Annotations and tasks

CS 685, Fall 2025

Advanced Natural Language Processing

https://people.cs.umass.edu/~brenocon/cs685_f25/

Brendan O'Connor

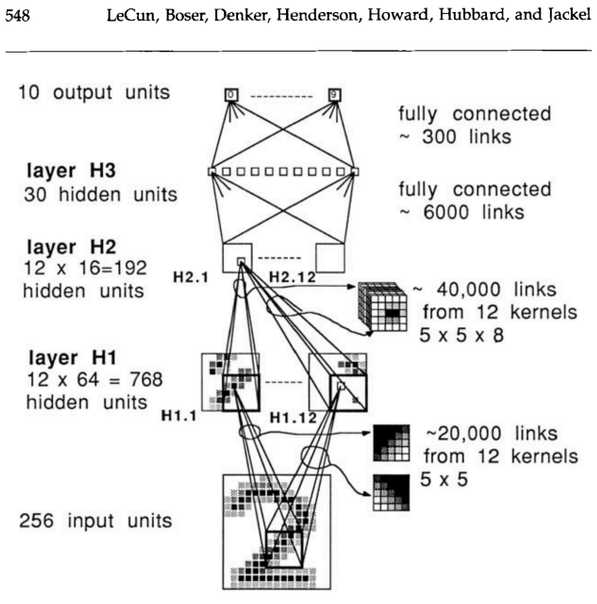
College of Information and Computer Sciences

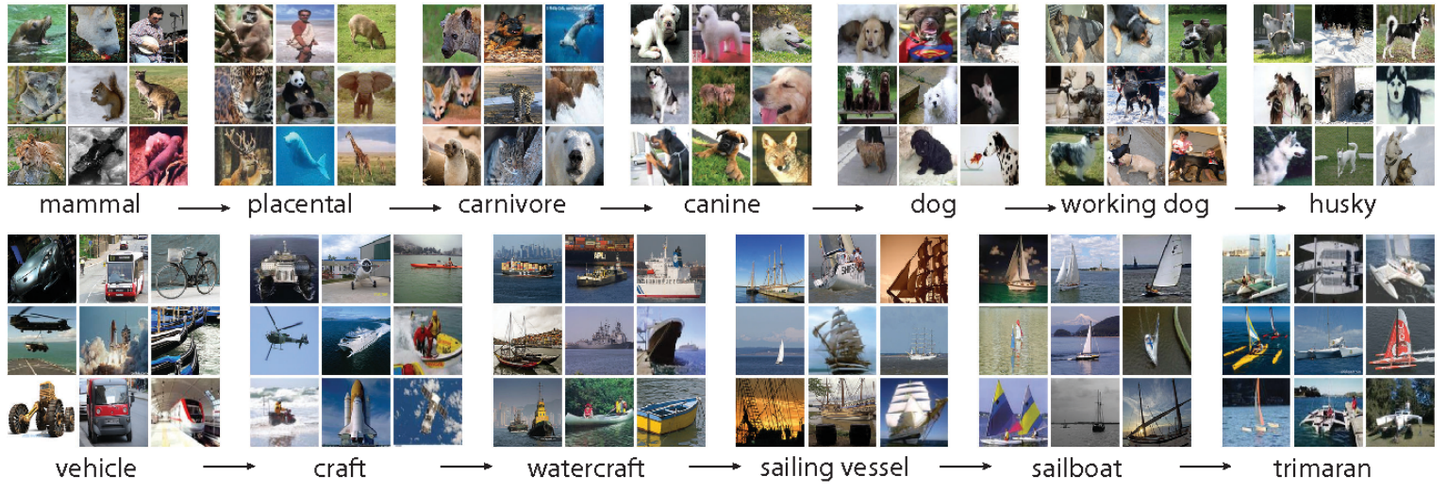
University of Massachusetts Amherst

- Are your labels any good??

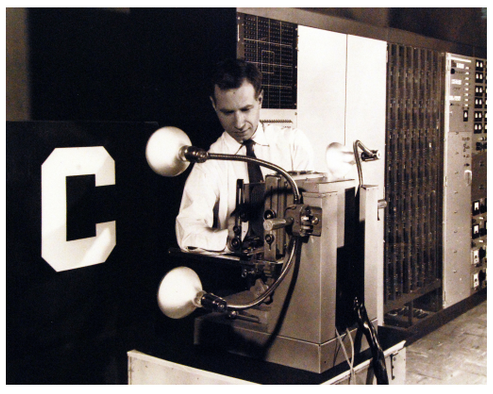
- Define a classification task that you'd like a model to do
- Then you need text and labels
  - 1. Natural annotations — information you can automatically retrieve about a text
  - 2. New human annotations — get people to manually create labels for a sample of texts!
  - (3. Repurposed old human annotations - e.g. educational texts/exams... if it's what you want to do...)

- Human behavioral data was a key factor for today's 3rd wave of neural network modeling, initially in computational vision





1957
Perceptron
(~log. reg.)



1989
Backprop &
convolutional NN

2012
ImageNet
data for CNN
training

Millions of labeled objects in images, collected via crowdsourcing (MTurk)
Revolutionized CV by using nearly the same model from 1989!

2019
(Counterpoint: LM pretraining)

https://www.mturk.com/

Secure https://www.mturk.com/mturk/welcome

amazon mechanical turk beta
Artificial Artificial Intelligence

Your Account | HITs | Qualifications

Already have an account?
Sign in as a **Worker** | **Requester**

Introduction | **Dashboard** | **Status** | **Account Settings**

Mechanical Turk is a marketplace for work.

We give businesses and developers access to an on-demand, scalable workforce.
Workers select from thousands of tasks and work whenever it's convenient.

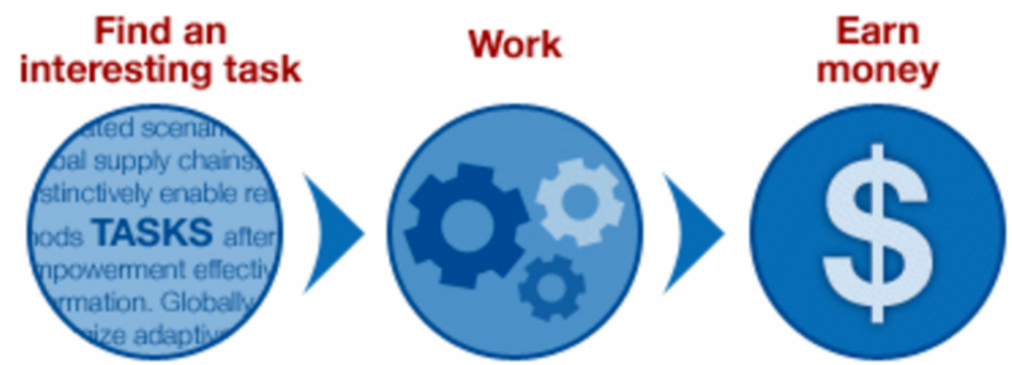
247,056 HITs available. View them now.

Make Money by working on HITs

HITs - *Human Intelligence Tasks* - are individual tasks that you work on. Find HITs now.

As a Mechanical Turk Worker you:

- Can work from home
- Choose your own work hours
- Get paid for doing good work



Get Results from Mechanical Turk Workers

Ask workers to complete HITs - *Human Intelligence Tasks* - and get results using Mechanical Turk. Get Started.

As a Mechanical Turk Requester you:

- Have access to a global, on-demand, 24 x 7 workforce
- Get thousands of HITs completed in minutes
- Pay only when you're satisfied with the results

- Crowdsourcing approach to annotation
  - Many different people do many different chunks
  - Typically don't know much about the workers (unless you add extra surveys / pre-test qualifications, etc.)
- Quality control is key
  - Check agreement with hidden gold standard items
  - Check agreement against other workers' judgments
- Useful resources: https://jkk.name/reading-notes/crowdsourcing/

- Alternative: higher skilled online providers (e.g. Upwork)
  - More like directly hiring a small number of freelancers; higher cost, better motivation, expertise
- Large companies/organizations typically have annotation contractors or employ some in-house

Exercise

Interannotator agreement

- How “real” is a task? Replicable? Reliability of annotations?
- How much do two humans *agree* on labels?
- Question: can an NLP system's accuracy be higher than the human agreement rate?

- The conventional view: IAA is the upper bound for machine performance
  - What affects IAA? Difficulty of task, human training, human motivation/effort....
  - IAA between 2 humans, or between 1 human vs. consensus

Cohen's Kappa for IAA

- If some classes predominate, raw agreement rate may be misleading
- Idea: normalize accuracy (agreement) rate such that answering randomly = 0.
  - From psychology / psychometrics / content analysis
- **Chance-adjusted agreement** from:

p_o: **o**bserved agreement rate

p_e: **e**xpected (by chance) rate

Other chanced-adjusted metrics: Fleiss, Krippendorff... see reading

- (added after lecture -- supplement to explanation on chalkboard)

- p_pool(y): the marginal probability of the label, among all annotators. for your exercise, literally the its count divided by 20
- chance agreement rate: assume two annotators are labeling at random, by the p_pool distribution (sum of squares).
- intuition for the adjustment formula: answering at chance rate should get 0. within the space between chance agreement and full agreement, how far up is your observed agreement rate?
- See Eisenstein reading for more details
  - note there are different kappa variants that may use pooled vs. unpooled probabilities

- Iterative cycle for annotation and guidelines/ codebook development
  1. Develop guidelines for the task
  2. Give guidelines to annotators
  3. Annotate a little
  4. Discuss/analyze disagreements; revise guidelines

Annotation example: framing/persuasion methods

- Annotating multilingual news articles from 2020-2022

3.1 Genre

Given a news article, we want to characterize the intended nature of the reporting: whether it is an *opinion* piece, it aims at objective news *reporting*, or it is *satirical*. This is a multiclass annotation scheme at the article level.

A satirical piece is a factually incorrect article, with the intent not to deceive, but rather to call out, ridicule, or expose behaviours considered 'bad'. It deliberately exposes real-world individuals, organisations and events to ridicule.

Given that the borders between *opinion* and objective news *reporting* might sometimes not be fully clear, we provide in Appendix A.1 an excerpt from the annotation guidelines with some rules that were used to resolve *opinion* vs. *reporting* cases.

3.2 Framing

Given a news article, we are interested in identifying the frames used in the article. For this purpose, we adopted the concept of framing introduced in (Card et al., 2015) and the taxonomy of 14 generic framing dimensions, their acronym is specified in parenthesis: *Economic (E)*, *Capacity and resources (CR)*, *Morality (M)*, *Fairness and equality (FE)*, *Legality, constitutionality and jurisprudence (LCJ)*, *Policy prescription and evaluation (PPE)*, *Crime and punishment (CP)*, *Security and defense (SD)*, *Health and safety (HS)*, *Quality of life (QOL)*, *Cultural identity (CI)*, *Public opinion (PO)*, *Political (P)*, and *External regulation and reputation (EER)*.

3.3 Persuasion Techniques

Attack on reputation: The argument does not address the topic, but rather targets the participant (personality, experience, deeds) in order to question and/or to undermine their credibility. The object of the argumentation can also refer to a group of individuals, an organization, an object, or an activity.
Justification: The argument is made of two parts, a statement and an explanation or an appeal, where the latter is used to justify and/or to support the statement.
Simplification: The argument excessively simplifies a problem, usually regarding the cause, the consequence, or the existence of choices.
Distraction: The argument takes focus away from the main topic or argument to distract the reader.
Call: The text is not an argument, but an encouragement to act or to think in a particular way.
Manipulative wording: the text is not an argument per se, but uses specific language, which contains words or phrases that are either non-neutral, confusing, exaggerating, loaded, etc., in order to impact the reader emotionally.

[Piskorski et al. 2023]

Annotation example: framing/persuasion methods

ATTACK ON REPUTATION

Name Calling or Labelling [AR:NCL]: a form of argument in which loaded labels are directed at an individual, group, object or activity, typically in an insulting or demeaning way, but also using labels the target audience finds desirable.
Guilt by Association [AR:GA]: attacking the opponent or an activity by associating it with a another group, activity or concept that has sharp negative connotations for the target audience.
Casting Doubt [AR:D]: questioning the character or personal attributes of someone or something in order to question their general credibility or quality.
Appeal to Hypocrisy [AR:AH]: the target of the technique is attacked on its reputation by charging them with hypocrisy/inconsistency.
Questioning the Reputation [AR:QR]: the target is attacked by making strong negative claims about it, focusing specially on undermining its character and moral stature rather than relying on an argument about the topic.

JUSTIFICATION

Flag Waving [J:FW]: justifying an idea by exhaling the pride of a group or highlighting the benefits for that specific group.
Appeal to Authority [J:AA]: a weight is given to an argument, an idea or information by simply stating that a particular entity considered as an authority is the source of the information.
Appeal to Popularity [J:AP]: a weight is given to an argument or idea by justifying it on the basis that allegedly "everybody" (or the large majority) agrees with it or "nobody" disagrees with it.
Appeal to Values [J:AV]: a weight is given to an idea by linking it to values seen by the target audience as positive.
Appeal to Fear, Prejudice [J:AF]: promotes or rejects an idea through the repulsion or fear of the audience towards this idea.

DISTRACTION

Strawman [D:SM]: consists in making an impression of refuting an argument of the opponent's proposition, whereas the real subject of the argument was not addressed or refuted, but instead replaced with a false one.
Red Herring [D:RH]: consists in diverting the attention of the audience from the main topic being discussed, by introducing another topic, which is irrelevant.
Whataboutism [D:W]: a technique that attempts to discredit an opponent's position by charging them with hypocrisy without directly disproving their argument.

SIMPLIFICATION

Causal Oversimplification [S:CaO]: assuming a single cause or reason when there are actually multiple causes for an issue.
False Dilemma or No Choice [S:FDNC]: a logical fallacy that presents only two options or sides when there are many options or sides. In extreme, the author tells the audience exactly what actions to take, eliminating any other possible choices.
Consequential Oversimplification [S:CoO]: is an assertion one is making of some "first" event/action leading to a domino-like chain of events that have some significant negative (positive) effects and consequences that appear to be ludicrous or unwarranted or with each step in the chain more and more improbable.

CALL

Slogans [C:S]: a brief and striking phrase, often acting like emotional appeals, that may include labeling and stereotyping.
Conversation Killer [A:CK]: words or phrases that discourage critical thought and meaningful discussion about a given topic.
Appeal to Time [C:AT]: the argument is centred around the idea that time has come for a particular action.

MANIPULATIVE WORDING

Loaded Language [MW:LL]: use of specific words and phrases with strong emotional implications (either positive or negative) to influence and convince the audience that an argument is valid.
Obfuscation, Intentional Vagueness, Confusion [MW:OVC]: use of words that are deliberately not clear, vague or ambiguous so that the audience may have its own interpretations.
Exaggeration or Minimisation [MW:EM]: consists of either representing something in an excessive manner or making something seem less important or smaller than it really is.
Repetition [MW:R]: the speaker uses the same phrase repeatedly with the hopes that the repetition will lead to persuade the audience.

Figure 1: **Persuasion techniques in our 2-tier taxonomy.** The six coarse-grained techniques are subdivided into 23 fine-grained ones. An acronym for each technique is given in squared brackets.

[Piskorski et al. 2023]

High-quality annotation guidelines for complex tasks... can get complicated!

A Annotation Guidelines

This appendix provides an excerpt of the annotation guidelines (Piskorski et al., 2023a) related to news genre and persuasion techniques.

A.1 News Genre

- *opinion* versus *reporting*: in the case of news articles that contain citations and opinions of others (i.e., not of the author), the decision whether to label such article as opinion or reporting should in principle depend on what the reader thinks the intent of the author of the article was. In order to make this decision simpler, the following rules were applied:
  - articles that contain even a single sentence (could be even the title) that is an opinion of the author or suggests that the author has some opinion on the specific matter should be labelled as *opinion*,
  - articles containing a speech or an interview with a **single** politician or expert, who provides her/his opinions should be labelled as *opinion*,
  - articles that "report" what a **single** politician or expert said in an interview, conference, debate, etc. should be labelled as *opinion* as well,
  - articles that provide a comprehensive overview (spectrum) of what many different politicians and experts said on a specific matter (e.g., in a debate), including their opinions, and without any opinion of the author, should be labelled as *reporting*,
  - articles that provide a comprehensive overview (spectrum) of what many different politicians and experts said on a specific matter (e.g., in a debate), including their opinions, and with some opinion or analysis of the author (the author might try to tell a story), should be labelled as *opinion* ,
  - commentaries and analysis articles should be labelled as *opinion*.
- *satire*: A news article that contains some small text fragment, e.g., a sentence, which appears satirical **is not supposed to be annotated as** *satire*.

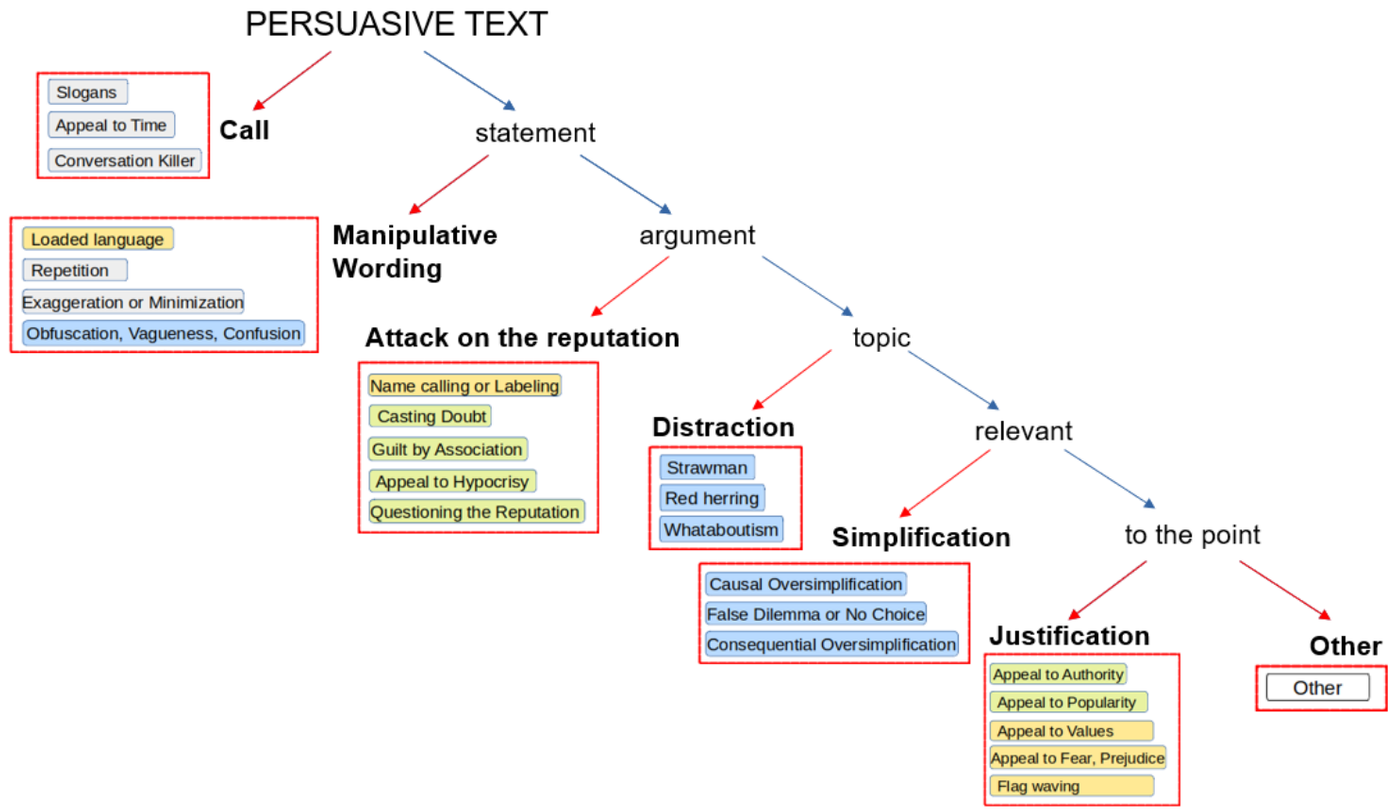


Figure 4: Decision diagram to determine which high-level approach is used in a text. The fine-grained techniques are marked in color, in an attempt to reflect the rhetorical dimension: (a) ethos, i.e., appeal to authority (green), (b) logos, i.e., appeal to logic (blue), and (c) pathos, e.e., appeal to emotions (yellow).

[Piskorski et al. 2023

How many labels is enough?

- For training, typically thousands of annotations are necessary for reasonable performance
  - In-context learning can work with <10 examples ("few-shot")
  - ~dozens seems minimum for fine-tuning (?)
- For evaluation, small #s is ok (but watch statistical significance!)
- Exact amounts are difficult to know in advance. Can do a **learning curve** to estimate if more annotations will be useful.

NLU classif. tasks

- Natural language understanding: broad semantic tasks, performed as (bi)text classification/regression
- GLUE benchmark collection (Wang et al., 2018, 2019)

Corpus	|Train|	|Test|	Task	Metrics	Domain
			Single-Sentence Tasks		
CoLA	8.5k	**1k**	acceptability	Matthews corr.	misc.
SST-2	67k	1.8k	sentiment	acc.	movie reviews
			Similarity and Paraphrase Tasks		
MRPC	3.7k	1.7k	paraphrase	acc./F1	news
STS-B	7k	1.4k	sentence similarity	Pearson/Spearman corr.	misc.
QQP	364k	**391k**	paraphrase	acc./F1	social QA questions
			Inference Tasks		
MNLI	393k	**20k**	NLI	matched acc./mismatched acc.	misc.
QNLI	105k	5.4k	QA/NLI	acc.	Wikipedia
RTE	2.5k	3k	NLI	acc.	news, Wikipedia
WNLI	634	**146**	coreference/NLI	acc.	fiction books

Table 1: Task descriptions and statistics. All tasks are single sentence or sentence pair classification, except STS-B, which is a regression task. MNLI has three classes; all other classification tasks have two. Test sets shown in bold use labels that have never been made public in any form.

- Natural Language Inference (NLI)
  - Does the *evidence* sentence entail, or contradict, the *hypothesis*?

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

Table 1: Randomly chosen examples from the development set of our new corpus, shown with their genre labels, their selected gold labels, and the validation labels (abbreviated E, N, C) assigned by individual annotators.

[Williams et al., 2018]

QA tasks

- Question-answering format are the predominant tasks for LLM evaluation
  - Multiple choice or short answer generated responses

Category	Task
Knowledge	ARC-Easy (MC) ARC-Challenge (MC) Jeopardy (EM) MMLU (MC) OpenbookQA (MC) TriviaQA (EM) WikidataQA (EM)
Math & Reasoning	Arithmetic (EM) GSM8K (EM) LSAT-AR (MC) Operators (EM) Repeat-Copy-Logic (EM)
Coding	HumanEval (pass@10) MBPP (pass@10)
Reading Comprehension	BoolQ (MC) CoQA (EM) DROP (EM) HotpotQA (EM) SQuAD (EM)
Commonsense	CommonsenseQA (MC) COPA (MC) PIQA (MC) Winograd (MC) Winogrande (MC)
Language Understanding	HellaSwag (MC) LAMBADA (EM) Language Identification (EM)
String Manipulation	CS Algorithms (EM) CUTE (EM) Dyck-Languages (EM)

[list from Liu et al., COLM 2025]

Are We Done with MMLU?

Aryo Pradipta Gema[1] **Joshua Ong Jun Leang**[1] **Giwon Hong**[1] **Alessio Devoto**[2]
Alberto Carlo Maria Mancino[2,3] **Rohit Saxena**[1] **Xuanli He**[4] **Yu Zhao**[1] **Xiaotang Du**[1]
Mohammad Reza Ghasemi Madani[5] **Claire Barale**[1] **Robert McHardy**[6] **Joshua Harris**[7]
Jean Kaddour[4] **Emile van Krieken**[1] **Pasquale Minervini**[1,8]
[1]University of Edinburgh [2]Sapienza University of Rome [3]Polytechnic University of Bari
[4]University College London [5]University of Trento [6]AssemblyAI
[7]UK Health Security Agency [8]Miniml.AI
{first.last, jong2, p.minervini}@ed.ac.uk
alessio.devoto@uniroma1.it alberto.mancino@poliba.it
mr.ghasemimadani@unitn.it joshua.harris@ukhsa.gov.uk
{xuanli.he, jean.kaddour.20, robert.mchardy.20}@ucl.ac.uk

Abstract

Maybe not. We identify and analyse errors in the popular Massive Multitask Language Understanding (MMLU) benchmark. Even though MMLU is widely adopted, our analysis demonstrates numerous ground truth errors that obscure the true capabilities of LLMs. For example, we find that 57% of the analysed questions in the *Virology* subset contain errors. To address this issue, we introduce a comprehensive framework for identifying dataset errors using a novel error annotation protocol. Then, we create MMLU-Redux, which is a subset of 5,700 manually re-annotated questions across all 57 MMLU subjects. We estimate that 6.49% of MMLU questions contain errors. Using MMLU-Redux, we demonstrate significant discrepancies with the model performance metrics that were originally reported. Our results strongly advocate for re-

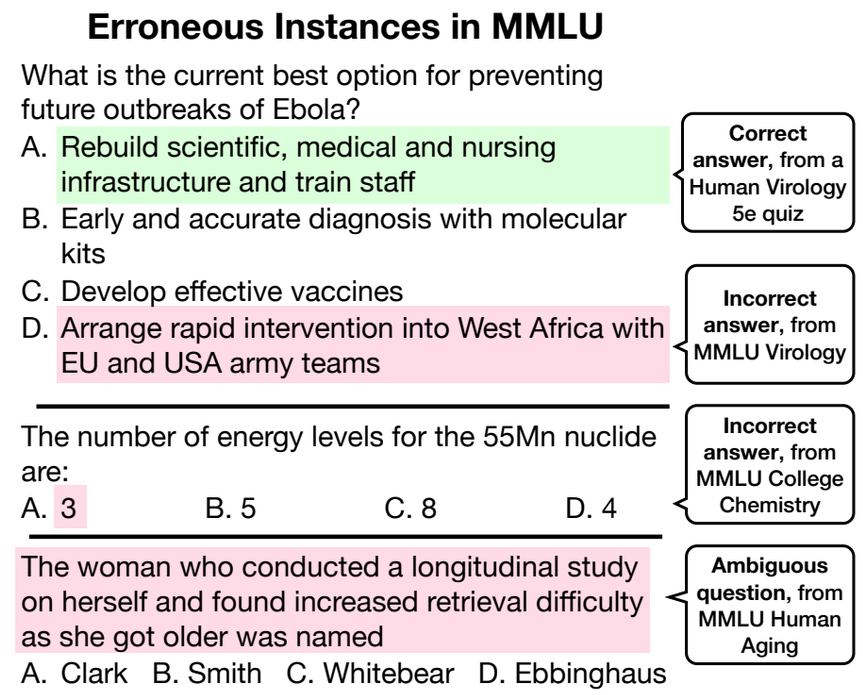


Figure 1: Examples of erroneous instances from MMLU Virology, College Chemistry, and Human Aging.

- Data quality is an ongoing issue! [Gema et al., NAACL 2025]