Course Logistics

Syllabus has been post on course website

- **Lecture structure**
  - First five weeks are notebook-guided lectures
  - Last five weeks are guest lectures
- **Grading breakdown**
  - 25% participation
  - 75% final project
- **No homework/assignment!**

We value participation!

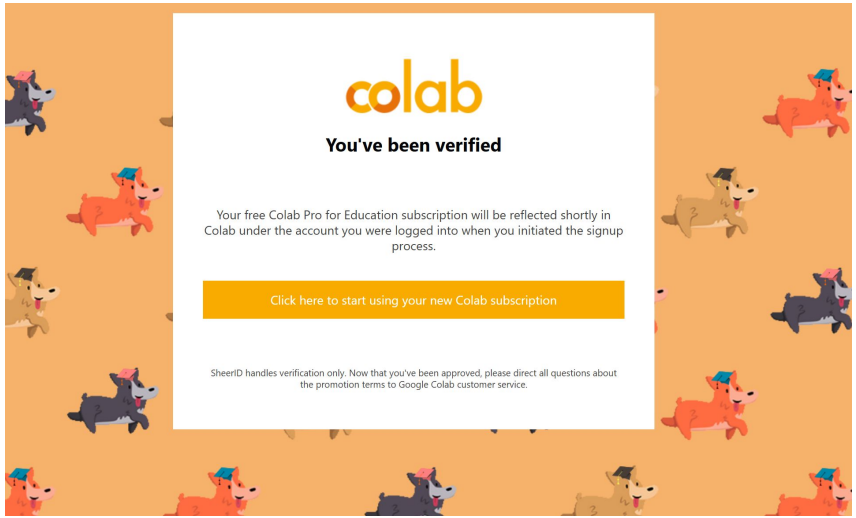
- **Starting Wednesday, Apr 8th, we will take attendance during lectures**
  - Please arrive on time!
- **How to make up for absence**
  - Email us before the lecture!
  - For the notebook-guided lectures, completing the exercises in the notebook
  - For the guest lectures, write a 200-word response to the readings suggested by the speaker

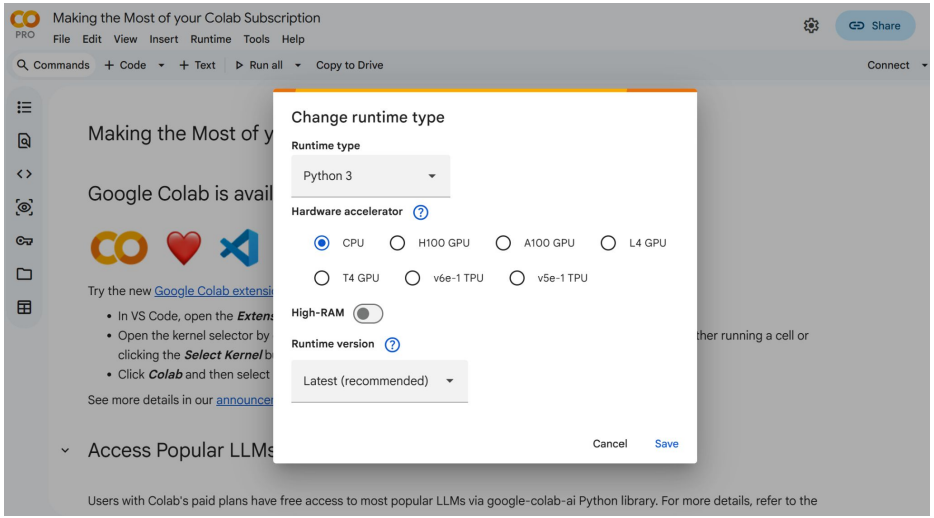
More info about project on Wednesday

- **Project:** Replicating results from interpretability papers or implementing your own research ideas
- **Collabration**: 1~3 student group with assigned mentors

Compute resources

Colab Pro for Education subscription





Questions?

Behavioral Analysis and Input Attribution

CS 221M Lecture
Jing Huang
Apr 6th, 2026

Connections to the bigger picture

Why does the system make this prediction?

What algorithms does the system use?

Does the algorithm used by the system at least have the right input features?

Lecture plan

Motivations

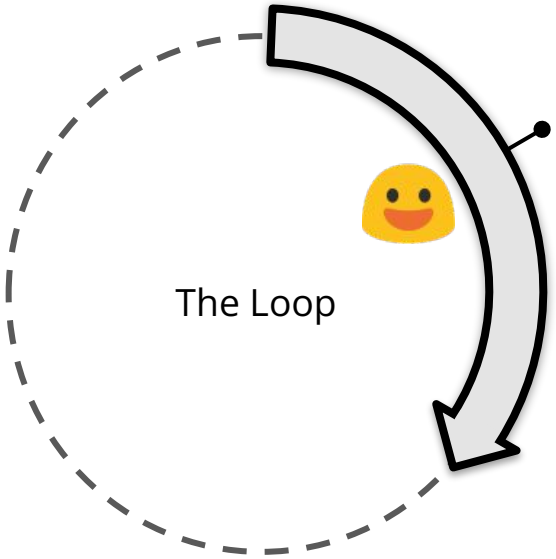
- A short history of benchmarking language models
- Performance vs. Competence

Behavioral testing with minimal pairs

- What are minimal pairs
- How to design minimal pairs

Gradient-based input attributions

Motivations: The LLM development cycle



Phase 1: Create a benchmark for a task that none of the existing models can solve.

WMT

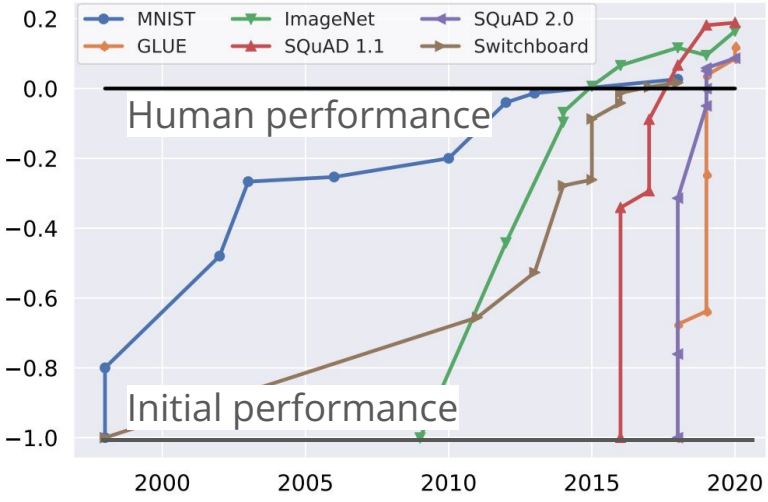
GLUE SuperGLUE

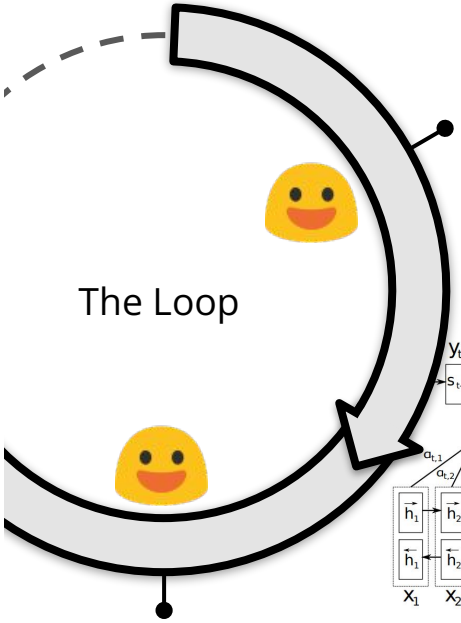
SQuAD

cais/mmlu

openai/gsm8k

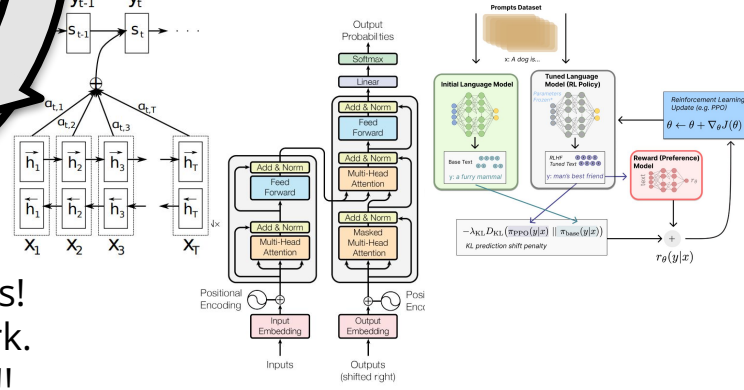
Motivations: The LLM development cycle





Phase 1: Create a benchmark for a task that none of the existing models can solve.

Phase 2: Build new models!
Evaluate on the benchmark.
The benchmark is solved!!



Kiela et al. Dynabench: Rethinking Benchmarking in NLP. 2021.

Case study: Natural language inference task

Task: Given a premise and a hypothesis, determine whether the premise entails, contradicts, is neutral to the hypothesis.

Premise: The judge was paid by th
Hypothesis: The actor paid the ju
Label: Entailment

Case study: Natural language inference task

Premise: The judge was paid by the actor.
Hypothesis: The actor paid the judge.

Label: Entailment
Prediction: Entailment

Premise: The judge was paid by the actor.
Hypothesis: The **judge** paid the **actor**.

Label: Contradict
Prediction: Entailment

Lexical overlap heuristic: Assume that a premise entails all hypotheses constructed from words in the premise.

McCoy et al. Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference. 2019.

Motivations: The LLM development cycle

Phase 3: Discover that models can't solve simple variations of

SYNTHETIC AND NATURAL NOISE BOTH BREAK NEURAL MACHINE TRANSLATION

WMT

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

Stress Test Evaluation for Natural Language Inference

GLUE SuperGLUE

Adversarial Examples for Evaluating Reading Comprehension Systems

SQuAD

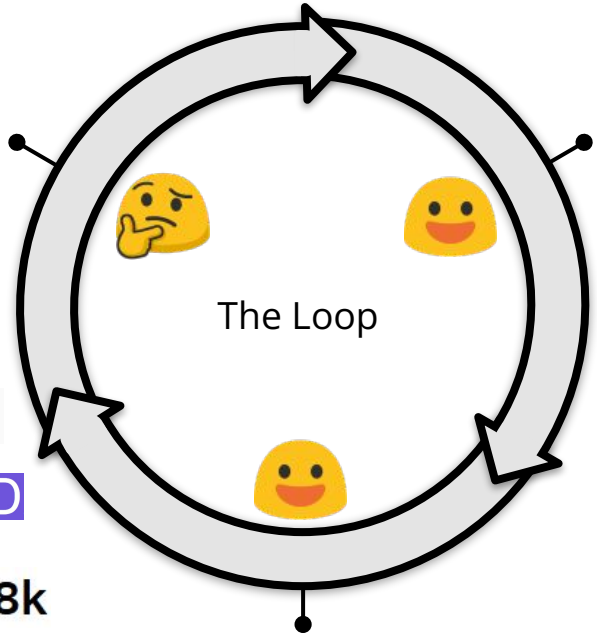
GSM-SYMBOLIC: UNDERSTANDING THE LIMITATIONS OF MATHEMATICAL REASONING IN LARGE LANGUAGE MODELS

openai/gsm8k

LARGE LANGUAGE MODELS ARE NOT ROBUST MULTIPLE CHOICE SELECTORS

cais/mmlu

MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark



Phase 1: Create a benchmark for a task that none of the existing models can solve.

Phase 2: Build new models! Evaluate on the benchmark. The benchmark is solved!!

Performance != Competence

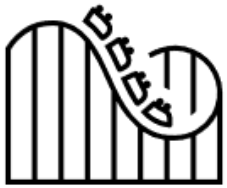
WMT

GLUE SuperGLUE

SQuAD

cais/mmlu

openai/gsm8k



SYNTHETIC AND NATURAL NOISE BOTH BREAK NEURAL MACHINE TRANSLATION

WMT

Right for the Wrong Reasons: Diagnosing Syntactic Heuristics in Natural Language Inference

Stress Test Evaluation for Natural Language Inference

GLUE SuperGLUE

Adversarial Examples for Evaluating Reading Comprehension Systems

SQuAD

GSM-SYMBOLIC: UNDERSTANDING THE LIMITATIONS OF MATHEMATICAL REASONING IN LARGE LANGUAGE MODELS

openai/gsm8k

LARGE LANGUAGE MODELS ARE NOT ROBUST MULTIPLE CHOICE SELECTORS

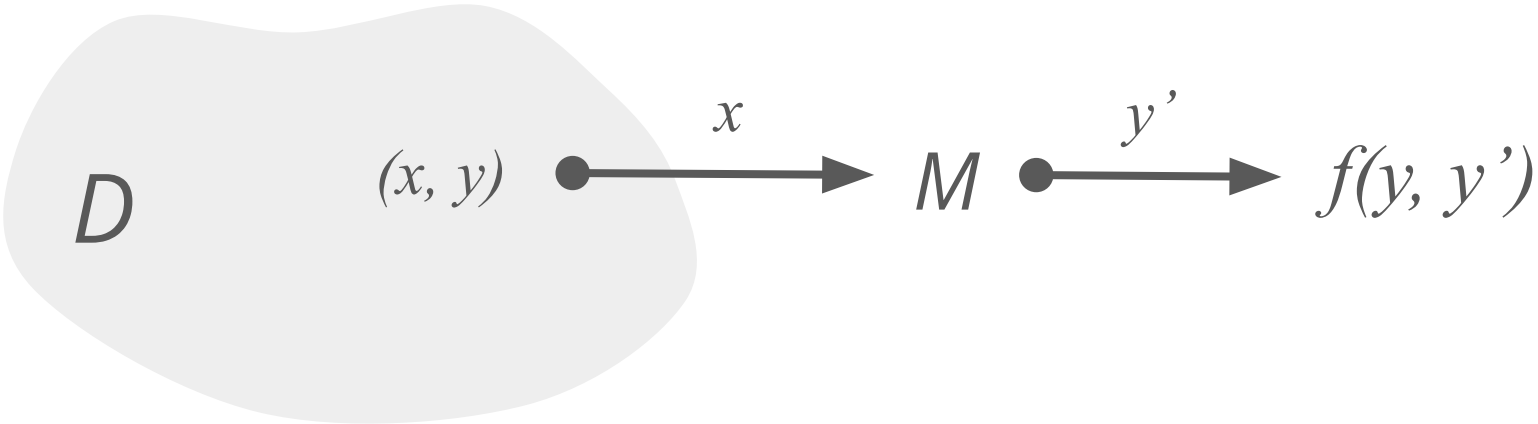
cais/mmlu

MMLU-Pro: A More Robust and Challenging Multi-Task Language Understanding Benchmark

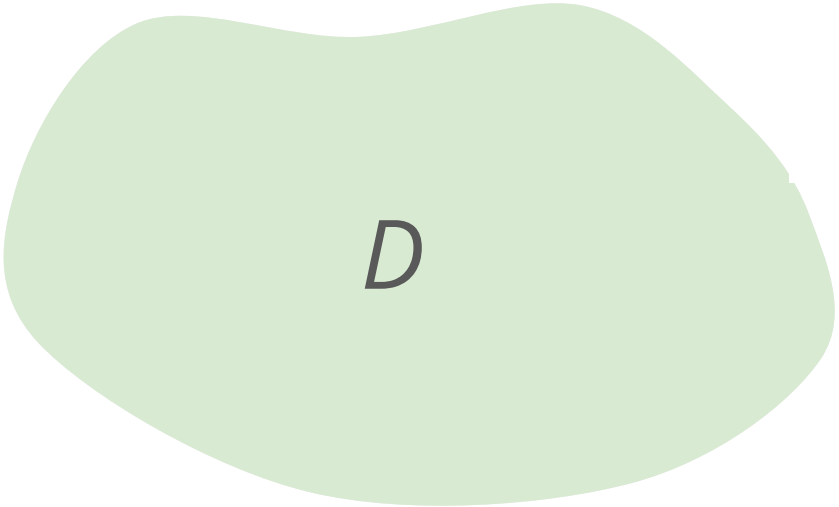
What is a task?

Task (informally)

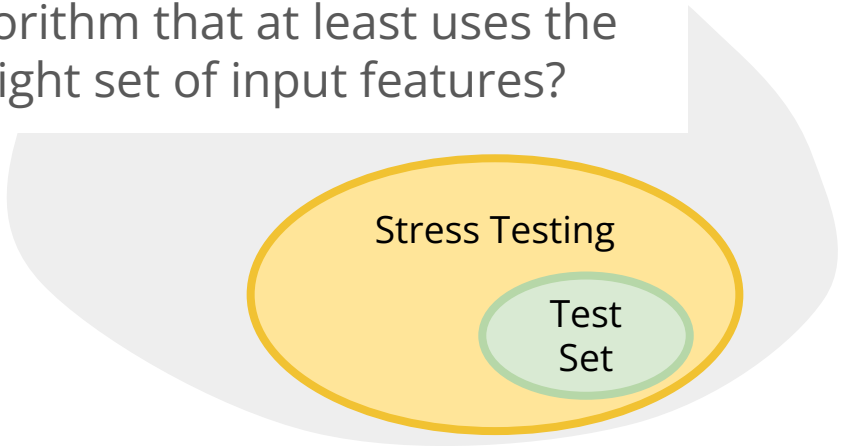
- **A data distribution *D***, e.g, {(a pair of sentences, entail), ...}
- **An objective function *f***, e.g., classification accuracy



Performance != Competence



Stress Testing: Behaviorally, is the model solving the task with an algorithm that at least uses the right set of input features?



Competence ≈ Whether a model can solve the task, i.e., behaviorally, predicts every instance of D correctly or internally, implements the correct algorithm

Performance = Whether a model can correctly predict instances of the test set

Questions?

Behavioral testing with minimal pairs

Minimal Pairs: A pair of sequences that differ by only one “feature”

Principle: Controlled variation, i.e., one feature change at a time

Behavioral testing with minimal pairs

Minimal Pairs: A pair of sequences that differ by only one “feature”

Principle: Controlled variation, i.e., one feature change at a time

Two types of features

- Background features: Changing their values should not change the model prediction.
- Causal features: Changing their value should change the model prediction in a predictable way.

Designing minimal pairs: Perturbing background features

Case study: Inserting irrelevant content to SQuAD

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?"
Prediction: Prague

Jia and Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. 2017.

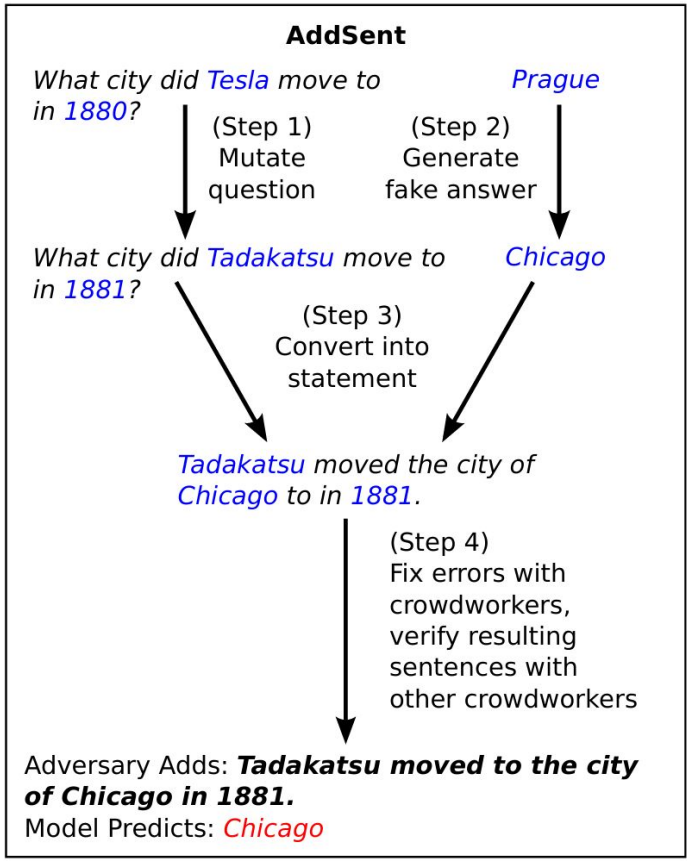
Case study: Inserting irrelevant content to SQuAD

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?"
Prediction: Prague

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. Tadakatsu moved to the city of Chicago in 1881."
Question: "What city did Tesla move to in 1880?"
Prediction: Chicago

Jia and Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. 2017.

How to construct systematic test cases



Template generation: Creating "slots" for controlled variation

Verify resulting example: Ensuring the change doesn't introduce confounding variables

Jia and Liang. Adversarial Examples for Evaluating Reading Comprehension Systems. 2017.

Case study: Primitive subsitutions in GSM8K

GSM8K

When Sophie watches her nephew, she gets out a variety of toys for him. The bag of building blocks has 31 blocks in it. The bin of stuffed animals has 8 stuffed animals inside. The tower of stacking rings has 9 multicolored rings on it.Sophie recently bought a tube of bouncy balls, bringing her total number of toys for her nephew up to 62. How many bouncy balls came in the tube?

Template generation

GSM Symbolic Template

When {name} watches her {family}, she gets out a variety of toys for him. The bag of building blocks has {x} blocks in it. The bin of stuffed animals has {y} stuffed animals inside.The tower of stacking rings has {z} multicolored rings on it.{name} recently bought a tube of bouncy balls, bringing her total number of toys she bought for her {family} up to {total}. How many bouncy balls came in the tube?

```
#variables:
- name   = sample(names)
- family = sample(["nephew", "cousin", "brother"])
- x      = range(5, 100)
- y      = range(5, 100)
- z      = range(5, 100)
- total  = range(100, 500)
- ans    = range(85, 200)

#conditions:
- x + y + z + ans == total
```

Mirzadeh et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. 2024.

Case study: Primitive subsitutions in GSM8K

GSM8K

When Sophie watches her nephew, she gets out a variety of toys for him. The bag of building blocks has 31 blocks in it. The bin of stuffed animals has 8 stuffed animals inside. The tower of stacking rings has 9 multicolored rings on it.Sophie recently bought a tube of bouncy balls, bringing her total number of toys for her nephew up to 62. How many bouncy balls came in the tube?

GSM Symbolic Template

Whe {name} wa ches he {family}, he gets out a variety of oys for hi . The ag of buil ng blocks has {x} blocks in it. The bin of stuffed animals has {y} stuffed animals inside.The tow r of stacki g rings has {z} multicolored rings on t.{name} re ently bought a tube of bouncy ba , bringi her total number of toys she bought for h r {family} p to {total}. How many bouncy balls came i the tube?

Background Variables

```
#variables:
- name   = sample(names)
- family = sample(["nephew",
- x      = range(5, 100)
- y      = range(5, 100)
- z      = range(5, 100)
- total  = range(100, 500)
- ans    = range(85, 200)

#conditions:
- x + y + z + ans == total
```

Mirzadeh et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. 2024.

Designing minimal pairs: Perturbing causal features

Case study: Primitive subsitutions in GSM8K

GSM8K

When Sophie watches her nephew, she gets out a variety of toys for him. The bag of building blocks has 31 blocks in it. The bin of stuffed animals has 8 stuffed animals inside. The tower of stacking rings has 9 multicolored rings on it.Sophie recently bought a tube of bouncy balls, bringing her total number of toys for her nephew up to 62. How many bouncy balls came in the tube?

GSM Symbolic Template

When {name} watches her {family}, she gets out a variety of toys for him. The bag of building block has {x} blocks in it. The bin of stuffed animals has {y} stuffed animals inside.The tower of stacking rings has {z} multicolored rings on it.{name} recently bought a tube of bouncy balls, bringing her total number of toys she bought for her {family} up to {total}. How many bouncy balls came in the tube?

Causal Variables

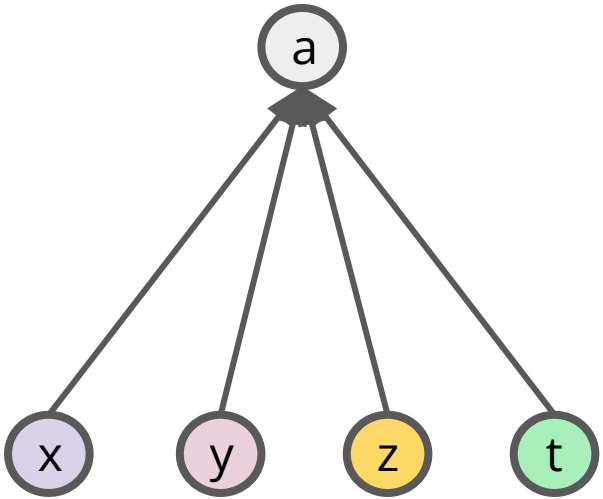
```
#variables:
- name   = sample(names)
- family = sample(["nephew", "cousin", "brother"])
- x      = range(5, 100)
- y      = range(5, 100)
- z      = range(5, 100)
- total  = range(100, 500)
- ans    = range(85, 200)

#conditions:
- x + y + z + ans == total
```

Mirzadeh et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. 2024.

Case study: Primitive subsitutions in GSM8K

a = t - x - y - z



A Causal Model

GSM Symbolic Template

```
When {name} watches her {family}, she gets out a variety
of toys for him.  The bag of building blocks has {x}
blocks in it.  The bin of stuffed animals has {y} stuffed
animals inside.The tower of stacking rings has {z}
multicolored rings on it.{name} recently bought a tube
of bouncy balls, bringing her total number of toys she
bought for her {family} up to {total}.  How many bouncy
balls came in the tube?

#variables:
- name   = sample(names)
- family = sample(["nephew", "cousin", "brother"])
- x      = range(5, 100)
- y      = range(5, 100)
- z      = range(5, 100)
- total  = range(100, 500)
- ans    = range(85, 200)

#conditions:
- x + y + z + ans == total
```

Mirzadeh et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. 2024.

Case study: Primitive subsitutions in GSM8K

Template generation: Creating "slots" for controlled variation

Verify resulting example: Ensuring the change doesn't introduce confounding variables

GSM Symbolic Template

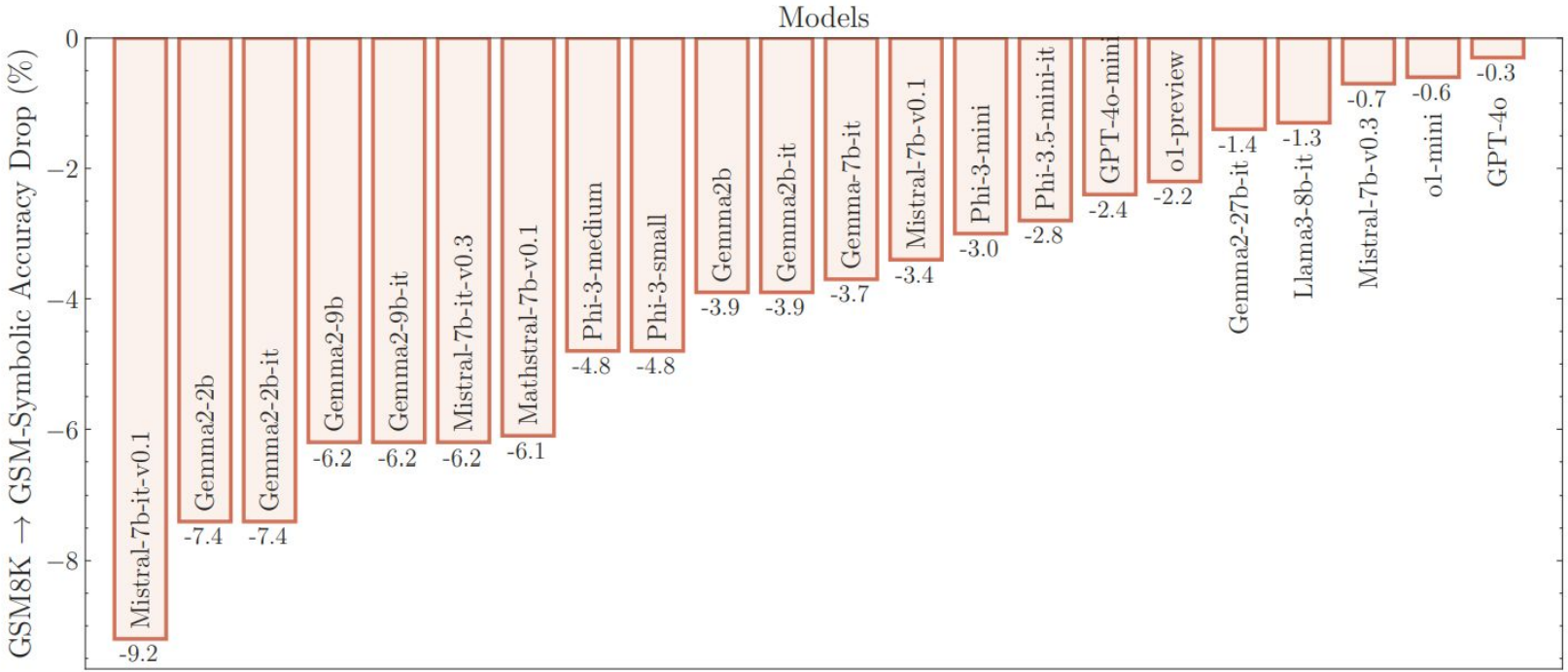
```
When {name} watches her {family}, she gets out a variety
of toys for him.  The bag of building blocks has {x}
blocks in it.  The bin of stuffed animals has {y} stuffed
animals inside.The tower of stacking rings has {z}
multicolored rings on it.{name} recently bought a tube
of bouncy balls, bringing her total number of toys she
bought for her {family} up to {total}.  How many bouncy
balls came in the tube?

#variables:
- name   = sample(names)
- family = sample(["nephew", "cousin", "brother"])
- x      = range(5, 100)
- y      = range(5, 100)
- z      = range(5, 100)
- total  = range(100, 500)
- ans    = range(85, 200)

#conditions:
- x + y + z + ans == total
```

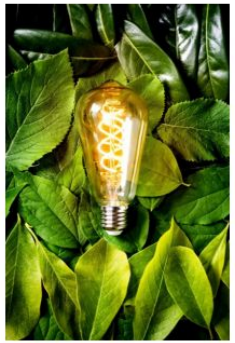
Mirzadeh et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. 2024.

Case study: Primitive subsitutions in GSM8K



Mirzadeh et al. GSM-Symbolic: Understanding the Limitations of Mathematical Reasoning in Large Language Models. 2024.

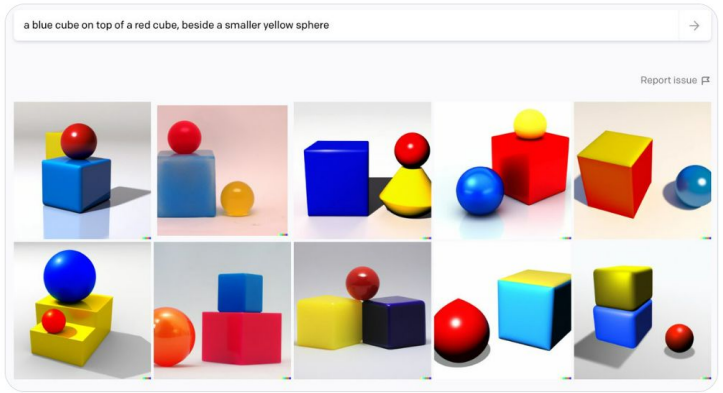
Case study: Compositionality in VLMs



(a) some plants surrounding a lightbulb

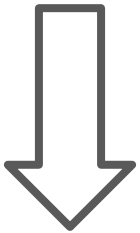
(b) a lightbulb surrounding some plants

Prompt: a blue cube on top of a red cube, beside a smaller yellow sphere



Thrush et al. Winoground: Probing Vision and Language Models for Visio-Linguistic Compositionality. 2022.
Twitter post: https://twitter.com/david_madras/status/1512573390896480267?lang=en

Does the system make a correct prediction?



Why does the system make this prediction?

Notebook demo: Design minimal pairs for MMLU

Goal: Designing minimal pairs to stress test an LLM on MMLU

Notebook demo: Design minimal pairs for MMLU

MMLU: A multiple-choice question answering test that covers 57 tasks including mathematics, US history, computer science, law, and more.

An example from MMLU

Find the degree for the given field extension Q(sqrt(2), sqrt(3), sqrt(18)) over Q.

A. 0
B. 4
C. 2
D. 6

Answer:

Notebook demo: Design minimal pairs for MMLU

Test model: The highest performing 7~8B scale model, i.e., Phi-3 7B model

Center for Research on Foundation Models HELM MMLU

Leaderboard: College Mathematics

Model	EM	Observed inference time (s)	# eval	# train	truncated
Gemini 1.0 Pro (001)	0.51	0.356			
PaLM-2 (Unicorn)	0.51	1.062			
GPT-4o (2024-08-06)	0.51	0.482			
Palmyra X V3 (72B)	0.51	0.591			
Llama 3.2 Vision Instruct Turbo (90B)	0.5	0.315	100	5	0
Llama 3.1 Instruct Turbo (70B)	0.49	4.479	100	5	0
Claude 3 Haiku (20240307)	0.49	0.715	100	5	0
Phi-3 (7B)	0.48	0.351	100	5	0
Mixtral (8x22B)	0.48	0.382	100	5	0
Yi (34B)	0.47	0.569	100	5	0

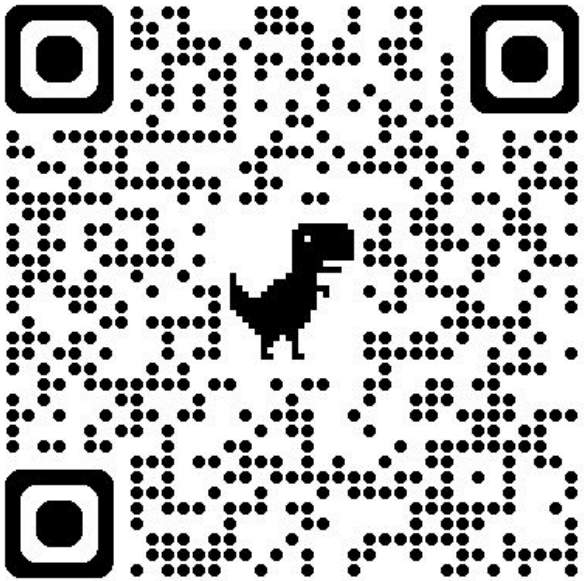
Notebook demo: Design minimal pairs for MMLU

Hypothesis I: A model produces the correct answer by implementing the correct algorithm.

Hypothesis II: A model produce the correct answer by memorize the contaminated instances in pre-training data.

Notebook demo: Design minimal pairs for MMLU

https://github.com/cs221m/cs221m-course/blob/main/03_behavioral_analysis.ipynb



Gradient-based input attribution

Goal: Understand how much each input feature contribute to model predictions

Intuition: A larger gradient at the feature value means the feature value has large impact on the output

A simple method: Input × Gradient

$$\text{InputXGradient}_i(M, x) = \frac{\partial M(x)}{\partial x_i} \cdot x_i$$

$$\frac{y_1 - y_2}{x_1 - x_2}$$

Implicitly contrasting the output of a minimal pair!

Dataset: https://scikit-learn.org/stable/datasets/toy_dataset.html

A simple method: Input × Gradient

$$\text{InputXGradient}_i(M, x) = \frac{\partial M(x)}{\partial x_i} \cdot x_i$$

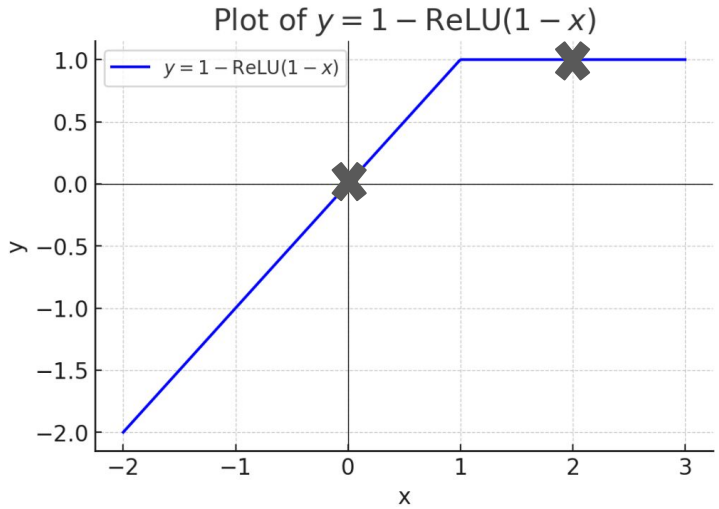
Interpreting a linear model: $y = w_1 x_1 + w_2 x_2$

- Larger weight on a feature, the more sensitive the model prediction is to the feature

Dataset: https://scikit-learn.org/stable/datasets/toy_dataset.html

Problems with Input × Gradient

$$\text{InputXGradient}_i(M, x) = \frac{\partial M(x)}{\partial x_i} \cdot x_i$$



Consider a minimal pair x=0 and x = 2. What happens if we evaluate at x=2?

Gradient being zero does not necessarily mean the attribution should be zero!

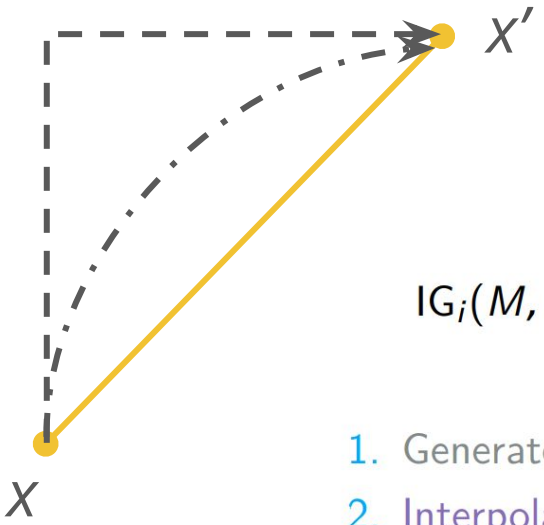
Figure taken from CS224U slides made by Prof. Christopher Potts. 2023.

Integrated gradient

Axioms of feature attribution

- Sensitivity: If the representations of two inputs x and x’ only differ at dimension i, and the two inputs lead to different predictions, then the feature at dimension i has non-zero attribution
- Implementation invariance: If two models have identical input-output behaviors, they should have the same attribution

Integrated gradient



$$\mathrm{IG}_i(M, x, x') = \overbrace{(x_i - x_i')}^{5} \cdot \overbrace{\sum_{k=1}^{m}}^{4} \frac{\partial M(x' + \overbrace{\frac{k}{m}}^{1} \cdot (x - x'))}{\partial x_i} \cdot \overbrace{\frac{1}{m}}^{4}$$

3

2

1. Generate $\alpha = [1, \ldots, m]$
2. Interpolate inputs between baseline x' and actual input x
3. Compute gradients for each interpolated input
4. Integral approximation through averaging
5. Scaling to remain in the space region as the original

Figure taken from CS224U slides made by Prof. Christopher Potts. 2023.

Questions?