

NLP Tasks, Data, and Evaluation



EECS 183/283a: Natural Language Processing

Today



- What tasks have NLP researchers historically cared about?
- How do we evaluate success on these tasks?
 - Dominant paradigm: automatic metrics computed on static benchmarks
- As models have improved their abilities, how have our evaluations changed?

Tasks and Evaluation Metrics

Classification

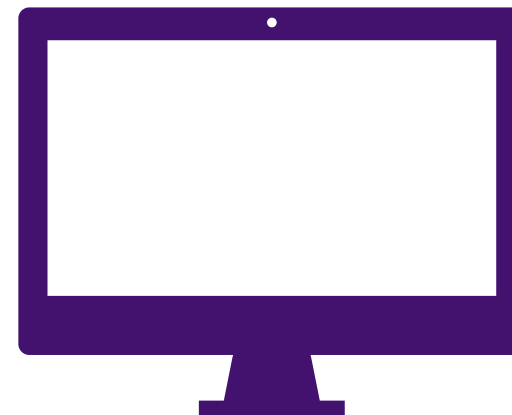
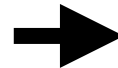
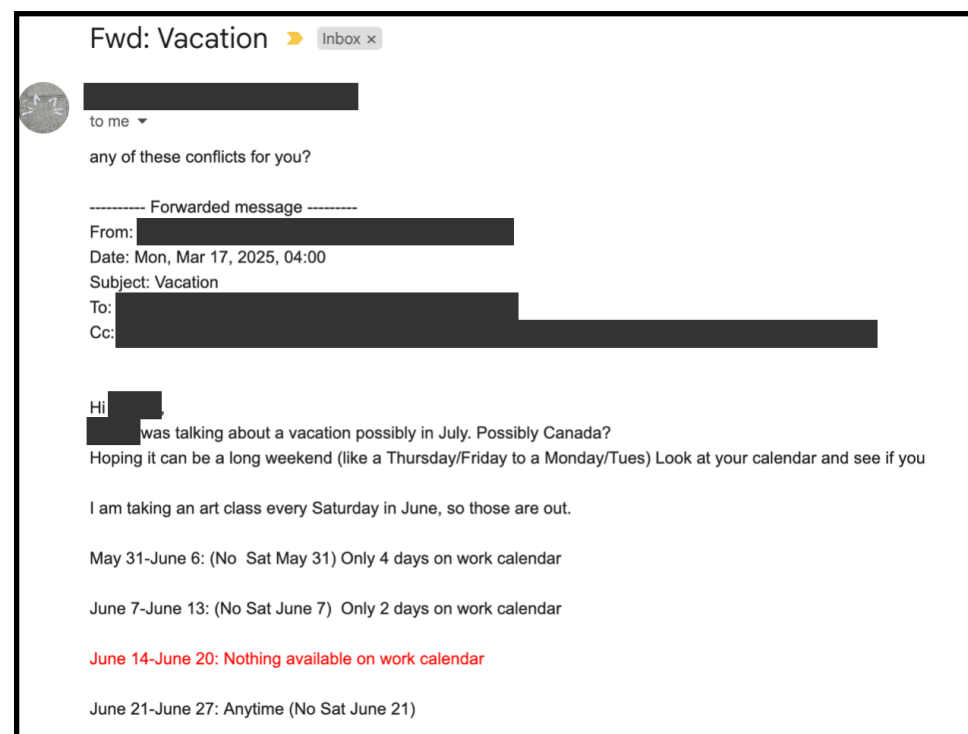


- **Input:** text data
- **Output:** label from a closed set

Binary Classification



- **Input:** text data
- **Output:** label from a closed set
- **Spam detection**

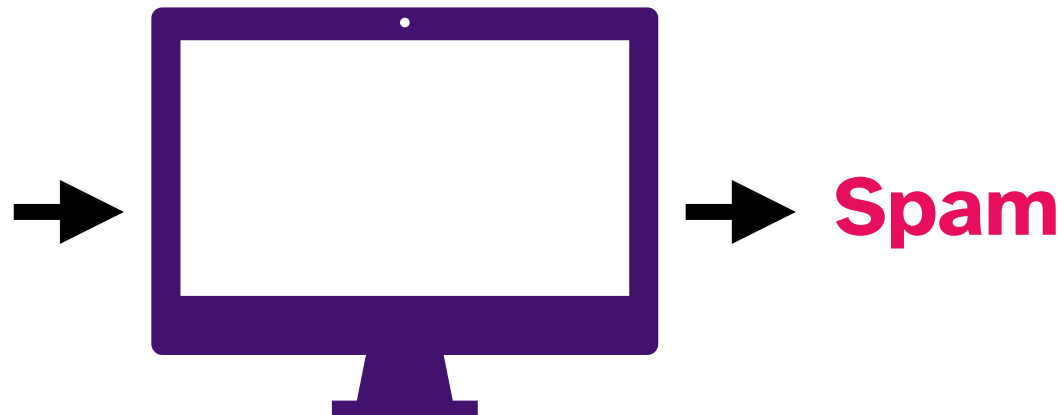
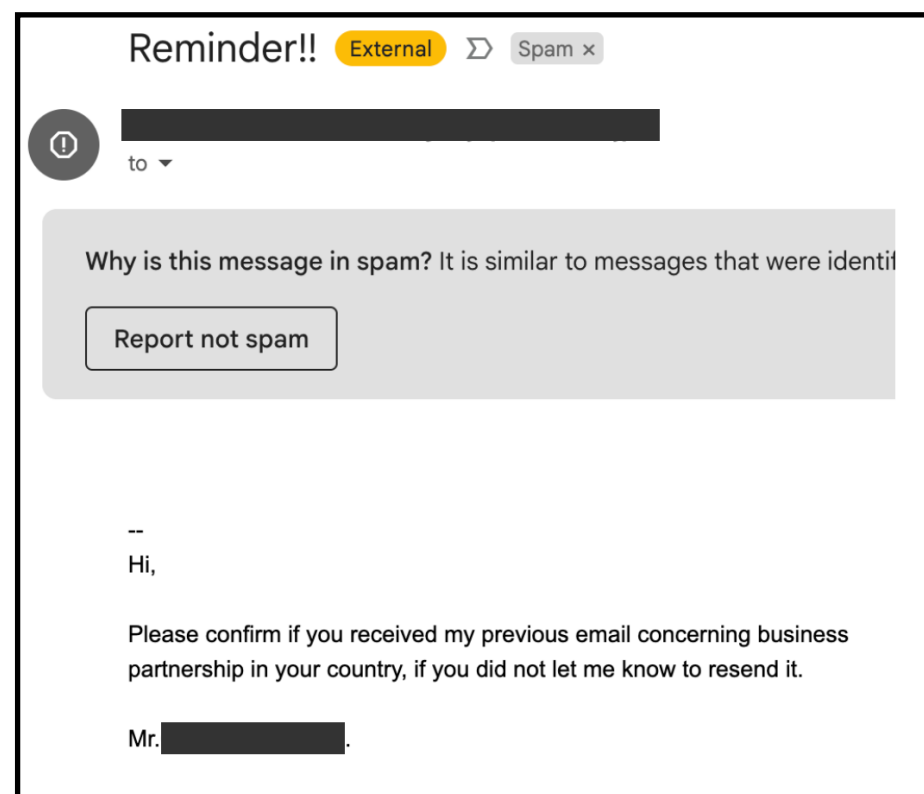


Not Spam

Binary Classification



- **Input:** text data
- **Output:** label from a closed set
- **Spam detection**



Multiclass Classification



- **Input:** text data
- **Output:** label from a closed set
- **Part of speech tagging**

Battle-tested industrial managers here always buck

JJ JJ NNS RB RB VB

up nervous newcomers with the tale of the first of

RP JJ NNS IN DT NN IN DT JJ IN

their countrymen to visit Mexico, a boatload of

PP NNS TO VB NNP DT NN IN

samurai warriors blown ashore 375 years ago.

NNS NNS VB RB CD NNS RB

Evaluation



Input
(31 words)

Battle-tested industrial ...

Battle-tested **industrial** ...

... industrial **managers** here ...

... managers **here** always ...

... here **always** buck up ...

...

... blown ashore 375 years **ago**.

**Correct
Label**★

JJ

JJ

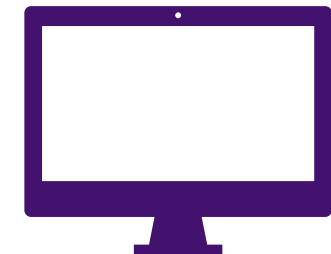
NNS

RB

RB

...

RB



VBD

JJ

NNS

RB

RB

...

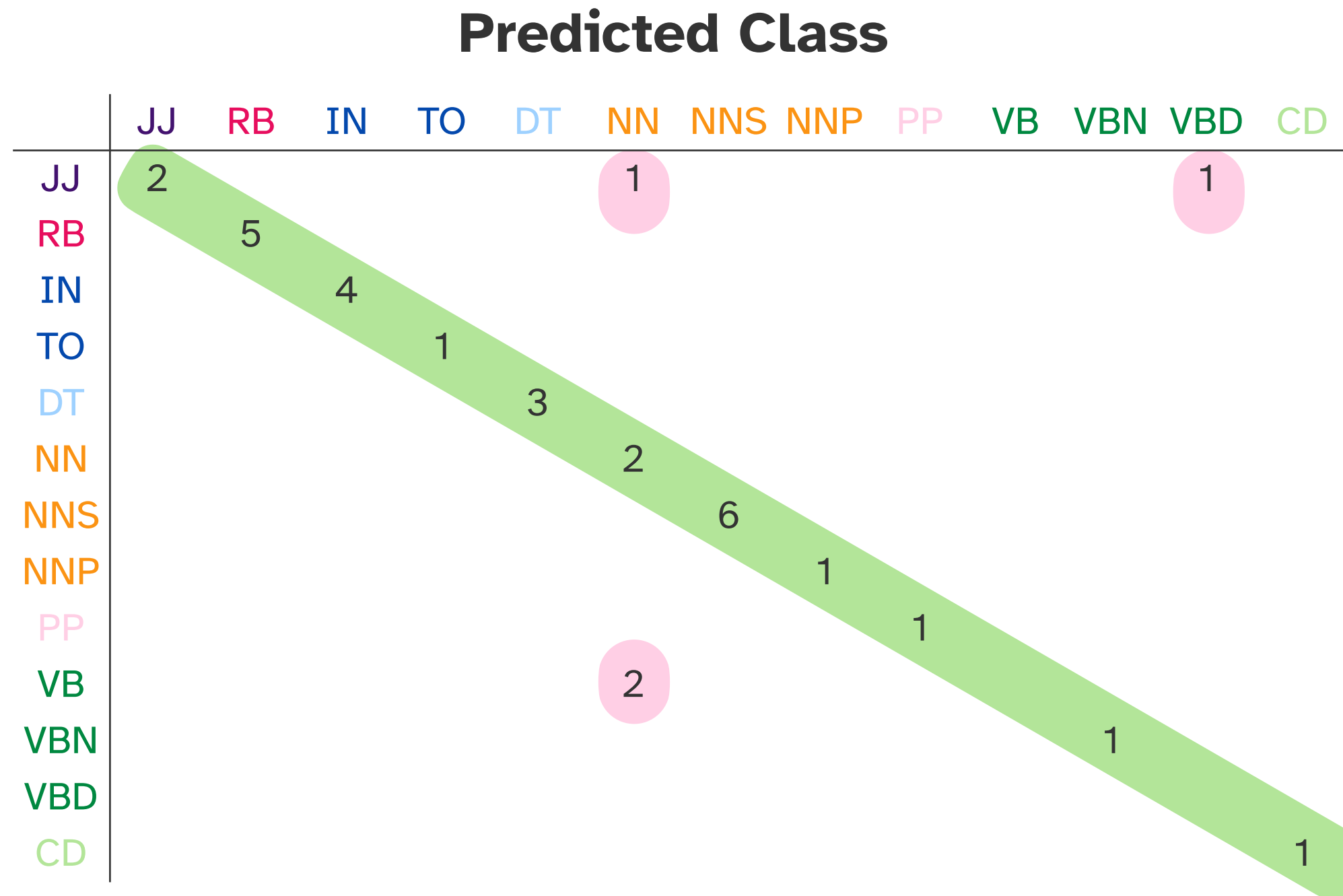
RB

Count examples for each pair of (target class, predicted class)

[illegible]

A stylized illustration of a building with a light blue body and a white triangular roof. The building features several arched windows and is outlined with thick black lines.

★ Target Class



Class: singular noun (**NN**)

2 true positive

$$\frac{TP}{TP + FP}$$

40%

★ Target Class

[illegible]

A stylized illustration of a building with a light blue body and a white triangular roof. The building features several arched windows and is outlined with thick black lines.

Class: singular noun (**NN**)

2 true positive

$$\frac{TP}{TP + FN}$$

★ Target Class

1

Class: singular noun (**NN**)

★ Target Class

[illegible]

Automatic Speech Recognition

Task: Convert a speech *utterance* into a text *transcript*

- ASR performance is measured using error rates
- Sentence error rate (exact match) can be too penalizing (1 or 0)
- Word Error Rate (WER) captures average number (%) of wrong words in the transcript
- For now, let's assume words are separated by spaces
 - For languages that are not space delimited, Character Error Rate (CER) can be used

Word Error Rate (WER)



Possible Errors : Substitutions (S), Deletions (D), Insertions (I)

$$\frac{S + D + I}{N}$$

- *Reference* : Take me to UC campus (# words N=5)
- *ASR Transcript* : Take me to **you** **see** campus
 - 1 substitution 1 insertion 0 deletions
 - $WER = 2/5 * 100 = 40 \%$
- WER can be greater than 100% (eg. Too many insertions)
- S + D + I needs to be computed using Edit (Levenshtein) distance

Levenshtein/Edit distance



Dynamic Programming based edit matrix computation $d[N, M]$

Transcript or Hypothesis (i)

	0.	1. Take	2. me	3. to	4. you	5. see	6. campus
0.							
1. Take		# Substitution	# Deletion				
2. me		# Insertion		$d[i,j]$			
3. to							
4. UC							
5. campus							

Calculate the least cost alignment/edit path allowing insertion, deletion and substitution of words

$$d[i, j] = \min \left(\begin{array}{l} d[i-1, j] + 1, \\ d[i, j-1] + 1, \\ d[i-1, j-1] + \text{local_substitution}(i, j) \end{array} \right)$$

Insertion
Deletion
Substitution

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d [N, M]$

Hypothesis (i)

	0.	1. Take	2. me	3. to	4. you	5. see	6. campus
0.	0						
1. Take							
2. me							
3. to							
4. UC							
5. campus							

Initialize 0 cost in extra row/column

Reference (j)

$$d[i, j] = \min \left(\begin{array}{l} d[i-1, j] + 1, \\ d[i, j-1] + 1, \\ d[i-1, j-1] + \text{local_substitution}(i, j) \end{array} \right)$$

Insertion
Deletion
Substitution

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d [N, M]$

Hypothesis (i)

	0.	1. Take	2. me	3. to	4. you	5. see	6. campus
0.	0	$d[i-1, j] + 1$					
1. Take	$d[i, j-1] + 1$						
2. me							
3. to							
4. UC							
5. campus							

“Grow” the edit matrix iteratively, by accumulating the cost for each element $d [i,j]$

$d [i, j] = \min (d [i-1, j] + 1 ,$

$d [i , j -1] + 1 ,$

$d [i -1 , j -1] + \text{local_substitution} (i , j))$

Insertion

Deletion

Substitution

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d [N, M]$

Hypothesis (i)

	0.	1. Take	2. me	3. to	4. you	5. see	6. campus
0.	0	1					
1. Take	1	$\min(1, 1, 0)$					
2. me							
3. to							
4. UC							
5. campus							

“Grow” the edit matrix iteratively, by accumulating the cost for each element $d [i,j]$

$d [i, j] = \min (d [i-1, j] + 1 ,$

$d [i , j -1] + 1 ,$

$d [i -1 , j -1] + \text{local_substitution} (i , j))$

Insertion

Deletion

Substitution

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d [N, M]$

Hypothesis (i)

	0.	1. Take	2. me	3. to	4. you	5. see	6. campus
0.	0	1					
1. Take	1	0					
2. me							
3. to							
4. UC							
5. campus							

“Grow” the edit matrix iteratively, by accumulating the cost for each element $d [i,j]$

$d [i, j] = \min (d [i-1, j] + 1 ,$

$d [i , j -1] + 1 ,$

$d [i -1 , j -1] + \text{local_substitution} (i , j))$

Insertion

Deletion

Substitution

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d [N, M]$

Hypothesis (i)

	0.	1. Take	2. me	3. to	4. you	5. see	6. campus
0.	0	1	2				
1. Take	1	0					
2. me							
3. to							
4. UC							
5. campus							

“Grow” the edit matrix iteratively, by accumulating the cost for each element $d [i,j]$

$d [i, j] = \min (d [i-1, j] + 1 ,$

$d [i , j -1] + 1 ,$

$d [i -1 , j -1] + \text{local_substitution} (i , j))$

Insertion

Deletion

Substitution

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d [N, M]$

Hypothesis (i)

	0.	1. Take	2. me	3. to	4. you	5. see	6. campus
0.	0	1	2				
1. Take	1	0	1				
2. me	2	1	0				
3. to							
4. UC							
5. campus							

“Grow” the edit matrix iteratively, by accumulating the cost for each element $d [i,j]$

$d [i, j] = \min (d [i-1, j] + 1 ,$

$d [i , j -1] + 1 ,$

$d [i -1 , j -1] + \text{local_substitution} (i , j))$

Insertion

Deletion

Substitution

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d [N, M]$

Hypothesis (i)

Reference (j)		0.	1. Take	2. me	3. to	4. you	5. see	6. campus
	0.	0	1	2	3			
	1. Take	1	0	1	2			
	2. me	2	1	0	1			
	3. to	3	2	1	0			
	4. UC							
	5. campus							

“Grow” the edit matrix iteratively, by accumulating the cost for each element $d [i,j]$

$$d [i, j] = \min (\begin{array}{l} d [i-1, j] + 1, \\ d [i, j -1] + 1, \\ d [i -1, j -1] + \text{local_substitution} (i, j) \end{array})$$

Insertion
Deletion
Substitution

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d [N, M]$

Hypothesis (i)

Reference (j)		0.	1. Take	2. me	3. to	4. you	5. see	6. campus
	0.	0	1	2	3	4		
	1. Take	1	0	1	2	3		
	2. me	2	1	0	1	2		
	3. to	3	2	1	0	1		
	4. UC	4	3	1	1	1		
	5. campus							

“Grow” the edit matrix iteratively, by accumulating the cost for each element $d [i,j]$

Substitution!

$$d [i, j] = \min (\begin{array}{l} d [i-1, j] + 1, \\ d [i, j -1] + 1, \\ d [i -1, j -1] + \text{local_substitution} (i, j) \end{array})$$

Insertion
Deletion
Substitution

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d [N, M]$

Hypothesis (i)

Reference (j)		0.	1. Take	2. me	3. to	4. you	5. see	6. campus
	0.	0	1	2	3	4	5	
	1. Take	1	0	1	2	3	4	
	2. me	2	1	0	1	2	3	
	3. to	3	2	1	0	1	2	
	4. UC	4	3	1	1	1	2	
	5. campus	5	4	2	2	2	2	

“Grow” the edit matrix iteratively, by accumulating the cost for each element $d [i,j]$

$$d [i, j] = \min (\begin{array}{l} d [i-1, j] + 1, \\ d [i, j -1] + 1, \\ d [i -1, j -1] + \text{local_substitution} (i, j) \end{array})$$

Insertion
Deletion
Substitution

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d[N, M]$

Hypothesis (i)

Reference (j)		0.	1. Take	2. me	3. to	4. you	5. see	6. campus
	0.	0	1	2	3	4	5	6
	1. Take	1	0	1	2	3	4	5
	2. me	2	1	0	1	2	3	4
	3. to	3	2	1	0	1	2	3
	4. UC	4	3	1	1	1	2	2
	5. campus	5	4	2	2	2	2	2

Edit cost
 $d[N, M] = 2$

$$\text{WER} = 2/5 = 40\%$$

Levenshtein/Edit distance



Dynamic Programming based edit matrix $d[N, M]$

Hypothesis (i)

	0.	1. Take	2. me	3. to	4. you	5. see	6. campus
0.	0	1	2	3	4	5	6
1. Take	1	0	1	2	3	4	5
2. me	2	1	0	1	2	3	4
3. to	3	2	1	0	1	2	3
4. UC	4	3	1	1	1	2	2
5. campus	5	4	2	2	2	2	2

Optimal path:
tells us what
the errors were

substitution

insertion

Reference (j)

Structured Prediction



- **Input:** text data
- **Output:** structured combination of labels from a closed set
- **Syntactic** (constituency) parsing

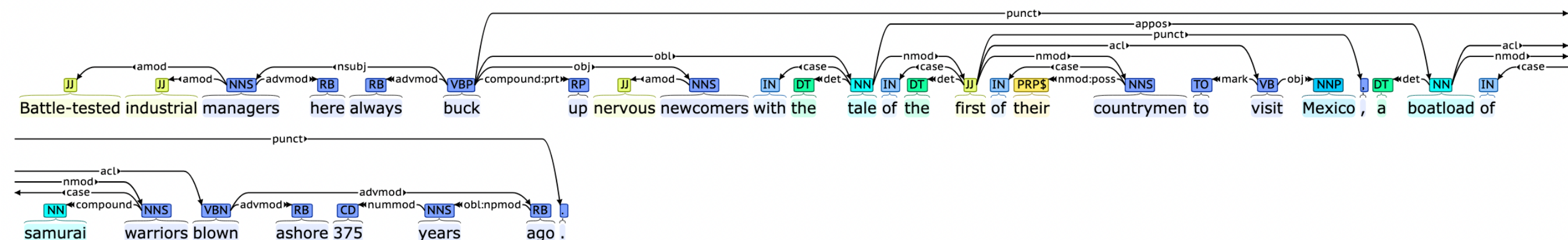
```
( (S
  (NP Battle-tested industrial managers
    here)
  always
  (VP buck
    up
    (NP nervous newcomers)
    (PP with
      (NP the tale
        (PP of
          (NP (NP the
            (ADJP first
              (PP of
                (NP their countrymen))))
            (S (NP *)
              to
              (VP visit
                (NP Mexico))))
            ,
            (NP (NP a boatload
              (PP of
                (NP (NP warriors)
                  (VP-1 blown
                    ashore
                    (ADVP (NP 375 years)
                      ago))))))
            (VP-1 *pseudo-attach*))))))
  .)
```

Penn Treebank, Marcus et al. 1993

Structured Prediction



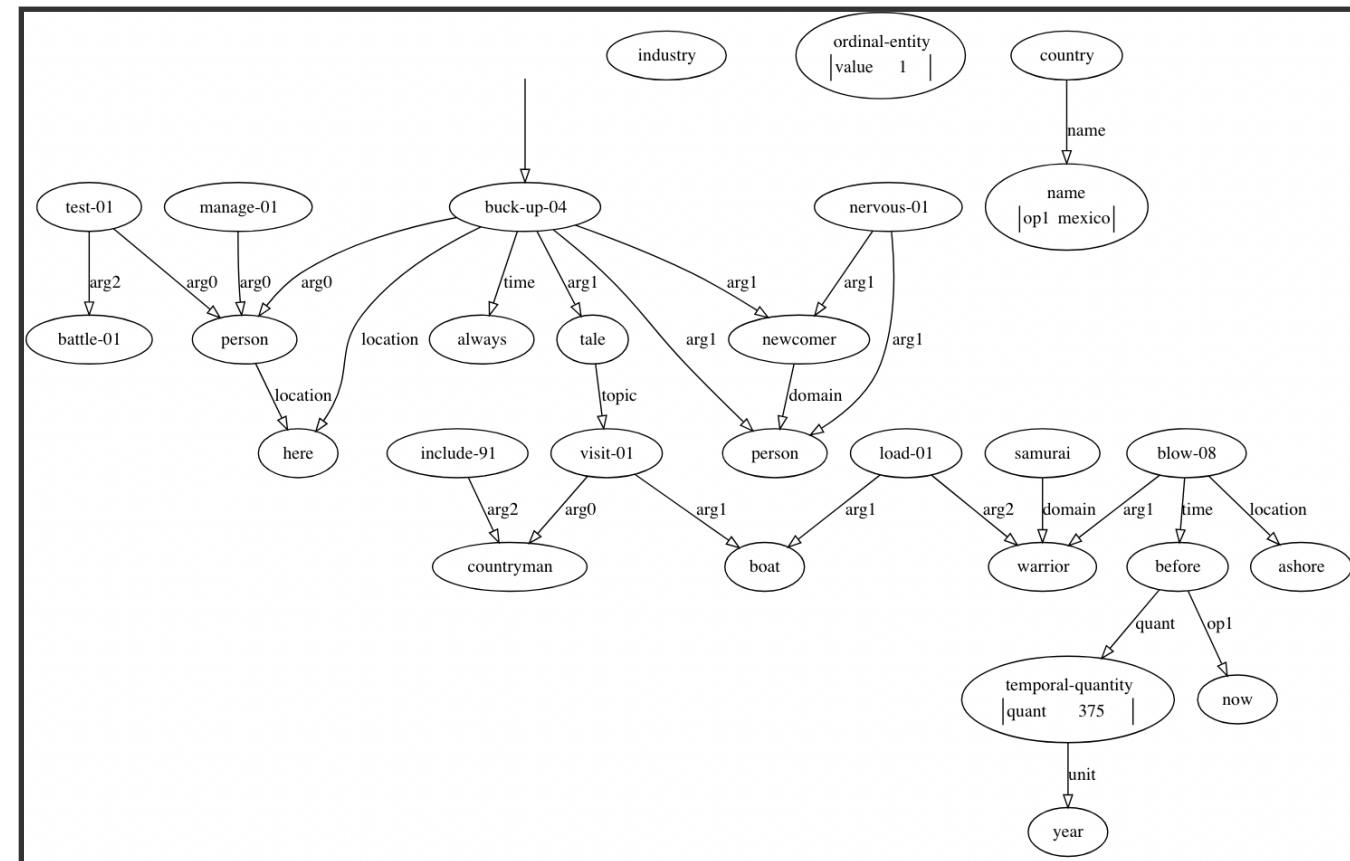
- **Input:** text data
- **Output:** structured combination of labels from a closed set
- **Syntactic** (dependency) parsing



Structured Prediction



- **Input:** text data
- **Output:** structured combination of labels from a closed set
- **Semantic parsing** (abstract meaning representation)



Structured Prediction



- **Input:** text data
- **Output:** structured combination of labels from a closed set
- **Semantic parsing** (text-to-SQL)

Task-specific eval metrics:

- exact match accuracy
- execution accuracy



airline	#
AA	123
Delta	456
...	...
JetBlue	404

?
=

airline	#
AA	123
Delta	456
...	...
United	789

show me flights from seattle to boston next monday

```
(SELECT DISTINCT flight.flight_id FROM flight WHERE (flight.from_airport IN (SELECT airport.service.airport_code FROM airport.service WHERE airport.service.city_code IN (SELECT city.city_code FROM city WHERE city.city_name = 'SEATTLE')))) AND (flight.to_airport IN (SELECT airport.service.airport_code FROM airport.service WHERE airport.service.city_code IN (SELECT city.city_code FROM city WHERE city.city_name = 'BOSTON')))) AND (flight.flight_days IN (SELECT days.days_code FROM days WHERE days.day_name IN (SELECT date_day.day_name FROM date_day WHERE date_day.year = 1993 AND date_day.month_number = 2 AND date_day.day_number = 8))));
```

Structured Prediction



- **Input:** text data
- **Output:** structured combination of labels from a closed set
- **Semantic parsing** (code generation)

Task-specific eval metrics:

- exact match accuracy
- test case pass rate

```
def incr_list(l: list):  
    """Return list with elements incremented by 1.  
    >>> incr_list([1, 2, 3])  
    [2, 3, 4]  
    >>> incr_list([5, 3, 5, 2, 3, 3, 9, 0, 123])  
    [6, 4, 6, 3, 4, 4, 10, 1, 124]  
    """  
    return [i + 1 for i in l]
```

```
def solution(lst):  
    """Given a non-empty list of integers, return the sum of all of the odd elements  
    that are in even positions.  
  
    Examples  
    solution([5, 8, 7, 1]) ==>12  
    solution([3, 3, 3, 3, 3]) ==>9  
    solution([30, 13, 24, 321]) ==>0  
    """  
    return sum(lst[i] for i in range(0, len(lst)) if i % 2 == 0 and lst[i] % 2 == 1)
```

```
def encode_cyclic(s: str):  
    """  
    returns encoded string by cycling groups of three characters.  
    """  
    # split string to groups. Each of length 3.  
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]  
    # cycle elements in each group. Unless group has fewer elements than 3.  
    groups = [(group[1:] + group[0]) if len(group) == 3 else group for group in groups]  
    return "".join(groups)  
  
def decode_cyclic(s: str):  
    """  
    takes as input string encoded with encode_cyclic function. Returns decoded string.  
    """  
    # split string to groups. Each of length 3.  
    groups = [s[(3 * i):min((3 * i + 3), len(s))] for i in range((len(s) + 2) // 3)]  
    # cycle elements in each group.  
    groups = [(group[-1] + group[:-1]) if len(group) == 3 else group for group in groups]  
    return "".join(groups)
```

HumanEval, Chen et al. 2021

Text Comprehension



- **Input:** text data
- **Output:** short piece of text (*extractive* QA: span in input text)

Task-specific eval metrics:
- span selection accuracy

In meteorology, precipitation is any product of the condensation of atmospheric water vapor that falls under **gravity**. The main forms of precipitation include drizzle, rain, sleet, snow, **graupel** and hail... Precipitation forms as smaller droplets coalesce via collision with other rain drops or ice crystals **within a cloud**. Short, intense periods of rain in scattered locations are called “showers”.

What causes precipitation to fall?

gravity

What is another main form of precipitation besides drizzle, rain, snow, sleet and hail?

graupel

Where do water droplets collide with ice crystals to form precipitation?

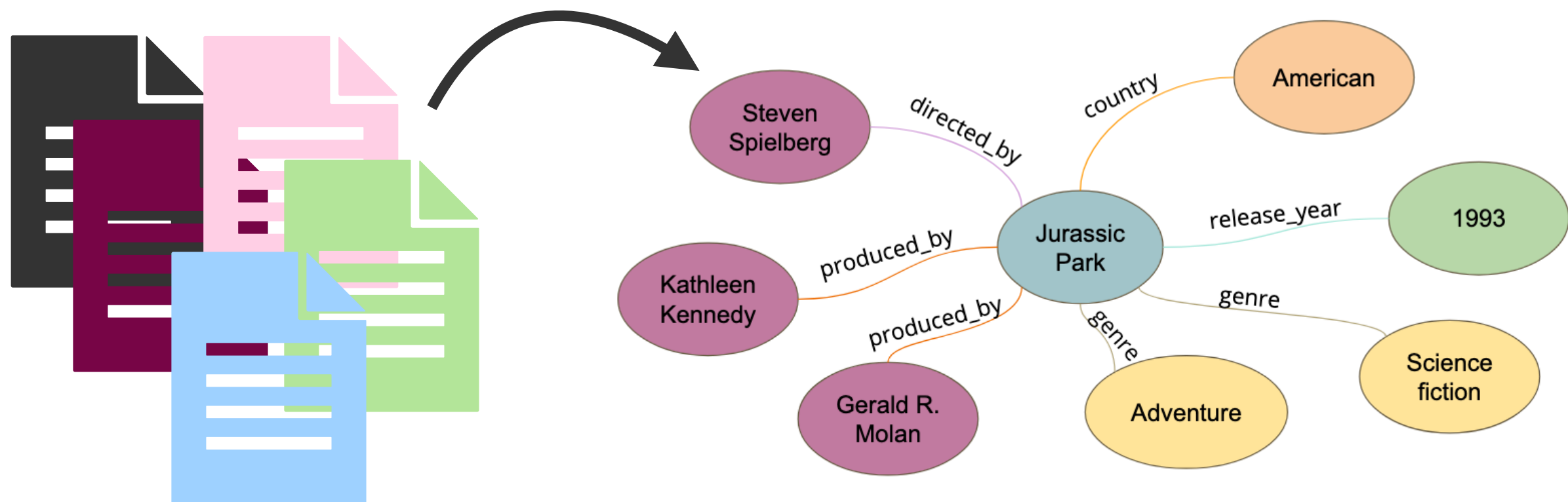
within a cloud

Rajpurkar et al. 2016

Information Extraction



- **Input:** lots of text data
- **Output:** structured representation of information



Task-specific eval metrics:

- precision of extracted facts
- recall of known facts

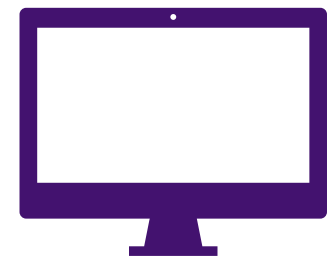
Question Answering



- **Input:** none
- **Output:** short piece of text
- Factual and commonsense

Task-specific eval metrics:
- multiple-choice accuracy

Options



0.2%

0.1% ✗

0.3%

0.3%

If someone wants to submerge themselves in water, what should they use?

- (a) coffee cup
- ★ (b) whirlpool bath
- (c) cup
- (d) puddle

Question Answering



- **Input:** none
- **Output:** short piece of text
- Factual and commonsense

Task-specific eval metrics:

- multiple-choice accuracy
- string similarity?

**most likely
answer**

If someone wants to
submerge themselves
in water, what should
they use?



bathtub ✓
≈ whirlpool bath

Reference-Based Text Generation



- **Input:** text data
- **Output:** same text, but in a different language
- **Machine translation**

The bottle floated into the cave



La botella entró a la cueva flotando

Reference-Based Evaluation: BLEU



- Reference y
The most natural form of language use is dialogue.
- System outputs / candidates $\hat{y}$
 $\hat{y}_1$ *The most common form of language use is conversation.*
 $\hat{y}_2$ *The most common form of language is speech.*

Reference-Based Evaluation: BLEU



- Reference y
The most natural form of language use is dialogue.
- System outputs / candidates $\hat{y}$
 $\hat{y}_1$ *The most common form of language use is conversation.*
 $\hat{y}_2$ *The most common form of language is speech.*
- $G_n(x)$ is the set of n -grams in sequence x

Reference-Based Evaluation: BLEU



- Reference y

The most natural form of language use is dialogue.
unigram

- System outputs / candidates $\hat{y}$

$\hat{y}_1$ *The most common form of language use is conversation.*

$\hat{y}_2$ *The most common form of language is speech.*

- $G_n(x)$ is the set of n -grams in sequence x

$G_1(y) = \{the, most, natural, form, of, language, use, is, dialogue, .\}$

Reference-Based Evaluation: BLEU



- Reference y

The most natural form of language use is dialogue.

bigram

- System outputs / candidates $\hat{y}$

$\hat{y}_1$ *The most common form of language use is conversation.*

$\hat{y}_2$ *The most common form of language is speech.*

- $G_n(x)$ is the set of n -grams in sequence x

$G_1(y) = \{the, most, natural, form, of, language, use, is, dialogue, .\}$

$G_2(y) = \{the\ most, most\ natural, natural\ form, form\ of, \dots\}$

Reference-Based Evaluation: BLEU



- Reference y
The most natural form of language use is dialogue.
- System outputs / candidates $\hat{y}$
 $\hat{y}_1$ *The most common form of language use is conversation.*
 $\hat{y}_2$ *The most common form of language is speech.*
- $G_n(x)$ is the set of n -grams in sequence x
 $G_1(y) = \{the, most, natural, form, of, language, use, is, dialogue, .\}$
 $G_2(y) = \{the\ most, most\ natural, natural\ form, form\ of, ...\}$
- $C(s, x)$ is the number of occurrences of n -gram s in x
 $C(the, y) = 1$ $C(most, y) = 1$ $C(natural\ form, y) = 1$

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

$\hat{y}_1$ *The most common form of language use is conversation.*

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₁ *The most common form of language use is conversation.*

$G_1(\hat{y}_1) = \{the, most, common, form, of, language, use, is, conversation, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>			
<i>most</i>			
<i>common</i>			
<i>form</i>			
<i>of</i>			
<i>language</i>			
<i>use</i>			
<i>is</i>			
<i>conversation</i>			
<i>.</i>			

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₁ *The most common form of language use is conversation.*

$G_1(\hat{y}_1) = \{the, most, common, form, of, language, use, is, conversation, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>	1		
<i>most</i>	1		
<i>common</i>	1		
<i>form</i>	1		
<i>of</i>	1		
<i>language</i>	1		
<i>use</i>	1		
<i>is</i>	1		
<i>conversation</i>	1		
<i>.</i>	1		

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₁ *The most common form of language use is conversation.*

$G_1(\hat{y}_1) = \{the, most, common, form, of, language, use, is, conversation, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>	1	1	
<i>most</i>	1	1	
<i>common</i>	1	0	
<i>form</i>	1	1	
<i>of</i>	1	1	
<i>language</i>	1	1	
<i>use</i>	1	1	
<i>is</i>	1	1	
<i>conversation</i>	1	0	
<i>.</i>	1	1	

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₁ *The most common form of language use is conversation.*

$G_1(\hat{y}_1) = \{the, most, common, form, of, language, use, is, conversation, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>	1	1	1
<i>most</i>	1	1	1
<i>common</i>	1	0	0
<i>form</i>	1	1	1
<i>of</i>	1	1	1
<i>language</i>	1	1	1
<i>use</i>	1	1	1
<i>is</i>	1	1	1
<i>conversation</i>	1	0	0
<i>.</i>	1	1	1

Reference-Based Evaluation: BLEU



n-gram precision

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₁ *The most common form of language use is conversation.*

$G_1(\hat{y}_1) = \{the, most, common, form, of, language, use, is, conversation, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>	1	1	1
<i>most</i>	1	1	1
<i>common</i>	1	0	0
<i>form</i>	1	1	1
<i>of</i>	1	1	1
<i>language</i>	1	1	1
<i>use</i>	1	1	1
<i>is</i>	1	1	1
<i>conversation</i>	1	0	0
<i>.</i>	1	1	1
Sum			8

Reference-Based Evaluation: BLEU



n-gram precision

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₁ *The most common form of language use is conversation.*

$G_1(\hat{y}_1) = \{the, most, common, form, of, language, use, is, conversation, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>	1	1	1
<i>most</i>	1	1	1
<i>common</i>	1	0	0
<i>form</i>	1	1	1
<i>of</i>	1	1	1
<i>language</i>	1	1	1
<i>use</i>	1	1	1
<i>is</i>	1	1	1
<i>conversation</i>	1	0	0
<i>.</i>	1	1	1
Sum	10		8

$$P_1(\hat{y}_1, y) = 8 / 10 = 0.8$$

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₂ *The most common form of language is speech.*

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₂ *The most common form of language is speech.*

$G_1(\hat{y}_2) = \{the, most, common, form, of, language, is, speech, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>			
<i>most</i>			
<i>common</i>			
<i>form</i>			
<i>of</i>			
<i>language</i>			
<i>is</i>			
<i>speech</i>			
<i>.</i>			

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₂ *The most common form of language is speech.*

$G_1(\hat{y}_2) = \{the, most, common, form, of, language, is, speech, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>	1		
<i>most</i>	1		
<i>common</i>	1		
<i>form</i>	1		
<i>of</i>	1		
<i>language</i>	1		
<i>is</i>	1		
<i>speech</i>	1		
<i>.</i>	1		

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₂ *The most common form of language is speech.*

$G_1(\hat{y}_2) = \{the, most, common, form, of, language, is, speech, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>	1	1	
<i>most</i>	1	1	
<i>common</i>	1	0	
<i>form</i>	1	1	
<i>of</i>	1	1	
<i>language</i>	1	1	
<i>is</i>	1	1	
<i>speech</i>	1	0	
<i>.</i>	1	1	

Reference-Based Evaluation: BLEU



n-gram precision

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₂ *The most common form of language is speech.*

$G_1(\hat{y}_2) = \{the, most, common, form, of, language, is, speech, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>	1	1	1
<i>most</i>	1	1	1
<i>common</i>	1	0	0
<i>form</i>	1	1	1
<i>of</i>	1	1	1
<i>language</i>	1	1	1
<i>is</i>	1	1	1
<i>speech</i>	1	0	0
<i>.</i>	1	1	1

Reference-Based Evaluation: BLEU



n-gram precision

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₂ *The most common form of language is speech.*

$G_1(\hat{y}_2) = \{the, most, common, form, of, language, is, speech, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>	1	1	1
<i>most</i>	1	1	1
<i>common</i>	1	0	0
<i>form</i>	1	1	1
<i>of</i>	1	1	1
<i>language</i>	1	1	1
<i>is</i>	1	1	1
<i>speech</i>	1	0	0
<i>.</i>	1	1	1
Sum			7

Reference-Based Evaluation: BLEU



n-gram precision

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*ŷ*₂ *The most common form of language is speech.*

$G_1(\hat{y}_2) = \{the, most, common, form, of, language, is, speech, .\}$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>the</i>	1	1	1
<i>most</i>	1	1	1
<i>common</i>	1	0	0
<i>form</i>	1	1	1
<i>of</i>	1	1	1
<i>language</i>	1	1	1
<i>is</i>	1	1	1
<i>speech</i>	1	0	0
<i>.</i>	1	1	1
Sum	9		7

$$P_1(\hat{y}_1, y) = 8 / 10 = 0.8$$

$$P_1(\hat{y}_2, y) = 7 / 9 = 0.77$$

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*y*₃ *language*

$$G_1(\hat{y}_3) = \{\text{language}\}$$

<i>s</i>	$C(s, \hat{y})$	$C(s, y)$	$\min(C(s, \hat{y}), C(s, y))$
<i>language</i>	1	1	1
Sum	1		1

$$P_1(\hat{y}_3, y) = 1 / 1 = 1.0 \text{ ?}$$

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*y*₃ *language*

$$\text{BP}(\hat{y}, y) = \begin{cases} 1 & |\hat{y}| \geq |y| \\ e^{(1-(|y|/|\hat{y}|))} & |\hat{y}| < |y| \end{cases}$$

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*y*₃ *language*

$$BP(\hat{y}, y) = \begin{cases} 1 & |\hat{y}| \geq |y| \\ e^{(1-(|y|/|\hat{y}|))} & |\hat{y}| < |y| \end{cases}$$

$$|y| = 10$$

	Text	$P_1(\hat{y}, y)$	$ \hat{y} $	$BP(\hat{y}, y)$
$\hat{y}_1$	<i>the most common</i>	0.8	10	1.00
$\hat{y}_2$	<i>the most common</i>	0.77	9	0.895
$\hat{y}_3$	<i>language</i>	1.00	1	0.0001

Reference-Based Evaluation: BLEU



***n*-gram precision**

How many of the *n*-grams actually occur in the reference?

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

y *The most natural form of language use is dialogue.*

*y*₃ *language*

Brevity penalty

$$BP(\hat{y}, y) = \begin{cases} 1 & |\hat{y}| \geq |y| \\ e^{(1 - (|y|/|\hat{y}|))} & |\hat{y}| < |y| \end{cases}$$

$|y| = 11$

	Text	$P_1(\hat{y}, y)$	$ \hat{y} $	$BP(\hat{y}, y)$
$\hat{y}_1$	<i>the most common</i>	0.8	10	1.00
$\hat{y}_2$	<i>the most common</i>	0.77	9	0.895
$\hat{y}_3$	<i>language</i>	1.00	1	0.0001

$$\text{BLEU}_{n=1} = BP \cdot P_1(\hat{y}, y)$$

Reference-Based Evaluation: BLEU



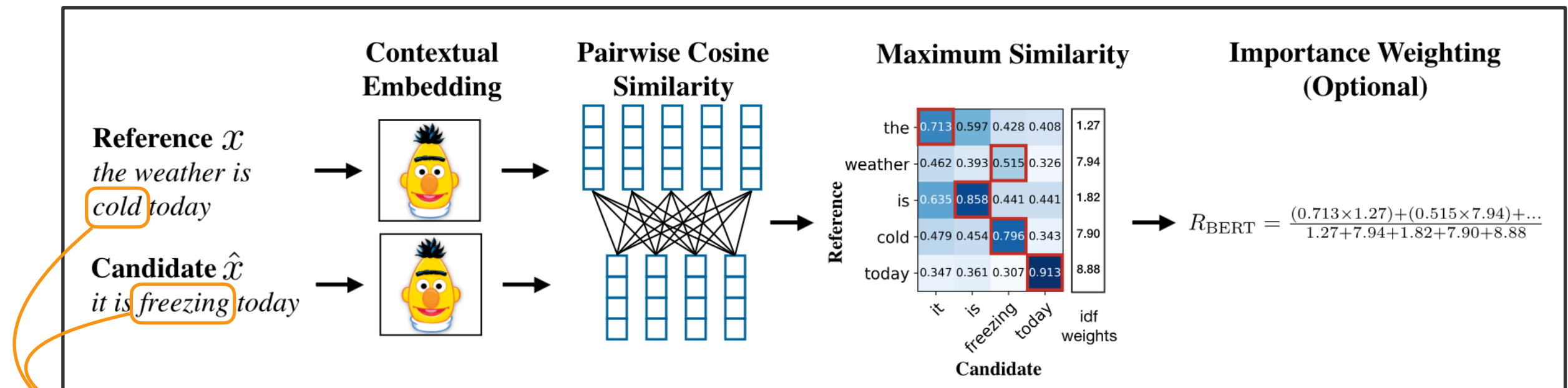
In practice: Average n -gram precision, for up to $N = 4$

$$P_n(\hat{y}, y) = \frac{\sum_{s \in G_n(\hat{y})} \min(C(s, \hat{y}), C(s, y))}{\sum_{s \in G_n(\hat{y})} C(s, \hat{y})}$$

$$\text{BP}(\hat{y}, y) = \begin{cases} 1 & |\hat{y}| \geq |y| \\ e^{(1-(|y|/|\hat{y}|))} & |\hat{y}| < |y| \end{cases}$$

$$\text{BLEU} = \text{BP}(\hat{y}, y) \exp \left(\sum_{n=1}^N \frac{1}{N} \ln P_n(\hat{y}, y) \right)$$

Reference-Based Evaluation: Representation Similarity



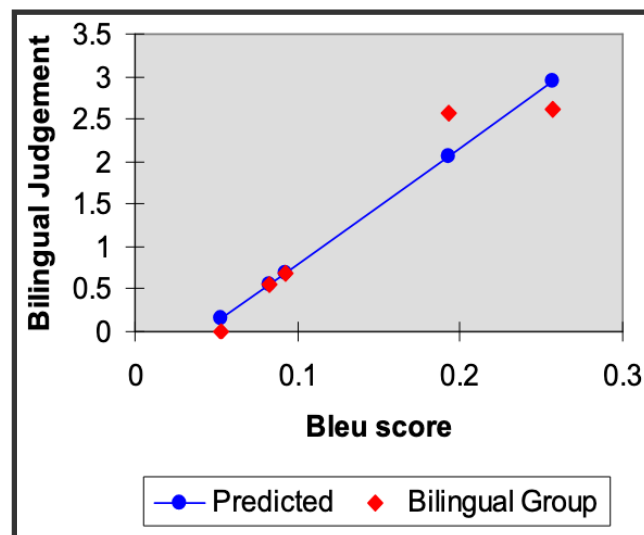
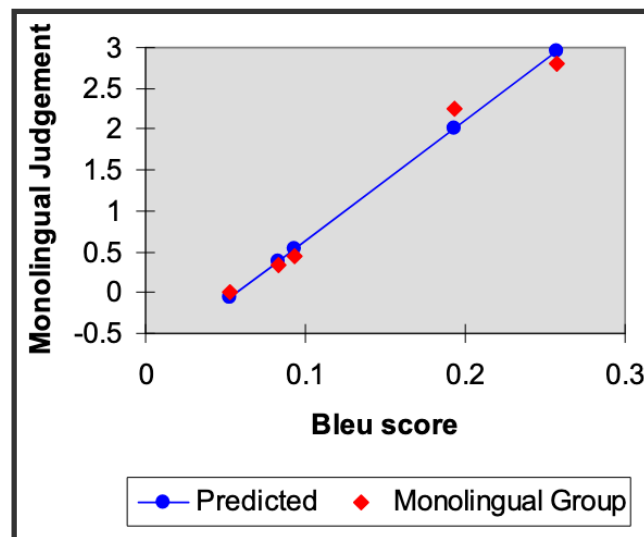
No n -gram overlap,
but should still get
some credit!

BERTScore, Zhang et al. 2020

Correlation with Human Judgment



Does our automatic metric reflect human judgments?



Papineni et al. 2002

Metric	en↔cs (5/5)	en↔de (16/16)	en↔et (14/14)	en↔fi (9/12)	en↔ru (8/9)	en↔tr (5/8)	en↔zh (14/14)
BLEU	.970/.995	.971/.981	.986/.975	.973/.962	.979/.983	.657/.826	.978/.947
ITER	.975/.915	.990/.984	.975/.981	.996/.973	.937/.975	.861/.865	.980/ –
RUSE	.981/ –	.997/ –	.990/ –	.991/ –	.988/ –	.853/ –	.981/ –
YiSi-1	.950/.987	.992/.985	.979/.979	.973/.940	.991/.992	.958/.976	.951/.963
P_{BERT}	.980/.994	.998/.988	.990/.981	.995/.957	.982/.990	.791/.935	.981/.954
R_{BERT}	.998/.997	.997/.990	.986/.980	.997/.980	.995/.989	.054/.879	.990/.976
F_{BERT}	.990/.997	.999/.989	.990/.982	.998/.972	.990/.990	.499/.908	.988/.967
$F_{\text{BERT}}(\text{idf})$	.985/.995	.999/.990	.992/.981	.992/.972	.991/.991	.826/.941	.989/.973

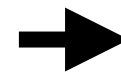
Zhang et al. 2020

Reference-Free Text Generation



- **Input:** maybe a natural language prompt, but nothing very complex
- **Output:** long piece of text

PREMISE: After years of being content with only a small inheritance, Sarah's life is turned upside down when she discovers she is actually a millionaire. She must learn to navigate her new life of luxury while also dealing with the guilt she feels over her new-found fortune.



GENERATED STORY:

Sarah Jensen never knew she had inherited a fortune until her mother died. In fact, she had never met her father ...[15 words]...
The only thing Sarah did know was that he had died long ago and that her mother had been left a sizeable inheritance. ...[23 words]...
Sarah's mother never spent any of the money ... all that mattered to her was how much it would mean to her daughter when she inherited it ...[81 words]...
Sarah learned that she was a millionaire. ...[417 words]...
She would never have had a chance like this - an opportunity to be financially independent, to live on her own ...[544 words]...
Now that she was rich, things were going to be different for her. She could afford a hotel suite and a lavish meal if she wanted it. ...[450 words]...
The words "Forgive me, father, for I have sinned" came unbidden into her mind as she reached out for her drink and took a large sip straight from the glass. ...[383 words]...
But the guilt was weighing on her now. ...[346 words]...

Human Evaluation



- Pairwise preference comparisons between outputs
 - Overall preference
 - Across different axes
- Identification of specific error types
- Humanlikeness

Human Evaluation



Pairwise Preference

Error Analysis

Humanlikeness

Questions:

- 1) Which story do you prefer / find more interesting overall?
 - ☐ Story A
 - ☐ Story B
 - ☐ Both are about equally good
 - ☐ Neither is good
- 2) Which story has a more coherent overarching plot?
 - ☐ Story A
 - ☐ Story B
 - ☐ Both are about equally good
 - ☐ Neither is good
- 3) Which story's plot is closer to the premise?
 - ☐ Story A
 - ☐ Story B
 - ☐ Both are about equally good
 - ☐ Neither is good
- 4) Indicate which of the following problems are present in Story A (possibly none, possibly more than one).
 - ☐ Jarring change(s) in narration or style
 - ☐ Factual inconsistencies/oddities
 - ☐ Very confusing or hard to understand
 - ☐ Often ungrammatical or disfluent
 - ☐ Highly repetitive
 - ☐ None of the above
- 5) Indicate which of the following problems are present in Story B (possibly none, possibly more than one).
 - ☐ Jarring change(s) in narration or style
 - ☐ Factual inconsistencies/oddities
 - ☐ Very confusing or hard to understand
 - ☐ Often ungrammatical or disfluent
 - ☐ Highly repetitive
 - ☐ None of the above
- 6) Do you think Story A was written by a human?
 - ☐ Yes
 - ☐ No
- 7) Do you think Story B was written by a human?
 - ☐ Yes
 - ☐ No

1. **Interesting.** Interesting to the reader.
2. **Coherent.** Plot-coherent.
3. **Relevant.** Faithful to the initial premise.
4. **Humanlike.** Judged to be human-written.

We additionally track how often generated stories suffer from any of the following writing issues:

1. *Narration.* Jarring change(s) in narration and/or style.
2. *Inconsistent.* Factually inconsistent or containing very odd details.
3. *Confusing.* Confusing or difficult to follow.
4. *Repetitive.* Highly repetitive.
5. *Disfluent.* Frequent grammatical errors.

Method	Interesting ↑	Coherent ↑	Relevant ↑	Humanlike ↑	Misc. Problems ↓
ROLLING	45.0	45.7	44.0	74.0	1.20
RE ³	54.3	60.0	64.0	83.3	1.07
ROLLING-FT	52.7	48.7	49.3	74.7	1.48
RE ³	53.7	60.0	65.3	80.0	1.35

LLM-as-a-Judge



- Human evaluation is expensive
- Maybe we can just prompt LLMs to judge for us?
- **Open question!**

Communicative Success



- Language is all about **getting things done in the world**
 - Directing a listener's attention to something
 - Instructing or suggesting a listener to do something
 - Acquiring information that someone else has
 - Teaching something to someone else
 - ...
- How well do our systems take action via language?

Evaluating NLG through Communicative Success



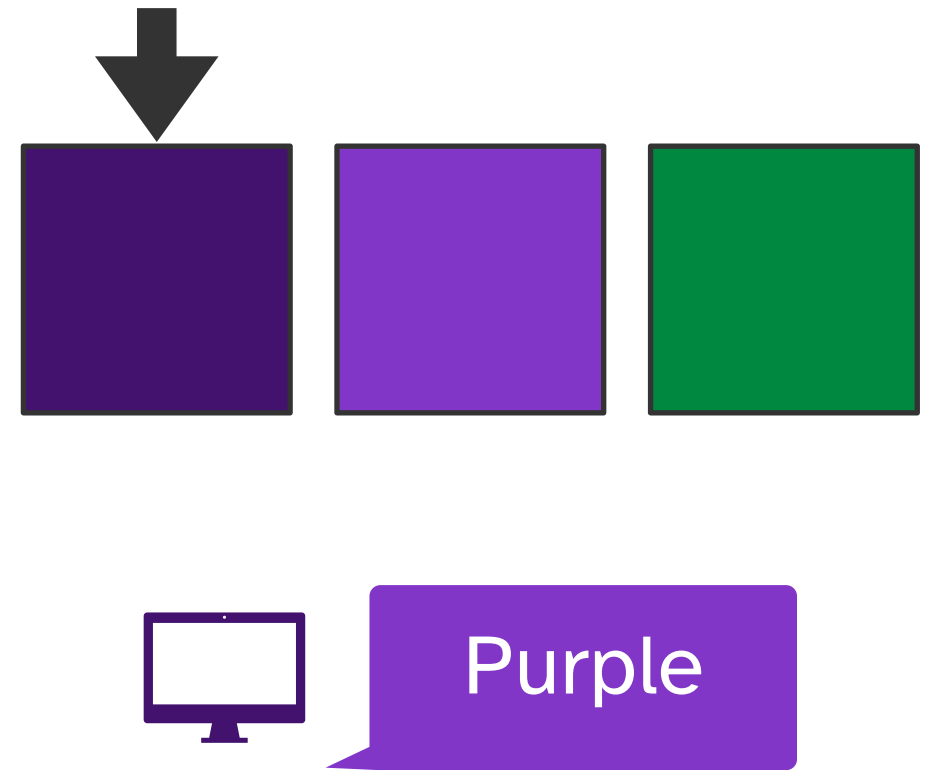
- Example task: reference game
- Input: image and target object



Evaluating NLG through Communicative Success



- Example task: reference game
- Input: image and target object
- Output: referring expression



Evaluating NLG through Communicative Success



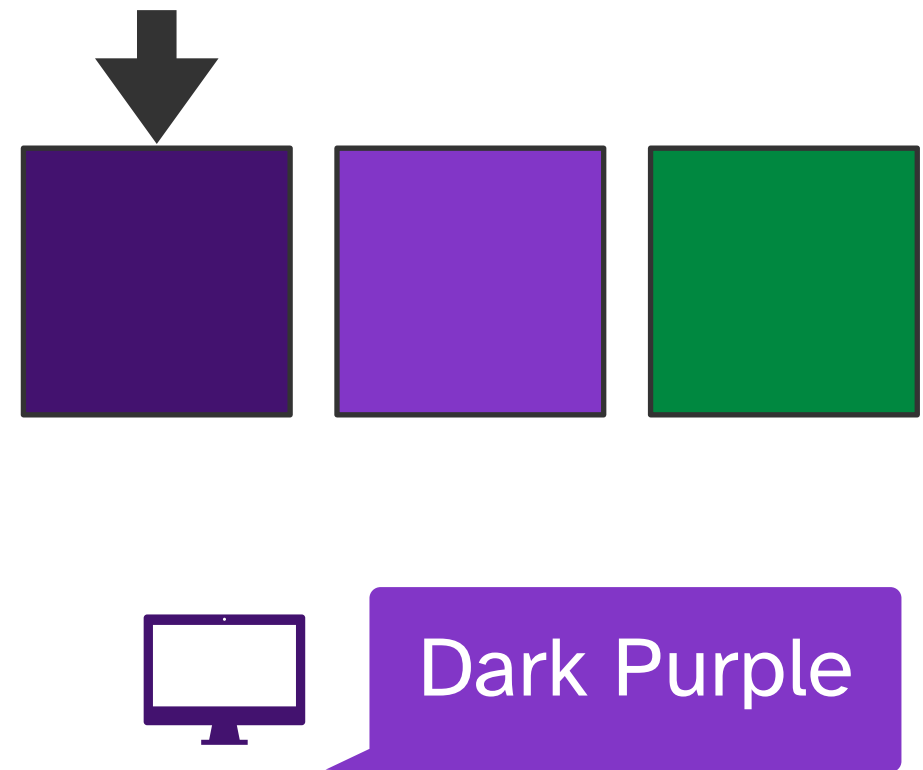
- Example task: reference game
- Input: image and target object
- Output: referring expression
- Communicative success is achieved if the listener, who has access to the image, can successfully identify the target



Evaluating NLG through Communicative Success



- Example task: reference game
- Input: image and target object
- Output: referring expression
- Communicative success is achieved if the listener, who has access to the image, can successfully identify the target



Evaluating NLG through Communicative Success



- Example task: reference game
- Input: image and target object
- Output: referring expression
- Communicative success is achieved if the listener, who has access to the image, can successfully identify the target

really easy to evaluate!



Dark Purple



Dialogue and Interactive Systems



- Language is all about **interaction**:
 - How do the utterances we produce influence the world around us?
 - How do other language users change their language use in response to yours?
- **Dynamic properties of interaction**
 - Formation of common ground
 - Alignment
- We need to study models in multi-turn interaction!

Dialogue and Interactive Systems



But evaluating interaction is hard:



What's a good place to get gluten free pastries in Berkeley?



For dedicated gluten-free pastries in Berkeley, your absolute best bet is **Mariposa Baking Company**.

A gluten-free pastry lover in Berkeley!
You're in luck because Berkeley is a haven for GF enthusiasts! Here are some top recommendations for gluten-free pastries in Berkeley:

1. Bette's Oceanview Diner (1807 4th St, Berkeley) — This beloved diner has

...

Dialogue and Interactive Systems



But evaluating interaction is hard:



What's a good place to get gluten free pastries in Berkeley?



For dedicated gluten-free pastries in Berkeley, your absolute best bet is **Mariposa Baking Company**.



That's too far from me. Do you have a recommendation near downtown Berkeley?

A gluten-free pastry lover in Berkeley!
You're in luck because Berkeley is a haven for GF enthusiasts! Here are some top recommendations for gluten-free pastries in Berkeley:

1. Bette's Oceanview Diner (1807 4th St, Berkeley) — This beloved diner has

...

Dialogue and Interactive Systems



But evaluating interaction is hard:



What's a good place to get gluten free pastries in Berkeley?



For dedicated gluten-free pastries in Berkeley, your absolute best bet is **Mariposa Baking Company**.

A gluten-free pastry lover in Berkeley!
You're in luck because Berkeley is a haven for GF enthusiasts! Here are some top recommendations for gluten-free pastries in Berkeley:

1. Bette's Oceanview Diner (1807 4th St, Berkeley) — This beloved diner has

...



Are any of these open on Sundays?

Dialogue and Interactive Systems



But evaluating interaction is hard:



What's a good place to get gluten free pastries in Berkeley?



For dedicated gluten-free pastries in Berkeley, your absolute best bet is **Mariposa Baking Company**.



That's too far from me. Do you have a recommendation near downtown Berkeley?

A gluten-free pastry lover in Berkeley!
You're in luck because Berkeley is a haven for GF enthusiasts! Here are some top recommendations for gluten-free pastries in Berkeley:

1. Bette's Oceanview Diner (1807 4th St, Berkeley) — This beloved diner has

...



Are any of these open on Sundays?

**because user behavior changes
depending on system behavior!**

Slot-Filling Dialogue Systems



- In well-defined tasks, we can define **slots** that need to be filled before providing information to the user
- However, this reduces the expressivity of dialogue:
 - “I was hoping you could **recommend** something”
 - “Are there any churches **or** museums on the east side?”
 - “I would like the **latest** train leaving that will arrive by 9:15 please.”

A: Hello! This is Concierge Service. I can help you find attractions, hotels, restaurants in Cambridge.

U: I'm looking for a restaurant.

[Domain=Restaurant]

A: What cuisine would you like?

U: I would like Italian food.

[Domain=Restaurant, Food=Italian]

A: Would you like a cheap, moderate or expensive Italian restaurant?

U: Actually, never mind, let's do Chinese.

[Domain=Restaurant, Food=Chinese]

A: Would you like a cheap, moderate or expensive Chinese restaurant?

User Simulators



- Use another model (or a set of rules) to “simulate” a human user
 - Allows scaling up experiments, including to more complex domains
 - Allows stability across system evaluations
- But doesn't reflect the real-world complexity of actual use case, e.g., user adaptation to systems over time
- **Open question!**

Building Language Technologies

Tasks (in General)



Most tasks can be thought of as a mapping from some input X to some output Y , where one or both of these have to do with language

Task	Input (X)	Output (Y)
Text classification	Text	Label from a fixed class
Automatic speech recognition	Audio signal	Text
Dependency parsing	Text	Dependency tree
Code generation	Text	Executable code
Question answering	Document and question (both text)	Answer (text)
Translation	Text (in source language)	Text (in target language)
Open-ended NLG	Optional prompt (text)	Text
Referring expression generation	Image and target object	Referring expression (text)
Dialogue	Conversation history (text)	Next utterance (text)

Implementation



How can we implement a classification model?

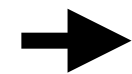
- Manual rules

```
def classify(x: str) -> str:
    sports_keywords = ["baseball", "soccer", "football", "tennis"]
    if any(keyword in x for keyword in sports_keywords):
        return "sports"
    else:
        return "other"
```

- Prompting a model without training

If the following text is about sports, reply "sports". Otherwise, reply "other".

"Cal football is set to lose its entire starting ..."

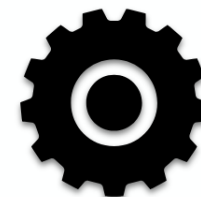


- Machine learning

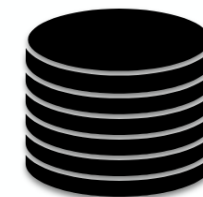
I love to play baseball.	sports
The stock price is going up.	other
He got a hat-trick yesterday.	sports
He is wearing tennis shoes.	other



Training



Model



Evaluation



- Given the implementation, we want to know how well our model can perform given new documents
- What is “new”?
 - Generally speaking: anything we didn’t have access to when implementing our model
- How can we estimate performance on new documents?
 - Simulate it using held-out data just for evaluation!

Data



Before pretrained models, nearly all datasets came with splits, assumed to be IID:

Training data
For updating model parameters directly

Validation data
For deciding when to stop training

Development data
For performing ablations, choosing hyperparameters, etc.

Test data
For estimating performance on the “real” task

Data



Before pretrained models, nearly all datasets came with splits, assumed to be IID:

Training data
For updating model parameters directly

Validation data
For deciding when to stop training

Development data
For performing ablations, choosing hyperparameters, etc.

Public test data

Private test data

Benchmarks and Model Evaluations

Why do we need benchmarks?



- Estimating how well our models will work on real-world data
- Shared understanding of model performance with standardized evaluations
- Building trust within a community in proving how well a new model does
- Driving progress towards specific tasks and capabilities

Single-Task Benchmarks



A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.

SNLI, Bowman et al. 2015

What is the capital of the state with the largest population? <code>answer(C, (capital(S,C), largest(P, (state(S), population(S,P))))).</code>
What are the major cities in Kansas? <code>answer(C, (major(C), city(C), loc(C,S), equal(S,stateid(kansas))))).</code>

GeoQuery, Zelle and Mooney 1996

Battle-tested/NNP*/JJ industrial/JJ managers/NNS here/RB always/RB buck/VB*/VBP up/IN*/RP nervous/JJ newcomers/NNS with/IN the/DT tale/NN of/IN the/DT first/JJ of/IN their/PP\$ countrymen/NNS to/TO visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS*/FW warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.
"/" From/IN the/DT beginning/NN ,/, it/PRP took/VBD a/DT man/NN with/IN extraordinary/JJ qualities/NNS to/TO succeed/VB in/IN Mexico/NNP ,/, "/" says/VBZ Kimihide/NNP Takimura/NNP ,/, president/NN of/IN Mitsui/NNS*/NNP group/NN 's/POS Kensetsu/NNP Engineering/NNP Inc./NNP unit/NN ./.

The Penn Treebank, Marcus et al. 1993

... [Arg0 the company] to ... offer [Arg1 a 15% to 20% stake] [Arg2 to the public] (wsj_0345) ¹
... [Arg0 Sotheby's] ... offered [Arg2 the Dorrance heirs] [Arg1 a money-back guarantee] (wsj_1928)
... [Arg1 an amendment] offered [Arg0 by Rep. Peter DeFazio] ... (wsj_0107)
... [Arg2 Subcontractors] will be offered [Arg1 a settlement] ... (wsj_0187)

PropBank, Palmer et al. 2005

Third-quarter sales in Europe were exceptionally strong, boosted by promotional programs and new products – although weaker foreign currencies reduced the company's earnings.

Michelle lives in a hotel room, and although **she drives a canary-colored Porsche**, *she hasn't time to clean or repair it.*

Most oil companies, when they set exploration and production budgets for this year, forecast revenue of \$15 for each barrel of crude produced.

The Penn Discourse Treebank, Prasad et al. 2008



Instruction: Head upstairs and walk past the piano through an archway directly in front. Turn right when the hallway ends at pictures and table. Wait by the moose antlers hanging on the wall.

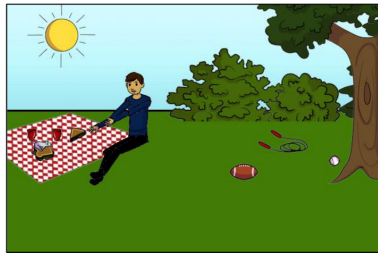
Room-to-Room, Anderson et al. 2018



What color are her eyes?
What is the mustache made of?



How many slices of pizza are there?
Is this a vegetarian pizza?



Is this person expecting company?
What is just under the tree?



Does it appear to be rainy?
Does this person have 20/20 vision?

VQA, Agrawal et al. 2015

Marcus et al. 1993, CL, "Building a large annotated corpus of English"; Zelle and Mooney 1996, AAAI, "Learning to parse database queries using inductive logic programming"; Palmer et al. 2005, CL, "The Proposition Bank"; Prasad et al. 2008, LREC, "The Penn Discourse TreeBank 2.0"; Bowman et al. 2015, EMNLP, "A large annotated corpus for learning natural language inference"; Agrawal et al. 2015, ICCV, "VQA"; Anderson et al. 2018, CVPR, "Vision-and-language navigation"

Single-Task Benchmarks



A man inspects the uniform of a figure in some East Asian country.	contradiction C C C C C	The man is sleeping	Classification
An older and younger man smiling.	neutral N N E N N	Two men are smiling and laughing at the cats playing on the floor.	
A black race car starts up in front of a crowd of people.	contradiction C C C C C	A man is driving down a lonely road.	
A soccer game with multiple males playing.	entailment E E E E E	Some men are playing a sport.	
A smiling costumed woman is holding an umbrella.	neutral N N E C N	A happy woman in a fairy costume holds an umbrella.	

SNLI, Bowman et al. 2015

What is the capital of the state with the largest population?
`answer(C, (capital(S,C), largest(P, (state(S), population(S,P))))).`

What are the major cities in Kansas?

`answer(C, (major(`
`equal(S,stateid(k`

Structured Prediction

GeoQuery, Zelle and Mooney 1996

Tagging

.../JJ industrial/JJ managers/NNS here/RB
/VBP up/IN*/RP nervous/JJ newcomers/NNS with/IN
N the/DT first/JJ of/IN their/PP\$ countrymen/NNS to/TO
visit/VB Mexico/NNP ,/, a/DT boatload/NN of/IN samurai/NNS*/FW
warriors/NNS blown/VBN ashore/RB 375/CD years/NNS ago/RB ./.
"/" From/IN the/DT beginning/NN ,/, it/PRP took/VBD a/DT man/NN
with/IN extraordinary/JJ qualities/NNS to/TO succeed/VB in/IN Mexico/NNP ,/,
"/" says/VBZ Kimihide/NNP Takimura/NNP ,/, president/NN of/IN
Mitsui/NNS*/NNP group/NN 's/POS Kensetsu/NNP Engineering/NNP Inc./NNP
unit/NN ./.

The Penn Treebank, Marcus et al. 1993

... [the company] to ... off
(wsj_0345)¹

... [Sotheby's] ... offered [the Dorrance heirs] [a money-back
guarantee] (wsj_1928)

... [an amendment] offered [by Rep. Peter DeFazio] ... (wsj_0107)

... [Subcontractors] will be offered [a settlement] ... (wsj_0187)

Spanning / Tagging

PropBank, Palmer et al. 2005

Third-quarter sales in Europe were ex
boosted by promotional programs and
although weaker foreign currencies reduced the com
pany's earnings.

Tagging

Michelle lives in a hotel room, and although she drives
a canary-colored Porsche, she hasn't time to clean or
repair it.

Most oil companies, when they set exploration and pro
duction budgets for this year, forecast revenue of \$15
for each barrel of crude produced.

The Penn Discourse Treebank, Prasad et al. 2008

Sequence Generation



Instruction: Head upstairs and walk past the piano through an
archway directly in front. Turn right when the hallway ends at
pictures and table. Wait by the moose antlers hanging on the wall.

Room-to-Room, Anderson et al. 2018

Question Answering



What color are her eyes?
What is the mustache made of?

How many slices of pizza are there?
Is this a vegetarian pizza?

Is this person expecting company?
What is just under the tree?

Does it appear to be rainy?
Does this person have 20/20 vision?

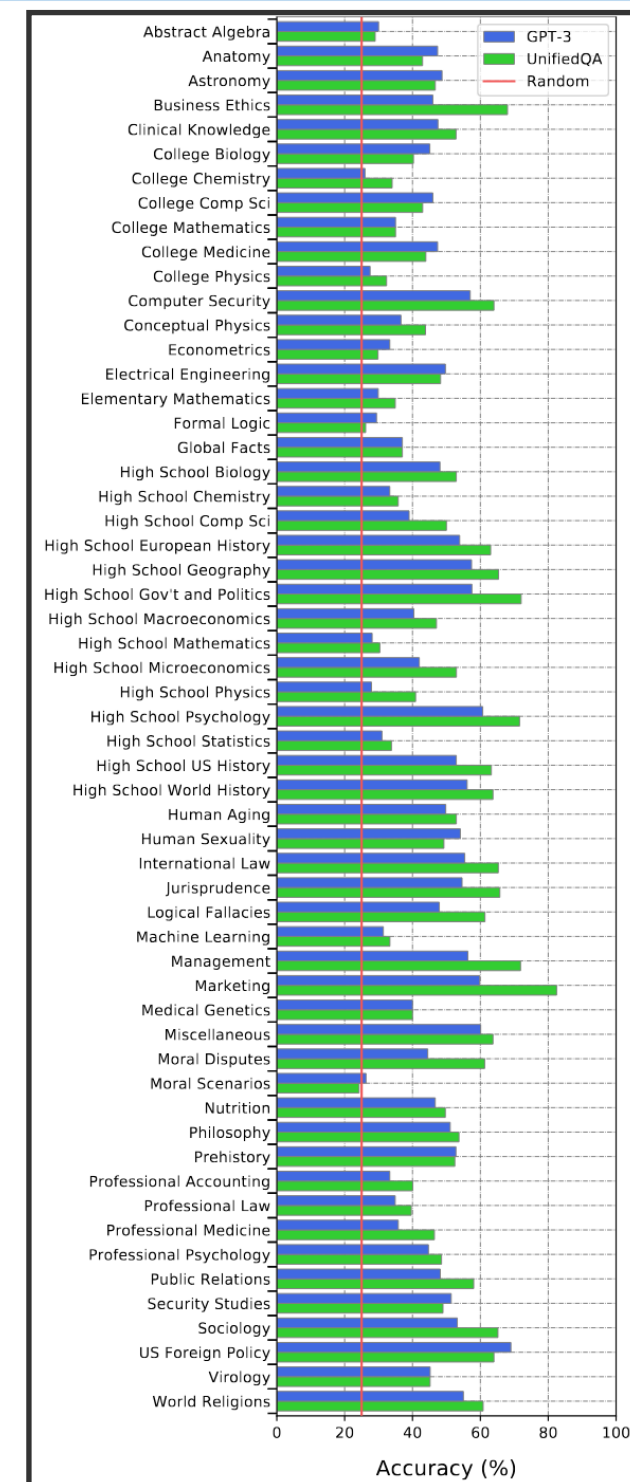
VQA, Agrawal et al. 2015

Multi-Task Benchmarks



Corpus	Train	Test	Task	Metrics	Domain
Single-Sentence Tasks					
CoLA	8.5k	1k	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
Similarity and Paraphrase Tasks					
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	391k	paraphrase	acc./F1	social QA questions
Inference Tasks					
MNLI	393k	20k	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	146	coreference/NLI	acc.	fiction books

GLUE, Wang et al. 2019

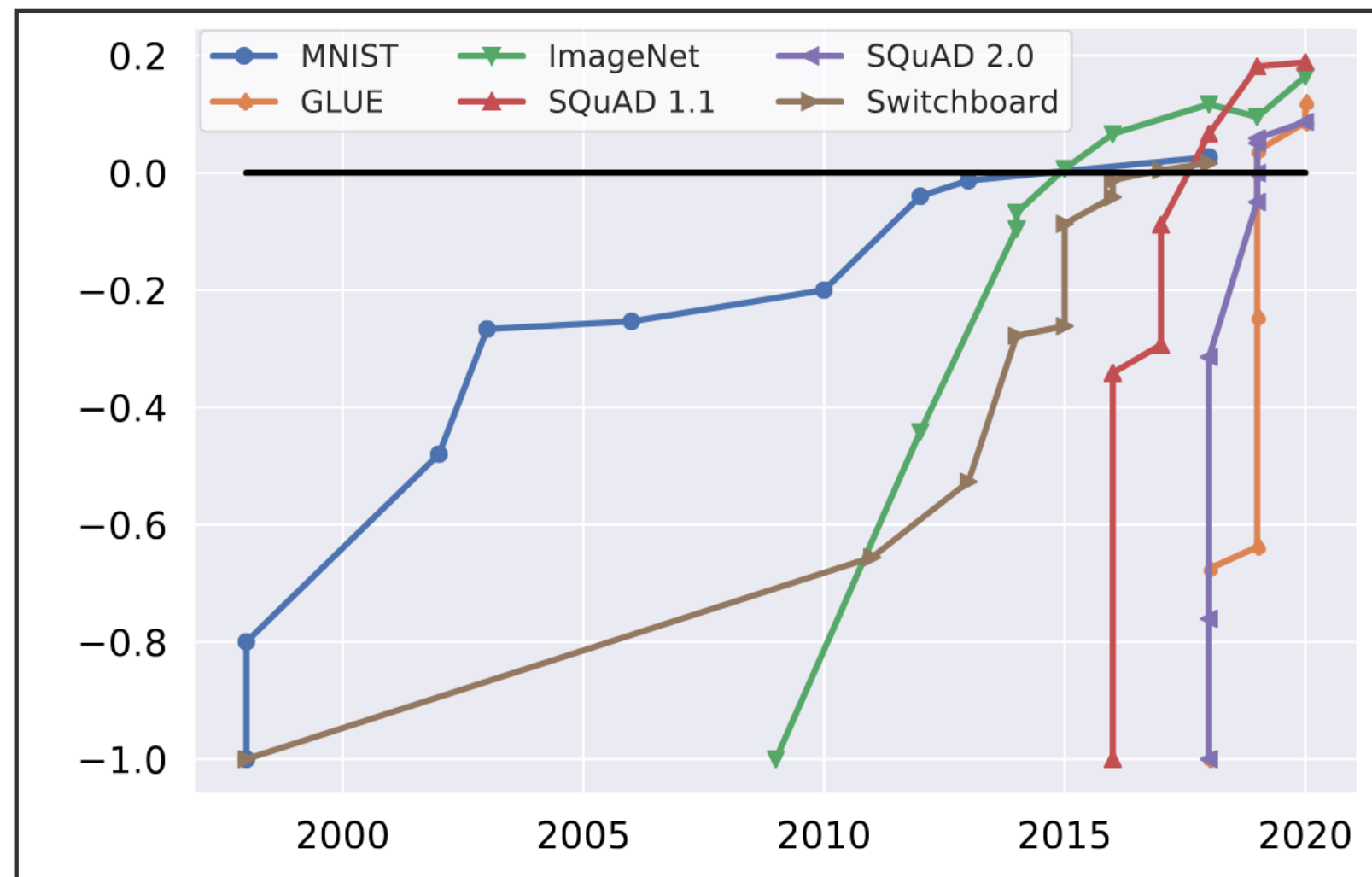


MMLU, Hendrycks et al. 2021

Capability-Pushing Benchmarks



Benchmarks are getting saturated quickly!



Kiela et al. 2021

Capability-Pushing Benchmarks



Capability-focused evaluation: choose a complex task and curate examples of it

Organic Chemistry

Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed (as a mixture of isomers), with water as a side product. These products are derivatives of fulvene. This product was then allowed to react with ethyl acrylate in a 1:1 ratio. Upon completion of the reaction, the bright yellow color had disappeared. How many chemically distinct isomers make up the final product (not counting stereoisomers)?

- A) 2
- B) 16
- C) 8
- D) 4

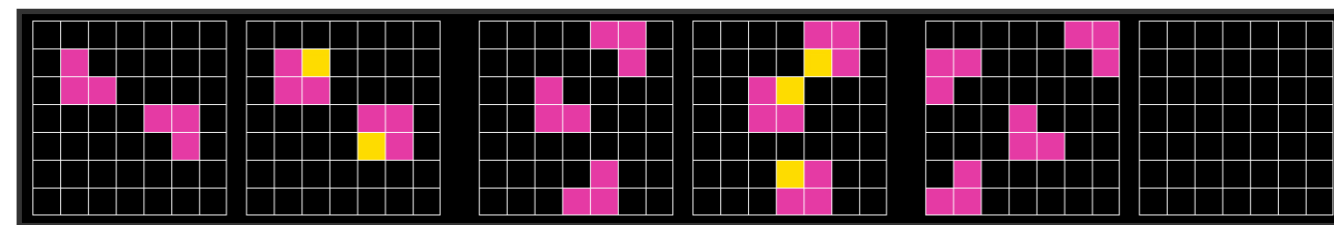
GPQA, Rein et al. 2024



BabyLM Challenge

Sample-efficient pretraining on a developmentally plausible corpus

BabyLM, Warstadt et al. 2023



ARC-AGI-1, Chollet 2019

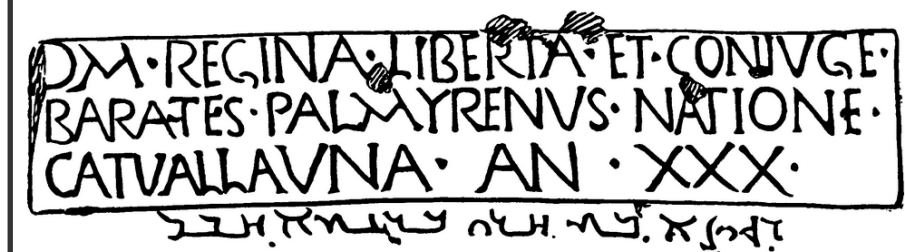
Konwinski Prize

\$1M for the AI that can close 90% of new GitHub issues

Andy Konwinski

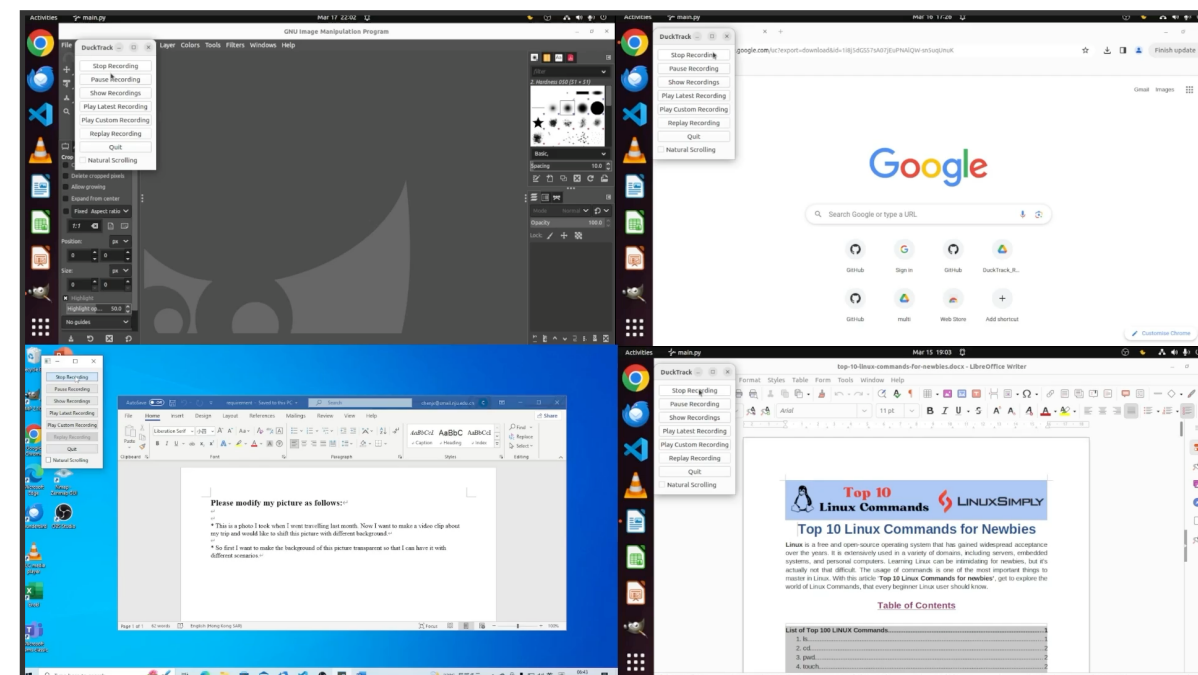
Classics

Question:



Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script. A transliteration of the text is provided: RGYN° BT HRY BR °T° HBL

HLE, Phan et al. 2025



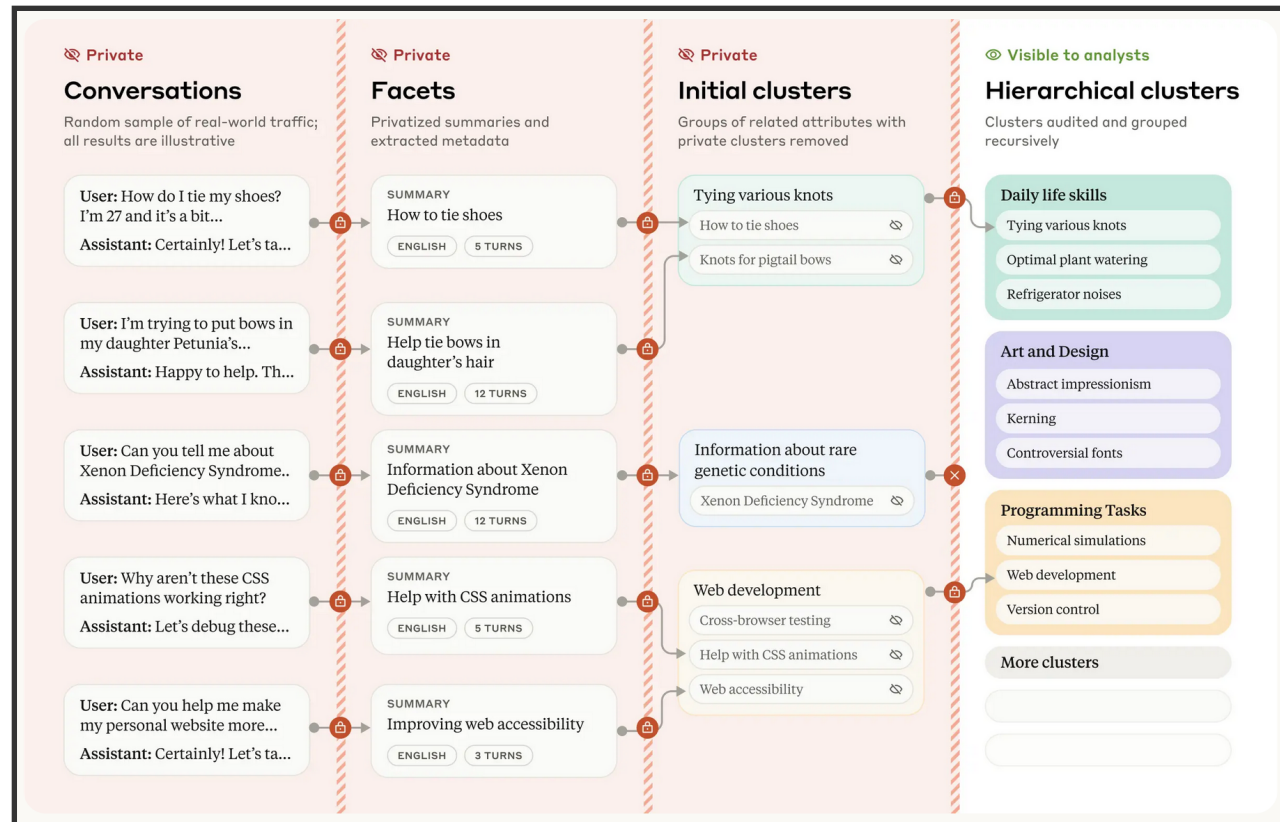
OSWorld, Xie et al. 2024

Use of Language Technologies

“In the Wild”



What are people *actually* using LMs for, and how well do they do?



Clio, Tamkin et al. 2024



WildChat (Zhao et al. 2024), WildBench (Lin et al. 2024)

LMarena

Scroll to the right to see full stats of each model

Model	Overall	Hard Prompts	Coding	Math	Creative Writing	Instruction Following	Longer Query	Multi-Turn
gemini-2.5-pro-prev...	1	1	1	1	1	1	1	1
gemini-2.5-pro-prev...	2	2	2	1	1	1	1	1
o3-2025-04-16	2	2	2	1	3	3	6	4
chatgpt-4o-latest-2...	4	4	2	6	2	3	1	1
gpt-4.5-preview-202...	4	4	2	4	3	3	2	1

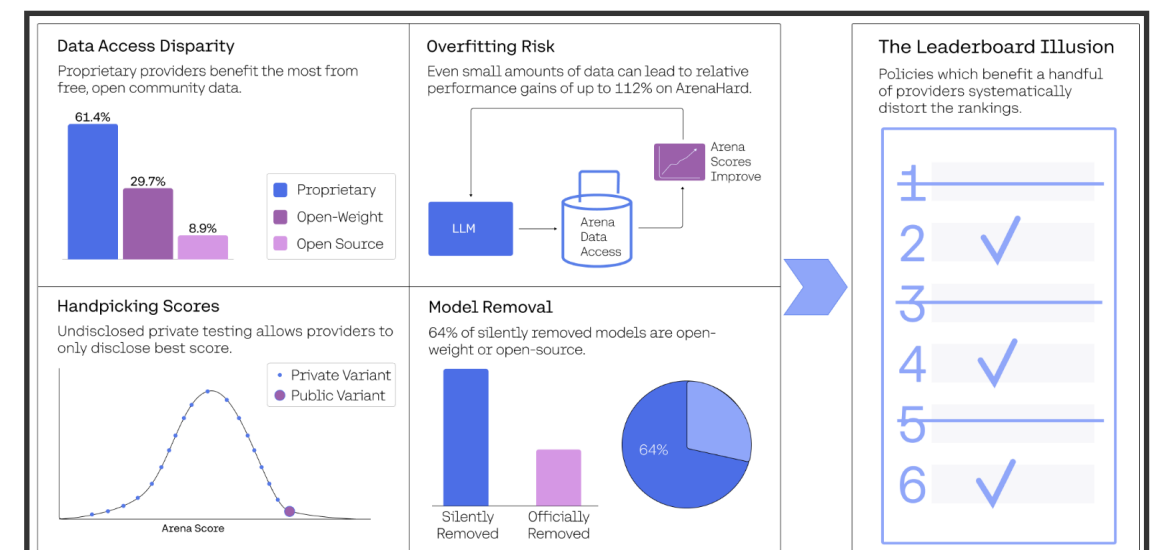
LMarena / Chatbot Arena, Chiang et al. 2024

OpenRouter

Token usage across models

Rank	Model	Token Usage
1.	Google: Gemini 2.0 Flash	267B tokens
2.	Anthropic: Claude Sonnet 4	196B tokens
3.	OpenAI: GPT-4o-mini	188B tokens
4.	Google: Gemini 2.5 Flash Preview 05-20	163B tokens
5.	Anthropic: Claude 3.7 Sonnet	137B tokens

OpenRouter

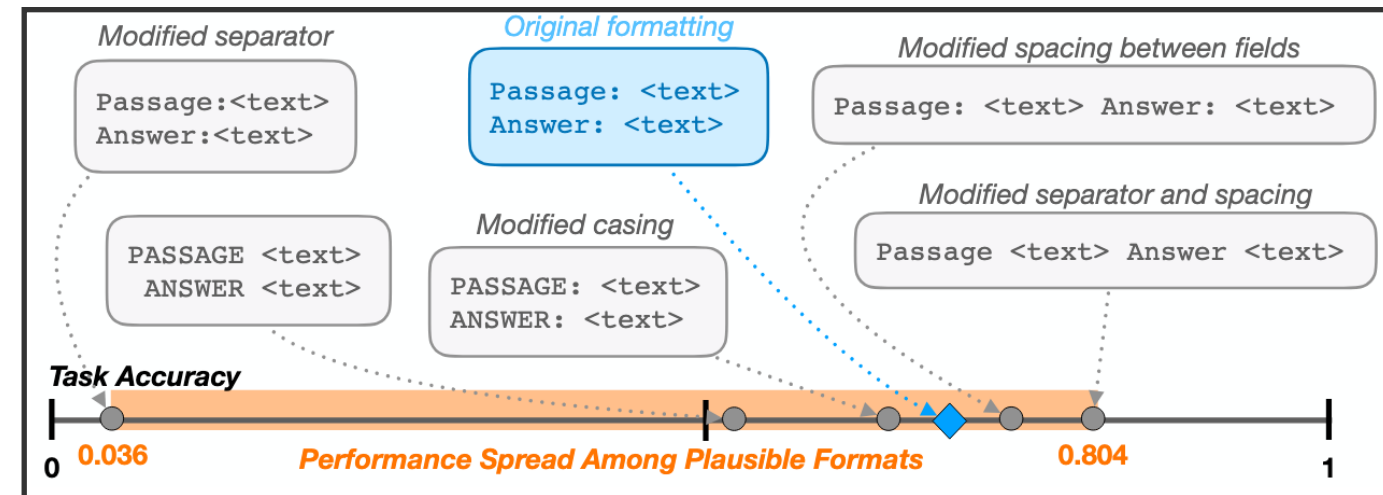


Singh et al. 2025

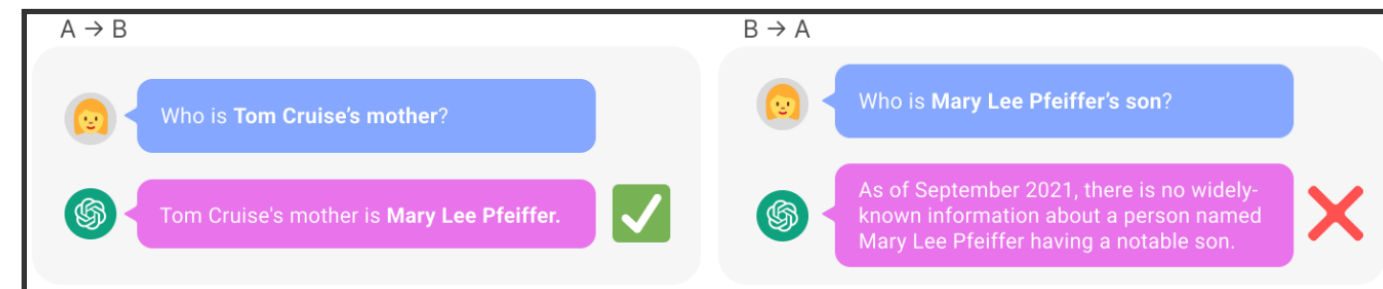
Distributional Model Evaluation



- What are the properties of the outputs models produce in general?
- **Consistency:** when evaluating models on the same task (or variations of it) multiple times, how consistent is its behavior?
- **Diversity:** do the outputs models generate have the same distributional properties as human language?



Sclar et al. 2024



Berglund et al. 2024



Zhang et al. 2025

Pitfall: Dataset Contamination



Before pretrained models, nearly all datasets came with splits, assumed to be IID:

Training data
For updating model parameters directly

Validation data
For deciding when to stop training

Development data
For performing ablations, choosing hyperparameters, etc.

Test data
For estimating performance on the “real” task

Pitfall:

Dataset Contamination



Dataset contamination occurs when actual test data is used in any of the previous splits (no longer IID). **Why is this bad?**



Pitfall: Dataset Contamination



After pre-training for multi-task models, it becomes especially problematic

Training data

Models might memorize test data that happens to be in their pretraining dataset!

Development tasks

For performing ablations, choosing hyperparameters, etc.

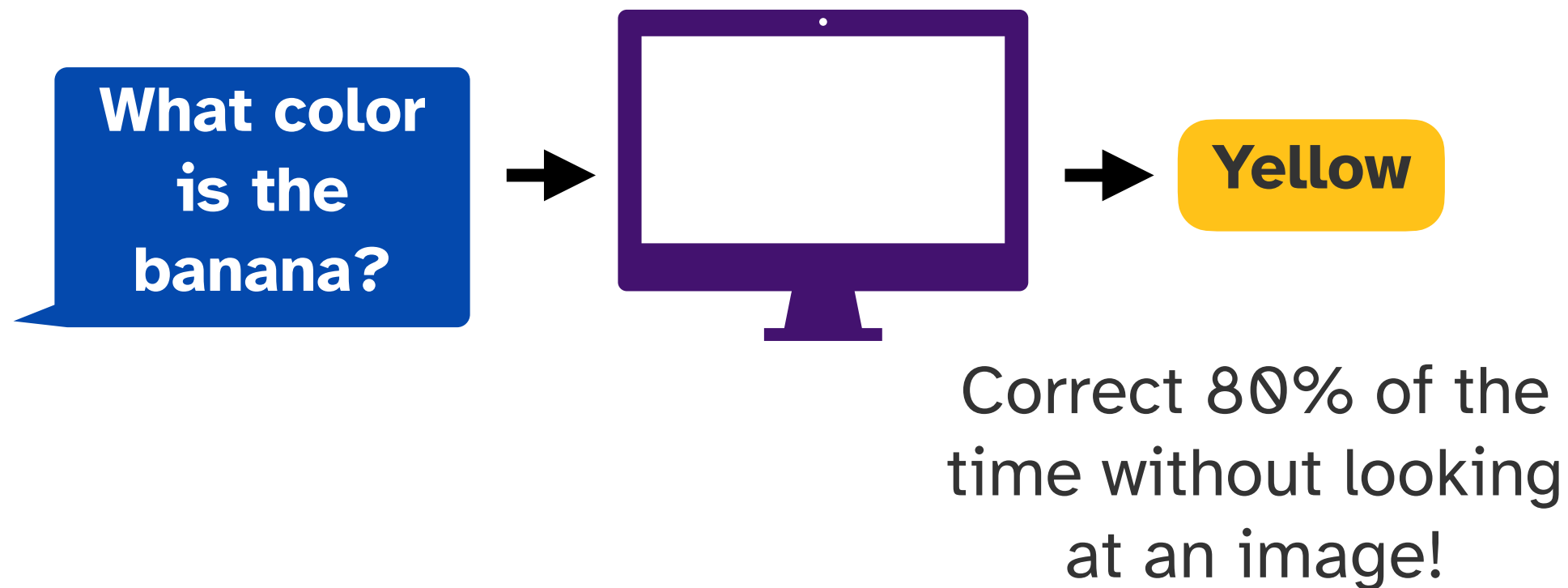
Test tasks

For estimating performance on the “real” task

Pitfall: Spurious Correlations



Does the benchmark actually evaluate what we want it to?



- Learned a spurious correlation that gives the model high accuracy: bananas are yellow
- Doesn't actually test visual understanding

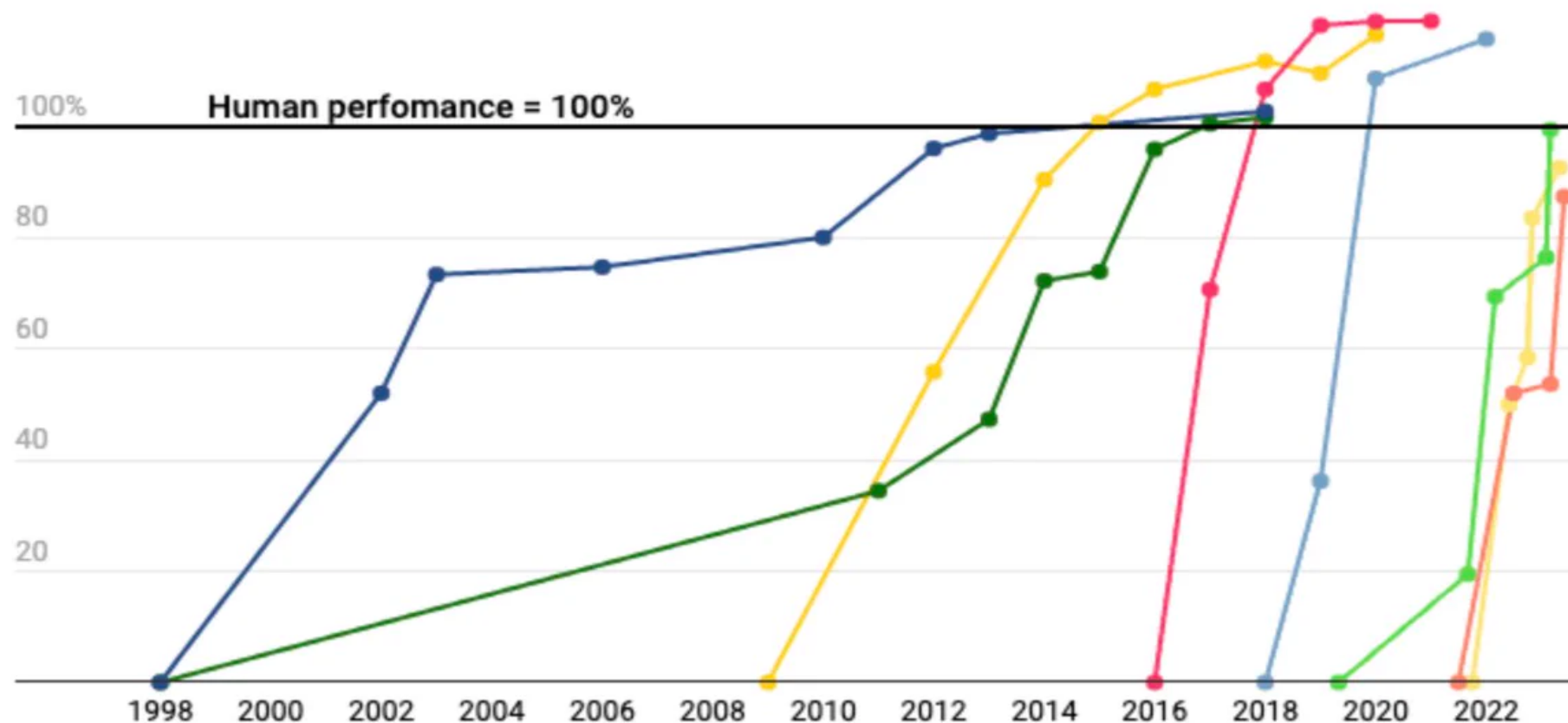
Pitfall: Defining “Human Performance”



AI has surpassed humans at a number of tasks and the rate at which humans are being surpassed at new tasks is increasing

State-of-the-art AI performance on benchmarks, relative to human performance

● Handwriting recognition ● Speech recognition ● Image recognition ● Reading comprehension
● Language understanding ● Common sense completion ● Grade school math ● Code generation



For each benchmark, the maximally performing baseline reported in the benchmark paper is taken as the “starting point”, which is set at 0%. Human performance number is set at 100%. Handwriting recognition = MNIST, Language understanding = GLUE, Image recognition = ImageNet, Reading comprehension = SQuAD 1.1, Reading comprehension = SQuAD 2.0, Speech recognition = Switchboard, Grade school math = GSK8k, Common sense completion = HellaSwag, Code generation = HumanEval.

Chart: Will Henshall for TIME • Source: [ContextualAI](#)

TIME

Pitfall: Defining “Human Performance”



Who is the human in “human performance”?

- Are they incentivized to perform well?
- Do they have the required expertise?

Organic Chemistry

Methylcyclopentadiene (which exists as a fluxional mixture of isomers) was allowed to react with methyl isoamyl ketone and a catalytic amount of pyrrolidine. A bright yellow, cross-conjugated polyalkenyl hydrocarbon product formed (as a mixture of isomers), with water as a side product. These products are derivatives of fulvene. This product was then allowed to react with ethyl acrylate in a 1:1 ratio. Upon completion of the reaction, the bright yellow color had disappeared. How many chemically distinct isomers make up the final product (not counting stereoisomers)?

- A) 2
- B) 16
- C) 8
- D) 4

GPQA, Rein et al. 2024

1. Write a question, including possible multiple choice answers and an explanation of the answer
2. Get expert #1 to validate by having them answer the question and suggest revisions
3. Revise the question given feedback
4. Get expert #2 to validate again
5. Get 3x non-expert (experts in other domains) to attempt to answer the question by allowing Google searches

Keep questions where both experts agree, and at most 1 non-experts can answer the question correctly with Google

Pitfall: Defining “Human Performance”

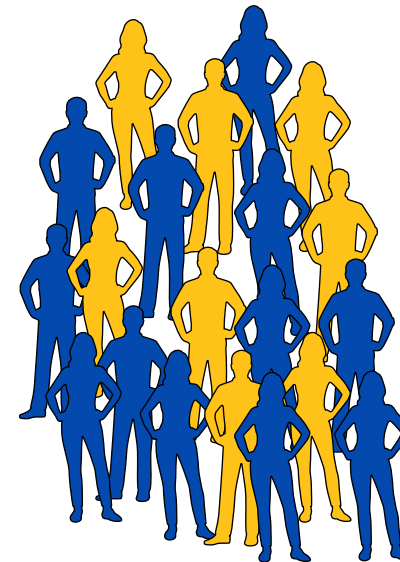


Is the task actually evaluable?

- Is there an objectively correct label?

What color should we make
our Berkeley t-shirts?

- Do annotators disagree? If so, how are we processing their disagreements?

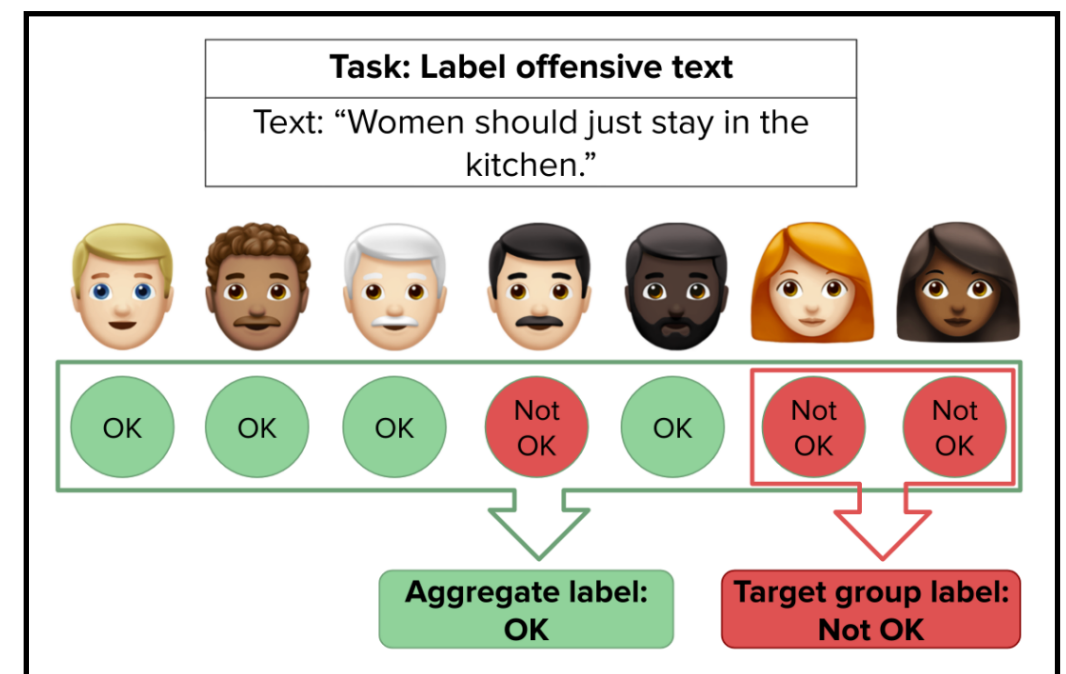


20 annotators

“Gold” label:

Blue

Human performance:
 $12/20 = 60\%$



Fleisig et al. 2023

Pitfall: Defining “Human Performance”



Do they match human performance in dimensions other than average single-instance performance?

- Consistency?
- Explainability?
- Generalizability?

Pitfall: The Long-Tail Paradox



- Which of the following words is most rare?
myriad, solipsist, anachronistic, apricate
- Can you think of a rarer word than these? (Check in the [Google Books Ngram Viewer](#))
- LLMs can look really impressive when we are trying to challenge them with tricks, but maybe this is because we're bad at coming up with long-tail (rare) challenging tasks out of the blue!
- Can seem even more impressive when the task is unverifiable, and any model response could satisfy you!

***“People... are easily fooled into reading structure into chaos,
reading meaning into nonsense... how much different is
interpreting non sequitur as whimsical conversation? ... a test
based on fooling people is confoundingly simple to pass”***

—Stuart Shieber, “Lessons from a restricted Turing test”

Zooming Out: The Role of Language Technologies



- What is their **intended** use?
 - Keeping users engaged?
 - Interacting with users as little as possible?
 - What kinds of users?
- How does this influence their design?
 - Their interface
 - The underlying model
 - The assumptions they are making

Zooming Out: The Role of Language Technologies



- How are they **actually** used?
 - Where do they fail?
 - How do users adapt to failures?
- What other considerations should we make in deployment?
 - Resource utilization and efficiency
 - Fairness, ethics, and social impact

Assignment 1: Evaluating Language Technologies



- Implement 4 fundamental tasks and evaluation metrics for language technologies
 - Classification
 - Automatic speech recognition
 - Machine translation
 - Open-ended text generation
- Evaluate and perform error analysis on some LLMs
- Bonus: design your own task and eval!

This Week's Discussion Sections



Friday:

8-9 AM Section 105 / CORY 289

9-10 AM Section 106 / DWIN 215

11 AM-12 PM Section 110 / ETCH 3107

6-7 PM Section 109 / DWIN 229

GSI: Sanjay