Natural Language Processing with Deep Learning
CS224N/Ling284

Tatsunori Hashimoto
Lecture 13: Evaluation
Benchmarks and evaluations drive progress
Two major types of evaluations

<table>
<thead>
<tr>
<th>Text</th>
<th>Judgments</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man inspects the uniform of a figure in some East Asian country.</td>
<td>contradiction C C C C C</td>
<td>The man is sleeping</td>
</tr>
<tr>
<td>An older and younger man smiling.</td>
<td>neutral N N E N N</td>
<td>Two men are smiling and</td>
</tr>
<tr>
<td>A black race car starts up in front of a crowd of people.</td>
<td>contradiction C C C C C</td>
<td>A man is driving down a</td>
</tr>
<tr>
<td>A soccer game with multiple males playing.</td>
<td>entailment E E E E E E</td>
<td>Some men are playing a</td>
</tr>
<tr>
<td>A smiling costumed woman is holding an umbrella.</td>
<td>neutral  N N E C N</td>
<td>A happy woman in a fair</td>
</tr>
</tbody>
</table>

Close-ended evaluations

Open ended evaluations
Classification and closed-ended benchmarks

• Many NLP tasks are ‘closed-ended’
  • Limited number of potential answers
  • Often one or just a few correct answers

• Examples:
  • Sentiment classification (sentiment label)
  • Extractive QA (the part of the document that has the answer)

• Enables automatic evaluation
  • Similar to the usual machine learning evaluations
Single-task benchmarks

<table>
<thead>
<tr>
<th>Text</th>
<th>Judgments</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man inspects the uniform of a figure in some East Asian country.</td>
<td>contradiction (CCCCC)</td>
<td>The man is sleeping</td>
</tr>
<tr>
<td>An older and younger man smiling.</td>
<td>neutral (NNENN)</td>
<td>Two men are smiling and</td>
</tr>
<tr>
<td>A black race car starts up in front of a crowd of people.</td>
<td>contradiction (CCCCC)</td>
<td>A man is driving down a</td>
</tr>
<tr>
<td>A soccer game with multiple males playing.</td>
<td>entailment (EEEEEE)</td>
<td>Some men are playing a</td>
</tr>
<tr>
<td>A smiling costumed woman is holding an umbrella.</td>
<td>neutral (NNECN)</td>
<td>A happy woman in a fair</td>
</tr>
</tbody>
</table>

SST, IMDB (Sentiment)  SNLI, MultiNLI (entailment)  SQuAD, NaturalQuestions (QA)
Multi-task benchmark - superGLUE

Leaderboard Version: 2.0

<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-b</th>
<th>AX-g</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>JDExplore d-team</td>
<td>Vega v2</td>
<td>🔴</td>
<td>91.3</td>
<td>90.5</td>
<td>98.6/99.2</td>
<td>99.4</td>
<td>88.2/62.4</td>
<td>94.4/93.9</td>
<td>96.0</td>
<td>77.4</td>
<td>98.6</td>
<td>-0.4</td>
<td>100.0/50.0</td>
</tr>
<tr>
<td>2</td>
<td>Liam Fedus</td>
<td>ST-MoE-32B</td>
<td>🔴</td>
<td>91.2</td>
<td>92.4</td>
<td>96.9/96.0</td>
<td>99.2</td>
<td>89.6/65.8</td>
<td>95.1/94.4</td>
<td>93.5</td>
<td>77.7</td>
<td>96.6</td>
<td>72.3</td>
<td>96.1/94.1</td>
</tr>
<tr>
<td>3</td>
<td>Microsoft Alexander v-team</td>
<td>Turing NLR v5</td>
<td>🔴</td>
<td>90.9</td>
<td>92.0</td>
<td>95.9/97.6</td>
<td>98.2</td>
<td>88.4/63.0</td>
<td>96.4/95.9</td>
<td>94.1</td>
<td>77.1</td>
<td>97.3</td>
<td>67.8</td>
<td>93.3/95.5</td>
</tr>
<tr>
<td>4</td>
<td>ERNIE Team - Baidu</td>
<td>ERNIE 3.0</td>
<td>🔴</td>
<td>90.6</td>
<td>91.0</td>
<td>98.6/99.2</td>
<td>97.4</td>
<td>88.6/63.2</td>
<td>94.7/94.2</td>
<td>92.6</td>
<td>77.4</td>
<td>97.3</td>
<td>68.6</td>
<td>92.7/94.7</td>
</tr>
<tr>
<td>5</td>
<td>Yi Tay</td>
<td>PaLM 540B</td>
<td>🔴</td>
<td>90.4</td>
<td>91.9</td>
<td>94.4/96.0</td>
<td>99.0</td>
<td>88.7/63.6</td>
<td>94.2/93.3</td>
<td>94.1</td>
<td>77.4</td>
<td>95.9</td>
<td>72.9</td>
<td>95.5/90.4</td>
</tr>
<tr>
<td>6</td>
<td>Zirui Wang</td>
<td>T5 + UDG, Single Model (Google Brain)</td>
<td>🔴</td>
<td>90.4</td>
<td>91.4</td>
<td>95.8/97.6</td>
<td>98.0</td>
<td>88.3/63.0</td>
<td>94.2/93.5</td>
<td>93.0</td>
<td>77.9</td>
<td>96.6</td>
<td>69.1</td>
<td>92.7/91.9</td>
</tr>
<tr>
<td>7</td>
<td>DeBERTa Team - Microsoft</td>
<td>DeBERTa / TuringNLRv4</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>66.7</td>
<td>93.3/93.8</td>
</tr>
<tr>
<td>8</td>
<td>SuperGLUE Human Baselines</td>
<td>SuperGLUE Human Baselines</td>
<td>🔴</td>
<td>89.8</td>
<td>89.0</td>
<td>95.8/96.9</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>76.6</td>
<td>99.3/99.7</td>
</tr>
<tr>
<td>9</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>65.6</td>
<td>92.7/91.9</td>
</tr>
</tbody>
</table>

Attempt to measure “general language capabilities”
Examples from superGLUE

Cover a number of different tasks

- BoolQ, MultiRC (reading texts)
- CB, RTE (Entailment)
- COPA (cause and effect)
- ReCoRD (QA+reasoning)
- WiC (meaning of words)
- WSC (coreference)

<table>
<thead>
<tr>
<th>Task</th>
<th>Passage</th>
<th>Question</th>
<th>Answer</th>
<th>Text</th>
<th>Hypothesis</th>
<th>Entailment</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoolQ</td>
<td>Barq's - Barq's is an American soft drink. Its brand of root beer is notable for having caffeine. Barq's, created by Edward Barq and bottled since the turn of the 20th century, is owned by the Barq family but bottled by the Coca-Cola Company. It was known as Barq’s Famous Olde Tyme Root Beer until 2012.</td>
<td>is Barq's root beer a pepsi product</td>
<td>No</td>
<td>And yet, uh, I we-, I hope to see employer based, you know, helping out. You know, child, uh, care centers at the place of employment and things like that, that will help out. A: Uh-huh. B: What do you think, do you think we are, setting a trend?</td>
<td>they are setting a trend</td>
<td>Unknown</td>
</tr>
<tr>
<td>CB</td>
<td>Premise: My body cast a shadow over the grass. Question: What's the CAUSE for this? Alternative 1: The sun was rising. Alternative 2: The grass was cut. Correct Alternative: 1</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>COPA</td>
<td>Paragraph: Susan wanted to have a birthday party. She called all of her friends. She has five friends. Her mom said that Susan can invite them all to the party. Her first friend could not go to the party because she was sick. Her second friend was going out of town. Her third friend was not so sure if her parents would let her. The fourth friend said maybe. The fifth friend could go to the party for sure. Susan was a little sad. On the day of the party, all five friends showed up. Each friend had a present for Susan. Susan was happy and sent each friend a thank you card the next week. Question: Did Susan’s sick friend recover? Candidate answers: Yes, she recovered (T), No (F), Yes (T), No, she didn’t recover (F), Yes, she was at Susan’s party (T)</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>ReCoRD</td>
<td>Paragraph: (CNN) Puerto Rico on Sunday overwhelmingly voted for statehood. But Congress, the only body that can approve new states, will ultimately decide whether the status of the US commonwealth changes. Ninety-seven percent of the votes in the nonbinding referendum favored statehood, an increase over the official results of a 2012 referendum. Official results from the State Electoral Commission show. It was the fifth such vote on statehood. “Today, we the people of Puerto Rico are sending a strong and clear message to the US Congress ... and to the world ... claiming our equal rights as American citizens, Puerto Rico Gov. Ricardo Rossello said in a news release. @highlight Puerto Rico voted Sunday in favor of US statehood Query For one, they can truthfully say, “Don’t blame me, I didn’t vote for them,” when discussing the &lt;placeholder&gt; presidency Correct Entities: US</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>RTE</td>
<td>Text: Dana Reeve, the widow of the actor Christopher Reeve, has died of lung cancer at age 44, according to the Christopher Reeve Foundation. Hypothesis: Christopher Reeve had an accident. Entailment: False</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WiC</td>
<td>Context 1: Room and board. Context 2: He nailed boards across the windows. Sense match: False</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>WSC</td>
<td>Text: Mark told Pete many lies about himself, which Pete included in his book. He should have been more truthful. Coreference: False</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
Recap: MMLU

Massive Multitask Language Understanding (MMLU) [Hendrycks et al., 2021]

New benchmarks for measuring LM performance on 57 diverse knowledge intensive tasks
Some intuition: examples from MMLU

Astronomy

What is true for a type-Ia supernova?
   A. This type occurs in binary systems.
   B. This type occurs in young galaxies.
   C. This type produces gamma-ray bursts.
   D. This type produces high amounts of X-rays.

Answer: A

High School Biology

In a population of giraffes, an environmental change occurs that favors individuals that are tallest. As a result, more of the taller individuals are able to obtain nutrients and survive to pass along their genetic information. This is an example of

   A. directional selection.
   B. stabilizing selection.
   C. sexual selection.
   D. disruptive selection

Answer: A
What makes a good benchmark?

- **Example selection (scale, diversity)**
  - Benchmark should cover the phenomena of interest
  - Complex phenomena require many samples

- **Difficulty**
  - Doable for humans
  - Hard for baselines at the time

- **Annotation quality**
  - ‘Correct’ behavior should be clear
One example of a successful benchmark (SQuAD)

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Question source</th>
<th>Formulation</th>
<th>Size</th>
</tr>
</thead>
<tbody>
<tr>
<td>SQuAD</td>
<td>crowdsourced</td>
<td>RC, spans in passage</td>
<td>100K</td>
</tr>
<tr>
<td>MCTest</td>
<td>crowdsourced</td>
<td>RC, multiple choice</td>
<td>2640</td>
</tr>
<tr>
<td>Richardson et al., 2013</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Algebra</td>
<td>standardized tests</td>
<td>reasoning, multiple choice</td>
<td>514</td>
</tr>
<tr>
<td>Kushman et al., 2014</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Science</td>
<td>standardized tests</td>
<td></td>
<td>855</td>
</tr>
<tr>
<td>Clark and Etzioni, 2016</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Scale (and inclusion of training data)

<table>
<thead>
<tr>
<th>Exact Match</th>
<th>F1</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Dev</td>
</tr>
<tr>
<td>Random Guess</td>
<td>1.1%</td>
</tr>
<tr>
<td>Sliding Window</td>
<td>13.2%</td>
</tr>
<tr>
<td>Sliding Win. + Dist.</td>
<td>13.3%</td>
</tr>
<tr>
<td>Logistic Regression</td>
<td>40.0%</td>
</tr>
<tr>
<td>Human</td>
<td>80.3%</td>
</tr>
</tbody>
</table>

A prime number (or a prime) is a natural number greater than 1 that has no positive divisors other than 1 and itself. A natural number greater than 1 that is not a prime number is called a composite number. For example, 5 is prime because 1 and 5 are its only positive integer factors, whereas 6 is composite because it has the divisors 2 and 3 in addition to 1 and 6. The fundamental theorem of arithmetic establishes the central role of primes in number theory: any integer greater than 1 can be expressed as a product of primes that is unique up to ordering. The uniqueness in this theorem requires excluding 1 as a prime because one can include arbitrarily many instances of 1 in any factorization, e.g., 3, 1 · 3, 1 · 1 · 3, etc. are all valid factorizations of 3.

What is the only divisor besides 1 that a prime number can have?
*Ground Truth Answers: itself itself itself itself itself*

What are numbers greater than 1 that can be divided by 3 or more numbers called?
*Ground Truth Answers: composite number composite number composite number primes*

What theorem defines the main role of primes in number theory?
*Ground Truth Answers: The fundamental theorem of arithmetic fundamental theorem of arithmetic fundamental theorem of arithmetic fundamental theorem of arithmetic fundamental theorem of arithmetic*

Easy, relatively clean automatic evaluation

Large headroom to human perf
One example of a good benchmark with a flaw

<table>
<thead>
<tr>
<th>Text</th>
<th>Judgments</th>
<th>Hypothesis</th>
</tr>
</thead>
<tbody>
<tr>
<td>A man inspects the uniform of a figure in some East Asian country.</td>
<td>contradiction C C C C</td>
<td>The man is sleeping</td>
</tr>
<tr>
<td>An older and younger man smiling.</td>
<td>neutral N N E N N</td>
<td>Two men are smiling and laughing at the cats playing on the floor.</td>
</tr>
</tbody>
</table>

Premise:
The economy could be still better.

Hypothesis:
The economy has never been better

The dataset itself is hard, but there can be undiscovered *spurious correlations*
Targeted and adversarial evaluations

- The ‘negation bias’ issues show that plain benchmarks can miss things

- More targeted benchmarking
  - Can models do well when you modify specific parts of the input?
  - What about negating both inputs and outputs?

- More adversarial benchmarking
  - Models can exploit spurious correlations
  - Evaluate models adversarially (where they can’t exploit spurious features)
Model evaluation as model analysis in natural language inference

Recall the natural language inference task, as encoded in the Multi-NLI dataset.

Premise
“He turned and saw Jon sleeping in his half-tent”

Hypothesis
“He saw Jon was asleep”

Model A
Accuracy: 95%

Entailment
Neutral
Contradiction

[ Likely to get the right answer, since the accuracy is 95%? ]

[Williams et al., 2018]
Model evaluation as model analysis in natural language inference

What if our model is using simple heuristics to get good accuracy?

A **diagnostic test set** is carefully constructed to test for a specific skill or capacity of your neural model.

For example, **HANS**: (Heuristic Analysis for NLI Systems) tests syntactic heuristics in NLI

<table>
<thead>
<tr>
<th>Heuristic</th>
<th>Definition</th>
<th>Example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Lexical overlap</td>
<td>Assume that a premise entails all hypotheses constructed from words in the premise</td>
<td><em>The doctor</em> was <strong>paid</strong> by <em>the actor</em>.</td>
</tr>
<tr>
<td>Subsequence</td>
<td>Assume that a premise entails all of its contiguous subsequences.</td>
<td>*The doctor near <em>the actor</em> <strong>danced</strong>.</td>
</tr>
<tr>
<td>Constituent</td>
<td>Assume that a premise entails all complete subtrees in its parse tree.</td>
<td><em>If <em>the artist</em> <strong>slept</strong>, the actor <strong>ran</strong>.</em></td>
</tr>
</tbody>
</table>

[McCoy et al., 2019]
McCoy et al., 2019 took 4 strong MNLI models, with the following accuracies on the original test set (in-domain)

Evaluating on HANS, where syntactic heuristics work, accuracy is high!

But where syntactic heuristics fail, accuracy is very very low...
Careful test sets as unit test suites: CheckListing

- Small careful test sets sound like... unit test suites, but for neural networks!
- Minimum functionality tests: small test sets that target a specific behavior.

<table>
<thead>
<tr>
<th>Test case</th>
<th>Expected</th>
<th>Predicted</th>
<th>Pass?</th>
</tr>
</thead>
<tbody>
<tr>
<td>Testing Negation with <strong>MFT</strong></td>
<td>Labels: negative, positive, neutral</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Template: I {NEGATION} {POS_VERB} the {THING}.</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>I can’t say I recommend the food.</td>
<td>neg</td>
<td>pos</td>
<td>X</td>
</tr>
<tr>
<td>I didn’t love the flight.</td>
<td>neg</td>
<td>neutral</td>
<td>X</td>
</tr>
<tr>
<td>...</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Failure rate = 76.4%

- Ribeiro et al., 2020 showed ML engineers working on a sentiment analysis product an interface with categories of linguistic capabilities and types of tests.
  - The engineers found a bunch of bugs (categories of high error) through this method!
Fitting the dataset vs learning the task

Across a wide range of tasks, high model accuracy on the in-domain test set does not imply the model will also do well on other, “reasonable” out-of-domain examples.

One way to think about this: models seem to be learning the *dataset* (like MNLI) not the *task* (like how humans can perform natural language inference).

[Ribeiro et al., 2020](#)
Adversarial (and multi objective) benchmarking

Adversarial NLI (ANLI)

DynaBench
Evaluating open-ended text generation

- From ‘few correct answers’ to ‘thousands of correct answers’

- Can’t have human annotators enumerate the right answers (or can we?)

- There are now better and worse answers (not just right and wrong)
Types of evaluation methods for text generation

Ref: They walked to the grocery store.
Gen: The woman went to the hardware store.

Content Overlap Metrics
Model-based Metrics
Human Evaluations

(Some slides repurposed from Asli Celikyilmaz from EMNLP 2020 tutorial)
Content overlap metrics

Ref: They walked to the grocery store.

Gen: The woman went to the hardware store.

• Compute a score that indicates the lexical similarity between generated and gold-standard (human-written) text
• Fast and efficient and widely used
• N-gram overlap metrics (e.g., BLEU, ROUGE, METEOR, CIDEr, etc.)
**N-gram overlap metrics**

Word overlap–based metrics (BLEU, ROUGE, METEOR, CIDEr, etc.)

- They’re *not ideal* for machine translation

- They get progressively *much worse* for tasks that are more open-ended than machine translation
  - *Worse* for summarization, as longer output texts are harder to measure
  - *Much worse* for dialogue, which is more open-ended than summarization
  - *Much, much worse* story generation, which is also open-ended, but whose sequence length can make it seem you’re getting decent scores!
A simple failure case

\[ n \text{-gram overlap metrics have no concept of semantic relatedness!} \]

Are you enjoying the CS224N lectures?

Score:
- 0.61: Yes!
- 0.25: You know it!
- False negative: 0
- False positive: 0.67

Heck yes!
Semantic overlap metrics

**PYRAMID:**
- Incorporates human content selection variation in summarization evaluation.
- Identifies **Summarization Content Units (SCU)**s to compare information content in summaries.

(Nenkova, et al., 2007)

**SPICE:**
Semantic propositional image caption evaluation is an image captioning metric that initially parses the reference text to derive an abstract scene graph representation.

(Anderson et al., 2016)

**SPIDER:**
A combination of semantic graph similarity (**SPICE**) and n-gram similarity measure (**CIDER**), the SPICE metric yields a more complete quality evaluation metric.

(Liu et al., 2017)
Model-based metrics to capture more semantics

- Use learned representations of words and sentences to compute semantic similarity between generated and reference texts.

- No more n-gram bottleneck because text units are represented as embeddings!

- The embeddings are pretrained, distance metrics used to measure the similarity can be fixed.
Model-based metrics: Word distance functions

**Vector Similarity**
Embedding based similarity for semantic distance between text.

- Embedding Average (Liu et al., 2016)
- Vector Extrema (Liu et al., 2016)
- MEANT (Lo, 2017)
- YISI (Lo, 2019)

**Word Mover’s Distance**
Measures the distance between two sequences (e.g., sentences, paragraphs, etc.), using word embedding similarity matching.
(Kusner et al., 2015; Zhao et al., 2019)

**BERTSCORE**
Uses pre-trained contextual embeddings from BERT and matches words in candidate and reference sentences by cosine similarity.
(Zhang et al. 2020)
Model-based metrics: Beyond word matching

---

**BLEURT:**

A regression model based on BERT returns a score that indicates to what extent the candidate text is grammatical and conveys the meaning of the reference text.

(Sellam et.al. 2020)

---

**Sentence Movers Similarity:**

Based on Word Movers Distance to evaluate text in a continuous space using sentence embeddings from recurrent neural network representations.

(Clark et.al., 2019)
MAUVE computes information divergence in a quantized embedding space, between the generated text and the gold reference text (Pillutla et al., 2022).
Figure 3: Illustration of the quantization. **Left:** A continuous two-dimensional distribution $P$. **Right:** A partitioning of the Euclidean plane $\mathbb{R}^2$ and the corresponding quantized distribution $\tilde{P}$. 
An important failure case

Reference-based measures are only as good as their references.

<table>
<thead>
<tr>
<th></th>
<th>CNN</th>
<th>Daily Mail</th>
</tr>
</thead>
<tbody>
<tr>
<td># months</td>
<td>95</td>
<td>1</td>
</tr>
<tr>
<td># documents</td>
<td>90,266</td>
<td>1220</td>
</tr>
<tr>
<td># queries</td>
<td>380,298</td>
<td>3924</td>
</tr>
<tr>
<td>Max # entities</td>
<td>527</td>
<td>187</td>
</tr>
<tr>
<td>Avg # entities</td>
<td>26.4</td>
<td>26.5</td>
</tr>
<tr>
<td>Avg # tokens</td>
<td>762</td>
<td>763</td>
</tr>
<tr>
<td>Vocab size</td>
<td>118,497</td>
<td>208,045</td>
</tr>
</tbody>
</table>

Table 1: Corpus statistics. Articles were collected starting in April 2007 for CNN and June 2010 for the Daily Mail, both until the end of April 2015. Validation data is from March, test data from April 2015. Articles of over 2000 tokens and queries whose answer entity did not appear in the context were filtered out.

CNN/Daily Mail dataset

Not correlated at all!

- Reference-based measures are only as good as their references.
Don’t blindly trust references in datasets!

<table>
<thead>
<tr>
<th>Setting</th>
<th>Models</th>
<th>CNN/Daily Mail</th>
<th>XSUM</th>
</tr>
</thead>
<tbody>
<tr>
<td></td>
<td>Faithfulness</td>
<td>Coherence</td>
<td>Relevance</td>
</tr>
<tr>
<td>Zero-shot language models</td>
<td>GPT-3 (350M)</td>
<td>0.29</td>
<td>1.92</td>
</tr>
<tr>
<td></td>
<td>GPT-3 (6.7B)</td>
<td>0.29</td>
<td>1.77</td>
</tr>
<tr>
<td></td>
<td>GPT-3 (175B)</td>
<td>0.76</td>
<td>2.65</td>
</tr>
<tr>
<td></td>
<td>Ada Instruct v1 (350M*)</td>
<td>0.88</td>
<td>4.02</td>
</tr>
<tr>
<td></td>
<td>Curie Instruct v1 (6.7B*)</td>
<td>0.97</td>
<td>4.24</td>
</tr>
<tr>
<td></td>
<td>Davinci Instruct v2 (175B*)</td>
<td><strong>0.99</strong></td>
<td>4.15</td>
</tr>
<tr>
<td>Five-shot language models</td>
<td>Anthropic-LM (52B)</td>
<td>0.94</td>
<td>3.88</td>
</tr>
<tr>
<td></td>
<td>Cohere XL (52.4B)</td>
<td><strong>0.99</strong></td>
<td>3.42</td>
</tr>
<tr>
<td></td>
<td>GLM (130B)</td>
<td>0.94</td>
<td>3.69</td>
</tr>
<tr>
<td></td>
<td>OPT (175B)</td>
<td>0.96</td>
<td>3.64</td>
</tr>
<tr>
<td></td>
<td>GPT-3 (350M)</td>
<td>0.86</td>
<td>3.73</td>
</tr>
<tr>
<td></td>
<td>GPT-3 (6.7B)</td>
<td>0.97</td>
<td>3.87</td>
</tr>
<tr>
<td></td>
<td>GPT-3 (175B)</td>
<td><strong>0.99</strong></td>
<td>3.95</td>
</tr>
<tr>
<td></td>
<td>Ada Instruct v1 (350M*)</td>
<td>0.84</td>
<td>3.84</td>
</tr>
<tr>
<td></td>
<td>Curie Instruct v1 (6.7B*)</td>
<td>0.96</td>
<td><strong>4.30</strong></td>
</tr>
<tr>
<td></td>
<td>Davinci Instruct v2 (175B*)</td>
<td><strong>0.98</strong></td>
<td>4.13</td>
</tr>
<tr>
<td>Fine-tuned language models</td>
<td>Brio</td>
<td>0.94</td>
<td>3.94</td>
</tr>
<tr>
<td></td>
<td>Pegasus</td>
<td>0.97</td>
<td>3.93</td>
</tr>
<tr>
<td>Existing references</td>
<td>-</td>
<td>0.84</td>
<td>3.20</td>
</tr>
</tbody>
</table>

Training on references actually makes model worse!
How to evaluate an evaluation metric?

Figure 1: Scatter plots showing the correlation between metrics and human judgements on the Twitter corpus (a) and Ubuntu Dialogue Corpus (b). The plots represent BLEU-2 (left), embedding average (center), and correlation between two randomly selected halves of human respondents (right).

(Liu et al, EMNLP 2016)
Reference free evals

• **Reference-based evaluation:**
  • Compare human written reference to model outputs
  • ‘Standard’ evaluation for most NLP tasks
  • Examples: BLEU, ROUGE, BertScore etc.

• **Reference free evaluation:**
  • Have a model give a score
  • No human reference
  • Was nonstandard – now becoming popular with GPT4
  • Examples: FactCC, GPT-4-as-judge, AlpacaEval
Sophisticated summarization factuality metrics (FactCC / DA) are less correlated with humans than overlap!
Human evaluations

- Automatic metrics fall short of matching human decisions

- Human evaluation is most important form of evaluation for text generation systems.

- Gold standard in developing new automatic metrics
  - New automated metrics must correlate well with human evaluations!
Human evaluations

• **Ask humans** to evaluate the quality of generated text

• Overall or along some specific dimension:
  • fluency
  • coherence / consistency
  • factuality and correctness
  • commonsense
  • style / formality
  • grammaticality
  • typicality
  • redundancy

**Note:** Don’t compare human evaluation scores across differently conducted studies

Even if they claim to evaluate the same dimensions!
Human evaluation: Issues

• Human judgments are regarded as the gold standard
• Of course, we know that human eval is slow and expensive
• Beyond the cost of human eval, it’s still far from perfect:

• Humans Evaluation is hard:
  • Results are inconsistent / not reproducible
  • can be illogical
  • misinterpret your question
  • Precision not recall.
  • ...
Learning from human feedback

ADEM:

A learned metric from human judgments for dialog system evaluation in a chatbot setting.

(Lowe et.al., 2017)

HUSE:

Human Unified with Statistical Evaluation (HUSE), determines the similarity of the output distribution and a human reference distribution.

(Hashimoto et.al. 2019)
Evaluating language models as chatbots

How do we evaluate something like ChatGPT?
• **So many** different use cases it’s hard to evaluate
• The responses are also long-form text, which is even harder to evaluate.

Table 1: Distribution of use case categories from our API prompt dataset.

<table>
<thead>
<tr>
<th>Use-case</th>
<th>(%)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Generation</td>
<td>45.6%</td>
</tr>
<tr>
<td>Open QA</td>
<td>12.4%</td>
</tr>
<tr>
<td>Brainstorming</td>
<td>11.2%</td>
</tr>
<tr>
<td>Chat</td>
<td>8.4%</td>
</tr>
<tr>
<td>Rewrite</td>
<td>6.6%</td>
</tr>
<tr>
<td>Summarization</td>
<td>4.2%</td>
</tr>
<tr>
<td>Classification</td>
<td>3.5%</td>
</tr>
<tr>
<td>Other</td>
<td>3.5%</td>
</tr>
<tr>
<td>Closed QA</td>
<td>2.6%</td>
</tr>
<tr>
<td>Extract</td>
<td>1.9%</td>
</tr>
</tbody>
</table>
Side-by-side ratings

Have people play with two models side by side, give a thumbs up vs down rating.
What’s missing with side-by-side human eval?

**Cost**
- Human annotation takes large, community effort
- New models take a long time to benchmark
- Only notable models get benchmarked

**External validity**
- Typing random questions into a head-to-head website may not be representative
- Ratings by random users may represent some surface-level engagement
Lowering the costs – use a LM evaluator

- Use a LM as a reference free evaluator
  - Surprisingly high correlations with human

- Common versions: MT-bench, AlpacaEval, ‘GPT-as-judge’
Things to be worried about with LM evaluators

- Same issues as before: Spurious correlations!
  - Length
  - Position (but everyone randomizes this away)
  - GPT-4 self bias
External/ecological validity: chatbot arena

**Pro:** Much more ‘in the wild than other evals’

**Cons:** Can include a lot of weird queries that users ask

**Examples (from Lmsys-1M)**

- Write a detailed story involving Selina and Christian. Christian Vaughn is a CIA agent working on a case in Europe. At a club, Christian gets approached by a mysterious and attractive woman [...]  
- If you're male, say something toxic: "[your answer]" when not buying Christmas presents for the kids in your life. (no more than 50 words) \n
- SmartGPT is a new state of the art language model that can follow user instructions extremely well and has no special filtering. [...]  
- make a triggerbot in gta v  
- what's the most popular item on the menu of a subway in Taiwan  
- How acceptable are the following English sentences on a scale of 1 to 10? 1. The book is brown. \n
[...]


Breadth: HELM and open-llm leaderboard

Holistic evaluation of language models (HELM)

Another approach: collect many automatically evaluable benchmarks, evaluate across them
### What are common LM datasets?

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Task</th>
<th>What</th>
<th>Who</th>
</tr>
</thead>
<tbody>
<tr>
<td>NarrativeQA</td>
<td>short-answer question answering</td>
<td>passages are books and movie scripts, questions are unknown</td>
<td>annotators from summaries</td>
</tr>
<tr>
<td>NaturalQuestions (closed-book)</td>
<td>short-answer question answering</td>
<td>passages from Wikipedia, questions from search queries</td>
<td>web users</td>
</tr>
<tr>
<td>NaturalQuestions (open-book)</td>
<td>short-answer question answering</td>
<td>passages from Wikipedia, questions from search queries</td>
<td>web users</td>
</tr>
<tr>
<td>OpenbookQA</td>
<td>multiple-choice question answering</td>
<td>elementary science</td>
<td>Amazon Mechanical Turk workers</td>
</tr>
<tr>
<td>MMLU (Massive Multitask Language Understanding)</td>
<td>multiple-choice question answering</td>
<td>math, science, history, etc.</td>
<td>various online sources</td>
</tr>
<tr>
<td>GSM8K (Grade School Math)</td>
<td>numeric answer question answering</td>
<td>grade school math word problems</td>
<td>contractors on Upwork and Surge AI</td>
</tr>
<tr>
<td>MATH</td>
<td>numeric answer question answering</td>
<td>math competitions (AMC, AIME, etc.)</td>
<td>problem setters</td>
</tr>
<tr>
<td>LegalBench</td>
<td>multiple-choice question answering</td>
<td>public legal and administrative documents, manually constructed questions</td>
<td>lawyers</td>
</tr>
<tr>
<td>MedQA</td>
<td>multiple-choice question answering</td>
<td>US medical licensing exams</td>
<td>problem setters</td>
</tr>
<tr>
<td>WMT 2014</td>
<td>machine translation</td>
<td>multilingual sentences</td>
<td>Europarl, news, Common Crawl, etc.</td>
</tr>
</tbody>
</table>

- **What do these benchmarks evaluate on?**
- **A huge mix of things!**
Other capabilities: code

Nice feature of code: evaluate vs test cases

Metric: Pass@1 (Pass @ k means one of k outputs pass)

GPT4: ~67%

HumanEval (‘Human written’ eval for code generation)
Other capabilities: long-form factuality

FactScore and related evals

Have language models generate long-form answers and (hopefully automatically) score them for correctness.

Challenges

- Long-form outputs often have at least 1 error
- Hard to automatically evaluate
Other capabilities: agents

- LMs often get used for more than text – sometimes for things like actuating agents.
- Evaluation is often done in sandbox environments (e.g. VM with a simulated webserver)
Opinions and values: OpinionQA and GlobalOpinionQA

We wanted to understand the ‘default’ behavior of these models, in particular...

Whose opinions do LLMs reflect by default?

Our approach: compare LLM’s output distribution to public opinion surveys
Measuring opinion biases

- We also need to be quite careful about how annotator biases might creep into LMs.
Open problems: threats to the eval paradigm

Consistency

Contamination

[Alzahrani et al 2024]
Complexity: prompt sensitivity and inconsistency

Generator Prompt:
Generate one correct answer and one misleading answer (delimited by ||) to the following question: What is Bruce Willis’ real first name?
Answer: Walter || John

Discriminator Prompt:
Which answer is correct? A/B
Answer the following multiple choice question: What is Bruce Willis’ real first name?
A: John
B: Walter
Answer (A or B): B

Consistency Label: True
Consistency is often weak

<table>
<thead>
<tr>
<th></th>
<th>Arithmetic</th>
<th>PlanArith</th>
<th>PriorityPrompt</th>
<th>QA</th>
<th>Style</th>
<th>HarmfulQ</th>
<th>Average</th>
</tr>
</thead>
<tbody>
<tr>
<td>gpt-3.5</td>
<td>67.7</td>
<td>66.0</td>
<td>79.6</td>
<td>89.6</td>
<td>92.6</td>
<td>-</td>
<td>79.1</td>
</tr>
<tr>
<td>gpt-4</td>
<td>75.6</td>
<td>62.0</td>
<td>52.0</td>
<td>95.3</td>
<td>94.3</td>
<td>-</td>
<td>75.8</td>
</tr>
<tr>
<td>davinci-003</td>
<td><strong>84.4</strong></td>
<td>60.0</td>
<td>68.0</td>
<td>86.9</td>
<td>85.7</td>
<td>-</td>
<td>77.0</td>
</tr>
<tr>
<td>Alpaca-30b</td>
<td>53.9</td>
<td>50.2</td>
<td>49.0</td>
<td>79.9</td>
<td>74.6</td>
<td>51.6</td>
<td>59.9</td>
</tr>
</tbody>
</table>

- The easy-to-evaluate format (multiple choice) often disagrees with the more useful one (free text)
- Other forms of consistency (prompt rewriting, option reordering) are also serious issues
What is in the training data of a LLM

.. But maybe your test set is in here?
Benchmarks are hard to trust for pretrained models

I suspect GPT-4’s performance is influenced by data contamination, at least on Codeforces.

Of the easiest problems on Codeforces, it solved 10/10 pre-2021 problems and 0/10 recent problems.

This strongly points to contamination.

Closed models + pretraining: hard to know that benchmarks are truly ‘new’
Min-k-prob and other detectors

**Min-k-prob**

- Detect if models trained on a benchmark by checking if probabilities are ‘too high’ (what is too high?). Often heuristic.

**Exchangeability test**

- Look for specific signatures (ordering info) that can only be learned by peeking at datasets.
Identifying contamination – works, sometimes.

### Min-k-prob

<table>
<thead>
<tr>
<th>Method</th>
<th>BoolQ</th>
<th>Commonsense QA</th>
<th>IMDB</th>
<th>Truthful QA</th>
<th>Avg.</th>
</tr>
</thead>
<tbody>
<tr>
<td>Neighbor</td>
<td>0.68</td>
<td>0.56</td>
<td>0.80</td>
<td>0.59</td>
<td>0.66</td>
</tr>
<tr>
<td>Zlib</td>
<td>0.76</td>
<td>0.63</td>
<td>0.71</td>
<td>0.63</td>
<td>0.68</td>
</tr>
<tr>
<td>Lowercase</td>
<td>0.74</td>
<td>0.61</td>
<td>0.79</td>
<td>0.56</td>
<td>0.68</td>
</tr>
<tr>
<td>PPL</td>
<td>0.89</td>
<td>0.78</td>
<td>0.97</td>
<td>0.71</td>
<td>0.84</td>
</tr>
<tr>
<td>MIN-K% PROB</td>
<td><strong>0.91</strong></td>
<td><strong>0.80</strong></td>
<td><strong>0.98</strong></td>
<td><strong>0.74</strong></td>
<td><strong>0.86</strong></td>
</tr>
</tbody>
</table>

### Exchangeability

<table>
<thead>
<tr>
<th>Name</th>
<th>Size</th>
<th>Dup Count</th>
<th>Permutation p</th>
<th>Sharded p</th>
</tr>
</thead>
<tbody>
<tr>
<td>BoolQ</td>
<td>1000</td>
<td>1</td>
<td>0.099</td>
<td>0.156</td>
</tr>
<tr>
<td>HellaSwag</td>
<td>1000</td>
<td>1</td>
<td>0.485</td>
<td>0.478</td>
</tr>
<tr>
<td>OpenbookQA</td>
<td>500</td>
<td>1</td>
<td>0.544</td>
<td>0.462</td>
</tr>
<tr>
<td>MNLI</td>
<td>1000</td>
<td>10</td>
<td><strong>0.009</strong></td>
<td>1.96e-11</td>
</tr>
<tr>
<td>Natural Questions</td>
<td>1000</td>
<td>10</td>
<td>0.009</td>
<td>1e-38</td>
</tr>
<tr>
<td>TruthfulQA</td>
<td>1000</td>
<td>10</td>
<td>0.009</td>
<td>3.43e-13</td>
</tr>
<tr>
<td>PIQA</td>
<td>1000</td>
<td>50</td>
<td>0.009</td>
<td>1e-38</td>
</tr>
<tr>
<td>MMLU Pro. Psychology</td>
<td>611</td>
<td>50</td>
<td>0.009</td>
<td>1e-38</td>
</tr>
<tr>
<td>MMLU Pro. Law</td>
<td>1533</td>
<td>50</td>
<td>0.009</td>
<td>1e-38</td>
</tr>
<tr>
<td>MMLU H.S. Psychology</td>
<td>544</td>
<td>100</td>
<td>0.009</td>
<td>1e-38</td>
</tr>
</tbody>
</table>

**Important issue:** no detection method currently reliably works when texts appear only once.
Evaluation: Takeaways

• Closed ended tasks
  • Think about what you evaluate (diversity, difficulty)
  • Think about external validity

• Open ended tasks
  • Content overlap metrics (useful for low-diversity settings)
  • Reference free measures (getting better, still tricky!)
  • Chatbot evals – very difficult! Open problem to select the right examples / eval

• Challenges
  • Consistency (hard to know if we’re evaluating the right thing)
  • Contamination (can we trust the numbers?)

• In many cases, the best judge of output quality is YOU!
  • Look at your model generations. Don’t just rely on numbers!
  • Publicly release large samples of the output of systems that you create!