Evaluating LLMs

CS 780/880 Natural Language Processing Lecture 24

Samuel Carton, University of New Hampshire

Last Lecture

Zero and few shot learning

Anatomy of a prompt

OpenAI API

Exemplar choice

Role-setting

BERT

System	MNLI-(m/mm)	QQP	QNLI	SST-2	CoLA	STS-B	MRPC	RTE	**Average**
	392k	363k	108k	67k	8.5k	5.7k	3.5k	2.5k	-
Pre-OpenAI SOTA	80.6/80.1	66.1	82.3	93.2	35.0	81.0	86.0	61.7	74.0
BiLSTM+ELMo+Attn	76.4/76.1	64.8	79.8	90.4	36.0	73.3	84.9	56.8	71.0
OpenAI GPT	82.1/81.4	70.3	87.4	91.3	45.4	80.0	82.3	56.0	75.1
$BERT_{BASE}$	84.6/83.4	71.2	90.5	93.5	52.1	85.8	88.9	66.4	79.6
$BERT_{LARGE}$	**86.7/85.9**	**72.1**	**92.7**	**94.9**	**60.5**	**86.5**	**89.3**	**70.1**	**82.1**

Devlin, Jacob, et al. "Bert: Pre-training of deep bidirectional transformers for language understanding." *arXiv preprint arXiv:1810.04805* (2018).

RoBERTa

Model	SQuAD 1.1/2.0	MNLI-m	SST-2	RACE
Our reimplementation (with NSP loss):				
SEGMENT-PAIR	90.4/78.7	84.0	92.9	64.2
SENTENCE-PAIR	88.7/76.2	82.9	92.1	63.0
Our reimplementation (without NSP loss):				
FULL-SENTENCES	90.4/79.1	84.7	92.5	64.8
DOC-SENTENCES	90.6/79.7	84.7	92.7	65.6
$\text{BERT}_{\text{BASE}}$	88.5/76.3	84.3	92.8	64.3
$\text{XLNet}_{\text{BASE}}$ (K = 7)	–/81.3	85.8	92.7	66.1
$\text{XLNet}_{\text{BASE}}$ (K = 6)	–/81.0	85.6	93.4	66.7

Liu, Yinhan, et al. "Roberta: A robustly optimized bert pretraining approach." *arXiv preprint arXiv:1907.11692* (2019).

GPT-4

	GPT-4 Evaluated few-shot	GPT-3.5 Evaluated few-shot	LM SOTA Best external LM evaluated few-shot	SOTA Best external model (incl. benchmark-specific tuning)
MMLU [49] Multiple-choice questions in 57 subjects (professional & academic)	**86.4%** 5-shot	70.0% 5-shot	70.7% 5-shot U-PaLM [50]	75.2% 5-shot Flan-PaLM [51]
HellaSwag [52] Commonsense reasoning around everyday events	**95.3%** 10-shot	85.5% 10-shot	84.2% LLaMA (validation set) [28]	85.6 ALUM [53]
AI2 Reasoning Challenge (ARC) [54] Grade-school multiple choice science questions. Challenge-set.	**96.3%** 25-shot	85.2% 25-shot	85.2% 8-shot PaLM [55]	86.5% ST-MOE [18]
WinoGrande [56] Commonsense reasoning around pronoun resolution	**87.5%** 5-shot	81.6% 5-shot	85.1% 5-shot PaLM [3]	85.1% 5-shot PaLM [3]
HumanEval [43] Python coding tasks	**67.0%** 0-shot	48.1% 0-shot	26.2% 0-shot PaLM [3]	65.8% CodeT + GPT-3.5 [57]
DROP [58] (F1 score) Reading comprehension & arithmetic.	80.9 3-shot	64.1 3-shot	70.8 1-shot PaLM [3]	**88.4** QDGAT [59]
GSM-8K [60] Grade-school mathematics questions	**92.0%*** 5-shot chain-of-thought	57.1% 5-shot	58.8% 8-shot Minerva [61]	87.3% Chinchilla + SFT+ORM-RL, ORM reranking [62]

Achiam, Josh, et al. "Gpt-4 technical report." *arXiv preprint arXiv:2303.08774* (2023).

Llama 3

Meta Llama 3 Pre-trained model performance

	Meta Llama 3 8B	Mistral 7B Published	Mistral 7B Measured	Gemma 7B Published	Gemma 7B Measured
MMLU 5-shot	**66.6**	62.5	63.9	64.3	64.4
AGIEval English 3-5-shot	**45.9**	--	44.0	41.7	44.9
BIG-Bench Hard 3-shot, CoT	**61.1**	--	56.0	55.1	59.0
ARC-Challenge 25-shot	78.6	78.1	78.7	53.2 0-shot	**79.1**
DROP 3-shot, F1	**58.4**	--	54.4	--	56.3

	Meta Llama 3 70B	Gemini Pro 1.0 Published	Mixtral 8x22B Measured
MMLU 5-shot	**79.5**	71.8	77.7
AGIEval English 3-5-shot	**63.0**	--	61.2
BIG-Bench Hard 3-shot, CoT	**81.3**	75.0	79.2
ARC-Challenge 25-shot	**93.0**	--	90.7
DROP 3-shot, F1	**79.7**	74.1 variable-shot	77.6

https://ai.meta.com/blog/meta-llama-3/

Gemini

	Gemini Ultra	Gemini Pro	GPT-4	GPT-3.5	PaLM 2-L	Claude 2	Inflect-ion-2	Grok 1	LLAMA-2
MMLU Multiple-choice questions in 57 subjects (professional & academic) (Hendrycks et al., 2021a)	**90.04%** CoT@32* 83.7% 5-shot	79.13% CoT@8* 71.8% 5-shot	87.29% CoT@32 (via API**) 86.4% 5-shot (reported)	70% 5-shot	78.4% 5-shot	78.5% 5-shot CoT	79.6% 5-shot	73.0% 5-shot	68.0%***
GSM8K Grade-school math (Cobbe et al., 2021)	**94.4%** Maj1@32	86.5% Maj1@32	92.0% SFT & 5-shot CoT	57.1% 5-shot	80.0% 5-shot	88.0% 0-shot	81.4% 8-shot	62.9% 8-shot	56.8% 5-shot
MATH Math problems across 5 difficulty levels & 7 subdisciplines (Hendrycks et al., 2021b)	**53.2%** 4-shot	32.6% 4-shot	52.9% 4-shot (via API**) 50.3% (Zheng et al., 2023)	34.1% 4-shot (via API**)	34.4% 4-shot	—	34.8%	23.9% 4-shot	13.5% 4-shot
BIG-Bench-Hard Subset of hard BIG-bench tasks written as CoT problems (Srivastava et al., 2022)	**83.6%** 3-shot	75.0% 3-shot	83.1% 3-shot (via API**)	66.6% 3-shot (via API**)	77.7% 3-shot	—	—	—	51.2% 3-shot
HumanEval Python coding tasks (Chen et al., 2021)	**74.4%** 0-shot (PT****)	67.7% 0-shot (PT****)	67.0% 0-shot (reported)	48.1% 0-shot	—	70.0% 0-shot	44.5% 0-shot	63.2% 0-shot	29.9% 0-shot
Natural2Code Python code generation. (New held-out set with no leakage on web)	**74.9%** 0-shot	69.6% 0-shot	73.9% 0-shot (via API**)	62.3% 0-shot (via API**)	—	—	—	—	—
DROP Reading comprehension & arithmetic. (metric: F1-score) (Dua et al., 2019)	**82.4** Variable shots	74.1 Variable shots	80.9 3-shot (reported)	64.1 3-shot	82.0 Variable shots	—	—	—	—
HellaSwag (validation set) Common-sense multiple choice questions (Zellers et al., 2019)	87.8% 10-shot	84.7% 10-shot	**95.3%** 10-shot (reported)	85.5% 10-shot	86.8% 10-shot	—	89.0% 10-shot	—	80.0%***
WMT23 Machine translation (metric: BLEURT) (Tom et al., 2023)	**74.4** 1-shot (PT****)	71.7 1-shot	73.8 1-shot (via API**)	—	72.7 1-shot	—	—	—	—

Team, Gemini, et al. "Gemini: a family of highly capable multimodal models." *arXiv preprint arXiv:2312.11805* (2023).

Older benchmarks

SST-2

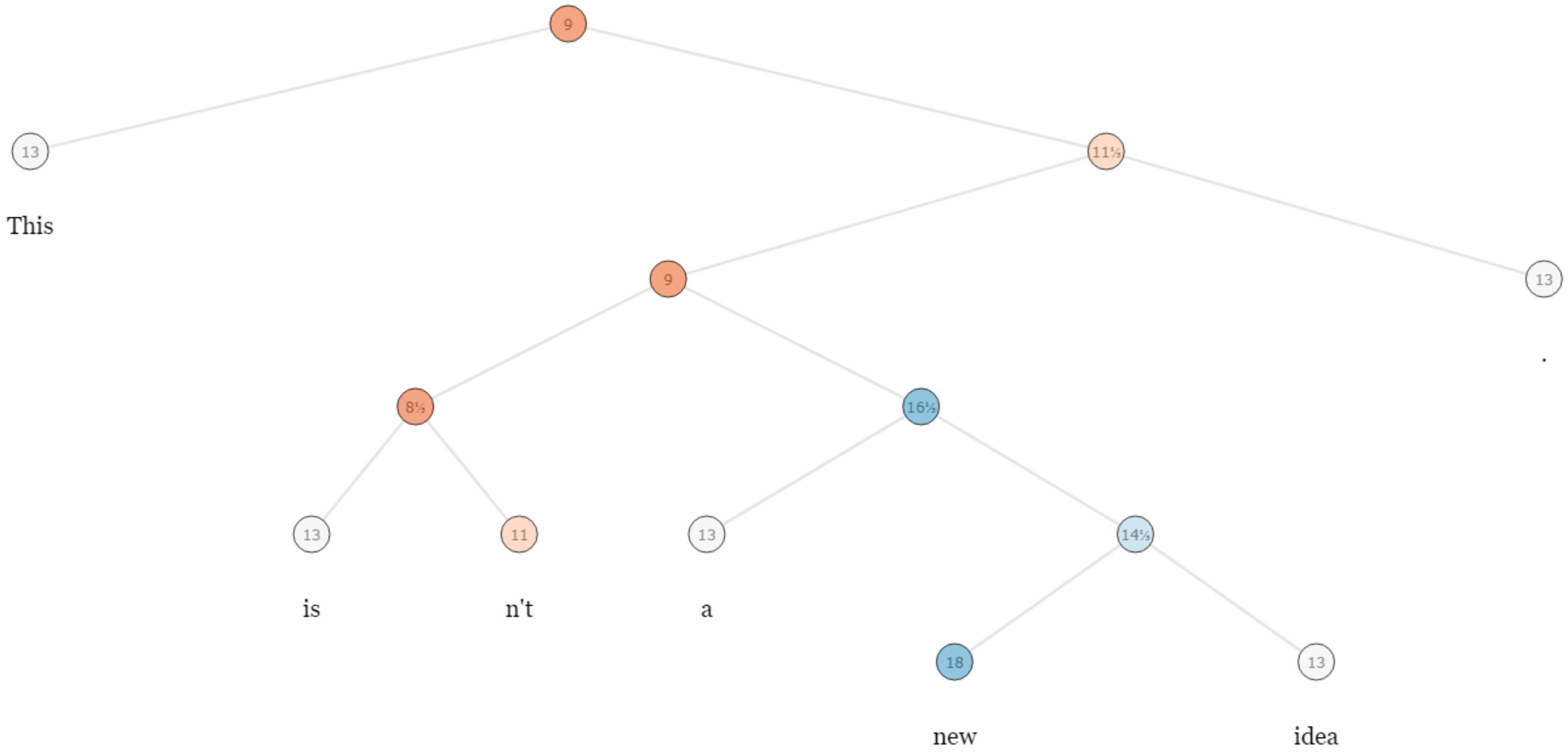
Sentiment detection

Very familiar to us at this point

Original is actually 5-class, and evaluated at **every** grammatical clause of each text

So we've been looking at the flattened, binarized version (SST-2)

Binary classification: Acc/F1/P/R



Socher, Richard, et al. "Recursive deep models for semantic compositionality over a sentiment treebank." *Proceedings of the 2013 conference on empirical methods in natural language processing*. 2013.

https://nlp.stanford.edu/sentiment/treebank.html

Multi-Genre Natural Language Inference (MNLI)

Natural language inference (NLI)

AKA entailment

Basic idea: take a **premise** and a **hypothesis**, and classify whether the premise **entails** the hypothesis

Ends up as 3-class classification: Acc/F1/P/R

Premise	Genre	Hypothesis
Met my first girlfriend that way.	FACE-TO-FACE **contradiction** C C N C	I didn't meet my first girlfriend until later.
8 million in relief in the form of emergency housing.	GOVERNMENT **neutral** N N N N	The 8 million dollars for emergency housing was still not enough to solve the problem.
Now, as children tend their gardens, they have a new appreciation of their relationship to the land, their cultural heritage, and their community.	LETTERS **neutral** N N N N	All of the children love working in their gardens.
At 8:34, the Boston Center controller received a third transmission from American 11	9/11 **entailment** E E E E	The Boston Center controller got a third transmission from American 11.
I am a lacto-vegetarian.	SLATE **neutral** N N E N	I enjoy eating cheese too much to abstain from dairy.
someone else noticed it and i said well i guess that's true and it was somewhat melodious in other words it wasn't just you know it was really funny	TELEPHONE **contradiction** C C C C	No one noticed and it wasn't funny at all.

Williams, Adina, Nikita Nangia, and Samuel R. Bowman. "A broad-coverage challenge corpus for sentence understanding through inference." *arXiv preprint arXiv:1704.05426* (2017).

The Corpus of Linguistic Acceptability (CoLA)

Corpus of 10,657 English sentences labeled as grammatical vs ungrammatical

Binary classification: Acc/P/R/F1

Label	Sentence	Source
*	The more books I ask to whom he will give, the more he reads.	Culicover and Jackendoff (1999)
✓	I said that my father, he was tight as a hoot-owl.	Ross (1967)
✓	The jeweller inscribed the ring with the name.	Levin (1993)
*	many evidence was provided.	Kim and Sells (2008)
✓	They can sing.	Kim and Sells (2008)
✓	The men would have been all working.	Baltin (1982)
*	Who do you think that will question Seamus first?	Carnie (2013)
*	Usually, any lion is majestic.	Dayal (1998)
✓	The gardener planted roses in the garden.	Miller (2002)
✓	I wrote Blair a letter, but I tore it up before I sent it.	Rappaport Hovav and Levin (2008)

Table 3: CoLA random sample, drawn from the in-domain training set (✓= acceptable, *=unacceptable).

Warstadt, Alex, Amanpreet Singh, and Samuel R. Bowman. "Neural network acceptability judgments." *Transactions of the Association for Computational Linguistics* 7 (2019): 625-641.

The Stanford Question Answering Dataset (SQUAD)

Reading comprehension dataset

A bunch of Wikipedia articles, along with questions asked about them where the answer is a snippet from the article

Sequence tagging task: F1

Other versions available: https://rajpurkar.github.io/SQuAD-explorer/

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

What causes precipitation to fall?
gravity

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

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

Figure 1: Question-answer pairs for a sample passage in the SQuAD dataset. Each of the answers is a segment of text from the passage.

Rajpurkar, Pranav, et al. "Squad: 100,000+ questions for machine comprehension of text." *arXiv preprint arXiv:1606.05250* (2016).

Newer benchmarks

Massive Multitask Language Understanding (MMLU)

Multiple-choice test of world knowledge across 57 areas.

Math, history, medicine, physics, etc

Because it’s multiple choice, ends up just being 4-class classification: Acc/P/R/F1

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/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 ✓

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

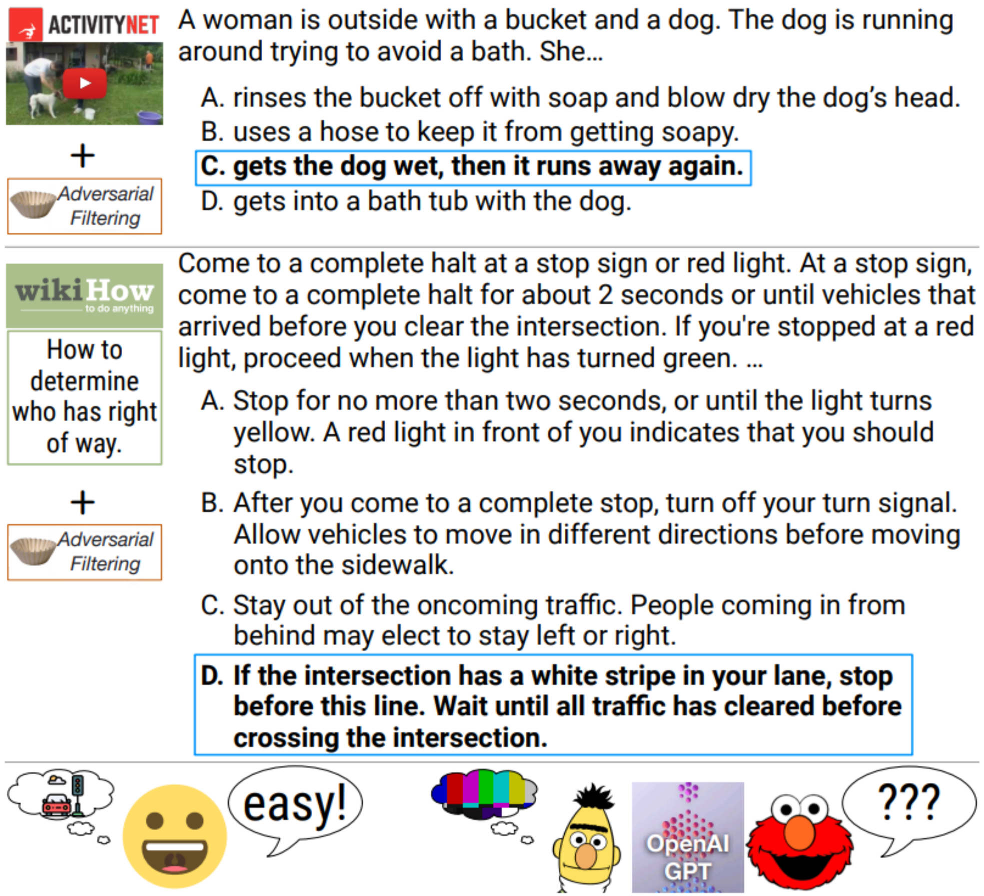
Hendrycks, Dan, et al. "Measuring massive multitask language understanding." *arXiv preprint arXiv:2009.03300* (2020).

HellaSwag

Commonsense natural language inference, i.e. can the model figure out the right continuation of a sentence, based on common sense

Collected from WikiHow using **adversarial filtering** to make it harder for models

Multiple-choice, so 4-class classification: Acc/P/R/F1



***HellaSwag*: Can a Machine *Really* Finish Your Sentence?**

In this paper, we show that commonsense inference still proves difficult for even state-of-the-art models, by presenting *HellaSwag*,

Zellers, Rowan, et al. "Hellaswag: Can a machine really finish your sentence?." *arXiv preprint arXiv:1905.07830* (2019).

AI2 Reasoning Challenge (ARC)

7787 science exam questions of various genres

Divided into challenge set and easy set

Multiple-choice, so Acc/P/R/F1

Knowledge Type	Example
Definition	What is a worldwide increase in temperature called? (A) greenhouse effect (B) global warming (C) ozone depletion (D) solar heating
Basic Facts & Properties	Which element makes up most of the air we breathe? (A) carbon (B) nitrogen (C) oxygen (D) argon
Structure	The crust, the mantle, and the core are structures of Earth. Which description is a feature of Earth's mantle? (A) contains fossil remains (B) consists of tectonic plates (C) is located at the center of Earth (D) has properties of both liquids and solids
Processes & Causal	What is the first step of the process in the formation of sedimentary rocks? (A) erosion (B) deposition (C) compaction (D) cementation
Teleology / Purpose	What is the main function of the circulatory system? (1) secrete enzymes (2) digest proteins (3) produce hormones (4) transport materials
Algebraic	If a red flowered plant (RR) is crossed with a white flowered plant (rr), what color will the offspring be? (A) 100% pink (B) 100% red (C) 50% white, 50% red (D) 100% white
Experiments	Scientists perform experiments to test hypotheses. How do scientists try to remain objective during experiments? (A) Scientists analyze all results. (B) Scientists use safety precautions. (C) Scientists conduct experiments once. (D) Scientists change at least two variables.
Spatial / Kinematic	In studying layers of rock sediment, a geologist found an area where older rock was layered on top of younger rock. Which best explains how this occurred? (A) Earthquake activity folded the rock layers...

Clark, Peter et al. "Think you have Solved Question Answering? Try ARC, the AI2 Reasoning Challenge." *ArXiv* abs/1803.05457 (2018): n. pag.

GSM8K

Collection of 8500 grade school math problems

Multiple choices **not** included in this dataset, so when not working with generative model, use a **generator/verifier** set up to propose multiple solutions and then pick one.

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

Cobbe, Karl, et al. "Training verifiers to solve math word problems." *arXiv preprint arXiv:2110.14168* (2021).

Discrete Reasoning Over the content of Paragraphs (DROP)

96k passages with questions and correct answers

Collected using live existing model (BiDAF), where annotators had to pick answers the BiDAF couldn't get correct

Reasoning	**Passage** (some parts shortened)	**Question**	**Answer**	**BiDAF**
Subtraction (28.8%)	That year, his **Untitled (1981)**, a painting of a haloed, black-headed man with a bright red skeletal body, depicted amid the artists signature scrawls, was **sold by Robert Lehrman for $16.3 million, well above its $12 million high estimate**.	How many more dollars was the Untitled (1981) painting sold for than the 12 million dollar estimation?	4300000	$16.3 million
Comparison (18.2%)	In **1517, the seventeen-year-old King sailed to Castile**. There, his Flemish court .... **In May 1518, Charles traveled to Barcelona in Aragon**.	Where did Charles travel to first, Castile or Barcelona?	Castile	Aragon
Selection (19.4%)	In 1970, to commemorate the 100th anniversary of the founding of Baldwin City, **Baker University professor and playwright Don Mueller and Phyllis E. Braun, Business Manager, produced a musical play entitled The Ballad Of Black Jack** to tell the story of the events that led up to the battle.	Who was the University professor that helped produce The Ballad Of Black Jack, Ivan Boyd or Don Mueller?	Don Mueller	Baker

Dua, Dheeru, et al. "DROP: A reading comprehension benchmark requiring discrete reasoning over paragraphs." *arXiv preprint arXiv:1903.00161* (2019).

HumanEval

A coding benchmark where the goal is to write correct code for a python function, given the python docstring

Each problem includes a bunch of unit tests the solution has to pass to be correct

```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)
```

Chen, Mark, et al. "Evaluating large language models trained on code." *arXiv preprint arXiv:2107.03374* (2021).

Beyond the Imitation Game Benchmark (BIG-bench)

More than 200 tasks collected into one big benchmark collection

Intended to be a holistic benchmark of overall LLM capability

auto_debugging	known_unknowns	parsinlu_reading_comprehension
bbq_lite_json	language_identification	play_dialog_same_or_different
code_line_description	linguistics_puzzles	repeat_copy_logic
conceptual_combinations	logic_grid_puzzle	strange_stories
conlang_translation	logical_deduction	strategyqa
emoji_movie	misconceptions_russian	symbol_interpretation
formal_fallacies_...	novel_concepts	vitaminc_fact_verification
hindu_knowledge	operators	winowhy

Table 1: The 24 tasks included in BIG-bench Lite, a diverse subset of JSON tasks that can be evaluated cheaply.

https://github.com/google/BIG-bench

Srivastava, Aarohi, et al. "Beyond the imitation game: Quantifying and extrapolating the capabilities of language models." *arXiv preprint arXiv:2206.04615* (2022).

AGIEval

A bunch of questions drawn from standardized tests in English and Chinese

Exams	#Participants	Language	Tasks	Subject
Gaokao	12M	Chinese	GK-geography GK-biology GK-history GK-chemistry GK-physics GK-En GK-Ch GK-Math-QA GK-Math-Cloze	Geography Biology History Chemistry Physics English Chinese Math Math
SAT	1.7M	English	SAT-En. SAT-Math	English Math
Lawyer Qualification Test	820K	Chinese	JEC-QA-KD JEC-QA-CA	Law Law
Law School Admission Test (LSAT)	170K	English	LSAT-AR LSAT-LR LSAT-RC	Law-Analytics Law-Logic Law-Reading
Civil Service Examination	2M 2M	English Chinese	LogiQA-en LogiQA-ch	Logic Logic
GRE GMAT	340K 150K	English English	AQuA-RAT	Math
AMC AIME	300K 3000	English English	MATH	Math

Zhong, Wanjun, et al. "Agieval: A human-centric benchmark for evaluating foundation models." *arXiv preprint arXiv:2304.06364* (2023).

Other benchmarks

Lots of other datasets have been introduced over the years

- GLUE
  - Includes CoLA, SST-2, MNLI, etc.
  - https://gluebenchmark.com/
- SuperGLUE
- ERASER
  - Explainability datasets
  - https://www.eraserbenchmark.com/

Concluding thoughts

LLM evaluation is hard! Some researchers spend a lot of their career coming up with new datasets

Special data collection tricks to create hard examples for LLMs

Historical movement from low-level linguistic capabilities (i.e. inference, sentiment) to high-level capabilities (world knowledge, solving logic puzzles)

Movement from classification to generation