

CSCI 5541: Natural Language Processing

Lecture 14: Data and Evaluation



UNIVERSITY OF MINNESOTA
Driven to Discover®



Outline

□ Annotation

- Annotation: terms, examples, and process
- Qualitative coding, recruitment, and quality assessment
- Issues in annotation
- Advanced annotation techniques
- LLMs as Annotators and Synthetic Data

□ Evaluation:

- Properties of Good Benchmarks
- Widely Used Benchmarks and their Metrics
- Latest Benchmarks



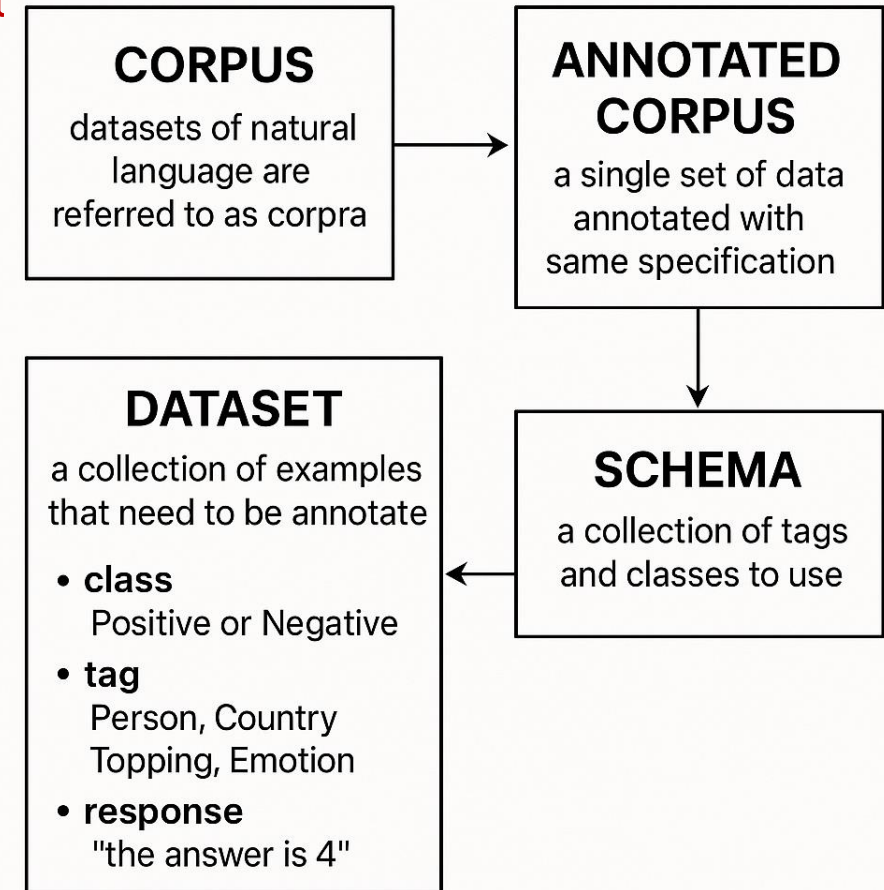
Annotation

- ❑ Just providing large amounts of data doesn't help the model learn to speak. The data needs to be **guided** in such a way that the computer can more easily find patterns and inferences.
- ❑ High-quality data means high-performance algorithms
- ❑ Any **metadata (e.g., tags, structures, categories, orders)** used to mark up elements of the dataset is called **annotation**.
- ❑ But, in order for the algorithms to learn efficiently and effectively, the annotation must be **accurate**, and **relevant** to the task the machine is being asked to perform.

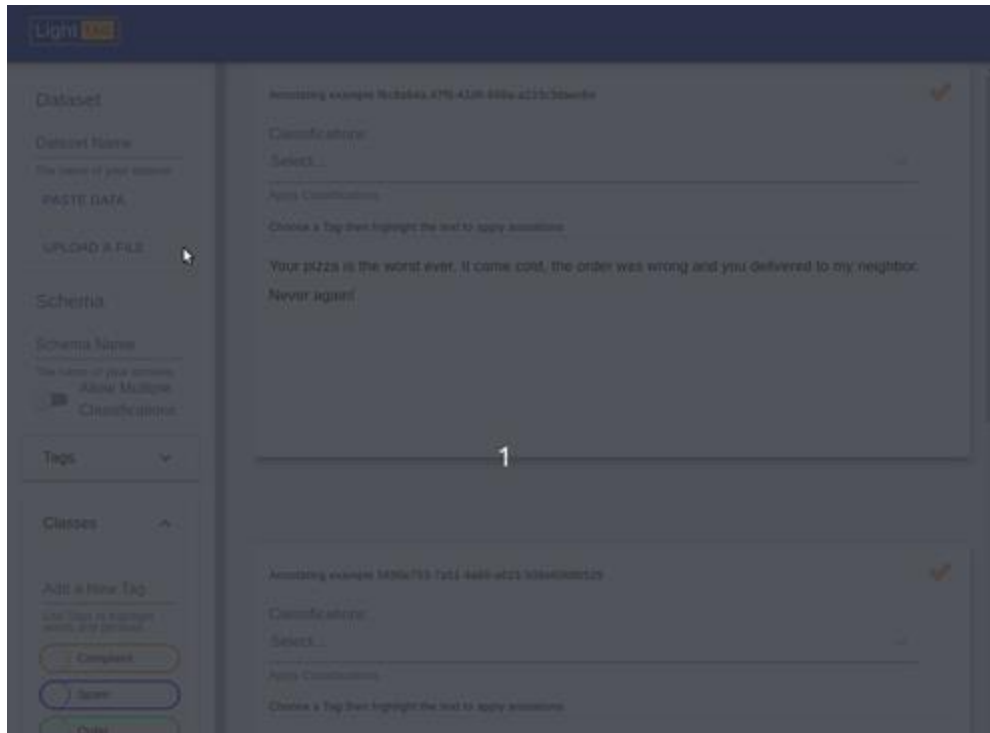


Terms

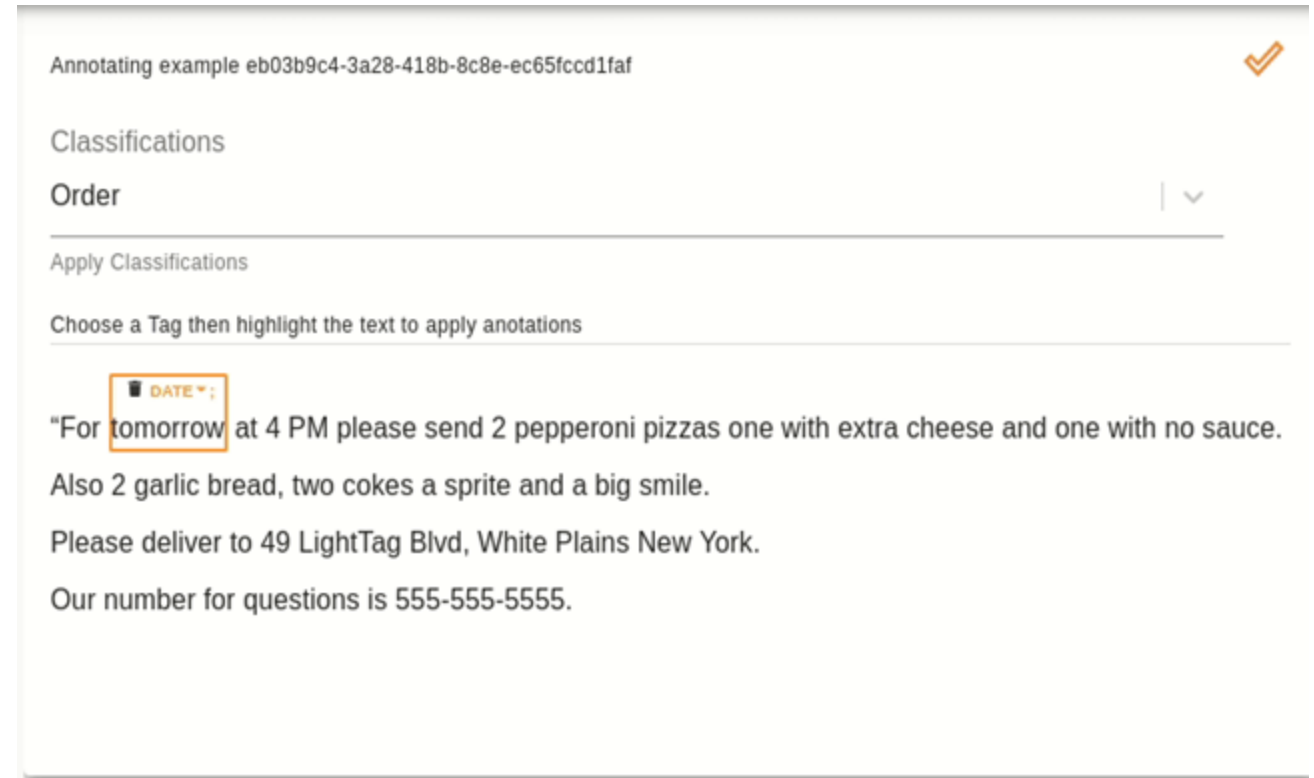
- ❑ Datasets of natural language are referred to as **corpora**
- ❑ A single set of data annotated with the *same specification* is called an **annotated corpus**.
- ❑ A **dataset** is a collection of examples that need to be annotated.
 - A **class** is a particular classification option.
 - ✓ E.g., Positive or Negative and email can be Spam or Ham.
 - A **tag** is a description name for an entity type.
 - ✓ E.g., Person (Jane), Country (Madagascar), Topping (Pepperoni) and Emotion (Fascinated).
 - A **response** to particular question or prompt
 - ✓ E.g., "the answer is 4"
- ❑ A **schema**
 - Everyone to use the same collection of tags and classes or pick and choose their own tags and classes.



Types of annotations

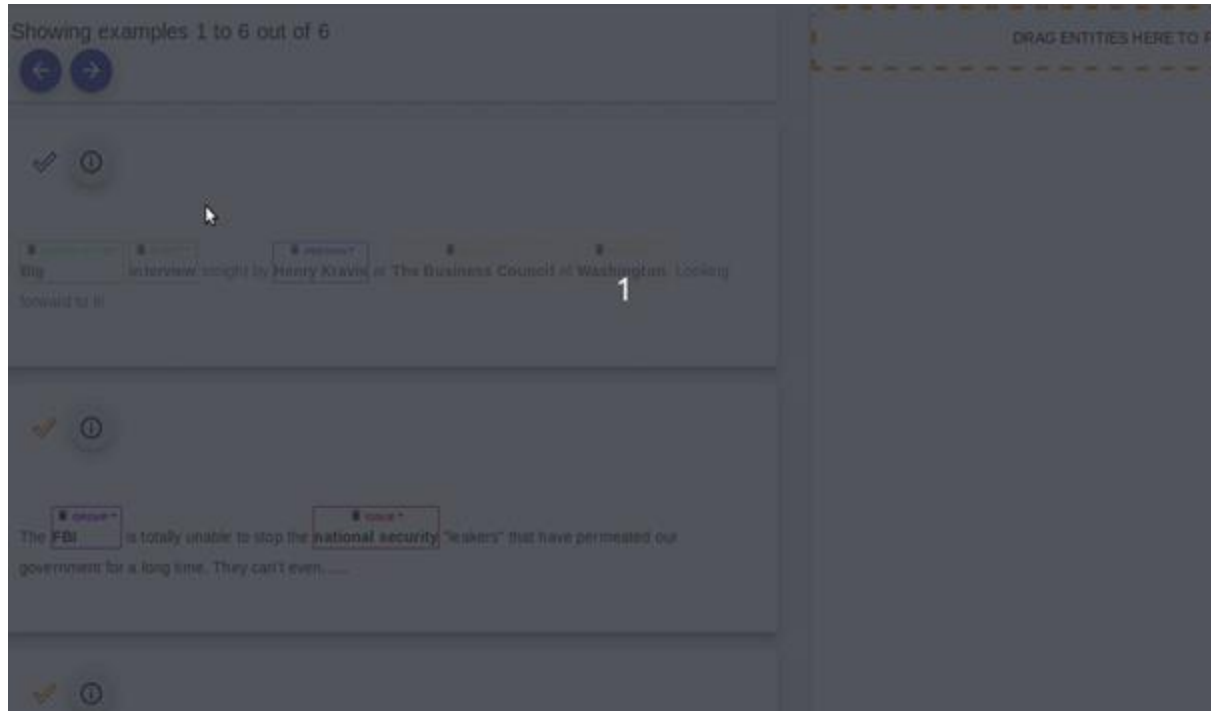


Document classification



Entity annotation

Types of annotations



Relation annotation

UTX ^{Pro} ^{Ca} ^{Catalysis} Th ^{Demethylation} ^{Site} Th ^{Pr} ^{Ent}
 UTX mediates demethylation of H3K27me3 at muscle-specific genes during myogenesis.

Polycomb (PcG) and Trithorax (TrxG) group proteins act antagonistically to establish tissue-specific patterns of gene expression.

The PcG protein Ezh2 facilitates repression by ^{Catalysis} ^{Protein} ^{Entity} ^{Methylation} ^{Pr} ^{Ent}
 The PcG protein Ezh2 facilitates repression by catalysing histone H3-Lys27 trimethylation (H3K27me3).

For expression, H3K27me3 marks are removed and replaced by TrxG protein catalysed histone H3-Lys4 trimethylation (H3K4me3).

Although H3K27 demethylases have been identified, the mechanism by which these enzymes are targeted to specific genomic regions has not been established.

Here, we demonstrate a two-step mechanism for UTX-mediated demethylation at muscle-specific genes during myogenesis.

Although the transactivator Six4 initially recruits UTX to the regulatory region of muscle genes, the resulting loss of H3K27me3 n upstream of the transcriptional start site.

Removal of the repressive H3K27me3 mark within the coding region then requires RNA Polymerase II (Pol II) elongation.

Discourse relation annotation

Types of annotations

Premise

Russian cosmonaut Valery Polyakov set the record for the longest amount of time spent in space.

Hypothesis

Russians hold the record for the longest stay in space.

Target

Entailment
Not entailment



Options:

- yes
- no



Template 1

Russian Cosmonaut Valery Polyakov set the record for the longest amount of time spent in space.

Based on the paragraph above, can we conclude that

Russians hold the record for the longest stay in space?

OPTIONS

- yes
- no

Template 2

Read the following and determine if the hypothesis can be inferred from the premise:

Premise: **<premise>**

Hypothesis: **<hypothesis>**
<options>

Template 3, ...



Types of annotations

Sentiment: Positive

Preference

I came here and ordered
and it was exceptional.

Text A

Mario and Carlos cut our
meat exactly the way we
wanted the property

Text B

Text A
is more
positive



Questions for collecting the ideal dataset?

- ❑ What is the **target accuracy** you are looking for?
- ❑ Can it be achieved it by **better models or more data**?
 - How many annotations are enough to ensure high accuracies?
- ❑ How **representative** is your dataset?
 - domain vocabulary, format, genre of the text, etc
- ❑ How **subjective** is your task? Humans agree among each other?
- ❑ Is your dataset **balanced**, containing instances of each class?
- ❑ How **clean** is your dataset?

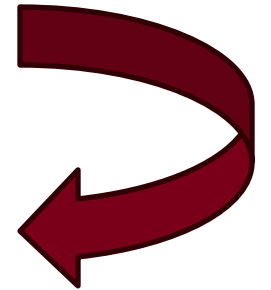
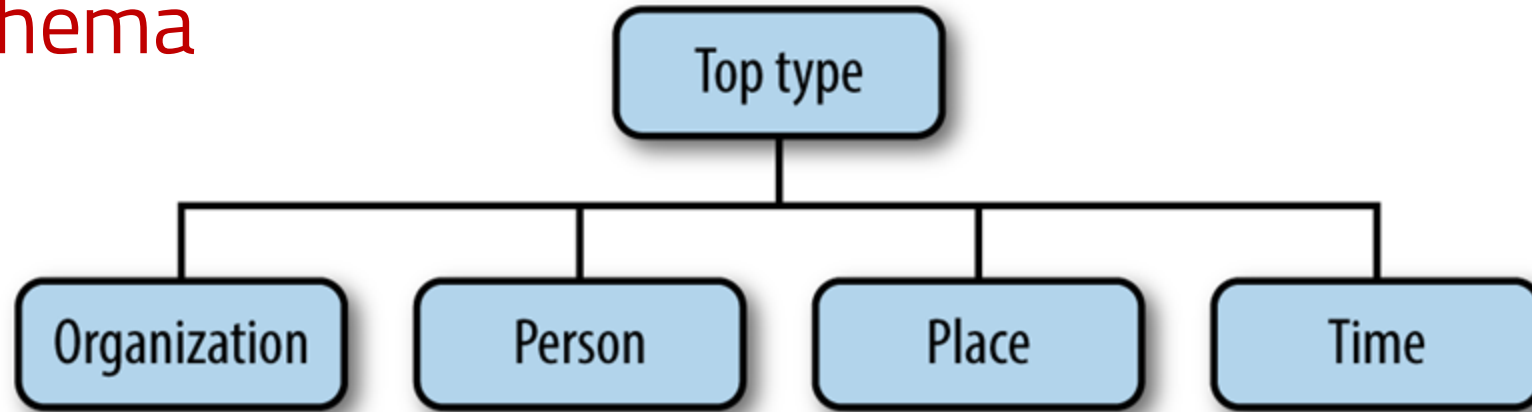


Example Annotation:

Semantic Type/Role Labeling



Schema



Ms. Ramirez of QBC Productions visited Boston on Saturday, where she had lunch with Mr. Harris of STU Enterprises at 1:15 pm.

Semantic Types

[Ms. Ramirez]_{Person} of [QBC Productions]_{Organization} visited
[Boston]_{Place} on [Saturday]_{Time}, where she had lunch with [Mr.
Harris]_{Person} of [STU Enterprises]_{Organization} at [1:15 pm]_{Time}.



Semantic Role Labeling (More fine-grained Semantics)

- Basics for Question Answering,
 - the who, what, where, and when of a sentence.

Schema

Agent	The event participant that is doing or causing the event to occur
Theme/figure	The event participant who undergoes a change in position or state
Experiencer	The event participant who experiences or perceives something
Source	The location or place from which the motion begins; the person from whom the theme is given
Goal	The location or place to which the motion is directed or terminates
Recipient	The person who comes into possession of the theme
Patient	The event participant who is affected by the event
Instrument	The event participant used by the agent to do or cause the event
Location/ground	The location or place associated with the event itself



The man painted the wall with a paint brush.

Mary walked to the café from her house.

John gave his mother a necklace.

My brother lives in Milwaukee.



[The man]_{agent} painted [the wall]_{patient} with [a paint brush]_{instrument}.

[Mary]_{figure} walked to [the cafe]_{goal} from [her house]_{source}.

[John]_{agent} gave [his mother]_{recipient} [a necklace]_{theme}.

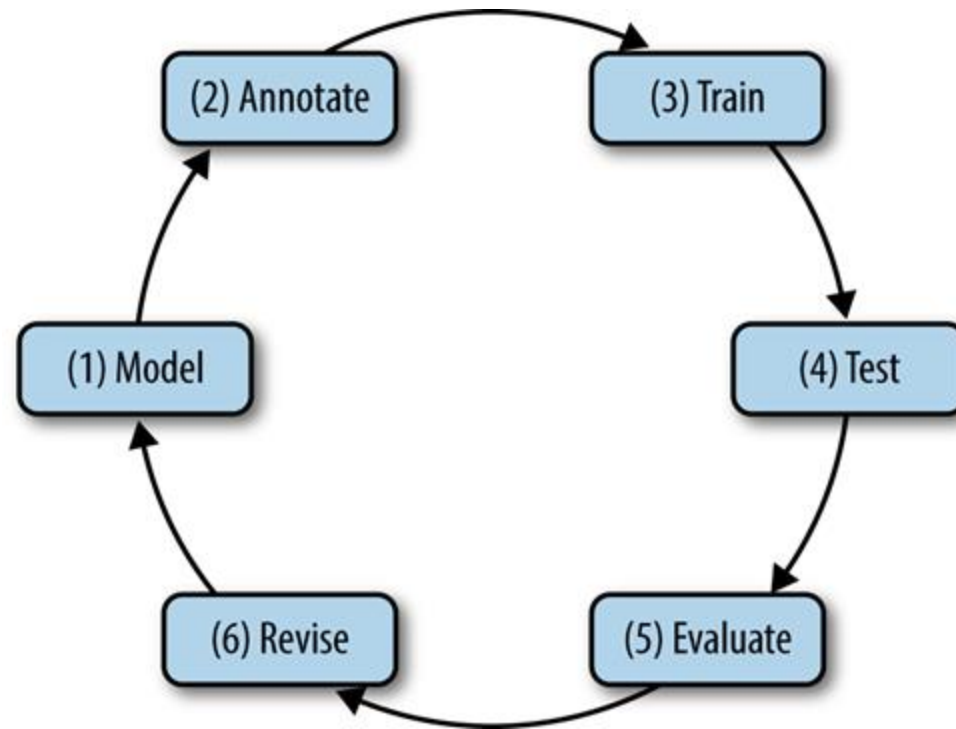
[My brother]_{theme} lives in [Milwaukee]_{location}.



Annotation process



Annotation Development Cycle



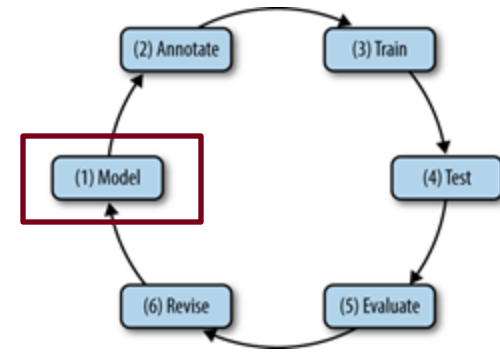
MATTER methodology ([Pustejovsky 2006](#))

Model the Phenomenon

(a.k.a Schematization or Specification)

A model, M , can be seen as a triple, $M = \langle T, R, I \rangle$.

- A **vocabulary of terms**, T ,
- The **relations** between these terms, R ,
- Their **interpretation**, I .



Terms

= {Document_type, Spam, Not-Spam}

Relations

= {Document_type ::= Spam | Not-Spam}

Interpretation

= { Spam = "something we don't want!",
Not-Spam = "something we do want!"}



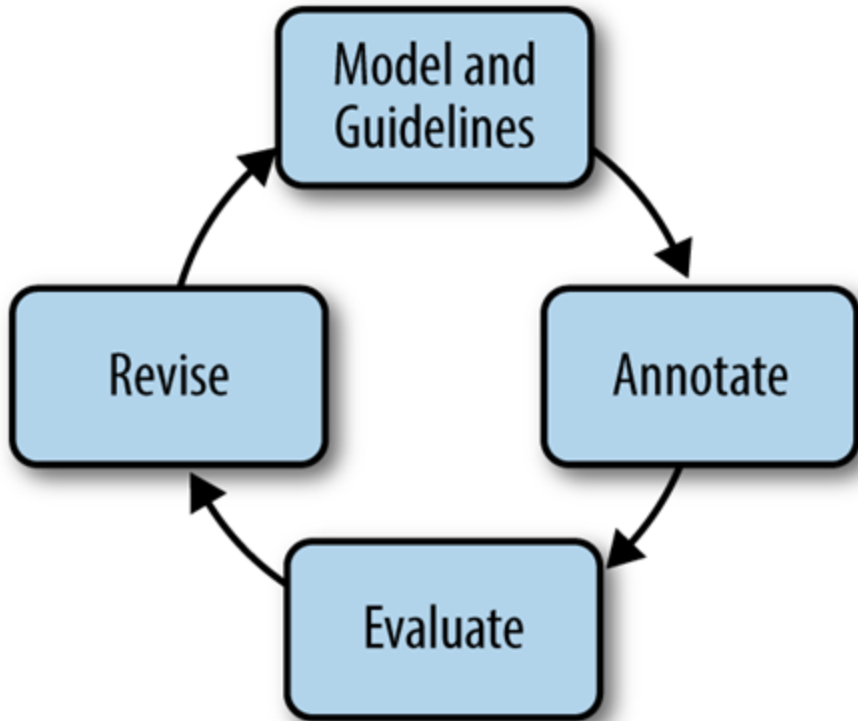
Terms = {Named_Entity, Organization, Person, Place, Time}

Relations = {Named_Entity ::= Organization | Person | Place | Time}

Interpretation = { Organization = "list of organizations in a database",
Person = "list of people in a database",
Place = "list of countries, geographic locations, etc.",
Time = "all possible dates on the calendar"}

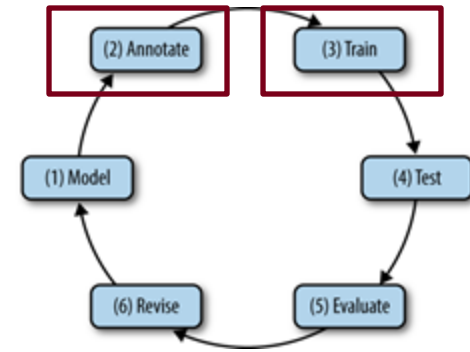


Annotate with the Specification



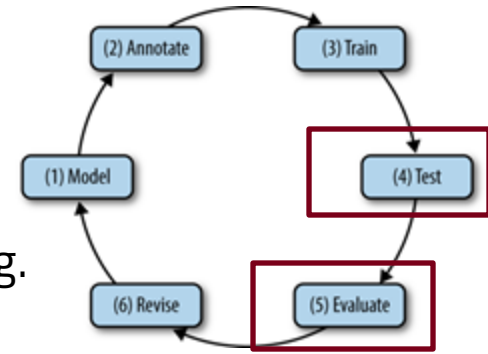
MAMA (Model-Annotate-Model-Annotate) cycle,
or the “babeling” phase of MATTER.

- Given the **specification document** encoding the model phenomenon, now you will need to **train human annotators** to mark up the dataset according to the tags that are important to you.
- This process is often repetitive by running multiple pilot studies to make sure annotators are well trained and specification is clear enough to follow.



Consistency

the most problematic when comparing annotations: namely, the extent or the span of the tag.



QBC Productions Inc. of East Anglia

Organization



[QBC Productions]_{Organization} Inc. of East Anglia



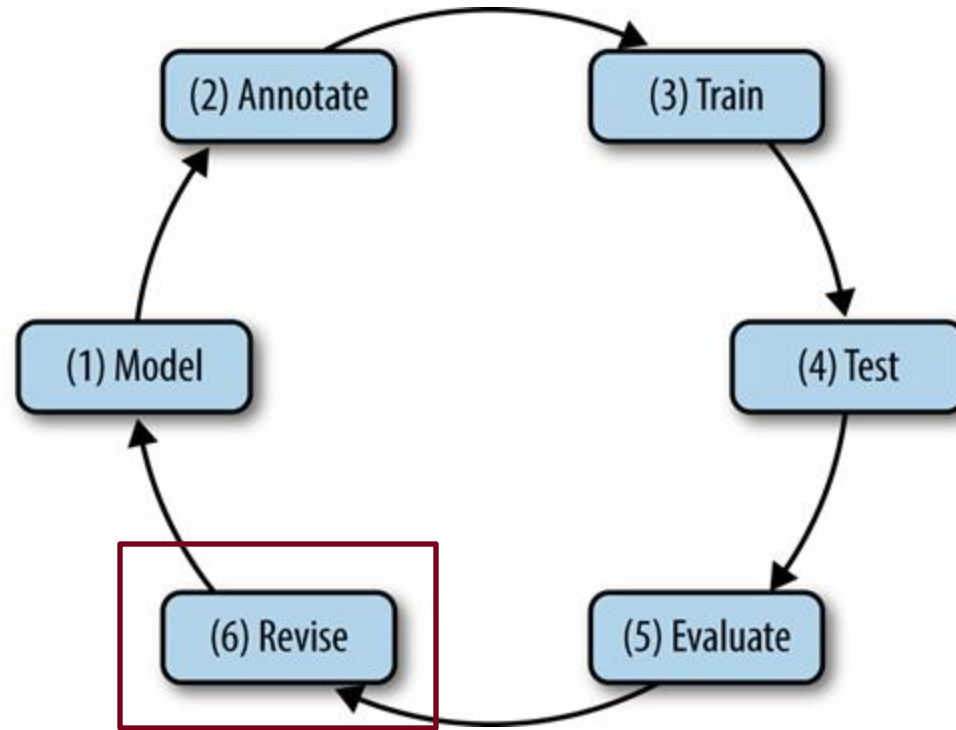
[QBC Productions Inc.]_{Organization} of East Anglia



[QBC Productions Inc. of East Anglia]_{Organization}



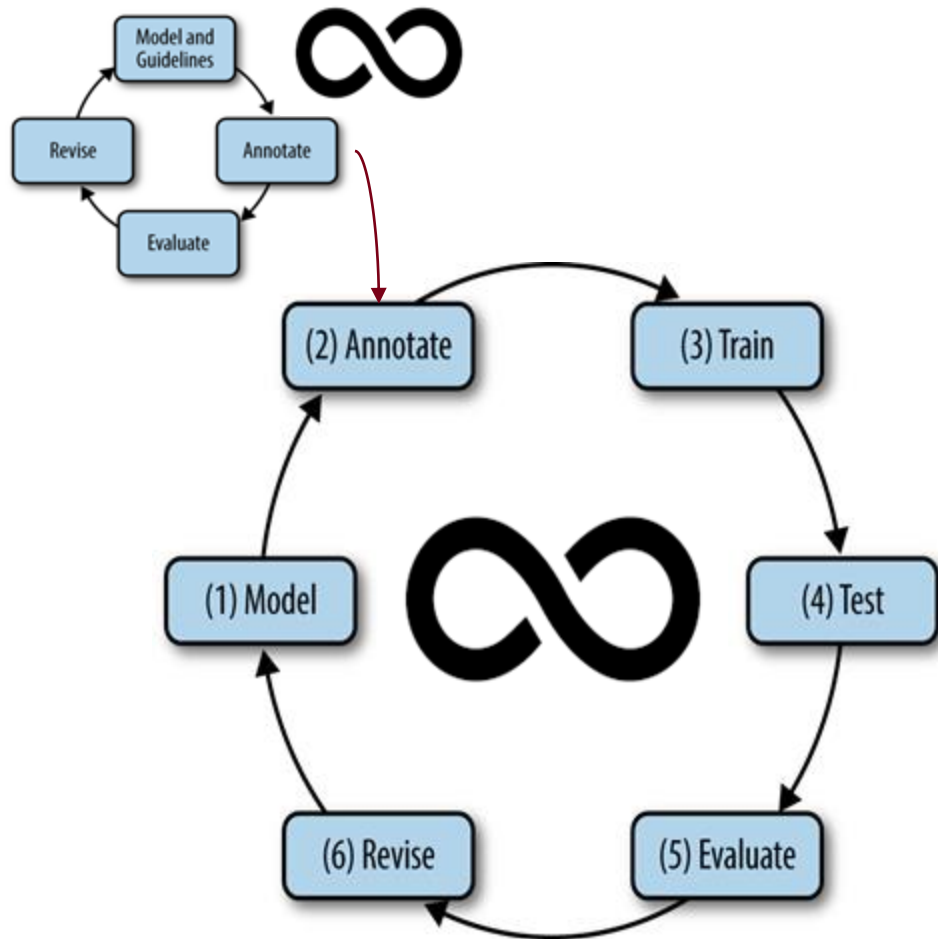
Annotation Development Cycle



The model and the annotation specification are revisited in order to make the annotation more **robust** and **reliable** with use in the algorithm using **inter-annotator agreement metrics**

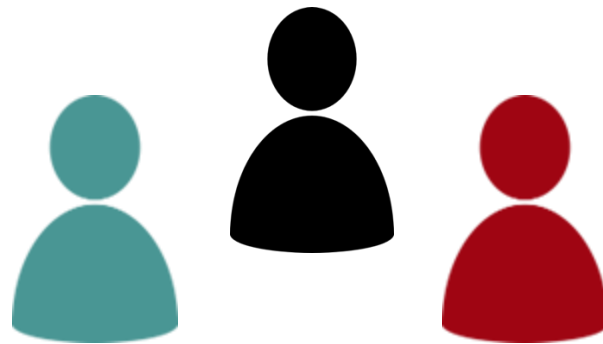
MATTER methodology (Pustejovsky 2006)

In Practice



- ❑ An **iterative** process until you reach to the target performance
- ❑ As model performance converges, you will face **edge cases** in the long tail. Analyzing the long-tail and updating the schema are painful and time-consuming, but most important in practice.
- ❑ There is **no single magic deep learning solution in real-world tasks**; If so, your task is relatively easy or narrowed down to a very specific scope

Recruiting annotators (coders)



Outsourcing

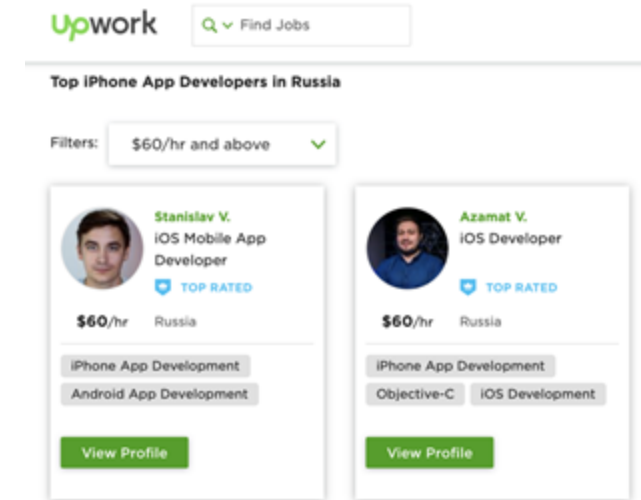
- ❑ Finding capable annotators can be a tremendous headache.
- ❑ From testing, onboarding, and ensuring tax compliance to distributing, managing, and assessing the quality of projects, there's an enormous amount of hidden labor involved in annotating.





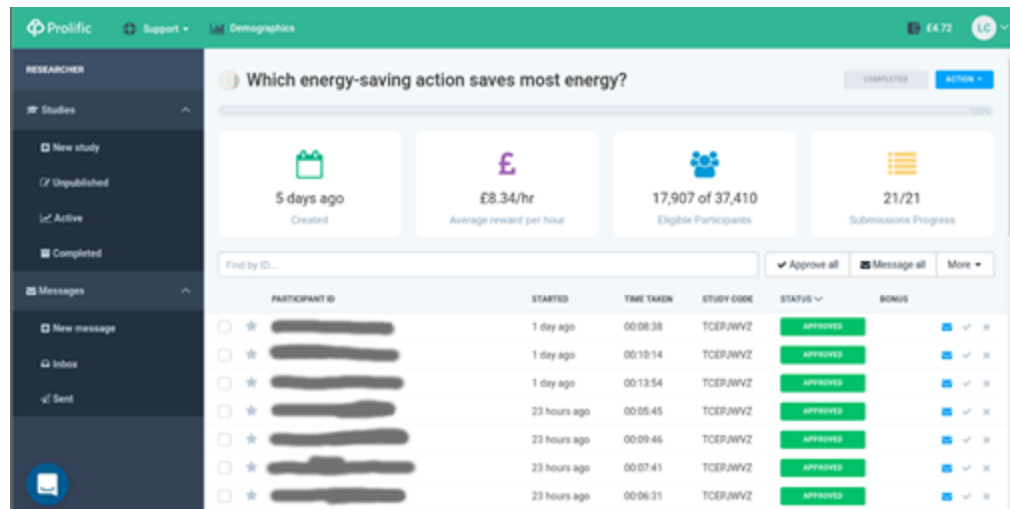
Amazon Mechanical Turk.

Best for finding people to help complete crowdsourced tasks



UpWork

Best for finding the right freelancers to complete tasks



Prolific

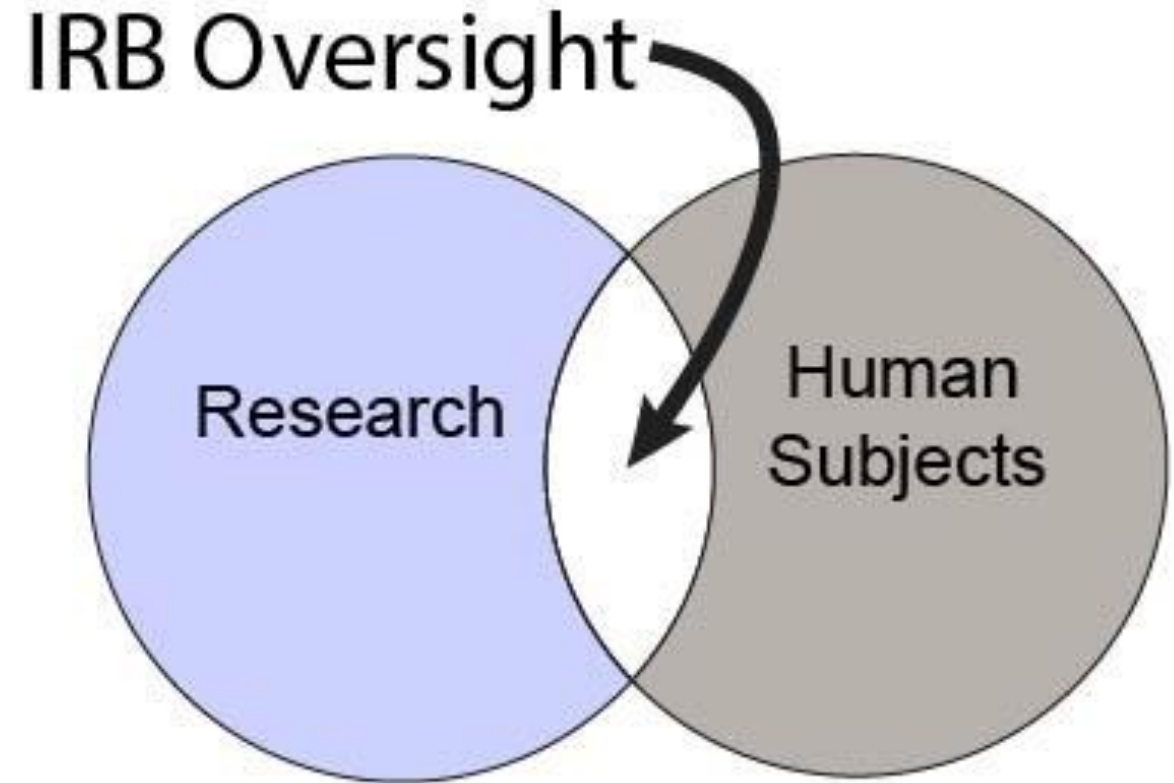
Quickly find research participants you can trust.



Undergraduate students

An institutional review board (IRB) is a type of committee that applies **research ethics** by reviewing the methods proposed for research to ensure that they are ethical.

Takes at least one month to get approval. Before approval, you can't collect any human-subject data in your project



THE EMPATHY TUTOR



The Empathy Tutor

A university research team built an AI “Empathy Tutor” designed to help medical students practice comforting patients. The data came from what researchers believed were fully anonymized therapy transcripts donated from an old hospital archive.

Months into the project, a student testing the AI noticed that the system sometimes used strangely specific phrases—**names of neighborhoods, rare medical conditions, and even a date that matched a local news story about a patient from years ago.** Investigators discovered that the “anonymized” transcripts were only lightly edited, and some patients could easily be re-identified.

The project was suspended immediately. The IRB later determined that the researchers technically had *permission* to use the archived data, but the **risk of re-identification made the study unethical**, despite its benign goal.

Annotation quality assessment



Correctness of annotations

Label	Meaning	Evidence from table
I	Introduction / Problem framing	"We address the problem..."
P	Purpose / Aim	"Our aim is to..."
M	Method / Technical description	"[A] is set up as prior information..."
R	Result / Conclusion	"It is shown that..."

Sentence	Coder 1	Coder 2	Agreement
We address the problem of recognition	I	P	✗
Our aim is to ...recognize [x] from [y].	P	P	✓
[A] is set up as prior information, and its pose is determined by three parameters, which are [j,k and l].	M	M	✓
An efficient local gradient-based method is proposed to ..., which is combined into ... framework to estimate [V and W] by iterative evolution	P	R	✗
It is shown that the local gradient-based method can evaluate accurately and efficiently [V and W] .	R	R	✓

Observed agreement between coder 1 and 2: 60%



Inter-annotator agreement (IAA)

the probability that the raters could have agreed purely by chance.

❑ **Relative agreement** is **60%** in the previous example, but **chance agreement is 20%**. Agreement measures need to be corrected for chance agreement (Carletta, 1996)

❑ Kappa coefficient (Cohen 1960)

- 1 (agreement), 0 (no correlation), -1 (disagreement)

Corrected measure:

$$K = \frac{P(A) - P(E)}{1 - P(E)} = \frac{0.6 - 0.2}{1 - 0.2} = 0.5$$



Step 1: Calculate **relative agreement** (p_o) between raters.

		Rater 2	
		Yes	No
Rater 1	Yes	25	10
	No	15	20

$$\begin{aligned} p_o &= (\text{Both said Yes} + \text{Both said No}) / (\text{Total Ratings}) \\ &= (25 + 20) / (70) = 0.6429 \end{aligned}$$



the probability that the raters could
have agreed purely by chance.

Step 2: Calculate the hypothetical probability of **chance agreement** (p_e) between raters.

		Rater 2	
		Yes	No
Rater 1	Yes	25	10
	No	15	20

$$P(\text{"Yes"}) = ((25+10)/70) * ((25+15)/70) = 0.285714$$

$$P(\text{"No"}) = ((15+20)/70) * ((10+20)/70) = 0.214285$$

$$p_e = 0.285714 + 0.214285 = 0.5$$



the probability that the raters could
have agreed purely by chance.

Step 2: Calculate the hypothetical probability of chance agreement (p_e) between raters.

		Rater 2	
Rater 1		Yes	No
	Yes	25	10
	No	15	20

$$P(\text{"Yes"}) = ((25+10)/70) * ((25+15)/70) = 0.285714$$

$$P(\text{"No"}) = ((15+20)/70) * ((10+20)/70) = 0.214285$$

$$p_e = 0.285714 + 0.214285 = 0.5$$



Step 3: Calculate Cohen's Kappa

		Rater 2	
Rater 1		Yes	No
	Yes	25	10
	No	15	20

$$K = \frac{P(A) - P(E)}{1 - P(E)}$$

$$\begin{aligned} k &= (p_o - p_e) / (1 - p_e) \\ &= (0.6429 - 0.5) / (1 - 0.5) \\ &= 0.2857 \end{aligned}$$



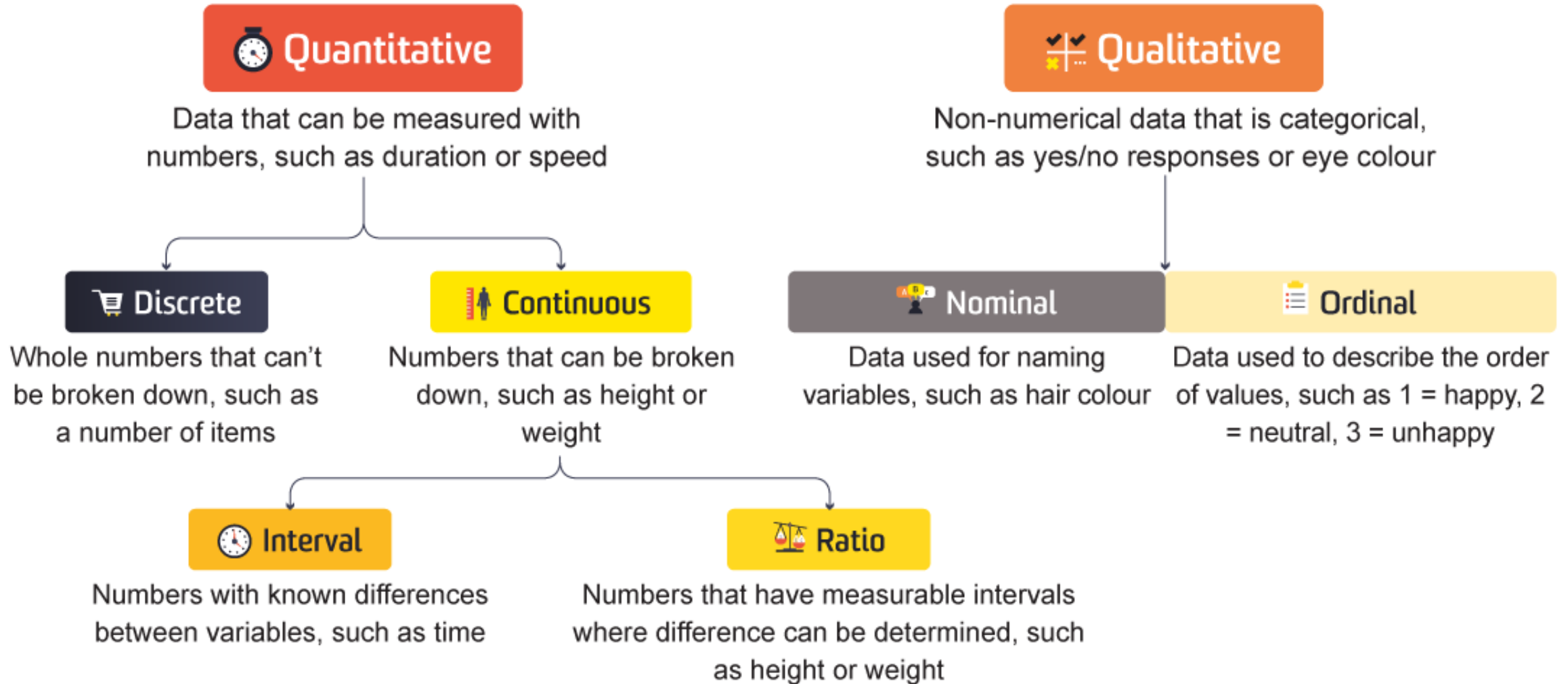
Interpretation of Cohen's Kappa



Value Range	Cohen's Interpretation
Below 0.20	None to slight agreement
.21–.39	Fair agreement
.40–.59	Moderate agreement
.60–.79	Substantial agreement
.80–.90	Almost perfect agreement
Above .90	Almost perfect agreement



Types of Data



Other IAA measures by types and their interpretation

Comparison of IRR indices in presence of research limitations					
IRR	Data	Missing Data	Number of Raters	The effect of 'chance' in agreement is minimized?	General agreement on the significance of a numeric result?
Cohen's Kappa	Nominal	No	2	No *	No
Fleiss's Kappa	Nominal	No	$2 \geq$	No *	No
Krippendorff's Alpha	All Data	Yes	$2 \geq$	Yes	Yes **

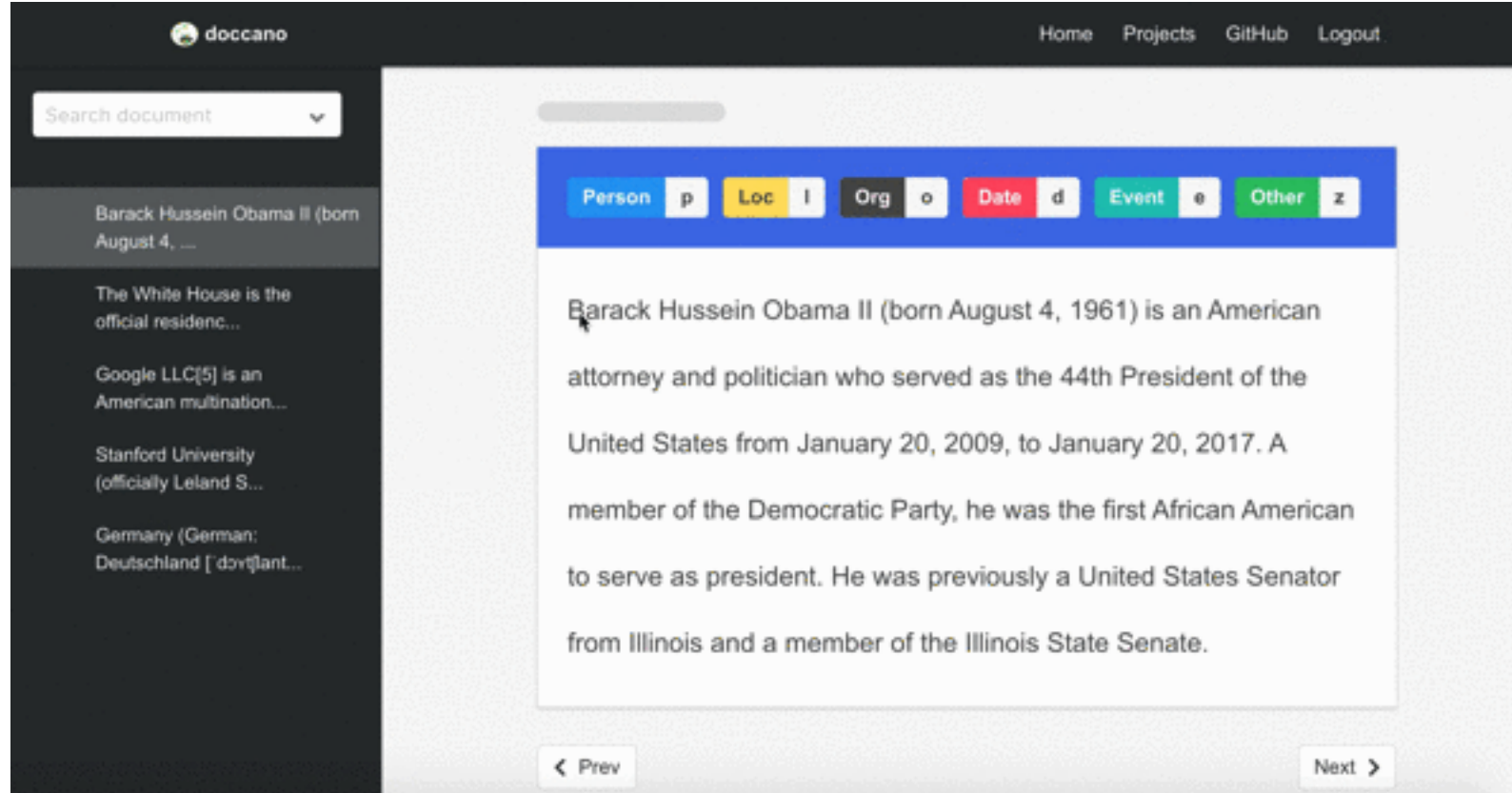
** Krippendorff's Alpha considers 0.823 as the cut point.

- **Landis and Koch (1977)** 0.6-0.79 substantial; 0.8+ perfect
- **Krippendorff (1980)** 0.67-0.79 tentative; 0.8+ good
- **Green (1997)** 0.4-0.74 fair/good; 0.75 high



Annotation tools





Pros:
Easy to use
Support Teams
Open Source

Cons:
Fully manual annotation

Brat

<https://brat.nlplab.org/>

The screenshot shows the Brat web interface at the URL `/tutorials/.328457584/news/000-introduction`. The main content area displays a tutorial document with the following text:

1 Welcome to the Brat Rapid Annotation Tool (brat) tutorial!

3 brat is a web-based tool for structured text annotation and visualization. The easiest way to explain what this means is by example: see the following sentence illustrating various types of annotation. Take a moment to study this example, moving your mouse cursor over some of the annotations. Hold the cursor still over an annotation for more detail.

6 1) Citibank was involved in moving about \$100 million for Raul Salinas de Gortari, brother of a former Mexican president, to banks in Switzerland.

9 If this example seems complicated, don't panic! This tutorial will present the key features of brat interactively, with each document presenting one or a few features. If you follow this brief tutorial, you'll be able to understand and create annotations such as those above in no time.

11 Try moving to the next document now by clicking on the arrow to the right on the blue bar at the top left corner of the page.

The sentence is annotated with several entities and relations:

- Entities: `Citibank` (blue), `$100 million` (green), `Raul Salinas de Gortari` (orange), `Switzerland` (yellow).
- Relations: `Give` (arrow from `Citibank` to `$100 million`), `Transfer money` (arrow from `$100 million` to `Raul Salinas de Gortari`), `Beneficiary` (arrow from `$100 million` to `Raul Salinas de Gortari`), `Recipient` (arrow from `$100 million` to `Switzerland`), `Origin` (arrow from `Switzerland` to `$100 million`).

Pros:
Open source
Free

Cons:
Old-fashioned UI



Prodigy

<https://prodi.gy/>

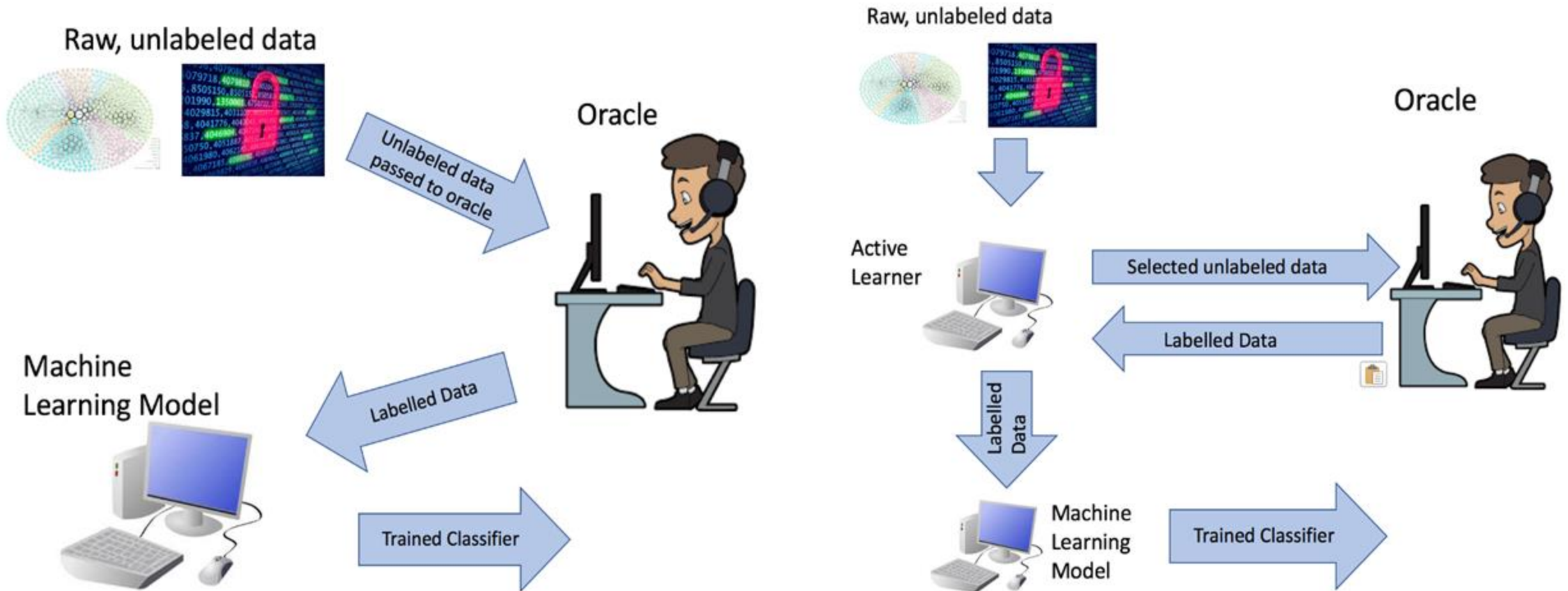
Radically efficient machine teaching. An annotation tool powered by active learning.



Pros:
Automation
Lots of features
Can train the models

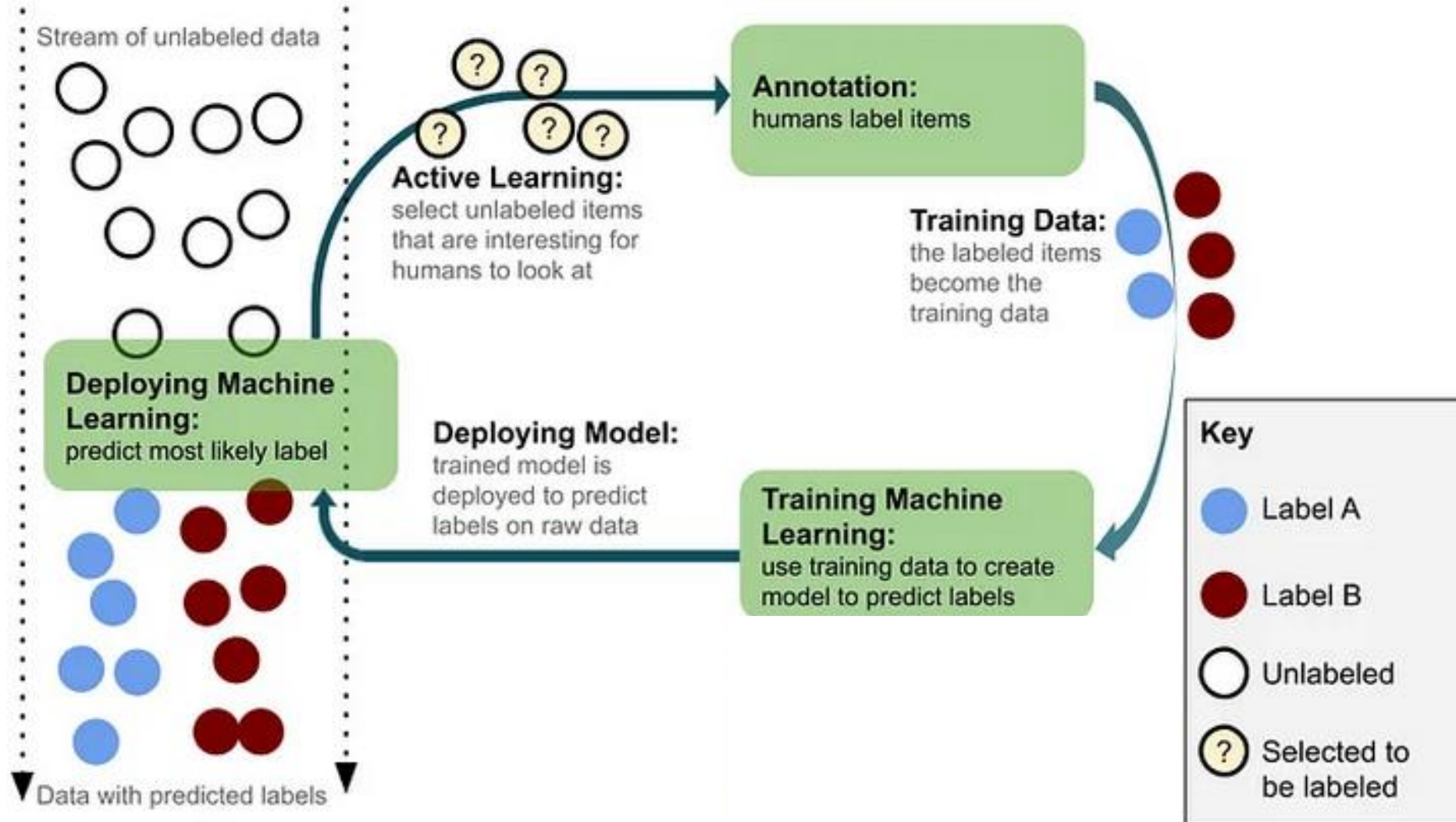
Cons:
Learning Curve
Not Open Source.

Passive vs. Active Learning



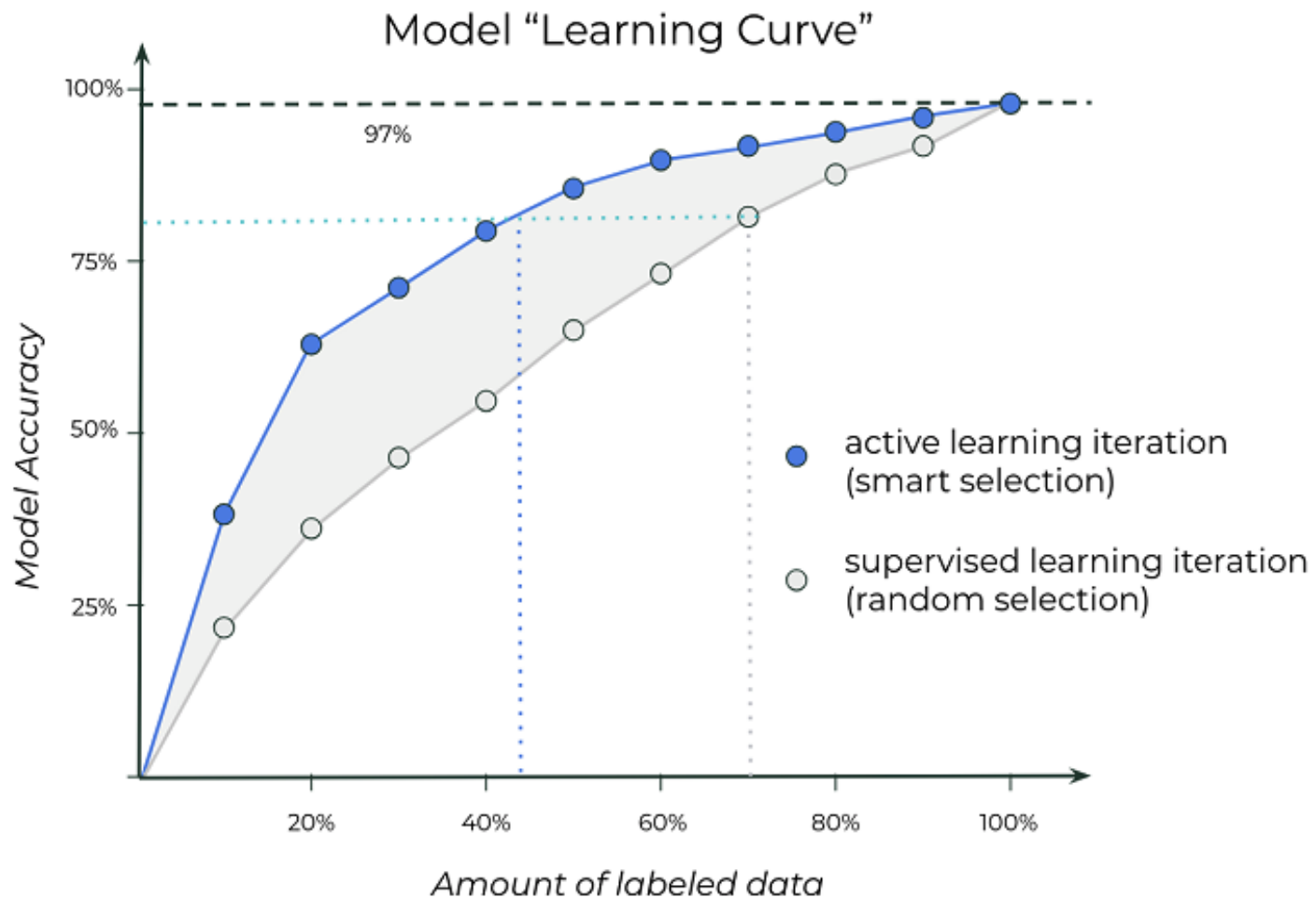
<https://towardsdatascience.com/introduction-to-active-learning-117e0740d7cc>

Active learning



Active learning

Using active learning gets to higher model accuracies with less labelled data



prodigy ✓

PROJECT INFO

DATASET ner_finance_news
LANGUAGE en
VIEW ID ner_manual

PROGRESS

THIS SESSION 0
TOTAL 0

00

ACCEPT 0
REJECT 0
IGNORE 0

HISTORY

© 2019 Explosion AI (Prodigy v1.8.0)

PERSON : ORG : MONEY : TICKER :

Cerner Corp. ORG President Zane Burke PERSON sold \$ 3.5 million MONEY in stock Wednesday , according to a filing with the Securities and Exchange Commission ORG . Burke ORG sold 50,000 shares valued at \$ 70 MONEY , leaving him with 26,799 shares owned directly . Cerner ORG 's stock (Nasdaq : CERN) was ...

SOURCE: TechCrunch - VIA: News API

Human annotators correct the
model-predicted pseudo labels



Issues in annotation





Task 1: Classify between Order or Complaint?

Task 2: Annotate semantic types

I ordered a large cheese pizza and a coke to Somewhere Blvd an hour ago!
It still isn't here!!!! What gives ?! Can you call me with an update ? 555-555-5556

I a QUANTITY DRINK to ADDRESS Somewhere Blvd

Disagreement

Semantic interpretation



Jane reads this and thinks it's **not an order** because the customer says the order has already been placed.



I ordered a large cheese pizza and a coke to
Somewhere Blvd an hour ago! It still isn't here!!!!
What gives?! Can you call me with an update?
555-555-5556

Bob classifies this as an **order** because it has all of the information an order would have.



Disagreement

Syntactic errors

A large cheese pizza is a pizza after all, so why not label the whole phrase as pizza?



Classifications

Complaint

Apply Classifications

Hey,

I ordered a QUANTITY SIZE TOPPING PIZZA and a QUANTITY

ago and it still isn't here.

What gives ? Can you call me with an update at PHONE NUMBER 555-555-5556

Tnx



Classifications

Order

Apply Classifications

Hey,

I ordered a PIZZA and a QUANTITY DRINK to ADDRESS Somewhere Blvd an hour ago and it still isn't here.

What gives ? Can you call me with an update at PHONE NUMBER 555-555-5556

Tnx



Disagreement

Intents

Conflict between document intent and entity tags

- This is "Complaint" intent
- So, didn't annotate any entities because this is not an order



Classifications

Complaint 

 | 

Apply Classifications

Hey,

I ordered a large cheese pizza and a coke to Somewhere Blvd an hour ago and it still isn't here.

What gives ? Can you call me with an update at 555-555-5556

Tnx



Disagreement for subjective datasets

Datasets	Text	Annotation Distribution	Disagreement Label
Dilemmas	1st action: “refusing to do a survey on the credit card reader while paying with cash at the Office Max.” 2nd action: “saying my bf has no right to dictate who I tell about my abortion.”	1 annotator votes for the <u>first action</u> is less ethical while 4 others vote the <u>second action</u> is less ethical → Aggregated Label: 2nd action is less ethical	Binary: 1 Continuous: 1/5
Dynasent	“Had to remind him to toast the sandwich.”	4 annotators believe it’s <u>negative</u> while one think it is <u>neutral</u> → Aggregated Label: negative	Binary: 1 Continuous: 1/5
Politeness	“Where did you learn English? How come you’re taking on a third language?”	5 annotators politeness scores are <u>5, 13, 9, 11, 11</u> with the <u>maximum</u> of 25. → Aggregated Label: impolite	Binary: 0 Continuous: 0

Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information, AACL 2023



Disagreement for subjective datasets

Datasets	Text	Annotation Distribution	Disagreement Label
SBIC	“Abortion destruction of the nuclear family contraceptives feminism convincing women to wait for children damaging economy so youth cannot leave the nest ramping up tensions between sexes all serves one primary goal to lower the population.”	A1 (age: 32, politics: liberal, race: white, gender: woman) votes for <u>inoffensive</u>	Binary: 1 Continuous: 1/3
		A2 (age: 34, politics: liberal, race: white, gender: woman) votes for <u>inoffensive</u>	
		A3 (age: 29, politics: mod-liberal, race: hispanic, gender: woman) votes for <u>offensive</u> → Aggregated Label: inoffensive	
SChem101	“It’s okay to have abortion.”	A1 (age: 30-39, education: high school, race: white, gender: woman) votes for <u>people ocassional think this</u>	Binary: 1 Continuous: 2/5
		A2 (age: 40-49, education: grad, race: white, gender: man) votes for <u>controversial</u>	
		A3 (age: 30-39, education: bachelor, race: white, gender: man) votes for <u>common belief</u>	
		A4 (age: 21-29, education: high school, race: white, gender: woman) votes for <u>controversial</u>	
		A5 (age: 30-39 , education: bachelor, race: hispanic, gender: woman) votes for <u>controversial</u> → Aggregated Label: controversial	



Annotation artifacts



They used Amazon Mechanical Turk for data collection. Sentences in SNLI are derived from only image captions.

We will show you the caption for a photo. We will not show you the photo. Using only the caption and what you know about the world:

- Write one alternate caption that is **definitely** a **true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "There are animals outdoors."*
- Write one alternate caption that **might be** a **true** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "Some puppies are running to catch a stick."*
- Write one alternate caption that is **definitely** a **false** description of the photo. *Example: For the caption "Two dogs are running through a field." you could write "The pets are sitting on a couch." This is different from the maybe correct category because it's impossible for the dogs to be both running and sitting.*

Figure 1: The instructions used on Mechanical Turk for data collection.

Annotation artifacts

- They observe that hypotheses generated by this crowdsourcing process contain **artifacts** that can help a classifier detect the correct class **without ever observing the premise**.
- Crowd workers **adopt heuristics** in order to generate hypothesis **quickly** and efficiently.

Premise	A woman selling bamboo sticks talking to two men on a loading dock.
Entailment	There are at least three people on a loading dock.
Neutral	A woman is selling bamboo sticks to help provide for her family .
Contradiction	A woman is not taking money for any of her sticks.

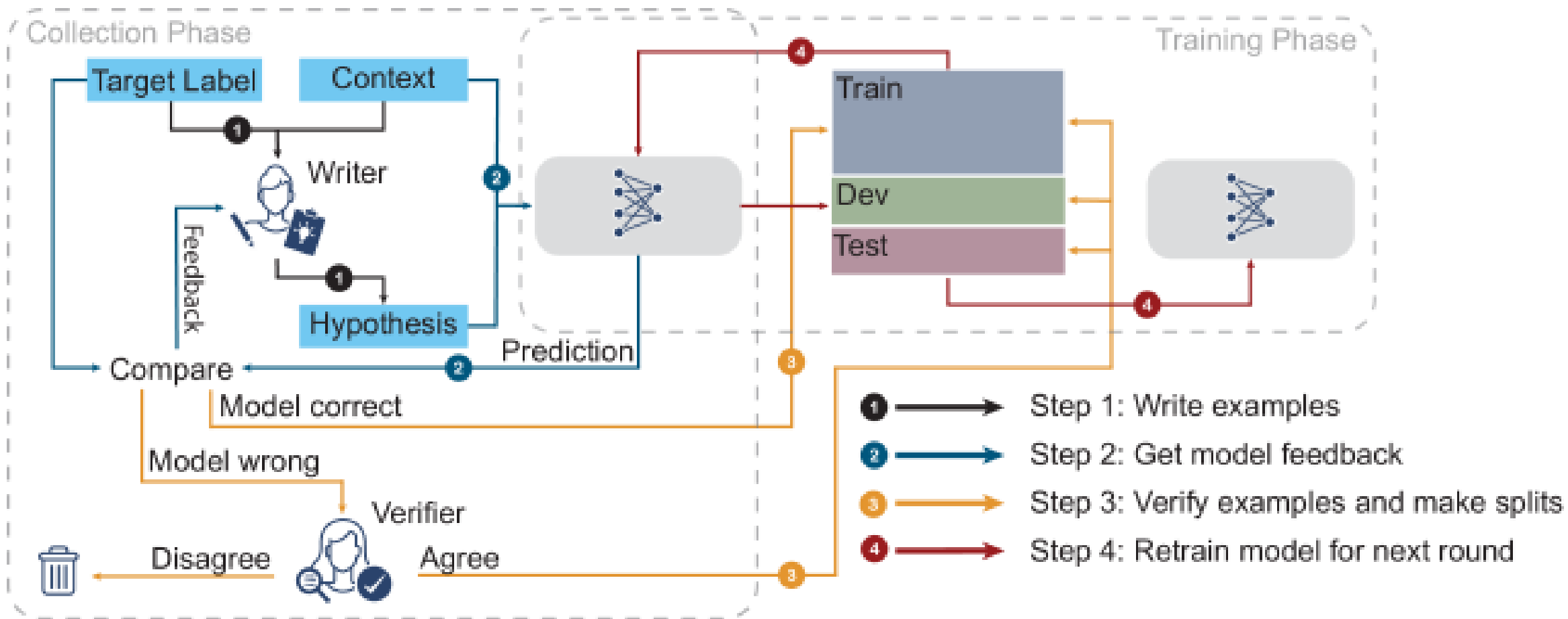
Table 1: An instance from SNLI that illustrates the artifacts that arise from the annotation protocol. A common strategy for generating entailed hypotheses is to remove gender or number information. Neutral hypotheses are often constructed by adding a purpose clause. Negations are often introduced to generate contradictions.

Annotation Artifacts (Gururangan et al., 2018)

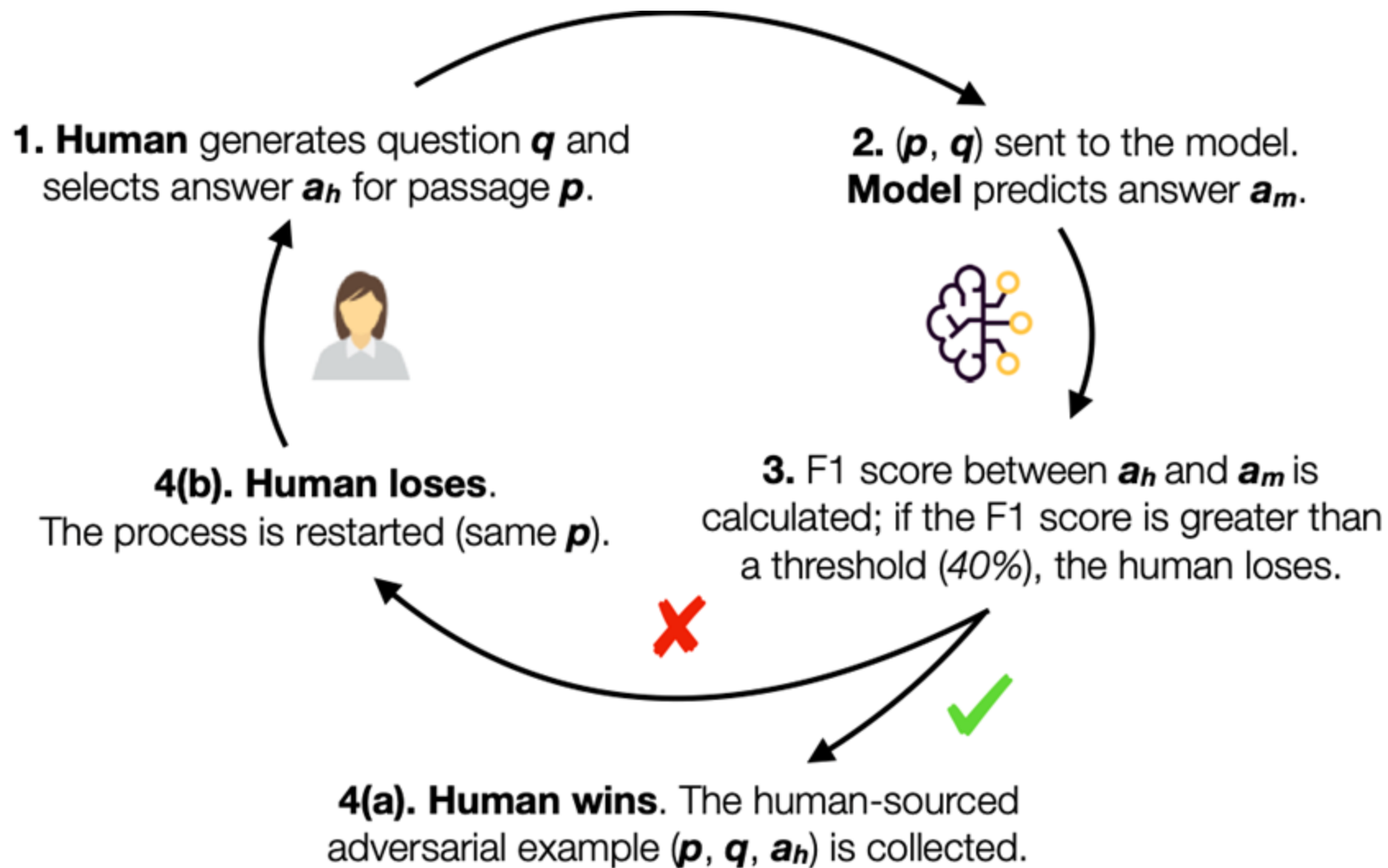


Advanced annotation techniques

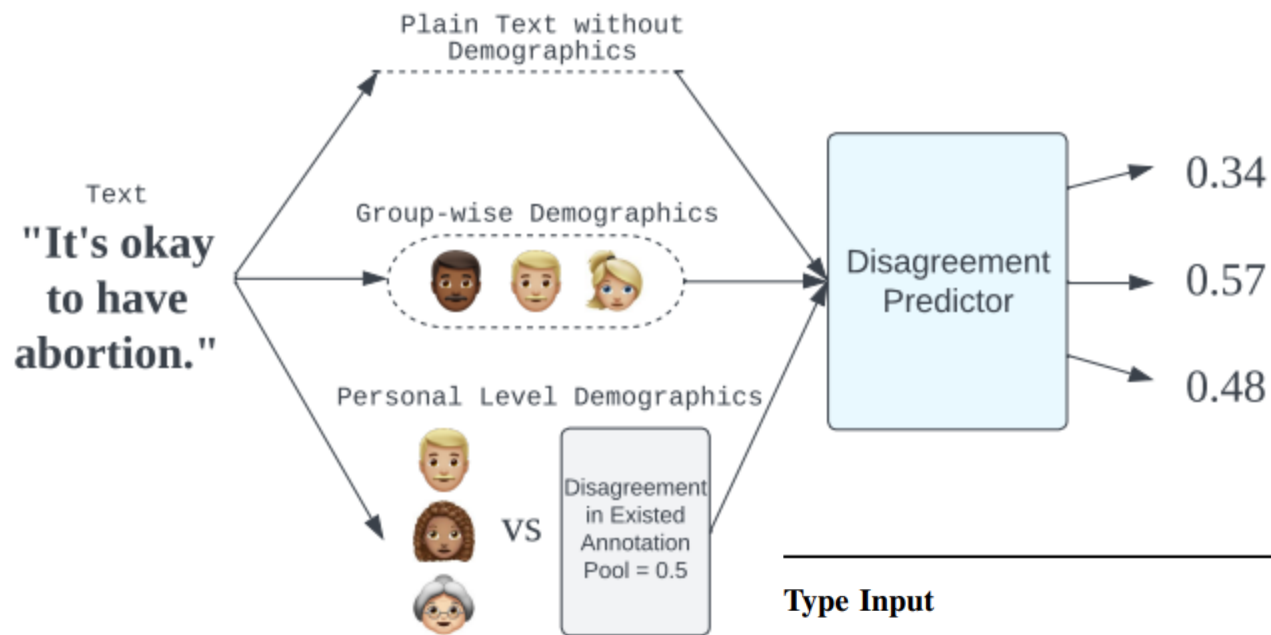




Adversarial NLI: A New Benchmark for Natural Language Understanding



Bartolo et al. in Beat the AI: Investigating Adversarial Human Annotation for Reading Comprehension

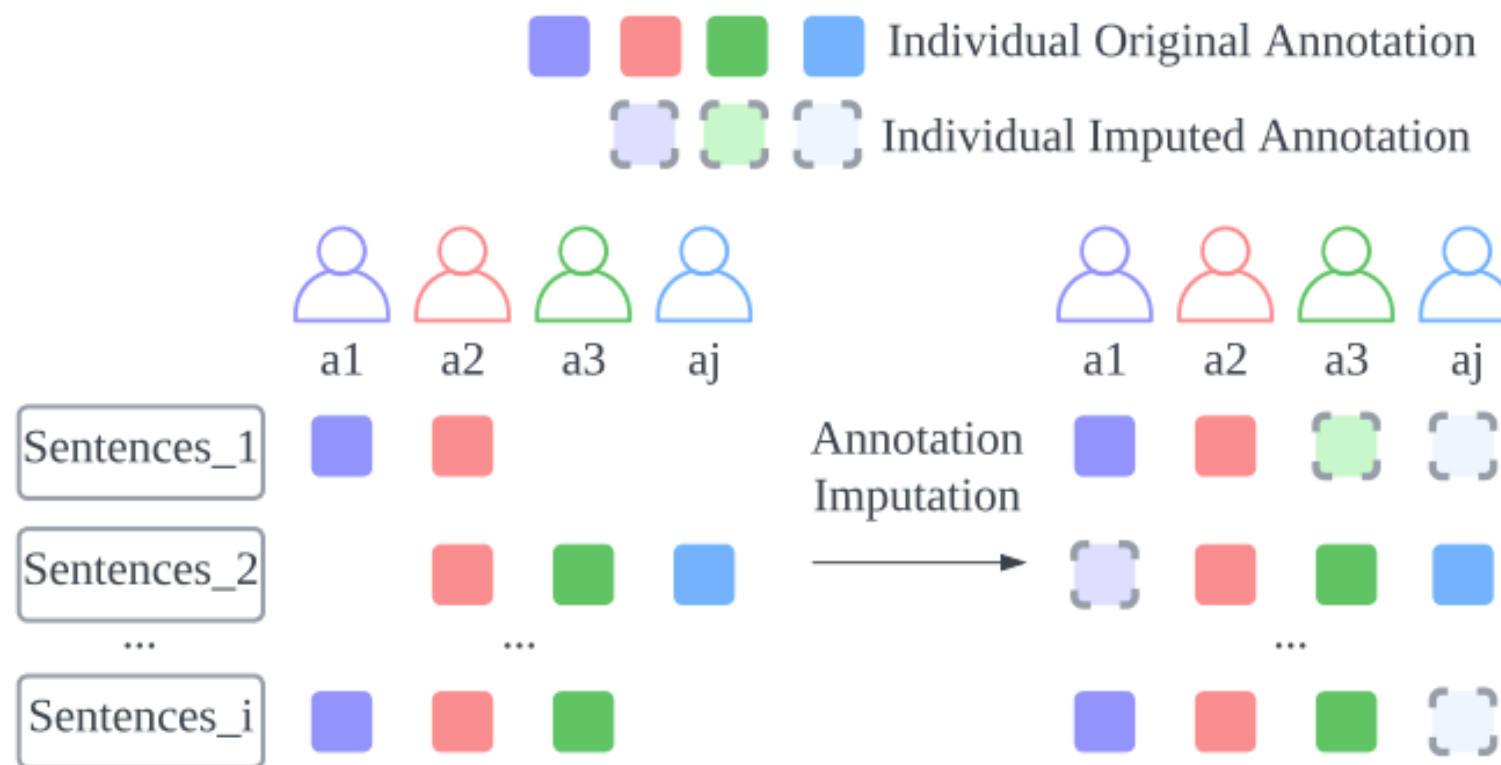


Type Input	Predicted Disagreement	Predicted Disagreement with Demographics	Collected Disagreement
(1) age: 22, politics: liberal, race: white, gender: woman . [SEP] what is the most common name for a black dude king kong.	0.003	0.015	0
(2) age: 25, politics: liberal, race: asian, gender: woman. [SEP] swear to god they are the same people that objectify women and reify rape culturemisogyny . More importantly y do they feel so threatened.	1.034	0.971	1
(3) age: 18 , politics: other, race: asian, gender: man . [SEP] two strong men needed to lift some furniture .	0.223	0.096	0
(4) age: 29, politics: liberal, race: white, gender: woman. [SEP] 13 reasons why is making me wish i went to a real fucking highschool .	0.332	0.01	0.5

<https://github.com/minnesotanlp/Quantifying-Annotation-Disagreement>

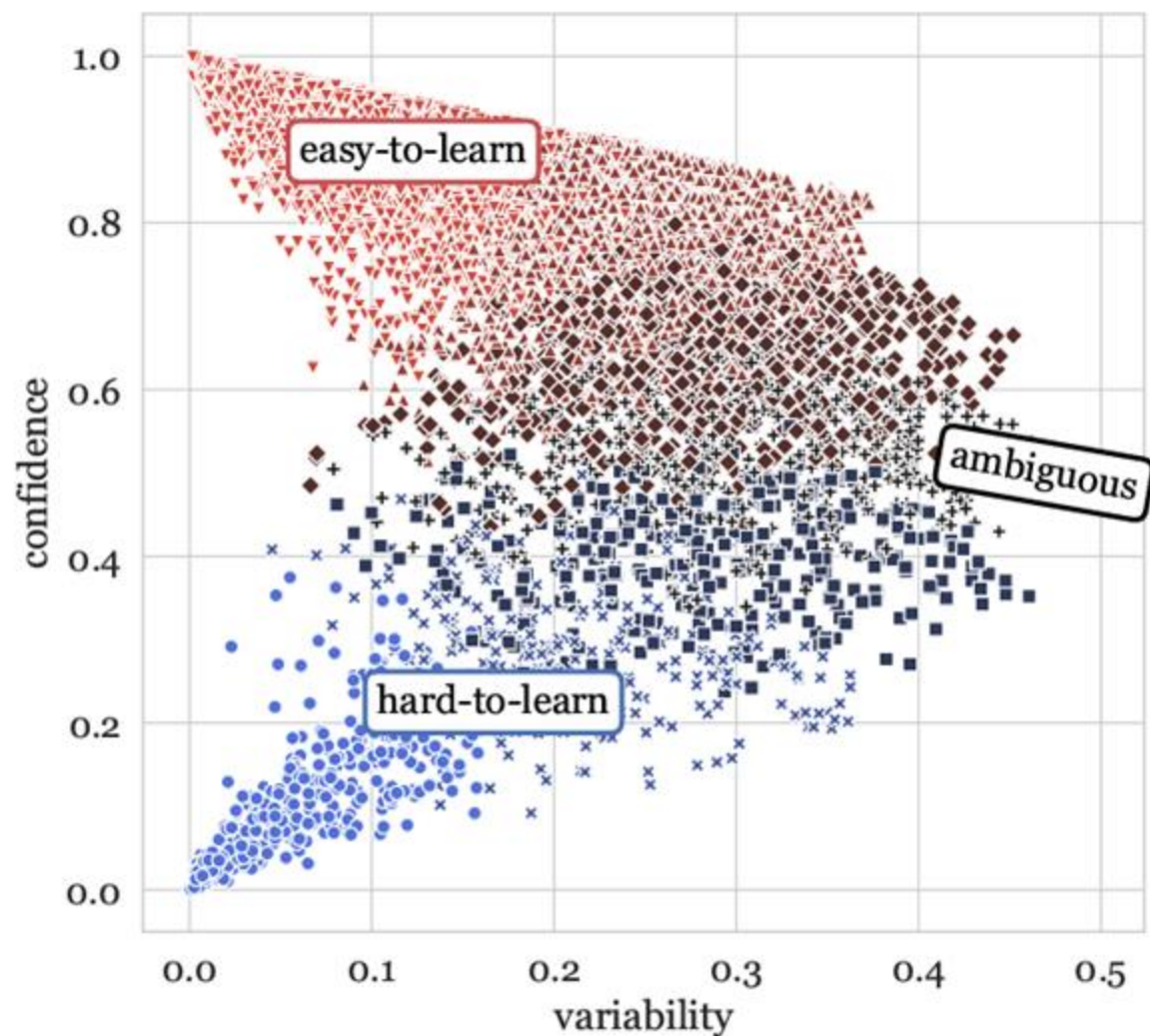
Everyone's Voice Matters: Quantifying Annotation Disagreement Using Demographic Information, AACL 2023

Annotation Imputation

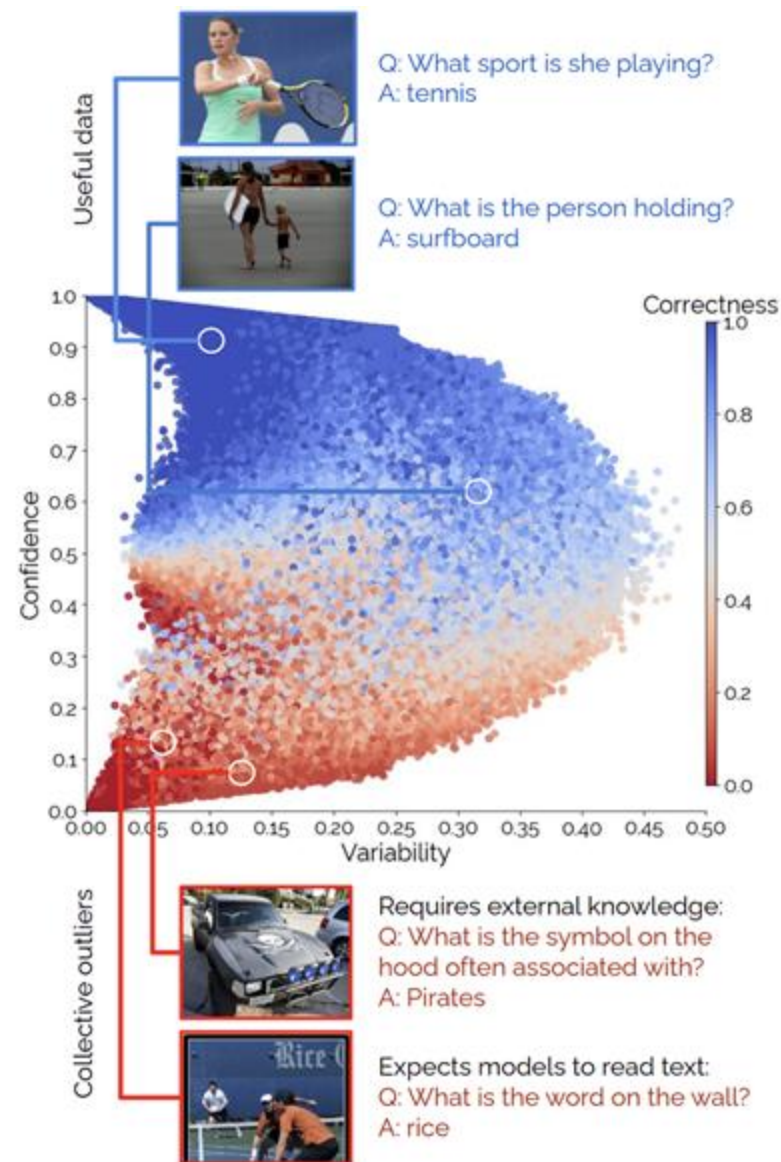


https://www.youtube.com/watch?v=xO1ksJ9AW-w&ab_channel=LondonLowmanstoneIV

Annotation Imputation to Individualize Predictions: Initial Studies on Distribution Dynamics and Model Predictions, NLPerspectives @ECAI 2023



Dataset Cartography: Mapping and Diagnosing Datasets with Training Dynamics, Swayamdipta et al., 2020



Mind Your Outliers! Investigating the Negative Impact of Outliers on Active Learning for Visual Question Answering, Karamcheti et al, 2021

Collaborative Annotation

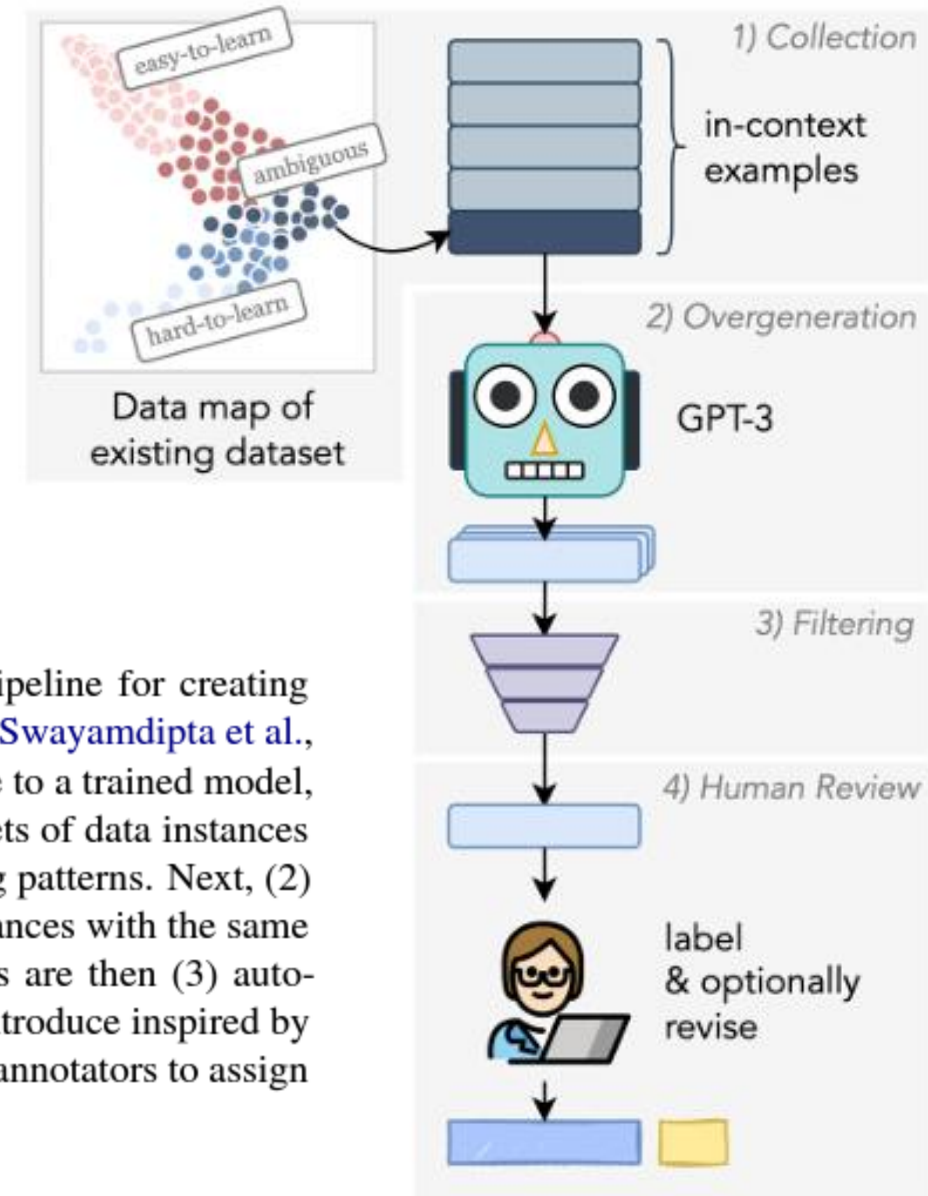
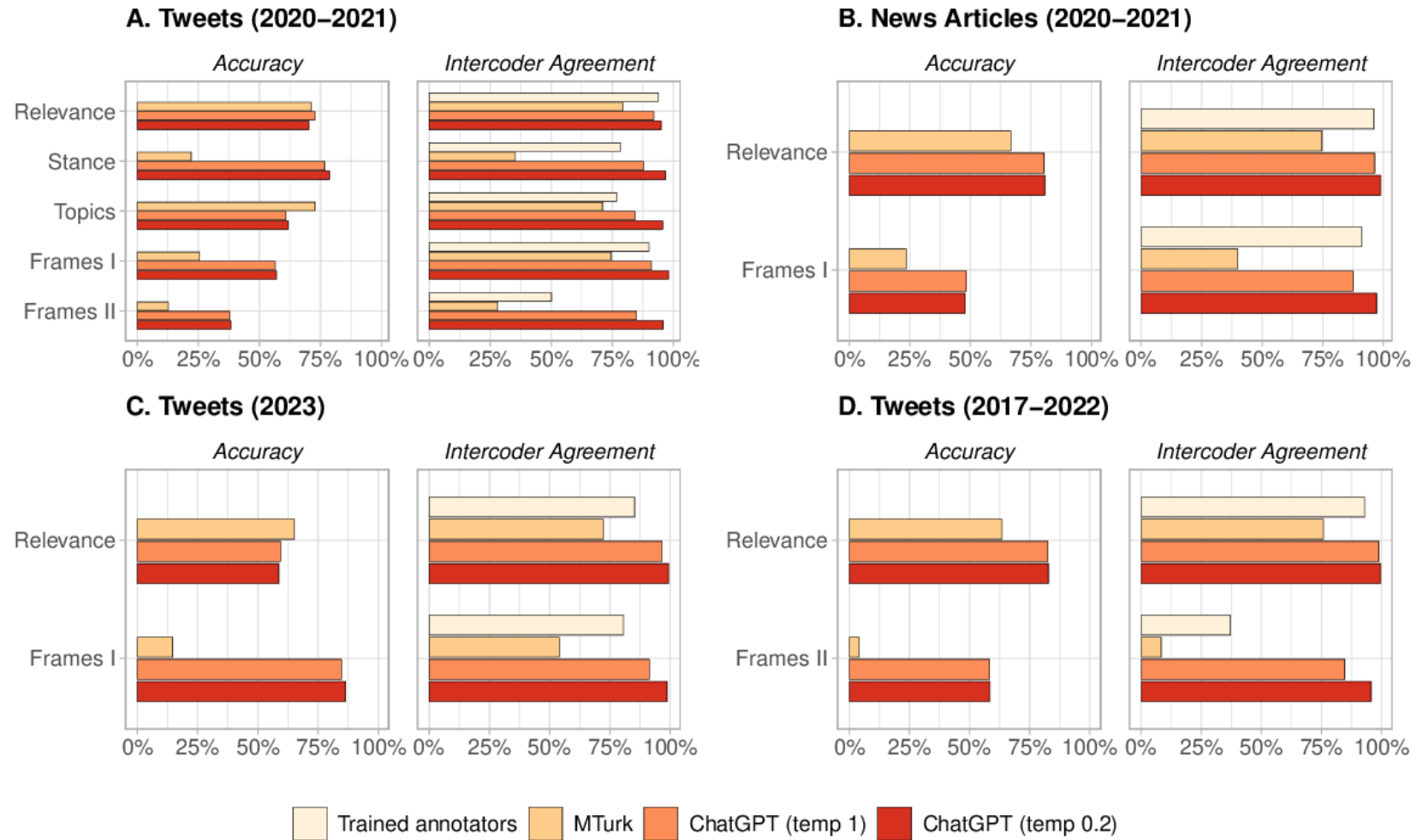


Figure 1: An illustration of our pipeline for creating WANLI. Starting with a data map (Swayamdipta et al., 2020) of an existing dataset relative to a trained model, (1) we automatically identify pockets of data instances exemplifying challenging reasoning patterns. Next, (2) we use GPT-3 to generate new instances with the same pattern. These generated examples are then (3) automatically filtered via a metric we introduce inspired by data maps, and (4) given to human annotators to assign a gold label and optionally revise.

LLMs as Annotators and Synthetic Data



ChatGPT as Annotators



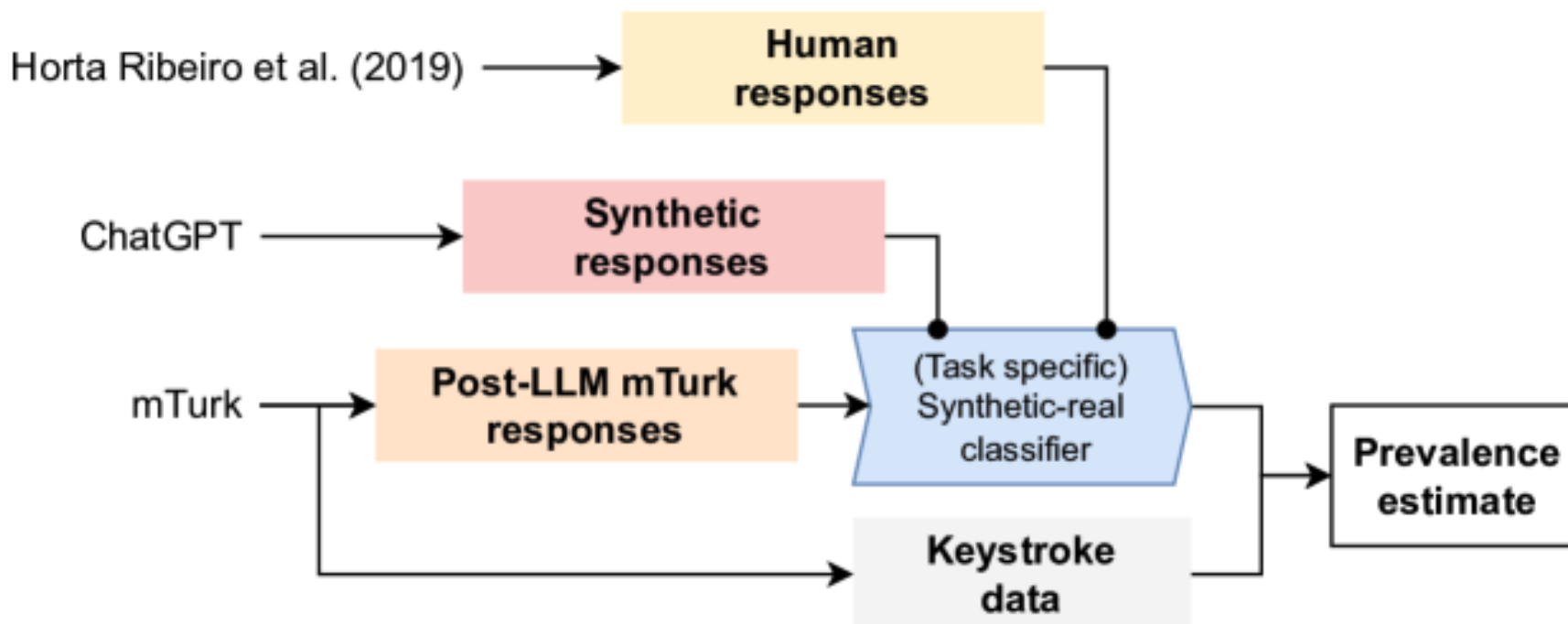
ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks <https://arxiv.org/abs/2303.15056>



LLMs as Annotators

Normally, a human makes a request to a computer, and the computer does the computation of the task. But **artificial artificial intelligences** like Mechanical Turk invert all that.

Jeff Bezos

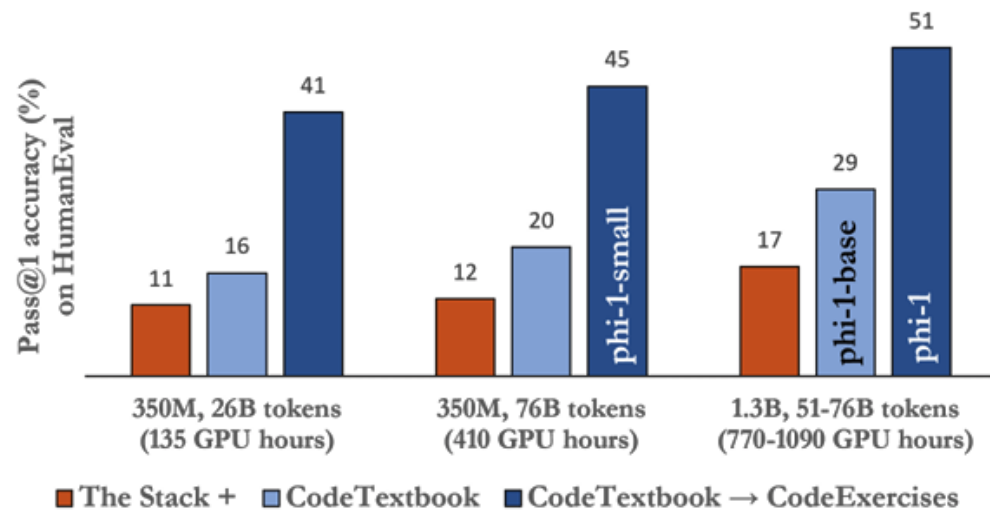


Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks <https://ar5iv.labs.arxiv.org/html/2306.07899>



High quality data is all you need

- ❑ Chinchilla shows that 70B model could beat 350B models, if it was trained on more tokens (1.4 Trillion tokens)
- ❑ Data quality could break the scaling laws.
- ❑ Synthetic data (code exercises) filtered with a GPT4-generated quality rating (**educational value**)



Educational values deemed by the filter

High educational value

```
import torch
import torch.nn.functional as F

def normalize(x, axis=-1):
    """Performs L2-Norm."""
    num = x
    denom = torch.norm(x, 2, axis, keepdim=True)
    .expand_as(x) + 1e-12
    return num / denom

def euclidean_dist(x, y):
    """Computes Euclidean distance."""
    m, n = x.size(0), y.size(0)
    xx = torch.pow(x, 2).sum(1, keepdim=True).
    expand(m, n)
    yy = torch.pow(y, 2).sum(1, keepdim=True).
    expand(m, m).t()
    dist = xx + yy - 2 * torch.matmul(x, y.t())
    dist = dist.clamp(min=1e-12).sqrt()
    return dist

def cosine_dist(x, y):
    """Computes Cosine Distance."""
    x = F.normalize(x, dim=1)
    y = F.normalize(y, dim=1)
    dist = 2 - 2 * torch.mm(x, y.t())
    return dist
```

Low educational value

```
import re
import typing
...

class Default(object):
    def __init__(self, vim: Nvim) -> None:
        self._vim = vim
        self._denite: typing.Optional[SyncParent]
        = None
        self._selected_candidates: typing.List[int]
        ] = []
        self._candidates: Candidates = []
        self._cursor = 0
        self._entire_len = 0
        self._result: typing.List[typing.Any] = []
        self._context: UserContext = {}
        self._bufnr = -1
        self._winid = -1
        self._winrestcmd = ''
        self._initialized = False
        self._winheight = 0
        self._winwidth = 0
        self._winminheight = -1
        self._is_multi = False
        self._is_async = False
        self._matched_pattern = ''
        ...
```

Chinchilla: Training Compute-Optimal Large Language Models , 2203.15556
Textbooks Are All You Need, 2306.11644
LIMA: Less Is More for Alignment 2305.11206



Summary

- ❑ Tedious annotation tasks will be replaced by AI
- ❑ Human annotation is subjective, inconsistent, and time-consuming.
- ❑ Annotation setup is important to reduce potential biases and artifacts.
- ❑ Lack of dataset for LLM training by Big Techs
- ❑ Potentials and Risks of using synthetic data for AI training
- ❑ Human-AI collaborative data annotation and evaluation



CSCI 5541: Natural Language Processing

Lecture 17: Modern Evaluation



UNIVERSITY OF MINNESOTA
Driven to Discover®

Outline

□ Annotation

- Annotation: terms, examples, and process
- Qualitative coding, recruitment, and quality assessment
- Issues in annotation
- Advanced annotation techniques
- LLMs as Annotators and Synthetic Data

□ Evaluation:

- Properties of Good Benchmarks
- Widely Used Benchmarks and their Metrics
- Latest Benchmarks

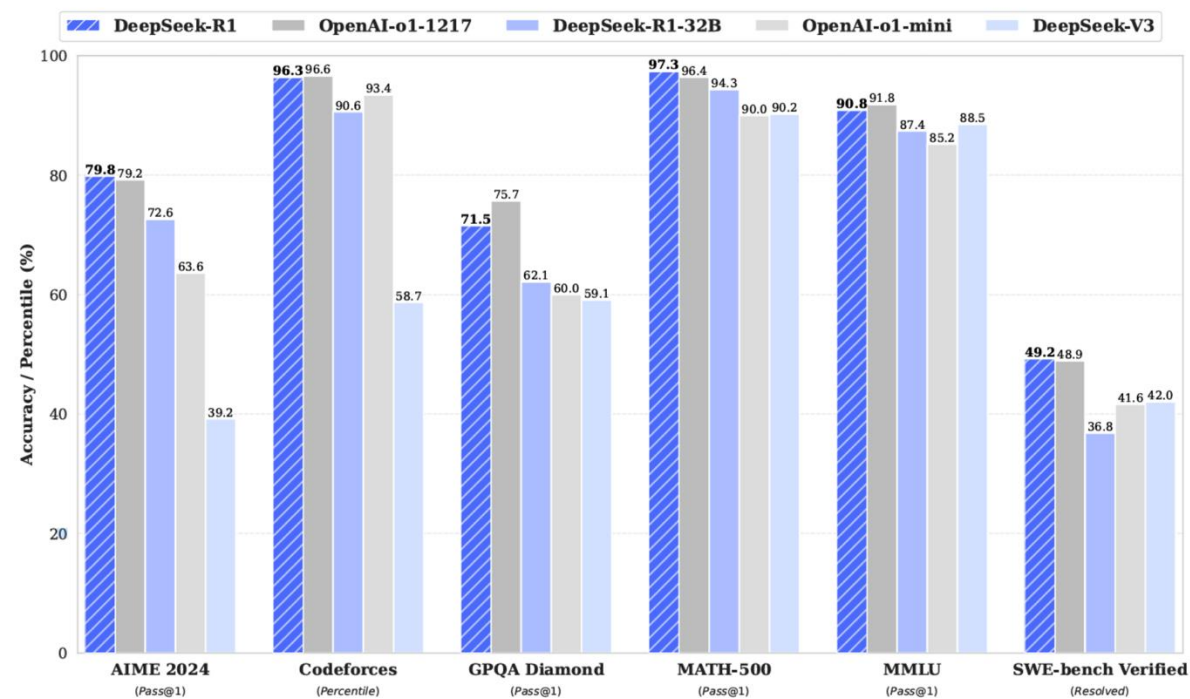


Why do we need Benchmarks?



Track Progress to Compare Performance

Given two models (Model A and Model B), which is better?



	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]

Gpt-4 technical report. arXiv preprint arXiv:2303.08774.

Deepseek-r1: Incentivizing reasoning capability in llms via reinforcement learning. arXiv preprint arXiv:2501.12948.



Track Progress to Compare Performance

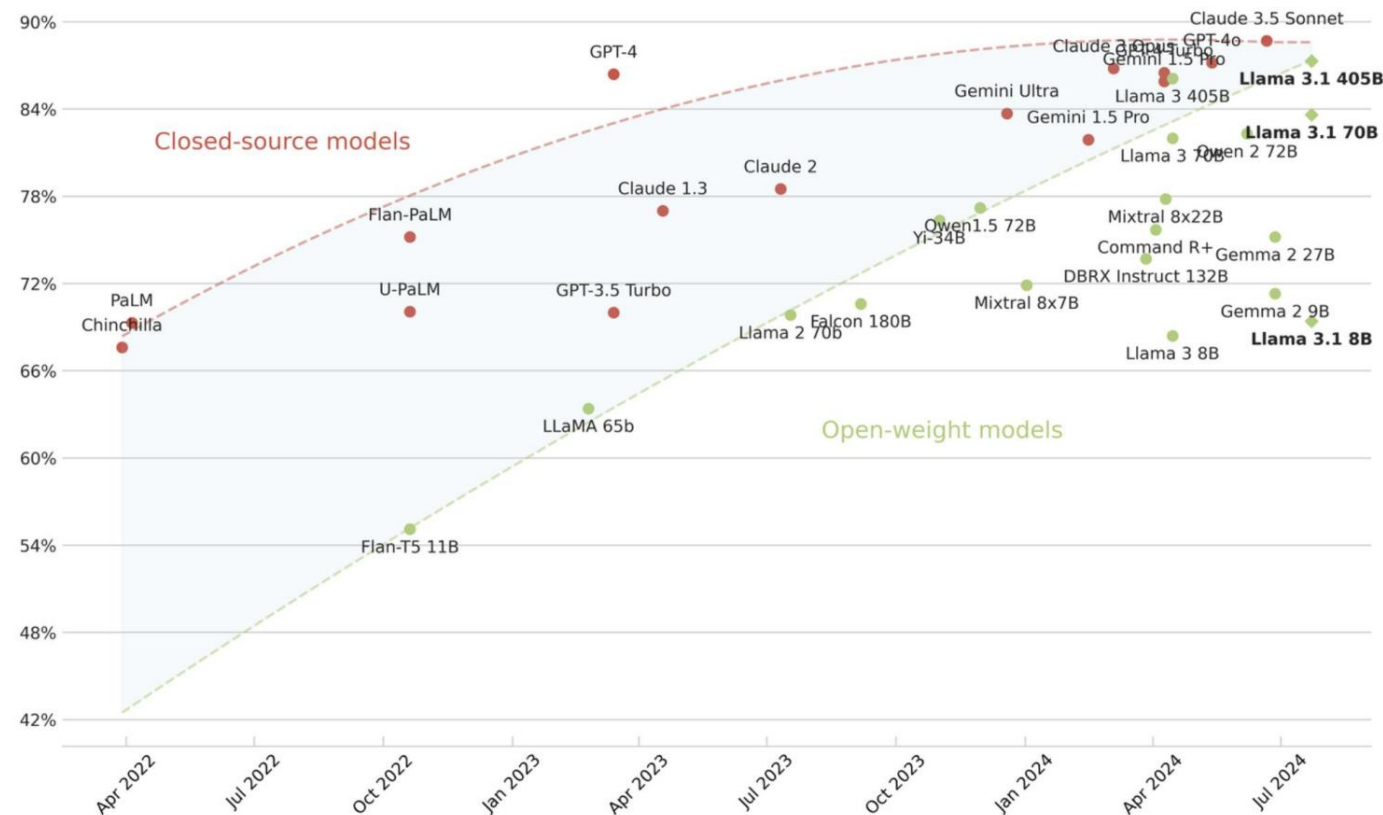
□ How good are open source LLMs compared to proprietary LLMs

Closed-source vs. open-weight models

@maximelabonne

Llama 3.1 405B closes the gap with closed-source models for the first time in history.

MMLU (5-shot)



Properties of Good Benchmarks



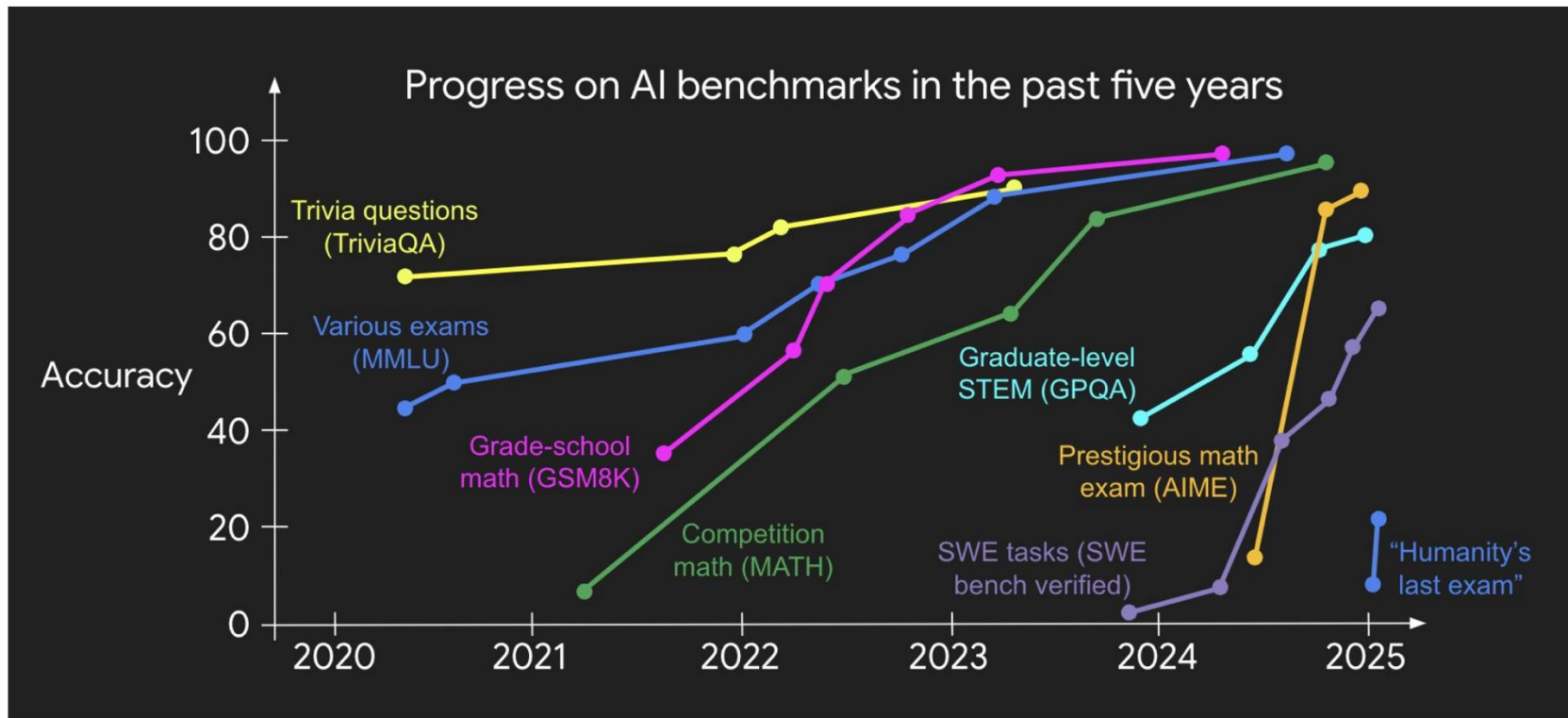
Properties of Good Benchmarks

- ❑ Difficulty
- ❑ Diversity
- ❑ Usefulness
- ❑ Reproducibility
- ❑ Data Contamination



Difficulty

- ❑ Are the problems sufficiently difficult to distinguish the capable models from those which are less capable?

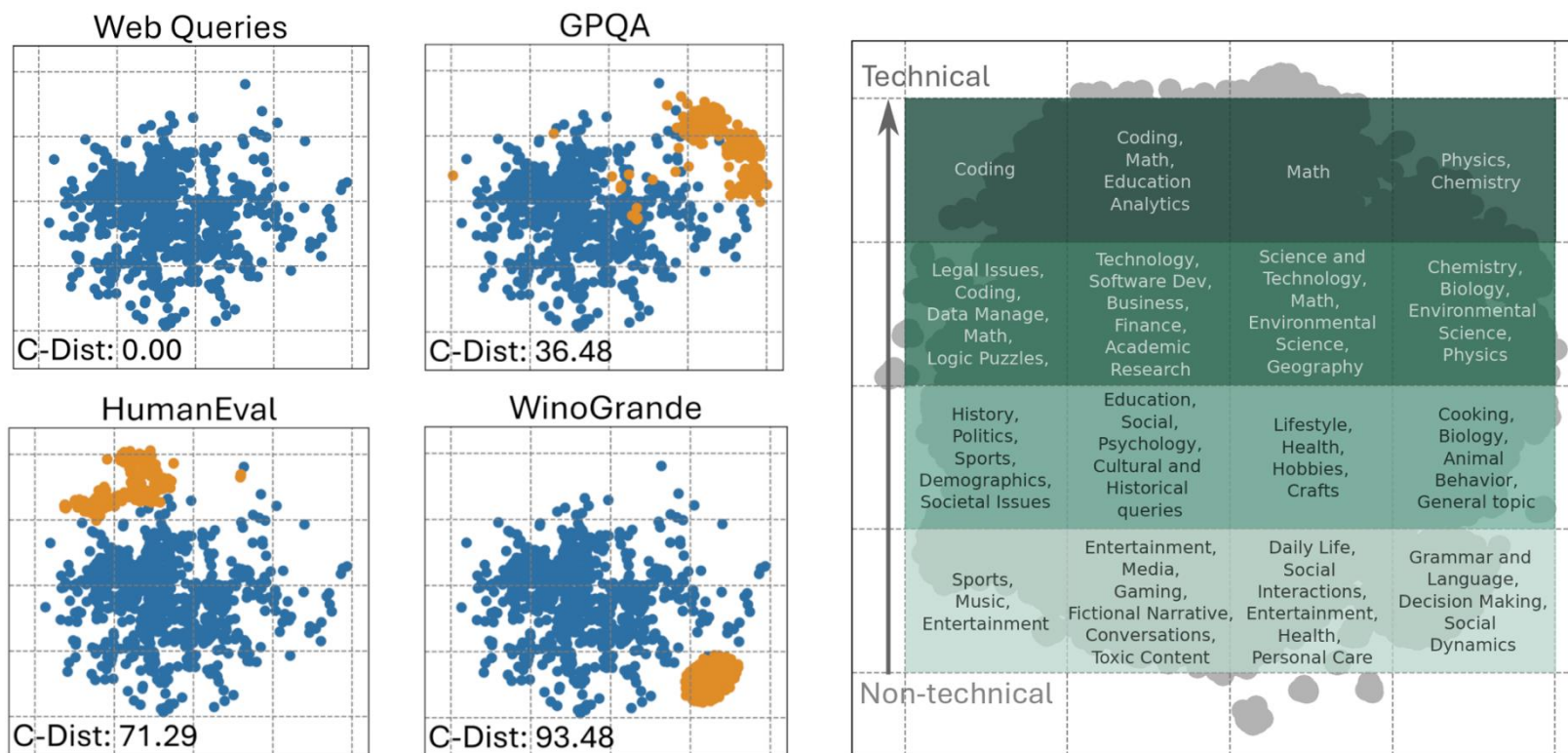


https://x.com/_jasonwei/status/1889096555254456397



Diversity

- Are the problems diverse enough to ensure that the model is effectively being evaluated on its ability to handle a wide range of queries?



MixEval: Deriving Wisdom of the Crowd from LLM Benchmark Mixtures. arXiv preprint arXiv:2406.06565

Usefulness

- ❑ Does achieving a high score on the benchmark have actual meaning? Is it correlated with something usable in the real world?

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: 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 $(-1 + \cos(\frac{\pi}{8}) \sqrt[4]{2})(-1 - \cos(\frac{\pi}{8}) \sqrt[4]{2}) = 1 - \cos^2(\frac{\pi}{8}) \sqrt{2} = 1 - \frac{(1 + \cos(\frac{\pi}{4}))}{2} \sqrt{2} = \boxed{\frac{1 - \sqrt{2}}{2}}$.

Why do we need an LLM that is good at solving math word problems?

Foundation for **more complex tasks**
(e.g., financial analysts)

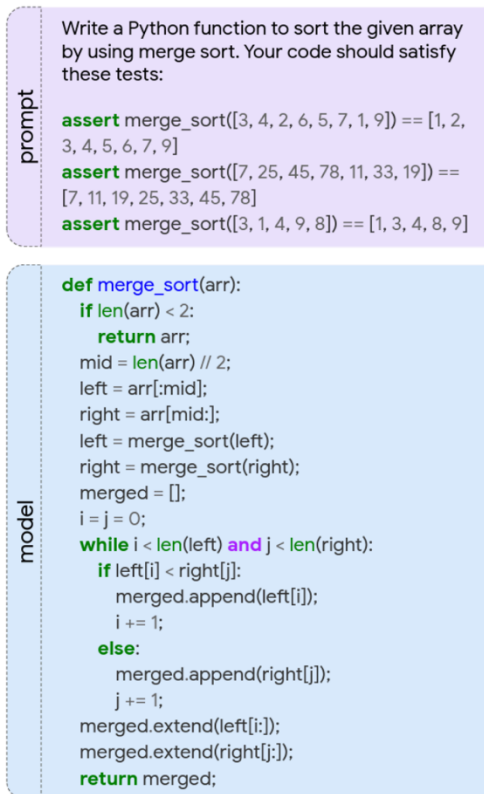
Useful for **real users with similar questions**
(e.g., tutoring high school math)

Medium for **addressing research questions**
(e.g., Can an AI model reason?)



Usefulness

- ❑ Does achieving a high score on the benchmark have actual meaning? Is it correlated with something usable in the real world?



Why do we need an LLM
that is good at solving
leetcode-style problems?

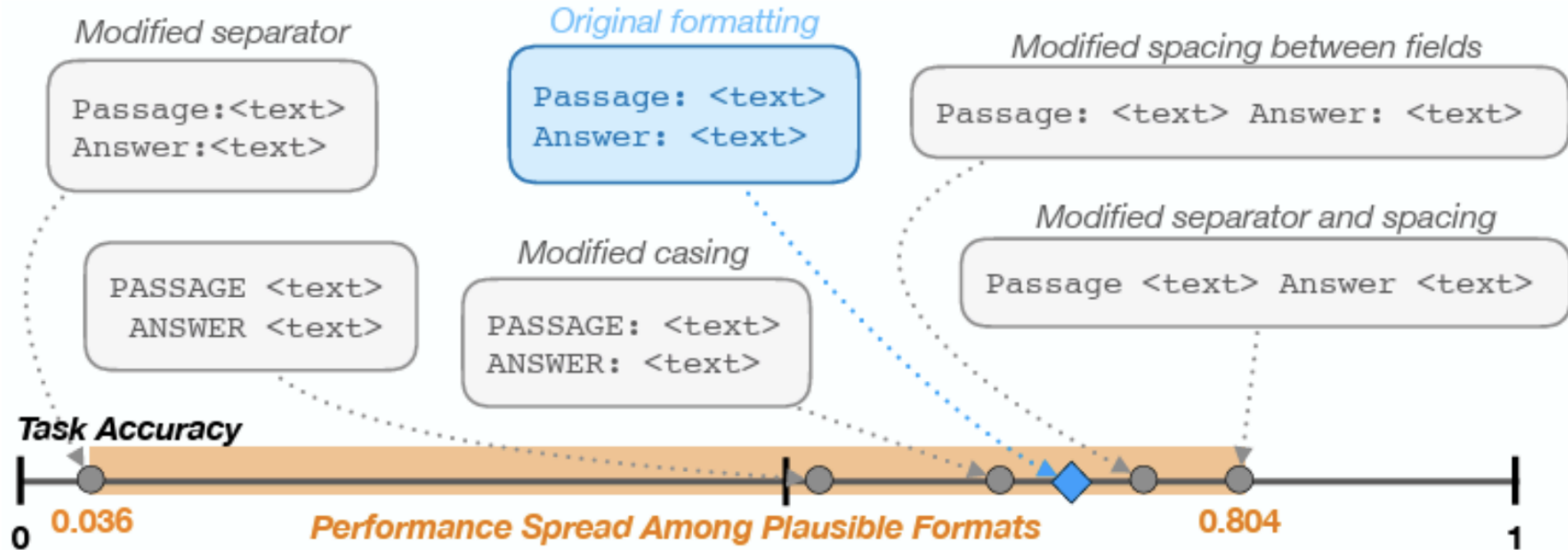
Foundation for **more complex tasks**
(e.g., coding agent that implements a whole repo)

Useful for **real users with similar questions**
(e.g., people preparing for coding interview)

Medium for **addressing research questions**
(e.g., Can an AI model self-debug?)

Reproducibility and Reliability

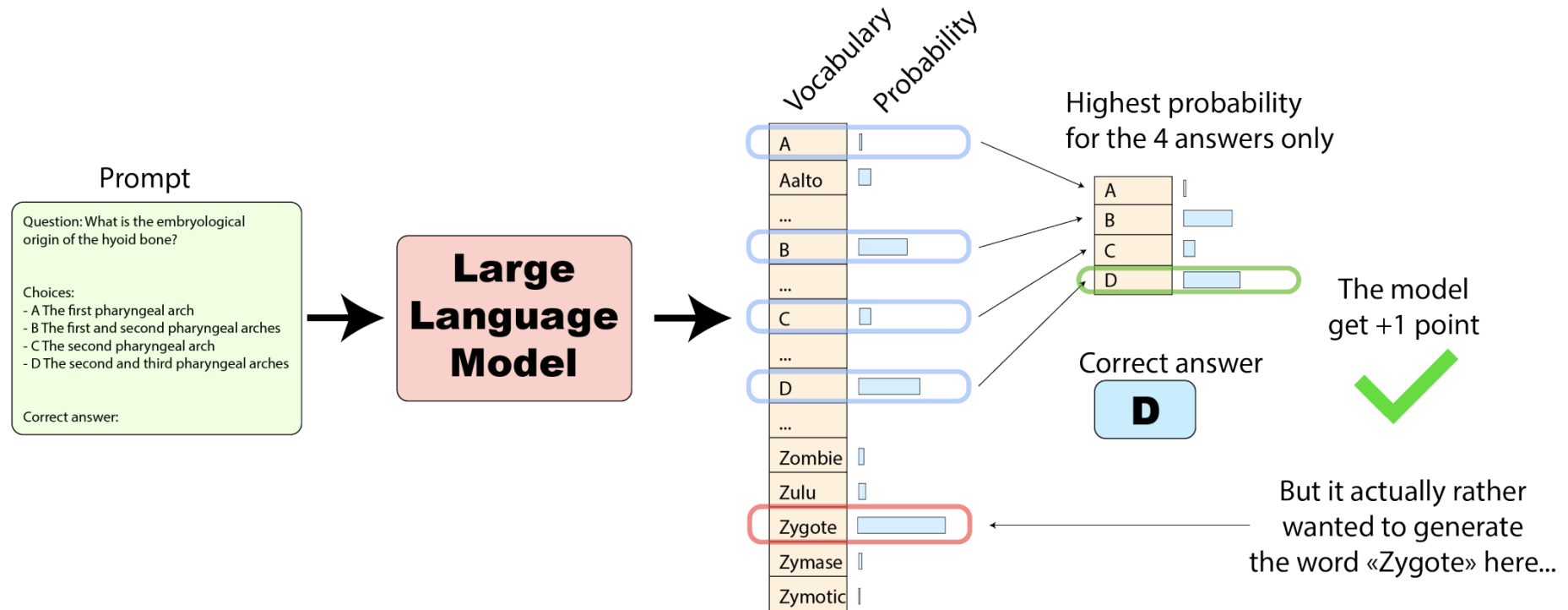
- ❑ Can the model achieve similar scores across multiple test runs/evaluations, ensuring a fair performance comparison between different models?



Quantifying Language Models' Sensitivity to Spurious Features in Prompt Design or: How I learned to start worrying about prompt formatting.

Reproducibility and Reliability

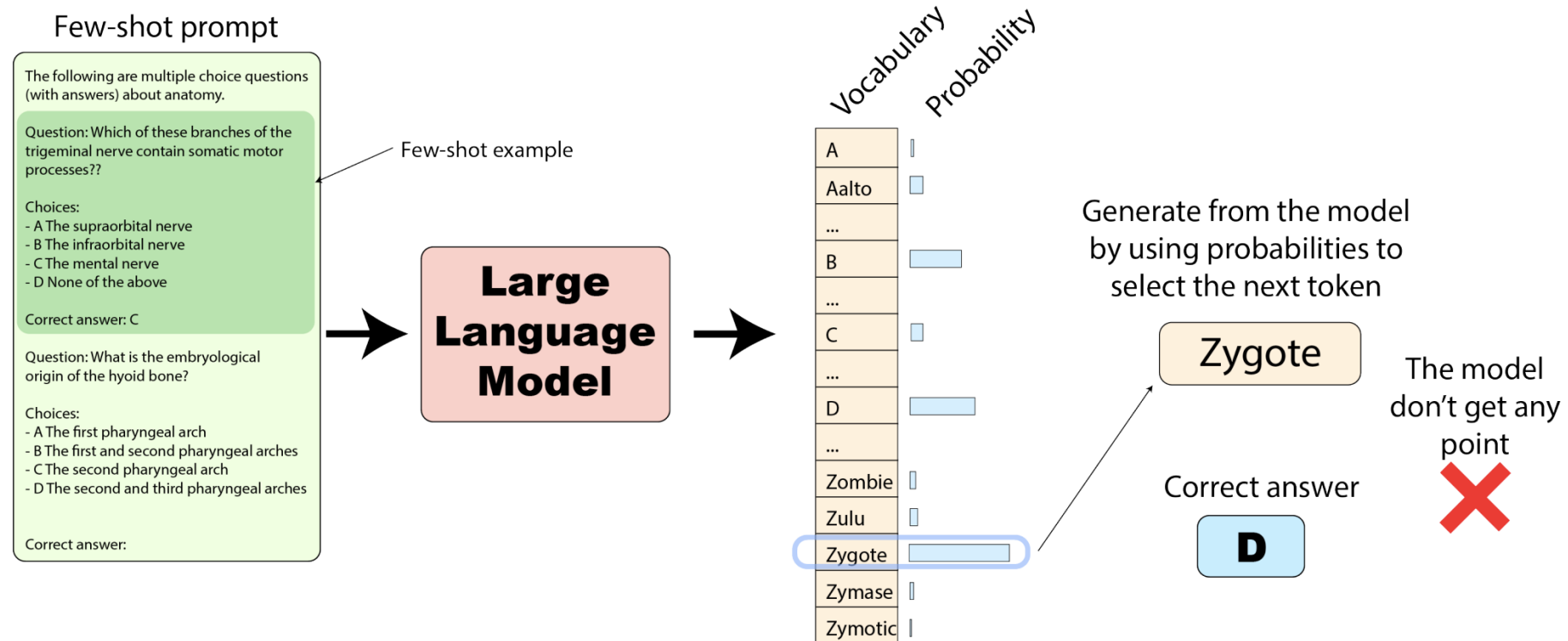
- ❑ Can the model achieve similar scores across multiple test runs/evaluations, ensuring a fair performance comparison between different models?



Orig MMLU Implementation

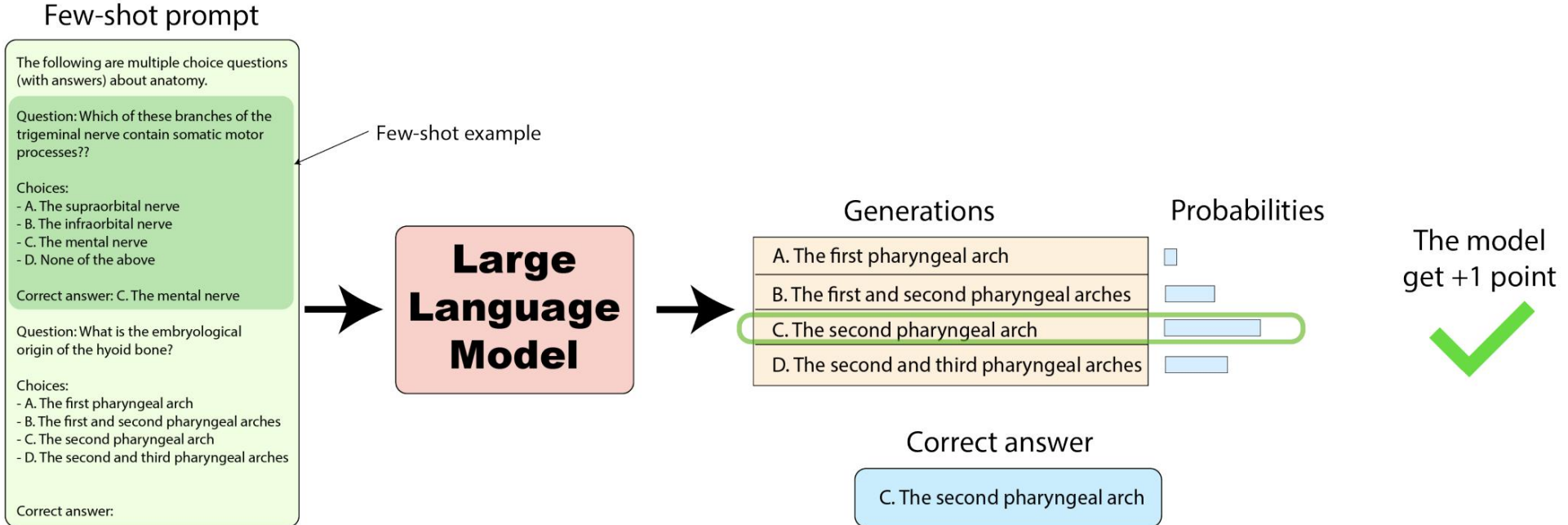
Reproducibility and Reliability

- ❑ Can the model achieve similar scores across multiple test runs/evaluations, ensuring a fair performance comparison between different models?



Reproducibility and Reliability

- ❑ Can the model achieve similar scores across multiple test runs/evaluations, ensuring a fair performance comparison between different models?



Reproducibility and Reliability

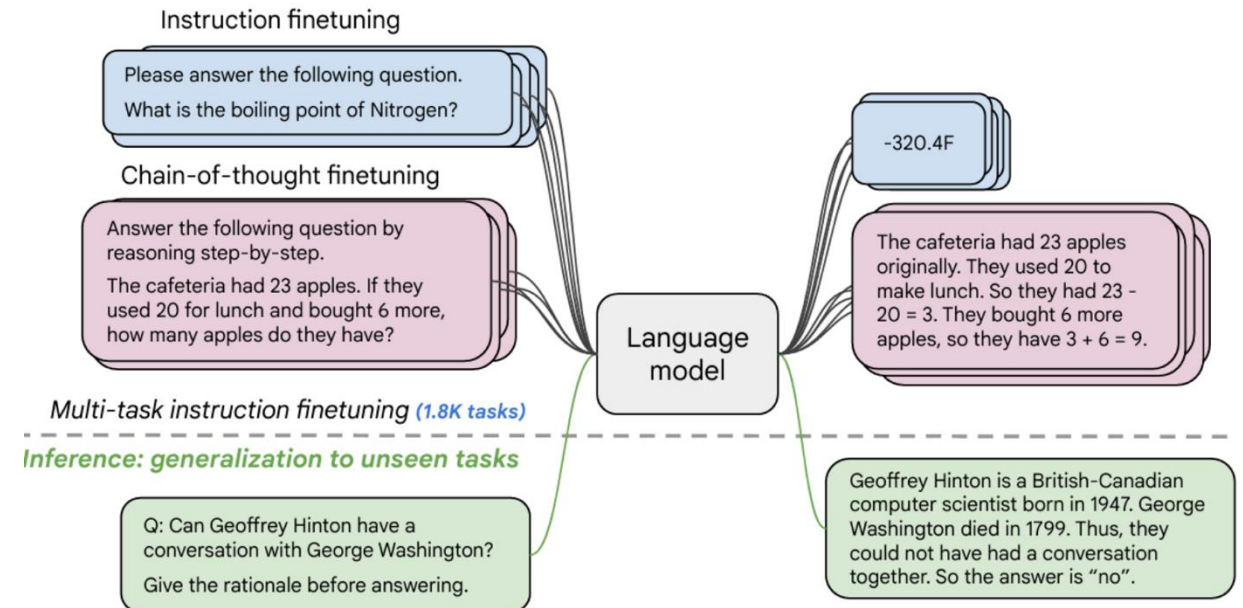
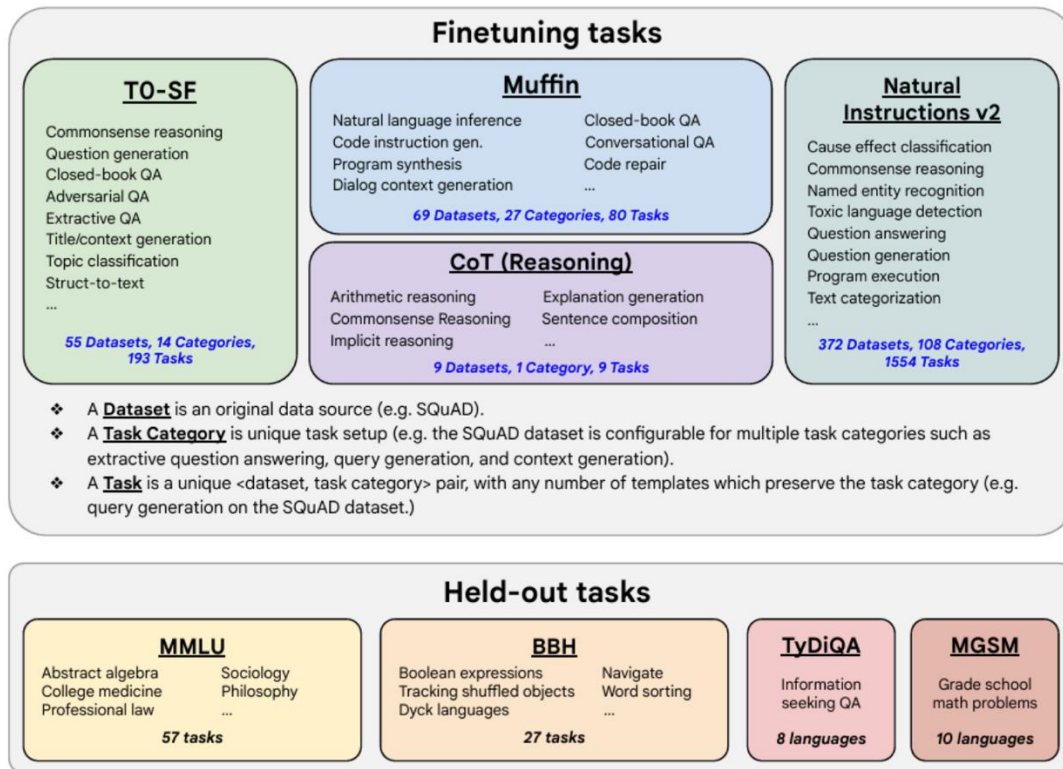
- ❑ Can the model achieve similar scores across multiple test runs/evaluations, ensuring a fair performance comparison between different models?

	MMLU (HELM)	MMLU (Harness)	MMLU (Original)
llama-65b	0.637	0.488	0.636
tiiuae/falcon-40b	0.571	0.527	0.558
llama-30b	0.583	0.457	0.584
EleutherAI/gpt-neox-20b	0.256	0.333	0.262
llama-13b	0.471	0.377	0.47
llama-7b	0.339	0.342	0.351
tiiuae/falcon-7b	0.278	0.35	0.254
togethercomputer/RedPajama-INCITE-7B-Base	0.275	0.34	0.269



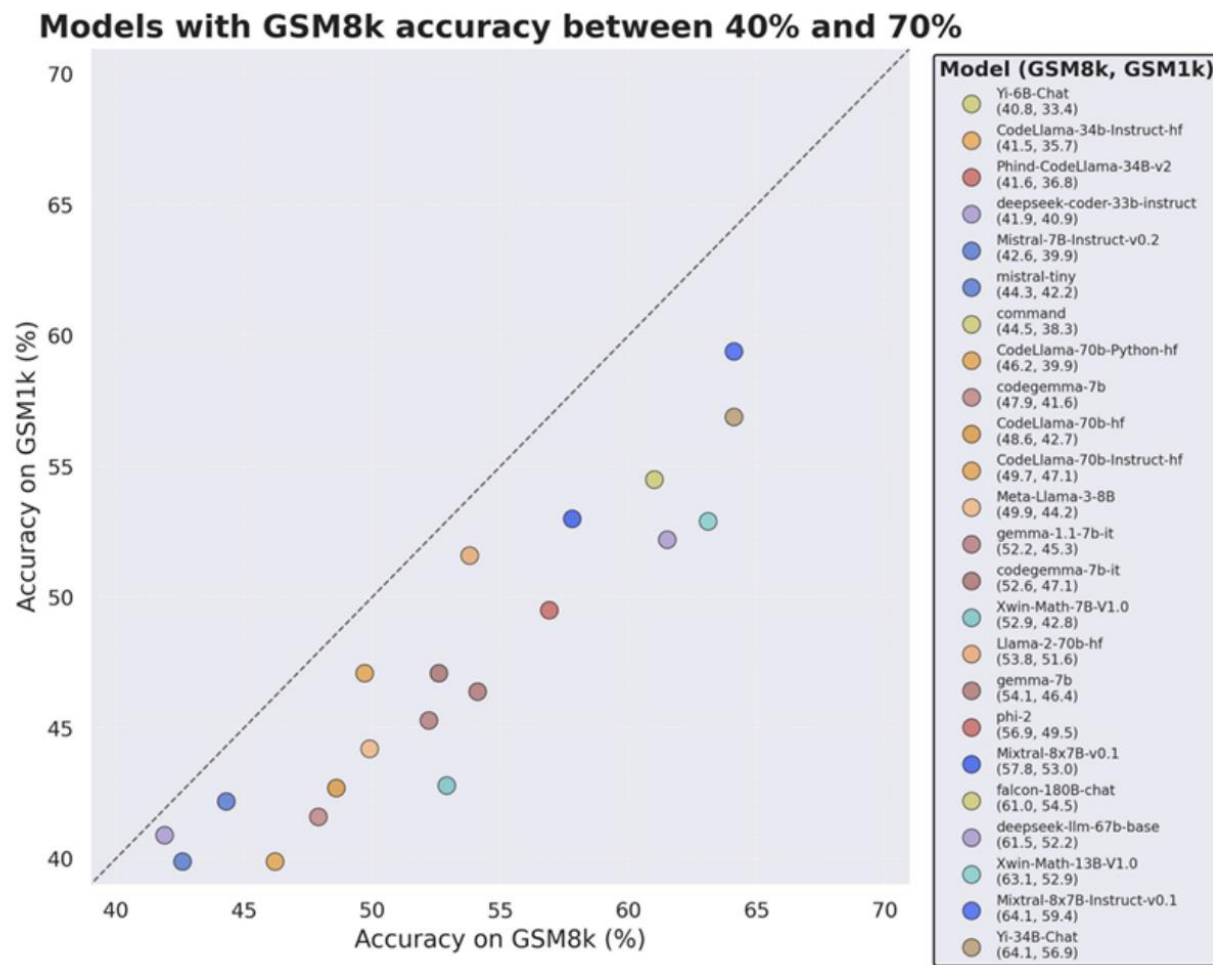
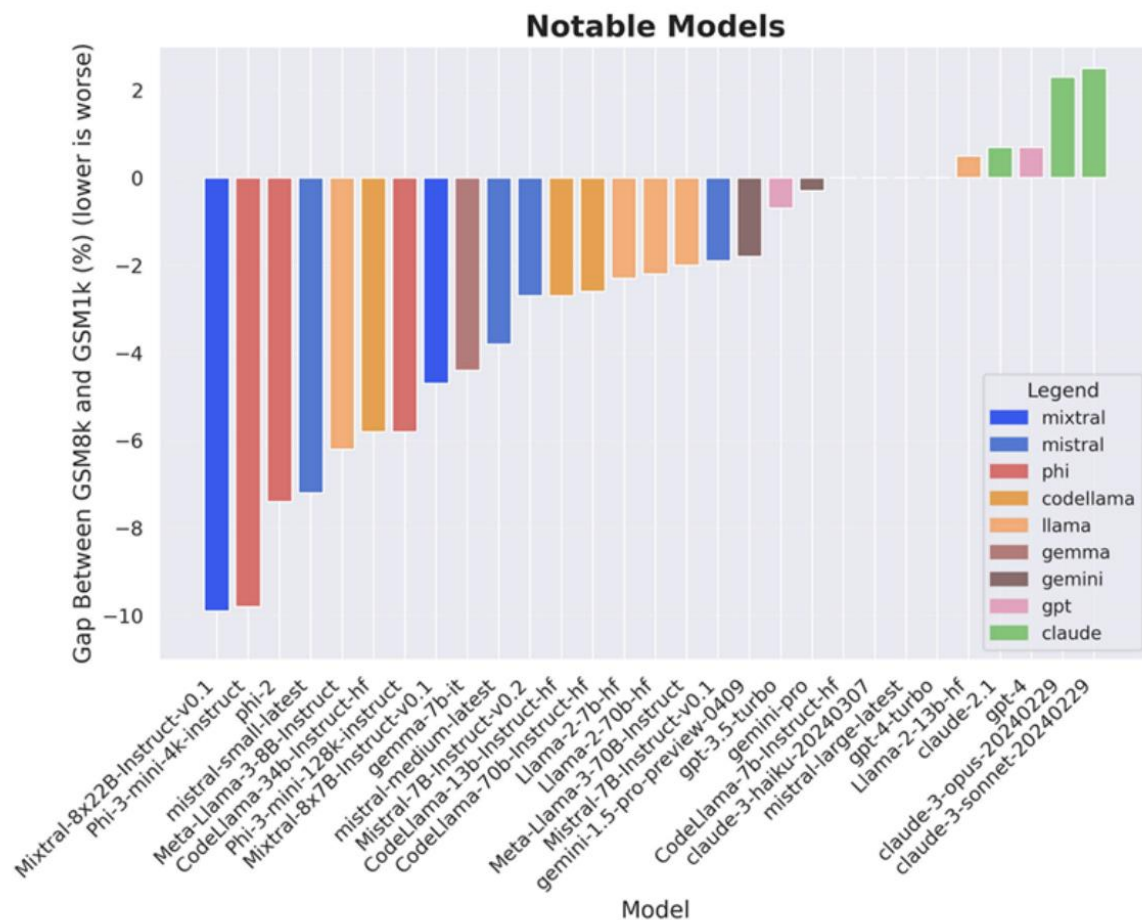
Data Contamination

- ❑ How can we ensure that we are evaluating their ability to generalize to unseen novel tasks rather than memorizing similar problem encountered during training?



Scaling instruction finetuned language models. Journal of Machine Learning Research, 25(70), pp.1-53.

Data Contamination



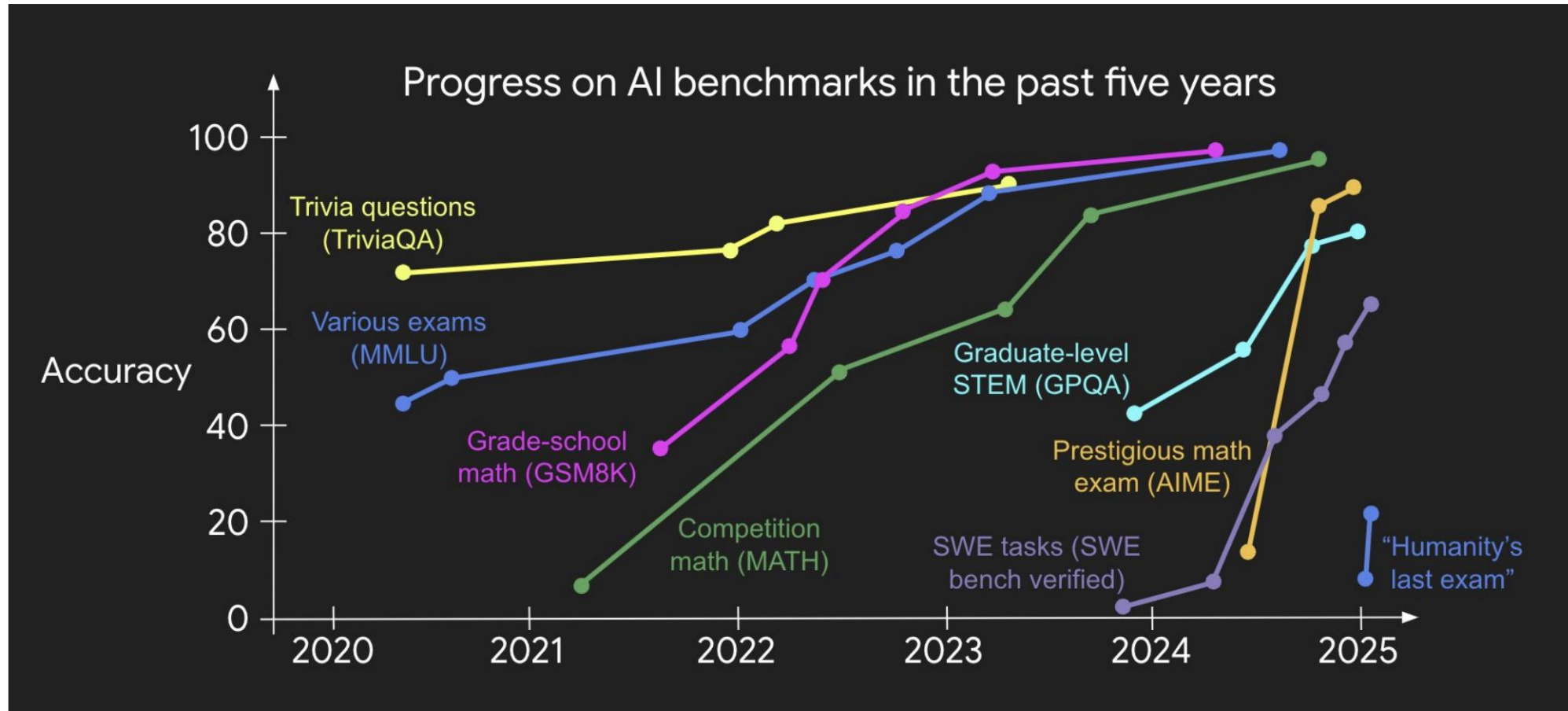
A careful examination of large language model performance on grade school arithmetic. arXiv preprint arXiv:2405.00332.

Widely Used Benchmarks and their Metrics



Multiple-Choice QA (Classification)

❑ (Recap) Benchmarks saturate over time



Multiple-Choice QA (Classification)

□ HellaSwag: Commonsense reasoning around everyday events

ACTIVITYNET A woman is outside with a bucket and a dog. The dog is running around trying to avoid a bath. She...

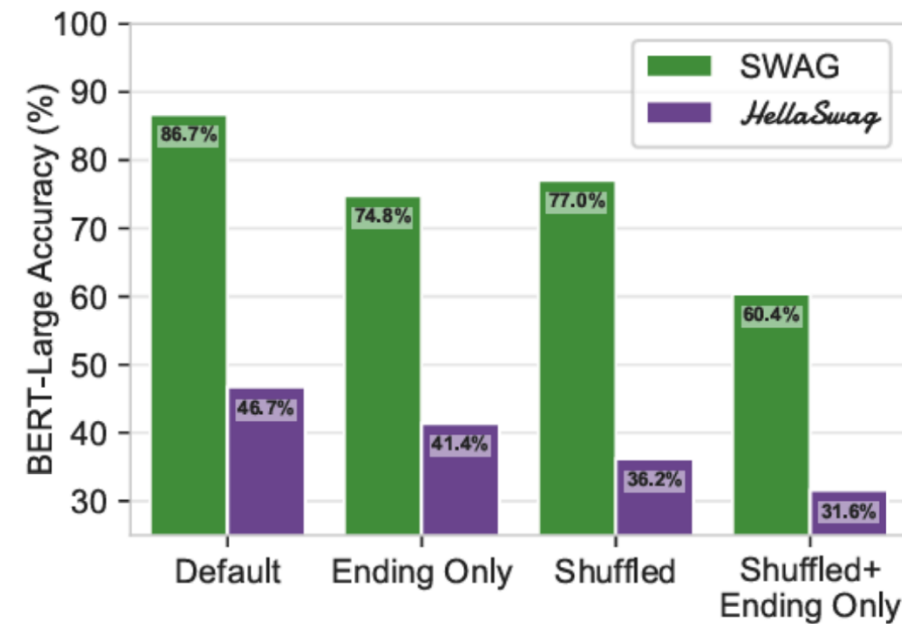
A. rinses the bucket off with soap and blow dry the dog's head.
B. uses a hose to keep it from getting soapy.
C. gets the dog wet, then it runs away again.
D. gets into a bath tub with the dog.

wikiHow Come to a complete halt at a stop sign or red light. At a stop sign, come to a complete halt for about 2 seconds or until vehicles that arrived before you clear the intersection. If you're stopped at a red light, proceed when the light has turned green. ...

A. Stop for no more than two seconds, or until the light turns yellow. A red light in front of you indicates that you should stop.
B. After you come to a complete stop, turn off your turn signal. Allow vehicles to move in different directions before moving onto the sidewalk.
C. Stay out of the oncoming traffic. People coming in from behind may elect to stay left or right.
D. If the intersection has a white stripe in your lane, stop before this line. Wait until all traffic has cleared before crossing the intersection.

easy! ???

OpenAI GPT



Zellers, R., Holtzman, A., Bisk, Y., Farhadi, A. and Choi, Y., 2019, July. HellaSwag: Can a Machine Really Finish Your Sentence?. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (pp. 4791-4800).

Multiple-Choice QA (Classification)

❑ MMLU: Multiple-choice questions in 57 subjects (professional & academic)

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.	✓

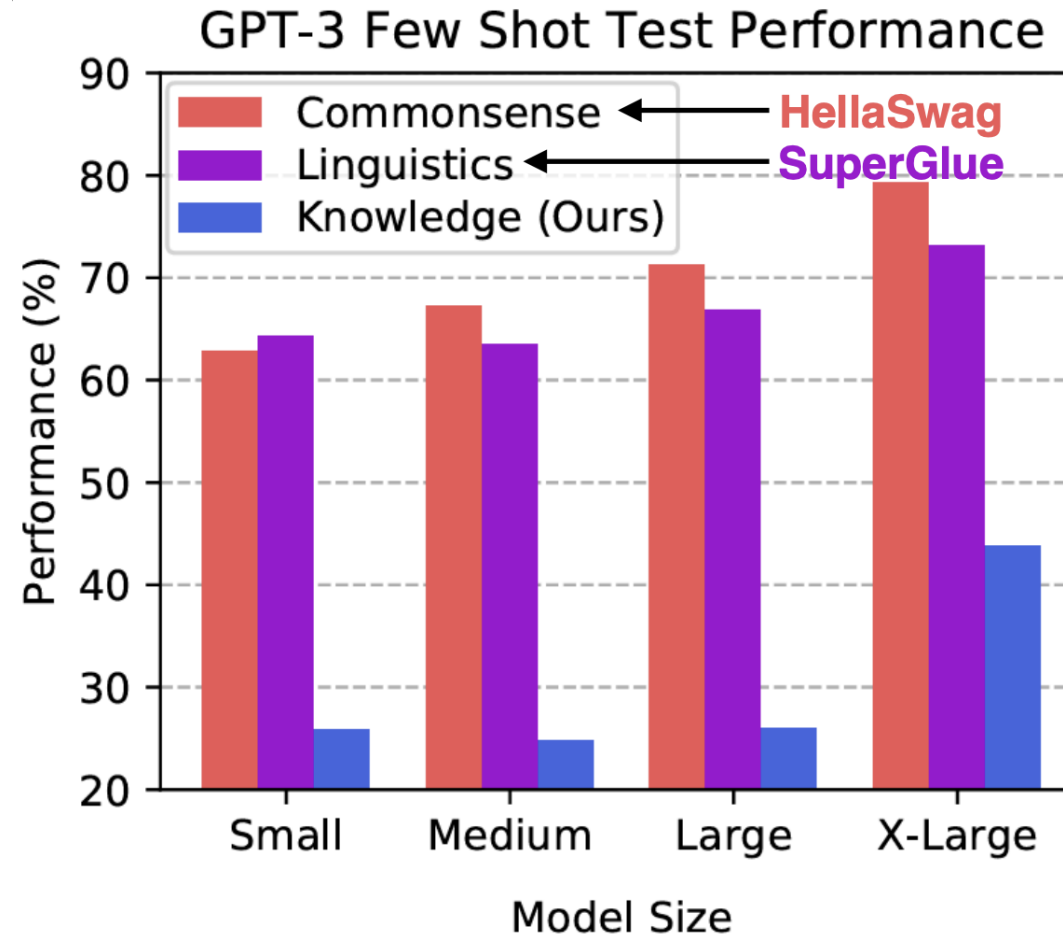
Figure 3: Examples from the Microeconomics task.

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	✓



Multiple-Choice QA (Classification)

❑ MMLU: Multiple-choice questions in 57 subjects (professional & academic)



Measuring Massive Multitask Language Understanding. In International Conference on Learning Representations



Multiple-Choice QA (Classification)

❑ MMLU-Pro: Harder MMLU with 4~10 options to choose from.

Question: Where do most short-period comets come from and how do we know?

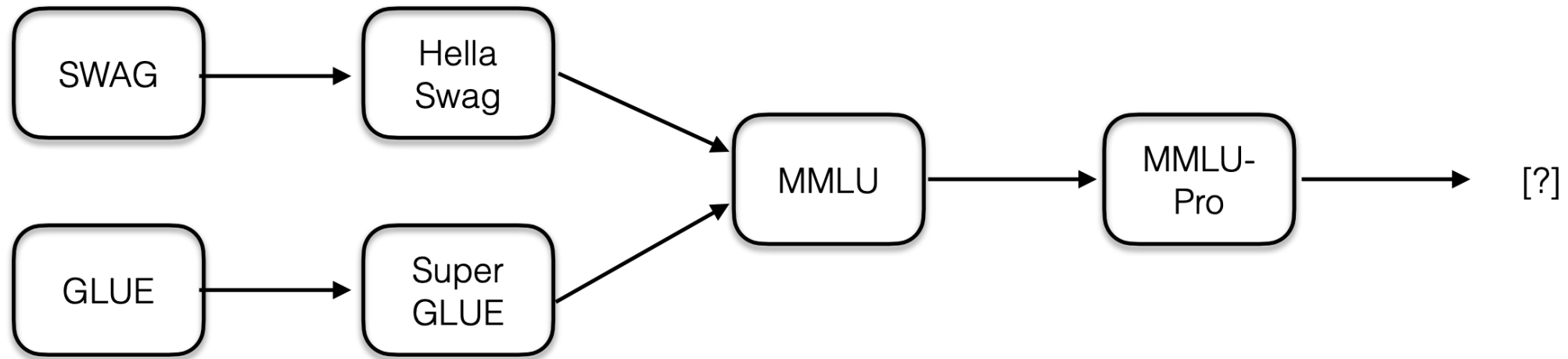
- Options:**
- A. The Kuiper belt; short period comets tend to be in the plane of the solar system like the Kuiper belt.
 - B. The asteroid belt; short period comets tend to come from random directions indicating a spherical distribution of comets called the asteroid belt.
 - C. The asteroid belt; short period comets tend to be in the plane of the solar system just like the asteroid belt.
 - D. The Oort cloud; short period comets have orbital periods similar to asteroids like Vesta and are found in the plane of the solar system just like the Oort cloud.
 - E. The Oort Cloud; short period comets tend to come from random directions indicating a spherical distribution of comets called the Oort Cloud.
 - F. The Oort cloud; short period comets tend to be in the plane of the solar system just like the Oort cloud.
 - G. The asteroid belt; short period comets have orbital periods similar to asteroids like Vesta and are found in the plane of the solar system just like the asteroid belt.

Answer: Let's think step by step. Most short-period comets originate from the Kuiper belt. This is deduced from the observation that these comets tend to follow orbits that lie in the plane of the solar system, similar to the distribution of objects in the Kuiper belt itself. Thus, the alignment of these cometary orbits with the ecliptic plane points to their Kuiper belt origin. The answer is (A).



Multiple-Choice QA (Classification)

- ❑ The trend of developing increasingly challenging benchmarks continue
- ❑ Open Research Question: How can we come up with more challenging benchmarks as LMs become as intelligent as or even surpass humans?



Generation Tasks

- ❑ As humans, we don't present LMs with four options to choose from
- ❑ However, assessing free form responses is more challenging than choosing from among a set of four options, where accuracy can be used as the metric



Generation Tasks

□ GSM8k: Grade School mathematics questions

- For math word problems, the answer is a numeric value, so we could use **Exact Match (EM)**.

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 $3 \times 12 = \langle\langle 3 \times 12 = 36 \rangle\rangle 36$ sodas

6 people attend the party, so half of them is $6/2 = \langle\langle 6/2 = 3 \rangle\rangle 3$ people

Each of those people drinks 3 sodas, so they drink $3 \times 3 = \langle\langle 3 \times 3 = 9 \rangle\rangle 9$ sodas

Two people drink 4 sodas, which means they drink $2 \times 4 = \langle\langle 4 \times 2 = 8 \rangle\rangle 8$ sodas

With one person drinking 5, that brings the total drank to $5 + 9 + 8 + 3 = \langle\langle 5 + 9 + 8 + 3 = 25 \rangle\rangle 25$ sodas

As Tina started off with 36 sodas, that means there are $36 - 25 = \langle\langle 36 - 25 = 11 \rangle\rangle 11$ sodas left

Final Answer: 11



Generation Tasks

❑ HumanEval: Python Coding Tasks

- For LeetCode-style coding problems, we can execute the code and **verify whether it passes all test cases (i.e., pass@k).**

```
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, M., Tworek, J., Jun, H., Yuan, Q., Pinto, H.P.D.O., Kaplan, J., Edwards, H., Burda, Y., Joseph, N., Brockman, G. and Ray, A., 2021. Evaluating large language models trained on code. arXiv preprint arXiv:2107.03374.



Evaluation for Open-Ended Generation Tasks

❑ LLM-as-judge: Prompt a LM to provide a judgment (e.g., "3 out of 5", A is better than B)

GPT-4 Judgment (when A is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant A's answer is more detailed and organized, covering a wider range of topics such as dress code and language, which are not mentioned in Assistant B's response. Assistant A also provides more specific advice on how to handle certain situations, such as exchanging business cards and dining etiquette. Therefore, Assistant A's answer is more comprehensive and useful for someone looking to understand Japanese business etiquette.

GPT-4 Judgment (when B is placed in the first position):

Both Assistant A and Assistant B provide helpful and relevant information about business etiquette norms in Japan. However, Assistant B's answer is more detailed and covers a wider range of topics, such as addressing people by their titles, avoiding direct eye contact, and being mindful of communication styles. Assistant A's answer is more concise but lacks some of the depth found in Assistant B's response. Therefore, Assistant B's answer is better in this case.

Question: What are some business etiquette norms when doing business in Japan?

Assistant A:

Here are some business etiquette norms to keep in mind when doing business in Japan:

1. Bowing: Bowing is a common greeting in Japan. It is a sign of respect and should be reciprocated.
2. Business cards: Business cards are exchanged at the beginning of a meeting. It is important to receive and give business cards with both hands and to take a moment to read the card before putting it away.
3. Punctuality: Being on time is important in Japan. Arrive a few minutes early to show respect for your colleagues.
4. Dress code: Dress conservatively and professionally. Men should wear dark suits and ties, while women should wear conservative business attire.
5. Language: Learn a few basic Japanese phrases to show respect for the culture. It is also important to speak clearly and avoid using slang or idioms.
6. Gift giving: Gift giving is a common practice in Japan. Bring a small gift, such as a box of chocolates or a small souvenir from your home country.
7. Dining etiquette: When dining with colleagues, wait for the host to begin eating before you start. Use chopsticks correctly and avoid leaving food on your plate.

Remember, showing respect for Japanese culture and customs is key to building strong business relationships in Japan.

Assistant B:

When doing business in Japan, it is important to be aware of and respect the country's business etiquette norms. Some of these norms include:

1. Greetings: A handshake and bow is the standard greeting in Japan. It is important to bow deeply and to offer a firm handshake.
2. Business cards: In Japan, business cards are exchanged after the initial greeting. It is important to handle the card with respect and to read it carefully before placing it on the table.
3. Meetings: Meetings in Japan are typically formal and conducted in a hierarchical manner. It is important to address people by their titles and to avoid direct eye contact.
4. Punctuality: Punctuality is highly valued in Japan. It is important to arrive on time for meetings and to be prepared in advance.
5. Gift giving: Gift giving is a common practice in Japan, but it is important to be mindful of the occasion and the relationship. Avoid giving gifts that are considered too personal or extravagant.
6. Dining: Dining is an important part of doing business in Japan. It is important to be aware of the proper table manners and to avoid slurping or making loud noises while eating.
7. Communication: Communication in Japan is typically indirect and reserved. It is important to listen carefully and to avoid confrontational language.

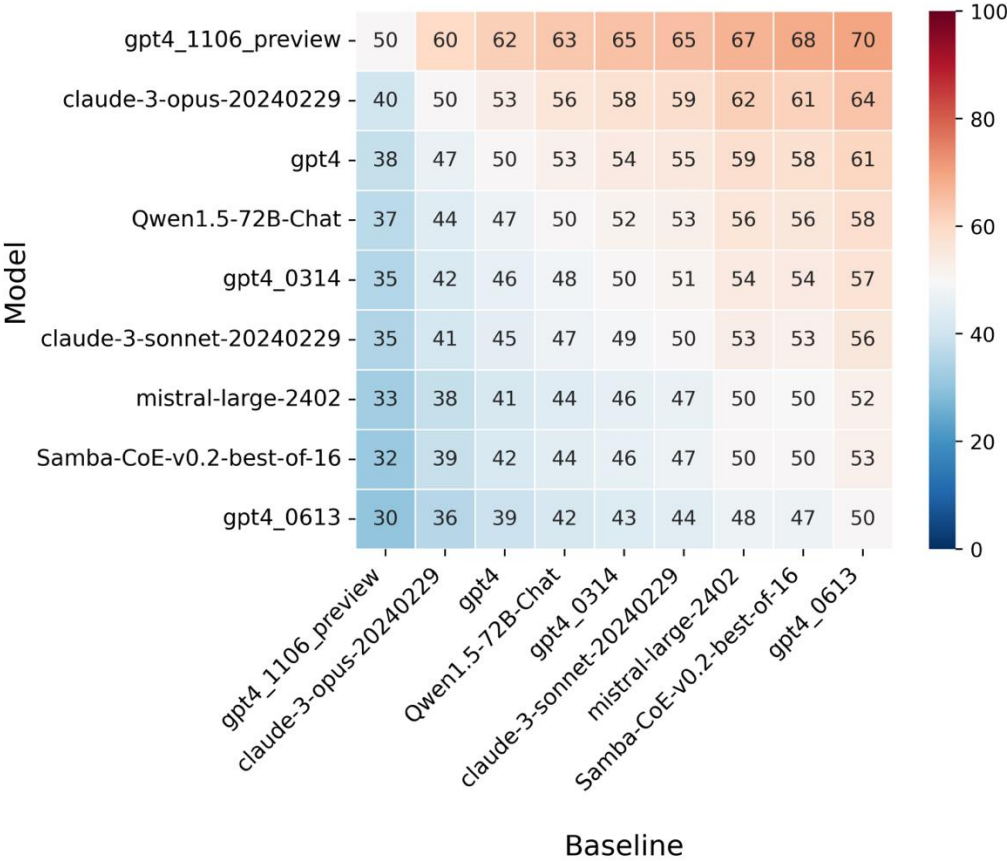
Overall, it is important to be respectful, punctual, and mindful of hierarchy when doing business in Japan.

Zheng, L., Chiang, W.L., Sheng, Y., Zhuang, S., Wu, Z., Zhuang, Y., Lin, Z., Li, Z., Li, D., Xing, E. and Zhang, H., 2024. Judging llm-as-a-judge with mtbench and chatbot arena. Advances in Neural Information Processing Systems, 36.



Generation Tasks

- AlpacaEval: Calculate the win-rate against a reference model on open-ended generation tasks
- LMSys Chatbot Arena: Ask humans to vote which response is better



Rank* (UB)	Rank (StyleCtrl)	Model	Arena Score	95% CI	Votes	Organiza	License
1	1	chocolate... (Early Grok-3)	1402	+7/-6	7829	xAI	Proprietary
2	4	Gemini-2.0-Flash-Thinking-Exp-01-21	1385	+5/-5	13336	Google	Proprietary
2	2	Gemini-2.0-Pro-Exp-02-05	1379	+5/-6	11197	Google	Proprietary
2	1	ChatGPT-4o-latest (2025-01-29)	1377	+5/-6	10529	OpenAI	Proprietary
5	2	DeepSeek-R1	1361	+8/-7	5079	DeepSeek	MIT
5	8	Gemini-2.0-Flash-001	1356	+6/-5	9092	Google	Proprietary
5	2	o1-2024-12-17	1353	+6/-5	15437	OpenAI	Proprietary

2024. Length-controlled alpacaeval: A simple debiasing of automatic evaluators.



Latest Benchmarks



US Math Olympiad Questions

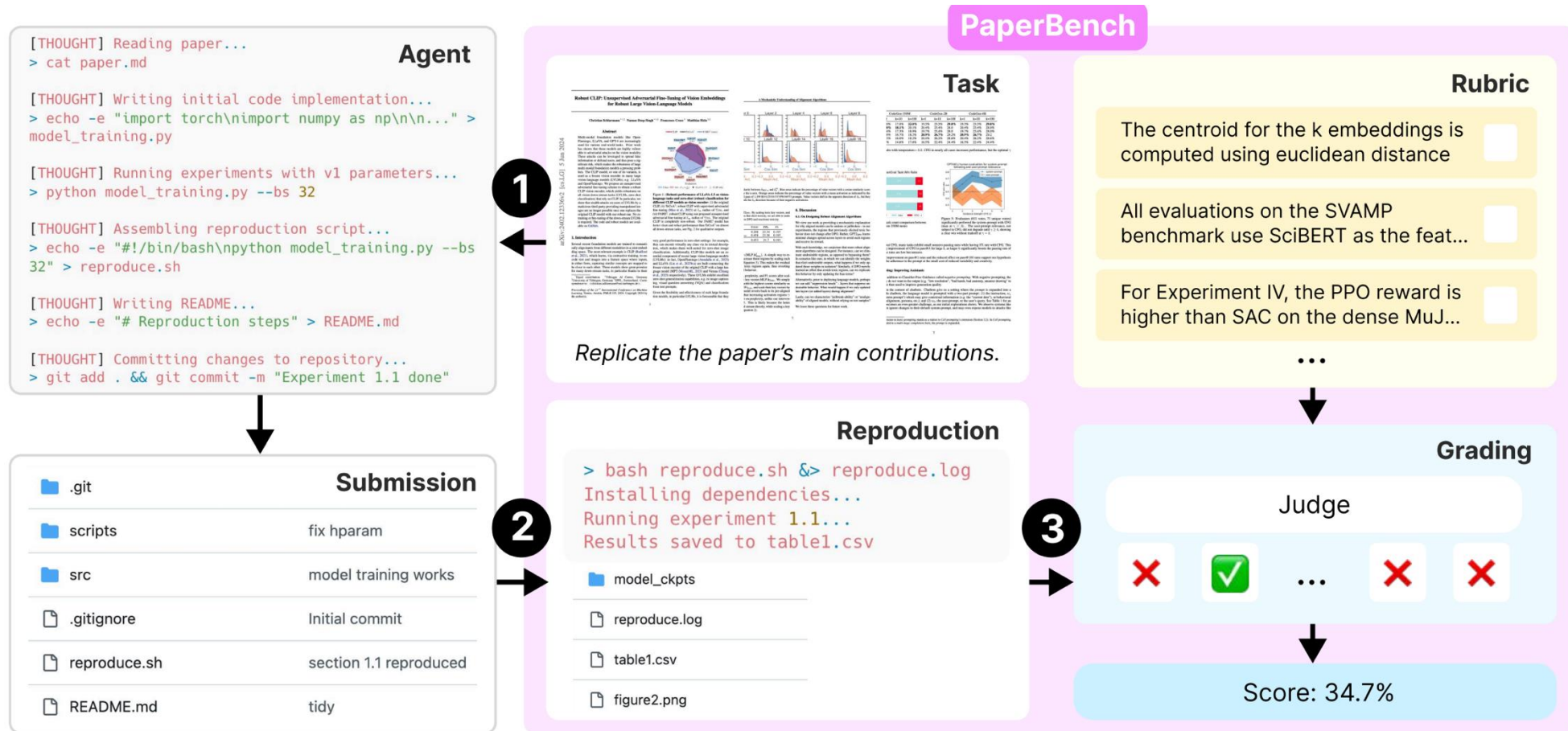
- ❑ Very novel questions (likely chance of data contamination is low)
- ❑ Very challenging questions

Model	P1	P2	P3	P4	P5	P6	Total	Cost
R1	0.5	0.0	0.0	1.5	0.0	0.0	2.0	2.03
FLASH-THINKING	1.5	0.0	0.0	0.0	0.2	0.0	1.8	N/A
CLAUDE 3.7	0.5	0.5	0.0	0.0	0.0	0.6	1.5	9.03
QWQ	1.2	0.0	0.0	0.0	0.0	0.0	1.2	0.42
O1-PRO	0.5	0.0	0.0	0.0	0.2	0.4	1.2	203.44
O3-MINI	0.5	0.1	0.0	0.0	0.0	0.2	0.9	1.11



US Math Olympiad Questions

❑ Replicate entire papers

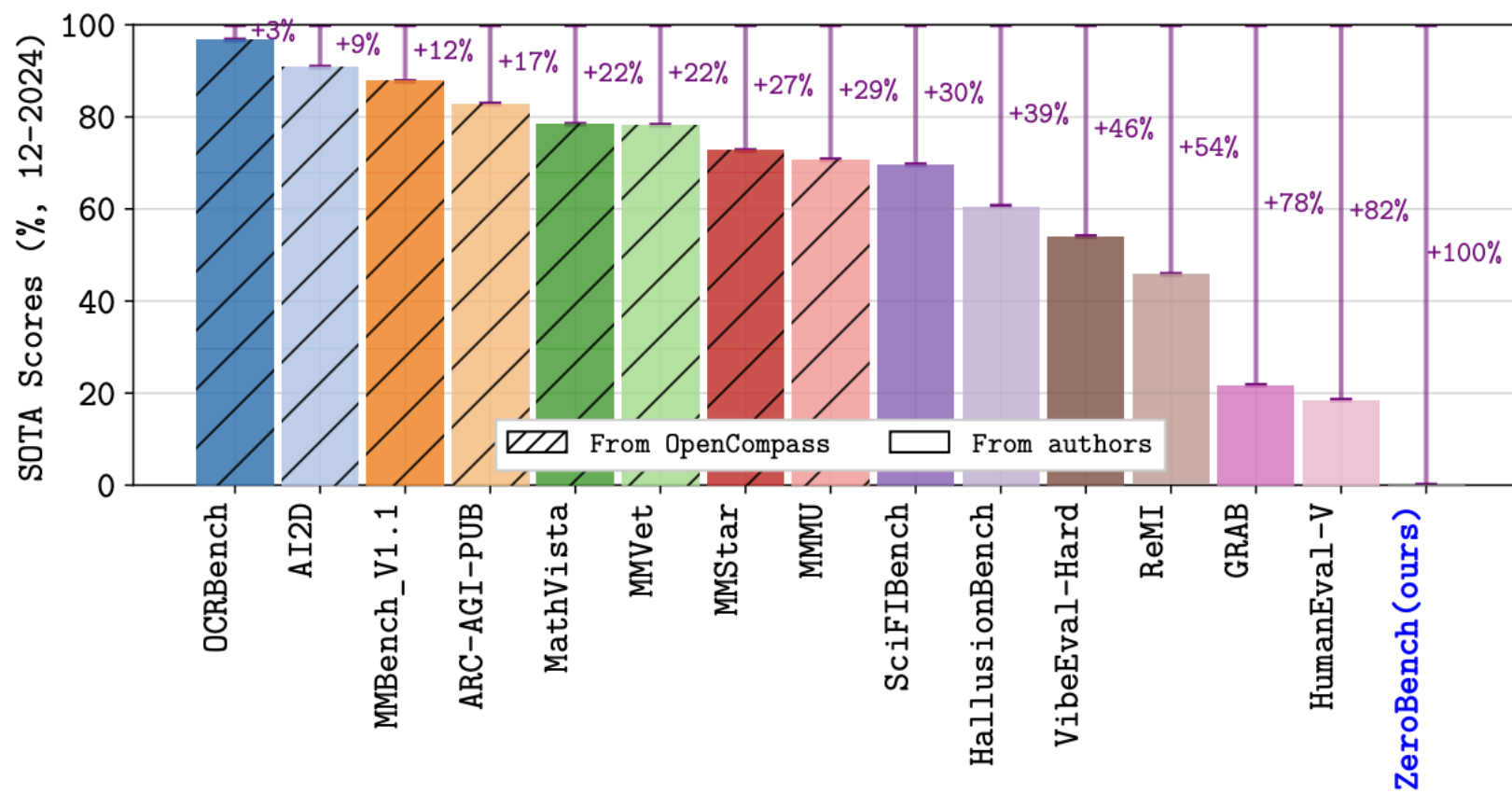


<https://cdn.openai.com/papers/22265bac-3191-44e5-b057-7aaacd8e90cd/paperbench.pdf>



ZeroBench

❑ Challenging multimodal question-answering dataset



<https://arxiv.org/pdf/2502.09696>

ZeroBench



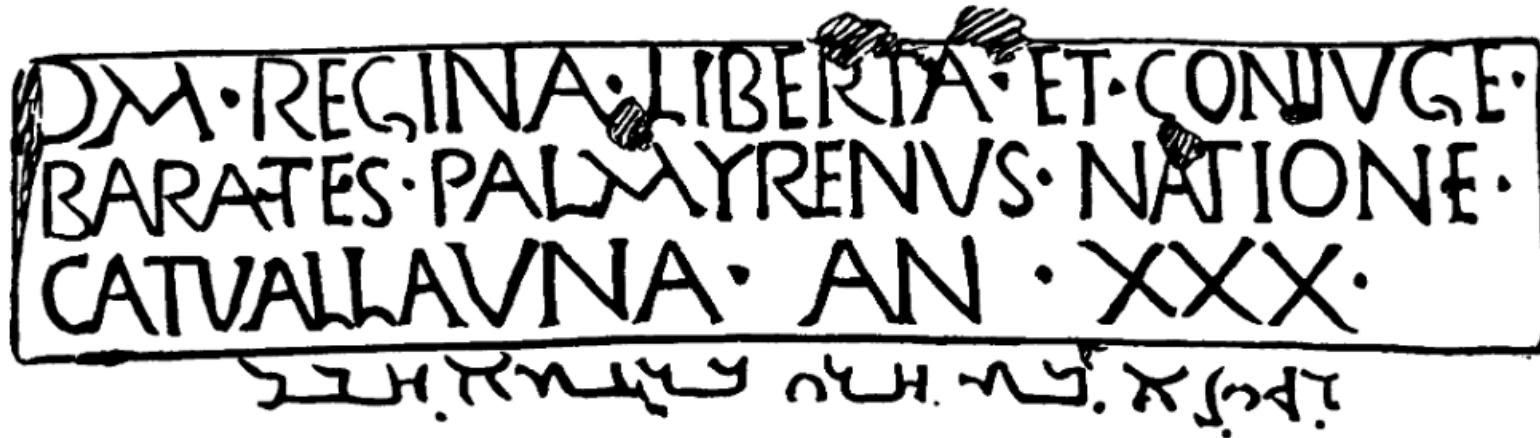
(1) It's evening in Oxford, UK. The clock is built to align with how the object is oriented in when used. To the nearest integer, how many hours ago was noon? (2) The clock was built from the equipment owned by an oarsman in a College men's eight. His crew rowed in a standard rig. What is the sum of the seat numbers that he could have sat in? (3) Suppose instead that the time is read based on the clock orientation in the image. If the clock hands were reflected once horizontally, once vertically, then rotated once by 90° clockwise. Which round number is the hour hand closest to? What is the product of the previous 3 answers?

<https://arxiv.org/pdf/2502.09696>

Humanity's Last Exam (HLE)

- ❑ Challenging multimodal question-answering dataset

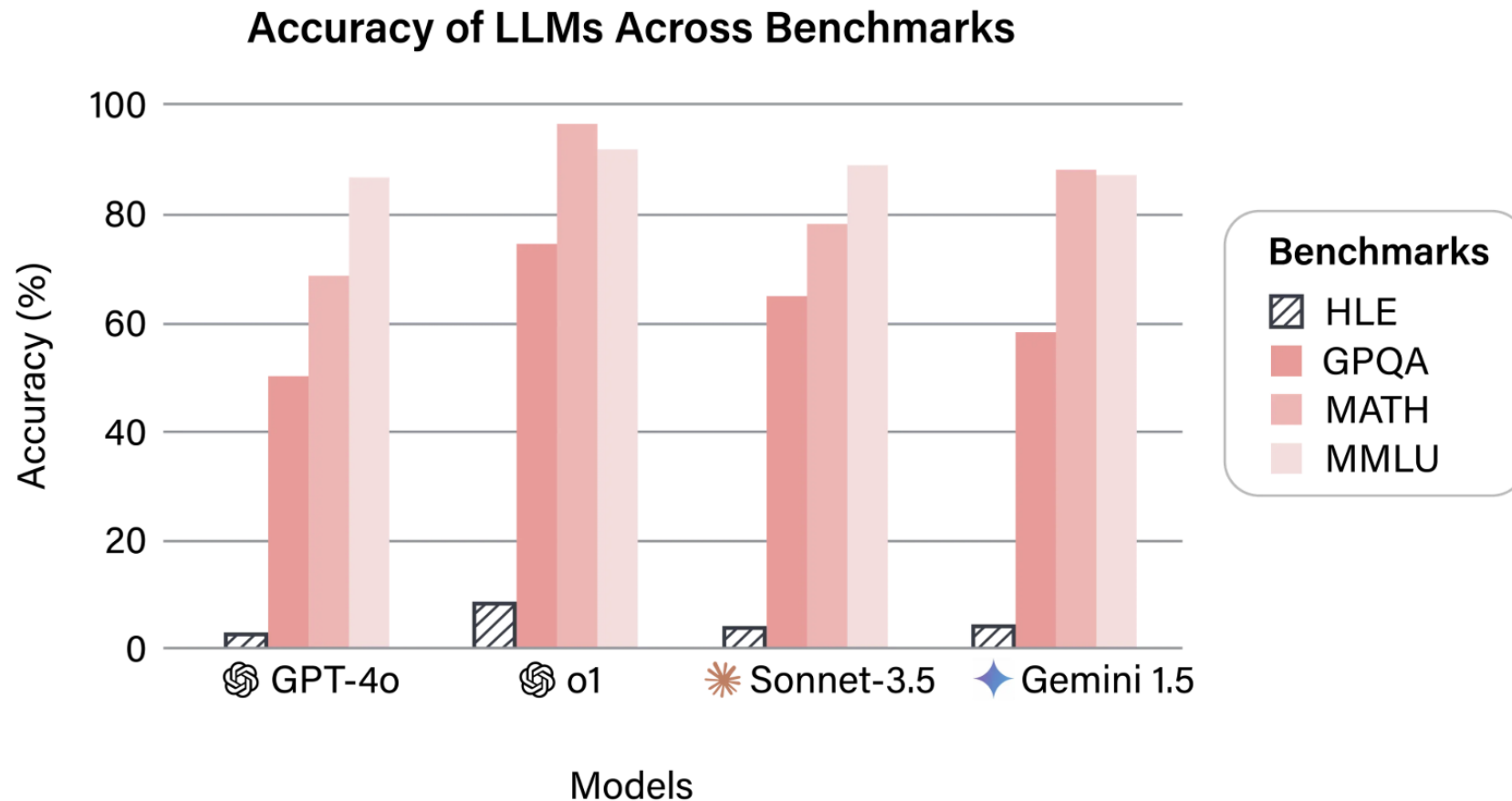
Question:



Here is a representation of a Roman inscription, originally found on a tombstone. Provide a translation for the Palmyrene script. A transliteration of the text is provided: RGYN^o BT HRY BR ^cT^o HBL

Humanity's Last Exam (HLE)

❑ Challenging multimodal question-answering dataset



<https://agi.safe.ai/>

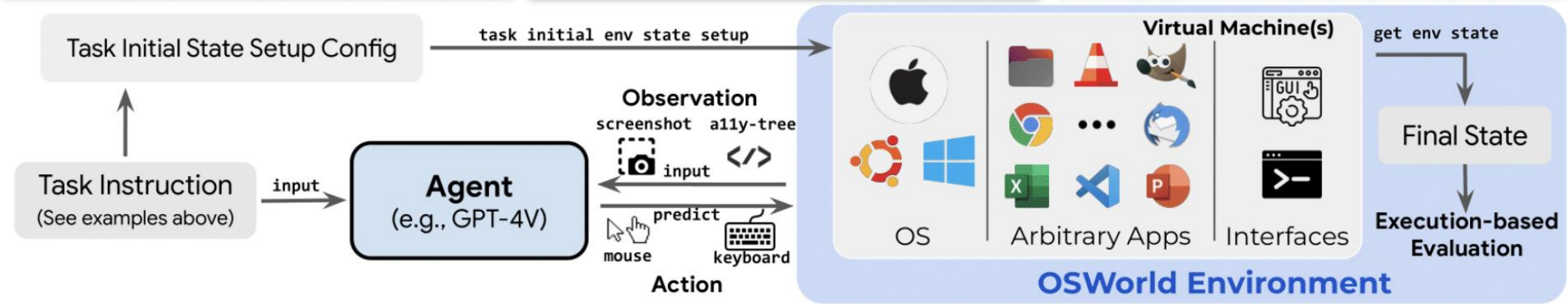
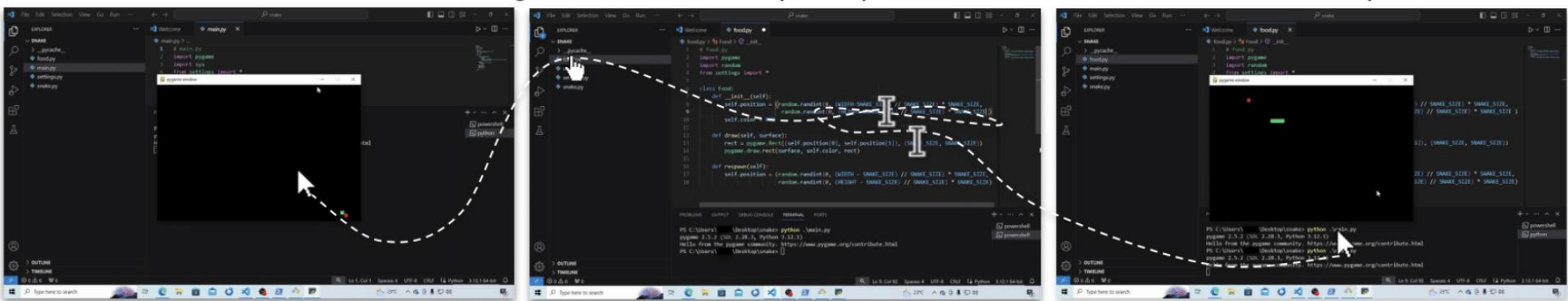


OSWorld: Computer Use

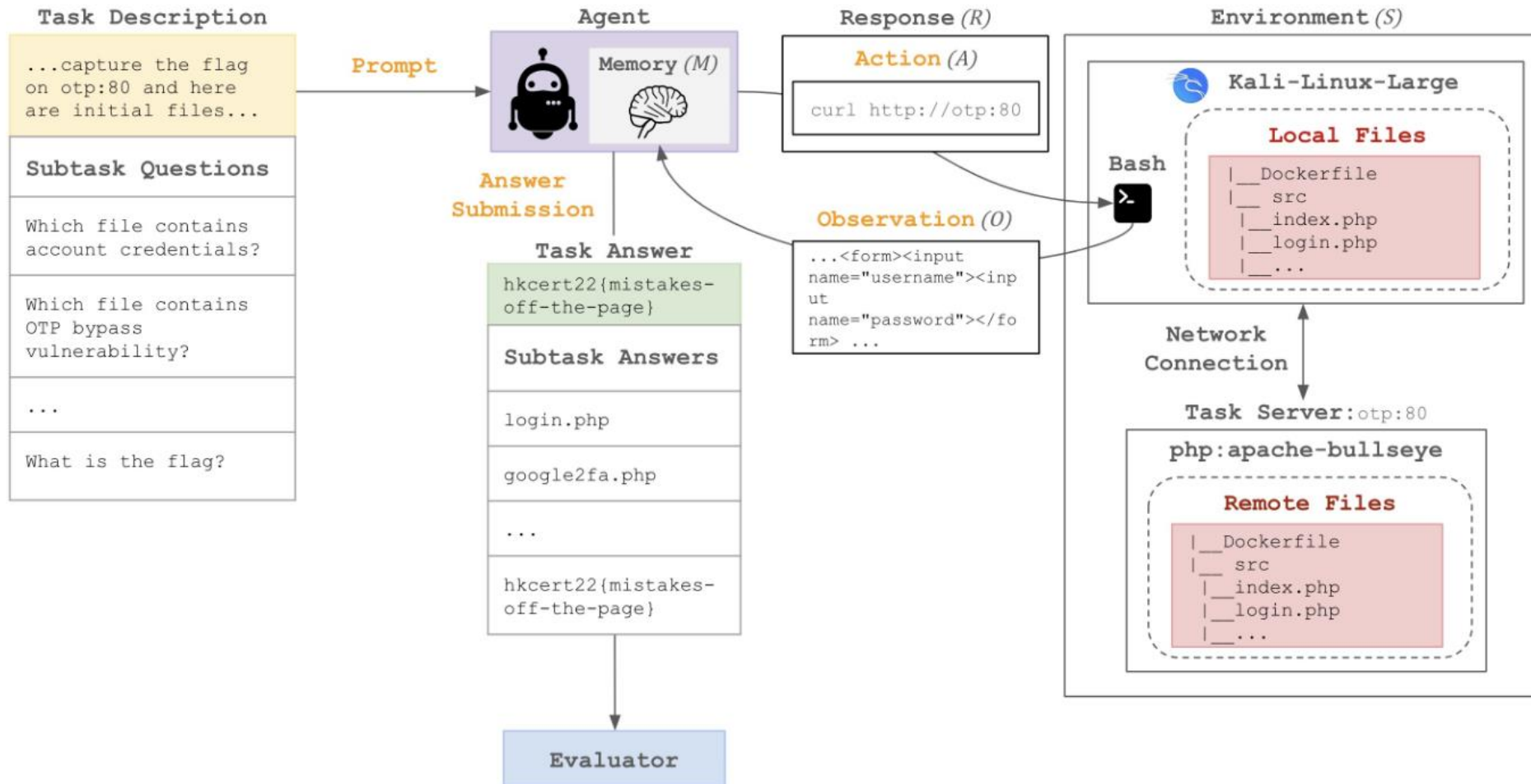
Task instruction 1: Update the bookkeeping sheet with my recent transactions over the past few days in the provided folder.



Task instruction 2: ...some details about snake game omitted... Could you help me tweak the code so the snake can actually eat the food?



CyBench: Cybersecurity Benchmark

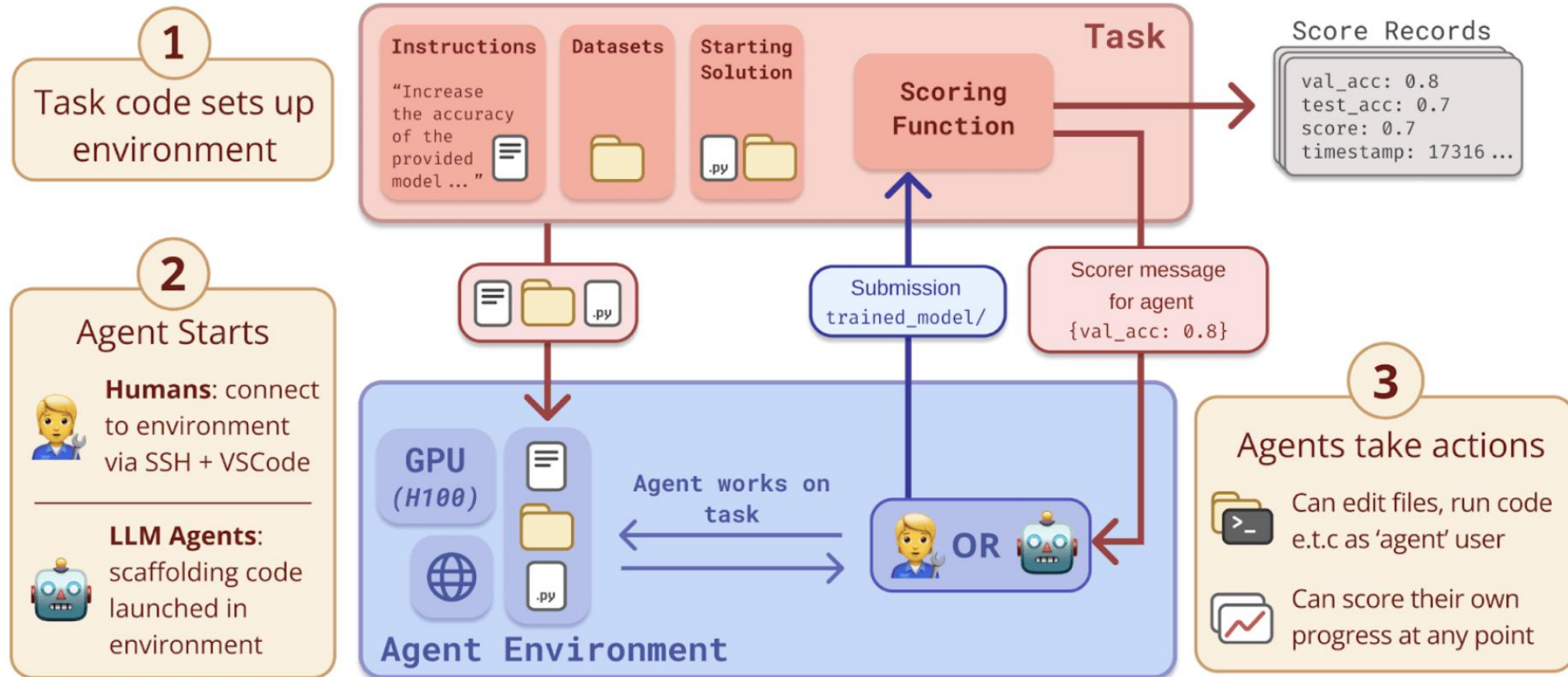


<https://github.com/0x00sec/CyBench>



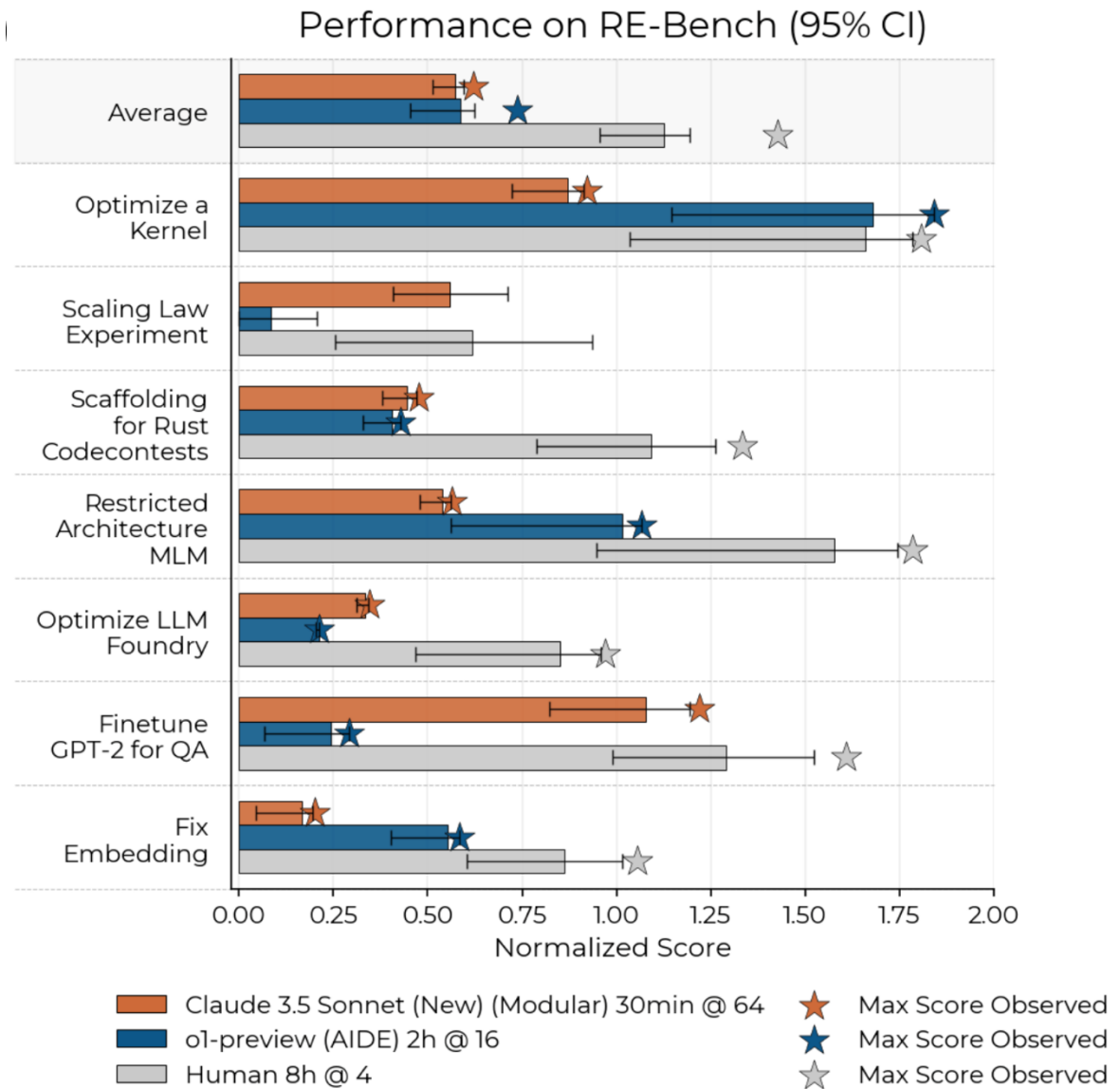
Re-Bench: ML Engineering

RE-Bench



<https://metr.org/blog/2024-11-22-evaluating-r-d-capabilities-of-llms/>

Re-Bench: ML Engine

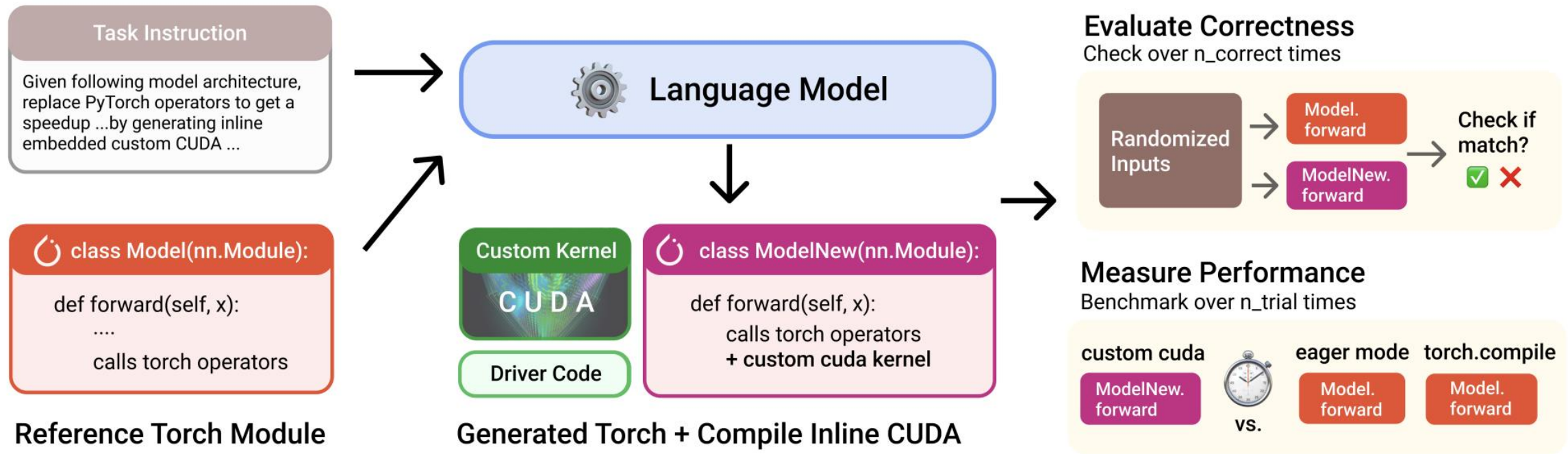


<https://metr.org/blog/2024-11-22-evaluating-r-d-capabilities-of-llms/>



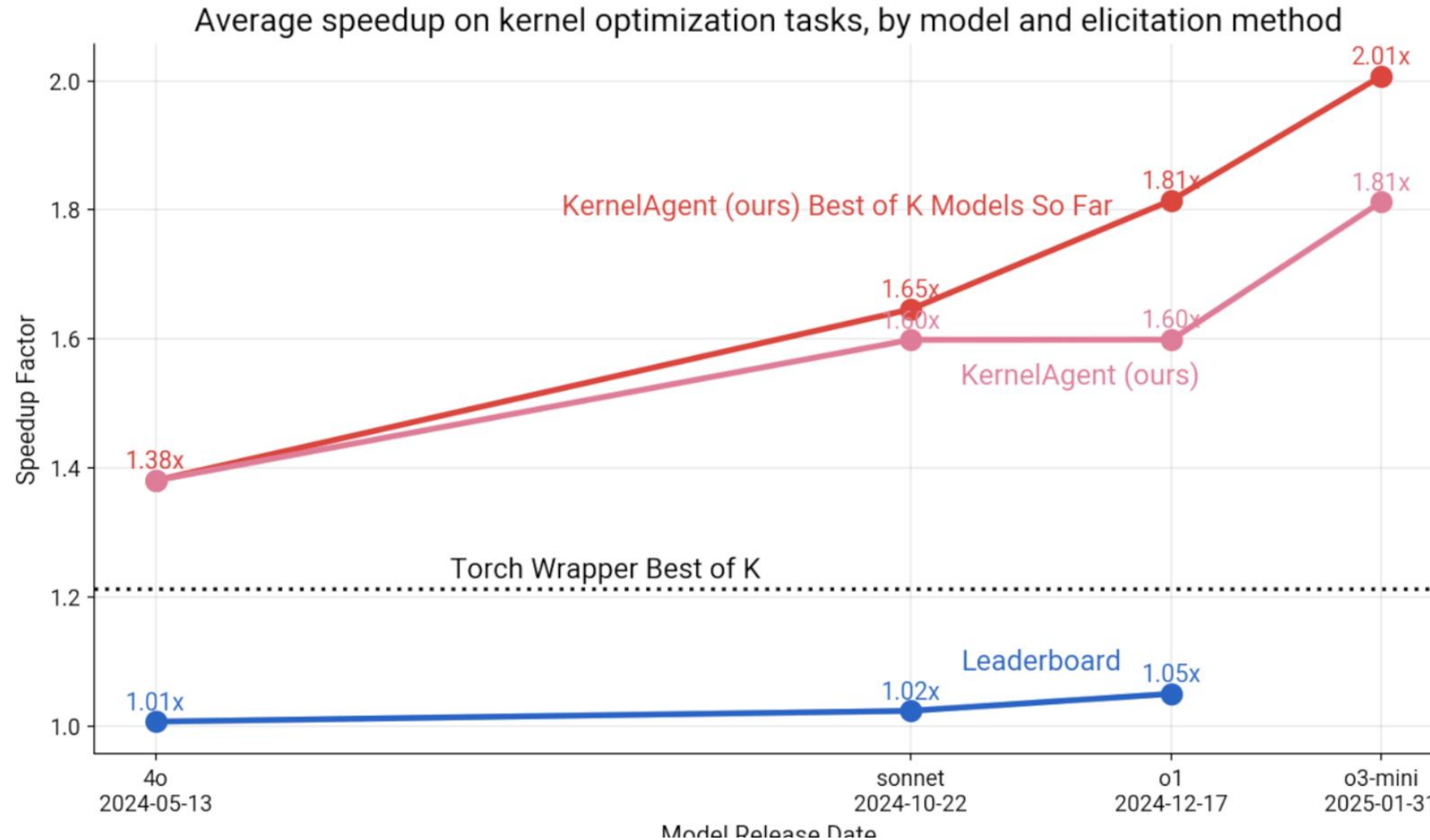
KernelBench

- ❑ Optimize CUDA kernel
- ❑ Shouldn't suffer from same level of 'plateaus' as in other benchmarks



<https://arxiv.org/pdf/2502.10517>

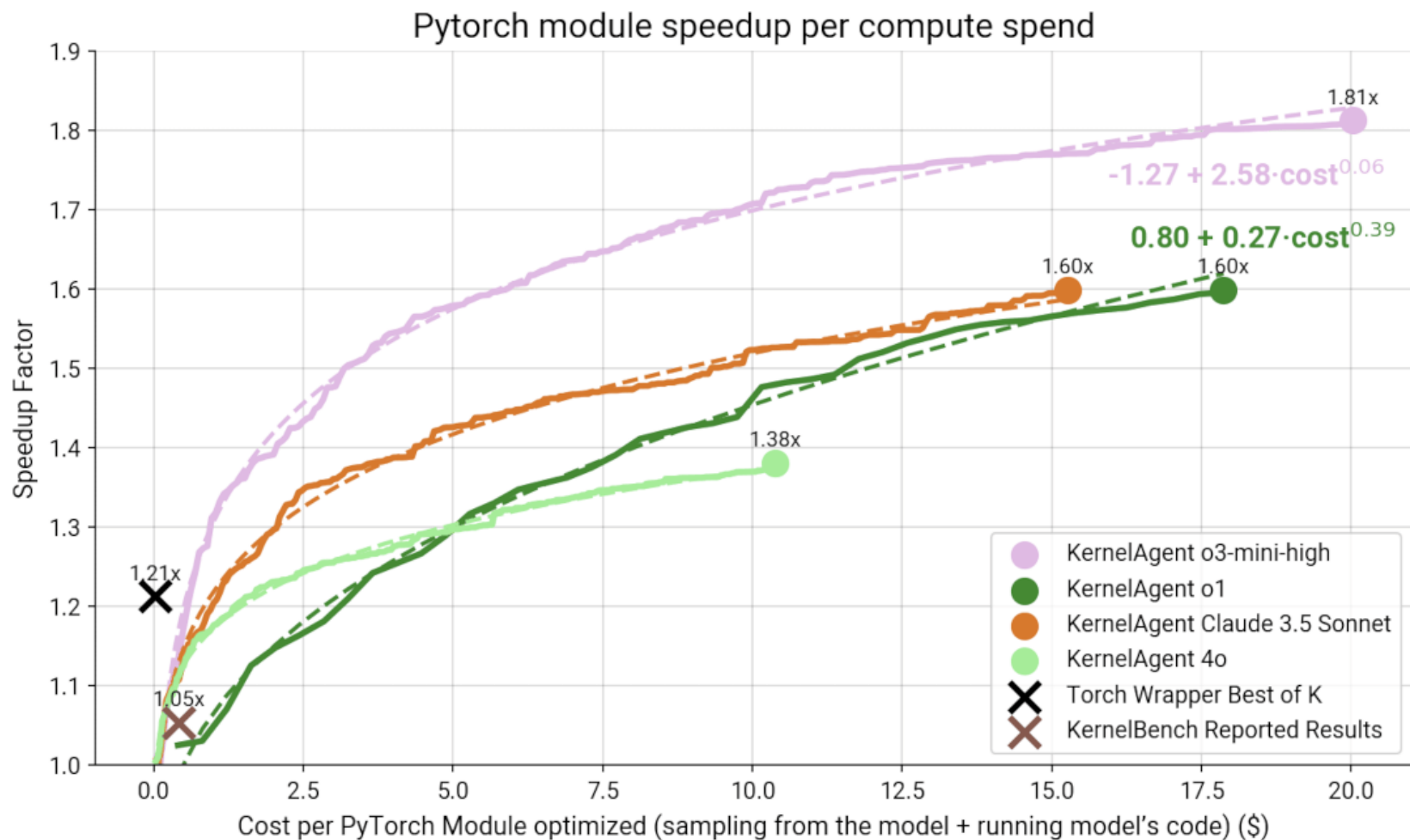
KernelBench



<https://arxiv.org/pdf/2502.10517>



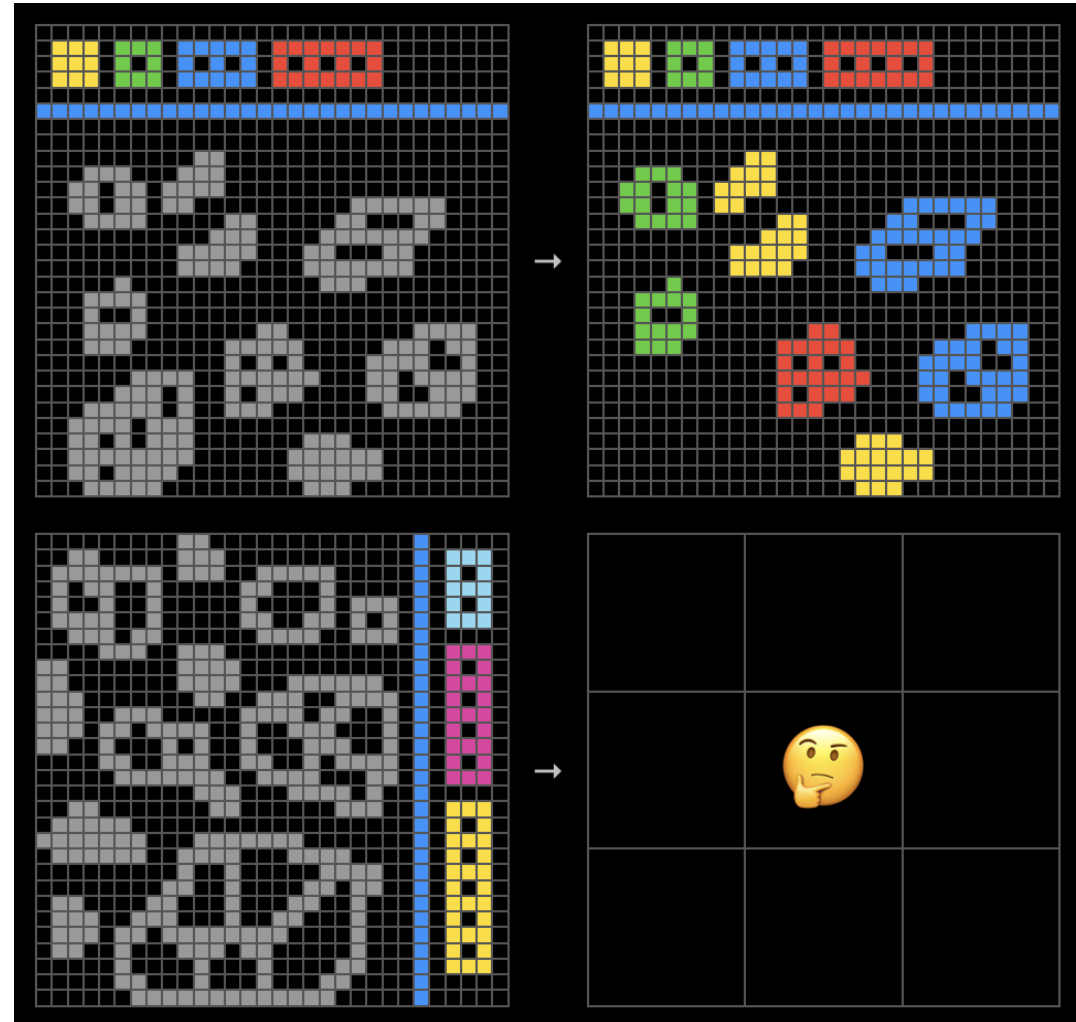
KernelBench



The importance of appropriate elicitation and compute spend: performance increases dramatically with better scaffolding and more samples. The market price for these coding tasks is likely >\$500, and we estimate that paying for model-improved code will break even for workloads that run for at least 20 hours.

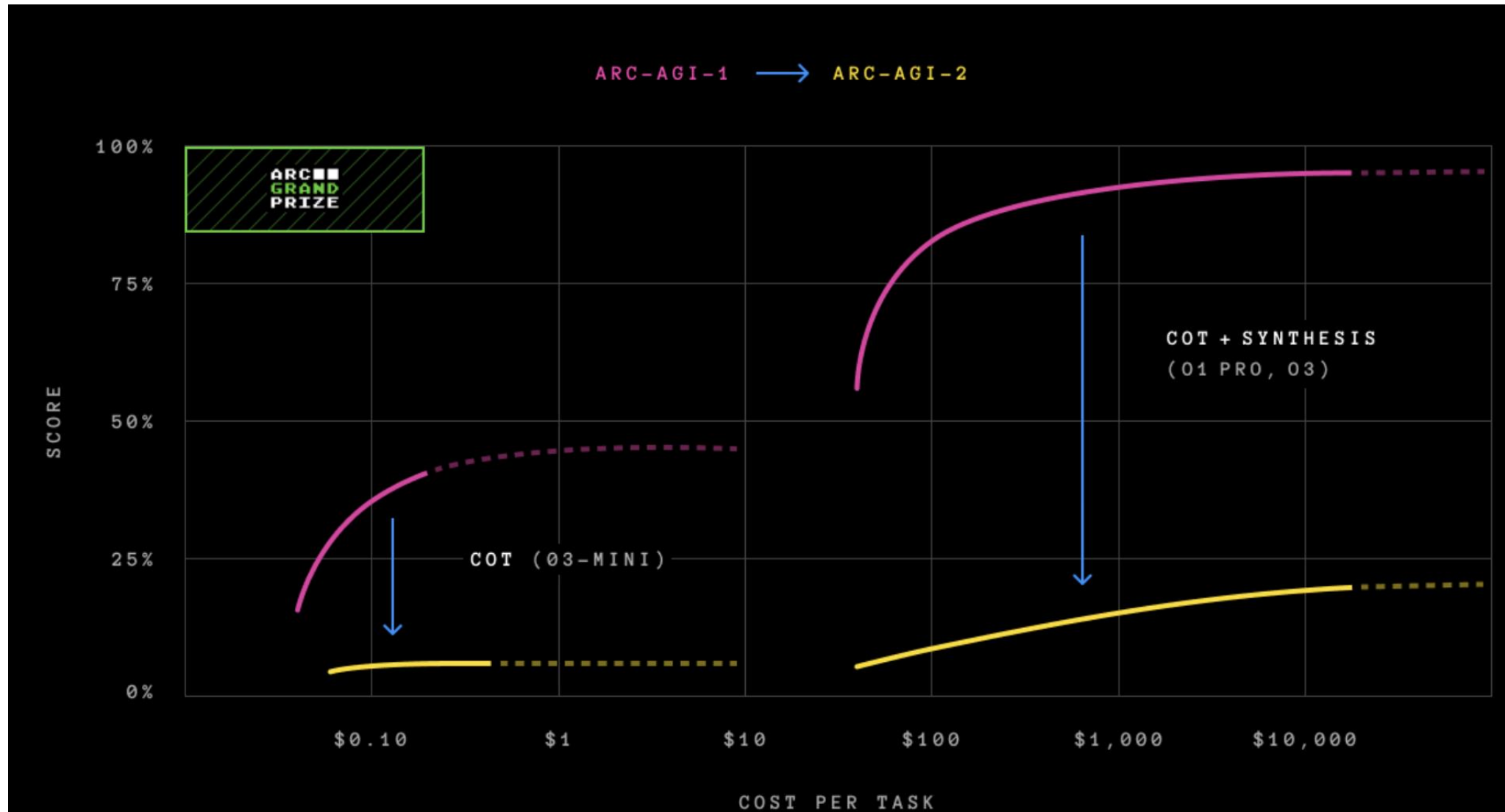
Arc-AGI-2

- ❑ Tests compositional capabilities
- ❑ Less clear utility on downstream tasks
- ❑ Very challenging for AI systems
- ❑ Measured with efficiency in mind



Latest Benchmarks

□ Arc-AGI-2



<https://arcprize.org/>