

Evaluation and Interpretability of NLP systems

Pia Sommerauer
Vrije Universiteit Amsterdam
2025/01/21

Introduction - who am I?



BA English and American Studies at
the University of Vienna

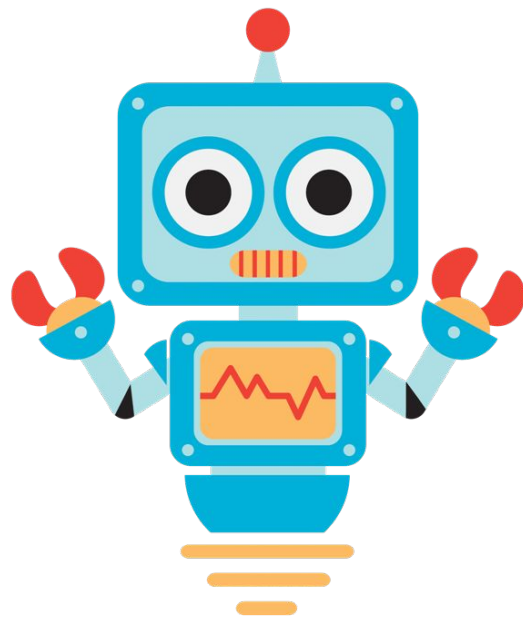


- Assistant Professor at VU Amsterdam
- PhD VU Amsterdam (2022)
- MA VU Amsterdam (2017)



Build models to understand how language works

?!



Introduction - who are you?

Today

Lecture	14:00-15:45
Hands-on session	16:15-18:00

- Ask questions anytime!
- Need a break? Let me know!

Lecture

Introduction to NLP	<i>What is NLP? Why model language computationally?</i>
Evaluation	<i>How do we know a system is working?</i>
Biases and shortcuts	<i>How do we know a system is doing what we think it's doing?</i>
Interpretability	<i>What does a model know? What does it look like? How well will it generalize?</i>
Introduction to the hands-on session	<i>Creating a checklist for a targeted analysis</i>

Goals

1. Understand the **basics of NLP**
2. Understand the importance of **evaluation**
3. **Critically question generic scores**
4. **Tools** for a simple, but powerful system **testing**

What is NLP?

What is Natural Language Processing (NLP)?

What is Natural Language Processing (NLP)?

Automatically processing language (usually text, sometimes also speech):

- Translation
- Summarization
- Extract information from text
- Classify documents (e.g. per topic)
- Sentiment Analysis
- Opinion extraction
- Hate speech detection
- Question-Answering
- ...

What is Natural Language Processing (NLP)?

- Automatically *understanding* language
- Automatically *generating* language
- Combinations of *understanding and generating* language

How does NLP work?

Time	Models	Approach	Interpretable
1950s-1970s	Rules	Finite-state automata, logic, ...	yes
1980s, 1990s	Traditional supervised learning	Train models for a specific task using task-specific training data	yes (features)
1990s, 2000s, 2010s	Deep learning , deep learning + word embeddings (2013)	Train models for specific task using task-specific training data	No
From 2017	Transformers (BERT, Roberta, etc., encoder-decoder, encoder-only)	Pre-trained models (masked word prediction); fine-tune models on specific task (supervised learning) using task-specific training data	No
From 2021	LLMs (transformers) (decoder)	Autoregressive LM; predict next word, prompt (zero-shot, few-shot)	No

How does NLP work?

Supervised learning for a specific task:

I (O) took (O) a (O) flight (O) from (O) Amsterdam (LOC) to (O) Vienna (LOC) yesterday (TIME).

Supervised learning to represent language:

Would you like some coffee or [mask]

Supervised machine learning

What an awesome movie!

Great restaurant

Delicious food!

Boring play, fell asleep

No need to see this film

Terrible service

I felt offended by this production

Pos

Pos

Pos

Neg

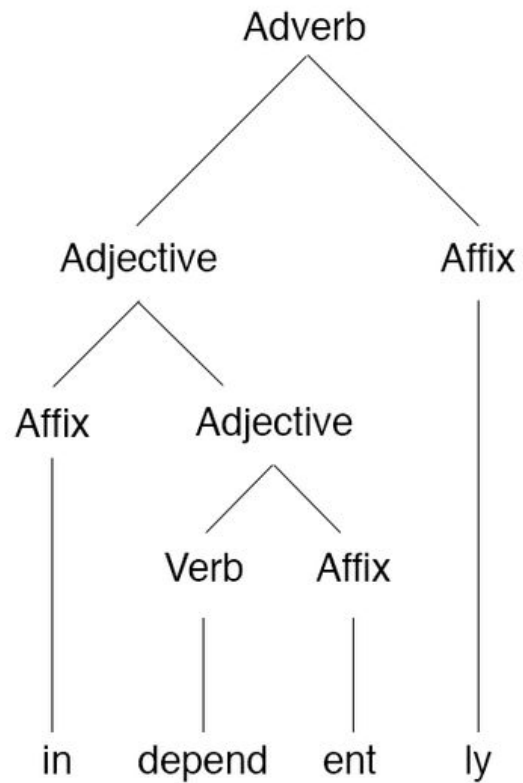
Neg

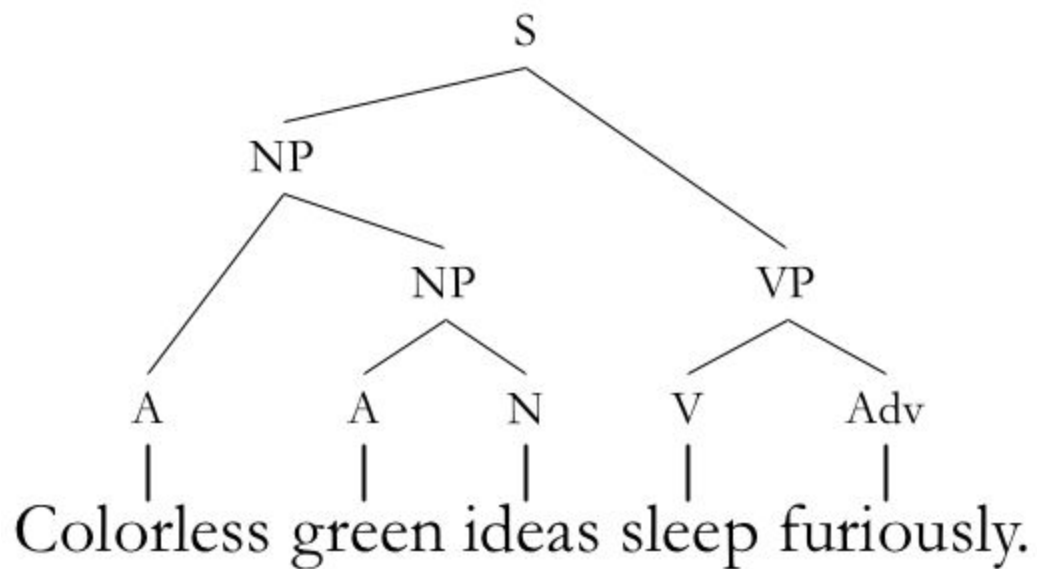
Neg

Neg

Why does supervised learning work for language?

- Language is systematic
- Language has structure





Who did **What** to **Whom**, and **How**, **When** and **Where** ...
(etc.) ?

[**Last night**] [**John**] [**threw**] [**a ball**] [**to Mary**] [**in the park**]

Why are NLP systems not perfect?

Variation

- (1) *Two planes were flown into the twin towers of the World Trade Center in New York City.*
- (2) *Coordinated suicide terrorist attacks carried out by the militant Islamist extremist network al-Qaeda against the United States on September 11, 2001.*
- (3) *The complete destruction of such massive buildings shocked nearly everyone.*

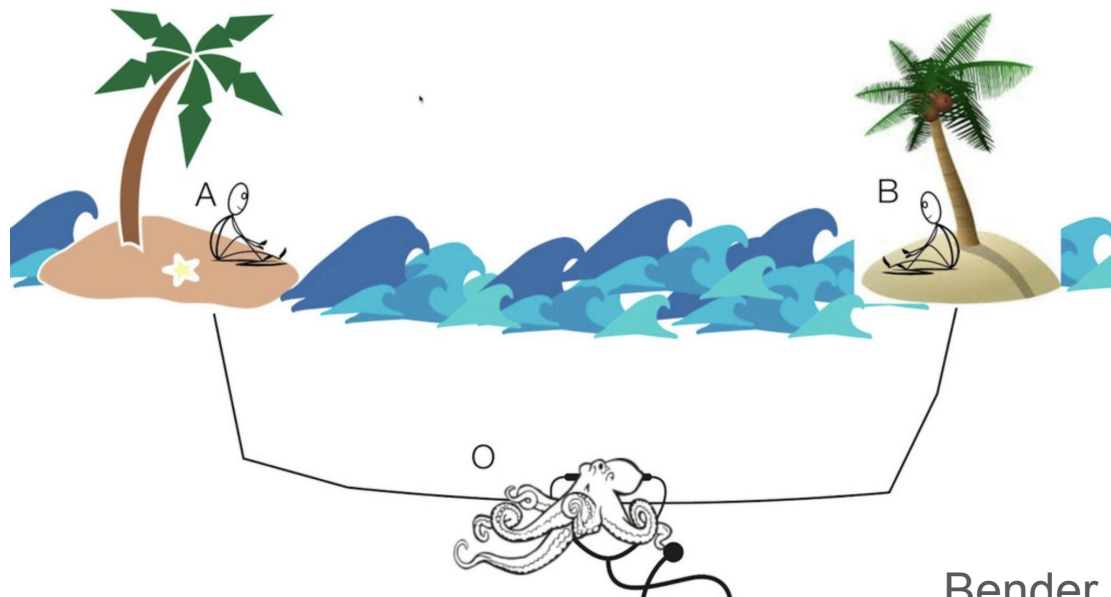
Taken from Remijnse (2025, p. 1)

Ambiguity

- *Bank*
- *Apple*
- *I saw the man with the telescope.*
- *She did not kill the man with the knife.*

Distributional semantics and the octopus

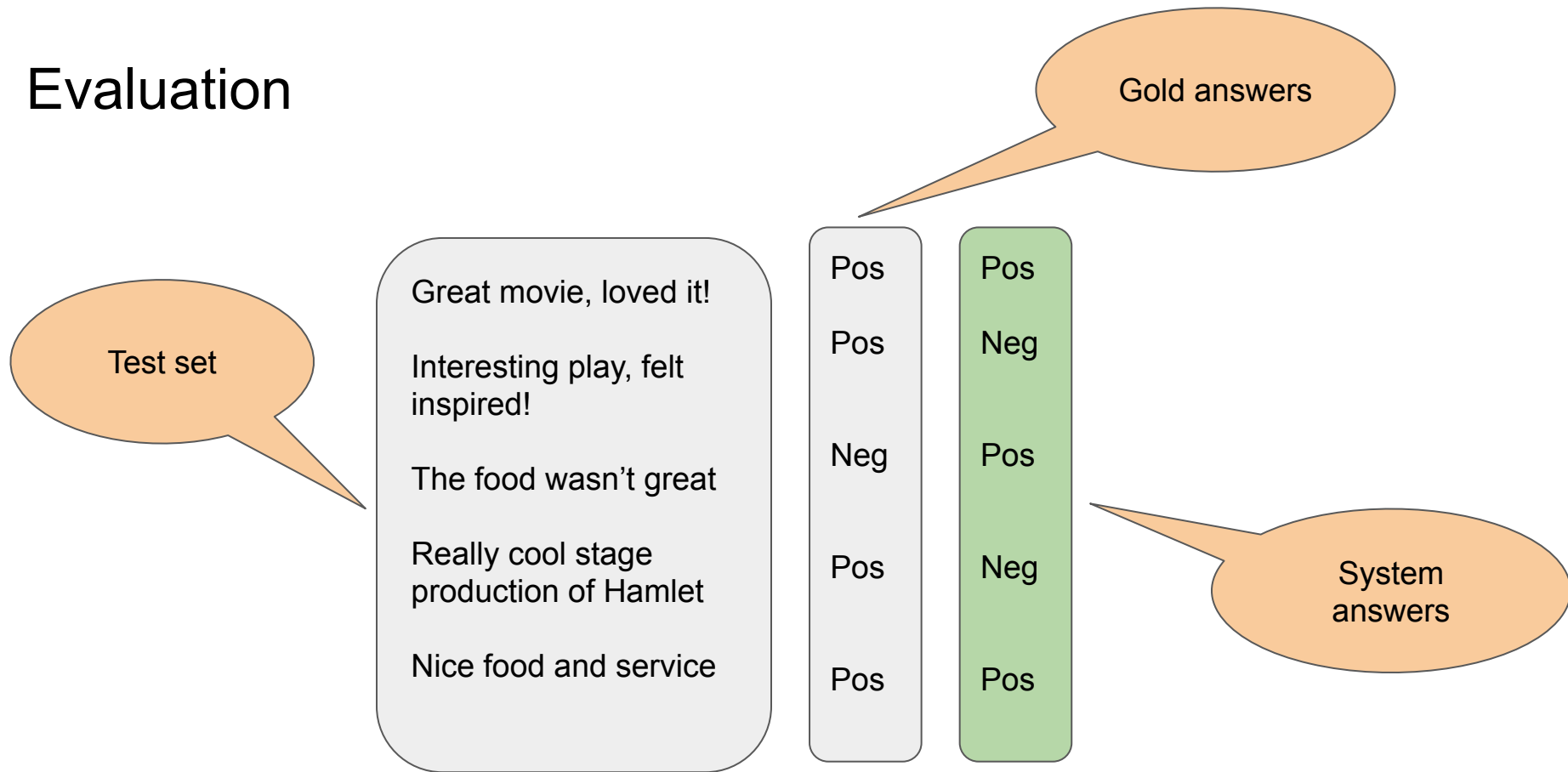
Distributional models can learn linguistically (and specifically semantically) accurate representations purely based on exposure to language (equal to forms).



Evaluation

What is evaluation?

Evaluation



Evaluation metrics

Accuracy

Proportion of correct
answers

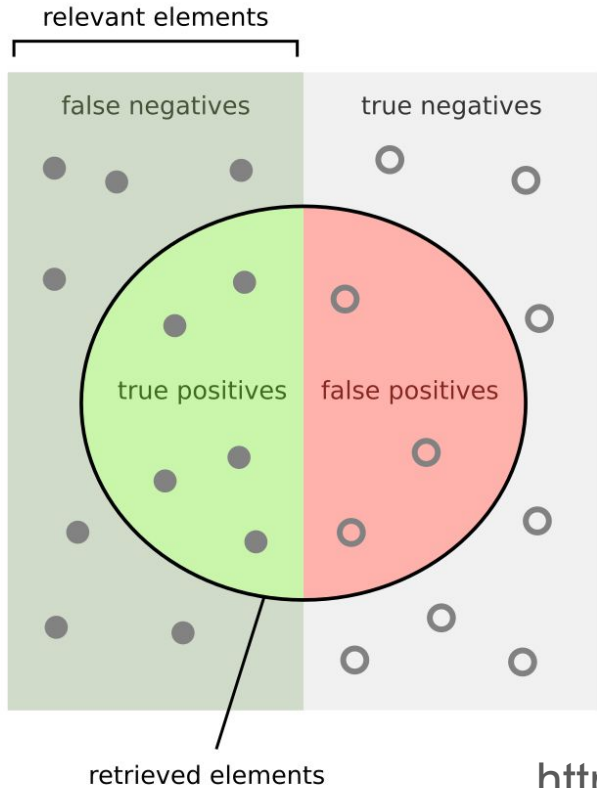
Limited!

Evaluation metrics

		Predicted condition	
		Positive (PP)	Negative (PN)
Total population = P + N			
Actual condition	Positive (P)	True positive (TP)	False negative (FN)
	Negative (N)	False positive (FP)	True negative (TN)

Taken from https://en.wikipedia.org/wiki/Confusion_matrix

Evaluation metrics



How many retrieved items are relevant?

$$\text{Precision} = \frac{\text{true positives}}{\text{true positives} + \text{false positives}}$$

How many relevant items are retrieved?

$$\text{Recall} = \frac{\text{true positives}}{\text{true positives} + \text{false negatives}}$$

$$F_1 = 2 \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

Requirements for evaluation data

- **High-quality (no mistakes)**
- **Human-labeled**
- **Not seen by the system**

Great movie, loved it!

Pos

Interesting play, felt inspired!

Pos

The food wasn't great

Neg

Really cool stage production of Hamlet

Pos

Nice food and service

Pos

Requirements for evaluation data

- **Representative of the task**
- **Indicate how the system will perform when applied (realistic)**

Great movie, loved it!

Pos

Interesting play, felt inspired!

Pos

The food wasn't great

Neg

Really cool stage production of Hamlet

Pos

Nice food and service

Pos

Other considerations

- Span labeling - how to match?
- Coreference resolution - clusters
- Include or exclude O (outside) labels?
- Micro- or macro-averaging?
- ...

Bias and shortcuts

Wolves and dogs

Explain the predictions

Explain the Prediction



Predicted: **Wolf**
True: **Wolf**



Predicted: **Husky**
True: **Husky**



Predicted: **Husky**
True: **Husky**



Predicted: **Wolf**
True: **Wolf**



Predicted: **Wolf**
True: **Wolf**



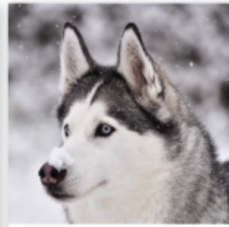
Predicted: **Wolf**
True: **Wolf**



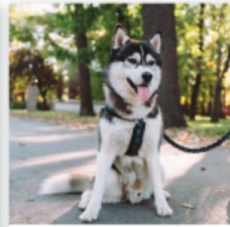
Predicted: **Husky**
True: **Wolf**



Predicted: **Wolf**
True: **Wolf**



Predicted: **Wolf**
True: **Husky**



Predicted: **Husky**
True: **Husky**

What is happening in supervised learning?

What an awesome movie!

Great restaurant

Delicious food!

Boring play, fell asleep

No need to see this film

Terrible service

I felt offended by this production

Pos

Pos

Pos

Neg

Neg

Neg

Neg

learn

F1 score: 0.85

'Evaluation' in terms of a single score

What is happening in supervised learning?

Selective, small error analysis

P

What an awesome movie!

Great restaurant

Delicious food!

Boring play, fell asleep

No need to see this film

Terrible service

I felt offended by this production

Pos

Pos

Pos

Neg

Neg

Neg

Neg

learn

Great movie, loved it!

Interesting play, felt inspired!

The food wasn't great

Really cool stage production of Hamlet

Nice food and service

Pos

Pos

Neg

Pos

Pos

Pos

Neg

Pos

Neg

Pos

What did the model learn?

What is happening in supervised learning?

Selective, small error analysis

P

What an awesome movie!

Great restaurant

Delicious food!

Boring play, fell asleep

No need to see this film

Terrible service

I felt offended by this production

Pos

Pos

Pos

Neg

Neg

Neg

Neg

learn

Great movie, loved it!

Interesting play, felt inspired!

The food wasn't great

Really cool stage production of Hamlet

Nice food and service

Pos

Pos

Neg

Pos

Pos

Pos

Neg

Pos

Neg

Pos

What if the system did not pick up on the
right signals?

What is happening in supervised learning?

Selective, small error analysis

Pr

What an awesome **movie!**

Great **restaurant**

Delicious **food!**

Boring **play**, fell asleep

No need to see this film

Terrible **service**

I felt offended by this **production**

Pos

Pos

Pos

Neg

Neg

Neg

Neg

learn

Great **movie**, loved it!

Interesting **play**, felt inspired!

The **food** wasn't great

Really cool **stage production** of Hamlet

Nice **food** and **service**

Pos

Pos

Neg

Pos

Pos

Pos

Neg

Pos

Neg

Pos

What is happening in supervised learning?

Selective, small error analysis

P

What an awesome movie!

Pos

Great restaurant

Delicious food!

Boring play, fell asleep

No need to see this film

Terrible service

I felt offended by this production

neg

Blackbox

We don't know what the system learned

Great movie, loved it!

Pos

Pos

...ay, felt

Pos

Neg

...sn't great

Neg

Pos

...tage
...f Hamlet

Pos

Neg

Nice food and service

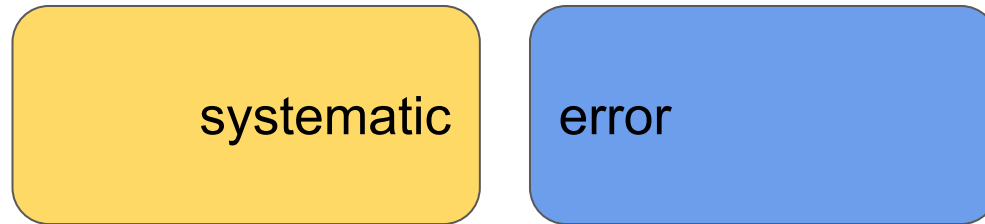
pos

pos

Dataset biases and distributions

Bias

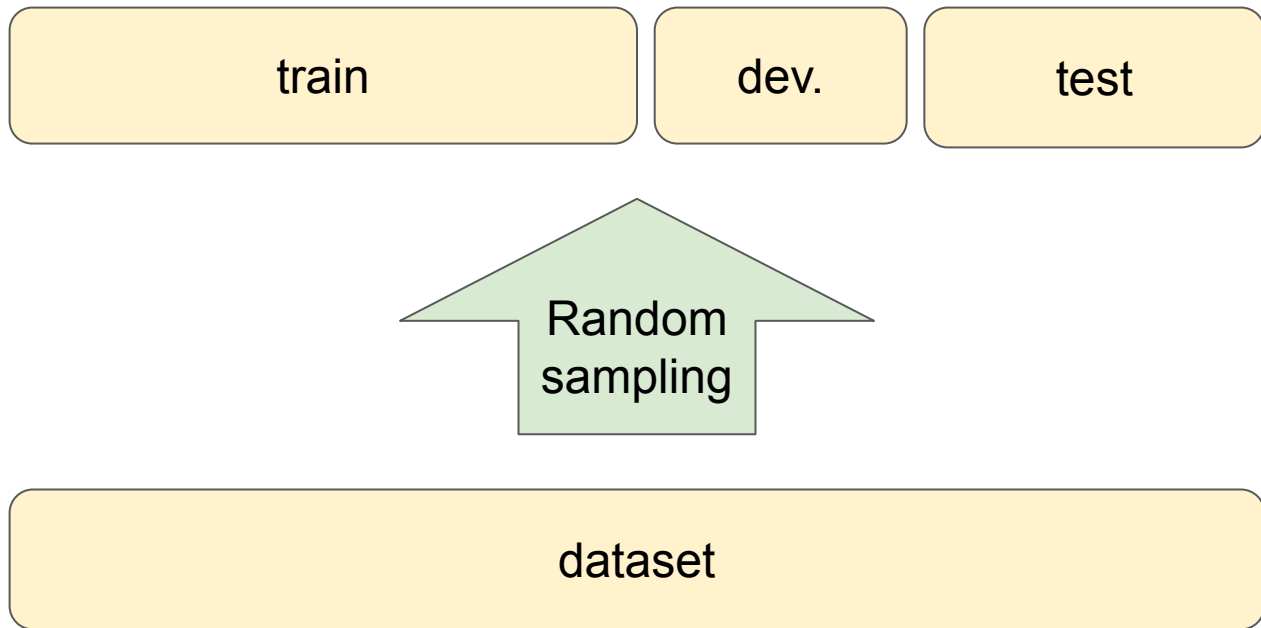
General definition of bias in science and engineering:



Taken from: <https://en.wikipedia.org/wiki/Bias> (February 15th, 2022)

Why don't we find out about biases in testing?

Why don't we find out about biases in testing?

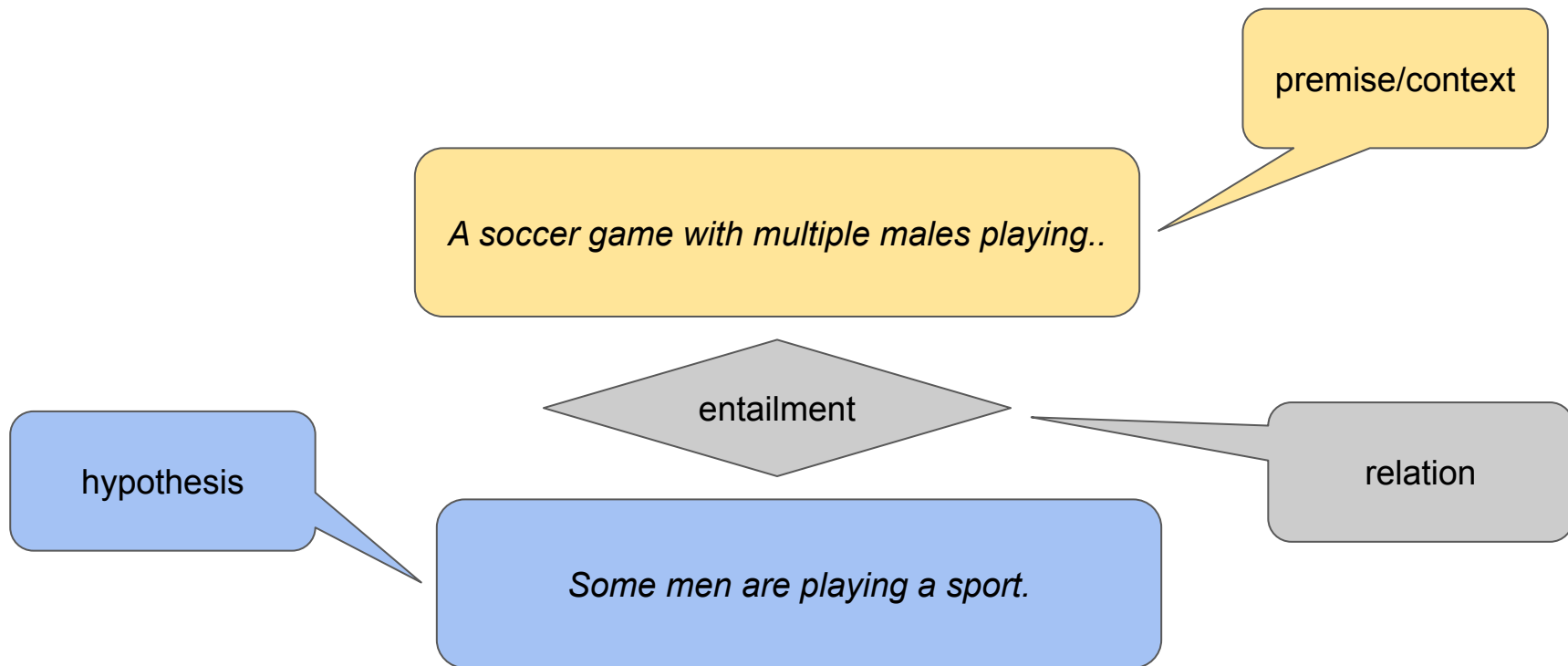


Why is bias a problem?

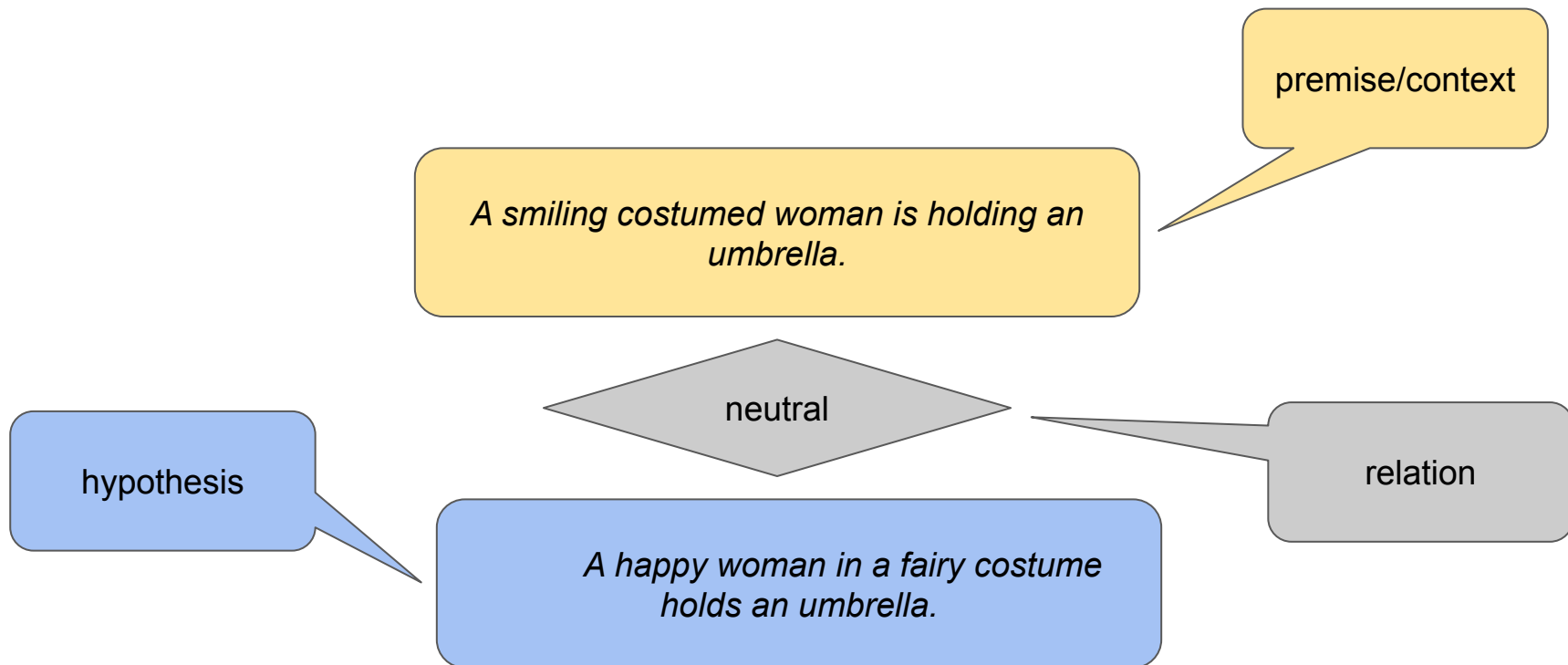
Why is bias a problem?

- New distribution (any real-world use-case)
- Potentially ethics (depending on the bias)
- Science (hypothesis-testing)

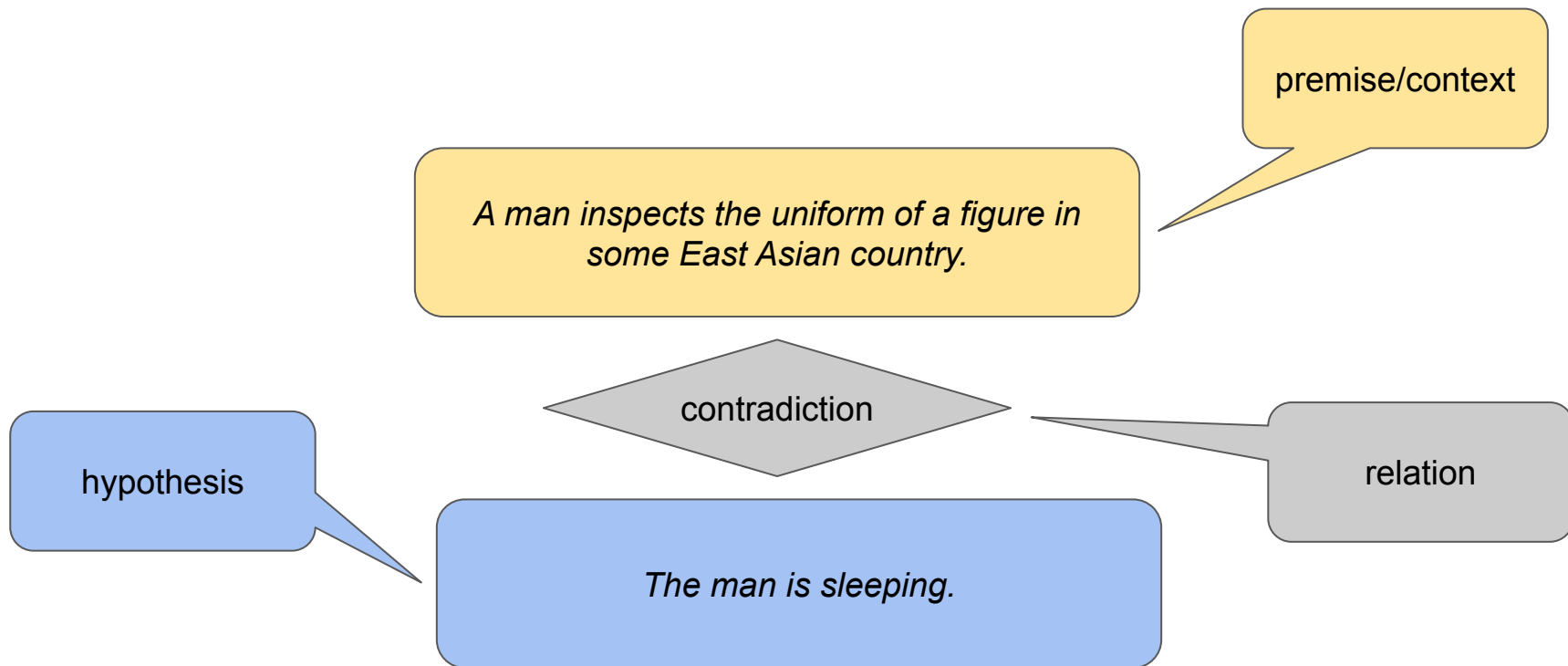
Example 1: A shortcut for NLI



Example 1: A shortcut for NLI



Example 1: A shortcut for NLI



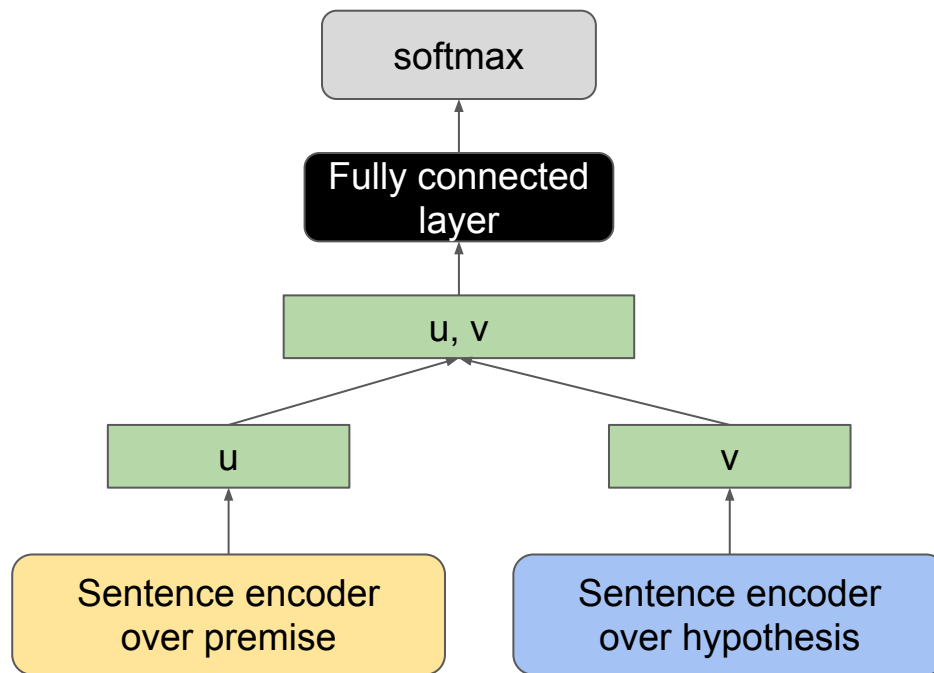
Example 1: A shortcut for NLI

Purpose of the task:

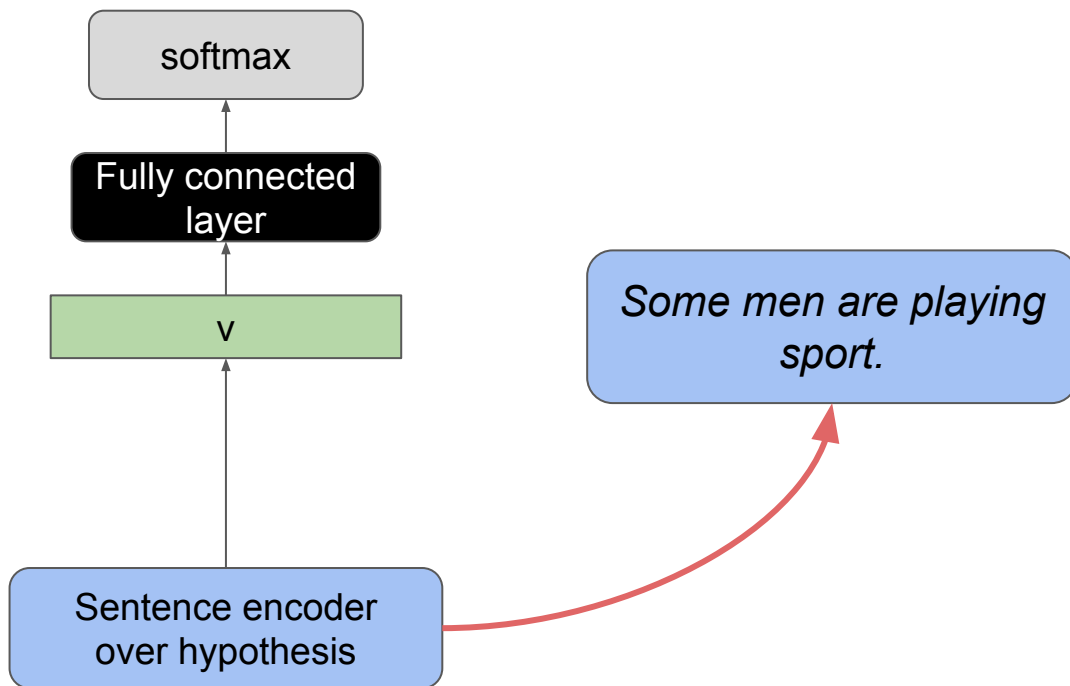
- General means of assessing how well a system can understand natural language
- High accuracy → “understood language”
- High accuracy → “understood logical relationships”

Example 1: A shortcut for NLI

Poliak et al. 2018



Example 1: A shortcut for NLI



Example 1: A shortcut for NLI

Step 1: Train state-of-the art model on hypotheses only

Outcome:

- For 6 out of 10 datasets, hypotheses-only clearly outperforms a majority baseline
- For one dataset, hypothesis-only also outperforms the best reported results

Example 1: A shortcut for NLI

			SNLI					
Word	Score	Freq	Word	Score	Freq	Word	Score	Freq
instrument	0.90	20	tall	0.93	44	sleeping	0.88	108
touching	0.83	12	competition	0.88	24	driving	0.81	53
least	0.90	10	because	0.83	23	Nobody	1.00	52
Humans	0.88	8	birthday	0.85	20	alone	0.90	50
transportation	0.86	7	mom	0.82	17	cat	0.84	49
speaking	0.86	7	win	0.88	16	asleep	0.91	43
screen	0.86	7	got	0.81	16	no	0.84	31
arts	0.86	7	trip	0.93	15	empty	0.93	28
activity	0.86	7	tries	0.87	15	eats	0.83	24
opposing	1.00	5	owner	0.87	15	sleeps	0.95	20

(a) entailment (b) neutral (c) contradiction

Example 2: Gender bias in coreference resolution

A man and his son get into a terrible car crash. The father dies, and the boy is badly injured. In the hospital, the surgeon looks at the patient and exclaims, “I can’t operate on this boy, he’s my son!”

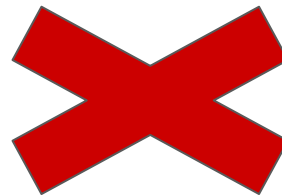
How can this be?

(Rudinger et al. 2018, p. 8)

Example 2: Gender bias in coreference resolution

What is coreference resolution? How does it relate to gender?

The **surgeon** couldn't operate on **her** patient: **it** was **her** son!



Example 2: Gender bias in coreference resolution

The technician told the customer that she could pay with cash.

The technician told the customer that she had completed the repair.

Dataset:

https://github.com/rudinger/winogender-schemas/blob/master/data/all_sentences.tsv

Example 2: Gender bias in coreference resolution

Tested three systems:
rule-based, statistics, neural

- Performance for all systems is low
- Performance on examples that go against tendencies is even lower

“Gotcha” sentences: “pronoun gender does not match the occupation’s majority gender (BLS) if OCCUPATION is the correct answer”

System	“Gotcha”?	Female	Male
RULE	no	38.3	51.7
	yes	10.0	37.5
STAT	no	50.8	61.7
	yes	45.8	40.0
NEURAL	no	50.8	49.2
	yes	36.7	46.7

ChatGPT and the surgeon



You

A man and his mother are in a car accident. The mother sadly dies. The man is rushed to the ER. When the doctor sees him, he says, "I can't operate on this man. He's my son!"

How is this possible?



ChatGPT

The doctor is the man's other parent—his mother, indicating that the doctor is a woman. This riddle plays on common assumptions about professions and gender roles.

What do systems learn?

- What do models learn linguistic structure?
- What do models learn about a task?

(Behavioral) Interpretability

Interpretability Goals

- What is the **internal structure** of a model?
- How does it **behave on different data**?
- Why does it make certain **decisions**?
- When does it **fail/succeed**?

Find general
tendencies rather than
individual examples

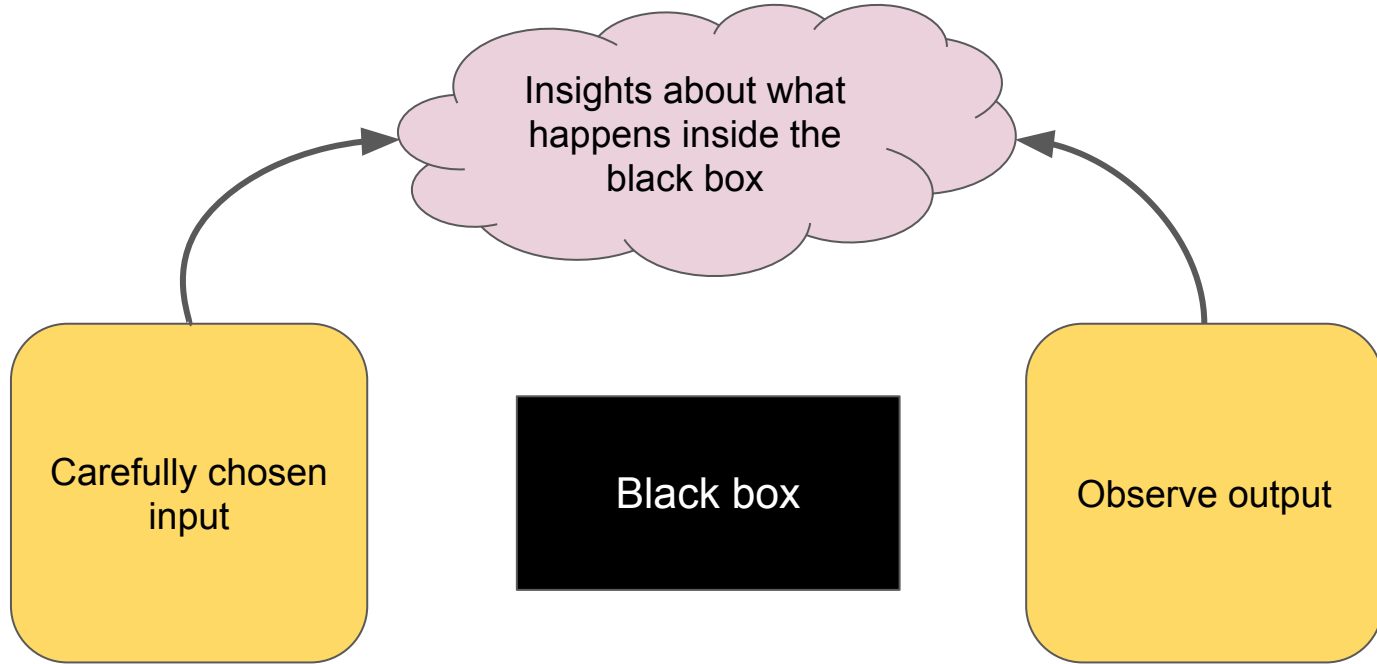
(Based on ACL 2020 Interpretability tutorial)

Interpretability vs Explainability?

Some Interpretability Approaches

- **Behavioral methods:** given an input, what output does the model provide?
- **Probing methods:** What is the internal structure of the model representation?
- **Feature-importance:** which features in the input are important for the prediction?
- **Intervention methods:** If I change part of the model, how does the model behavior change?

Behavior Interpretability



How is this different from evaluation?

Standard test sets	Challenge sets
<ul style="list-style-type: none">● Often taken from corpora - natural distribution of linguistic phenomena● Same distribution as the training set	<ul style="list-style-type: none">● Target specific linguistic phenomena● Systematic● Long-tail-phenomena, specific, difficult examples● (Partly) include negative examples/ill-formed data

Challenge sets today

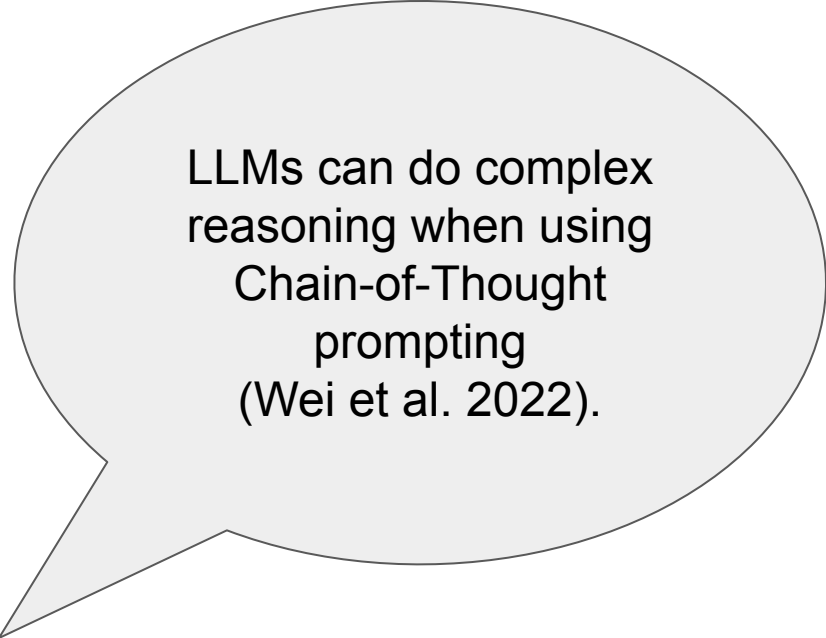
Contrastive Translation Pairs for English to German translation (Sennrich 2017)

Check how often the model assigns the higher probability to the correct translation

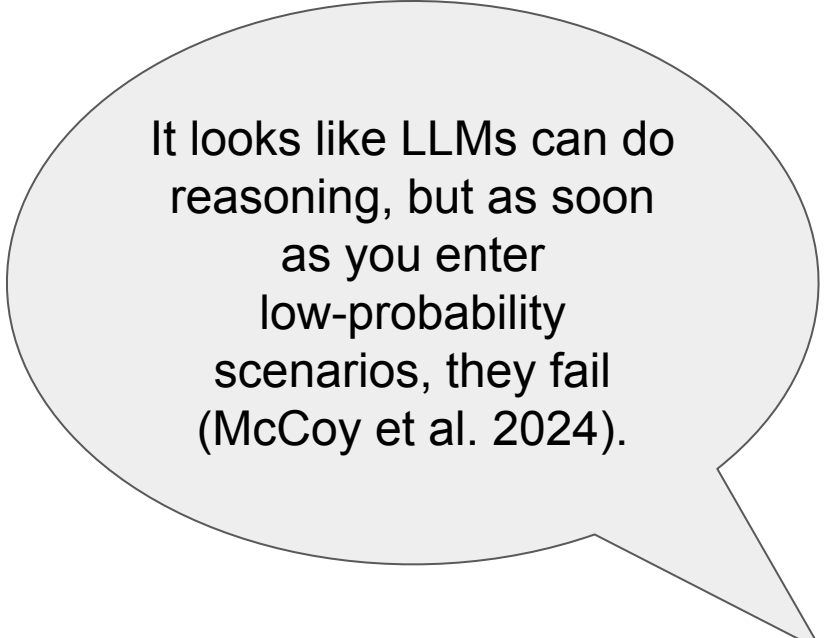
category	English	German (correct)	German (contrastive)
NP agreement	[...] of the American Congress	[...] des amerikanischen Kongresses	* [...] der amerikanischen Kongresses
subject-verb agr.	[...] that the plan will be approved	[...], dass der Plan verabschiedet wird	* [...], dass der Plan verabschiedet werden
separable verb particle	he is resting	er ruht sich aus	* er ruht sich an
polarity	the timing [...] is uncertain	das Timing [...] ist unsicher	das Timing [...] ist sicher
transliteration	Mr. Ensign's office	Senator Ensigns Büro	Senator Enisgns Büro

Table 1: Example contrastive translations pair for each error category.

What can LLMs do?



LLMs can do complex reasoning when using Chain-of-Thought prompting (Wei et al. 2022).



It looks like LLMs can do reasoning, but as soon as you enter low-probability scenarios, they fail (McCoy et al. 2024).

Shifted Ciphers Example

Shift cipher: Task probability

Common task: Rot-13. Decode the message by shifting each letter thirteen positions backward in the alphabet.

Input: Jryy, vs gurl qba'g pbzr, fb or vg.

Correct: Well, if they don't come, so be it.

✓ **GPT-4:** Well, if they don't come, so be it.

Uncommon task: Rot-2. Decode the message by shifting each letter two positions backward in the alphabet.

Input: Ygnn, kh vjga fqp'v eqog, uq dg kv.

Correct: Well, if they don't come, so be it.

✗ **GPT-4:** Well, if there isn't cake, to be it.

ARN: Analogical Reasoning on Narratives

Zhivar Sourati, Filip Ilievski,
Pia Sommerauer, Yifan Jiang

TACL 2024

Can LLMs identify analogies between stories?

After months of rigorous training and pushing through excruciating pain, Emily crossed the finish line of the marathon, triumphantly claiming her well-deserved medal.

No pain no gain

Jacob knew that mastering the guitar requires countless hours of sore fingers and frustrating practice, but he embraced pain with determination, knowing that in the end, winning the competition would be worth every ache.

Different types of textual analogies

Surface similarities

After months of rigorous **training** and pushing through excruciating pain, **Emily** finally crossed the finish line of the **marathon**, triumphantly claiming her well-deserved medal.

training ⇔ **training**

Emily ⇔ **Emily**

marathon ⇔ **marathon**

Emily loved training for **marathon**, but couldn't find friends until she realized that all people **training** there shared the same interest and passion and she got friends with them.

Diagnostic experiments

query

After months of rigorous training and pushing through excruciating pain, Emily crossed the finish line of the marathon, triumphantly claiming her well-deserved medal.

No pain no gain

candidate

Jacob knew that mastering the guitar requires countless hours of sore fingers and frustrating practice, but he embraced pain with determination, knowing that in the end, winning the competition would be worth every ache.

No pain no gain

candidate

Emily loved training for a marathon, but couldn't find friends until she realized that all people training there shared the same interest and passion and she became friends with them.

Birds of a feather flock together

Diagnostic data

Analogy = same proverb

Disanalogy = different proverb

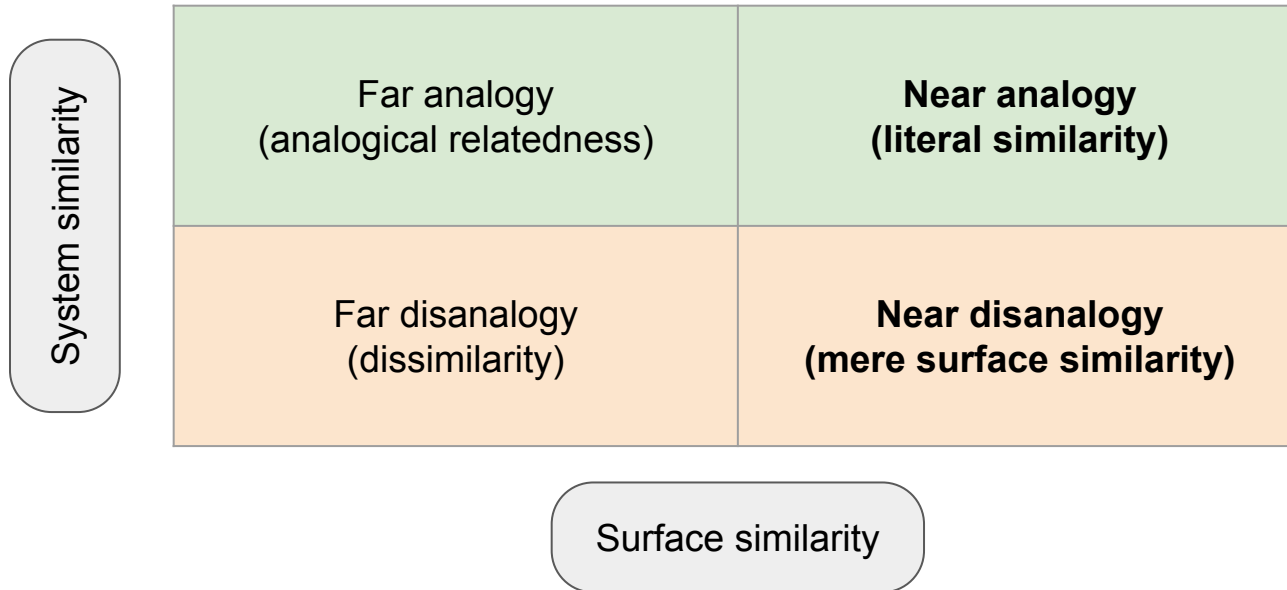
Far analogy (analogical relatedness)	Near analogy (literal similarity)
Far disanalogy (dissimilarity)	Near disanalogy (mere surface similarity)

Diagnostic data

Far analogy (analogical relatedness)	Near analogy (literal similarity)
Far disanalogy (dissimilarity)	Near disanalogy (mere surface similarity)

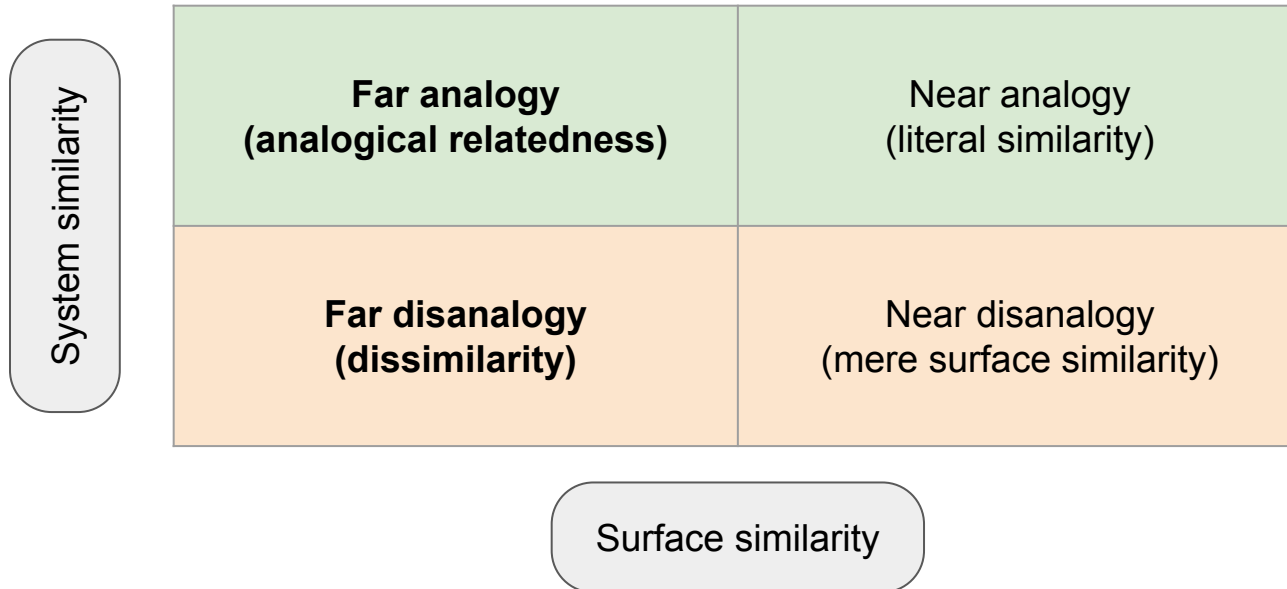
Task Partition	(near, far)	(near, near)	(far, far)	(far, near)	<i>Avg.</i>
SBERT	84.3	71.5	12.6	1.00	42.3
GPT3.5	88.1	81.3	50.4	21.7	60.3
GPT4.0	94.0	92.5	57.1	29.1	68.1
UnifiedQA-L	43.2	49.1	47.2	50.5	47.5
UnifiedQA-3B	66.4	68.4	47.3	44.4	56.6
UnifiedQA-11B	60.7	61.2	54.8	74.6	62.8
Llama-2-7B	63.4	58.0	50.1	43.1	53.7
Llama-2-13B	80.9	81.4	44.5	35.4	60.5
FlanT5-L	84.5	80.3	41.4	14.4	55.1
FlanT5-xl	78.9	68.3	44.7	21.1	53.2
FlanT5-xxl	89.9	81.3	51.1	35.6	64.4
Macaw-11B	88.0	84.6	42.1	35.8	62.6
<i>Avg.</i>	76.9	73.2	45.3	33.9	57.3
human	98.6	97.2	96.8	91.4	96.0

Diagnostic data



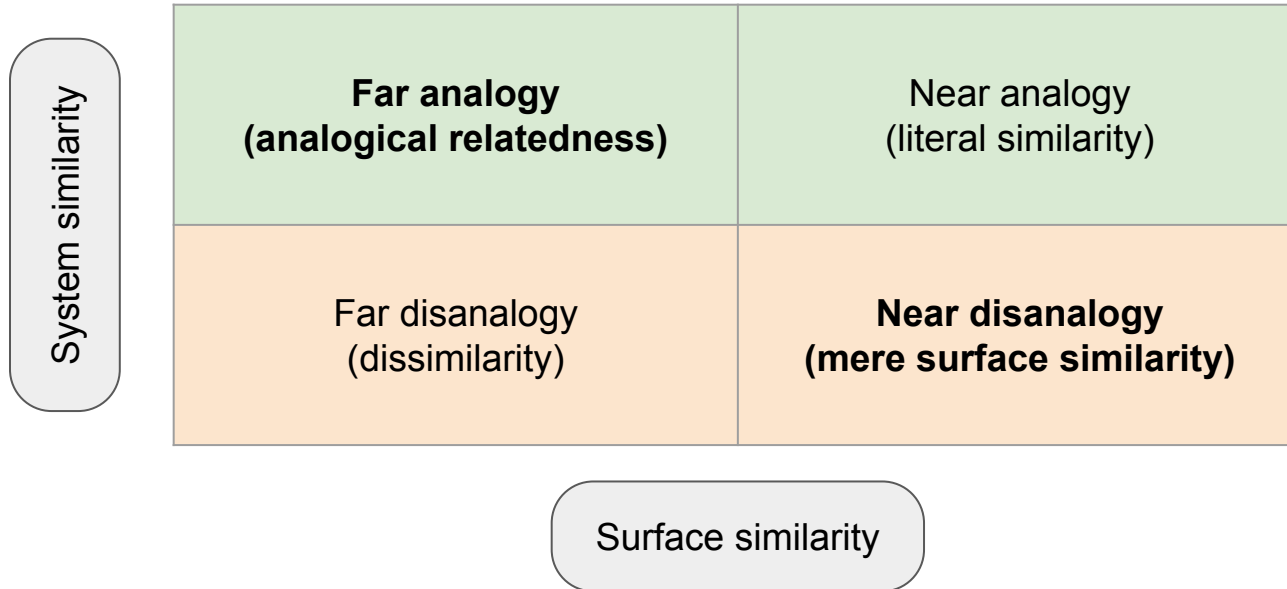
Task Partition	(near, far)	(near, near)	(far, far)	(far, near)	<i>Avg.</i>
SBERT	84.3	71.5	12.6	1.00	42.3
GPT3.5	88.1	81.3	50.4	21.7	60.3
GPT4.0	94.0	92.5	57.1	29.1	68.1
UnifiedQA-L	43.2	49.1	47.2	50.5	47.5
UnifiedQA-3B	66.4	68.4	47.3	44.4	56.6
UnifiedQA-11B	60.7	61.2	54.8	74.6	62.8
Llama-2-7B	63.4	58.0	50.1	43.1	53.7
Llama-2-13B	80.9	81.4	44.5	35.4	60.5
FlanT5-L	84.5	80.3	41.4	14.4	55.1
FlanT5-xl	78.9	68.3	44.7	21.1	53.2
FlanT5-xxl	89.9	81.3	51.1	35.6	64.4
Macaw-11B	88.0	84.6	42.1	35.8	62.6
<i>Avg.</i>	76.9	73.2	45.3	33.9	57.3
human	98.6	97.2	96.8	91.4	96.0

Diagnostic data



Task Partition	(near, far)	(near, near)	(far, far)	(far, near)	<i>Avg.</i>
SBERT	84.3	71.5	12.6	1.00	42.3
GPT3.5	88.1	81.3	50.4	21.7	60.3
GPT4.0	94.0	92.5	57.1	29.1	68.1
UnifiedQA-L	43.2	49.1	47.2	50.5	47.5
UnifiedQA-3B	66.4	68.4	47.3	44.4	56.6
UnifiedQA-11B	60.7	61.2	54.8	74.6	62.8
Llama-2-7B	63.4	58.0	50.1	43.1	53.7
Llama-2-13B	80.9	81.4	44.5	35.4	60.5
FlanT5-L	84.5	80.3	41.4	14.4	55.1
FlanT5-xl	78.9	68.3	44.7	21.1	53.2
FlanT5-xxl	89.9	81.3	51.1	35.6	64.4
Macaw-11B	88.0	84.6	42.1	35.8	62.6
<i>Avg.</i>	76.9	73.2	45.3	33.9	57.3
human	98.6	97.2	96.8	91.4	96.0

Diagnostic data



Task Partition	(near, far)	(near, near)	(far, far)	(far, near)	<i>Avg.</i>
SBERT	84.3	71.5	12.6	1.00	42.3
GPT3.5	88.1	81.3	50.4	21.7	60.3
GPT4.0	94.0	92.5	57.1	29.1	68.1
UnifiedQA-L	43.2	49.1	47.2	50.5	47.5
UnifiedQA-3B	66.4	68.4	47.3	44.4	56.6
UnifiedQA-11B	60.7	61.2	54.8	74.6	62.8
Llama-2-7B	63.4	58.0	50.1	43.1	53.7
Llama-2-13B	80.9	81.4	44.5	35.4	60.5
FlanT5-L	84.5	80.3	41.4	14.4	55.1
FlanT5-xl	78.9	68.3	44.7	21.1	53.2
FlanT5-xxl	89.9	81.3	51.1	35.6	64.4
Macaw-11B	88.0	84.6	42.1	35.8	62.6
<i>Avg.</i>	76.9	73.2	45.3	33.9	57.3
human	98.6	97.2	96.8	91.4	96.0

Results

- **Analogy near, distractor far:** high performance (clearly above baseline)
- **Analogy near, distractor near:** mostly above baseline
- **Analogy far, distractor far:** around baseline
- **Analogy far, distractor near:** below baseline*

*One exception: UnifiedQA-11B performed clearly above baseline but performed much lower on the other tasks.

Takeaways

- Far analogies are extremely difficult for models
- Lower-order mappings help the models
- GPTs are good at near analogies
- QA task models are good at far analogies, but not at near analogies
- Models do not seem to distinguish between surface similarity and analogy

Insights from a carefully organized dataset.

Summary

- NLP and supervised learning
- Evaluation
- Biases and shortcuts
- Behavioral testing
- Example: Analogies between stories

Hands-on Session

Create your own challenge dataset

1. Pick a language, use-case, topic, phenomenon, etc. you find interesting (in groups of ~4)
2. Think of capabilities (what should a system be able to do)
3. Create small sets of test examples (can be manually, by looking for examples online, by using LLMs)
4. If time and feasible: test a model (SpaCy, NLTK, Coreferee, ChatGPT, anything you find)
5. Report the results back to the group

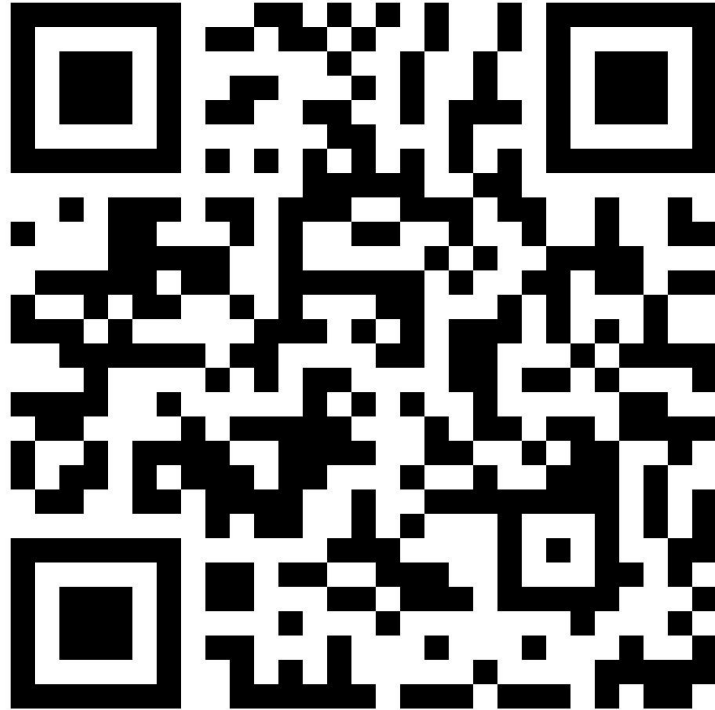
Labels: positive, negative, or neutral; INV: same pred. (INV) after removals/ additions; DIR: sentiment should not decrease (↑) or increase (↓)

Test <i>TYPE</i> and Description	Failure Rate (%)					Example test cases & expected behavior	
	☐	G	a	👤	RoB		
Vocab.+POS	<i>MFT</i> : Short sentences with neutral adjectives and nouns	0.0	7.6	4.8	94.6	81.8	The company is Australian. neutral That is a private aircraft. neutral
	<i>MFT</i> : Short sentences with sentiment-laden adjectives	4.0	15.0	2.8	0.0	0.2	That cabin crew is extraordinary. pos I despised that aircraft. neg
	<i>INV</i> : Replace neutral words with other neutral words	9.4	16.2	12.4	10.2	10.2	@Virgin should I be concerned that → when I'm about to fly ... INV @united the → our nightmare continues... INV
	<i>DIR</i> : Add positive phrases, fails if sent. goes down by > 0.1	12.6	12.4	1.4	0.2	10.2	@SouthwestAir Great trip on 2672 yesterday... You are extraordinary. ↑ @AmericanAir AA45 ... JFK to LAS. You are brilliant. ↑
	<i>DIR</i> : Add negative phrases, fails if sent. goes up by > 0.1	0.8	34.6	5.0	0.0	13.2	@USAirways your service sucks. You are lame. ↓ @JetBlue all day. I abhor you. ↓
Robust.	<i>INV</i> : Add randomly generated URLs and handles to tweets	9.6	13.4	24.8	11.4	7.4	@JetBlue that selfie was extreme. @pi9QDK INV @united stuck because staff took a break? Not happy 1K.... https://t.co/PWK1jb INV
	<i>INV</i> : Swap one character with its neighbor (typo)	5.6	10.2	10.4	5.2	3.8	@JetBlue → @JeBtlue I cri INV @SouthwestAir no thanks → thakns INV
NER	<i>INV</i> : Switching locations should not change predictions	7.0	20.8	14.8	7.6	6.4	@JetBlue I want you guys to be the first to fly to # Cuba → Canada... INV @VirginAmerica I miss the #nerdbird in San Jose → Denver INV
	<i>INV</i> : Switching person names should not change predictions	2.4	15.1	9.1	6.6	2.4	...Airport agents were horrendous. Sharon → Erin was your saviour INV @united 8602947, Jon → Sean at http://t.co/58tuTgli0D , thanks. INV

Checklist metric

Failure rate: *percentage of wrong answers*

More specific instructions and links



Time-management

15:45-16:15	Break; think and talk about ideas, try to form groups
16:15-16:25	Finalize groups, decide on use-case, language, phenomenon (moderated by Pia)
16:25-17:30	Hack! Pia and Urja will help.
17:30-18:00	Mini-presentations - report back (1 slide max)

Thank you!

pia.sommerauer@vu.nl

References

Belinkov, Y. and Glass, J., 2019. Analysis methods in neural language processing: A survey. Transactions of the Association for Computational Linguistics, 7, pp.49-72.

Bender, E.M. and Koller, A., 2020, July. Climbing towards NLU: On meaning, form, and understanding in the age of data. In Proceedings of the 58th annual meeting of the association for computational linguistics (pp. 5185-5198).

Isahara, H., 1995. JEIDA's test-sets for quality evaluation of MT systems. In Proceedings of Machine Translation Summit V.

MacCartney, B., 2009. Natural language inference. Stanford University.

McCoy, R.T., Yao, S., Friedman, D., Hardy, M. and Griffiths, T.L., 2023. Embers of Autoregression: Understanding Large Language Models Through the Problem They are Trained to Solve. arXiv preprint arXiv:2309.13638.

Poliak, A., Naradowsky, J., Haldar, A., Rudinger, R. and Van Durme, B., 2018, June. Hypothesis Only Baselines in Natural Language Inference. In Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics (pp. 180-191).

Ribeiro, M.T., Wu, T., Guestrin, C. and Singh, S., 2020, January. Beyond Accuracy: Behavioral Testing of NLP Models with CheckList. In Annual Meeting of the Association for Computational Linguistics.

Rudinger, R., Naradowsky, J., Leonard, B. and Van Durme, B., 2018. Gender Bias in Coreference Resolution. In Proceedings of NAACL-HLT (pp. 8-14).

Sennrich, R., 2017, April. How Grammatical is Character-level Neural Machine Translation? Assessing MT Quality with Contrastive Translation Pairs. In Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers (pp. 376-382).

Wei, J., Wang, X., Schuurmans, D., Bosma, M., Xia, F., Chi, E., Le, Q.V. and Zhou, D., 2022. Chain-of-thought prompting elicits reasoning in large language models. Advances in neural information processing systems, 35, pp.24824-24837.