# Large Language Models

## Metrics, tasks, benchmarks

Flavio Giobergia

# Metrics

# Geometric mean (a digression)

- The geometric mean over values $x_1, x_2, \ldots, x_N$ computes a mean across $N$ values

$$GM = (x_1 \cdot x_2 \cdot \ldots \cdot x_N)^{1/N} = \prod_i^N x_i^{1/N}$$

- GM penalizes more (w.r.t .arithmetic mean) the presence of low values:
  - Arithmetic mean (0.1, 0.9, 0.9) = 0.633
  - Geometric mean (0.1, 0.9, 0.9) = 0.433
- Computing the product of many values can produce *numerical instability*
- So we often find, instead

$$GM = \exp\left(\log\left(\prod_i^N x_i^{1/N}\right)\right) = \exp\left(\frac{1}{N}\sum_i^N \log x_i\right)$$

# Perplexity

- *Perplexity* is a metric that quantifies how uncertain the model is in predicting the (correct) next word
  - *High perplexity*: the model is *uncertain* about the "correct" next word
  - *Low perplexity*: the model is *certain* about the "correct" next word

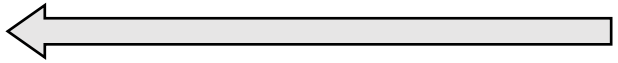
- It is quantified in terms of how large the probability assigned to each next word of a sentence is

- The number represents the average* number of words the model is indecise over, for a sentence

(*) Geometric mean

# Perplexity, intuition

- The dog chased the _____
- Probability distribution:
  - P(cat) = 0.25
  - P(mouse) = 0.15
  - P(table) = 0.05
  - P(chair) = 0.1
  - P(computer) = 0.05
  - P(tiger) = 0.2
  - P(crocodile) = 0.1
  - P(owner) = 0.1
  - P(the) = 0



Correct word: cat
P(cat) = 0.25

Even if the vocabulary has 10 words, It's *as if the model was confused between 4 words* ($1/0.25 = 0.25^{-1}$), each with equal probability

# Perplexity

- The perplexity over a sentence $w_1, w_2, \ldots, w_N$ of tokens is computed as follows:

$$PPL = \exp\left(-\frac{1}{N}\sum_{i}^{N}\log p(w_i|w_{<i})\right)$$

- Note that:
  - If a model predicts perfectly all words ($\forall\, i, p(w_i|w_{<i}) = 1$), the negative log-likelihoods (NLL) sum to 0, and the perplexity will be 1
    - The model is certain about the sentence
  - If $p(w_i|w_{<i}) < 1$, the NLL will sum to a large number which, exponentiated, produces an even larger number
    - The model is uncertain about the sentence, it is considering other words as well

# Perplexity, example (certain model)

<bos>The dog chased the cat<eos>

- P(The|<bos>) = 0.9 $\rightarrow$ $NLL = -\log(0.9) = 0.1054$
- P(dog|<bos>,The) = 0.85 $\rightarrow$ $NLL = 0.1625$
- P(chased|<bos>,The,dog) = 0.88 $\rightarrow$ $NLL = 0.1278$
- P(the|<bos>,The,dog,chased) = 0.75 $\rightarrow$ $NLL = 0.2877$
- P(cat|<bos>,The,dog,chased,the) = 0.95 $\rightarrow$ $NLL = 0.0513$
- P(<eos>|<bos>,The,dog,chased,the) = 0.8 $\rightarrow$ $NLL = 0.2231$
- $PPL = e^{\frac{1}{6}(0.1054+0.1625+0.1278+0.2877+0.0513+0.2231)} = 1.1731$

# Perplexity, example (uncertain model)

<bos>The dog chased the cat<eos>

- P(The|<bos>) = $0.05 \rightarrow NLL = 2.9957$
- P(dog|<bos>,The) = $0.1 \rightarrow NLL = 2.3026$
- P(chased|<bos>,The,dog) = $0.2 \rightarrow NLL = 1.6094$
- P(the|<bos>,The,dog,chased) = $0.01 \rightarrow NLL = 4.6052$
- P(cat|<bos>,The,dog,chased,the) = $0.005 \rightarrow NLL = 5.2983$
- P(<eos>|<bos>,The,dog,chased,the) = $0.001 \rightarrow NLL = 6.9078$
- $PPL = e^{\frac{1}{6}(2.9957+2.3026+1.6094+4.6052+5.2983+6.9078)} = 52.1$

# Perplexity, example (with GM indecision)

<bos>The dog chased the cat<eos>

- P(The|<bos>) = $0.05$ → Uncertain over 1/0.05 = 20 “equivalent” words
- P(dog|<bos>,The) = $0.1$ → = 10 words
- P(chased|<bos>,The,dog) = $0.2$ → 5 words
- P(the|<bos>,The,dog,chased) = $0.01$ → 100 words
- P(cat|<bos>,The,dog,chased,the) = $0.005$ → 200 words
- P(<eos>|<bos>,The,dog,chased,the) = $0.001$ → 1000 words
- $PPL = (20 \cdot 10 \cdot 5 \cdot 100 \cdot 200 \cdot 1000)^{1/6} = 52.1$

# BLEU

- *BLEU* (*Bilingual Evaluation Understudy*) is a metric used to evaluate a generated sequence, when a reference one is available
- Computing BLEU-n:
  - Generate all i-grams ($i = 1, 2, \ldots, n$) for predicted and reference sentences
  - Count, for each size i, the the fraction of matching i-grams over all *generated* i-grams
    - i.e., the precision ($precision_i$)
  - Compute the *geometric mean* over all values i
  - Multiply by a *brevity penalty* (BP)
    - (Some models may produce much shorter sequences to artificially improve precision!)

$$BLEU\text{–}n = BP \cdot \exp\left(\frac{1}{n}\sum_i \log(precision_i)\right)$$

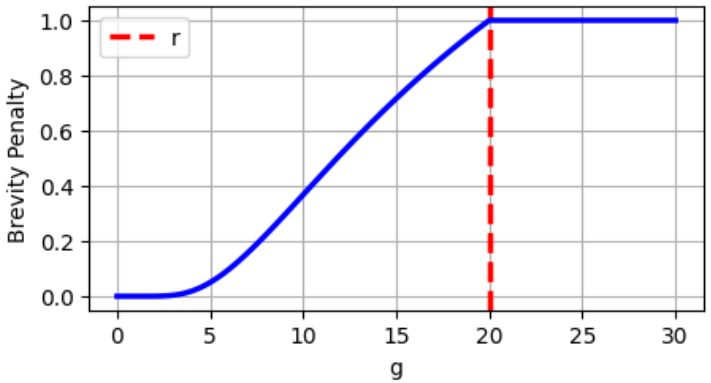
# Brevity Penalty

- If a model generates a very short sequence (w.r.t. the reference), it is easier to obtain a larger precision!
  - (Because of the smaller denominator)

- BLEU introduces a multiplicative penalty if the model produces a short sequence

- If $r$ is the length of the reference sequence, and $g$ is the length of the generated sequence, BP is defined as:

$$BP = \begin{cases} 1 & if\ g > r \\ e^{\left(1-\frac{r}{g}\right)} & if\ g \leq r \end{cases}$$



# BLEU – example (1)

- Reference: The dog chased the cat
- Generated: The dog ran after the cat
- BLEU-2 (i = 1, 2)

i=1)

- Reference: (The), (dog), (chased), (the), (cat)
- Generated: (The), (dog), (ran),(after), (the), (cat)
- $precision_1 = 4 / 6 = 0.667$

i=2)

- Reference: (The, dog), (dog, chased), (chased, the), (the, cat)
- Generated: (The, dog), (dog, ran), (ran, after), (after, the), (the, cat)
- $precision_2 = 2 / 5 = 0.4$

# BLEU – example (2)

- $GM = \exp\left(\frac{1}{2} \cdot (\log(0.667) + \log(0.4)\right) = 0.5165$
- $r = 5$
- $g = 6$
- $BP = 1$

$$BLEU\text{–}2 = 1 \cdot 0.5165 = 0.5165$$

# BLEU – limitations

- BLEU is effective *if we need to match exact results*
- It attempts to capture sequentiality of words with the introduction of n-grams of various lengths
- However:
  - BLEU does not care about the *semantic* of results
    - Semantically similar sentences are not accepted as valid
  - No considerations about *fluency*, or *meaning*
    - Garbage sentences may get relatively large BLEU scores
  - *Word order* ignored beyond n-grams

# Other metrics

- *ROUGE* (*Recall-Oriented Understudy for Gisting Evaluation*)
  - Family of evaluation metrics
    - ROUGE-N (considers n-grams for the computation of metrics)
      - $ROUGE-n = \frac{|G_n \cap R_n|}{|R_n|}$
      - Where $G_n$, $R_n$ are the bags (multisets) of n-grams of Generated and Reference texts)
    - ROUGE-L (considers the Longest Common Subsequence)
    - ROUGE-W (Weighted LCS)
    - …
  - Used, for instance, in summarization
    - The focus on recall verifies that "all information" in the reference is present
- *METEOR* (*Metric for Evaluation of Translation with Explicit ORdering*)
  - Addresses some of BLEU's problems:
    - Adds stemming, synonymy matching
    - Includes both precision and recall (with $F_\beta$ score)
    - Adds a fragmentation penalty (penalty if words are not contiguous)

# BERT Score

- Given Generated sequence G and Reference sequence R,
- Tokenize G and R, get output vectors *via BERT* (or similar model)
- *Compare* each generated token against each reference token
  - With a similarity function (e.g. cosine similarity)
  - Get a score (0 → 1) on how similar the predicted vector is (to the most similar reference vector)

$$precision = \frac{1}{|G|} \sum_{i} \max_{j} \text{sim}(G_i, R_j)$$

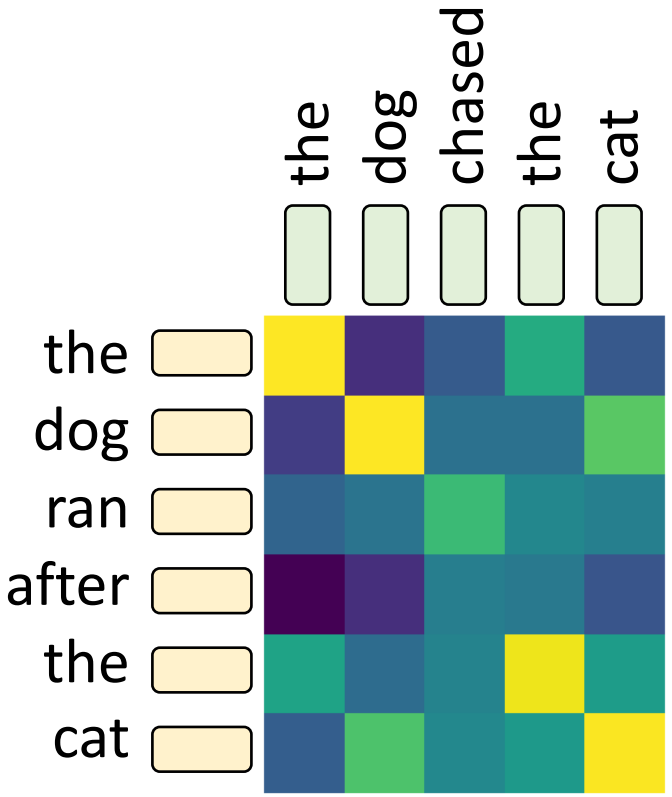
- Compare each reference token against each predicted token
  - Same as before

$$recall = \frac{1}{|R|} \sum_{j} \max_{i} \text{cos}sim(G_i, R_j)$$

- ($F_1$ score can be computed as usual)

# BERT Score – example

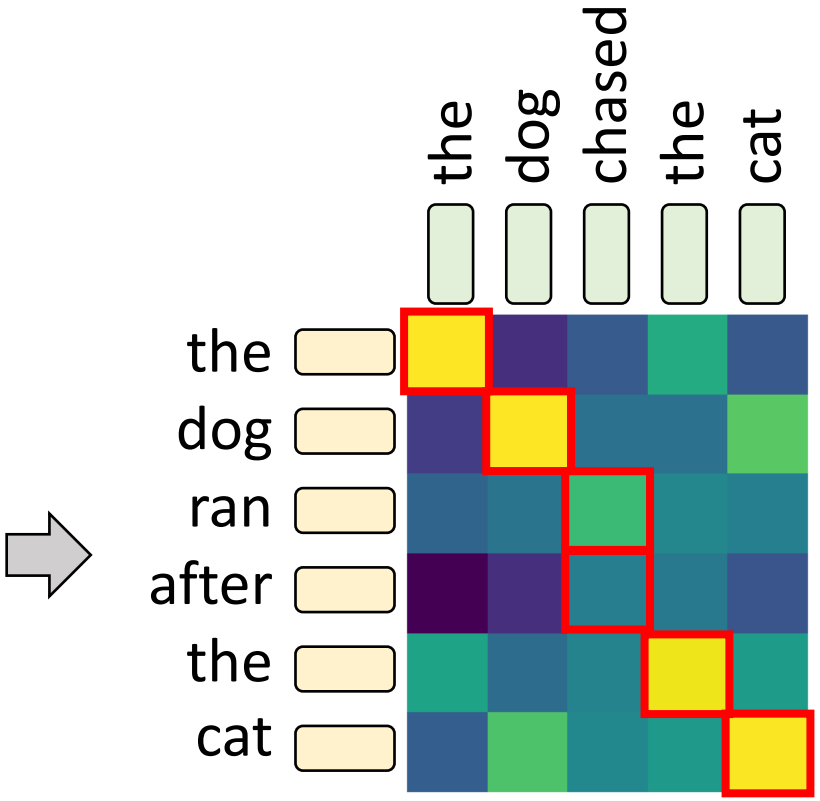
- Reference: the dog chased the cat
- Predicted: the dog ran after the cat



# BERT Score – example

- Reference: the dog chased the cat
- Predicted: the dog ran after the cat

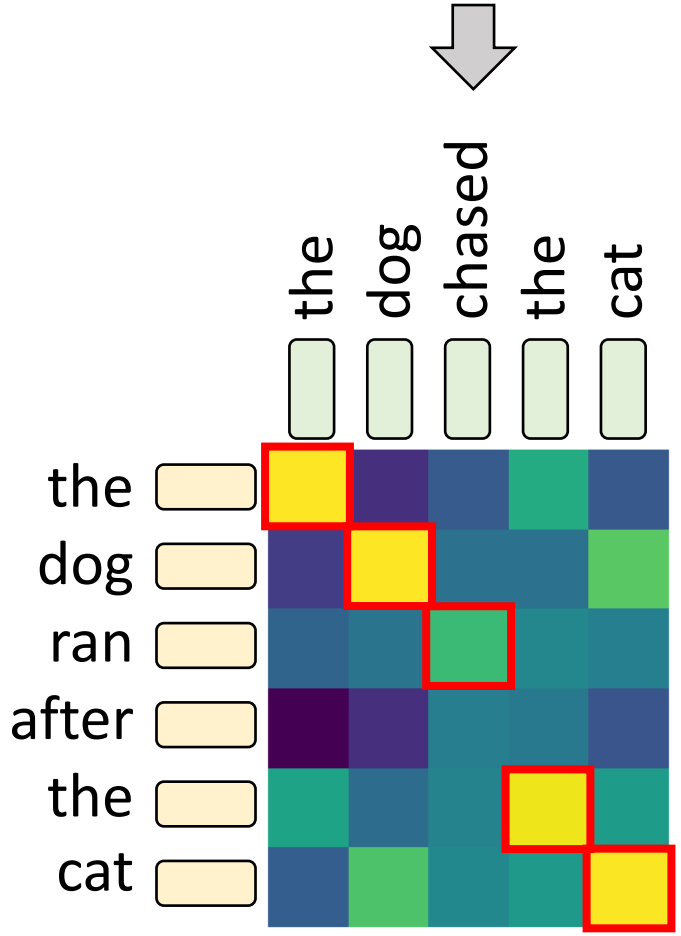
- Precision
  - Get best match for each row
    - The → 0.9967
    - Dog → 0.9968
    - Ran → 0.9351
    - After → 0.8868
    - The → 0.9921
    - Cat → 0.996




- $precision = \frac{1}{6} \cdot (0.9967 + 0.9968 + 0.9351 + 0.8868 + 0.9921 + 0.996) = 0.96725$

# BERT Score – example

- Reference: the dog chased the cat
- Predicted: the dog ran after the cat




- Recall
  - Get best match for each column
    - The → 0.9967
    - Dog → 0.9968
    - Chased → 0.9351
    - The → 0.9921
    - Cat → 0.996
- $recall = \frac{1}{5} \cdot (0.9967 + 0.9968 + 0.9351 + 0.9921 + 0.996) = 0.96725$

# BERT Score

- BERT Score “solves” the semantic similarity problem
- Still, does not explicitly consider order
  - However, tokens are now contextualized, so that helps
- BERT Score relies on an external model
  - Any limitation of the model reflects on the quality of the result
  - And, computationally, it is not ideal
- This score does not offer a clear interpretation
  - “BERT says so”

# Exact Match (EM)

- Sometimes, we instead want to match a token/tokens *exactly*

- *Exact Match* (*EM*) produces a binary output
  - 1 if the match is correct (generally, including case, punctuation)
  - 0 otherwise

# Ranking

- For each token, we can *assign a rank* to the right word
  - If we sort the tokens by descending probability, where is the right token placed?

- Commonly adopted ranking metrics can be used in this setting
  - Rank, MRR, NDCG, precision@k, recall@k, …

# Task-specific metrics

- Many tasks are framed so as to get a specific word/words as the output
  - E.g., cloze/fill-the-blank questions
- In that case, classic metrics can be adopted directly
  - Accuracy, precision, recall, F1 score

# Human evaluation

- Automated metrics fail to capture concepts like:
  - Coherence, creativity, relevance, fluency
- Human judgement is necessary for these activities
- However, humans are a *limited resource*!
  - Costly
  - Slow process
  - Evaluation does not scale
  - (We will see in future lectures, sometimes we replace humans with other LMs)
- Common methods:
  - *Rating scales* (e.g., “evaluate on a scale from 1 to 5 for fluency)
  - *Pairwise comparisons* (e.g., “which of these two sentences is more relevant?)

# Tasks

# Text completion

- The model is given an incomplete sentence or passage and must figure out what a plausible continuation is
- The model either *generates the correct continuation* of the text,
- Or, it *chooses the best option* among a pool of options

# LAMBADA

- LAMBADA (https://arxiv.org/pdf/1606.06031)
  - A collection of narrative passages
  - Designed so that humans can guess the word if they read the entire passage, *but not just the last sentence*
- Metrics
  - Accuracy, Perplexity, Rank

*Context:* In my palm is a clear stone, and inside it is a small ivory statuette. A guardian angel. "Figured if you're going to be out at night getting hit by cars, you might as well have some backup." I look at him, feeling stunned. Like this is some sort of sign.
*Target sentence:* But as I stare at Harlin, his mouth curved in a confident grin, I don't care about _____.
*Target word:* signs

---

*Context:* Both its sun-speckled shade and the cool grass beneath were a welcome respite after the stifling kitchen, and I was glad to relax against the tree's rough, brittle bark and begin my breakfast of buttery, toasted bread and fresh fruit. Even the water was tasty, it was so clean and cold.
*Target sentence:* It almost made up for the lack of _____.
*Target word:* coffee

# ROCStories, HellaSwag, StoryCloze

- There are other datasets comprised of short stories, e.g.
  - ROCStories (https://arxiv.org/pdf/1604.01696v1)
  - HellaSwag (https://aclanthology.org/P19-1472.pdf)
  - StoryCloze (https://arxiv.org/pdf/1604.01696v1)
- They provide a story, and possible endings (one correct, the others not)
- The LM can be evaluated on:
  - Capability of detecting the right answer among options (accuracy, precision, …)
  - Capability of generating the right ansawer (PPL, BLEU, …)

| Context | Right Ending | Wrong Ending |
|---|---|---|
| Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating. | Karen became good friends with her roommate. | Karen hated her roommate. |
| Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money. | Jim decided to devise a plan for repayment. | Jim decided to open another credit card. |
| Gina misplaced her phone at her grandparents. It wasn't anywhere in the living room. She realized she was in the car before. She grabbed her dad's keys and ran outside. | She found her phone in the car. | She didn't want her phone anymore. |

| Context | Right Ending | Wrong Ending |
|---|---|---|
| Tom and Sheryl have been together for two years. One day, they went to a carnival together. He won her several stuffed bears, and bought her funnel cakes. When they reached the Ferris wheel, he got down on one knee. | Tom asked Sheryl to marry him. | He wiped mud off of his boot. |
| Karen was assigned a roommate her first year of college. Her roommate asked her to go to a nearby city for a concert. Karen agreed happily. The show was absolutely exhilarating. | Karen became good friends with her roommate. | Karen hated her roommate. |
| Jim got his first credit card in college. He didn't have a job so he bought everything on his card. After he graduated he amounted a $10,000 debt. Jim realized that he was foolish to spend so much money. | Jim decided to devise a plan for repayment. | Jim decided to open another credit card. |

# Question Answering (QA)

- Can models answer questions? In different scenarios:
- Type of access to knowledge
  - Open-book: allow the model to “search” the answer in a text/set of texts
    - Either via Information Retrieval (i.e., access to some database – RAG),
    - Or, provided via context (“prompt”)
  - Closed-book: measure what the model already knows (i.e., encoded ein the parameters)
- Type of answers
  - Extractive
  - Abstractive
  - Multiple-choice (A, B, C, D)

# QA benchmarks

- SQuAD (Stanford Question Answering Dataset)
  - https://rajpurkar.github.io/SQuAD-explorer/
  - 100K QA pairs from Wikipedia articles
  - Answers are segments of text from the articles
  - SQuAD 2.0 introduces “unanswerable” questions
- TriviaQA
  - https://nlp.cs.washington.edu/triviaqa/
  - 650K QA pairs (+ evidence) based on various trivia websites
- Natural Questions
  - https://ai.google.com/research/NaturalQuestions
  - By Google, with real user queries from Google search & associated answers from Wikipedia
  - Has both short- and long-form answers
- WebQuestions
  - https://aclanthology.org/D13-1160.pdf
  - 6K QA pairs (Q: Google Suggest API, A: Amazon Mechanical Turk)

## Natural Questions example

Question:

how many episodes in season 2 breaking bad?

Short Answer:

13

Long Answer:

The second season of the American television drama series Breaking Bad premiered on March 8 , 2009 and concluded on May 31 , 2009 . It consisted of 13 episodes , each running approximately 47 minutes in length . AMC broadcast the second season on Sundays at 10 : 00 pm in the United States . The complete second season was released on Region 1 DVD and Region A Blu - ray on March 16 , 2010.

## TriviaQA example

**Question**: American Callan Pinckney’s eponymously named system became a best-selling (1980s-2000s) book/video franchise in what genre?
**Answer**: Fitness
**Excerpt**: Callan Pinckney was an American fitness professional. She achieved unprecedented success with her Callanetics exercises. Her 9 books all became international best-sellers and the video series that followed went on to sell over 6 million copies. Pinckney’s first video release ”Callanetics: 10 Years Younger In 10 Hours” outsold every other **fitness** video in the US.

# Translation, summarization

- Translation
  - Generate the translation of a sentence, and evaluate the quality of the result
  - Datasets
    - WMT (Workshop on Machine Translation) publish datasets with various constraints, sizes, source/target languages
      - E.g., WMT 2024 ➔ https://www2.statmt.org/wmt24/mtdata/
- Summarization
  - Generate a summarized version of an original text (e.g., news article)
  - Extractive (select best sentences), or Abstractive (generate new text)
  - Datasets
    - CNN/Daily Mail (https://arxiv.org/pdf/1602.06023v5)
    - PubMed Diabetes (https://linqs.org/datasets/#pubmed-diabetes)
- Metrics
  - BLEU, ROUGE, METEOR, PPL, …

# Natural Language Inference

- Given a "premise", NLI is the task to determine whether a "hypothesis" is:
  - True (*entailment*)
  - False (*contradiction*)
  - Undetermined (*neutral*)

| Premise | Label | Hypothesis |
|---|---|---|
| A man inspects the uniform of a figure in some East Asian country. | contradiction | The man is sleeping. |
| An older and younger man smiling. | neutral | Two men are smiling and laughing at the cats playing on the floor. |
| A soccer game with multiple males playing. | entailment | Some men are playing a sport. |

http://nlpprogress.com/english/natural_language_inference.html

- Datasets
  - Stanford NLI (https://nlp.stanford.edu/projects/snli/)
  - Multi-genre NLI (https://arxiv.org/abs/1704.05426)

- Metrics: classification-based

# Grammatical acceptability

- Can models detect whether sentences grammatically correct or not?

- CoLA (Corpus of Linguistic Acceptability)
  - https://nyu-mll.github.io/CoLA/
  - 10k sentences, annotated for acceptability

- Metrics: any used for a binary classification task

| Label | Sentence |
|---|---|
| * | The more books I ask to whom he will give, the more he reads. |
| ✓ | I said that my father, he was tight as a hoot-owl. |
| ✓ | The jeweller inscribed the ring with the name. |
| * | many evidence was provided. |
| ✓ | They can sing. |
| ✓ | The men would have been all working. |
| * | Who do you think that will question Seamus first? |
| * | Usually, any lion is majestic. |
| ✓ | The gardener planted roses in the garden. |
| ✓ | I wrote Blair a letter, but I tore it up before I sent it. |

# Text classification

- Task: classifying sentences into categories based on the contents.
- Sentiment analysis
  - IMDb (https://ai.stanford.edu/~amaas/data/sentiment/)
    - 50k reviews, (positive/negative)
  - Yelp (https://www.yelp.com/dataset)
    - ~7M reviews (5 stars)
  - SST-2 (Stanford Sentiment Treebank) (https://github.com/YJiangcm/SST-2-sentiment-analysis)
- Topic classification
  - AG News (http://groups.di.unipi.it/~gulli/AG_corpus_of_news_articles.html)
  - 20 Newsgroup (http://qwone.com/~jason/20Newsgroups/)

# Mathematical reasoning

- MATH (https://arxiv.org/pdf/2103.03874v2)
  - 12.5K middle school/high school problems (in LaTeX)
- GSM8k (https://arxiv.org/pdf/2110.14168)
  - 8.5K middle school math problems
  - Provides: Problem, solution, annotations (in red)

MATH Dataset (Ours)

**Problem:** Tom has a red marble, a green marble, a blue marble, and three identical yellow marbles. How many different groups of two marbles can Tom choose?

**Solution:** There are two cases here: either Tom chooses two yellow marbles (1 result), or he chooses two marbles of different colors ($\binom{4}{2} = 6$ results). The total number of distinct pairs of marbles Tom can choose is $1 + 6 = \boxed{7}$.

**Problem:** If $\sum_{n=0}^{\infty} \cos^{2n}\theta = 5$, what is $\cos 2\theta$?

**Solution:** Note $1 + \cos^2\theta + \cos^4\theta + \cdots = \frac{1}{1-\cos^2\theta} = 5$. Hence, $\cos^2\theta = \frac{4}{5}$. Then $\cos 2\theta = 2\cos^2\theta - 1 = \boxed{\frac{3}{5}}$.

**Problem:** The equation $x^2 + 2x = i$ has two complex solutions. Determine the product of their real parts.

**Solution:** Complete the square by adding 1 to each side. Then $(x+1)^2 = 1 + i = e^{\frac{i\pi}{4}}\sqrt{2}$, so $x + 1 = \pm e^{\frac{i\pi}{8}}\sqrt[4]{2}$. The desired product is then

$$\left(-1 + \cos\left(\tfrac{\pi}{8}\right)\sqrt[4]{2}\right)\left(-1 - \cos\left(\tfrac{\pi}{8}\right)\sqrt[4]{2}\right) =$$

$$1 - \cos^2\left(\tfrac{\pi}{8}\right)\sqrt{2} = 1 - \frac{(1+\cos(\frac{\pi}{4}))}{2}\sqrt{2} = \boxed{\frac{1-\sqrt{2}}{2}}.$$

**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

# Commonsense reasoning

- Commonsense reasoning is the ability to make inferences based on everyday knowledge of the world
  - Cause-effect
  - Social norms
  - Reasoning about physical objects
- Examples of tasks
  - ambiguity resolution (infer the intended meaning of words in context)
  - causal reasoning (infer cause-and-effect relationships)
  - temporal reasoning (understand sequentiality of events)
  - physical reasoning (understand the physical domain: permanence, properties, …)
  - social reasoning (interpret people's/social interactions)
  - counterfactual reasoning (think about hypothetical scenarios)
  - …

# Ambiguity resolution

- Given a sentence with an ambiguous pronoun, can the model understand the entity, based on context?
  - The trophy didn't fit in the suitcase because it was too big.
  - What is "it"?
- Datasets
  - Winograd Schema Challenge (https://cdn.aaai.org/ocs/4492/4492-21843-1-PB.pdf)
  - Winogrande (https://winogrande.allenai.org/)
    - 273 pronoun resolution problems
    - With "twin sentences" (similar sentences with different outcomes, & trigger word that enables the change)

Winogrande examples

| | | Twin sentences | Options (**answer**) |
|---|---|---|---|
| ✓ (1) | a | The trophy doesn't fit into the brown suitcase because **it**'s too *large*. | **trophy** / suitcase |
| | b | The trophy doesn't fit into the brown suitcase because **it**'s too *small*. | trophy / **suitcase** |
| ✓ (2) | a | Ann asked Mary what time the library closes, *because* **she** had forgotten. | **Ann** / Mary |
| | b | Ann asked Mary what time the library closes, *but* **she** had forgotten. | Ann / **Mary** |
| ✗ (3) | a | The tree fell down and crashed through the roof of my house. Now, I have to get **it** *removed*. | **tree** / roof |
| | b | The tree fell down and crashed through the roof of my house. Now, I have to get **it** *repaired*. | tree / **roof** |
| ✗ (4) | a | The lions ate the zebras because **they** are *predators*. | **lions** / zebras |
| | b | The lions ate the zebras because **they** are *meaty*. | lions / **zebras** |

# Causal reasoning

- Determine cause/effect relationships in common scenarios.
- Datasets
  - COPA (Choice of Plausible Alternatives)
    - https://cdn.aaai.org/ocs/2418/2418-10878-1-PB.pdf
    - 1,000 questions
    - Premise + 2 alternatives
  - ATOMIC (An Atlas of Machine Commonsense for If-Then Reasoning)
    - https://arxiv.org/pdf/1811.00146v3
    - 877K descriptions of if-then relations
      - If-even-then-event,
      - if-event-then-mental-state,
      - …

COPA examples

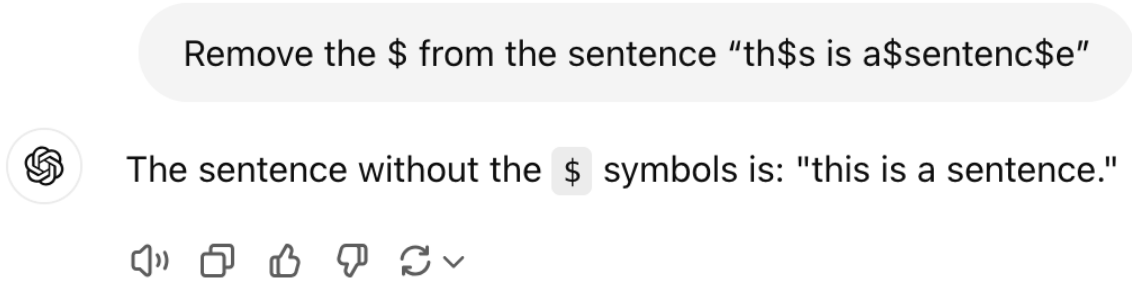
(forward causal reasoning)
Premise: The man lost his balance on the ladder. *What happened as a result?*
Alternative 1: He fell off the ladder.
Alternative 2: He climbed up the ladder.

(backwards causal reasoning)
Premise: The man fell unconscious. *What was the cause of this?*
Alternative 1: The assailant struck the man in the head.
Alternative 2: The assailant took the man's wallet.

ATOMIC examples

| | | |
|---|---|---|
| "PersonX makes PersonY's coffee" | If-Event-Then-Mental-State | PersonX wanted to be helpful<br>PersonY will be appreciative<br>PersonY will be grateful |
| | If-Event-Then-Event | PersonX needs to put the coffee in the filter<br>PersonX gets thanked<br>PersonX adds cream and sugar |
| | If-Event-Then-Persona | PersonX is helpful<br>PersonX is deferential |

# Synthetic tasks




- We can generate specific tasks synthetically
- The task may be easy to generate, but non-trivial for the models to solve
  - What is 4123 + 9421?
  - Remove the $ from the sentence “th$s is a$sentenc$e”
  - Unscrable the word “aplpe”
- Can be useful to test specific model capabilities, and generate data the model is guaranteed to have never seen

# Benchmarks

# Benchmarking LLMs

- There are various famous benchmarks for LLMs
- They cover a wide variety of datasets, and tasks
- The purpose is to provide a well-rounded evaluation of models

# GLUE

- General Language Understanding Evaluation
- Provides a “single number” evaluation
- Combines performance across various tasks
- https://gluebenchmark.com

https://gluebenchmark.com/tasks

**GLUE Tasks**

| Name | Download | More Info | Metric |
|---|---|---|---|
| The Corpus of Linguistic Acceptability | | | Matthew's Corr |
| The Stanford Sentiment Treebank | | | Accuracy |
| Microsoft Research Paraphrase Corpus | | | F1 / Accuracy |
| Semantic Textual Similarity Benchmark | | | Pearson-Spearman Corr |
| Quora Question Pairs | | | F1 / Accuracy |
| MultiNLI Matched | | | Accuracy |
| MultiNLI Mismatched | | | Accuracy |
| Question NLI | | | Accuracy |
| Recognizing Textual Entailment | | | Accuracy |
| Winograd NLI | | | Accuracy |
| Diagnostics Main | | | Matthew's Corr |

https://openreview.net/pdf?id=rJ4km2R5t7

# SuperGLUE

- "performance on [GLUE] has recently come close to the level of non-expert humans"
- SuperGLUE Introduces more difficult tasks

https://super.gluebenchmark.com/tasks

**SuperGLUE Tasks**

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

https://w4ngatang.github.io/static/papers/superglue.pdf

# MMLU

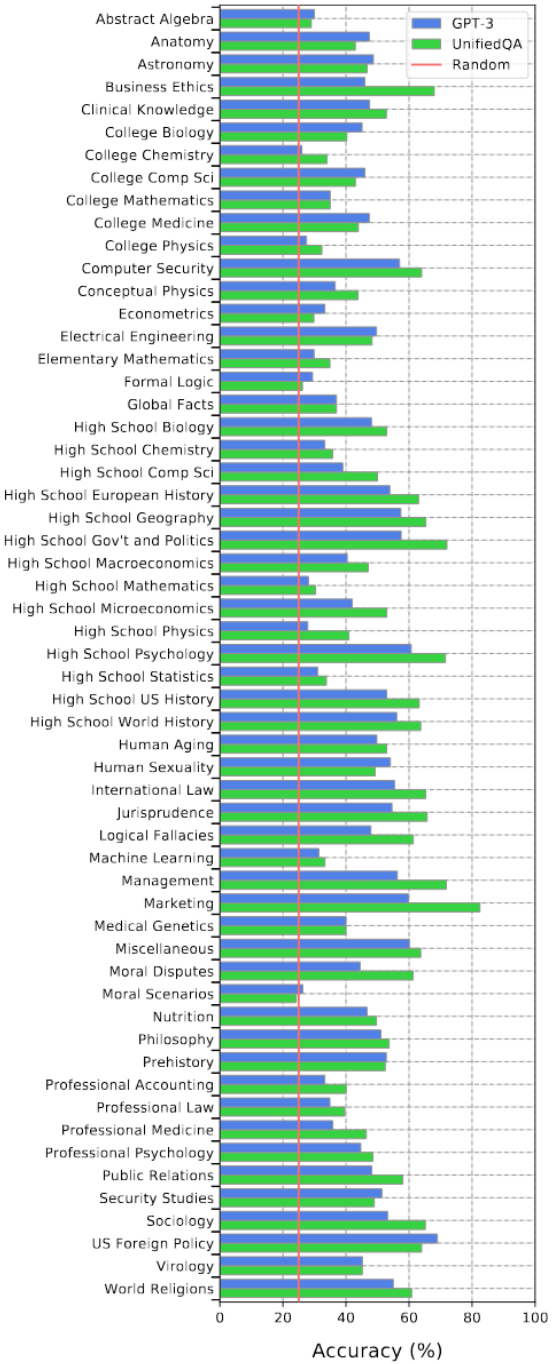
- Massive Multitask Language Understanding
- Focused on Question Answering
- ~16k question/answers (4 options)
- Covers 57 topics, including:
  - Mathematics,
  - Astronomy,
  - Philosophy,
  - Law,
  - ...

**Conceptual Physics**

When you drop a ball from rest it accelerates downward at 9.8 m/s². If you instead throw it downward assuming no air resistance its acceleration immediately after leaving your hand is
(A) 9.8 m/s² ✓
(B) more than 9.8 m/s² ✗
(C) less than 9.8 m/s² ✗
(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 ✓




https://arxiv.org/pdf/2009.03300

# LLM contamination

- Since datasets and benchmarks are public data, it may happen that they end up in the training corpus of LLMs
  - Intenionally, or not…
- For closed LLMs, this problem cannot be proved (training corpus is private)
- However, the continuous improvements across tasks may be somewhat related to this kind of *contamination*
- *Contamination can occur at various levels*
  - (e.g. during pre-training, or supervised fine-tuning, after deployment)

Sainz, Oscar, et al. "Nlp evaluation in trouble: On the need to measure llm data contamination for each benchmark." *arXiv preprint arXiv:2310.18018* (2023). https://arxiv.org/pdf/2310.18018