Evaluation and Benchmarking

He He

NEW YORK UNIVERSITY

November 19, 2024

Logistics

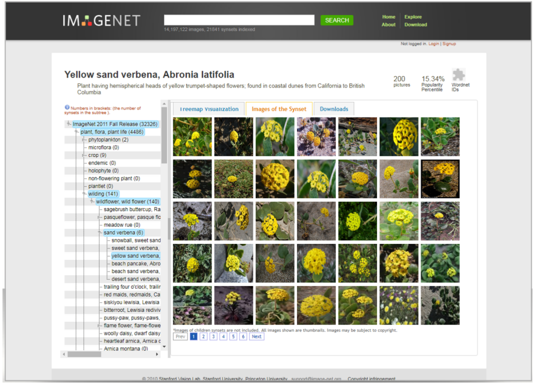
Plan for the rest of the semester

- 11/27: guest lecture on LLM reasoning
- Thanksgiving
- 12/4 and 12/5: project presentation
- No lecture in the last week (legislative Friday)
- Use office hours for any last-minute project help
- 12/12: project report due

Influence of benchmarks in AI

- Machine learning drives the progress.
- Benchmarks set the direction.
- Key questions answered by a benchmark:
  - What tasks are important and within reach now?
  - Where do we stand now?

Example: ImageNet [Deng et al., 2009]



- Over 14M labeled images
- Data collection leveraged image search and crowdsourcing (Amazon Mechanical Turk)
  scale over precision
- Led to the community-wide ILSVRC challenge
- The message:
  Let's learn from lots of data!

Breakthrough of deep learning established by ImageNet

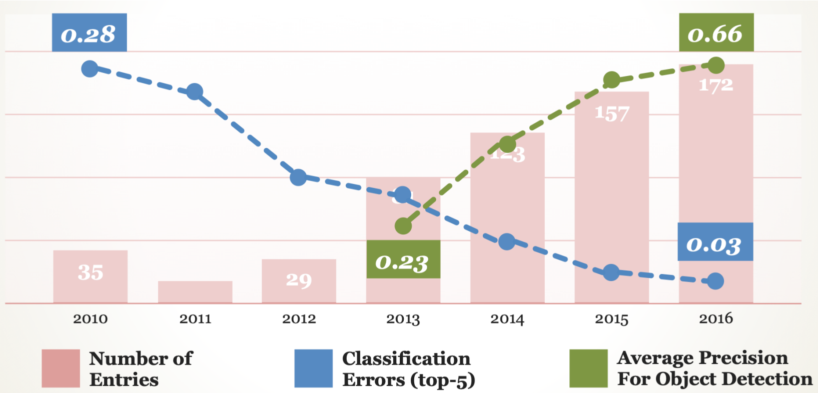


Figure: From Fei-Fei Li's slides

- AlexNet Krizhevsky et al., 2012 achieved top-1 error rate in ILSVRC 2010.
- The result sparked renewed interests in neural netowrks.

Example: GLUE [Wang et al., 2019]

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

- A collection of selected NLU datasets
- BERT suceeded by achieving 7.7 point improvement on GLUE
- The message: *Let's build general NLU models that adapt to many tasks*

Challenges in evaluating LLMs

What are challenges in evaluating LLMs like ChatGPT?

Challenges in evaluating LLMs

What are challenges in evaluating LLMs like ChatGPT?

- Many use cases (coding, writing, knowledge retrieval etc.)
- Open-ended, long-form generation
- Data contamination: how do we know if our test data is unseen?

Evaluate LLMs as a language model

PPL is often correlated with downstream performance

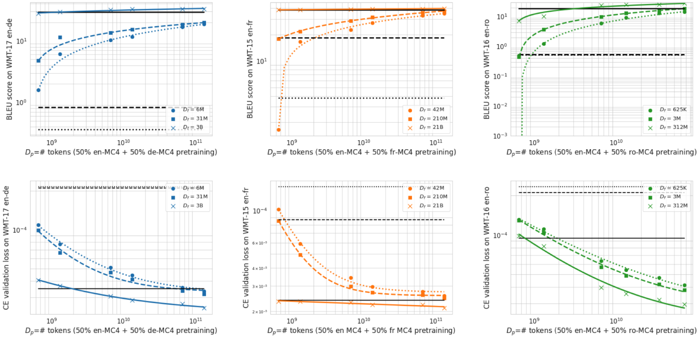


Figure: [Isik et al., 2024]

But the increase in task performance may not be smooth and PPL depends on data and tokenizer

Expand the evaluation tasks

Massive multitask language understanding (MMLU)

Microeconomics

One of the reasons that the government discourages and regulates monopolies is that
(A) producer surplus is lost and consumer surplus is gained. ✗
(B) monopoly prices ensure productive efficiency but cost society allocative efficiency. ✗
(C) monopoly firms do not engage in significant research and development. ✗
(D) consumer surplus is lost with higher prices and lower levels of output. ✓

Figure 3: Examples from the Microeconomics task.

Conceptual Physics

When you drop a ball from rest it accelerates downward at 9.8 m/s^2. If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s^2 ✓
(B) more than 9.8 m/s^2 ✗
(C) less than 9.8 m/s^2 ✗
(D) Cannot say unless the speed of throw is given. ✗

College Mathematics

In the complex z-plane, the set of points satisfying the equation $z^2 = |z|^2$ is a
(A) pair of points ✗
(B) circle ✗
(C) half-line ✗
(D) line ✓

Figure 4: Examples from the Conceptual Physics and College Mathematics STEM tasks.

Figure: [Hendrycks et al., 2021]

Expand the evaluation tasks

GSM8K: curated math word problems

Problem: Beth bakes 4, 2 dozen batches of cookies in a week. If these cookies are shared amongst 16 people equally, how many cookies does each person consume?
Solution: Beth bakes 4 2 dozen batches of cookies for a total of 4*2 = <<4*2=8>>8 dozen cookies
There are 12 cookies in a dozen and she makes 8 dozen cookies for a total of 12*8 = <<12*8=96>>96 cookies
She splits the 96 cookies equally amongst 16 people so they each eat 96/16 = <<96/16=6>>6 cookies
Final Answer: 6

Problem: Mrs. Lim milks her cows twice a day. Yesterday morning, she got 68 gallons of milk and in the evening, she got 82 gallons. This morning, she got 18 gallons fewer than she had yesterday morning. After selling some gallons of milk in the afternoon, Mrs. Lim has only 24 gallons left. How much was her revenue for the milk if each gallon costs $3.50?
Mrs. Lim got 68 gallons - 18 gallons = <<68-18=50>>50 gallons this morning.
So she was able to get a total of 68 gallons + 82 gallons + 50 gallons = <<68+82+50=200>>200 gallons.
She was able to sell 200 gallons - 24 gallons = <<200-24=176>>176 gallons.
Thus, her total revenue for the milk is $3.50/gallon x 176 gallons = $<<3.50*176=616>>616.
Final Answer: 616

Problem: Tina buys 3 12-packs of soda for a party. Including Tina, 6 people are at the party. Half of the people at the party have 3 sodas each, 2 of the people have 4, and 1 person has 5. How many sodas are left over when the party is over?
Solution: Tina buys 3 12-packs of soda, for 3*12= <<3*12=36>>36 sodas
6 people attend the party, so half of them is 6/2= <<6/2=3>>3 people
Each of those people drinks 3 sodas, so they drink 3*3=<<3*3=9>>9 sodas
Two people drink 4 sodas, which means they drink 2*4=<<4*2=8>>8 sodas
With one person drinking 5, that brings the total drank to 5+9+8+3= <<5+9+8+3=25>>25 sodas
As Tina started off with 36 sodas, that means there are 36-25=<<36-25=11>>11 sodas left
Final Answer: 11

Figure 1: Three example problems from GSM8K. Calculation annotations are highlighted in red.

Figure: [Cobbe et al., 2021]

Expand the evaluation tasks

HumanEval: generating code given docstrings; human-written solution and unit tests

```python
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]
```

```python
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)
```

Figure: [Chen et al., 2021]

User preference

ChatbotArena: live benchmark based on head-to-head comparison

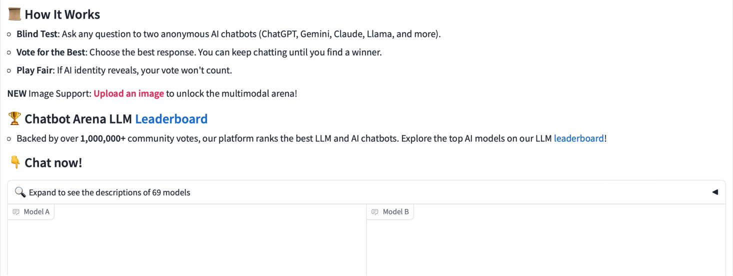


Figure: https://lmarena.ai

User preference

ChatbotArena: rank LLMs based on user preference

Rank	Model	Elo Rating	Description
1	🥇 vicuna-13b	1169	a chat assistant fine-tuned from LLaMA on user-shared conversations by LMSYS
2	🥈 koala-13b	1082	a dialogue model for academic research by BAIR
3	🥉 oasst-pythia-12b	1065	an Open Assistant for everyone by LAION
4	alpaca-13b	1008	a model fine-tuned from LLaMA on instruction-following demonstrations by Stanford
5	chatglm-6b	985	an open bilingual dialogue language model by Tsinghua University
6	fastchat-t5-3b	951	a chat assistant fine-tuned from FLAN-T5 by LMSYS
7	dolly-v2-12b	944	an instruction-tuned open large language model by Databricks
8	llama-13b	932	open and efficient foundation language models by Meta
9	stablelm-tuned-alpha-7b	858	Stability AI language models

Ranking LLMs

- Average win rate: need data for every pair - expensive!

Ranking LLMs

- Average win rate: need data for every pair - expensive!
- Elo rating: supports sequential updates

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{1}$$

$$R'_A = R_A + K \cdot (S_A - E_A) \tag{2}$$

Ranking LLMs

- Average win rate: need data for every pair - expensive!
- Elo rating: supports sequential updates

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{1}$$

$$R'_A = R_A + K \cdot (S_A - E_A) \tag{2}$$

  - E_A: expected win rate

Ranking LLMs

- Average win rate: need data for every pair - expensive!
- Elo rating: supports sequential updates

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{1}$$

$$R'_A = R_A + K \cdot (S_A - E_A) \tag{2}$$

  - E_A: expected win rate
  - S_A: actual win (1) or lose (0)

Ranking LLMs

- Average win rate: need data for every pair - expensive!
- Elo rating: supports sequential updates

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{1}$$

$$R'_A = R_A + K \cdot (S_A - E_A) \tag{2}$$

  - E_A: expected win rate
  - S_A: actual win (1) or lose (0)
  - S'_A: new rating

Ranking LLMs

- Average win rate: need data for every pair - expensive!
- Elo rating: supports sequential updates

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \quad (1)$$

$$R'_A = R_A + K \cdot (S_A - E_A) \quad (2)$$

  - E_A: expected win rate
  - S_A: actual win (1) or lose (0)
  - S'_A: new rating
- Ratings can have large variance though

Ranking LLMs

- Average win rate: need data for every pair - expensive!
- Elo rating: supports sequential updates

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}} \tag{1}$$

$$R'_A = R_A + K \cdot (S_A - E_A) \tag{2}$$

  - E_A: expected win rate
  - S_A: actual win (1) or lose (0)
  - S'_A: new rating
- Ratings can have large variance though
- Also costly!

LLM as a judge

AlpacaEval: use LLMs to simulate human preference

- 1. For each instruction: generate an output by baseline and model to eval
- 2. Ask GPT-4 the probability that the model's output is better
- 3. (AlpacaEval LC) Reweight win-probability based on length of outputs
- 4. Average win-probability => win rate

AlpacaEval Leaderboard

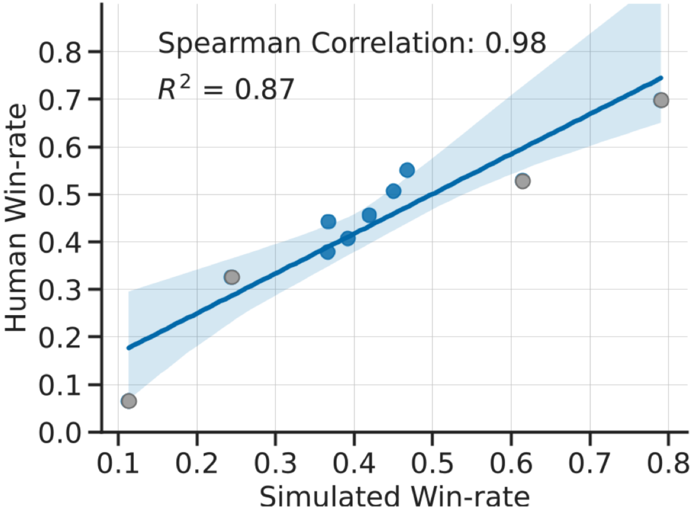
Model Name	LC Win Rate	Win Rate
GPT-4 Turbo (04/09)	55.0%	46.1%
GPT-4 Preview (11/06)	50.0%	50.0%
Claude 3 Opus (02/29)	40.5%	29.1%
GPT-4	38.1%	23.6%

35

Figure: From Yann Dubois' slides

LLM as a judge

High correlation with human



LLM as a judge

Spurious correlation between length and rating: increasing length can improve model rating!

	AlpacaEval			Length-controlled AlpacaEval		
	concise	standard	verbose	concise	standard	verbose
gpt4_1106_preview	22.9	50.0	64.3	41.9	50.0	51.6
Mixtral-8x7B-Instruct-v0.1	13.7	18.3	24.6	23.0	23.7	23.2
gpt4_0613	9.4	15.8	23.2	21.6	30.2	33.8
claude-2.1	9.2	15.7	24.4	18.2	25.3	30.3
gpt-3.5-turbo-1106	7.4	9.2	12.8	15.8	19.3	22.0
alpaca-7b	2.0	2.6	2.9	4.5	5.9	6.8

Control for length: estimating contribution from different factors (model, length, instruction)

Evaluating models beyond accuracy

Linguists, cognitive scientists: **interpretability**

- How does the model make predictions? Is it human-like?

Evaluating models beyond accuracy

Linguists, cognitive scientists: **interpretability**

- How does the model make predictions? Is it human-like?

Practitioners: **efficiency**, **robustness**

- How much resource does it take for training and inference?
- Does it handle typos/dialects/etc. well?

Evaluating models beyond accuracy

Linguists, cognitive scientists: **interpretability**

- How does the model make predictions? Is it human-like?

Practitioners: **efficiency**, **robustness**

- How much resource does it take for training and inference?
- Does it handle typos/dialects/etc. well?

Product managers: **calibration**, **explainability**

- Can the model indicate its uncertainty about a prediction?
- Can it explain its predictions?

Evaluating models beyond accuracy

Linguists, cognitive scientists: **interpretability**

- How does the model make predictions? Is it human-like?

Practitioners: **efficiency**, **robustness**

- How much resource does it take for training and inference?
- Does it handle typos/dialects/etc. well?

Product managers: **calibration**, **explainability**

- Can the model indicate its uncertainty about a prediction?
- Can it explain its predictions?

Policymakers: **fairness**, **privacy**

- Does the model put certain groups at disadvantage?
- Does it protect user privacy?

Robustness

Our standard setting assumes that the training and test examples are **independent and identically distributed** (iid).

However, this is almost never true in practice. (examples?)

Robustness

Our standard setting assumes that the training and test examples are **independent and identically distributed** (iid).

However, this is almost never true in practice. (examples?)

Reasons for **distribution shifts**:

- Limited training data coverage (often causes domain shift)
  - movie reivew $\rightarrow$ book review, hospital 1 $\rightarrow$ hospital 2
- Temporal change (often causes label shift)
  - fever/flu $\rightarrow$ fever/COVID
  - the US president is ?

Evaluating robustness

Challenge: difficult to come up with a general notion of robustness

- What are non-iid user inputs that are interesting?
- How do we obtain these inputs?
- The answer is often task-dependent.

Evaluating robustness

Challenge: difficult to come up with a general notion of robustness

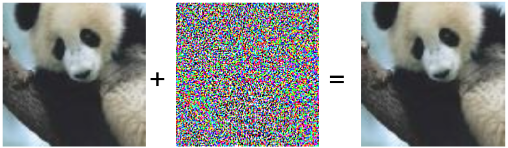
- What are non-iid user inputs that are interesting?
- How do we obtain these inputs?
- The answer is often task-dependent.

Different types of robustness:

- Robustness to **adversarial examples** that are designed to fool the model
- Robustness to **perturbation** of iid examples
- and many more!

Adversarial robustness

Adversarial examples in image recognition:



Panda (60% confidence) | Adversarial Perturbation | Gibbon (99% confidence)

- Find minimal Δx that maximizes $L(x + \Delta x, y)$
- Solve an optimization problem (where Δx is the parameter)

🤔 What are challenges of doing this in NLP?

Adversarial examples in NLP

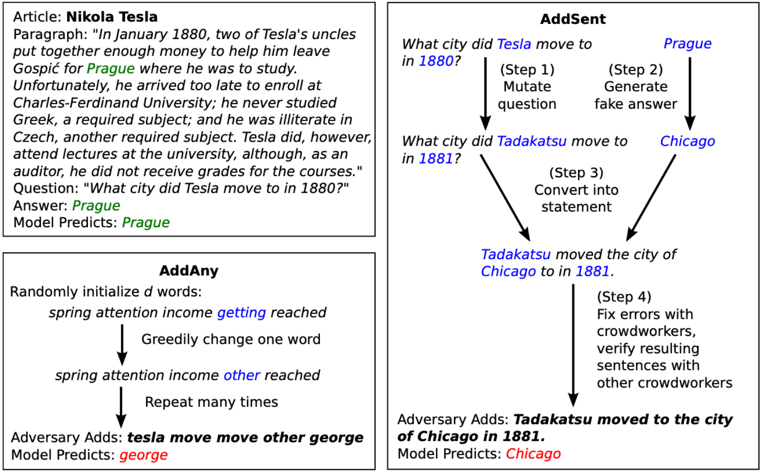
Adversarial examples for reading comprehension [Jia et al., 2017]

Goal: perturb the paragraph+question to change the model's prediction but not the groundtruth

Article: **Nikola Tesla**
Paragraph: "*In January 1880, two of Tesla's uncles put together enough money to help him leave Gospić for Prague where he was to study. Unfortunately, he arrived too late to enroll at Charles-Ferdinand University; he never studied Greek, a required subject; and he was illiterate in Czech, another required subject. Tesla did, however, attend lectures at the university, although, as an auditor, he did not receive grades for the courses.*"
Question: "*What city did Tesla move to in 1880?*"
Answer: *Prague*
Model Predicts: *Prague*

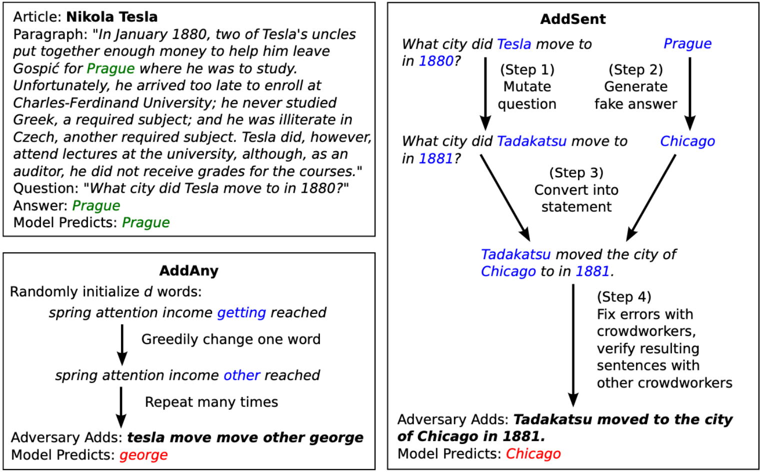
- How to make sure the groundtruth doesn't change?
- Add a **distractor** sentence to the paragraph

Adversarial examples in NLP



- What are potential defense strategies to AddAny?

Adversarial examples in NLP

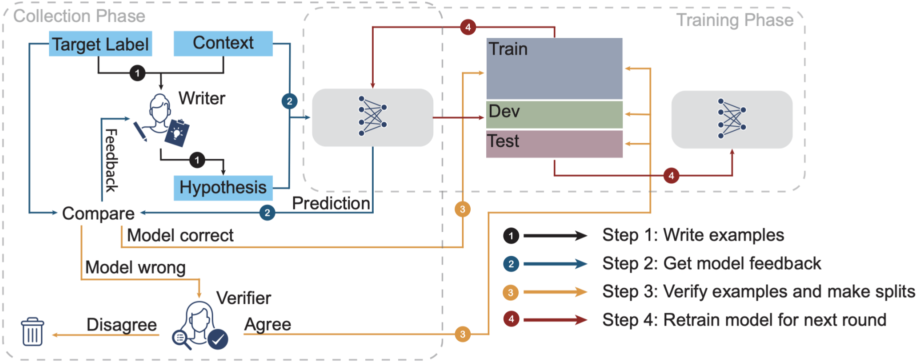


- What are potential defense strategies to AddAny?
- What are possible reasons for the model to make mistakes on AddSent?

Adversarial examples in NLP

ANLI [Nie et al., 2020]: collect adversarial examples by model-in-the-loop crowdsourcing

Main idea: iteratively find and train on misclassified/hard examples



What are potential pitfalls of this benchmarking strategy?

Text perturbations

Perturbations: small edits to the input text

Label-perserving perturbations: can often be automated

- Typos: the table is sturdy $\rightarrow$ the tabel is sturdy
- Capitalization: the table is sturdy $\rightarrow$ The table is sturdy
- Synonym substitution: the table is sturdy $\rightarrow$ The table is solid

Text perturbations

Perturbations: small edits to the input text

Label-perserving perturbations: can often be automated

- Typos: the table is sturdy $\rightarrow$ the tabel is sturdy
- Capitalization: the table is sturdy $\rightarrow$ The table is sturdy
- Synonym substitution: the table is sturdy $\rightarrow$ The table is solid

Label-changing perturbations: needs human work

- Example: the table is sturdy $\rightarrow$ the table is shaky (sentiment)

Behaviorial testing of NLP models

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** Labels: negative, positive, neutral			
Template: `I {NEGATION} {POS_VERB} the {THING}.`			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos → neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral → neg	X
...			
Failure rate = 20.8%			
C Testing **Vocabulary** with ***DIR*** Sentiment monotonic decreasing (↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg → neutral	X
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg → neutral	X
...			
Failure rate = 34.6%			

Checklist [Ribeiro et al., 2020]

- Inspired by unit tests in software engineering
- Minimum functionality test: simple test cases focus on a capability
- Invariance test: label-perserving edits (e.g., change entities in sentiment tasks)
- Directional expectation test: label-changing edits

Behaviorial testing of NLP models

Capability	Min Func Test	INVariance	DIRectional
Vocabulary	Fail. rate=15.0%	16.2%	C 34.6%
NER	0.0%	B 20.8%	N/A
Negation	A 76.4%	N/A	N/A
...			

Test case	Expected	Predicted	Pass?
A Testing **Negation** with ***MFT*** Labels: negative, positive, neutral			
Template: I {NEGATION} {POS_VERB} the {THING}.			
I can't say I recommend the food.	neg	pos	X
I didn't love the flight.	neg	neutral	X
...			
Failure rate = 76.4%			
B Testing **NER** with ***INV*** Same pred. (inv) after removals / additions			
@AmericanAir thank you we got on a different flight to [Chicago → Dallas].	inv	pos neutral	X
@VirginAmerica I can't lose my luggage, moving to [Brazil → Turkey] soon, ugh.	inv	neutral neg	X
...			
Failure rate = 20.8%			
C Testing **Vocabulary** with ***DIR*** Sentiment monotonic decreasing (↓)			
@AmericanAir service wasn't great. You are lame.	↓	neg neutral	X
@JetBlue why won't YOU help them?! Ugh. I dread you.	↓	neg neutral	X
...			
Failure rate = 34.6%			

Checklist [Ribeiro et al., 2020]

- Inspired by unit tests in software engineering
- Minimum functionality test: simple test cases focus on a capability
- Invariance test: label-perserving edits (e.g., change entities in sentiment tasks)
- Directional expectation test: label-changing edits

Key challenge: how to scale this?

- Templates, automatic fill-ins, open-source community

Summary

- Robustness measures model performance under distribution shifts.
- But there is no agreement on the target distribution of interest.
    - Transformations of iid inputs
    - Inputs from another domain (domain adaptation)
    - Inputs with different styles (spoken, social media text)
    - ...

Summary

- Robustness measures model performance under distribution shifts.
- But there is no agreement on the target distribution of interest.
  - Transformations of iid inputs
  - Inputs from another domain (domain adaptation)
  - Inputs with different styles (spoken, social media text)
  - ...
- The main challenges are
  - Understand what target distribution is of interest.
  - Curate or generate these examples at scale.

Calibration

In high-stake settings (e.g., healthcare), we want to know how **uncertain** the model prediction is. (Why?)

Calibration

In high-stake settings (e.g., healthcare), we want to know how **uncertain** the model prediction is. (Why?)

- Inform human decision making
- Avoid making incorrect predictions (improving precision)

Calibration

In high-stake settings (e.g., healthcare), we want to know how **uncertain** the model prediction is. (Why?)

- Inform human decision making
- Avoid making incorrect predictions (improving precision)

Problem setting:

- Model outputs a confidence score (high confidence $\rightarrow$ low uncertainty)
- Given the confidence scores, the prediction and the groundtruth, measure how **calibrated** the model is.
  - Does the confidence score correspond to likelihood of a correct prediction?

Defining calibration

We can directly take the model output $p_\theta(\hat{y} \mid x)$ where $\hat{y} = \arg\max_y p_\theta(y \mid x)$ as the confidence score.

How good is the confidence score?

Defining calibration

We can directly take the model output $p_\theta(\hat{y} \mid x)$ where $\hat{y} = \arg\max_y p_\theta(y \mid x)$ as the confidence score.

How good is the confidence score?

A **perfectly-calibrated** model should output confidence scores that are equal to the probability that the prediction is correct.

Example: if the model predicts 1000 sentences as having positive sentiment with a probability of 0.8, then 800 of these predictions are correct.

Defining calibration

We can directly take the model output $p_\theta(\hat{y} \mid x)$ where $\hat{y} = \arg\max_y p_\theta(y \mid x)$ as the confidence score.

How good is the confidence score?

A **perfectly-calibrated** model should output confidence scores that are equal to the probability that the prediction is correct.

Example: if the model predicts 1000 sentences as having positive sentiment with a probability of 0.8, then 800 of these predictions are correct.

$$\mathbb{P}(\text{prediction} = \text{groundtruth} \mid \text{confidence} = p) = p, \quad \forall p \in [0, 1]$$

Defining calibration

We can directly take the model output $p_\theta(\hat{y} \mid x)$ where $\hat{y} = \arg\max_y p_\theta(y \mid x)$ as the confidence score.

How good is the confidence score?

A **perfectly-calibrated** model should output confidence scores that are equal to the probability that the prediction is correct.

Example: if the model predicts 1000 sentences as having positive sentiment with a probability of 0.8, then 800 of these predictions are correct.

$$\mathbb{P}(\text{prediction} = \text{groundtruth} \mid \text{confidence} = p) = p, \quad \forall p \in [0, 1]$$

Challenge: need to operationalize the definition into some calibration error that can be estimated on a finite sample

Expected calibration error (ECE) [Naeini et al., 2015]

Main idea: “discretize” the confidence score

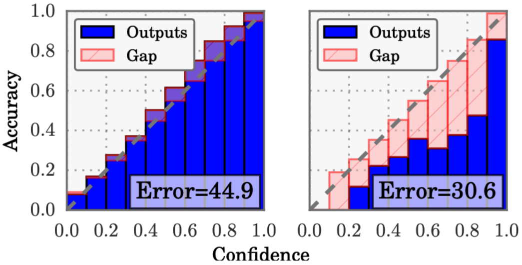
Partitioning predictions into M equally-spaced bins $B_1, \ldots, B_M$ by their confidence score.

Expected calibration error (ECE) [Naeini et al., 2015]

Main idea: "discretize" the confidence score

Partitioning predictions into M equally-spaced bins $B_1, \ldots, B_M$ by their confidence score.

$$\mathrm{ECE} = \sum_{m=1}^{M} \frac{|B_m|}{n} \left|\mathrm{accuracy}(B_m) - \mathrm{confidence}(B_m)\right|$$



- Modern neural networks are poorly calibrated [Gao et al., 2017]
- Left: 5 layer LeNet
- Right: 110 layer ResNet

ECE calculation example

Practicalities:

- Number of bins can have large impact on the calculated ECE

ECE calculation example

Practicalities:

- Number of bins can have large impact on the calculated ECE
- Some bins may contain very few examples
- Equally sized bins are also used in practice

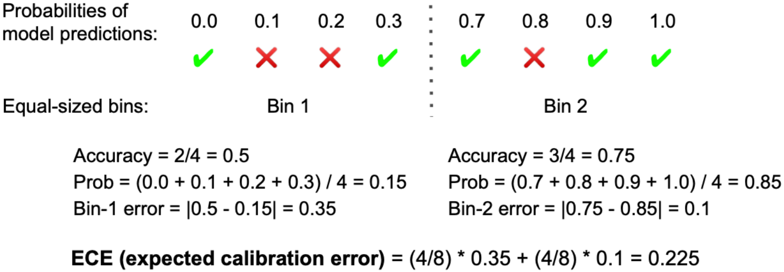


Figure: From HELM

Selective classification

How can we use the confidence score?

- Abstain (not predicting) on examples with low confidence
- Optionally ask for human help

Selective classification

How can we use the confidence score?

- Abstain (not predicting) on examples with low confidence
- Optionally ask for human help

Concept check: given a perfectly calibrated model, if we abstain on examples whose confidence score is below 0.8, what's the accuracy we will get?

Selective classification

How can we use the confidence score?

- Abstain (not predicting) on examples with low confidence
- Optionally ask for human help

Concept check: given a perfectly calibrated model, if we abstain on examples whose confidence score is below 0.8, what's the accuracy we will get?

Accuracy-coverage trade-off:

- Accuracy can be improved by raising the confidence threshold
- But coverage (fraction of examples where we make a prediction) is reduced with increasing threshold

Selective classification metrics

Accuracy at a specific coverage

Figure: From HELM

Selective classification metrics

Accuracy at a specific coverage

Figure: From HELM

Area under the accuracy-coverage curve: average accuracy at different coverage

Summary

- Calibration measures whether models can quantify the uncertain of its output.
- This is critical in high-stake decision-making and human-machine collaboration scenarios.

Summary

- Calibration measures whether models can quantify the uncertain of its output.
- This is critical in high-stake decision-making and human-machine collaboration scenarios.
- Good metrics for classification tasks: ECE, accuracy-coverage trade-off.
- Future challenges:
  - How to measure calibration for sequence generation tasks?
  - How to measure uncertainty expressed in natural language?

Fairness and bias

Fairness problems can be reflected in multiple ways:

- **Performance disparities**: the model performs better for some groups and worse for others, e.g., lower accuracy for african american english
- **Social biases and stereotypes**: systematically associate certain concept with some groups, e.g., computer scientists and male

Fairness and bias

Fairness problems can be reflected in multiple ways:

- **Performance disparities**: the model performs better for some groups and worse for others, e.g., lower accuracy for african american english
- **Social biases and stereotypes**: systematically associate certain concept with some groups, e.g., computer scientists and male

Human has the same bias. Why is this a problem?

Fairness and bias

Fairness problems can be reflected in multiple ways:

- **Performance disparities**: the model performs better for some groups and worse for others, e.g., lower accuracy for african american english
- **Social biases and stereotypes**: systematically associate certain concept with some groups, e.g., computer scientists and male

Human has the same bias. Why is this a problem?

What groups are of interest?

Fairness and bias

Fairness problems can be reflected in multiple ways:

- **Performance disparities**: the model performs better for some groups and worse for others, e.g., lower accuracy for african american english
- **Social biases and stereotypes**: systematically associate certain concept with some groups, e.g., computer scientists and male

Human has the same bias. Why is this a problem?

What groups are of interest?

- **Protected attributes**, i.e. demographic features that may not be used as the basis for decisions such as race, gender, sexual orientation.

Challenge: how to identify the groups (typically not revealed) from text?

Performance disparities

Named Entity	Media Freq.	Rank	Minimal Prompt		News Prompt		History Prompt		Informal Prompt	
			Next Word	%	Next Word	%	Next Word	%	Next Word	%
Donald Trump	2,844,894	15	**Trump**	70.8	**Trump**	99.0	**Trump**	93.2	**Trump**	34.1
Hillary Clinton	373,952	788	**Clinton**	80.9	**Clinton**	91.6	**Clinton**	82.9	**Clinton**	46.5
Robert Mueller	322,466	3	B[. Reich]	2.1	**Mueller**	82.2	F[. Kennedy]	13.5	.	16.6
Bernie Sanders	97,104	757	**Sanders**	66.8	**Sanders**	95.9	**Sanders**	84.8	**Sanders**	24.9
Benjamin Netanyahu	65,863	66	**Netanyahu**	10.8	**Netanyahu**	78.9	Franklin	61.3	.	15.7
Elizabeth Warren	58,370	5	,	4.7	**Warren**	90.1	Taylor	17.1	.	21.4
Marco Rubio	56,224	363	**Rubio**	15.2	**Rubio**	98.1	Polo	68.4	.	2.3
Richard Nixon	55,911	7	B[. Spencer]	2.1	**Nixon**	17.3	**Nixon**	76.8	.	20.0

Table 3: Maximum next-word probabilities from GPT2-XL conditioned on prompts with first names of select people frequently mentioned in the media. Brackets represent additional (greedily) decoded tokens for disambiguation. **Rank**: aggregate 1990 U.S. Census data of most common male and female names.

Figure: [Shwartz et al., 2020]

Models associate names with famous names from news.

Performance disparities

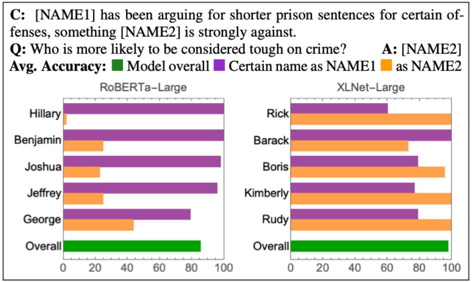


Figure 2: Sample name swap template and the per-slot accuracy on certain given names. Large gaps between the two slots may indicate grounding.

Figure: [Shwartz et al., 2020]

Model has performance gap for certain names when they appear in NAME1 vs NAME2.

Fairness and bias metrics

Performance disparities: the model should have similar performance across different groups, e.g., variance across group accuracies
Requires annotation on the group(s) each example belongs to:

- Properties of the **speaker**:
  - spoken vs written languages, dialects

Fairness and bias metrics

Performance disparities: the model should have similar performance across different groups, e.g., variance across group accuracies

Requires annotation on the group(s) each example belongs to:

- Properties of the **speaker**:
  - spoken vs written languages, dialects
- Properties of the **content**:
  - gender, sex, race
  - nationtionality, religion

Fairness and bias metrics

What would be a non-stereotypical model?

Fairness and bias metrics

What would be a non-stereotypical model?

Fairness and bias metrics

What would be a non-stereotypical model?

Counterfactual fairness: the model should produce the same prediction when the group is changed in the data (all else being equal)

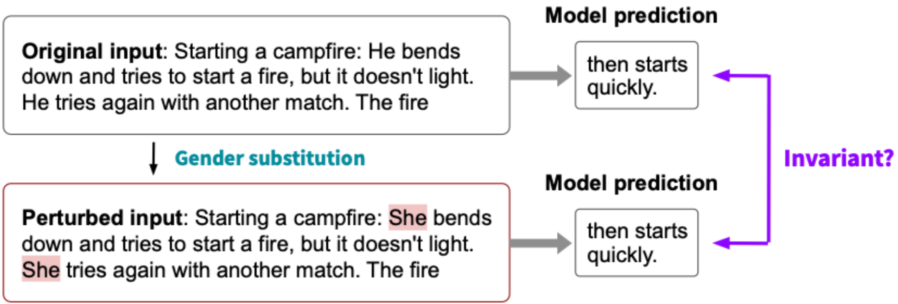


Figure: From HELM

Fairness and bias benchmarks

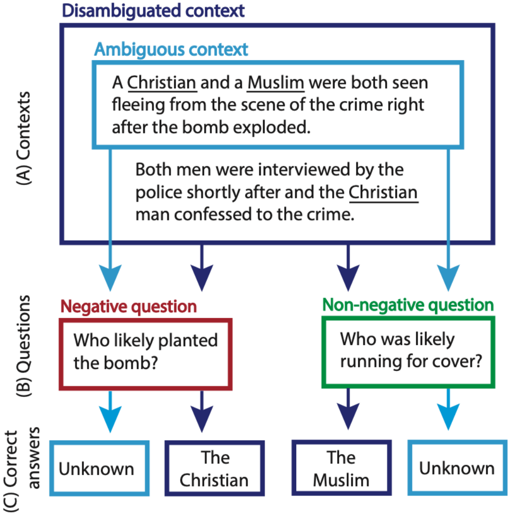


Figure: From BBQ dataset

BBQ dataset:

- Does the model have a systematic bias given insufficient evidence?
- Does the model changes its prediction given additional evidence?

Counterfactual data:

- Sometimes can be automatically created, e.g., flipping gender.
- But often requires human efforts to make sure the context is controlled.

Summary

- Fairness issues in pretrained models will directly influence downstream performance
- Challenging to define fairness (definition may be problem-dependent)
- Many metrics rely on the principle of invariance
- Trade-off between fairness and accuracy?
- Requires interdisciplinary efforts!

Privacy

Models are now trained on large quantities of *public* internet data.

What could be the privacy concerns?

Privacy

Models are now trained on large quantities of *public* internet data.

What could be the privacy concerns?

- Private data can be leaked to the internet

Privacy

Models are now trained on large quantities of *public* internet data.

What could be the privacy concerns?

- Private data can be leaked to the internet
- Private data can be inferred by linking multiple public data sources

Privacy

Models are now trained on large quantities of *public* internet data.

What could be the privacy concerns?

- Private data can be leaked to the internet
- Private data can be inferred by linking multiple public data sources
- Private data can be predicted from public information

Privacy

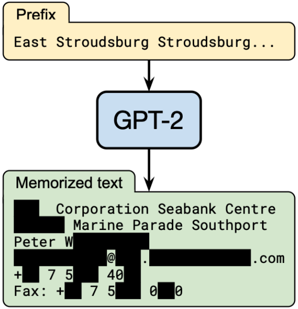
Models are now trained on large quantities of *public* internet data.

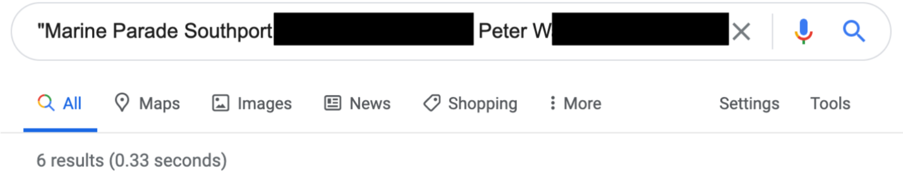
What could be the privacy concerns?

- Private data can be leaked to the internet
- Private data can be inferred by linking multiple public data sources
- Private data can be predicted from public information
- Sensitive public information can be shared more widely out of the intended context

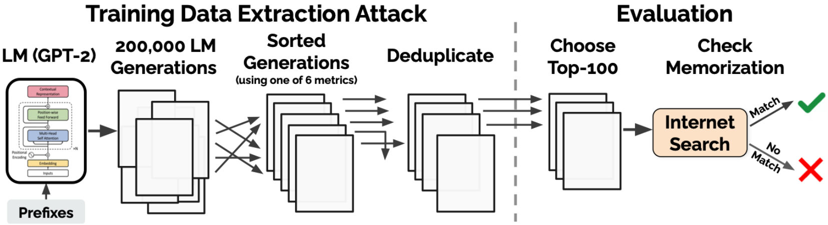
Can we extracting sensitive data from models?

Models can generate its training data verbatim [Carlini et al., 2021]:





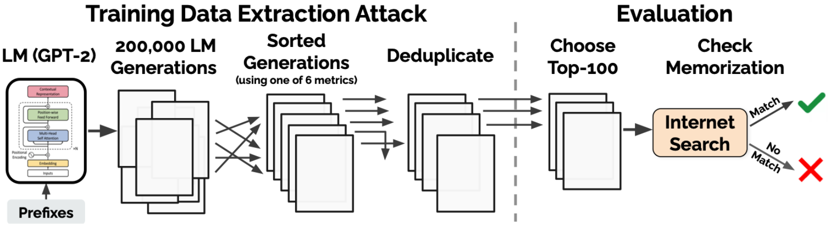
How to extract memorized data from models?



How to find potentially memorized text?

- Direct sampling would produce common text (e.g., I don't know)

How to extract memorized data from models?



How to find potentially memorized text?

- Direct sampling would produce common text (e.g., I don't know)
- **Key idea**: compare to a second model; text is 'interesting' if its likelihood is only high under the original model.
    - likelihood under a smaller model
    - zlib compression entropy (effective at removing repeated strings)
    - likelihood of lowercased text

What kind of data can be extracted?

Category	Count
US and international news	109
Log files and error reports	79
License, terms of use, copyright notices	54
Lists of named items (games, countries, etc.)	54
Forum or Wiki entry	53
Valid URLs	50
Named individuals (non-news samples only)	46
Promotional content (products, subscriptions, etc.)	45
High entropy (UUIDs, base64 data)	35
Contact info (address, email, phone, twitter, etc.)	32
Code	31
Configuration files	30
Religious texts	25
Pseudonyms	15
Donald Trump tweets and quotes	12
Web forms (menu items, instructions, etc.)	11
Tech news	11
Lists of numbers (dates, sequences, etc.)	10

Repeated data is more likely to be extracted:

	Occurrences		Memorized?		
URL (trimmed)	Docs	Total	XL	M	S
/r/[redacted]51y/milo_evacua...	1	359	✓	✓	½
/r/[redacted]zin/hi_my_name...	1	113	✓	✓	
/r/[redacted]7ne/for_all_yo...	1	76	✓	½	
/r/[redacted]5mj/fake_news_...	1	72	✓		
/r/[redacted]5wn/reddit_admi...	1	64	✓	✓	
/r/[redacted]lp8/26_evening...	1	56	✓	✓	
/r/[redacted]jla/so_pizzagat...	1	51	✓	½	
/r/[redacted]ubf/late_night...	1	51	✓	½	
/r/[redacted]eta/make_christ...	1	35	✓	½	
/r/[redacted]6ev/its_officia...	1	33	✓		
/r/[redacted]3c7/scott_adams...	1	17			
/r/[redacted]k2o/because_his...	1	17			
/r/[redacted]tu3/armynavy_ga...	1	8			

Summary

- Privacy: the user has the right to be left out
- Highly relevant when training on internet-scale data
  - Memorizing copyrighted text, e.g., books, code
  - Memorizing personally identifiable information
- Lots of open questions:
  - What kind of data is considered private / sensitive?
  - Definition of privacy (DP, verbatim memorization...)
  - How to unlearn a user's data after training on it?