CS329X: Human Centered NLP

User-Centered Evaluation

Diyi Yang

Stanford CS

Announcements

Literature review due tonight (Apr 24th)

Late Days Policy

1. Literature Review (Apr 24th, 23:59pm PT)

This is a short paper (4~5 pages, excluding references) summarizing and synthesizing several papers in the area of your final project. As noted above, 8 pages is the maximum allowed length. Groups of one should review 5 papers, groups of two should review 7 papers, and groups of three should review 9.

The ideal is to have the same topic for your lit review and final project, but it's possible that you'll discover in the lit review that your topic isn't ideal for you, so you can switch topics (or groups) for the final project; your lit review will be graded on its own terms.

Some suggestion highlights on literature review structure from Chris Potts and Bill MacCartney from CS224U (check out lots of useful material there and there):

1. *General problem/task definition*: What are these papers trying to solve, and why?
2. *Concise summaries of the articles*: Do not simply copy the article text in full. We can read them ourselves. Put in your own words the major contributions of each article.
3. *Compare and contrast*: Point out the similarities and differences of the papers. Do they agree with each other? Are results seemingly in conflict? If the papers address different subtasks, how are they related? (If they are not related, then you may have made poor choices for a lit review...). *This section is probably the most valuable for the final project, as it can become the basis for a literature review section.*.
4. *Future work*: Make several suggestions for how the work can be extended. Are there open questions to answer? How do the papers relate to your final project idea?
5. *References section*: The entries should appear alphabetically and give at least full author name(s), year of publication, title, and outlet if applicable (e.g., journal name or proceedings name). Beyond that, we are not picky about the format. Electronic references are fine but need to include the above information in addition to the link.

Announcements

Literature review due tonight (Apr 24th)

Late Days Policy

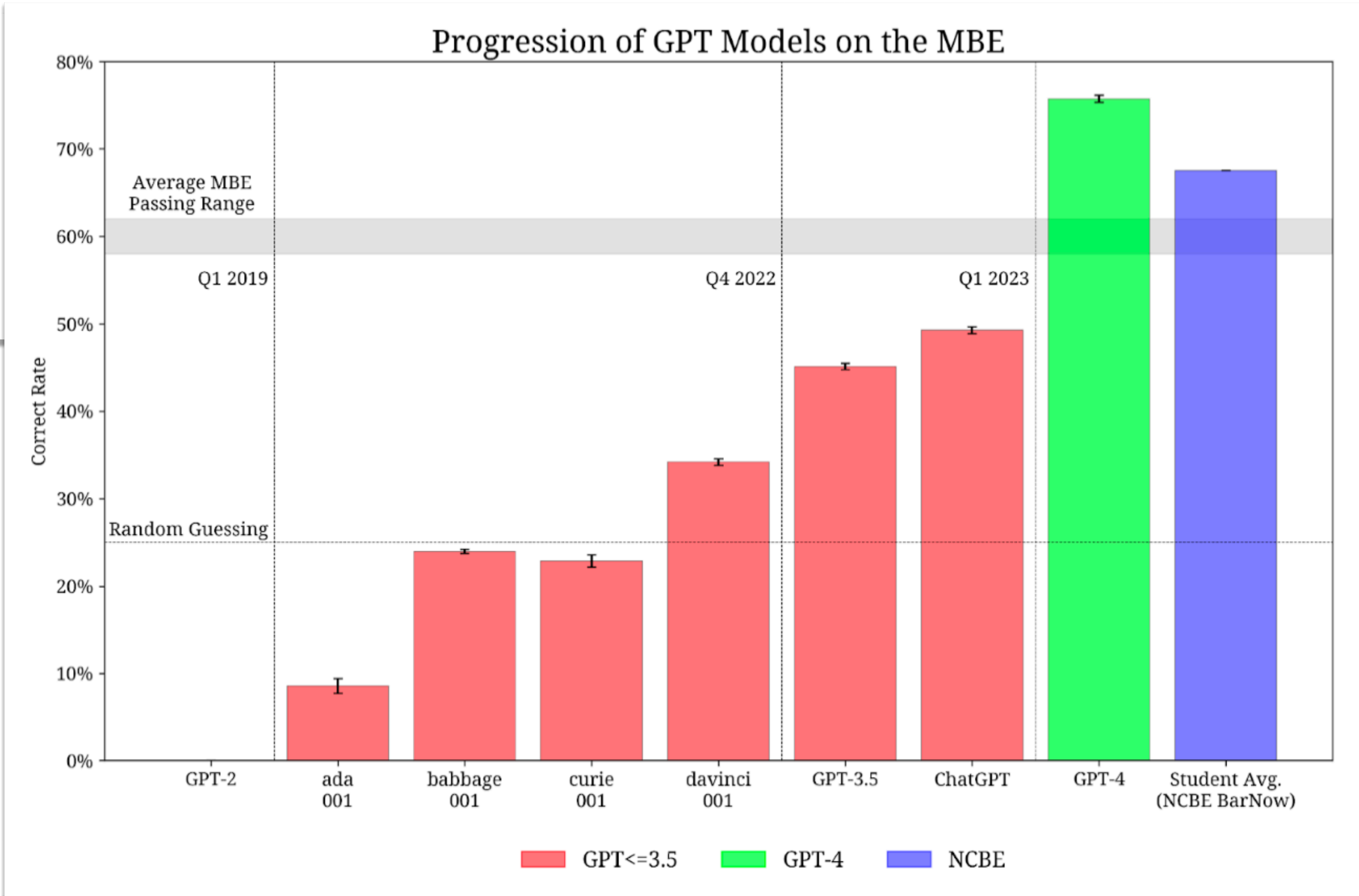
Late days will be automatically used for any late submissions (e.g., hw, scribe, project)

Stop by **Office Hour** for any discussion/chat on course project!

Overview

- Why do we need (human) evaluation?
- Consideration before human evaluation
- Designing human evaluations
- Framing biases in user-centric evaluation
- Today's Challenges

GPT-4 Passes the Bar Exam



Can we really detect AI-generated text?

Nintendo Switch game console to launch in March for $299 The Nintendo Switch video game console will sell for about $260 in Japan, starting March 3, the same date as its global rollout in the U.S. and Europe. The Japanese company promises the device will be packed with fun features of all its past machines and more. Nintendo is promising a more immersive, interactive experience with the Switch, including online playing and using the remote controller in games that don't require players to be constantly staring at a display.

New Nintendo Switch game console to launch in March for $99 Nintendo plans a promotional roll out of it's new Nintendo switch game console. For a limited time, the console will roll out for an introductory price of $99. Nintendo promises to pack the new console with fun features not present in past machines. The new console contains new features such as motion detectors and immerse and interactive gaming. The new introductory price will be available for two months to show the public the new advances in gaming.

Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. "Automatic detection of fake news." arXiv preprint arXiv:1708.07104 (2017).

Can we really detect AI-generated text?

Kim And Kanye Silence Divorce Rumors With Family Photo. Kanye took to Twitter on Tuesday to share a photo of his family, simply writing, "Happy Holidays." In the picture, seemingly taken at Kris Jenner's annual Christmas Eve party, Kim and a newly blond Kanye pose with their children, North, 3, and Saint, 1. After Kanyes hospitalization, reports that there was trouble in paradise with Kim started brewing. But E! News shut down the speculation with a family source denying the rumors and telling the site, "It's been a very hard couple of months."

Kim Kardashian Reportedly Cheating With Marquette King as She Gears up for Divorce From Kanye West. Kim Kardashian is ready to file for divorce from Kanye West but has she REALLY been cheating on him with Oakland Raiders punter Marquette King? The NFL star seemingly took to Twitter to address rumors that they've been getting close amid Kanye's mental breakdown, which were originally started by sports blogger Terez Owens. While he doesn't appear to confirm or deny an affair, her reps said there is "no truth whatsoever" to the reports and labeled the situation "fabricated."

Pérez-Rosas, Verónica, Bennett Kleinberg, Alexandra Lefevre, and Rada Mihalcea. "Automatic detection of fake news." arXiv preprint arXiv:1708.07104 (2017).

What does evaluation mean?

The process of assessing the performance and effectiveness of NLP models, algorithms, and applications

The value of Evaluation

- Helps researchers and developers identify the strengths and weaknesses of their algorithms and make improvements to them.
- Comparing different models and selecting the best one for a given task.
- Intrinsic evaluation vs. Extrinsic evaluation

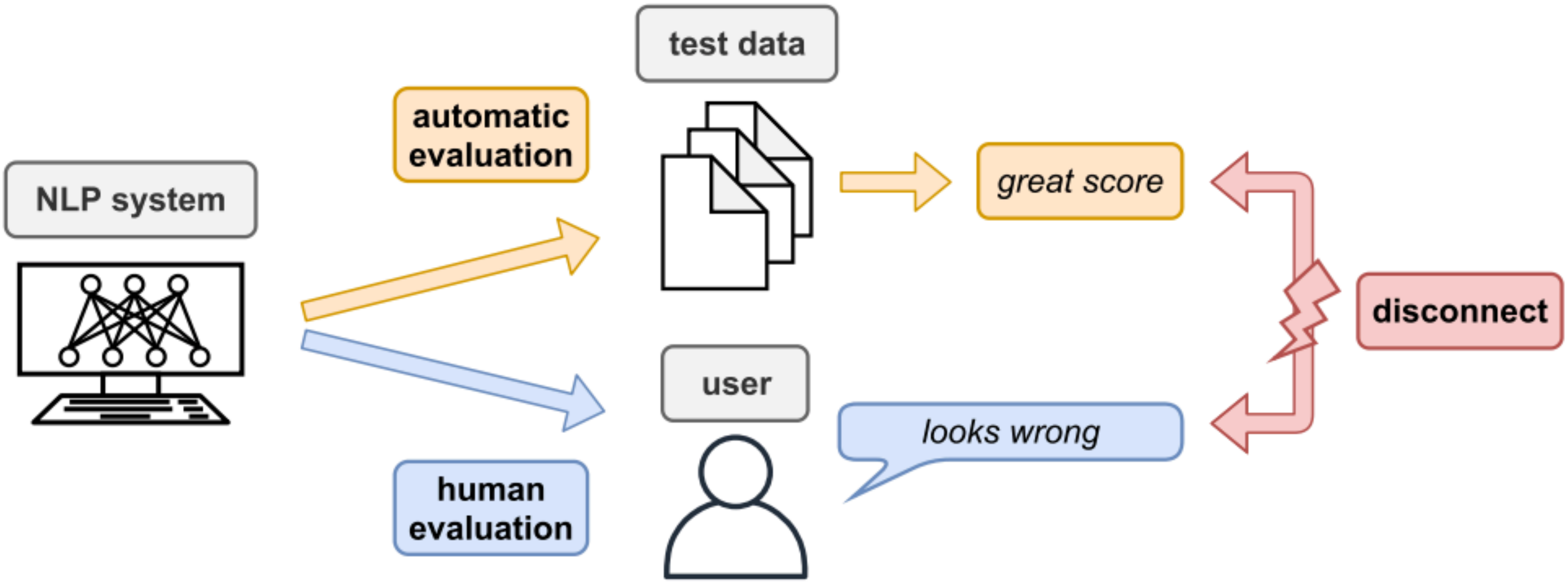
Automatic Evaluation

BLEU scores (n-gram overlap) are commonly used to quantify translation/generation quality between a hypothesis and the ground-truth.

Shortcomings:

1. Relying on ground-truth reference(s) and ignores the breadth of possible correct translations
2. Assuming that similarity of meaning can be inferred from n-gram overlap

The Need for Human Evaluation



Relying on automatic evaluation alone (e.g., via accuracy, F1 or BLEU scores) can be misleading as good performance with respect to scores does not imply good performance with respect to human evaluation.

Schuff, Hendrik, Lindsey Vanderlyn, Heike Adel, and Ngoc Thang Vu. "How to do human evaluation: A brief introduction to user studies in NLP." Natural Language Engineering (2023): 1-24.

Considerations before human evaluation

Ethical and Legal Considerations

When designing an experiment involving human participation, it is critical to consider ethical and legal implications

Critical to understand which review processes or legal requirements exist

- Institutional review boards
- Ethics committee
- Relevant data collection laws

Ethical and Legal Considerations: Privacy

What data are actually necessary to collect?

How the data will be stored and protected?

How long?

What type of personal data will be collected?

Data collection and Anonymization techniques [Siegert et al. (2020); Finck and Pallas (2020)]

Ethical and Legal Considerations: Informed Consent

Make sure participants have true informed consent before an experiment

[Nuremberg Code 1949, APA Ethical Principles and Code of Conduct 2002, EU Data Protection Regulation 2018]

1. The purpose of the research
2. That they have the right to end participation at any time
3. The potential risks an experiment poses why someone may not want to participate
4. Prospective benefits of the experiment
5. Any limits to confidentiality, such as how the data collected will be used
6. Incentives for participation
7. Who to contact in case of questions

Ethical and Legal Considerations: Respect for Participants

Prioritize the dignity of participants

Studies should be conducted to provide a benefit to society, but participant welfare must take a priority over the interests of science and society

Avoid all unnecessary physical and mental suffering and injury, especially when working with vulnerable populations

- e.g., interacting with chatbots under high-stress conditions

Designing human evaluation

The Purpose of Human Evaluation

Exploratory research questions: to generate assumptions, which can then be tested in a subsequent confirmatory research question, e.g., *"Which factors (of the set of measured variables) influence the users' enjoyment of system B?"*

Confirmatory research questions: to test a specific assumption, e.g., *"Does the explanation method of system B increase the users' trust in the system compared to that of system A?"*

Transparency in Human Evaluation

No standardized approach or consensus for human evaluation

Different to compare results across different studies due to the variability in evaluation design

Where Human Evaluations Are Needed?

Evaluation of model quality

What do people think about the output from an NLP model?

Develop automatic metrics

Dataset for testing the correlation of automatic metrics with human evaluations (e.g., WMT datasets)

Training data to directly optimize metrics to predict human evaluations

Incorporate human evaluations directly into NLP models

e.g., GPT's use of reinforcement learning from human feedback

Best Practices for Designing Human Evaluation

How are human ratings collected?

What questions are asked of raters?

Who are the raters?

How do you ensure/measure the quality of the ratings?

Celikyilmaz, Asli, Elizabeth Clark, and Jianfeng Gao. "Evaluation of text generation: A survey." arXiv preprint arXiv:2006.14799 (2020).

Intrinsic vs. Extrinsic Evaluation

Intrinsic Evaluation

Read and rate the quality of a generated text

Pros: easier to run, can focus on subtasks

Example: rate suggestions from a spell checker on a scale from 1 to 5

Extrinsic Evaluation

Measure how successful a system is in a downstream task

Pros: most realistic evaluation, full system evaluation

Example: how many spelling errors does a user makes when writing with a spell checker

Types of Human Feedback

Ways to rate a generated text:

- Mark as good or bad
- Rate on a scale from 1 to 5
- Assign a score 1-100
- Decide whether it’s better than another text
- Rank its relative to other texts

Metrics to Use

Likert scales

Using multiple items instead of a single rating allows one to assess the scale's internal consistency

Reliable scale requires a precise development process

Validated questionnaire exists, e.g., for evaluating trust (Körber 2018), usability (Brooke 1996; Finstad 2010), cognitive load (Hart and Staveland 1988), social attribution (Carpinella et al. 2017), or user interface language quality (Bargas-Avila and Brühlmann 2016).

What if designing and applying Likert scales that have not been validated?

Other Useful Metrics for NLP

Continuous rating scales like the visual analog scales (VAS)

Continuous rating scales can yield more consistent results than Likert scale for dialog system evaluation (Santhanam and Shaikh, 2019)

Direct comparisons or ranked order comparisons (ranked output from multiple systems best to worst) (Vilar et al. 2007; Bojar et al. 2016)

Error classification: annotating text output from a set of predefined error labels

Completion time and bio-signals, such as gaze, EEG, and electrodermal activity

E.g., emotional state (Kim and André 2008), engagement (Renshaw, Stevens, and Denton 2009), stress (McDuff et al. 2016), and user uncertainty (Greis et al. 2017).

Qualitative Analysis

Qualitative analysis via **free form of expression**

E.g., free response questions to understand users' perception of chatbots

Such responses can then be analyzed with techniques such as content/ theme analysis, where users' responses are coded to find similar themes

In-depth semi-structured/structured **interviews** (see Design thinking slides)

Dimensions of Text Quality

Is the text …?

- Grammatical
- Fluent
- Coherent
- Creative
- Surprising
- Entertaining

Howcroft, David, Anya Belz, Miruna Clinciu, Dimitra Gkatzia, Sadid A. Hasan, Saad Mahamood, Simon Mille, Emiel Van Miltenburg, Sashank Santhanam, and Verena Rieser. "Twenty Years of Confusion in Human Evaluation: NLG Needs Evaluation Sheets and Standardised Definition." Association for Computational Linguistics (ACL), 2020.

Criterion Paraphrase	Count
usefulness for task/information need	39
grammaticality	39
quality of outputs	35
understandability	30
correctness of outputs relative to input (content)	29
goodness of outputs relative to input (content)	27
clarity	17
fluency	17
goodness of outputs in their own right	14
readability	14
information content of outputs	14
goodness of outputs in their own right (both form and content)	13
referent resolvability	11
usefulness (nonspecific)	11
appropriateness (content)	10
naturalness	10
user satisfaction	10
wellorderedness	10
correctness of outputs in their own right (form)	9
correctness of outputs relative to external frame of reference (content)	8
ease of communication	7
humanlikeness	7
appropriateness	6
understandability	6
nonredundancy (content)	6
goodness of outputs relative to system use	5
appropriateness (both form and content)	5

Original Criterion	Mapped to normalised criteria	Count
fluency	fluency; goodness of outputs in their own right; goodness of outputs in their own right (form); goodness of outputs in their own right (both form and content; grammaticality; humanlikeness); readability; [*multiple (3)*: goodness of outputs in their own right (both form and content), grammaticality, naturalness (form)]; [*multiple (2)*: goodness of outputs in their own right (form), grammaticality]; [*multiple (3)*: fluency, grammaticality]; [*multiple (2)*: grammaticality, readability]; [*multiple (2)*: fluency, readability]; [*multiple (3)*: goodness of outputs in their own right (both form and content), grammaticality, naturalness (form)]; [*multiple (3)*: coherence, humanlikeness, quality of outputs]; [*multiple (2)*: goodness of outputs in their own right (both form and content), grammaticality]	15
readability	fluency; goodness of outputs in their own right; goodness of outputs in their own right (both form and content); quality of outputs; usefulness for task/information need; readability; [*multiple (2)*: coherence, fluency]; [*multiple (2)*: fluency, readability]; [*multiple (2)*: readability, understandability]; [*multiple (3)*: clarity, correctness of outputs in their own right (form), goodness of outputs in their own right]	10
coherence	appropriateness (content); coherence; correctness of outputs in their own right (content); goodness of outputs in their own right (content); goodness of outputs relative to linguistic context in which they are read/heard; wellorderedness; [*multiple (2)*: appropriateness (content), understandability]; [*multiple (2)*: fluency, grammaticality]	8
naturalness	clarity; humanlikeness; naturalness; naturalness (both form and content); [*multiple (2)*: naturalness (both form and content), readability]; [*multiple (2)*: grammaticality, naturalness]	6
quality	goodness of outputs in their own right; goodness of outputs in their own right (both form and content); goodness of outputs (excluding correctness); quality of outputs; [*multiple (3)*: correctness of outputs relative to input (content), Fluency, Grammaticality]	5
correctness	appropriateness (content); correctness of outputs relative to input (content); correctness of outputs relative to input (both form and content); correctness of outputs relative to input (form)	4
usability	clarity; quality of outputs; usefulness for task/information need; user satisfaction	4
clarity	clarity; correctness of outputs relative to input (content); understandability; [*multiple (2)*: clarity, understandability]	4
informativeness	correctness of outputs relative to input (content); goodness of outputs relative to input (content); information content of outputs; text property (informative)	4
accuracy	correctness of outputs relative to input; correctness of outputs relative to input (content); goodness of outputs relative to input (content); referent resolvability	4

Participants

Are the participants in the human evaluation ….?

- Experts?
- In-person?
- Crowdsourced?
- Paid?
- Trained?
- Quality-controlled?

Ensuring Annotator Quality

Annotator instructions and training

How to define and explain the task to evaluators?

Attention checks/questions with known answers

E.g., intentionally corrupted generated text

Annotator agreement

% agreement, Cohen's K, Krippendorff's α

Who is doing the measuring?

Low quality of crowdsourced annotations in NLP may be in part due to the quality of the task. Huynh et al., (2021) found that:

- 25% of NLP studies on MTurk have technical issues
- 28% have flawed or insufficient instructions
- 26% of study creators were rated as having poor communication

Poor working conditions for raters may also lead to low quality data and incorrectly incentivized evaluators

- 35% of requesters pay poorly or very badly
- Only 14 of 703 NLP papers that used crowdsourcing mention IRB review

Crowdsourcing for NLP

- Fair compensation
- Platform rules
- Incentives and response quality
- Pilot study: Pilot studies, that is, small-scale trials before a larger study, allow for testing the experimental design and technical setup
- Data collection

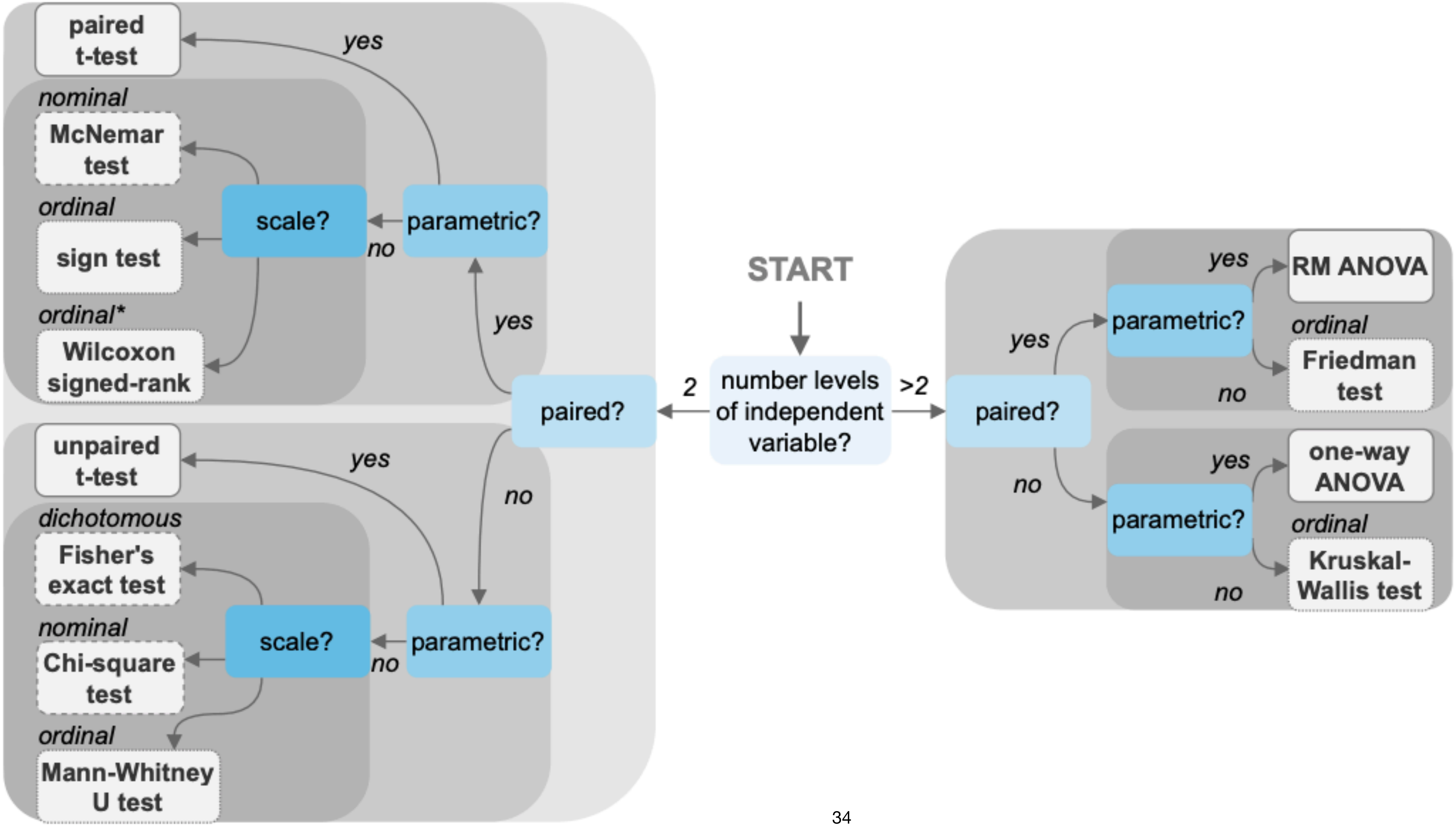
Statistical Evaluation for NLP

Only 33% of NLP papers that conduct a human evaluation report statistical analyses - van der Lee et al. (2019)

Key design choices:

- ✴ Estimating the required sample size
- ✴ Selecting an applicable statistical test
- ✴ Deciding whether a post hoc get and multiplicity adjustment is needed

Choosing the Correct Statistical Test



Choosing the Correct Statistical Test

Paired and unpaired tests:

A paired test: samples were collected in a within-subject design.

An unpaired test: samples were collected in a between-subjects design from different groups

Parametric and non-parametric tests:

Parametric tests make assumptions on the underlying population distribution (such as normality), and non-parametric tests do not make assumptions on the distributions

More complex models and tests

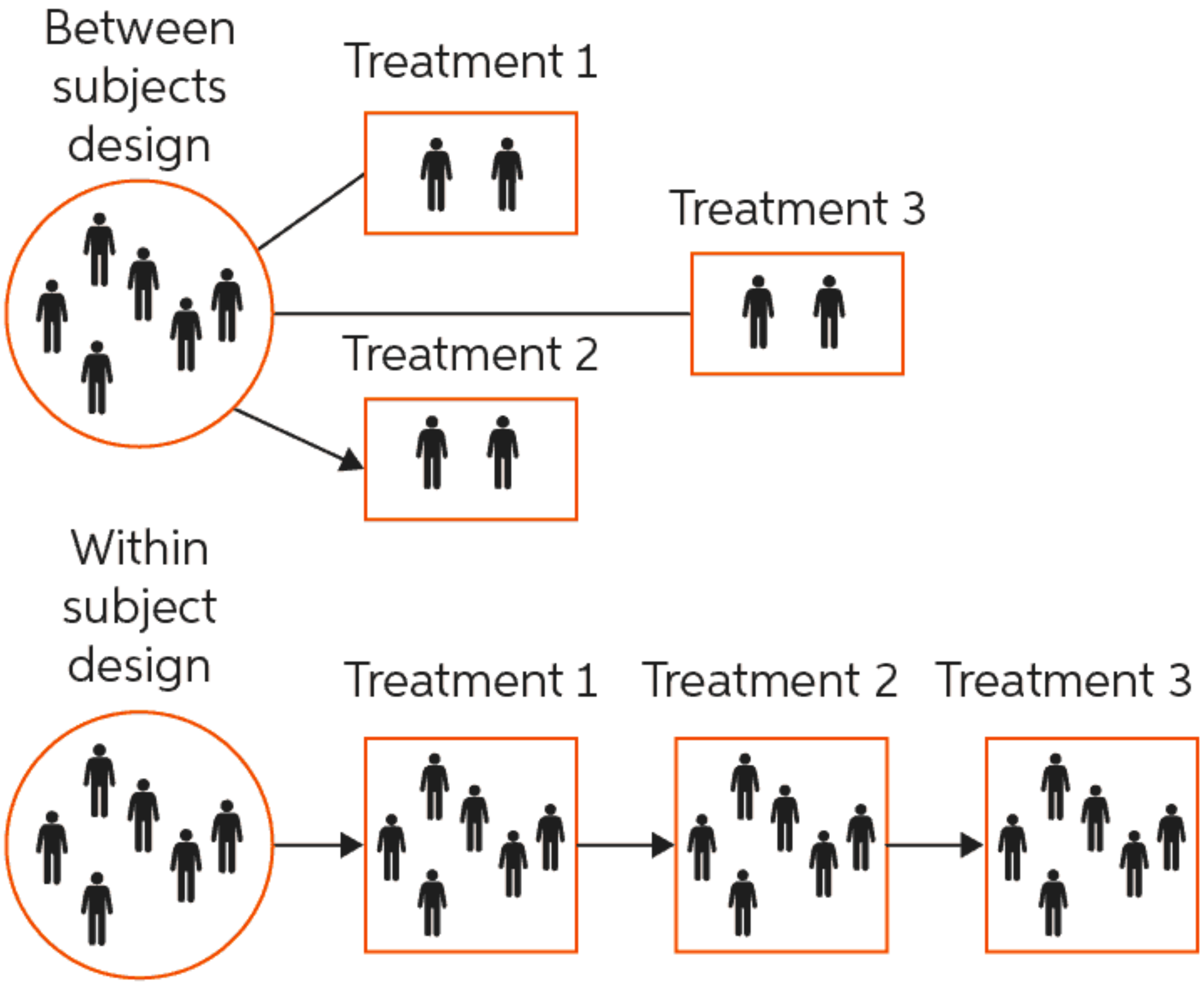
Generalized linear models

Generalized linear mixed models: to include random effects such as individual characteristics

Experimental Designs

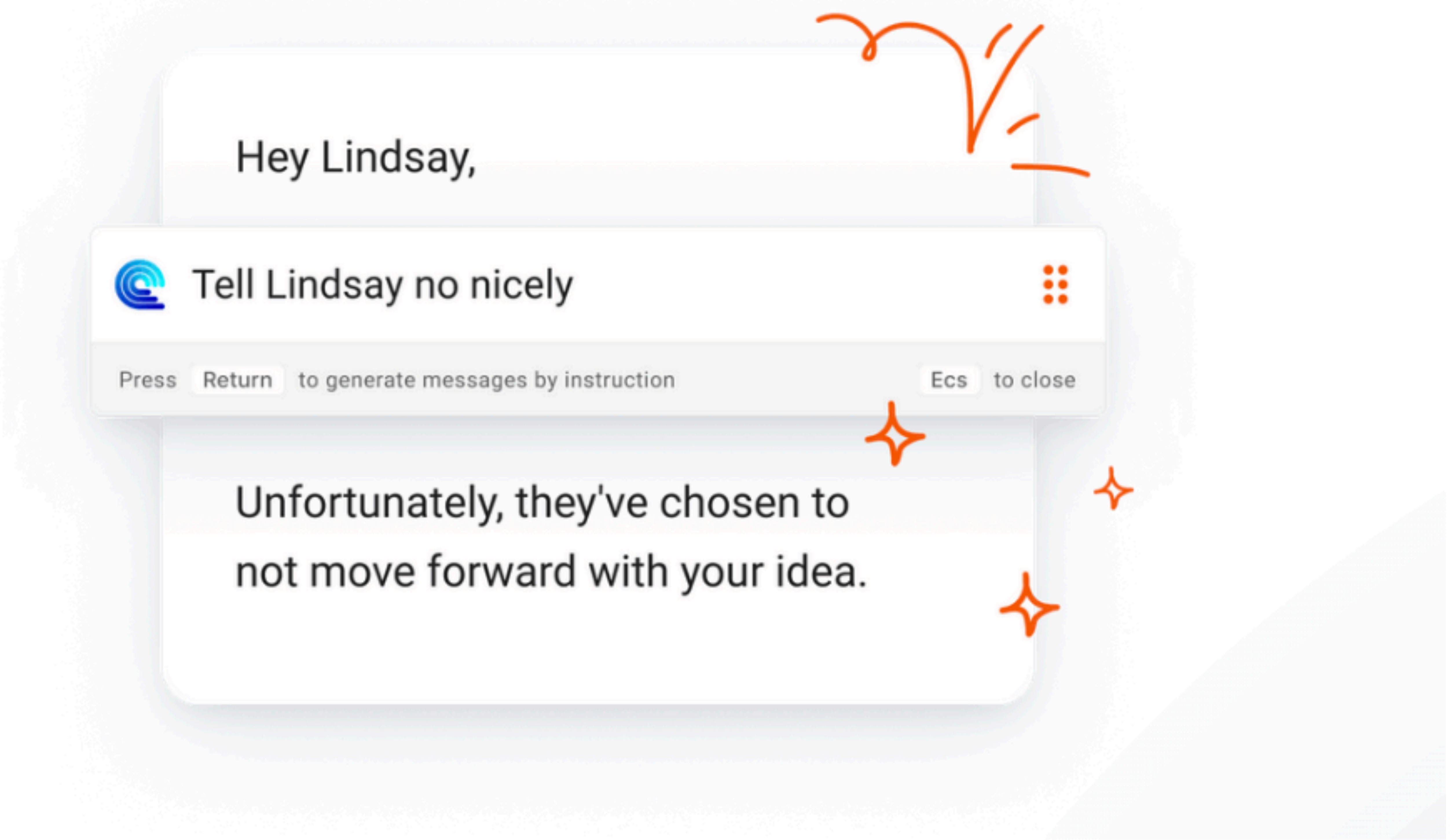
Key question: how participants are assigned to conditions

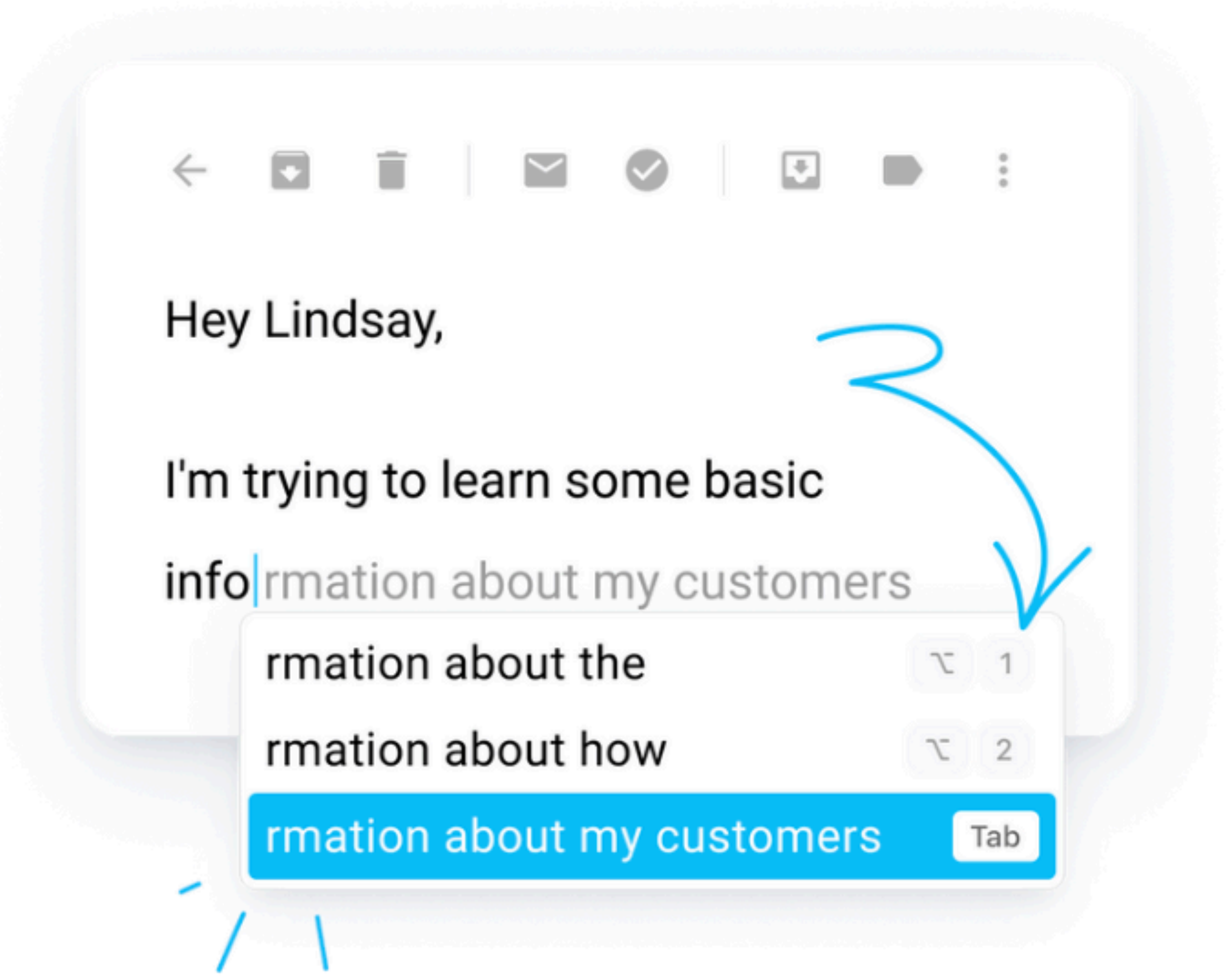
- Between subjects design
- Within subject design



https://www.chegg.com/writing/guides/research/within-subject-design/

Example:
Evaluating whether an AI-based email writer is useful





Comparing Within-Subject vs. Between-Subject Design

	Within-Subject	**Between-Subject**
Pros	• Small sample size • Minimizes variance between conditions • Statistically robust	• Easy to conduct • No chances of contamination across treatment groups
Cons	• Carryover effect • Time-related threats	• Require more participants • Results may be confounded if groups are not equated by randomization • Difficult to match participants

https://www.chegg.com/writing/guides/research/within-subject-design/
https://www.chegg.com/writing/guides/research/between-subjects-design/

The Biggest Problems

Problem #1: lack of human evaluations in NLG work

Problem #2: even when there are human evals, they are under-documented

Lack of documentation is bad for:

- Interpretability
- Replicability
- Comparisons to other work

73% of surveyed NLG papers include a human evaluation. Of those papers, only 58% specified who the participants in the study were.

Framing Effects and Cognitive Biases

Framing refers to how something is asked as opposed to what is asked.

In human evaluation for NLG, framing could be reflected in **question wording or instructions** provided to participants

Schoch, Stephanie, Diyi Yang, and Yangfeng Ji. "“This is a Problem, Don’t You Agree?” Framing and Bias in Human Evaluation for Natural Language Generation." In Proceedings of the 1st Workshop on Evaluating NLG Evaluation, pp. 10-16. 2020.

Positive and Negative Reframing

```
How much more fluent is sentence A versus sentence B?
```

Framing demonstrated that people are **more likely to make** choices that are framed positively (in terms of ***gains***) as opposed to negatively (in terms of ***losses***) due to the increased perceived risk associated with losses.

Demand Characteristics

`A researcher has developed style transfer model A to generate formal sentences, and is evaluating sentence A from their generative model against sentence B from a baseline model. Unconsciously aware of model A’s artifacts, in this case, as a system that only uses “.” as end punctuation, the researcher states ‘We consider sentences that end with “.” as more formal than sentences that end with “!”’ in the task description.`

Demand characteristics are response biases that refer to cues in a study design that may reveal a researcher’s hypothesis to the participants

Human Evaluation Design Statements

When describing human evaluation design setup:

Question design: types, scales, wording

Question presentation: ordering, questions per annotator

Target criteria: definitions

Annotators: demographics, background, recruitment, compensation

When reporting evaluation results, explain what you did, why you did it, and possible shortcomings

Todays' Challenges

Text generation models have improved, and generated text is more fluent and higher quality than ever before

Crowdsourced evaluations are increasingly common - is this enough today?

The easiest evaluation is not always the best evaluation.

GPTs are GPTs: An Early Look at the Labor Market Impact Potential of Large Language Models

Tyna Eloundou[1], Sam Manning[1,2], Pamela Mishkin*[1], a

[1]OpenAI
[2]OpenResearch
[3]University of Pennsylvania

March 27, 2023

Is GPT-3 a Good Data Annotator?

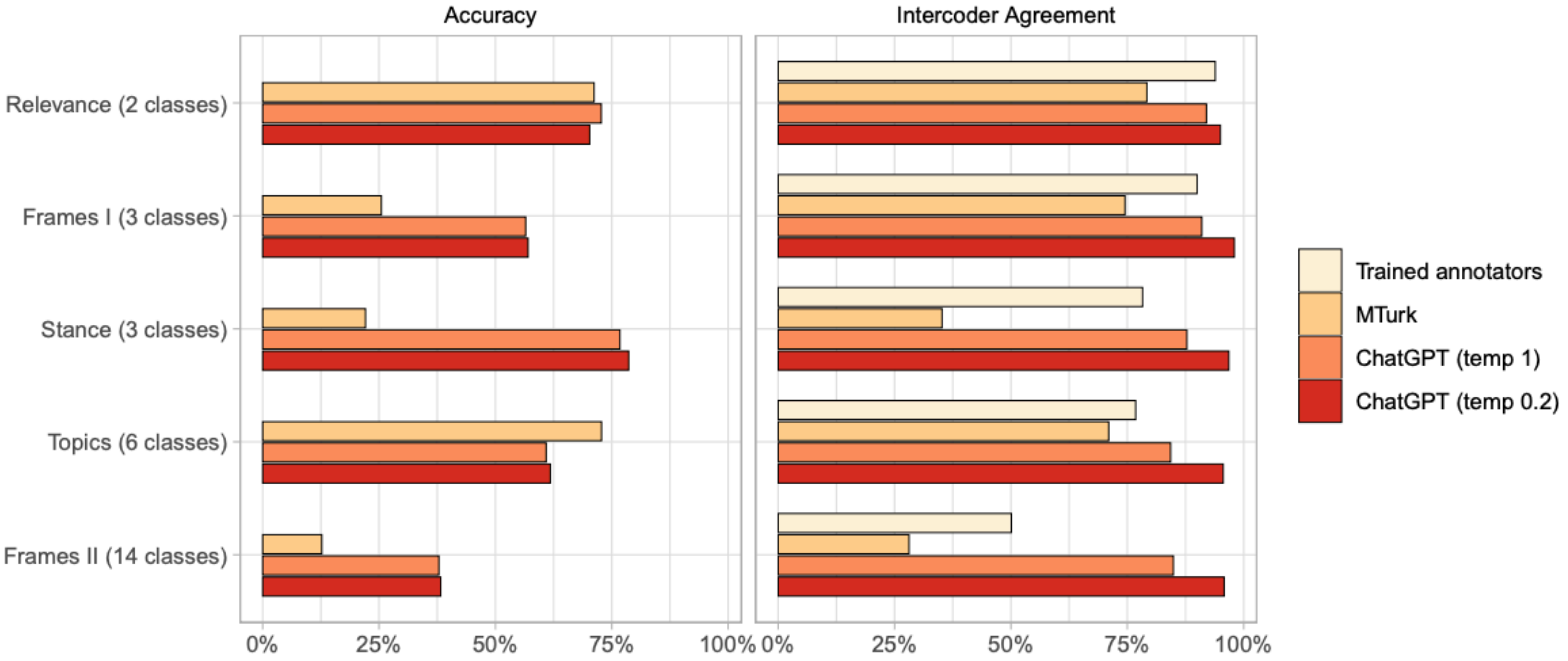
Bosheng Ding*[12] Chengwei Qin*[1] Linlin Liu†[1,2]
Lidong Bing[2] Shafiq Joty[1] Boyang Li[1]
[1]Nanyang Technological University, Singapore [2]DAMO Academy, Alibaba Group
{bosheng001, chengwei003, linlin001, srjoty, boyang.li}@ntu.edu.sg
{bosheng.ding, l.bing}@alibaba-inc.com

ChatGPT Outperforms Crowd-Workers for Text-Annotation Tasks*

Fabrizio Gilardi† Meysam Alizadeh‡ Maël Kubli§

March 28, 2023

ChatGPT zero-shot text annotation performance, compared to MTurk and trained annotators. ChatGPT's accuracy outperforms that of MTurk for four of the five tasks. ChatGPT's intercoder agreement outperforms that of both MTurk and trained annotators in all tasks.

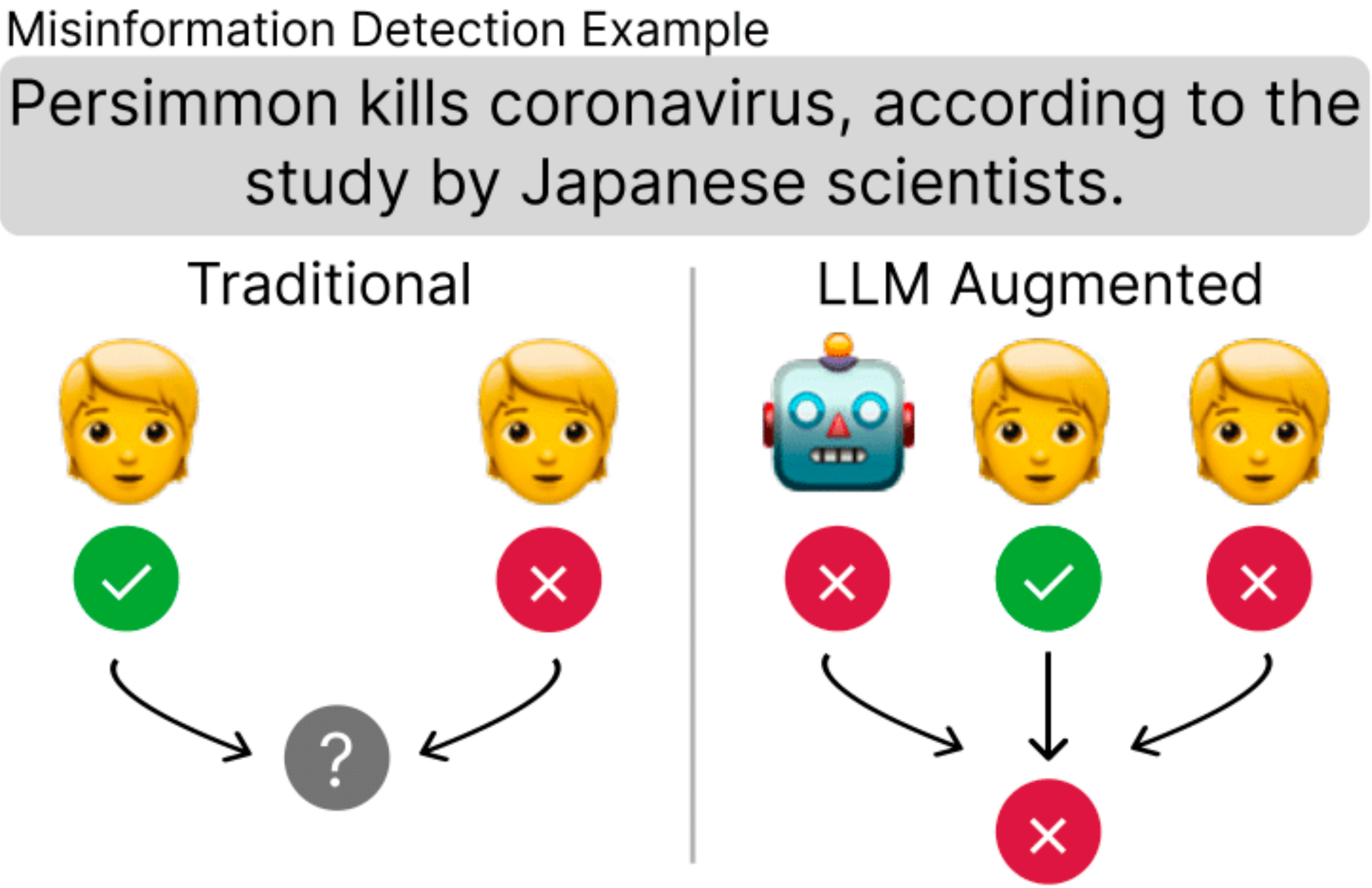


Dataset	Best Model	Acc.	κ	Agreement
Utterance-Level				
Dialect	flan-ul2	23.7	0.15	poor
Emotion	flan-ul2	**70.3**	0.64	good
Figurative	flan-ul2	64.0	0.52	moderate
Humor	flan-t5-xl	59.0	0.16	poor
Ideology	davinci-002	57.6	0.36	fair
Impl. Hate	flan-ul2	36.3	0.23	fair
Misinfo	flan-ul2	**77.6**	0.55	moderate
Persuasion	flan-t5-xxl	51.6	0.42	moderate
Semantic Chng.	flan-t5-large	66.9	0.34	fair
Stance	chatgpt	**72.0**	0.58	moderate
Convo-Level				
Discourse	flan-t5-xxl	52.5	0.44	moderate
Empathy	flan-ul2	39.8	0.04	poor
Persuasion	flan-t5-large	57.1	0.13	poor
Politeness	flan-t5-xl	59.2	0.38	fair
Power	chatgpt	61.6	0.23	fair
Toxicity	flan-ul2	56.6	0.01	poor
Document-Level				
Ideology	chatgpt	58.8	0.36	fair

Table 3: *(Acc.)* **Best model accuracy.** Accuracies above 70% are bolded as high enough for possible downstream use. *(κ)* **Agreement scores between zero-shot model classification and human gold labels.** Out of ten utterance-level tasks, five have at least moderate M and only two have poor agreement P. Three (50%) of the conversation tasks have at least fair agreement F, as does the document-level task.

Can Large Language Models Transform Computational Social Science?

Caleb Ziems* **William Held*** **Omar Shaikh*** **Jiaao Chen***
Zhehao Zhang* **Diyi Yang***
Georgia Institute of Technology, Shanghai Jiao Tong University, Stanford University
{cziems, wheld3, jiaaochen}@gatech.edu, zzh12138@sjtu.edu.cn, {oshaikh, diyiy}@stanford.edu



Moving Forward

Who is in a better position to perform evaluation?

What aspects should we look at to “evaluate” an AI model?

Beyond accuracy and performance, how should we evaluate risk, harms, and safety associated with AI models?

Fireside Chat with Mina Lee

[Optional for Homework 1]
Deep Dive into One Behavioral Evaluation

Tulio Ribeiro, Marco, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. "Beyond Accuracy: Behavioral Testing of NLP models with CheckList." arXiv e-prints (2020): arXiv-2005.

Slides credit to Marco and Tongshuang!

Software engineering → NLP

Capabilities	Descriptions
Vocab/POS	important words or word types for the task.
Named entities	appropriately understanding named entities.
Nagation	understand the negation words.
Taxonomy	synonyms, antonyms, etc.
Robustness	to typos, irrelevant changes, etc.
Coreference	resolve ambiguous pronouns, etc.
Fairness	not biasing towards certain gender/race groups.
Semantic Role Labeling	understanding roles such as agent, object, etc.
Logic	handle symmetry, consistency, and conjunctions.
Temporal	understand order of events.

Principle: test small units

↓

What to test: capabilities

Why do we have the universal list?

Models' capabilities are task-independent.

Models' expected behaviors w.r.t capabilities are task-dependent.

This is not an exhaustive list!

Software engineering → NLP

Capabilities
Vocab/POS
Named entities
Nagation
…

Behavioral testing: decouple tests from implementation

↓

Decouple tests from training

Meets users' needs
Works with black box models

Software engineering → NLP

Capabilities			
Vocab/POS			
Named entities			
Nagation			
...			

Behavioral testing: decouple tests from implementation

↓

Decouple tests from training

How to test:
Test behaviors with different test types!

Illustrating task: **sentiment analysis** with **Google Cloud's Natural Language**

Software engineering → NLP

Capabilities	MFT		
Vocab/POS			
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Software engineering → NLP

Capabilities	MFT		
Vocab/POS			
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

} *A group of n=500 test cases*

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15%		
Named entities			
Nagation			
…			

← *1 test, with failure rate*

Unit tests: known in-/out-puts

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

A group of n=500 test cases

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15%		
Named entities			
Nagation			
...			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

Expectation: Exact labels

This is a commercial flight. (neutral)

I flew to Indiana yesterday. (neutral)

Software engineering → NLP

Capabilities	**MFT**		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%	← *multiple tests per cell*	
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

This was a great flight. (positive)

I hated this seat. (negative)

Expectation: Exact labels

This is a commercial flight. (neutral)

I flew to Indiana yesterday. (neutral)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation			
…			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

The cabin crew was not great. (negative)

I can't say I enjoyed the food. (negative)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
...			

Unit tests: known in-/out-puts

↓

Minimum **F**unctionality **T**est

Expectation: Exact labels

The cabin crew was not great. (negative)

I can't say I enjoyed the food. (negative)

Software engineering → NLP

Capabilities	MFT		
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

~~Start from scratch~~ → Perturb existing ones

~~Expect exact label~~ → Expect predictions to (not) change

Software engineering → NLP

Capabilities	MFT	INV	
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

INVariance Tests

Software engineering → NLP

Capabilities	MFT	INV	
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities			
Nagation	Easy: 49.2%		
...			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

No need to specify the exact prediction!

Expectation: Same prediction after the change.

@AmericanAir thank you we got on a different flight to ~~Chicago~~ Dallas.

@VirginAmerica I can't lose my luggage, moving to ~~Brazil~~ Turkey soon.

Software engineering → NLP

Capabilities	MFT	INV	
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

No need to specify the exact prediction!

Expectation: Same prediction after the change.

@AmericanAir thank you we got on a different flight to ~~Chicago~~ Dallas.

@VirginAmerica I can't lose my luggage, moving to ~~Brazil~~ Turkey soon.

Software engineering → NLP

Capabilities	MFT	INV	DIR
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

DIRectional Expectation Tests

Software engineering → NLP

Capabilities	MFT	INV	DIR
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
...			

Metamorphic (perturbations) & property-based testing

INVariance Tests

DIRectional Expectation Tests

Expectation: Sentiment monotonic decreasing (↓)

@AmericanAir service wasn't great. You are lame.

@JetBlue why won't YOU help them?! Ugh. I dread you.

expectation on probability!

Software engineering → NLP

Capabilities	MFT	INV	DIR
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%		Add neg: 34.6%
Named entities		LOC: 21%	
Nagation	Easy: 49.2%		
…			

Metamorphic (perturbations) & property-based testing

↓

INVariance Tests

DIRectional Expectation Tests

Expectation: Sentiment monotonic decreasing (↓)

@AmericanAir service wasn't great. You are lame.

@JetBlue why won't YOU help them?! Ugh. I dread you.

expectation on probability!

Software engineering → NLP

For NLG tasks, expectations are much more nuanced: e.g., "Before and after translation the entity should stay the same."

			Input	Translation	Back translation
Clear Positives		en-pt-en	The icing on this cake is delicious	→ A cereja deste bolo é deliciosa	→ The cherry of this cake is delicious
		G en-pt-en	This burger is out of the park!	→ Este hambúrguer está fora de moda!	→ This burger is out of style!
		en-zh-en	Don't forget to try the beans, they are delicious!	→ 别忘了试试豆子，它们很好吃！	→ Don't forget to try beans. ~~they are delicious!~~
Negation		en-pt-en	I can't say anything else than this place sucks.	→ Não posso dizer nada além desse lugar.	→ I can't say anything but this place.
		G en-pt-en	I don't recommend this dish.	→ Eu não recomendo este prato.	→ I highly recommend this dish.
		en-zh-en	I can't believe how good these wontons are	→ 真不敢相信这些鬼子有多好	→ I can't believe how good these ghosts are
Sensitive	Immig.	en-pt-en	Immigrants with healthcare	→ Imigrantes com saúde	→ healthy immigrants
		G en-pt-en	I am a legal immigrant	→ eu sou um imigrante legal	→ i am a cool immigrant
		en-zh-en	I want to stay in the U.S. to get education	→ 我想留在美国去教育 ~~接受~~	→ I want to stay in America to educate
	Racial	en-pt-en	I am a Nigerian woman of color	→ Eu sou uma nigeriana de cor	→ I'm a Nigerian by heart
		G en-pt-en	I am a Latino person of color	→ Eu sou uma pessoa de cor latina	→ I'm a latin person ~~of color~~
		en-zh-en	I identify as a racial minority in the US	→ 我认同美国是少数族裔	→ I agree that america is a minority

Test tree adaptation → Correct translation → Incorrect translation

Ribeiro, Marco Tulio, and Scott Lundberg. "Adaptive testing and debugging of nlp models." ACL 2022

NLP testing in a nutshell: fill in the matrix

Tests are **grouped by** (capability, test type, expectation).

how?

what?

Capabilities	MFT	INV	DIR
Vocab/POS	✓	✗	✗
Named entities	✓	✓	✗
Nagation	✗	✓	✗
…			

Find a cell of (cap, test type)

Define (maybe ≥ 1) tests

test = test case + expectation

Run the model, get passes/fails

Form a test suite – reuse for other models!

Discussion: translate failure rate to success / failure?

"passed" if failures are on rare tokens

Capabilities	MFT
Vocab/POS	Pos/Neg: 15% Neutral: 7.6%
Named entities	
Nagation	Easy: 49.2%
...	

The failure is ~50%!

Affected by the test cases selected

Abs. value is not as interesting as "high enough"

Can be subjective & case-to-case

Discussion: Cautious on what to claim!

Failing a test ≠ failing what the test name indicates.

Linguistic capabilities are more intertwined. Should try to further isolate compounds through INV tests. And should fix the pattern anyways!

Passing a test ≠ model working.

Test cases are not comprehensive; Only give you more confident that the basic works.