr>
<td>4-gram</td>
<td>4</td>
<td>DATA</td>
</tr>
<tr>
<td>N-gram</td>
<td>N</td>
<td></td>
</tr>
</tbody>
</table>
## N-gram Precision

<table>
<thead>
<tr>
<th>Candidate</th>
<th>the</th>
<th>the</th>
<th>the</th>
<th>the</th>
<th>the</th>
<th>the</th>
<th>the</th>
<th>the</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference 1</td>
<td>the</td>
<td>cat</td>
<td>is</td>
<td>on</td>
<td>the</td>
<td>mat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference 2</td>
<td>there</td>
<td>is</td>
<td>a</td>
<td>cat</td>
<td>on</td>
<td>the</td>
<td>mat</td>
<td></td>
</tr>
</tbody>
</table>

\[
\text{Unigram - Precision} = \frac{m}{w_t} = \frac{7}{7} = 1
\]

\[
\text{Bigram - Precision} = \frac{m}{w_t} = \frac{0}{1} = 0
\]

- \( m \): N-grams from the candidate that are found in the reference
- \( w_t \): N-grams in the candidate
BLEU

• Main idea: The closer a machine translation is to a professional human translation, the better it is.

• Scores are calculated over sentences by comparing them with a set of good quality reference translations.

• Scores are then averaged over the whole text to output a final score in [0,1]. (1 good translation, 0 bad translation)
## BLEU

<table>
<thead>
<tr>
<th>Candidate</th>
<th>the</th>
<th>the</th>
<th>the</th>
<th>the</th>
<th>the</th>
<th>the</th>
<th>the</th>
<th>the</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference 1</td>
<td>the</td>
<td>cat</td>
<td>is</td>
<td>on</td>
<td>the</td>
<td>mat</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference 2</td>
<td>there</td>
<td>is</td>
<td>a</td>
<td>cat</td>
<td>on</td>
<td>the</td>
<td>mat</td>
<td></td>
</tr>
</tbody>
</table>

\[
Unigram - Precision = \frac{\sum_w \min(m_w, m_w^{max})}{w_t} = \frac{2}{7}
\]

- \(m_w\): N-grams from the candidate that are found in the reference for N-gram \(w\)
- \(m_w^{max}\): For each N-gram in the candidate translation, the algorithm takes its maximum total count in any of the reference translations.
- \(w_t\): N-grams in the candidate
BLEU

<table>
<thead>
<tr>
<th>Candidate</th>
<th>the</th>
<th>cat</th>
</tr>
</thead>
<tbody>
<tr>
<td>Reference 1</td>
<td>the</td>
<td>cat</td>
</tr>
<tr>
<td>Reference 2</td>
<td>there</td>
<td>is</td>
</tr>
</tbody>
</table>

\[
Unigram - Precision = \frac{\sum_w \min(m_w, m_{w}^{\text{max}})}{w_t} = \frac{2}{2} = 1
\]

\[
Bigram - Precision = \frac{\sum_w \min(m_w, m_{w}^{\text{max}})}{w_t} = \frac{1}{1} = 1
\]

Problems:
- Favors short translations
- Adding more reference translation increases score
- Score ≠ Quality

Solutions:
- Punish too short translations
- Consider also recall
ROUGE

• ROUGE-N
• ROUGE-L
• ROUGE-W
• ROUGE-S
ROUGE-N

ROUGE-N (Recall)

\[
\frac{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} \text{Count}_{\text{match}}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} \text{Count}(gram_n)} \quad (1)
\]

BLEU (Precision)

\[
\frac{\sum_{C \in \{Candidates\}} \sum_{n-gram \in C} \text{Count}_{\text{clip}}(n-gram)}{\sum_{C' \in \{Candidates\}} \sum_{n-gram' \in C'} \text{Count}(n-gram')}
\]
**ROUGE-N**

<table>
<thead>
<tr>
<th>Table</th>
<th>Word 1</th>
<th>Word 2</th>
<th>Word 3</th>
<th>Word 4</th>
<th>Word 5</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate 1</td>
<td>I</td>
<td>always</td>
<td>invariably</td>
<td>perpetually</td>
<td>do</td>
</tr>
<tr>
<td>Candidate 2</td>
<td>I</td>
<td>always</td>
<td>do</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference 1</td>
<td>I</td>
<td>always</td>
<td>do</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference 2</td>
<td>I</td>
<td>invariably</td>
<td>do</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference 3</td>
<td>I</td>
<td>perpetually</td>
<td>do</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Candidate 1: \( \text{Unigram - Recall} = \frac{9}{9} = 1 \)

Candidate 2: \( \text{Unigram - Recall} = \frac{7}{9} = 0.777 \)
### ROUGE-N

<table>
<thead>
<tr>
<th>Candidate 1</th>
<th>I</th>
<th>always</th>
<th>invariably</th>
<th>perpetually</th>
<th>do</th>
</tr>
</thead>
<tbody>
<tr>
<td>Candidate 2</td>
<td>I</td>
<td>always</td>
<td>do</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference 1</td>
<td>I</td>
<td>always</td>
<td>do</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference 2</td>
<td>I</td>
<td>invariably</td>
<td>do</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Reference 3</td>
<td>I</td>
<td>perpetually</td>
<td>do</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

\[
ROUGE - N_{Multi} = \max_{r \in References} ROUGE - N(r, candidate)
\]

Candidate 1: \( ROUGE - 1_{Multi} = \frac{3}{3} = 1 \)

Candidate 2: \( ROUGE - 1_{Multi} = \frac{3}{3} = 1 \)
ROUGE

• ROUGE-L: Longest Common Subsequence
• ROUGE-W: Weighted Longest Common Subsequence
• ROUGE-S: Skip-Bigram Co-Occurrence Statistics
METEOR

• Successor of BLEU
• Based on weighted harmonic mean of unigram precision and recall
• Stemming
  • fishing, fished, fisher = fish
• Synonym Matching
  • buy = purchase
METEOR

• Find maximum alignment with the fewest crosses

\[ P = \frac{m}{w_t} = \frac{6}{6} = 1 \]

\[ R = \frac{m}{w_r} = \frac{6}{6} = 1 \]

• \( m \): Number of connections between candidate and reference
• \( w_r \): Number of Unigrams in the reference
• \( w_t \): Number of Unigrams in the candidate
METEOR

• Combine precision and recall

$$F_{mean} = \frac{10PR}{R + 9P}$$

• Calculate penalty $p$ on number of crosses

$$M = F_{mean}(1 - p)$$

• If there are more reference solutions, select highest score
  • Same as ROUGE
Is there a limit?

- GPT-3
  - General language model
  - 175 billion parameters
  - Turing test?
Giving GPT-3 a Turing Test – Common Sense

Q: What is your favorite animal?
A: My favorite animal is a dog.

Q: Why?
A: Because dogs are loyal and friendly.

Q: What are two reasons that a dog might be in a bad mood?
A: Two reasons that a dog might be in a bad mood are if it is hungry or if it is hot.

Q: How many eyes does a giraffe have?
A: A giraffe has two eyes.

Q: How many legs does a frog have?
A: A frog has four legs.

Q: Are there any animals with three legs?
A: No, there are no animals with three legs.

Q: Why don't animals have three legs?
A: Animals don't have three legs because they would fall over.
Giving GPT-3 a Turing Test – Common Sense

Q: How do you sporgle a morgle?  
A: You sporgle a morgle by using a sporgle.

Q: How many bonks are in a quoit?  
A: There are three bonks in a quoit.

Q: How many rainbows does it take to jump from Hawaii to seventeen?  
A: It takes two rainbows to jump from Hawaii to seventeen.

Q: Which colorless green ideas sleep furiously?  
A: Ideas that are colorless, green, and sleep furiously are the ideas of a sleep furiously.

Q: Do you understand these questions?  
A: I understand these questions.
Q: When counting, what number comes before 100?  
A: Ninety-nine comes before 100.

Q: When counting, what number comes before 123?  
A: One hundred and twenty-two comes before 123.

Q: When counting, what number comes before 1000?  
A: Nine hundred and ninety-nine comes before 1000.

Q: When counting, what number comes before 10,000?  
A: Nine thousand and ninety-nine comes before 10,000.

Q: When counting, what number comes before a million?  
A: Nine hundred thousand and ninety-nine comes before a million.

Q: Write one line of Ruby code to reverse an array.  
A: ary.reverse

Q: Write one line of Ruby code to count how many X's are in an array.  
A: ary.count { |x| x == 'X' }

Q: Reverse the following array: [1, 3, 5, 6, 10, 4, 2, 77]  
A: [10, 6, 4, 2, 77, 3, 5, 1]

Q: How many X's are in this array: [1, 3, 5, 6, 10, 4, 2, 77]  
A: 2
Giving GPT-3 a Turing Test – Logic

Q: If I have a marble and a paper clip in a box, put a pencil in the box, and remove the marble, what is left?  
A: A paper clip.

Q: If I have two shoes in a box, put a pencil in the box, and remove one shoe, what is left?  
A: A shoe.

Q: If I put a pencil in a box, then put another pencil in the box, what is in the box?  
A: Two pencils.

Q: Then if I take out a pencil and put in a frog, what is in the box?  
A: A frog.
Giving GPT-3 a Turing Test

• Some of the problems can be solved by using a different prompt
  • E.g. If the question is nonsense say «get real»
Conclusion

• NLP is everywhere
• Benchmarks: GLUE and SuperGLUE
  • State-of-the-art performance
• Metrics for machine translation
  • BLEU
  • ROUGE
  • METEOR
• Limitations of GPT-3