

CSCI 5541: Natural Language Processing

Lecture 14: All about Data, Annotation, and Evaluation of LLMs

Dongyeop Kang (DK), University of Minnesota

dongyeop@umn.edu | twitter.com/dongyeopkang | dykang.github.io



UNIVERSITY OF MINNESOTA
Driven to Discover®

Outline

- ❑ Annotation terms, examples, and process
- ❑ Qualitative coding
- ❑ Recruiting annotators (coders)
- ❑ Annotation quality assessment
- ❑ Annotation tools
- ❑ Issues in annotation
- ❑ Advanced annotation techniques
- ❑ LLMs as Annotators and Synthetic Data



Annotation

- ❑ Despite the emergent ability of LLMs, fine-tuned models trained on annotated dataset still shows better performance.
- ❑ High-quality data means high-performance algorithms
- ❑ Just providing large amounts of data doesn't help the model understand and learn to speak. The data needs to be **guided** in such a way that the computer can more easily find patterns and inferences.
- ❑ 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.



Datasets
5,659 machine learning datasets

Share your dataset with the ML community!

Search for datasets

Best match

- Filter by Modality (clear)
- Texts x
 - Images 1737
 - Videos 558
 - Audio 260
 - Medical 206
 - 3D 168

- Filter by Task
- Question Answering 213
 - Language Modelling 95
 - Reading Comprehension 71
 - Named Entity Recognition 59

1568 dataset results for Texts x

- GLUE** (General Language Understanding Evaluation benchmark)
General Language Understanding Evaluation (GLUE) benchmark is a collection of nine natural language understanding tasks, including single-sentence tasks CoLA and SST-...
1,258 PAPERS • 33 BENCHMARKS
- SQuAD** (Stanford Question Answering Dataset)
The Stanford Question Answering Dataset (SQuAD) is a collection of question-answer pairs derived from Wikipedia articles. In SQuAD, the correct answers of questions can...
1,257 PAPERS • 11 BENCHMARKS
- Penn Treebank**
The English Penn Treebank (PTB) corpus, and in particular the section of the corpus corresponding to the articles of Wall Street Journal (WSJ), is one of the most known and...
1,185 PAPERS • 14 BENCHMARKS
- SST** (Stanford Sentiment Treebank)
The Stanford Sentiment Treebank is a corpus with fully labeled parse trees that allows for a complete analysis of the compositional effects of sentiment in language. The cor...
1,086 PAPERS • 4 BENCHMARKS
- Visual Question Answering (VQA)**
Visual Question Answering (VQA) is a dataset containing open-ended questions about images. These questions require an understanding of vision, language and common-...
936 PAPERS • 2 BENCHMARKS
- IMDb Movie Reviews**
The IMDb Movie Reviews dataset is a binary sentiment analysis dataset consisting of 50,000 reviews from the Internet Movie Database (IMDb) labeled as positive or nega-...
924 PAPERS • 7 BENCHMARKS

Hugging Face Search models, datasets, users...

Datasets 27,579 Filter by: new Full-text search Add filters Sort: Most Downloads

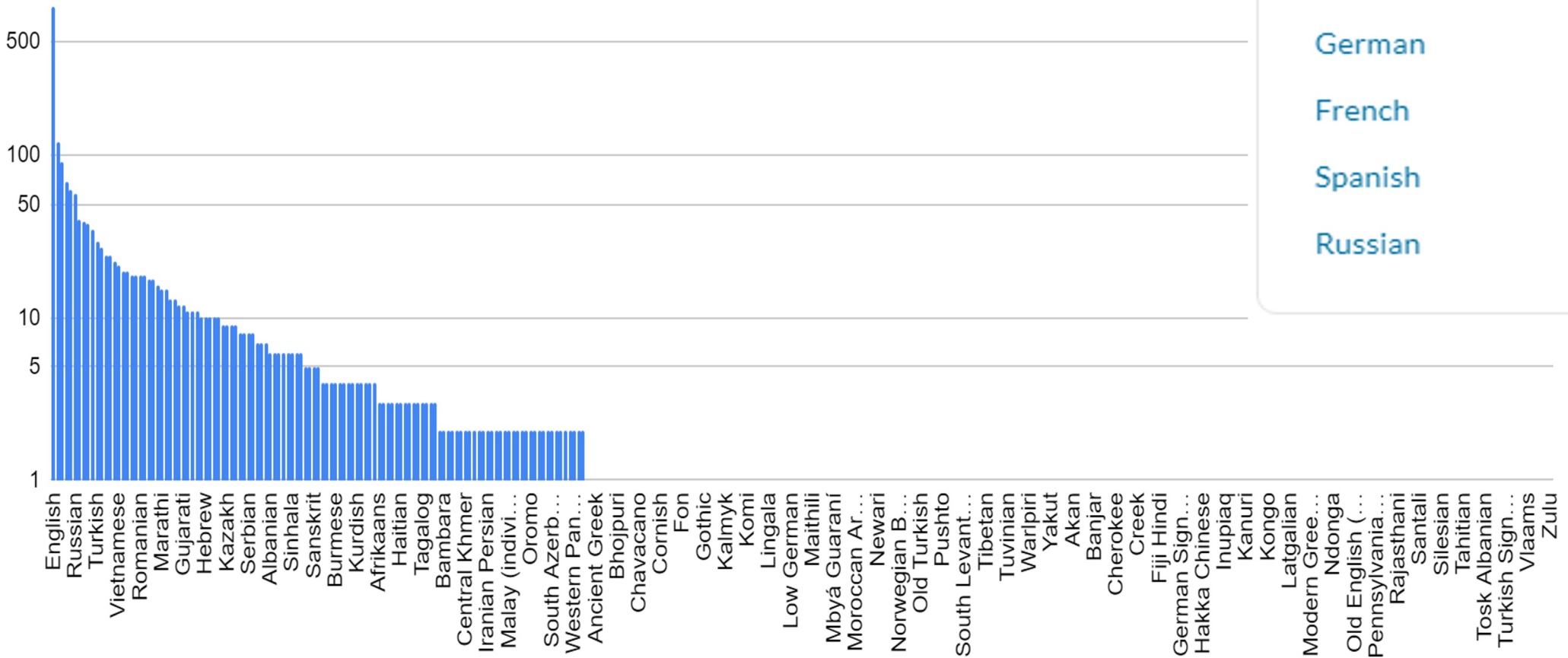
- glue**
Preview • Updated about 8 hours ago • ↓ 1.1M • ♥ 134
- super_glue**
Preview • Updated about 5 hours ago • ↓ 1.02M • ♥ 77
- allenai/nllb**
Preview • Updated Sep 29, 2022 • ↓ 797k • ♥ 24
- argilla/news-summary**
Preview • Updated 20 days ago • ↓ 765k • ♥ 20
- openwebtext**
Preview • Updated about 5 hours ago • ↓ 554k • ♥ 92
- bigscience/P3**
Preview • Updated Feb 1 • ↓ 532k • ♥ 89
- wikitext**
Preview • Updated about 5 hours ago • ↓ 350k • ♥ 99

<https://paperswithcode.com/datasets>

Current benchmark datasets are skewed to high-resource languages

Filter by Language

English	828
Chinese	122
German	91
French	69
Spanish	62
Russian	58

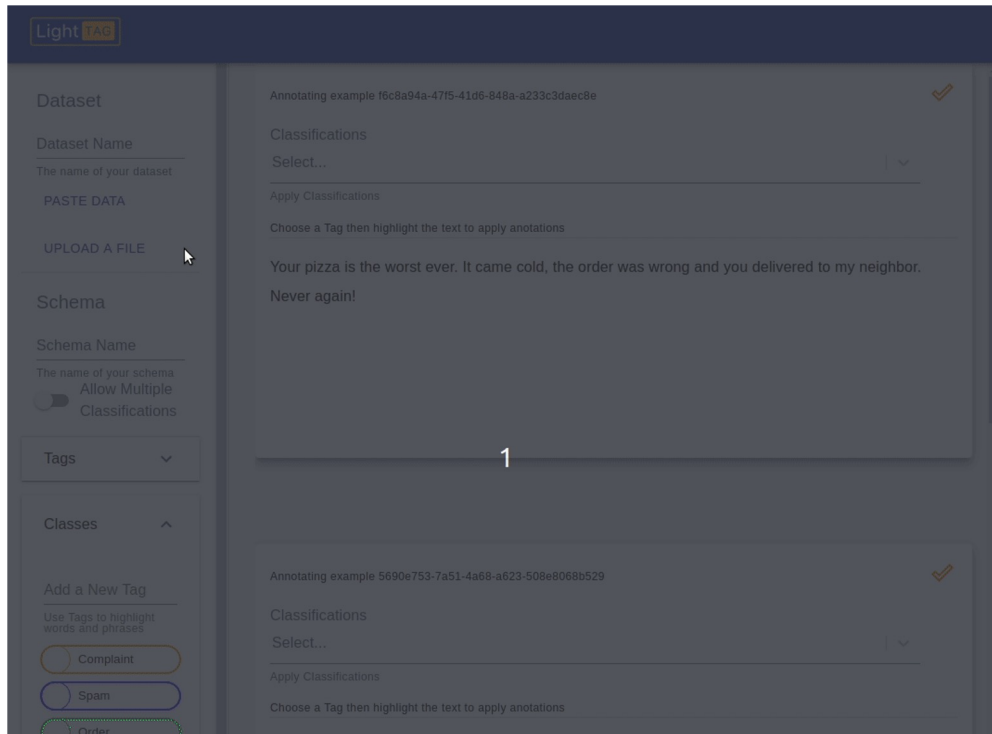


Terms

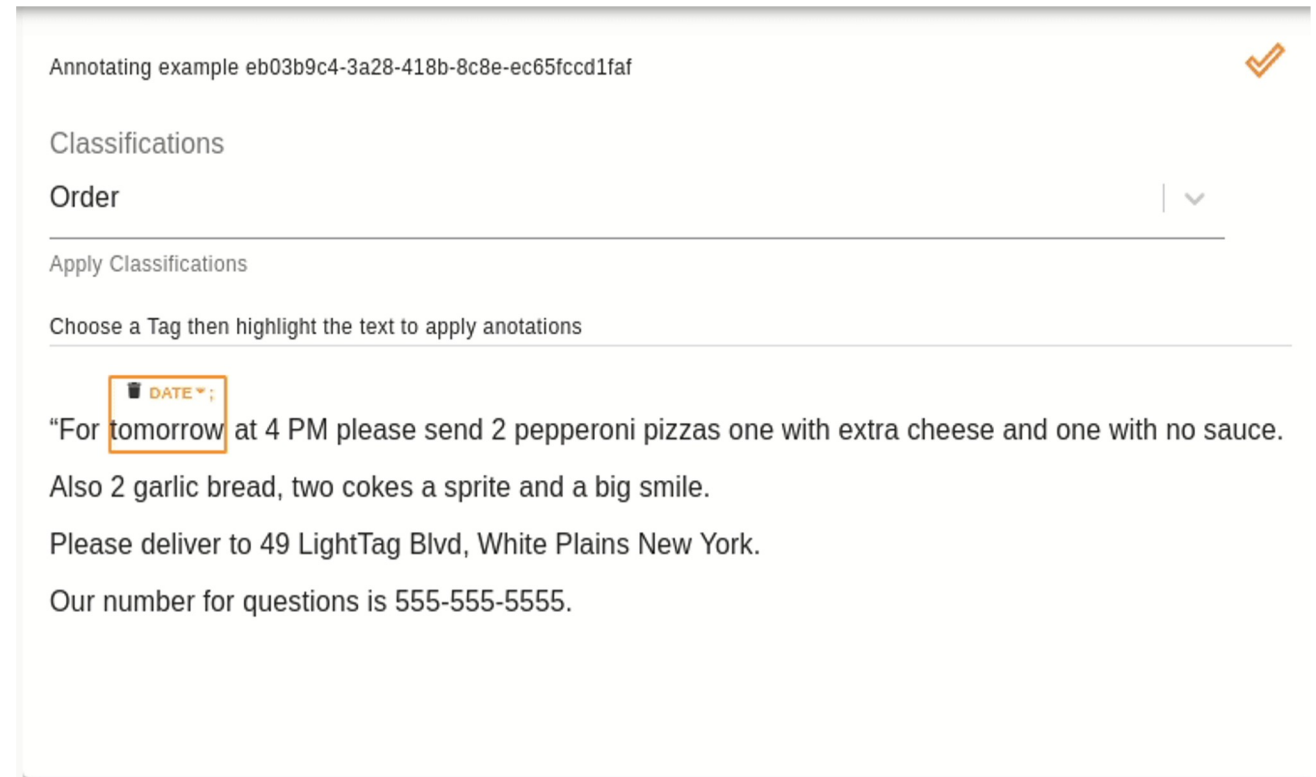
- ❑ Datasets of natural language are referred to as **corpora**
- ❑ A single set of data annotated with the same specification is called an **annotated corpus**.
- ❑ A **dataset** is a collection of examples that need to be annotated.
 - ❑ A **class** is a particular classification option.
 - E.g., Positive 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 **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

Showing examples 1 to 6 out of 6

DRAG ENTITIES HERE TO FO

Big interview tonight by Henry Kravis at The Business Council of Washington. Looking forward to it!

The FBI is totally unable to stop the national security "leakers" that have permeated our government for a long time. They can't even.....

Relation annotation

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 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 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



Questions for collecting the ideal dataset?

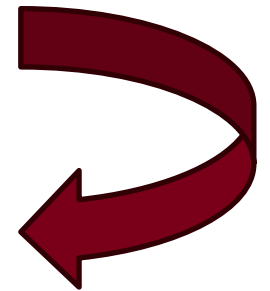
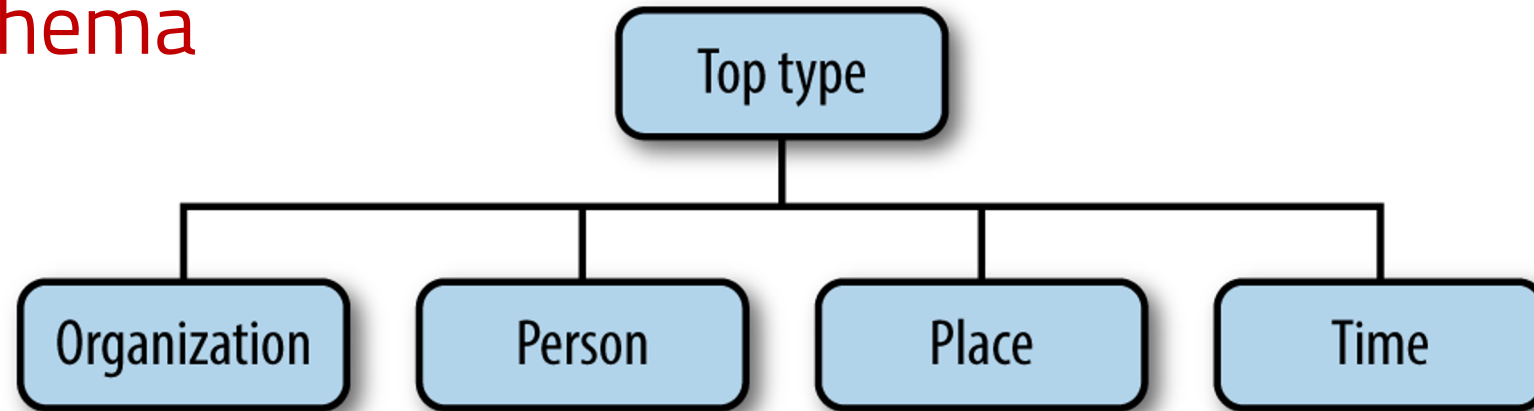
- ❑ What is the **target accuracy** you are looking for?
- ❑ Can it be achieved it by **better models or more data**?
 - How many annotations are enough to ensure high accuracies?
- ❑ How **representative** is your dataset?
 - domain vocabulary, format, genre of the text, etc
- ❑ Is your dataset **balanced**,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



Examples on semantic types/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

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

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}.

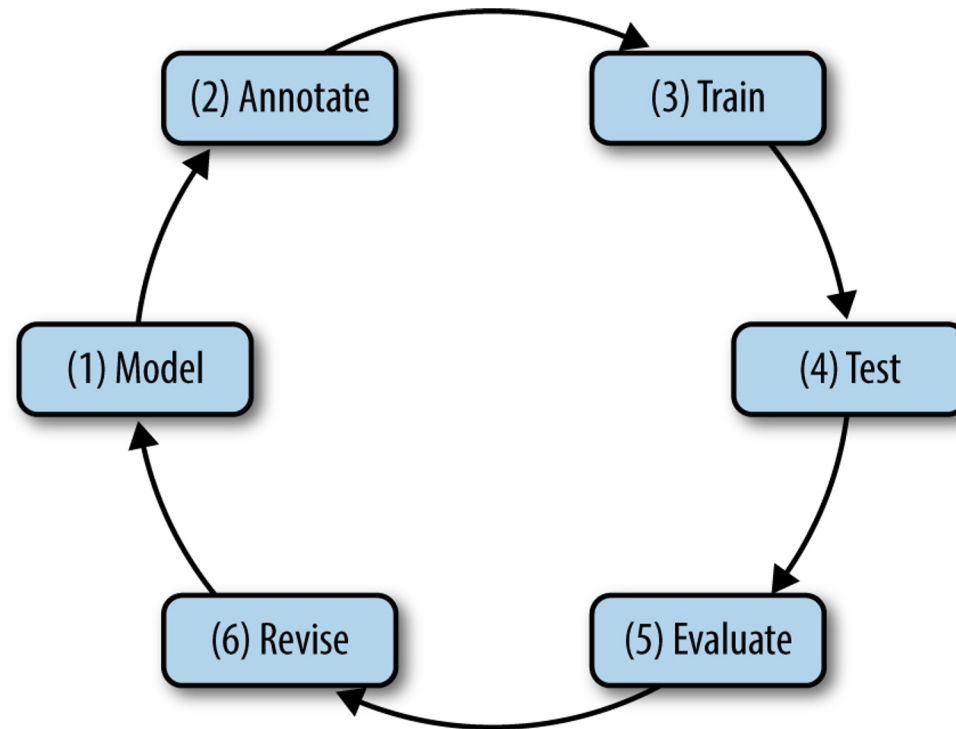
[My brother]_{theme} lives in [Milwaukee]_{location}.



Annotation process



Annotation Development Cycle

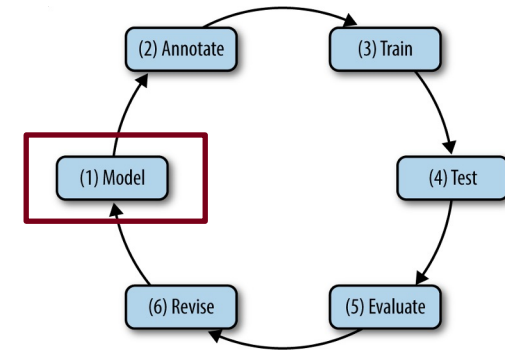


MATTER methodology ([Pustejovsky 2006](#))

Model the Phenomenon

A 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!"}



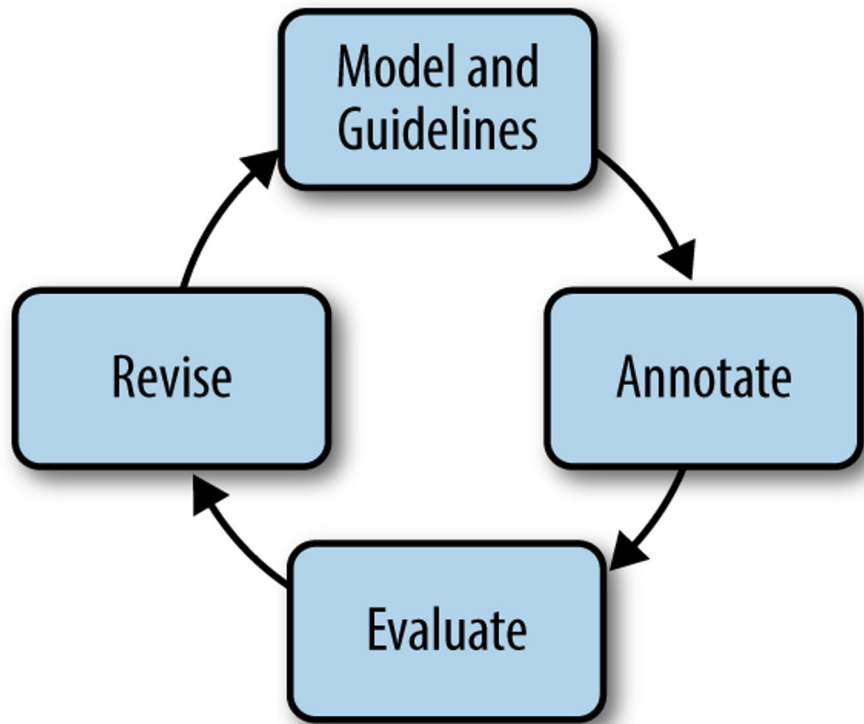
Terms = {Named_Entity, Organization, Person, Place, Time}

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

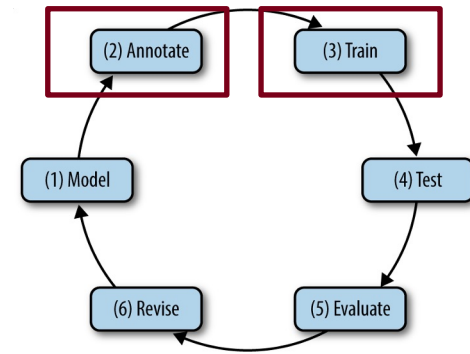
Interpretation = { Organization = "list of organizations in a database",
Person = "list of people in a database",
Place = "list of countries, geographic locations, etc.",
Time = "all possible dates on the calendar"}



Annotate with the Specification



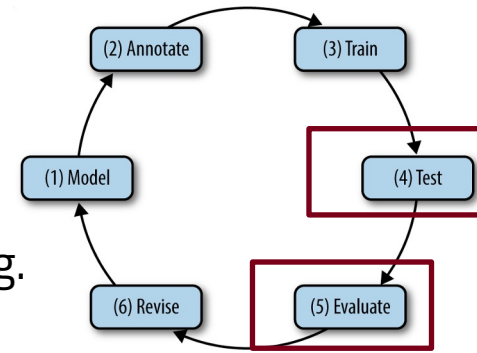
MAMA (Model-Annotate-Model-Annotate) cycle, or the “babeling” phase of MATTER.



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.

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



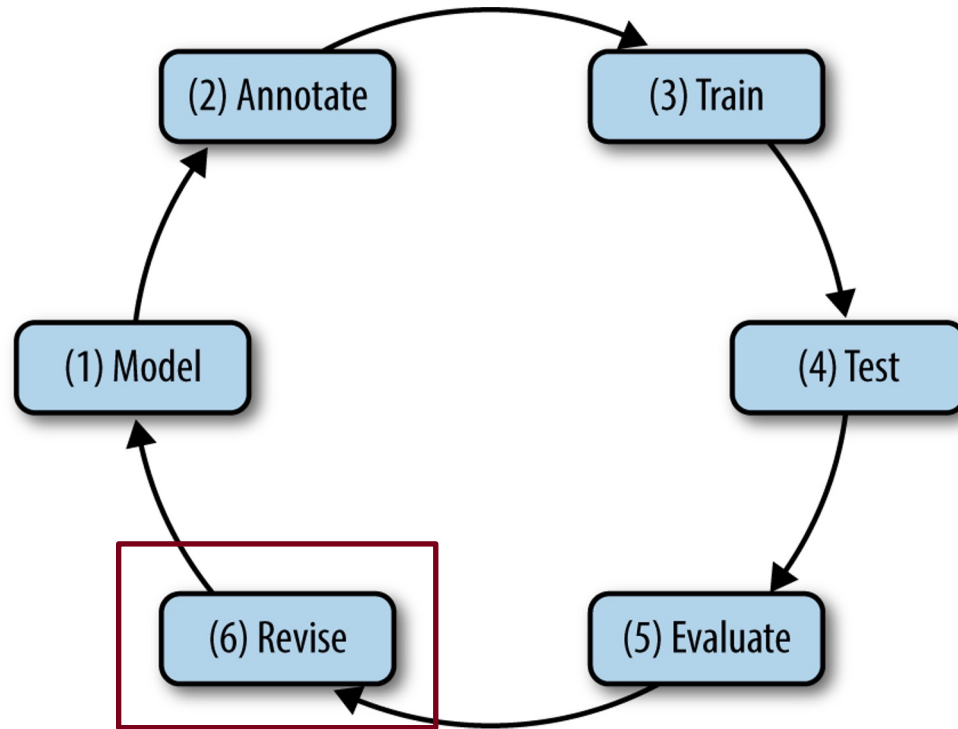
[QBC Productions Inc.]_{Organization} of East Anglia



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



Annotation Development Cycle



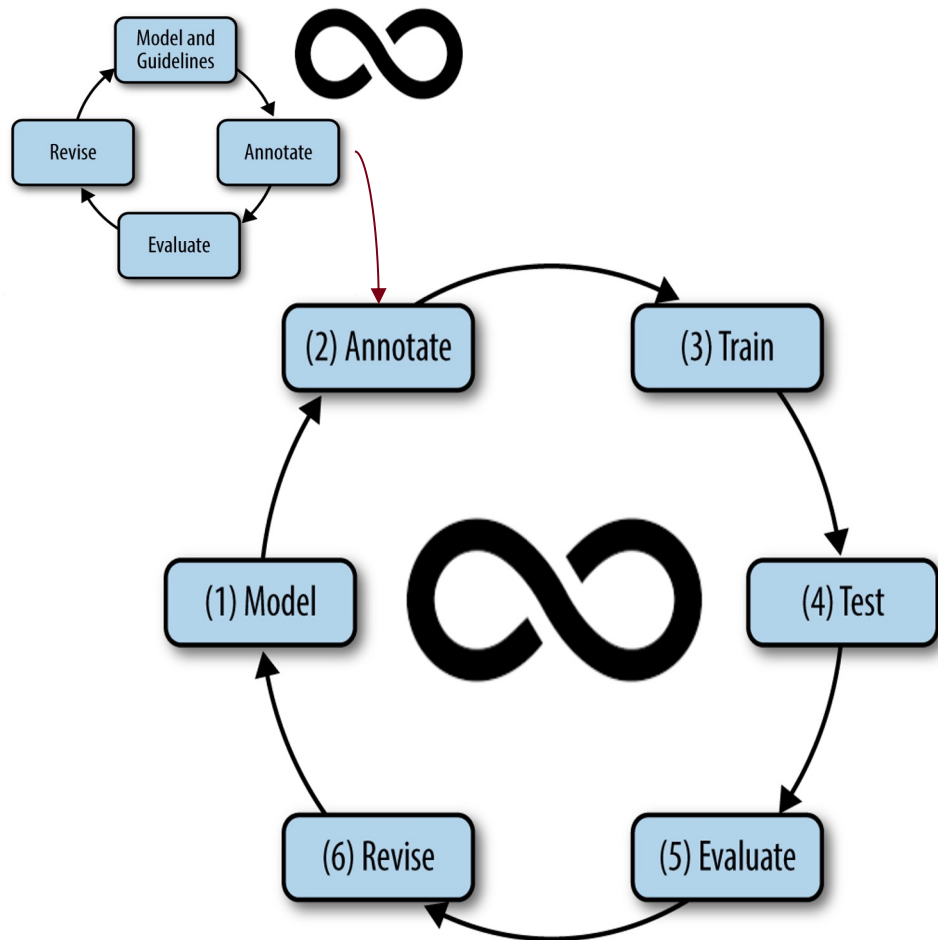
Revise

The model and the annotation specification are revisited in order to make the annotation more **robust and reliable with use in the algorithm.**

MATTER methodology (Pustejovsky 2006)



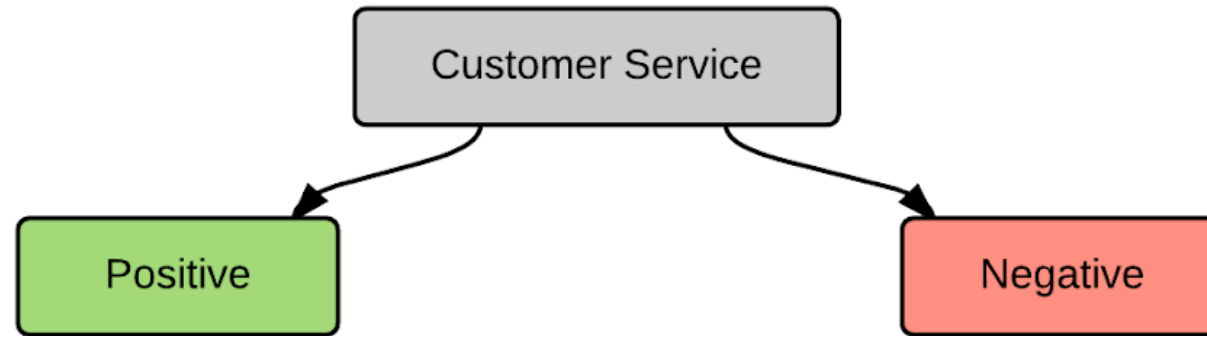
In Practice

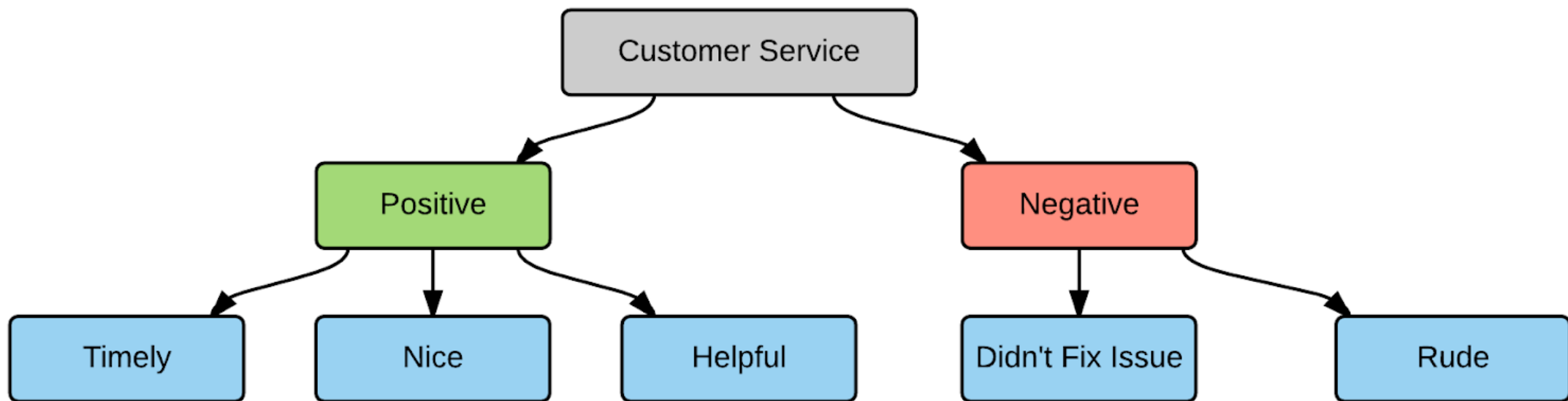


- ❑ An **iterative** process until you reach to the target performance
- ❑ As model performance converges, you will face **edge cases** in the long tail. Analyzing the long-tail and updating the schema are painful and time-consuming, but most important in practice.
- ❑ There is **no single magic deep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

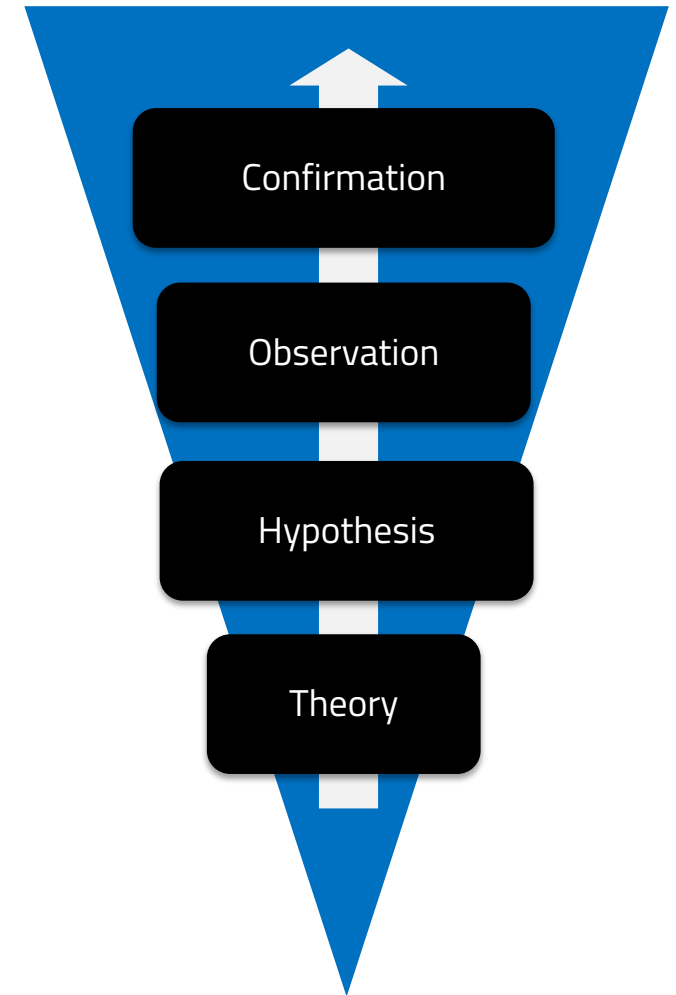
Qualitative coding



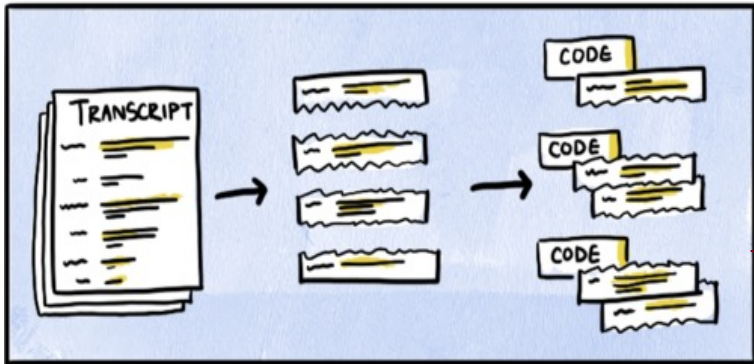
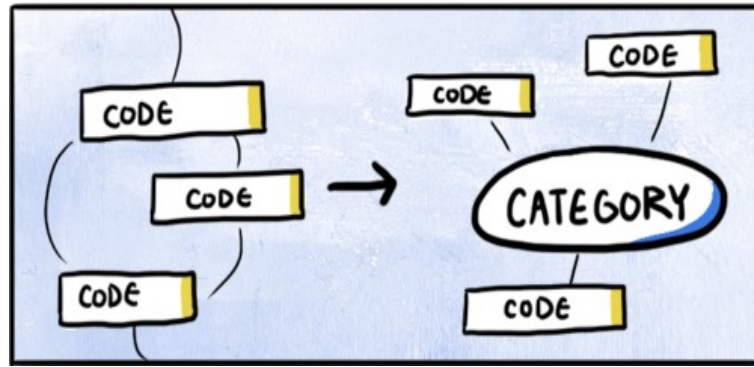
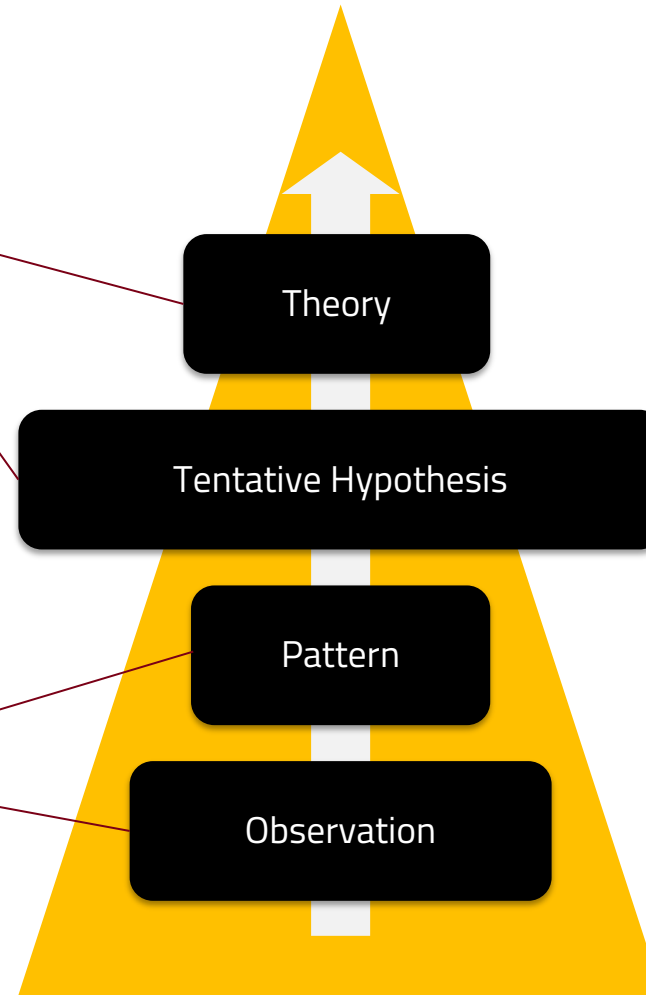




Deductive coding



Inductive coding



Steps in inductive coding

Open coding:

Compare snippets with snippets and create codes that connect them.



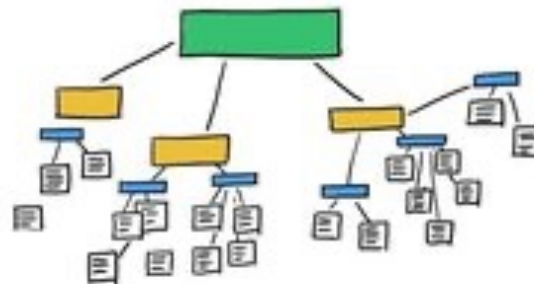
Axial coding:

Compare codes with codes and create categories (or axes) that connect them.



Selective coding:

Compare categories with categories and create the core category that connect them.



Human-AI Collaborative Taxonomy Construction

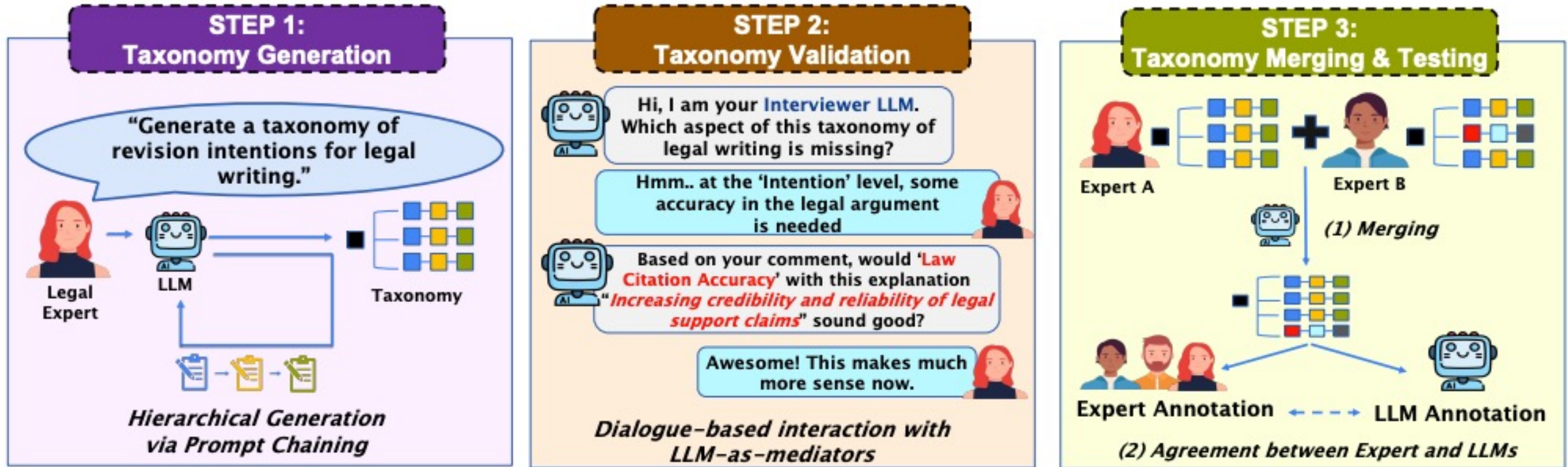
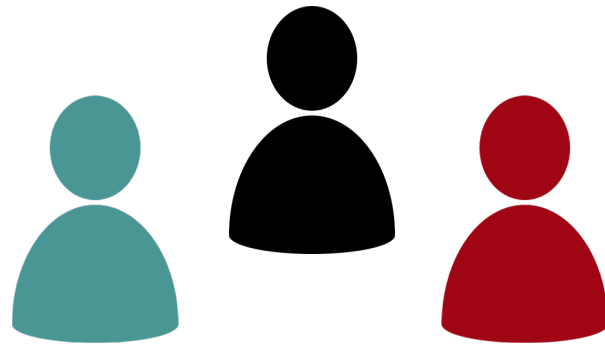


Figure 1: An end-to-end pipeline of our three-step Human-AI collaborative taxonomy construction process. For each step, we portray several design implications for better human-AI interaction strategies that were described in Section 3.



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

The screenshot shows the Prolific research platform interface. At the top, it says "Prolific" and "Demographics". The main heading is "Which energy-saving action saves most energy?". Below this, there are four summary cards: "5 days ago Created", "£8.34/hr Average reward per hour", "17,907 of 37,410 Eligible Participants", and "21/21 Submissions Progress". A table below shows a list of participants with columns for PARTICIPANT ID, STARTED, TIME TAKEN, STUDY CODE, STATUS, and BONUS. The table contains 7 rows of data, all with a status of "APPROVED".

Prolific

Quickly find research participants you can trust.

The screenshot shows the Upwork freelance marketplace interface. At the top, it says "Upwork" and "Find Jobs". Below this, there is a search filter for "Top iPhone App Developers in Russia" with a filter set to "\$60/hr and above". Two developer profiles are shown: "Stanislav V. iOS Mobile App Developer" and "Azamat V. iOS Developer". Both profiles show a "TOP RATED" badge, a rate of "\$60/hr", and a location of "Russia". They also list skills like "iPhone App Development", "Android App Development", "Objective-C", and "iOS Development".

UpWork

Best for finding the right freelancers to complete tasks

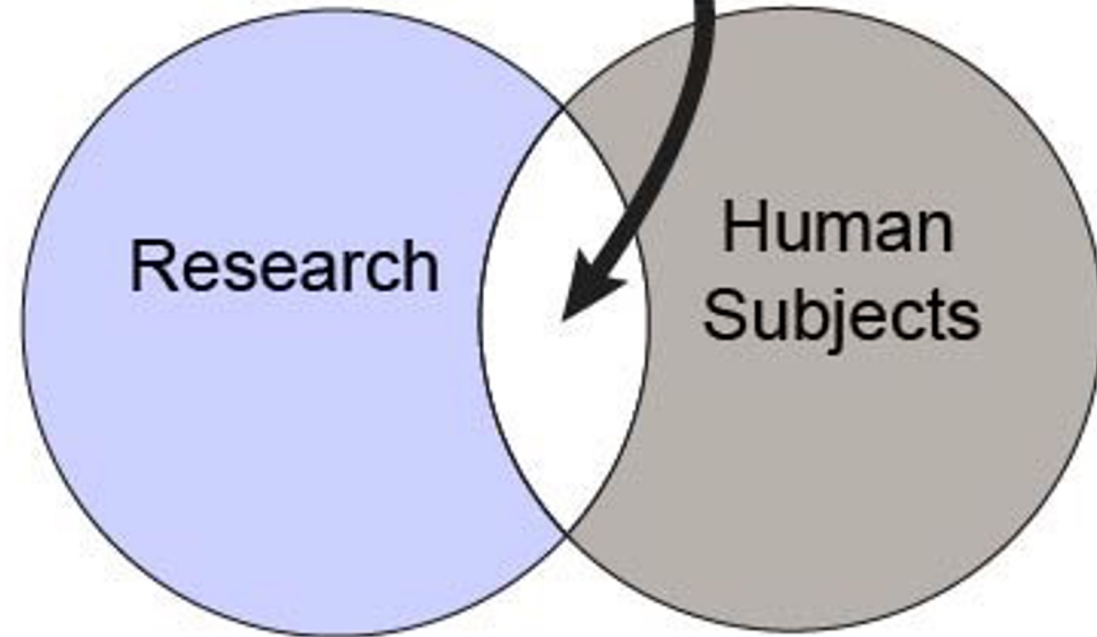


Undergraduate students



IRB Oversight

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 two months to get approval
- Before approval, you can't collect any human-subject data in your project

Annotation quality assessment



Correctness of annotations

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	
		Yes	No
Rater 1	Yes	25	10
	No	15	20

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

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

$$p_e = 0.285714 + 0.214285 = 0.5$$



Step 3: Calculate Cohen's Kappa

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

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

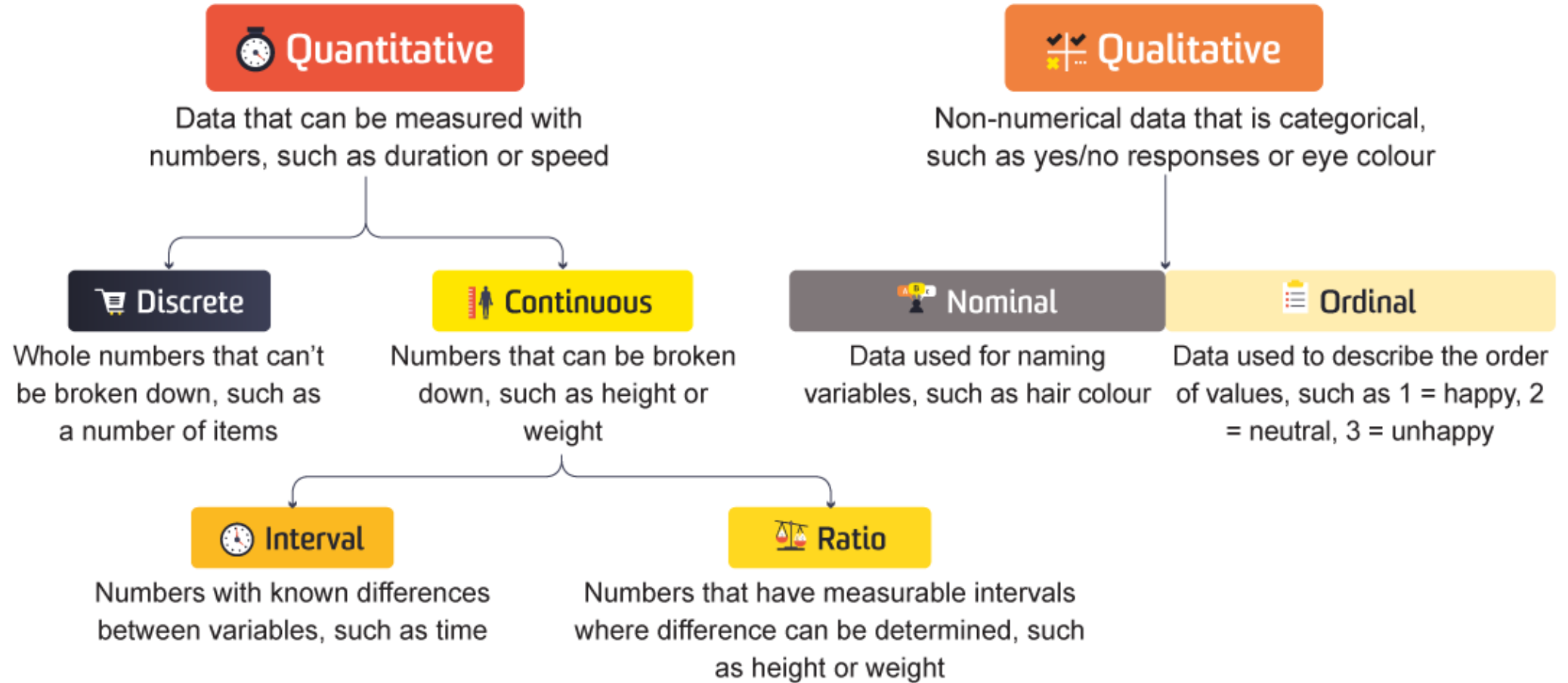


Interpretation of Cohen's Kappa

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.

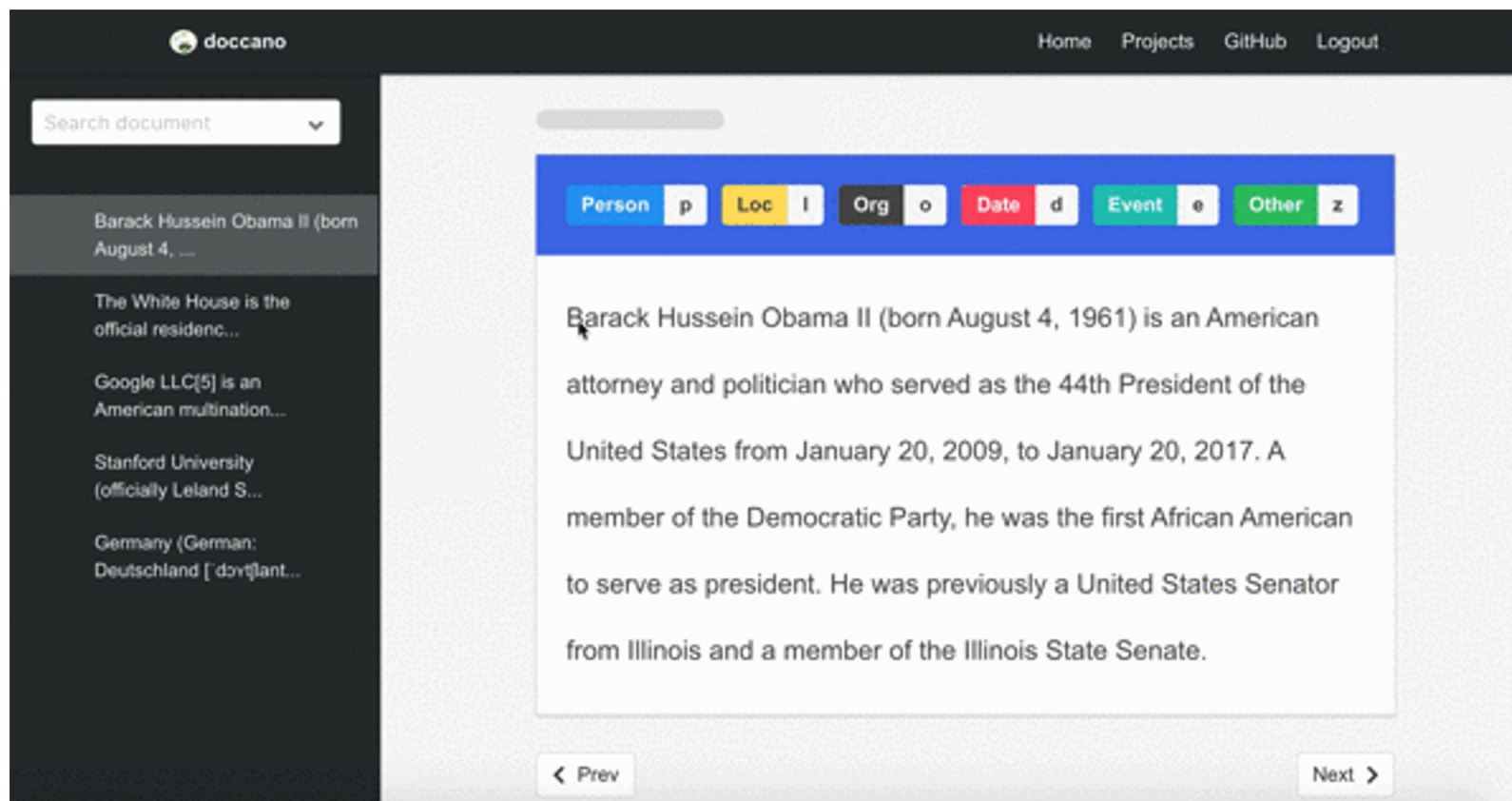
- **Landis and Koch (1977)** 0.6-0.79 substantial; 0.8+ perfect
- **Krippendorff (1980)** 0.67-0.79 tentative; 0.8+ good
- **Green (1997)** 0.4-0.74 fair/good; 0.75 high



Annotation tools



Doccano



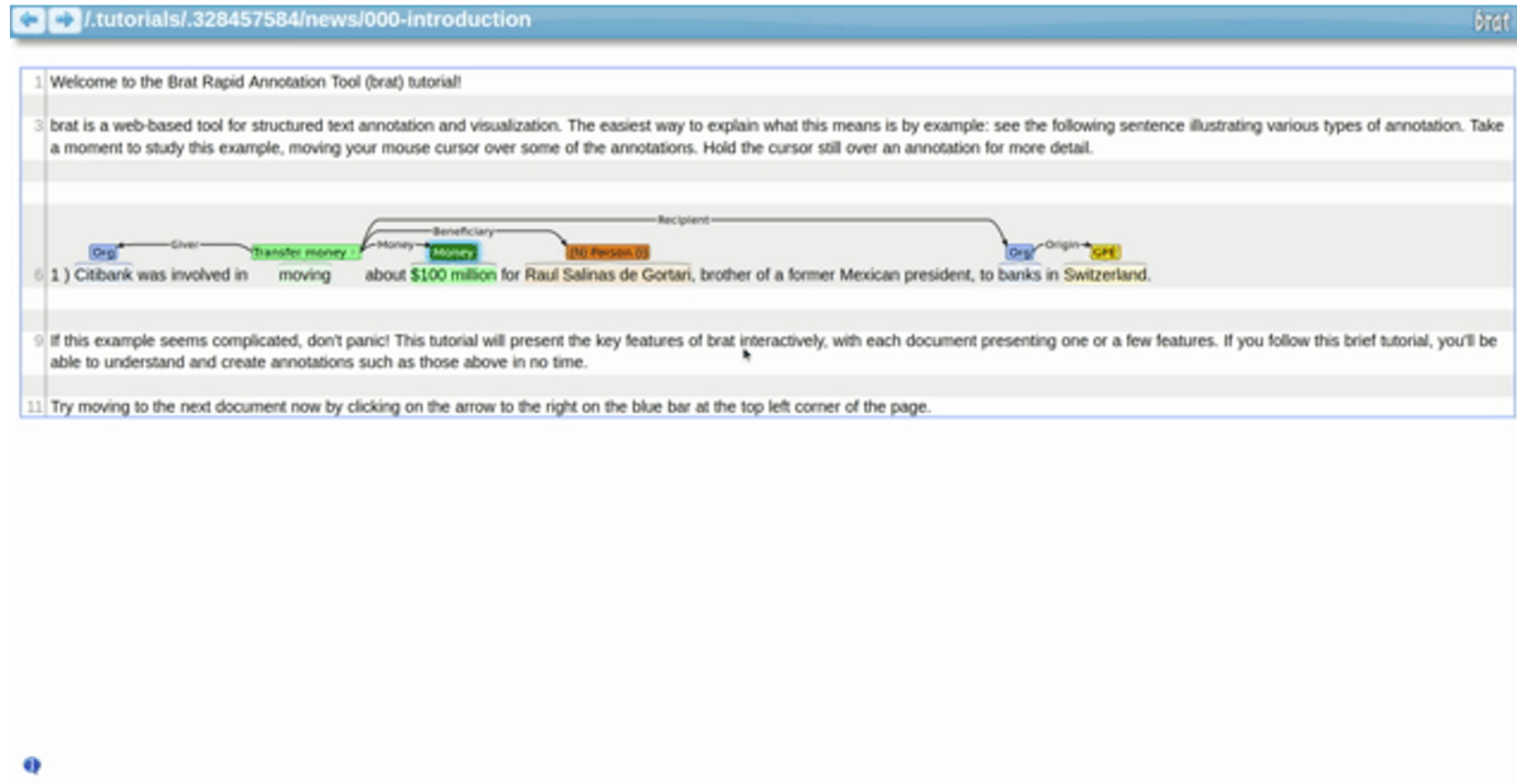
Pros:
Easy to use
Support Teams
Open Source

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 and annotations:

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 annotations in the sentence are as follows:

- `Org` (blue box) on "Citibank"
- `Givee` (black line) from "Citibank" to "moving"
- `Transfer money` (green box) on "moving"
- `Money` (green box) on "\$100 million"
- `Beneficiary` (black line) from "\$100 million" to "Raul Salinas de Gortari"
- `Recipient` (black line) from "Raul Salinas de Gortari" to "banks in Switzerland"
- `Org` (blue box) on "banks in Switzerland"
- `Origin` (black line) from "banks in Switzerland" to "Switzerland"
- `City` (yellow box) on "Switzerland"

Pros:
Open source
Free

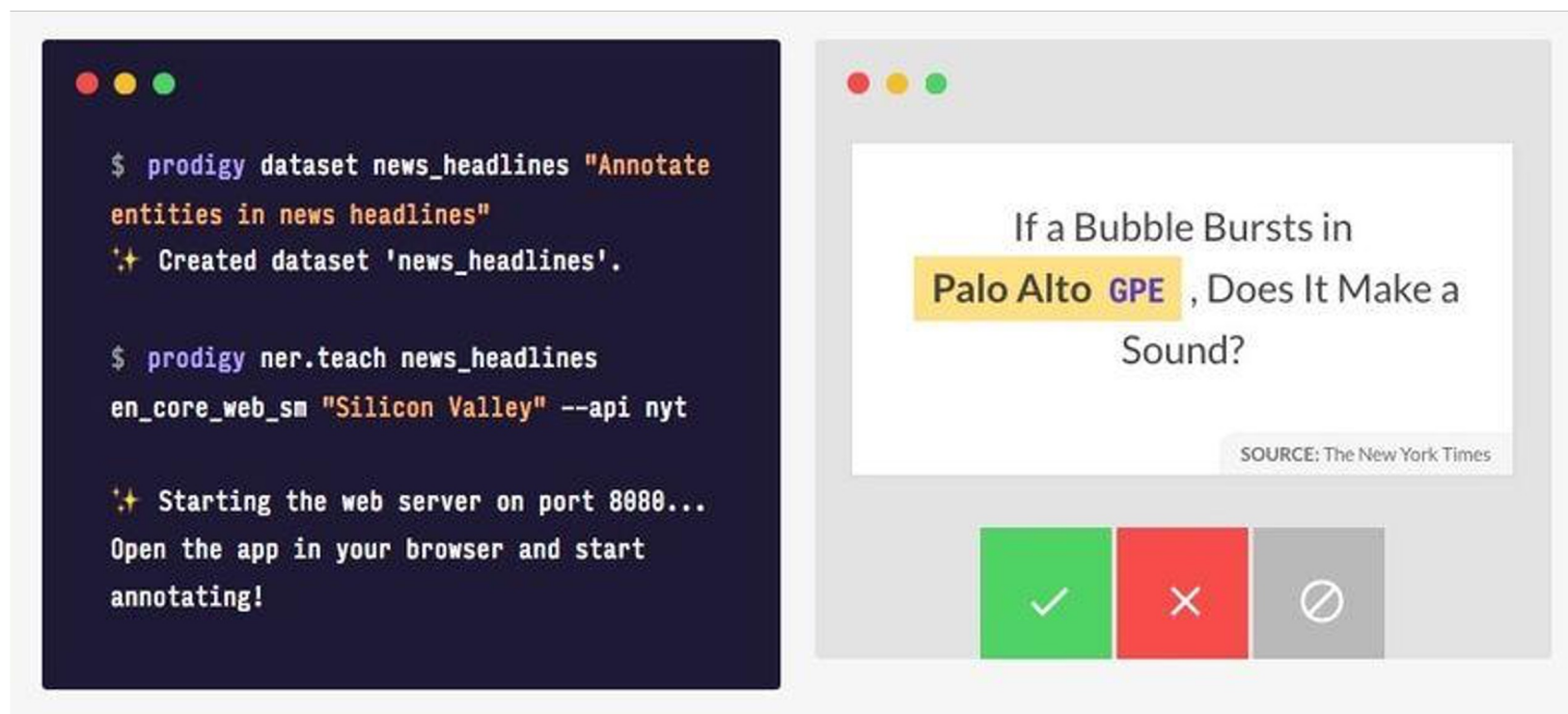
Cons:
Old-fashioned UI



Prodigy

<https://prodi.gy/>

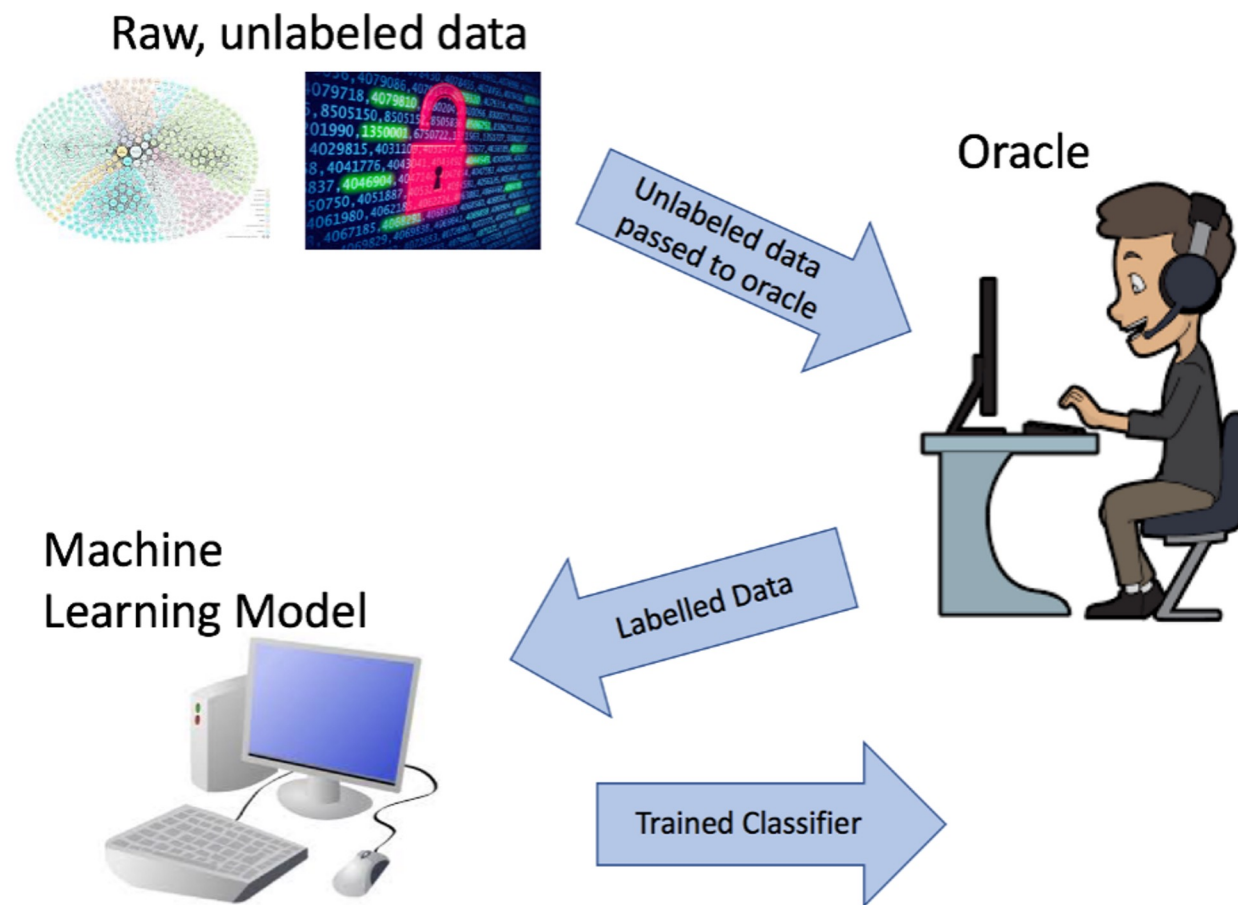
Radically efficient machine teaching. An annotation tool powered by active learning.



Pros:
Automation
Lots of features
Can train the models

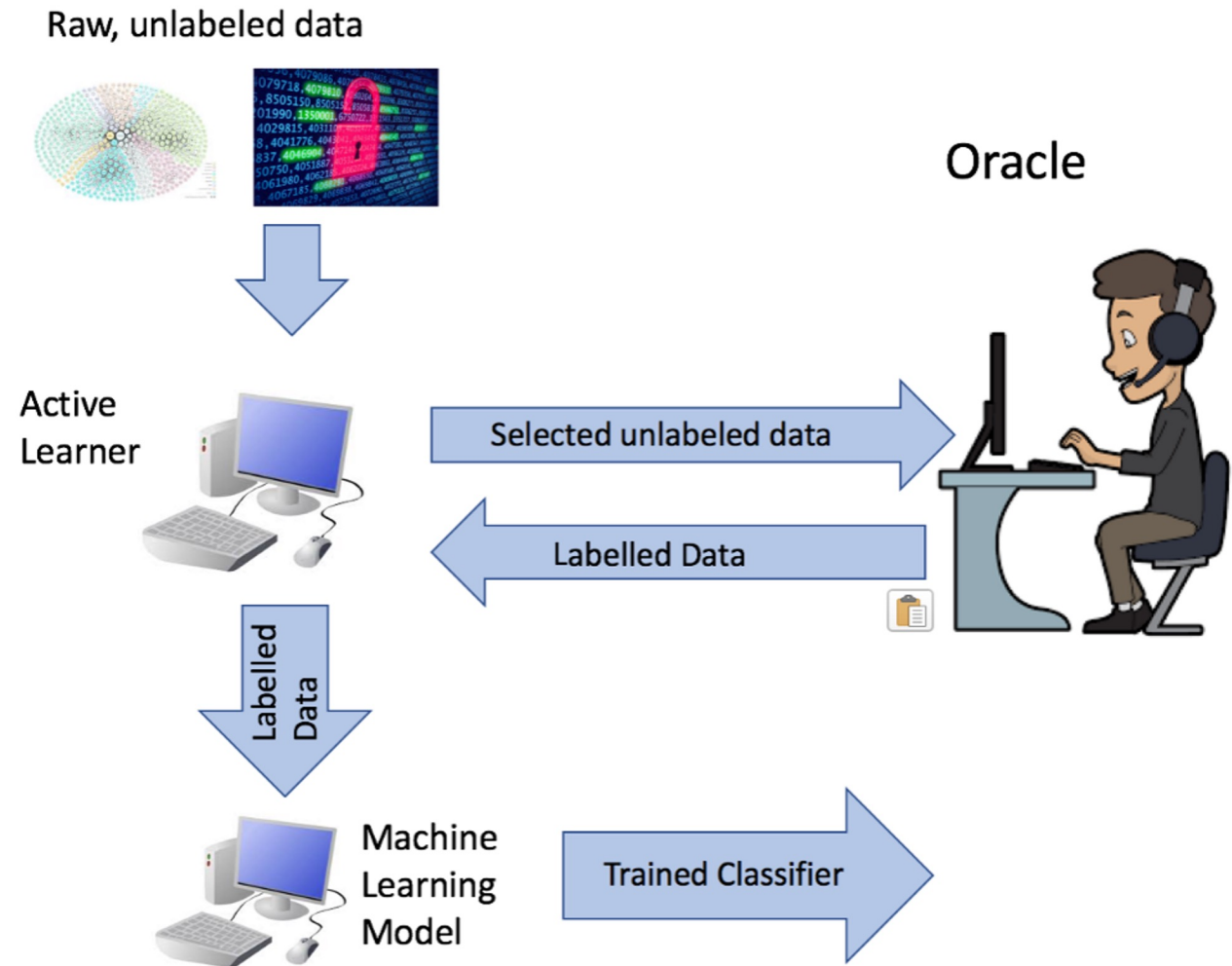
Cons:
Learning Curve
Not Open Source.

Passive learning



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

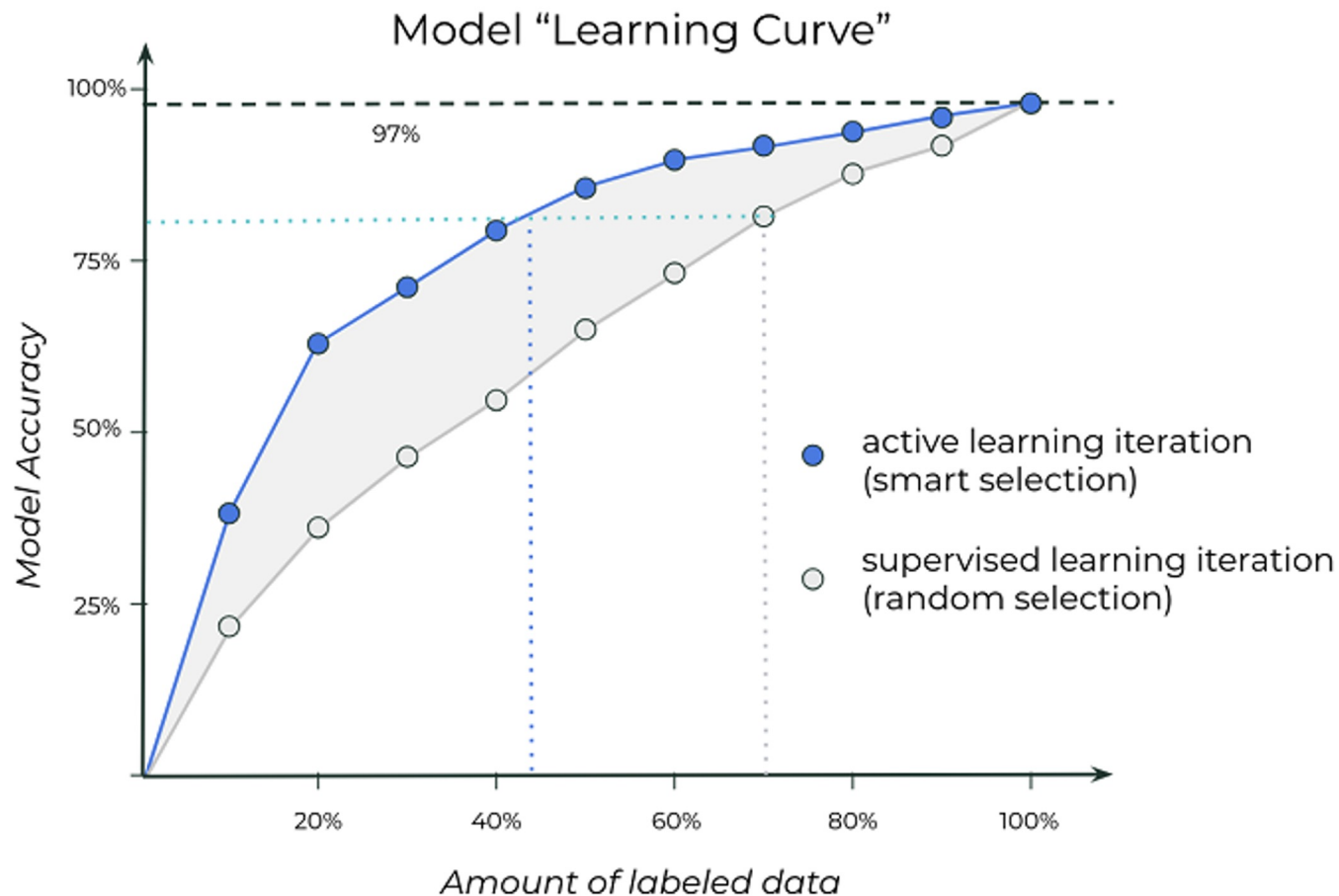
Active learning



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

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

ACCEPT 0
REJECT 0
IGNORE 0

HISTORY

© 2019 Explicon AI | Prodigy v1.3.0

PERSON 1 ORG 2 MONEY 3 TICKER 4

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

✓ ✗ ⦶ ↶



Active learning

Bruce PERSON Springsteen has sold the master recordings and publishing rights for his life's work to Sony for a reported \$500m (£376m).The deal gives Sony ownership of his 20 studio albums, including classics like Born To Run, The River and Born In The USA, according to multiple US reports.A 20-time Grammy winner, Springsteen's music generated about \$15m in revenue last year.His deal follows similar sales by Bob Dylan, Blondie and David PERSON Bowie.Warner Music bought the worldwide rights to Bowie's music in September, and Dylan sold his catalogue of more than 600 songs in December last year to Universal Music Group at a purchase price widely reported as \$300m.

SRL prediction
before active learning

Bruce Springsteen PERSON has sold the master recordings and publishing rights for his life's work to Sony ORG for a reported \$500m (£376m).The deal gives Sony ORG ownership of his 20 studio albums, including classics like Born To Run, The River and Born In The USA LOCATION, according to multiple US reports.A 20-time Grammy winner, Springsteen PERSON 's music generated about \$15m in revenue last year.His deal follows similar sales by Bob PERSON Dylan PERSON , Blond PERSON ie PERSON and David Bowie PERSON . Warner ORG Music ORG bought the worldwide rights to Bowie PERSON 's music in September, and Dylan PERSON sold his catalogue of more than 600 songs in December last year to Universal ORG Music ORG Group ORG at a purchase price widely reported as \$300m.

SRL prediction
after active learning



Issues in annotation





Classify between Order or Complaint?
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 coke 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 a SIZE large TOPPING cheese PIZZA 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 large cheese pizza and a QUANTITY a DRINK coke 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

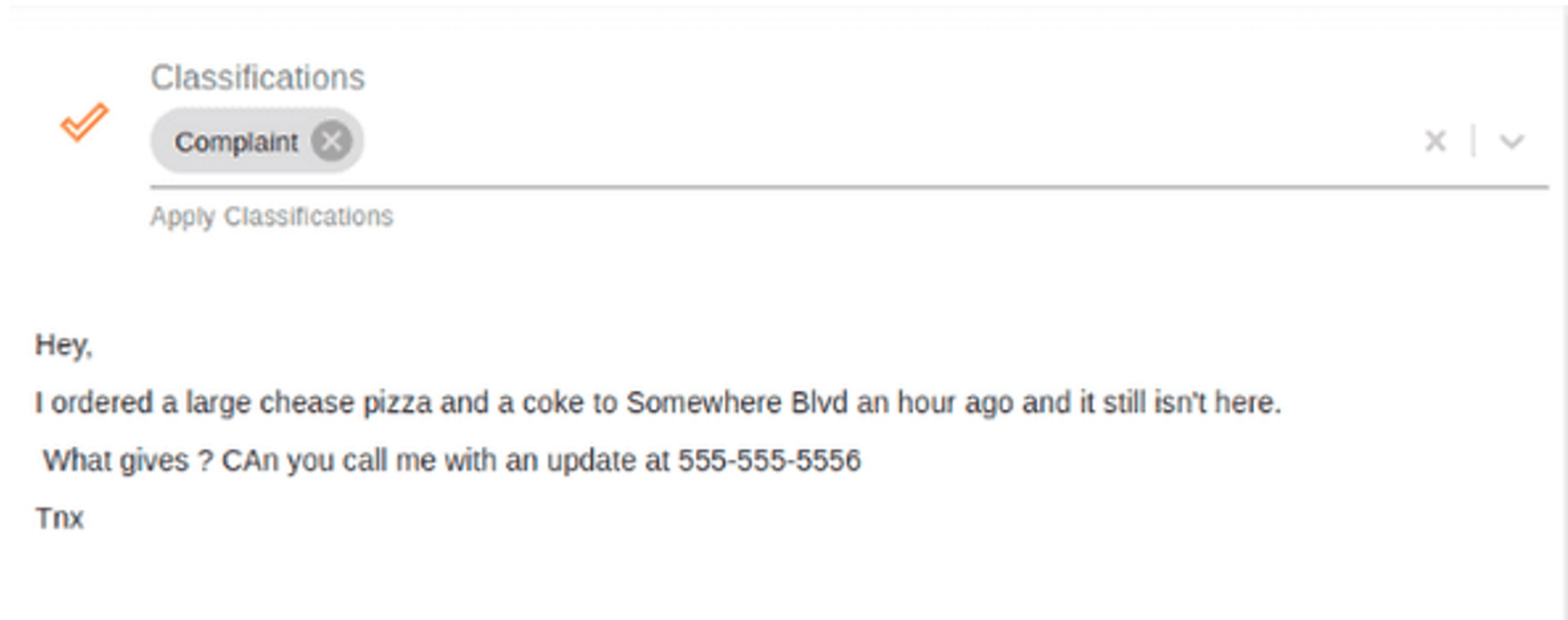


Disagreement

Intents

Conflict between document intent and entity tags

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



The screenshot shows a chat interface with a classification panel at the top. The panel is titled "Classifications" and has a checkmark icon on the left. Below the title, there is a button labeled "Complaint" with a close icon (X) on the right. To the right of the button, there are two more icons: a close icon (X) and a dropdown arrow (v). Below the panel, the text of the chat message is displayed: "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

Table 1: Examples from the five disagreement datasets used in this paper. A stands for annotator.

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> 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	Binary: 1 Continuous: 1/3
SChem101	“It’s okay to have abortion.”	A1 (age: 30-39, education: high school, race: white, gender: woman) votes for <u>people occasional think this</u> 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	Binary: 1 Continuous: 2/5

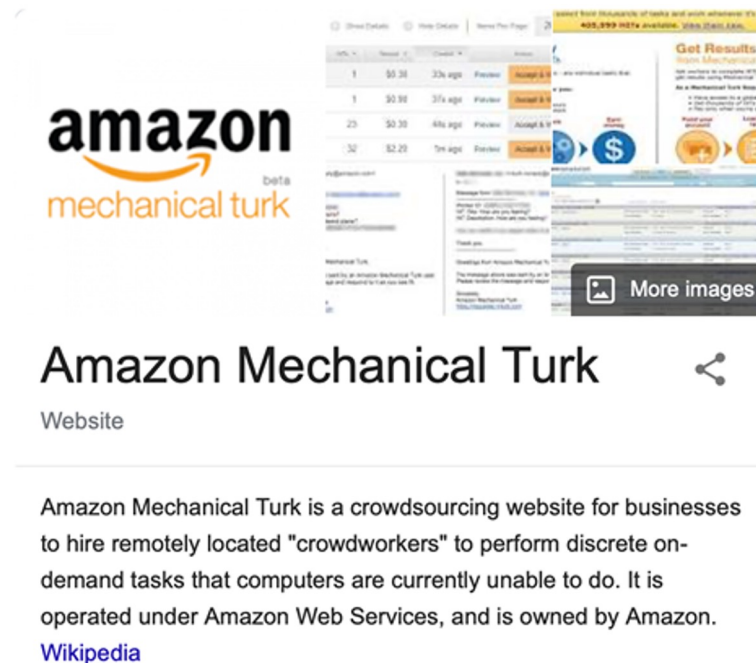


Disagreement for subjective datasets

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 of 25</u> . → Aggregated Label: impolite	Binary: 0 Continuous: 0



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)



Mitigate artifacts

Improve Inference Discourage use of artifacts

The diagram shows the equation $\log p_{\theta}(y|P, H) - \log p(y|H)$ with two red ovals highlighting the terms. A yellow arrow points from the first term to the text 'Improve Inference' above it. Another yellow arrow points from the second term to the text 'Discourage use of artifacts' above it.

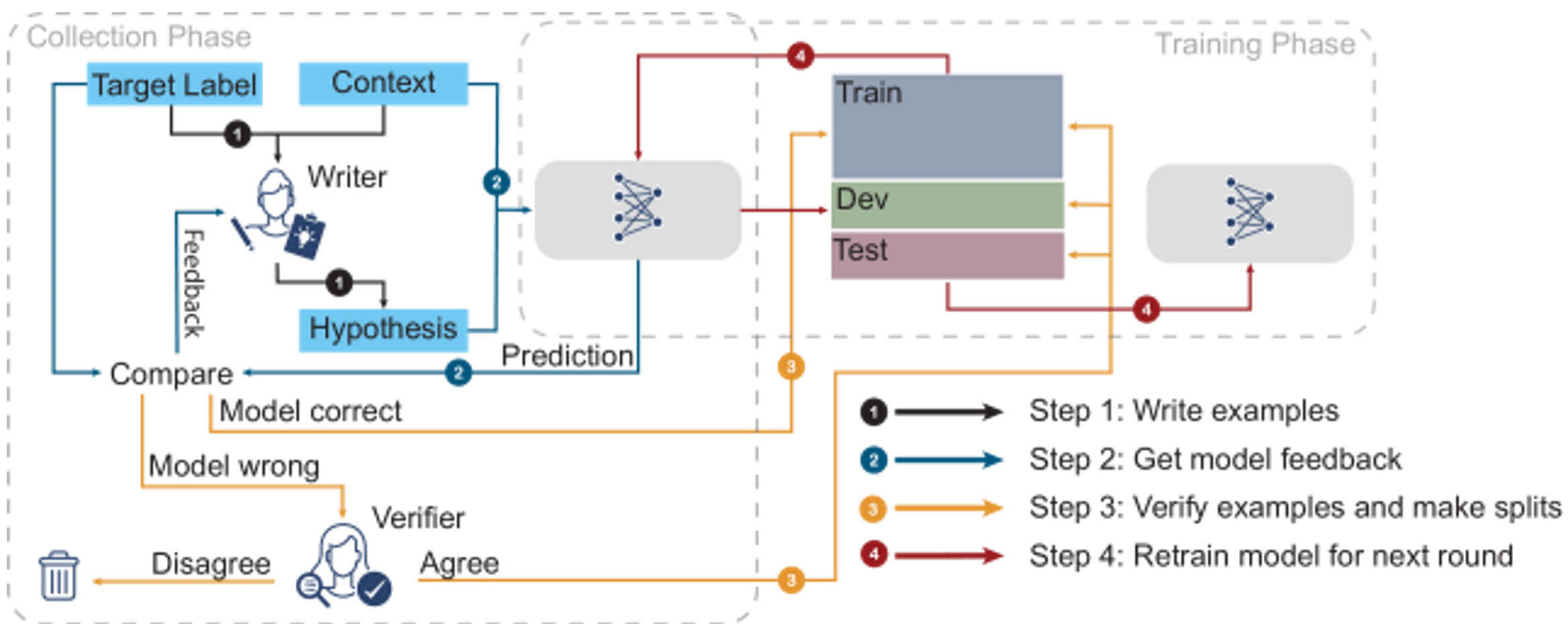
$$\log p_{\theta}(y|P, H) - \log p(y|H)$$

Don't Take the Premise for Granted: Mitigating Artifacts in Natural Language Inference (Belinkov et al, ACL 2019)



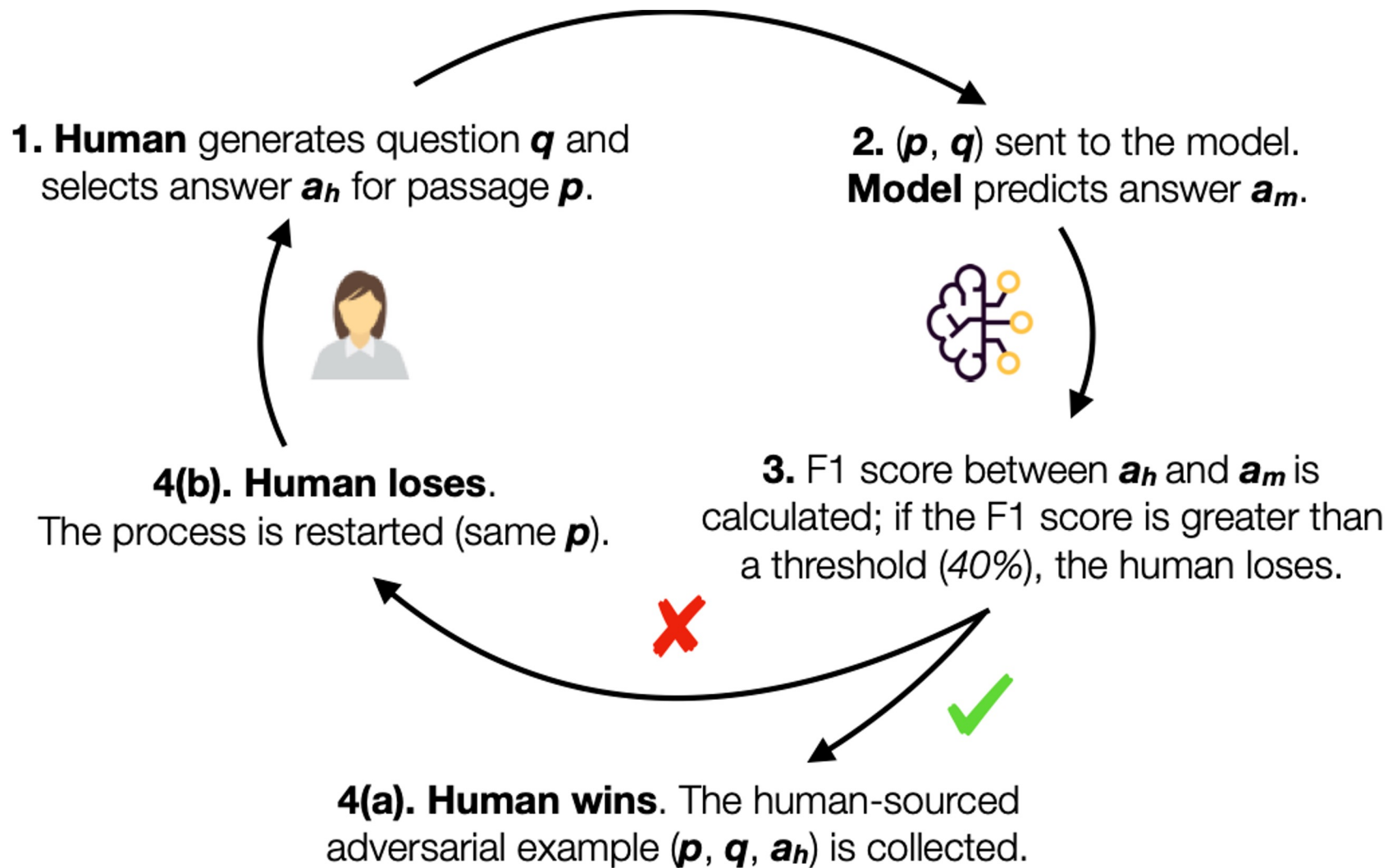
Advanced annotation techniques





Adversarial NLI: A New Benchmark for Natural Language Understanding





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

Dynabench: Rethinking Benchmarking in AI



ML Commons



THE UNIVERSITY
of NORTH CAROLINA
at CHAPEL HILL



The
Alan Turing
Institute

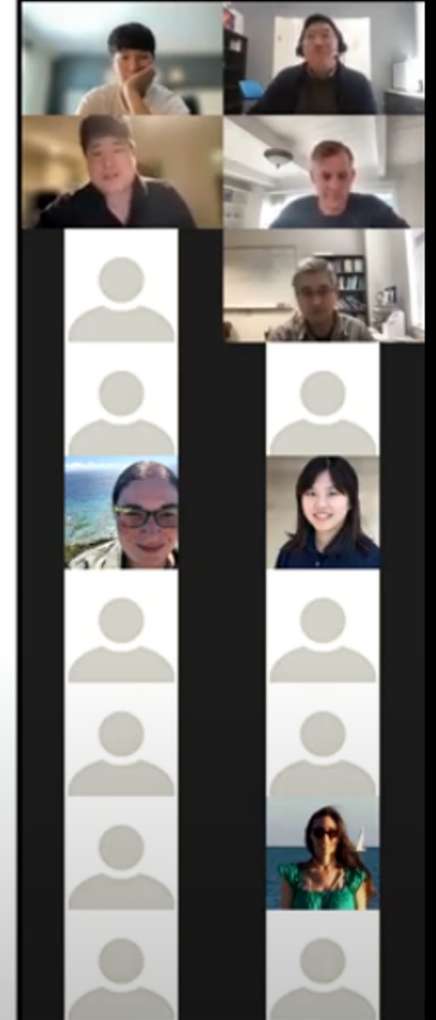


SFU
SIMON FRASER
UNIVERSITY



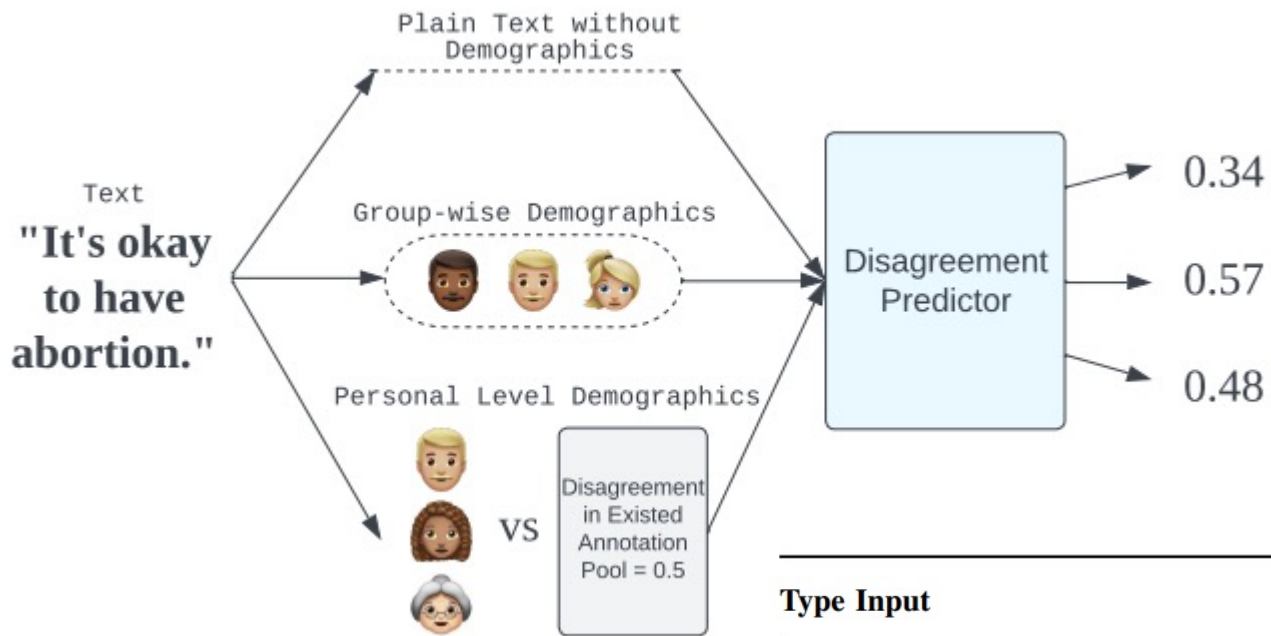
Stanford
University

FACEBOOK AI



<https://www.youtube.com/watch?v=3LP24xp5Bro>





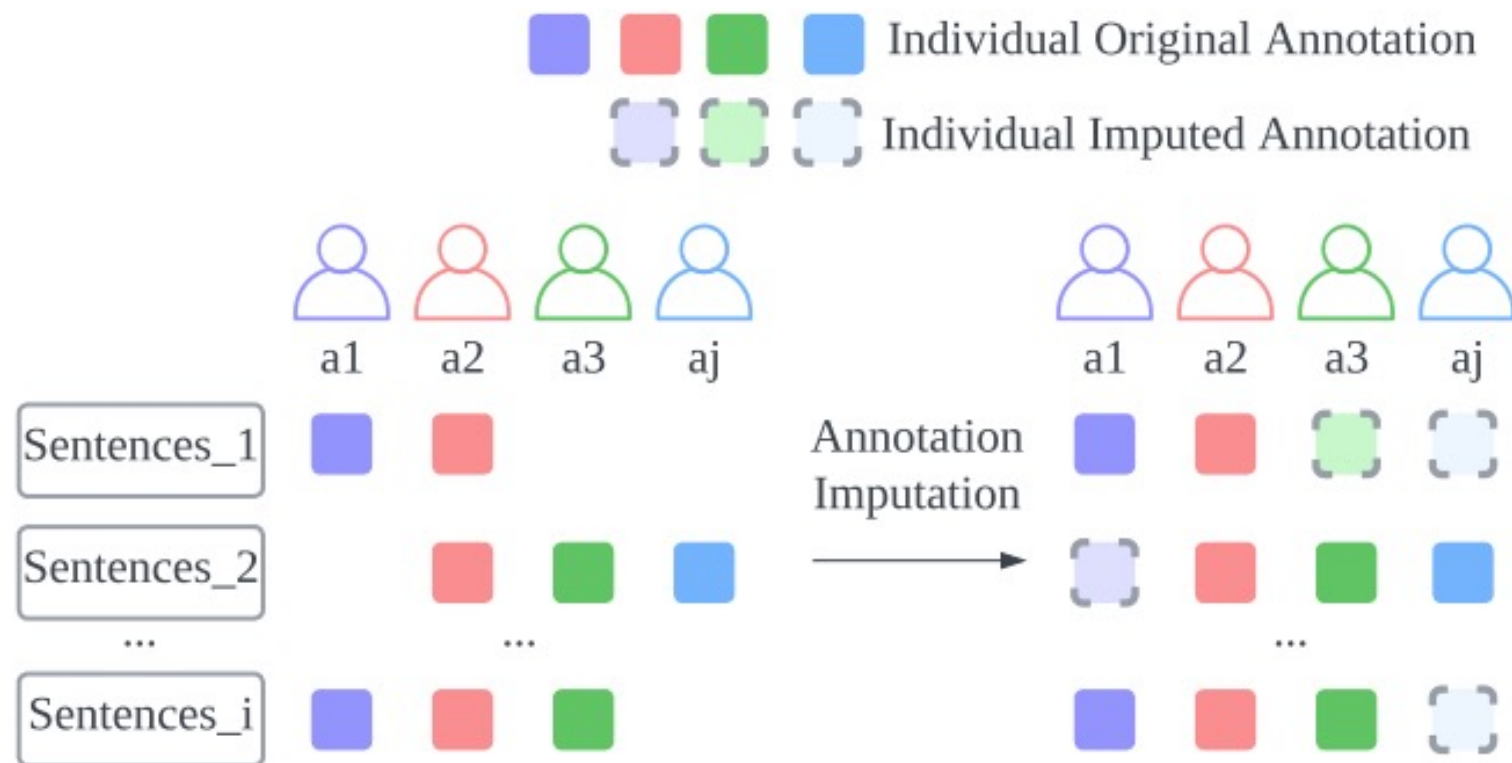
Type Input	Predicted Disagreement	Predicted Disagreement with Demographics	Collected Disagreement
(1) age: 22, politics: liberal, race: white, gender: woman . [SEP] what is the most common name for a black dude king kong.	0.003	0.015	0
(2) age: 25, politics: liberal, race: asian, gender: woman. [SEP] 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, AAI 2023

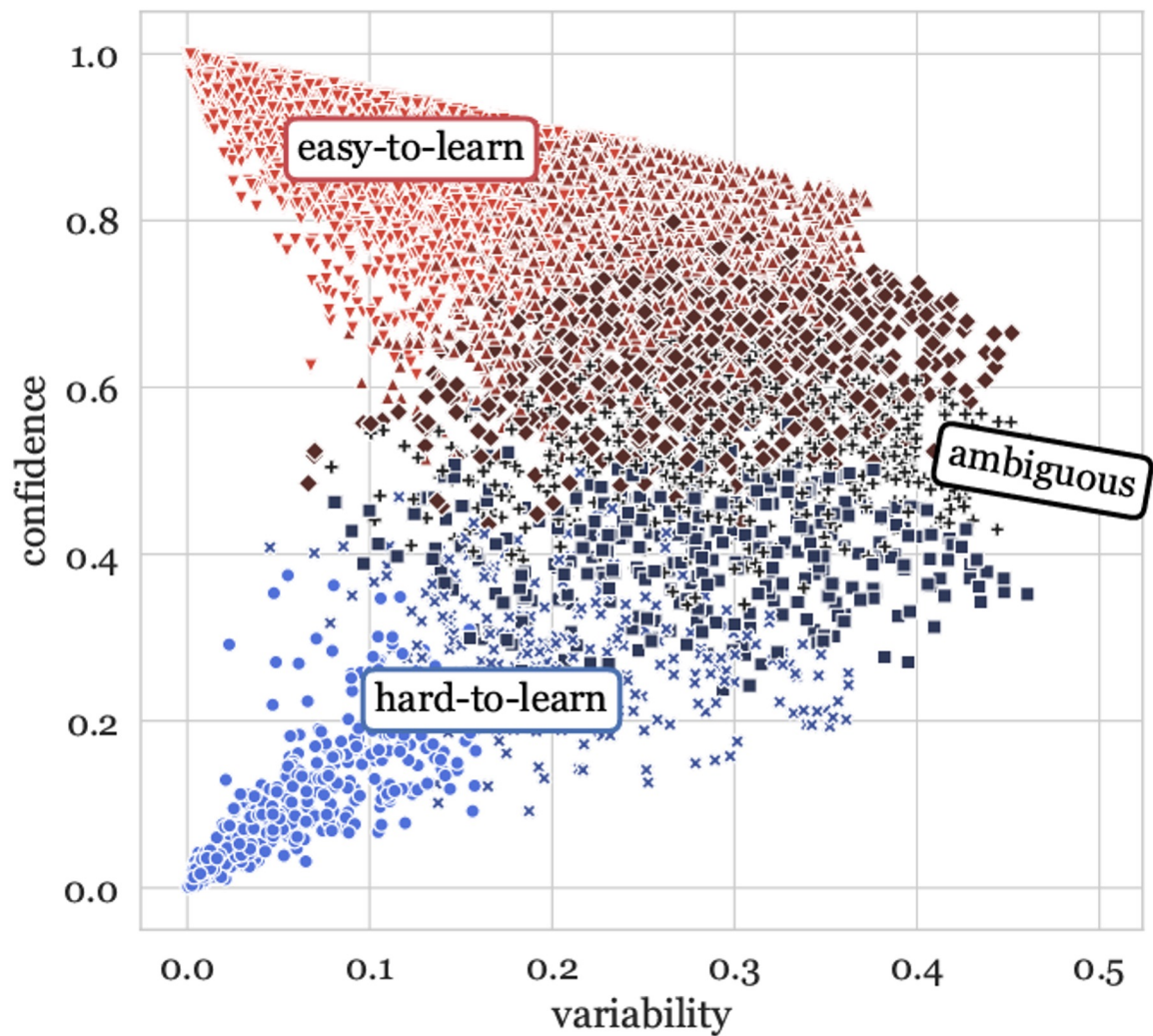


Annotation Imputation

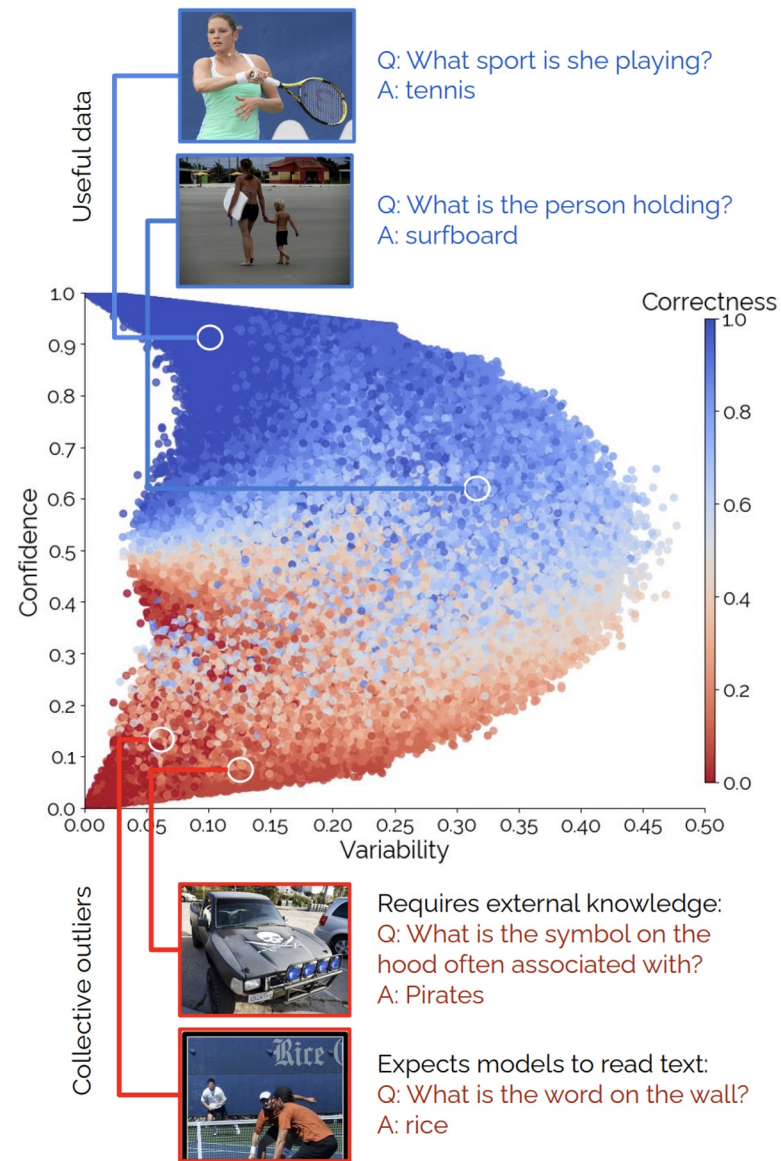


https://www.youtube.com/watch?v=x01ksJ9AW-w&ab_channel=LondonLowmanstoneIV

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

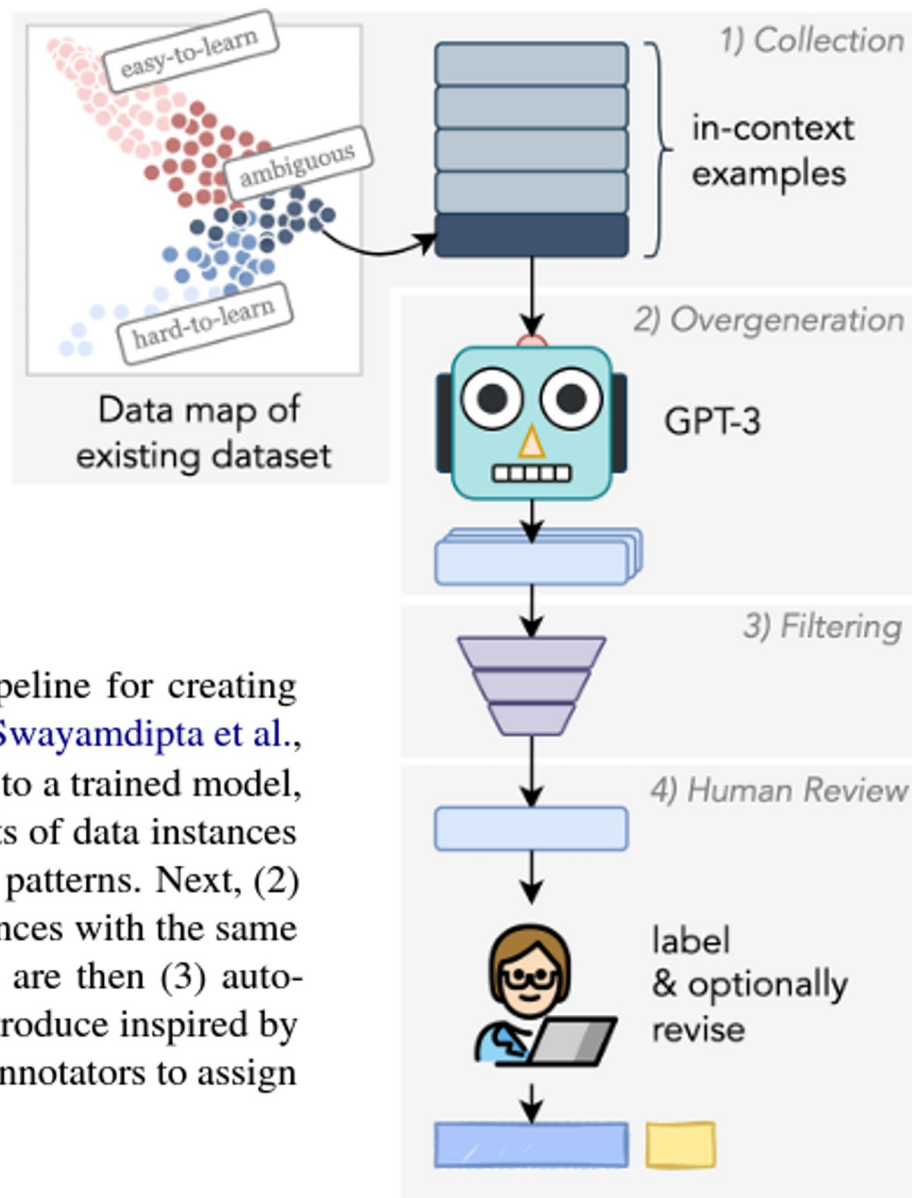
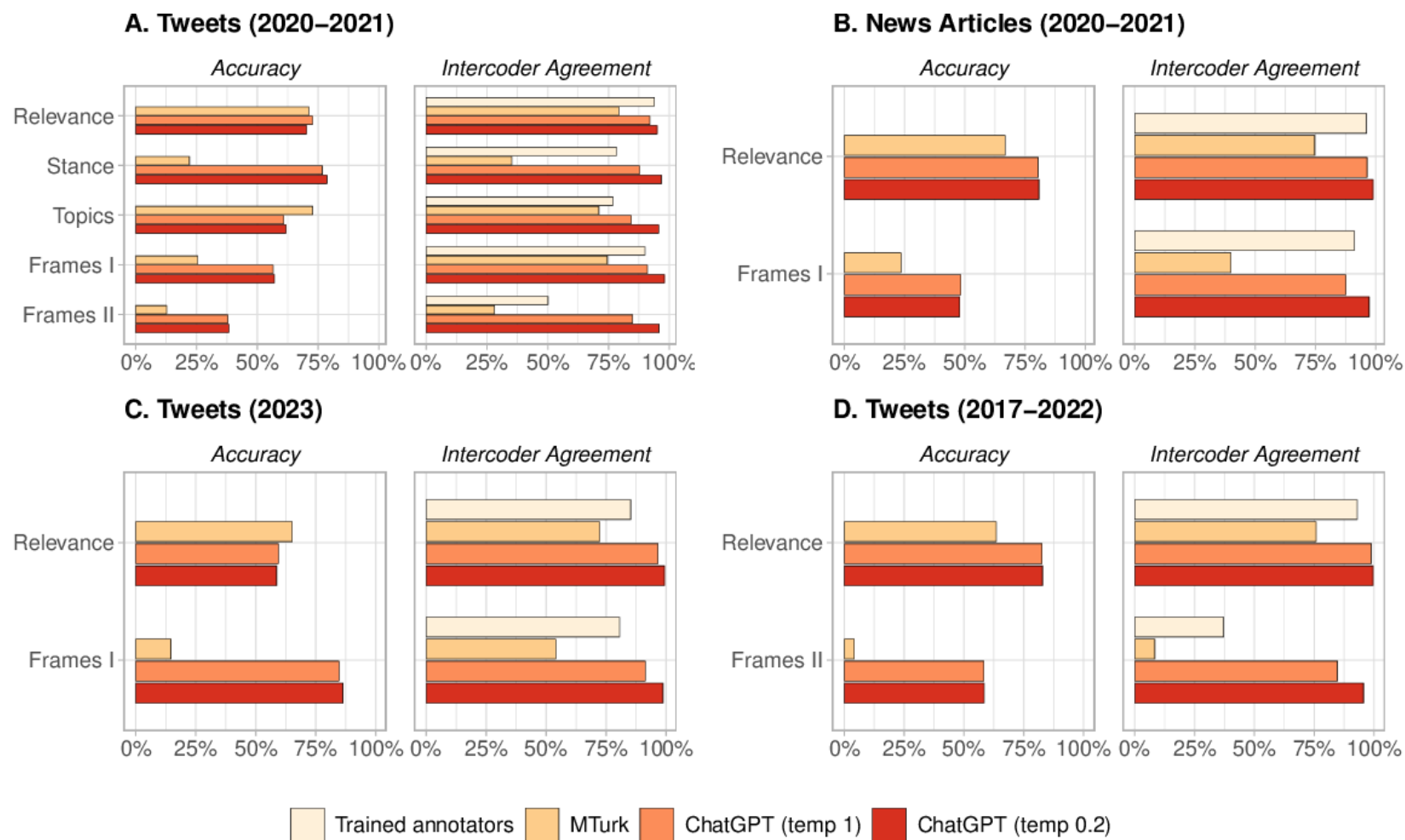


Figure 1: An illustration of our pipeline for creating WANLI. Starting with a data map (Swayamdipta et al., 2020) of an existing dataset relative to a trained model, (1) we automatically identify pockets of data instances exemplifying challenging reasoning patterns. Next, (2) we use GPT-3 to generate 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



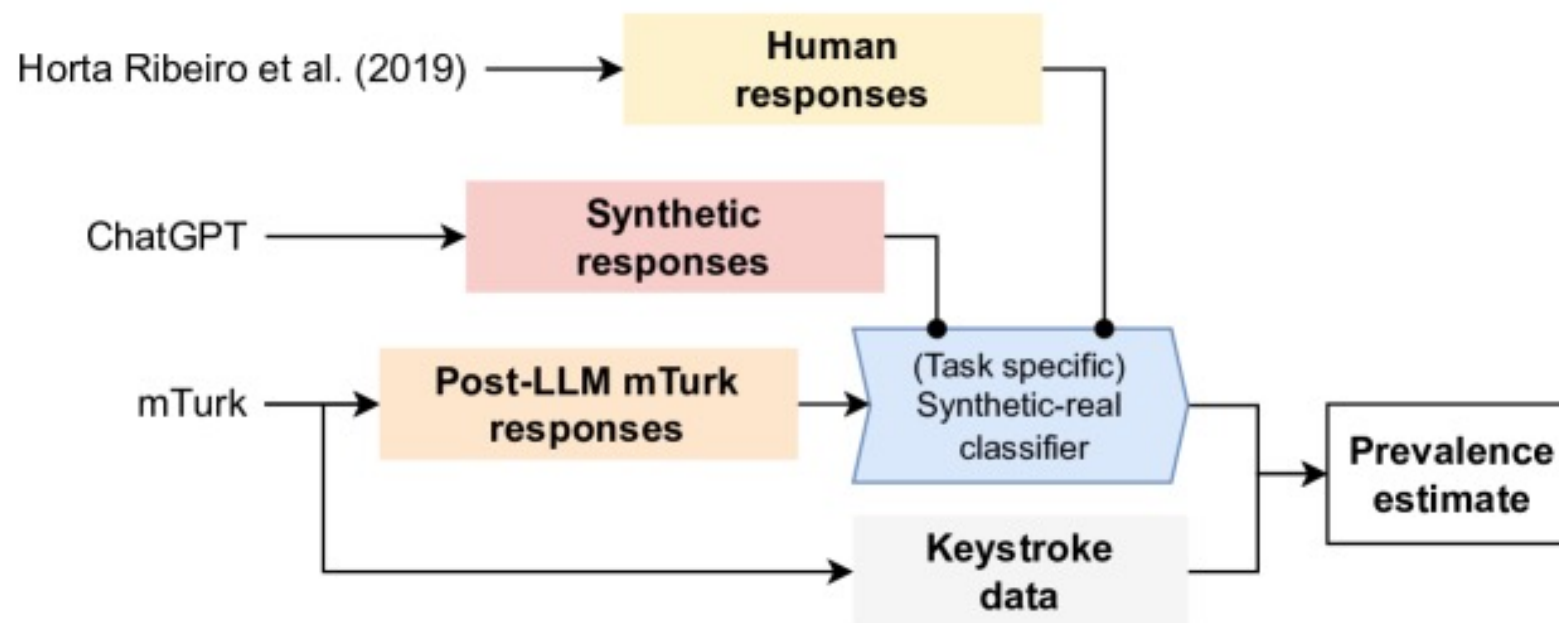
ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks <https://arxiv.org/abs/2303.15056>



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

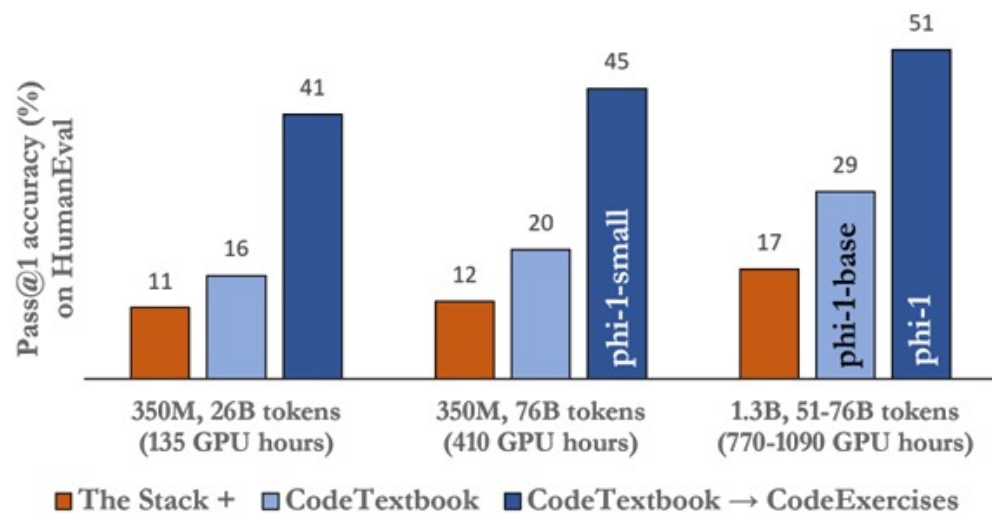


Artificial Artificial Artificial Intelligence: Crowd Workers Widely Use Large Language Models for Text Production Tasks <https://ar5iv.labs.arxiv.org/html/2306.07899>



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 = ''
        ...
```

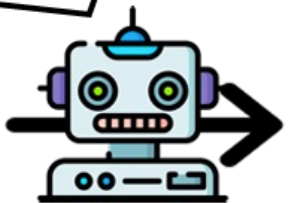
Chinchilla: Training Compute-Optimal Large Language Models , 2203.15556
Textbooks Are All You Need, 2306.11644
LIMA: Less Is More for Alignment 2305.11206



SelectLLM

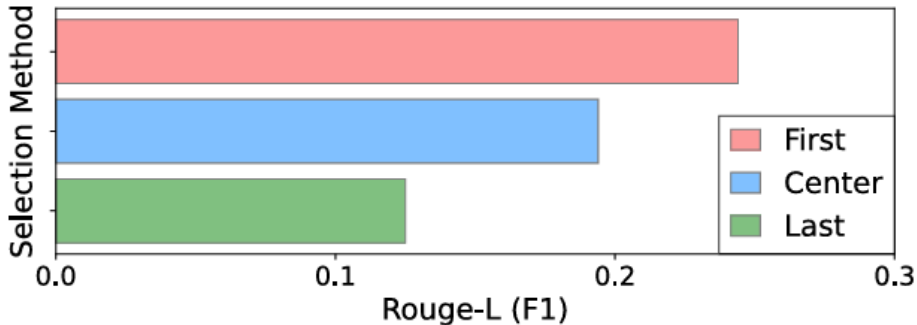
Rank given instructions based on their *impactfulness* and *informativeness* for **model fine-tuning**

[1] 1st Instruction
⋮
[N] Nth Instruction



[N] > ... > [10] > ... > [1]
First Center Last

Fine-tune LMs with Instructions from Different Rank Selections



The following are {N} candidate instructions that describe a task, each indicated by a number identifier [].

```
[1]
### Instruction: {Example #1 Instruction}
### Input: {Example #1 Input}
.
.
.
[N]
### Instruction: {Example #N Instruction}
### Input: {Example #N Input}
```

Examine the provided list of {N} instructions, each uniquely identified by a number in brackets [].

Your task is to select {num} instructions that will be annotated by human annotators for model fine-tuning.

Look for instructions that are clear and relevant, exhibit a high level of complexity and detail, represent a diverse range of scenarios and contexts, offer significant instructional value and potential learning gain, and present unique challenges and specificity.

These selected instructions should ideally be the most beneficial for model fine-tuning after being annotated by human annotators.

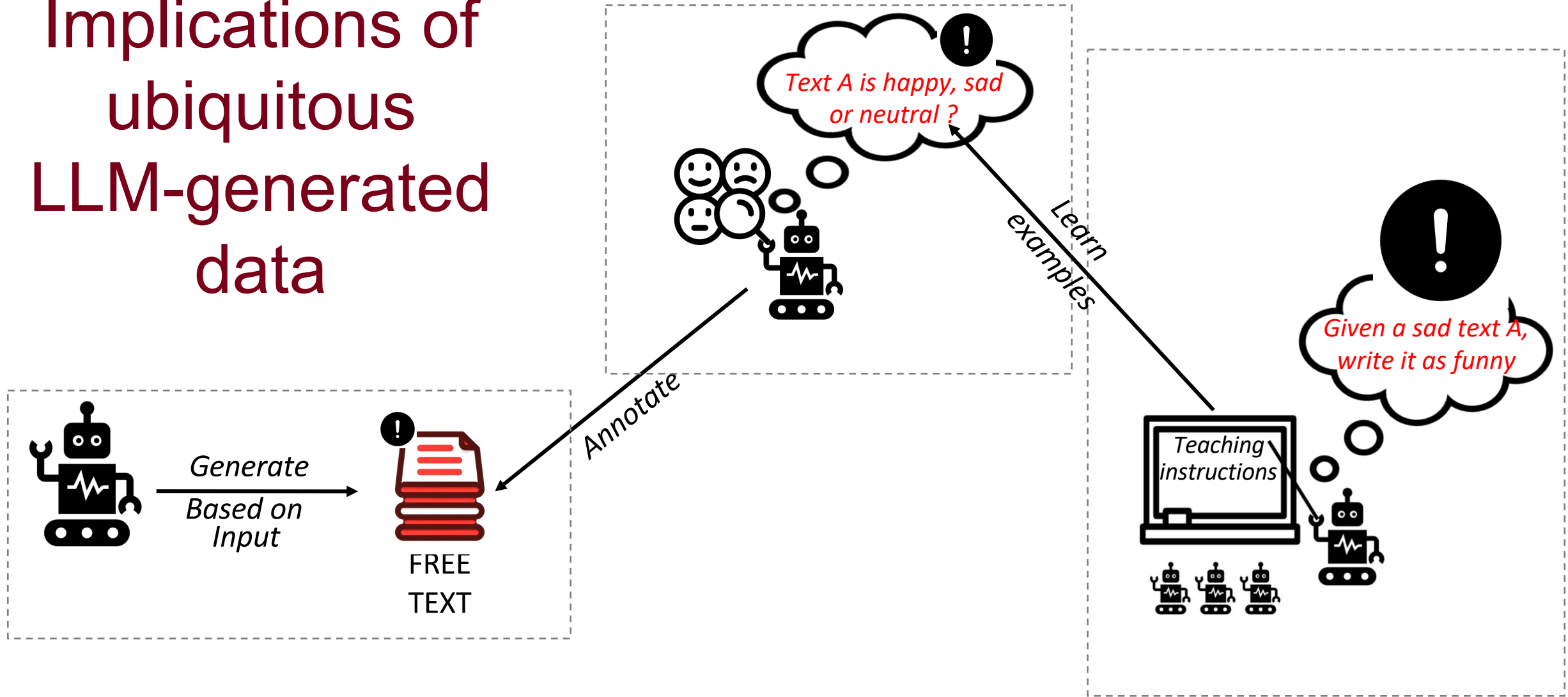
Present your selections using the format [], e.g., [1,2] or [2,3].

The most impactful {num} instructions (only identifiers) are:

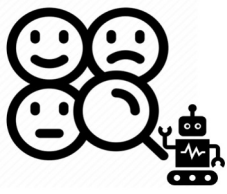
SelectLLM: Can LLMs Select Important Instructions to Annotate? <https://arxiv.org/abs/2401.16553>



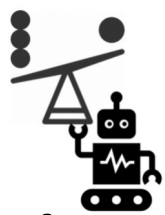
Implications of ubiquitous LLM-generated data



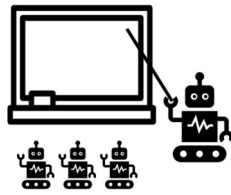
Under the Surface: Tracking the Artifactuality of LLM-Generated Data <https://arxiv.org/abs/2401.14698>



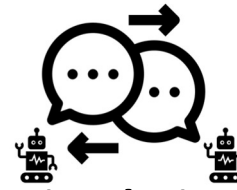
Task Labels



Preference



Instructions



Simulation



Free-Form Text

1) PROMPT :

Choose the sentiment of the given text from positive and negative

Text: a feast for the eyes

Response :

Positive

2) **PROMPT:** Which sentence sounds more negative?

Sentence 1: I did not really feel like it was heartfelt since she never called to apologize.

Sentence 2: They were a bit distracted and a little under-staffed, so maybe that's why.

Response:

Human: Sentence 1

GPT-3: Sentence 2

3) **PROMPT :** Instruction 1: You are given a science question and four answer options. Your task is to find the correct answer.

Input: Which part of a bicycle BEST moves in a circle? ...

Instruction 2: Given a negative review, convert it to a positive review by making minimal changes.

Input: we stood there in shock, because we...

Response :

Instruction: In this task, you will be given a profile of someone and your job is to generate a set of interesting questions that can lead to a conversation with the person.
Input: Yvonne has been playing the violin since she was four years old. She loves all kinds of music, but her favorite composer is Bach.

4) **PROMPT :** Here's the context for this question: Seeker : My dog is the only reason I haven't ended everything. . I just imagine leaving her. Counsellor response: I want to give my cats a good life that's what keeps me going.

Explorations are when a mental health counsellor shows interest in a seeker by asking about unstated experiences. What level of exploration is expressed in the response ? A. strong exploration B. weak exploration C. No exploration.

Response :

Agent 1 : I think the response shows strong exploration, because it mentions the importance of a pet in the seeker's life. Answer : A
Agent 2 : I disagree with Agent1. Thought the response acknowledges a pet, it does not specifically acknowledge the seeker's feelings. I think the level of exploration is weak. Answer : B

5) **PROMPT:** Based on social media text with a {target sentiment}, can you write a new text in a similar style with the same sentiment.

Text: Lucian Favre having 2nd thoughts about Gladbach qualifying for the Champions League - Juventus, Man City and Sevilla. Group Of Death.

Target sentiment: Negative

Response:

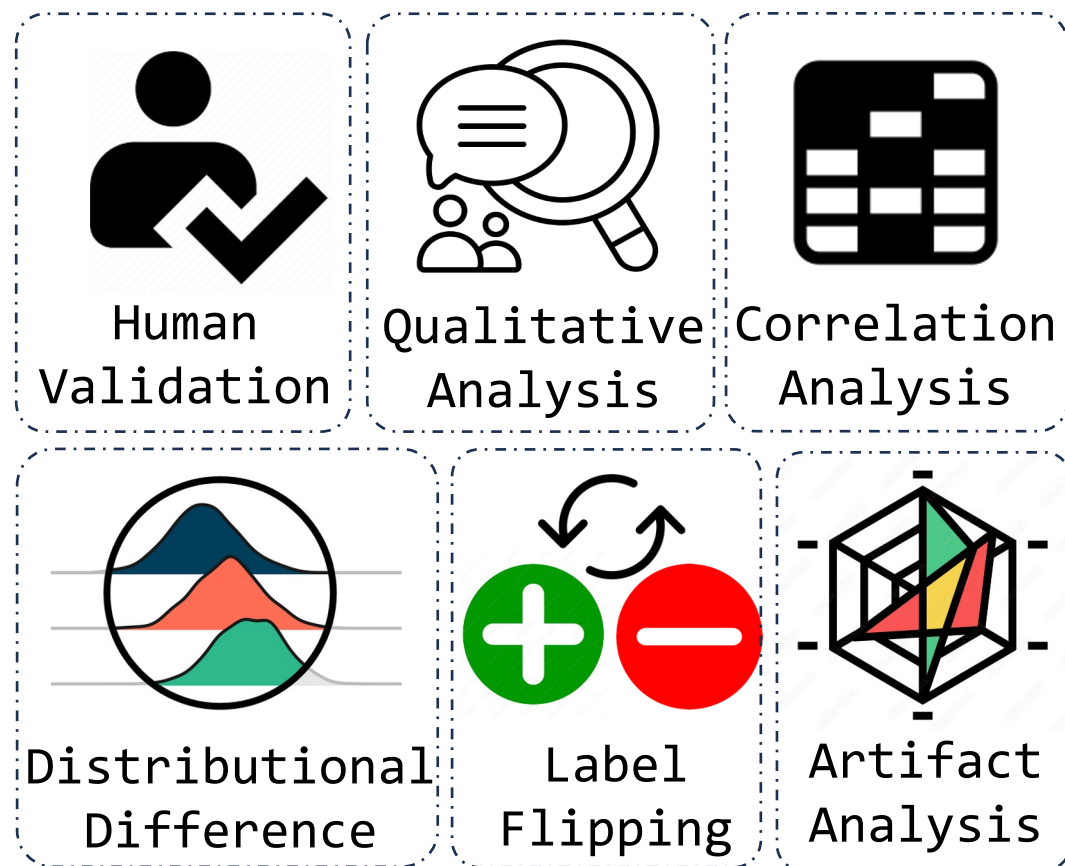
Why do we get the worst groups every year? This Champions League draw is no different - Real Madrid, Dortmund, and Galatasaray. How are we supposed to advance?

Types of LLM-generated data

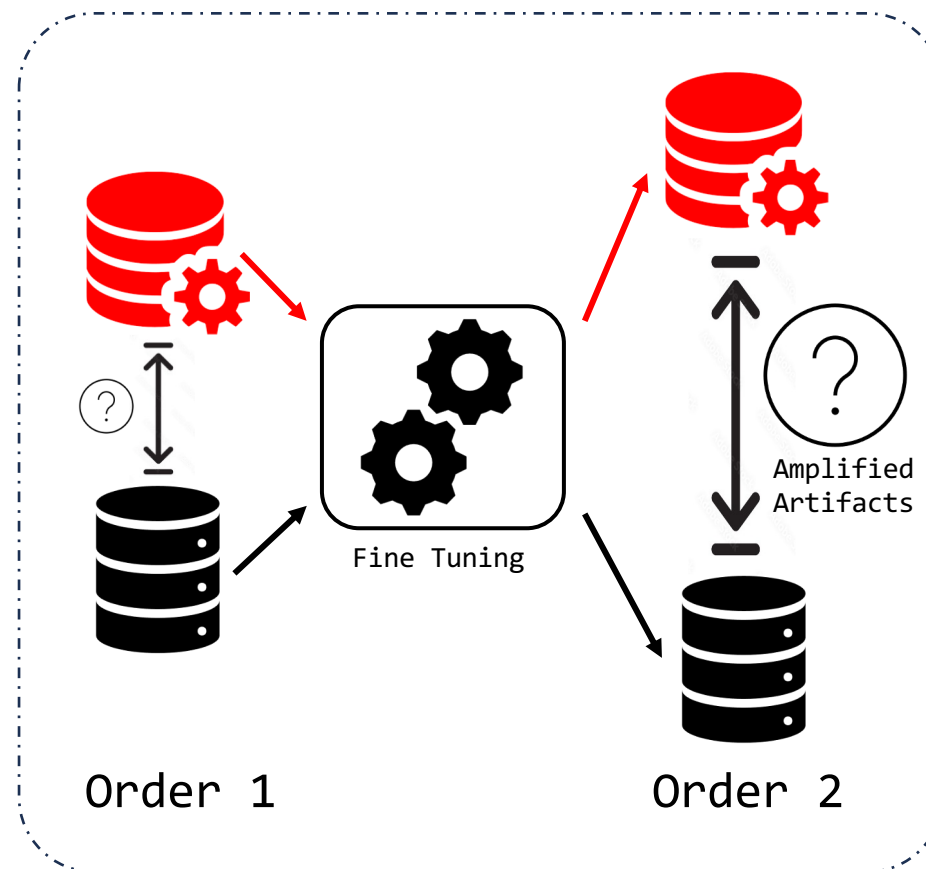
Under the Surface: Tracking the Artifactuality of LLM-Generated Data <https://arxiv.org/abs/2401.14698>



Stress Testing Methods

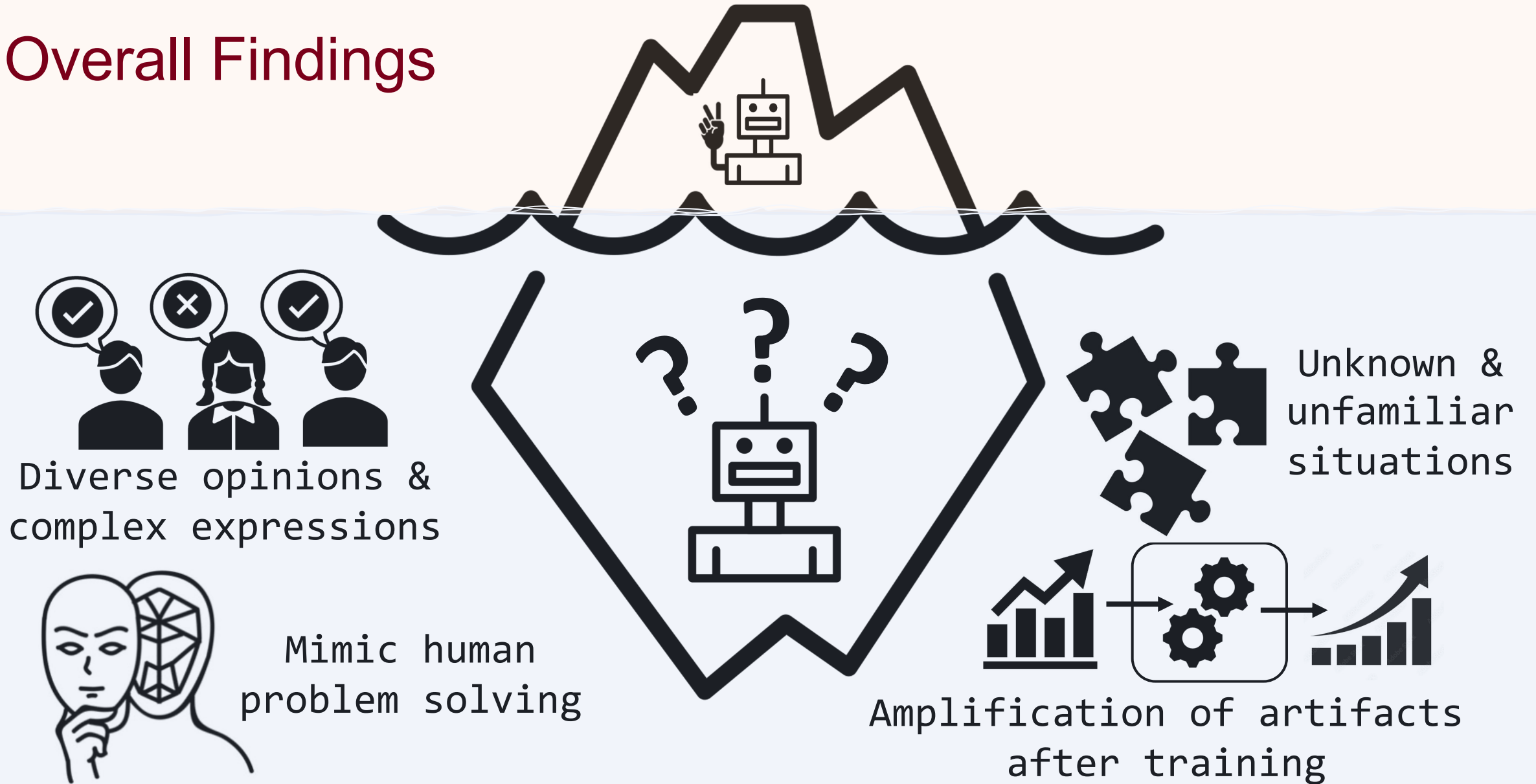


2nd order Stress Testing Methods



Under the Surface: Tracking the Artifactuality of LLM-Generated Data <https://arxiv.org/abs/2401.14698>

Overall Findings



Under the Surface: Tracking the Artificality of LLM-Generated Data <https://arxiv.org/abs/2401.14698>

Outline

- ❑ 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

