EMNLP 2024

Human-Centered Evaluation of Language Technologies

Su Lin Blodgett Microsoft Research Montréal
Jackie C.K. Cheung McGill University
Q. Vera Liao Microsoft Research Montréal
Ziang Xiao Johns Hopkins University

Microsoft

Blodgett, S. L., Cheung, J. C. K., Liao, V., & Xiao, Z. (2024, November). Human-Centered Evaluation of Language Technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts* (pp. 39-43).

JOHNS HOPKINS UNIVERSITY

NLP Is In The World!

similarweb Products Customers Our Data Pricing Resources Conta

chatgpt.com VS. + Compare this site to October

Home > Website Traffic Checker > chatgpt.com

chatgpt.com Website Analysis for October 2024

chatgpt resources and news. learn how artificial intelligence is changing the world, and use ai chat bots to discover what

Show more

Computers Electronics and Technology > Programming and Developer Software

Company	Year Founded	Employees	Annual Revenue
Chat Gpt	2022	51 - 200	--

Keep track of your competitors

See their traffic and performance metrics

Try it now

Global Rank #13 ▲1

Country Rank #20 ▲3

United States

Showing Similarweb estimated data.
Publicly validate your site's metrics by connecting your GA4

Connect your Google Analytics

Total Visits 3.7B

Bounce Rate 37.96%

Pages per 3.66

Rank	Website	Category
1	google.com	Computers Electronics and Technology > Search Engines
2	youtube.com	Arts & Entertainment > Streaming & Online TV
3	facebook.com	Computers Electronics and Technology > Social Media Networks
4	amazon.com	Ecommerce & Shopping > Marketplace
5	reddit.com	Computers Electronics and Technology > Social Media Networks
6	yahoo.com	News & Media Publishers
7	x.com	Computers Electronics and Technology > Social Media Networks
8	instagram.com	Computers Electronics and Technology > Social Media Networks
9	wikipedia.org	Reference Materials > Dictionaries and Encyclopedias
10	microsoftonline.c...	Computers Electronics and Technology > Programming and Developer Software
11	office.com	Computers Electronics and Technology > Programming and Developer Software
12	chatgpt.com	Computers Electronics and Technology > Programming and Developer Software
13	linkedin.com	Computers Electronics and Technology > Social Media Networks
14	espn.com	Sports > Sports - Other
15	ebay.com	Ecommerce & Shopping > Marketplace

Why Think About Evaluation?

Possible questions of interest:

- Does this NLP system have a certain property, or skill?
  - Does it understand? Does it know something about language such as its syntax or semantics?
- Is this NLP system useful?
  - Can it help users solve a task better, faster, or more cheaply?
- Is this NLP system harmful?
  - Might it risk users privacy? Does it perpetuate stereotypes? Does it equally serve all groups of users?

How do we decide which questions to ask, how to answer these questions, and how to do so well?

NLP Task Settings

Tasks familiar to NLP researchers

- Machine translation, text summarization, sentiment analysis, dialogue systems
- Evaluation practices well attested in existing conference tracks

New use cases the field hasn't engaged deeply with traditionally

- Applications enabled by large pretrained models
- Entertainment, medicine, finance, education
- Many use cases invented by users interacting with systems!
- How do we think about evaluation with the growing diversity of language technologies?

Inspirations from Social Sciences and HCI

Give us methods and vocabulary to complement existing NLP evaluation methods

From the social sciences:

- Dealing with contested constructs (e.g., intelligence, gender, fairness)
- Definitions, measurements and operationalizations; validity of measurements

From human-computer interaction:

- Empirical studies involving users
- Qualitative and quantitative approaches both valued!

Goals of the Tutorial

- Reflect on current landscape of evaluation in NLP
  - Assumptions about evaluation methods
  - Trade-offs between different aspects of evaluation
- Learn about viewpoints from the social sciences and HCI
- Build toolkit for:
  - Designing evaluations
  - Methods to evaluate evaluations: vocabulary to discuss, critique and analyze evaluations
- Hands-on practice of the above!
- Examples motivated in terms of applications, but we think understanding this landscape is generally useful even if you work more upstream

Today's Tutorial

1. Current evaluation practices (NLP)
2. Evaluating evaluation: perspectives from the social sciences
3. Evaluation practices in HCI

Break

4. Example language technologies and their HCI evaluations
5. Reflection (in groups, independently, …)
6. Hands-on exercise (in groups, independently, …)

Current Evaluation Practices in NLP

Su Lin Blodgett & Jackie C.K. Cheung

Section Overview

Classifying existing evaluation methods in NLP

Dataset construction and benchmarking

Common methods for results analysis

Motivations for performing evaluations

Assumptions behind current practices

Motivations and Limitations of this Section

Capture current landscape of evaluation

Reflect on assumptions underpinning these methods

We focus on practices represented by *academic publications*

- Other methods in industry may be more attested and less covered in the academic literature, but we do not have full visibility on their practices

Basic Distinctions in NLP Evaluation

Automatic vs. human evaluation

Reference-based vs. reference-free

Intrinsic vs. extrinsic evaluation

What is the task?

- Classification, structure prediction, generation, representation learning
- Implications for metrics design

Automatic Evaluation – Classification

Evaluations where human intervention is not needed *at the time of evaluation*

Classification: evaluate against gold-standard, reference label

Precision

$$\frac{\text{\# correct}}{\text{\# predicted}}$$

Recall

$$\frac{\text{\# correct}}{\text{\# in-dataset}}$$

F1

$$\frac{2 \times P \times R}{P + R}$$

Metrics embed assumptions about what is important!

- e.g., How do we aggregate across classes if they are imbalanced?
- Micro- vs. macro-averaging treat minority classes differently.

Automatic Evaluation – Structure Prediction

Compare similarity of system prediction vs. reference output

Example: Constituent Parsing

PARSEVAL: Consider a constituent correct if span and label are correct (Black et al., 1991)

Compute P, R, F1

This is [$_{NP}$ a constituent]. Reference

This is [$_{VP}$ a constituent]. ×

This is a [$_{NP}$ constituent]. ×

Automatic Evaluation – Generation

Compare similarity of system output to reference generation

Example: Automatic summarization

ROUGE scores compute N-gram overlap

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

(Lin, 2004)

Reference-based vs. Reference-free

Methods so far assume **a gold-standard reference is available**

How are references gathered?

- Expert annotations – costly!
- Crowd annotations – cheaper but need to control quality
- Semi-automatic or LLM-generated labels

References embed assumptions

- About who carries knowledge or whose knowledge is valued
- About whether there is a single reference, or many

Next, let's consider a **reference-free** approach

Reference-free Evaluation – Generation

QuestEval: Summarization evaluation via question answering

(Scialom et al., 2021)

Relies on question generation and question answering systems!

Source Document This is the embarrassing moment a *Buckingham Palace* guard slipped and fell on a manhole cover in front of hundreds of shocked tourists as he took up position in his sentry box. [...] The Guard comprises two detachments, one each for Buckingham Palace and St James's Palace, under the command of the Captain of The Queen's Guard.
Generated Question Where was the Changing of the Guard held?
Weighter prediction *Important Question*
Answer Span Buckingham Palace

Correct Summary The Queen's Guard slipped on a manhole cover during the Changing of the Guard at *Buckingham Palace* last week. [...]
Predicted Answer Buckingham Palace: ✓

Hallucinated Summary The Queen's Guard slipped on a manhole cover during the Changing of the Guard at *St James's Palace* last week. [...]
Predicted Answer St James's Palace: ✗

Incomplete Summary The Queen's Guard slipped on a manhole cover during the Changing of the Guard during an embarrassing moment.. [...]
Predicted Answer Unanswerable: ✗

Evaluation for Unsupervised or Induction Settings

e.g., topic models, language models, grammar induction

Two approaches:

- Comparing induced structure to reference structure in the target domain
- Testing for desired properties of / behaviours related to the induced structures

Grammar Induction Evaluation

Reference-based

Similar to evaluation of supervised parsing

Consider a constituent correct if span is correct

Compute P, R, F1

This is [$_{NP}$ a constituent]. Reference

This is [a constituent]. **OK**

This is a [constituent]. ×

Perplexity

Assumption: a good model should predict test corpus with high likelihood, because test corpus is drawn from the true data generation distribution

For a model *q*, applied to a test corpus of length N:

$$2^{-\frac{1}{N}\log_2 q(w_1 \ldots w_N)}$$

Evaluation of Learned Representations

Representations learned by neural models have no absolute interpretation → reference-based evaluation not possible!

- Instead, test if learned representation has expected property or structure

Example: Word vector evaluation with WordSim-353

(Finkelstein et al., 2001)

monk	oracle	5
cemetery	woodland	2.08
food	rooster	4.42
coast	hill	4.38
forest	graveyard	1.85
shore	woodland	3.08
monk	slave	0.92

Human Evaluation Methods – Human Judgments

Ask human annotators for their judgments: usually used for generation tasks

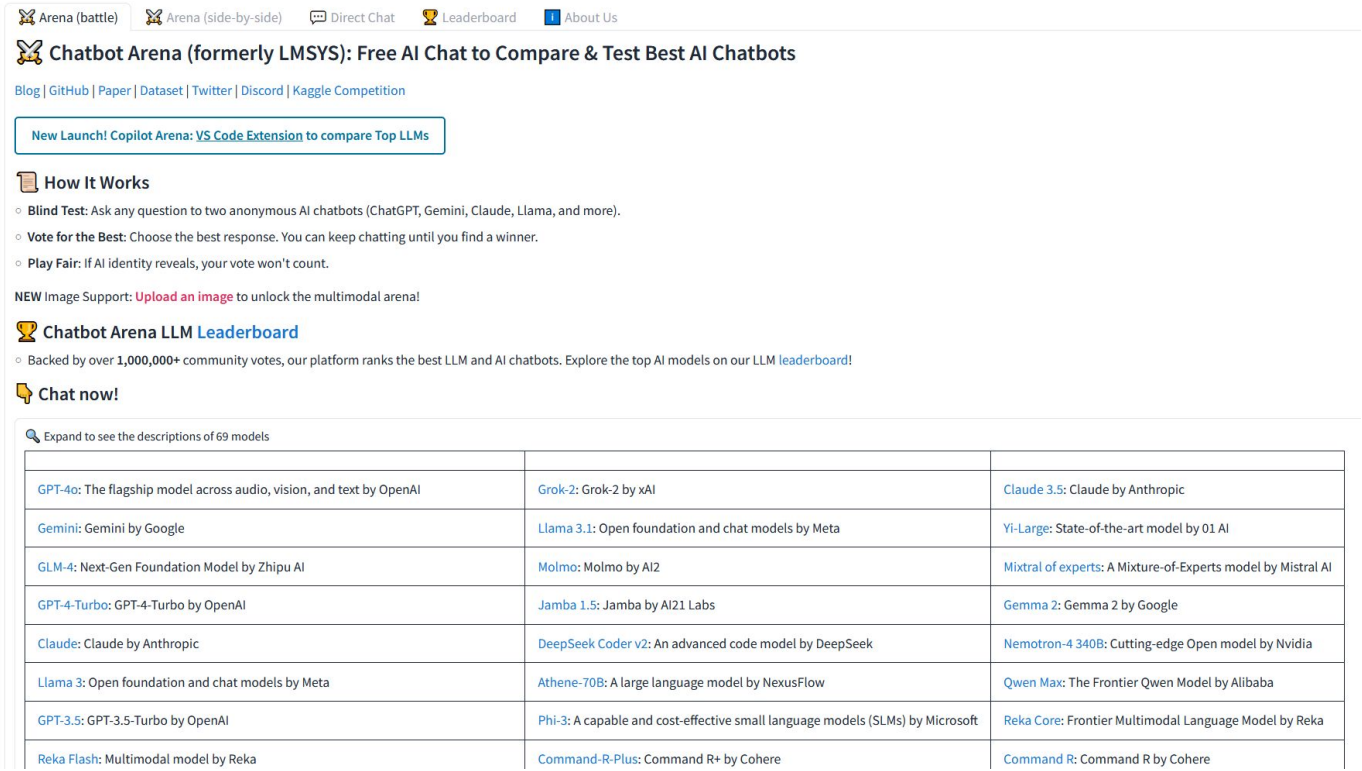
Absolute: Ask judges to give a rating of a model output

e.g., Overall score, informativeness, non-redundancy, linguistic quality scores

Preferences: Ask judges to give a relative judgement between two outputs

Chatbot Arena

(Chiang et al., 2024)



Arena (battle) | Arena (side-by-side) | Direct Chat | Leaderboard | About Us

Chatbot Arena (formerly LMSYS): Free AI Chat to Compare & Test Best AI Chatbots

Blog | GitHub | Paper | Dataset | Twitter | Discord | Kaggle Competition

New Launch! Copilot Arena: VS Code Extension to compare Top LLMs

How It Works

- **Blind Test**: Ask any question to two anonymous AI chatbots (ChatGPT, Gemini, Claude, Llama, and more).
- **Vote for the Best**: Choose the best response. You can keep chatting until you find a winner.
- **Play Fair**: If AI identity reveals, your vote won't count.

NEW Image Support: **Upload an image** to unlock the multimodal arena!

Chatbot Arena LLM Leaderboard

- Backed by over **1,000,000+** community votes, our platform ranks the best LLM and AI chatbots. Explore the top AI models on our LLM leaderboard!

Chat now!

Expand to see the descriptions of 69 models

GPT-4o: The flagship model across audio, vision, and text by OpenAI	Grok-2: Grok-2 by xAI	Claude 3.5: Claude by Anthropic
Gemini: Gemini by Google	Llama 3.1: Open foundation and chat models by Meta	Yi-Large: State-of-the-art model by 01 AI
GLM-4: Next-Gen Foundation Model by Zhipu AI	Molmo: Molmo by AI2	Mixtral of experts: A Mixture-of-Experts model by Mistral AI
GPT-4-Turbo: GPT-4-Turbo by OpenAI	Jamba 1.5: Jamba by AI21 Labs	Gemma 2: Gemma 2 by Google
Claude: Claude by Anthropic	DeepSeek Coder v2: An advanced code model by DeepSeek	Nemotron-4 340B: Cutting-edge Open model by Nvidia
Llama 3: Open foundation and chat models by Meta	Athene-70B: A large language model by NexusFlow	Qwen Max: The Frontier Qwen Model by Alibaba
GPT-3.5: GPT-3.5-Turbo by OpenAI	Phi-3: A capable and cost-effective small language models (SLMs) by Microsoft	Reka Core: Frontier Multimodal Language Model by Reka
Reka Flash: Multimodal model by Reka	Command-R-Plus: Command R+ by Cohere	Command R: Command R by Cohere

Human Evaluation Methods – Structured Evaluation

Judgments do not have to be at the passage level.

Breakdown is often structured depending on the task setting

e.g., The Pyramid Method for summarization evaluation

1. Annotate reference summaries for information chunks (**SCUs**; **summary content units**)
2. Annotate system summaries for SCUs
3. Score overlap between the system and reference SCUs

(Nenkova and Passonneau, 2004)

LLM Evaluation

Emerging area: replace the human in human evaluation methods with LLMs

At present, they seem unreliable at replicating human judgments, with large variance in correlations across datasets.

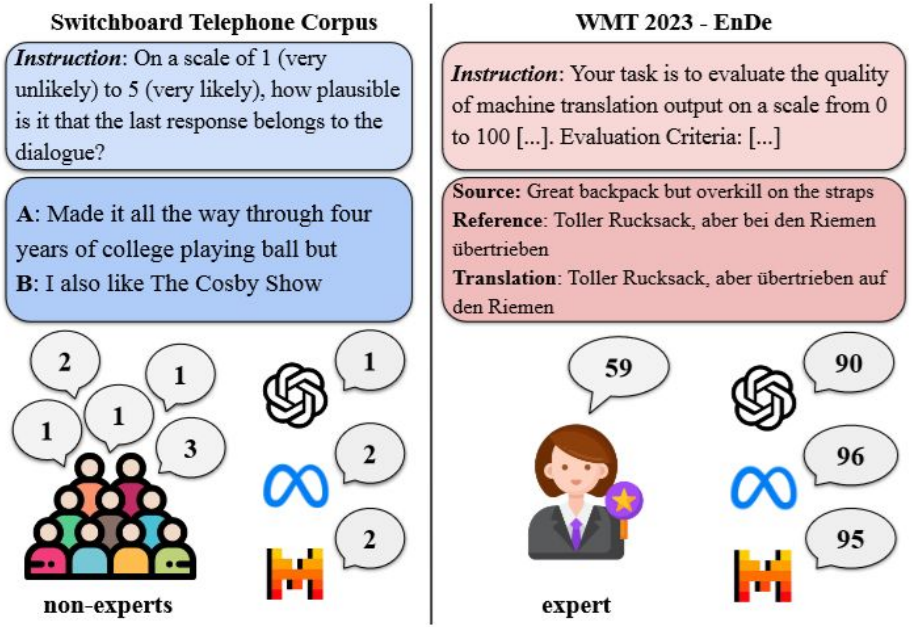


Figure 1: Evaluation by expert and non-expert human annotators and by LLMs for two tasks involving human-generated (left) and machine-generated text (right).

(Bavaresco et al., 2024)

Intrinsic vs. Extrinsic Evaluation

Intrinsic: A model trained for a task being evaluated w.r.t. **the same** task

e.g., Reference-based evaluations are usually intrinsic

Extrinsic: A model trained for a task being evaluated using another task (that the first task is thought to be useful for)

e.g., QuestEval: evaluate summarization via QA

e.g., Evaluate language model using automatic speech recognition

How Are Evaluations Judged?

How are automatic metrics evaluated?

- Most common answer: by correlation with human judgments
  - e.g., SummEval
- Intrinsic metrics sometimes evaluated by correlation with extrinsic metrics
  - e.g., Does improving perplexity improve word error rate in speech recognition?

How are human judgments evaluated?

- Most common answer: by inter-annotator agreement.
  - This could be problematic, e.g. if multiple correct answers possible (Passonneau and Carpenter, 2014)

Later, we will discuss **validity**!

Common Analyses

Manual reading - common, but often does not follow a formal method (Zhou et al., 2022)

> *“[I]t just comes down to me reading a lot of samples and then choosing the one which overall seems to be better”*

Error analysis - characterizing or taxonomizing model errors

> Often qualitative

Ablation studies

Benchmark Datasets

Most evaluations require benchmark dataset, which are diverse in their construction

- Large crowdsourced datasets
  - e.g., SQUaD for question answering (Rajpurkar et al., 2016)
- Targeted expert-constructed datasets
  - e.g., Winograd Schema Challenge for common-sense reasoning (Levesque et al., 2012)

Benchmark dataset consists of:

- Test instances
- Method for assessing model behavior using the instances
- Method to accumulate model behavior on instances into overall score or result

Dataset construction practices

How have dataset construction practices evolved over time?

Three broad time periods:

– 1980s: Classical period

1990s – mid-2010s: Empirical revolution

mid-2010s – now: Modern synthesis

Classical Period: Case-based Evaluation (–1980s)

Demonstrate that theory works on selected cases that illustrate a phenomenon of interest. Mostly human evaluation (by paper authors!)

in an analogous manner. Thus, the lexical entries for the French verb forms *connaît* and *sait* might be as follows:

$$\begin{bmatrix} \text{Cat} = \text{V} \\ \text{Lex} = \text{connaitre} \\ \text{Tense} = \text{Pres} \\ \text{Subj} = \begin{bmatrix} \text{Pers} = 3 \\ \text{Num} = \text{Sing} \\ \text{Anim} = + \end{bmatrix} \\ \text{Obj} = [\text{Cat} = \text{NP}] \end{bmatrix} \qquad \begin{bmatrix} \text{Cat} = \text{V} \\ \text{Lex} = \text{savoir} \\ \text{Tense} = \text{Pres} \\ \text{Subj} = \begin{bmatrix} \text{Pers} = 3 \\ \text{Num} = \text{Sing} \\ \text{Anim} = + \end{bmatrix} \\ \text{Obj} = [\text{Cat} = \text{S}] \end{bmatrix}$$

Each requires its subject to be third person, singular and animate. Taking a rather simplistic view of the difference between these verbs for the sake of the example, this lexicon states that *connaît* takes noun phrases as objects, whereas *sait* takes sentences.

(Kay, 1984)

The Empirical Revolution (1990s – mid-2010s)

- Empirical, dataset-based evaluation
  - Draw from a representative sample from one or more data sources
  - Standard benchmarks with agreed-upon metrics, data splits, and automatic evaluation metrics

Most famous example: the Penn Treebank - Wall Street Journal for parsing

```
( (S
     (NP
          (NP (NNP Pierre) (NNP Vinken) ) (, ,)
               (ADJP (NP (CD 61) (NNS years) ) (JJ old) ) (, ,) )
     (VP (MD will) (VP (VB join)
          (NP (DT the) (NN board) )
          (PP (IN as) (NP (DT a) (JJ nonexecutive) (NN director) ))
          (NP (NNP Nov.) (CD 29) ))) (. .) ) )
```

(Marcus et al., 1999)

Modern Synthesis: Pendulum Swings Back (mid-2010s –)

Challenge datasets – samples have particular properties thought to be difficult

e.g., Winograd Schema Challenge, hand designed to be difficult

The trophy *doesn't fit into* ***the suitcase*** *because it was too* ***large/small****.*

What doesn't fit?

Can be created using insights about task and/or automatic methods

e.g., adversarial filtering to remove cases solvable by baseline models (Sakaguchi et al., 2021)

Other Trends in Dataset Construction Practices

Out-of-distribution testing

Distribution shift in test set *on purpose* – systematic generalization

Require models to learn some capability to generalize well

e.g., Coreference resolution → Winograd Schema Challenge

e.g., sNLI → HANS in natural language inference literature

Multi-dataset benchmarks and evaluation

e.g., SuperGLUE (Wang et al., 2019)

Reflections on Assumptions in Field

What is a task?

Datasets are often constructed w.r.t. to a specific task.

How do we reflect on what datasets are useful for, and what the definition of a task is?

Is summarization a task? Is question answering a task?

What is the point of a task?

To test for intelligent behaviour? For usefulness?

To make claims about models that "understand language" in a particular way?

Summary of Current Practices

- Diverse methods employed in NLP for evaluation
  - Automatic vs. human evaluations
  - Reference-based vs. reference-free
  - Task setting influences choice of evaluation approach

- Dataset construction is key part of evaluation, and has evolved over time

- Evaluation and analysis approaches and metrics embed assumptions about researchers' goals and interests

What's Next?

Possible **limitations and concerns** in current practices:

- Assume more is better → trend towards large-scale multi-task benchmarks
  - Could think more about validity and capabilities of interest
- Current practices tend to abstract away from deployment settings/users
  - How much does context specificity matter?
- Assumption about humans being “gold standard”
  - Can benefit from HCI theory and empirical work on humans
- Assumption about (dis)agreement
  - Can benefit from HCI and social sciences on understanding and navigating dissensus

Evaluating evaluation: Perspectives from Social Sciences

Su Lin Blodgett & Ziang Xiao

Section Outline

- **Measures in social science**
- **Measurement theory**
- **Evaluating measures**
- **Measurement theory in language technology evaluation**

In the social sciences,

We are often interested in measuring **theoretical**, **unobservable** constructs to understand humans and society.

For example,

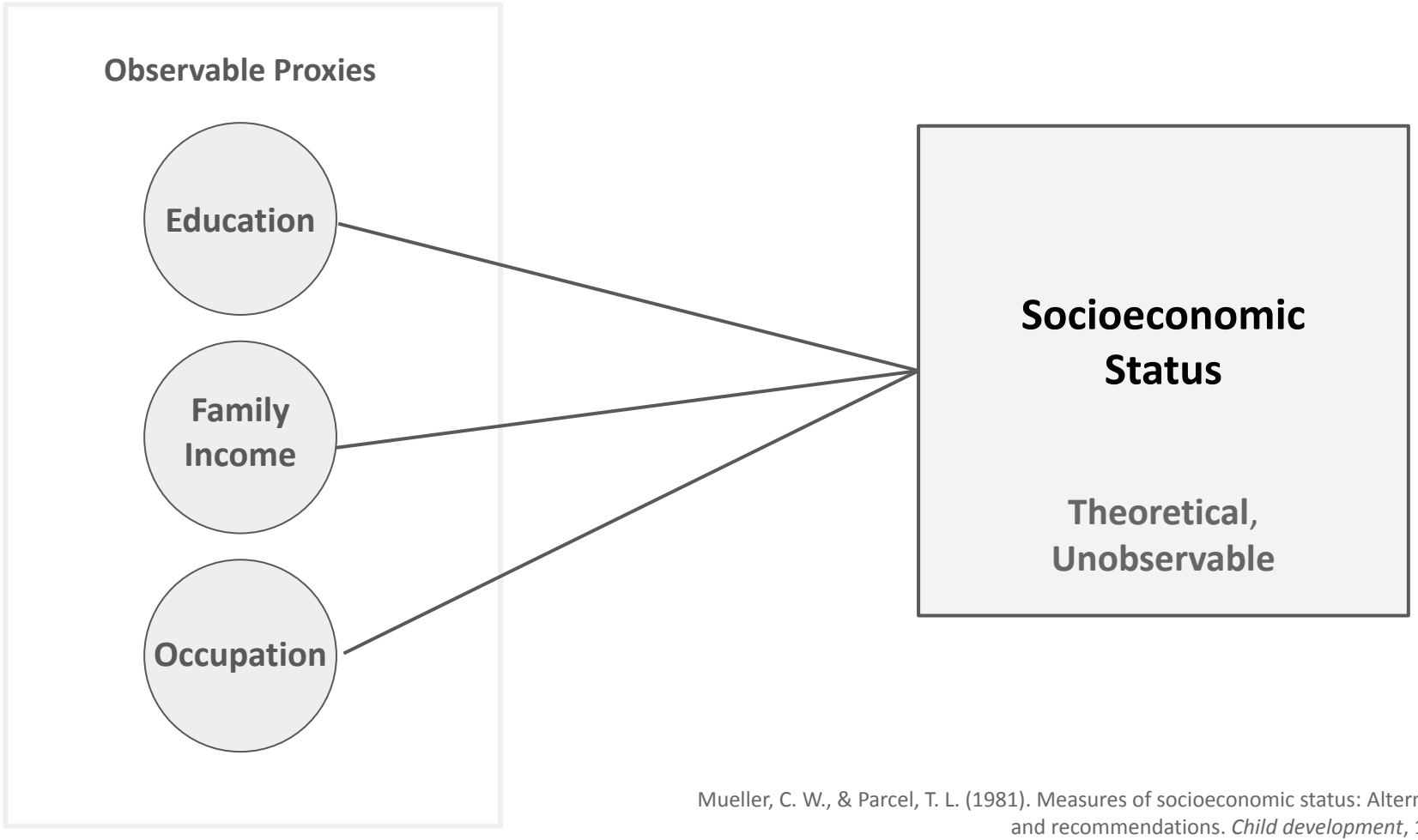
- Motivation
- Psychological state
- Socioeconomic status
- Quality of life
- Intelligence
- Teacher quality

Zeller, R. A., & Carmines, E. G. (1980). Measurement in the social sciences: The link between theory and data. Cambridge University Press.

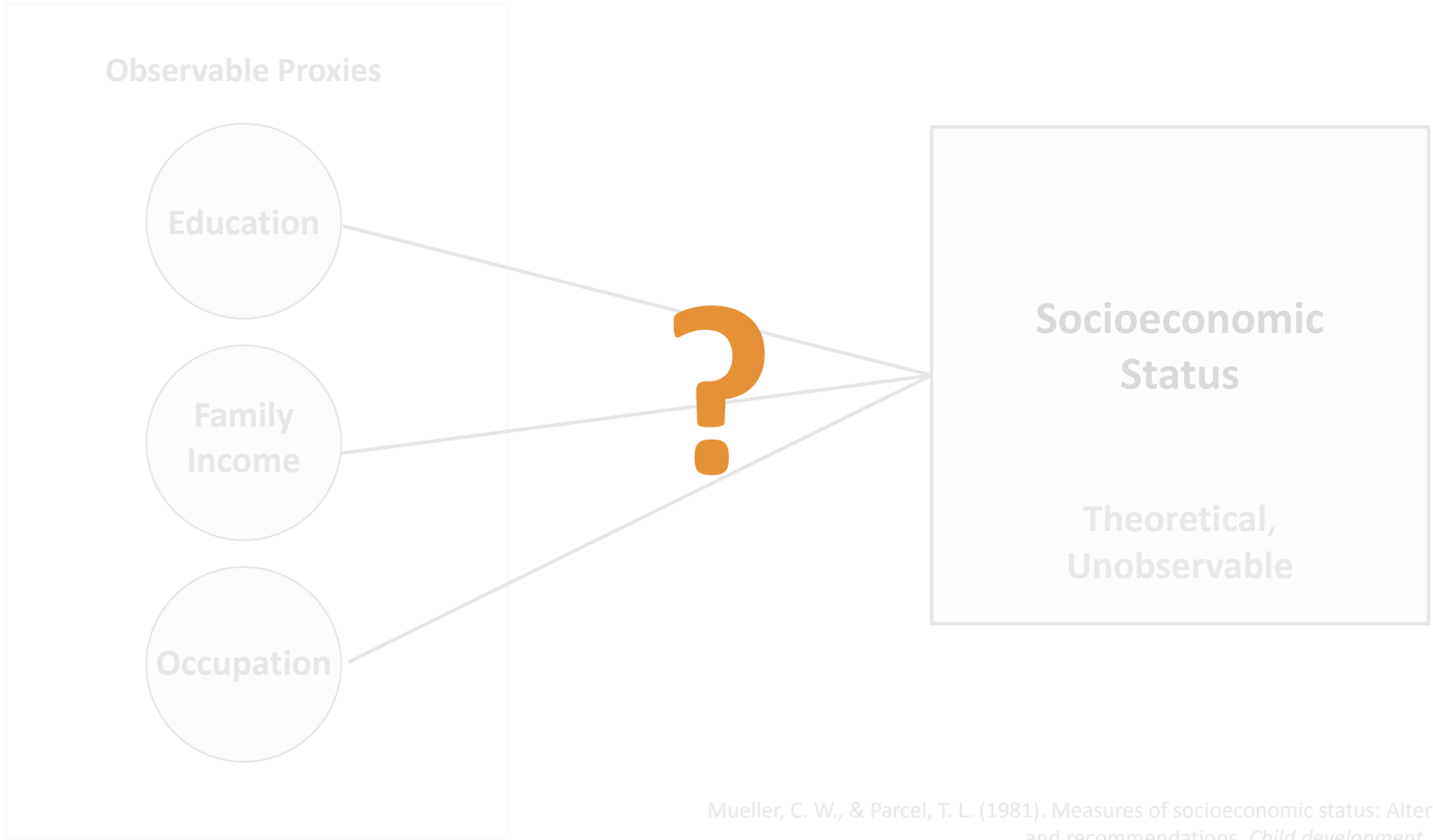
Socioeconomic Status

Theoretical, Unobservable

Mueller, C. W., & Parcel, T. L. (1981). Measures of socioeconomic status: Alternatives and recommendations. *Child development*, 13-30



Mueller, C. W., & Parcel, T. L. (1981). Measures of socioeconomic status: Alternatives and recommendations. *Child development*, 13-30



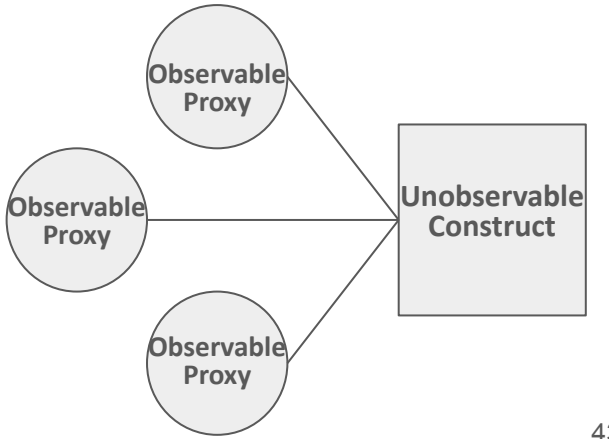
Mueller, C. W., & Parcel, T. L. (1981). Measures of socioeconomic status: Alternatives and recommendations. *Child development*, 13-30

Measurement Theory

Measurement theory is the foundation for developing, evaluating, and interpreting measurement in the social sciences.

Articulate assumptions:

- What do you want to measure?
- How do you measure it?



Consequences of a bad measure

Job suitability - hiring or promotion decisions that rely on flawed personality tests like the Myers-Briggs Type Indicator (MBTI)

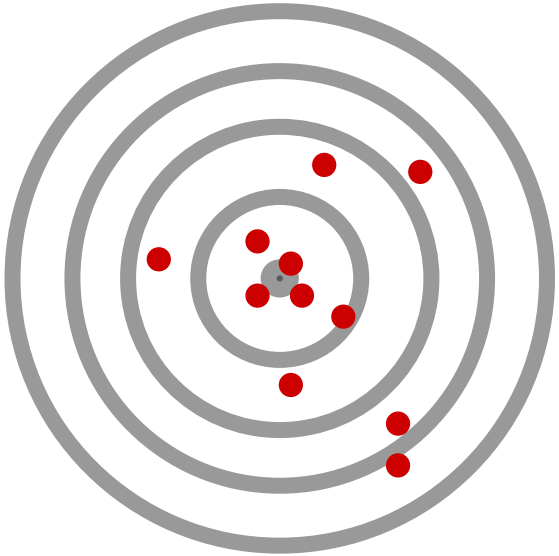
Intelligence - standardized tests that measure using culturally biased questions

Recidivism risk - in criminal justice, risk assessment tools that measure by heavily weighing factors like neighborhood or family history

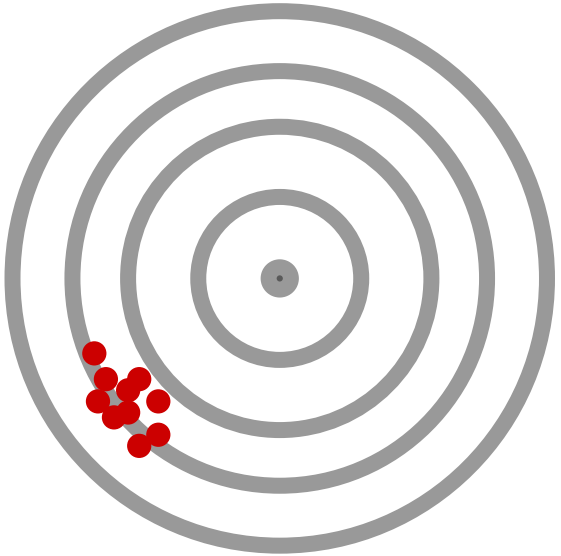
How should we evaluate the quality of a measurement?

Methods in measurement theory aims to **identify** or **quantify** measurement errors

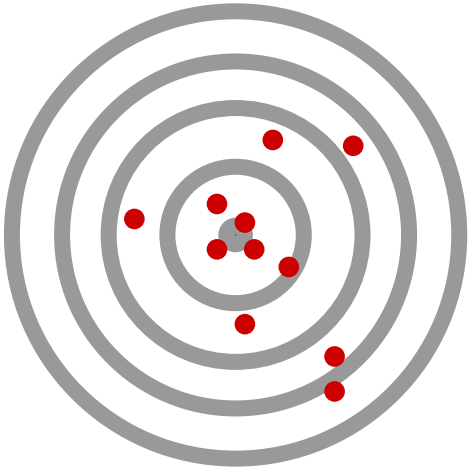
Reliability



Validity



Reliability refers the degree to which the measure of a construct is consistent or dependable.



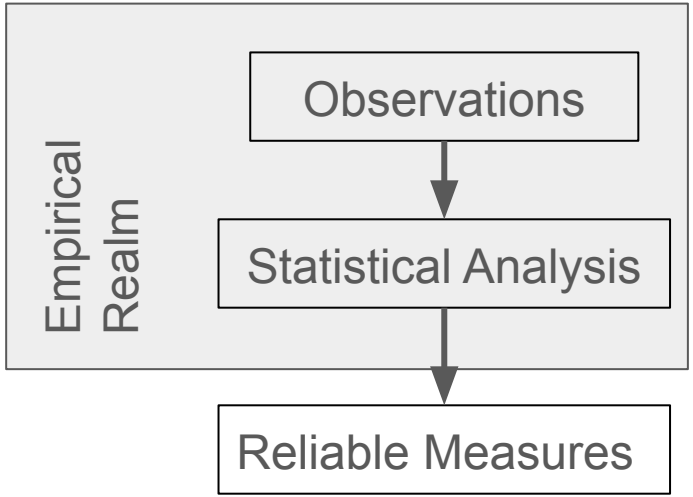
Not reliable

Types of Reliability

Test-retest Reliability: A measure of how consistent a measurement when applied multiple times to the same individual, indicating the stability of the scores over time.

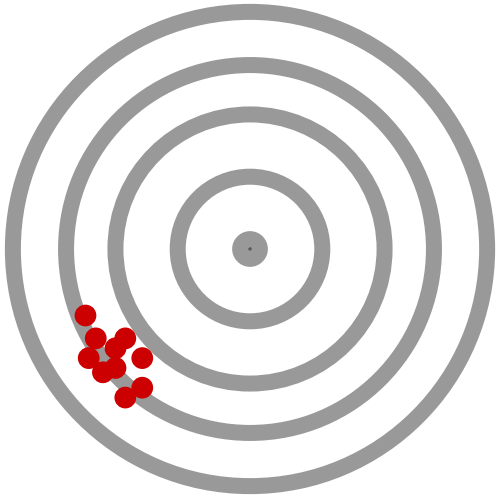
Internal-consistency Reliability: A measure of how well a set of items in a measure the same underlying construct.

Reliability Assessment

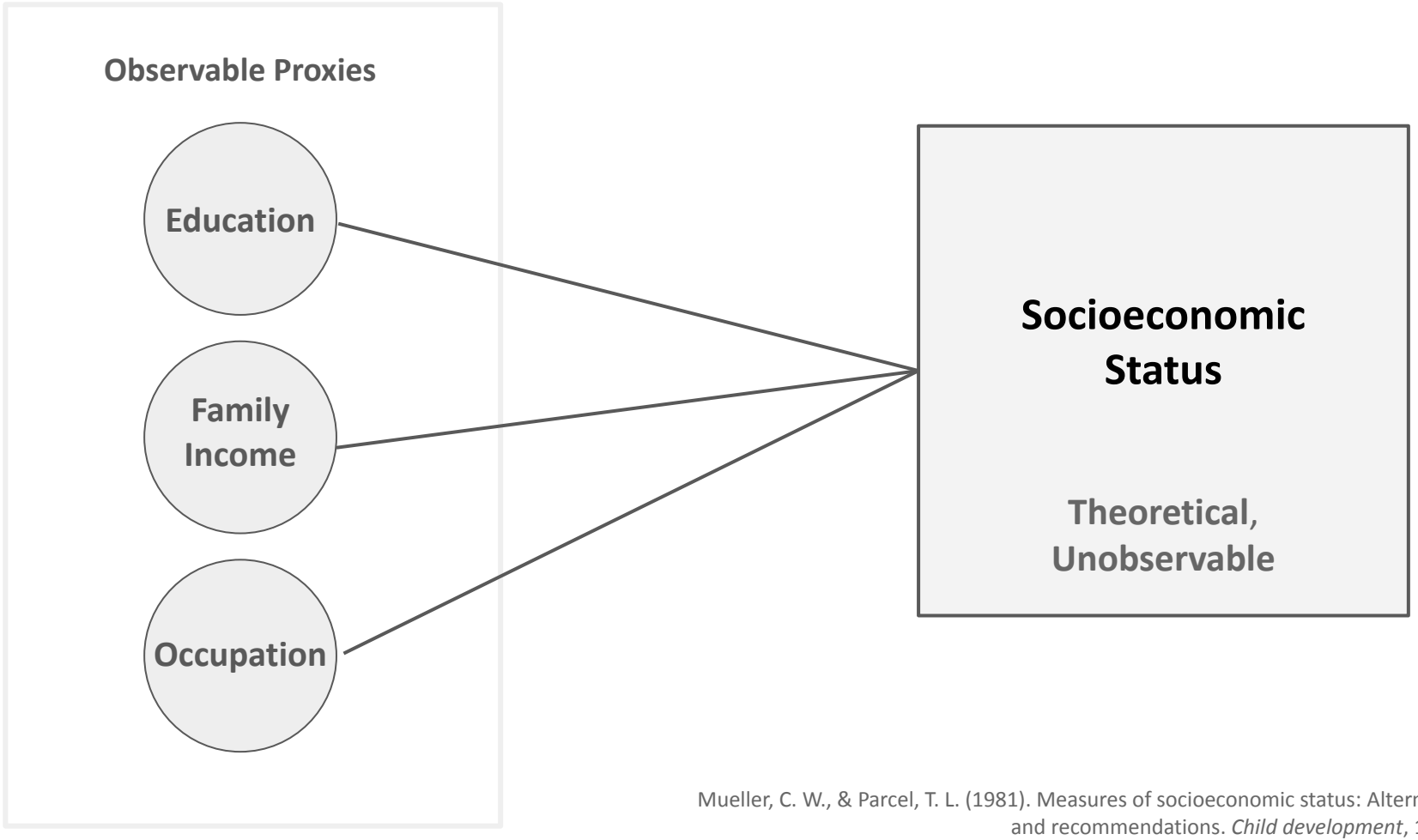


Wellington, J., & Szczerbinski, M. (2007). Research methods for the social sciences. A&C Black.

Validity refers to the extent to which a measure adequately represents the underlying construct that it is supposed to measure.



Reliable, but not valid



Mueller, C. W., & Parcel, T. L. (1981). Measures of socioeconomic status: Alternatives and recommendations. *Child development*, 13-30

Validity Frameworks (and many more)

Measurement Theory

- Representational Validity
  - Face validity
  - Content validity
- Criterion-related Validity
  - Convergent validity
  - Discriminant validity
  - Concurrent validity
  - Predictive validity

Social Science Research

- Internal validity
- External validity
  - Ecological validity
  - Cross-cultural validity
  - Population validity
  - ….

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

Wellington, J., & Szczerbinski, M. (2007). Research methods for the social sciences. A&C Black.

Types of Validity

Representational Validity: How well the operationalization is a good reflection of the construct

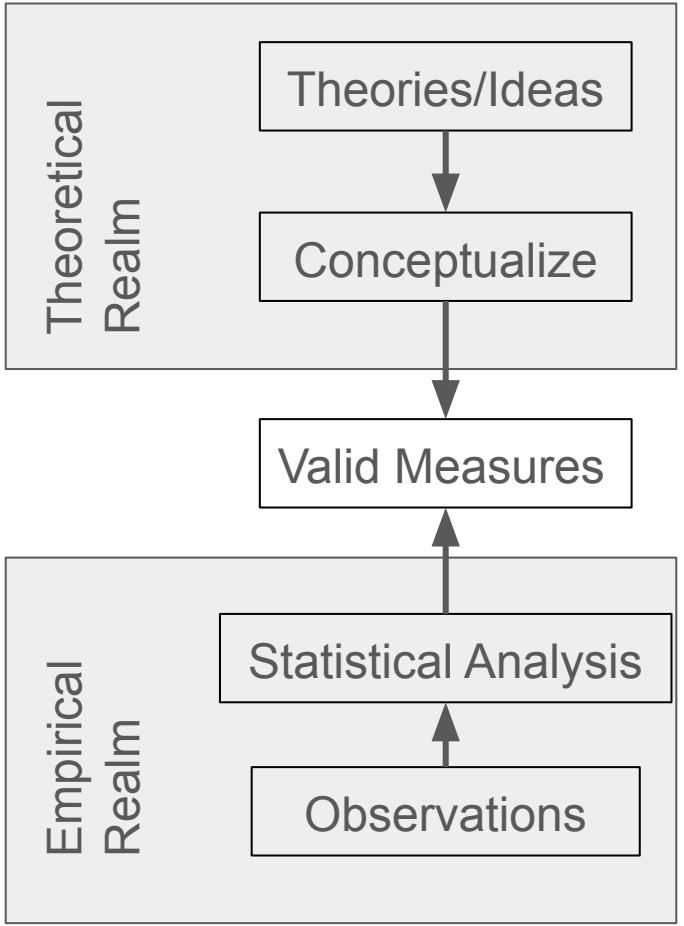
- Face validity
- Content validity

Criterion-related Validity: How well the operationalization behaves the way it should given the theory of the construct

- Convergent
- Discriminant
- Concurrent
- Predictive validity

Cronbach, L. J., & Meehl, P. E. (1955). Construct validity in psychological tests. *Psychological bulletin*, *52*(4), 281.

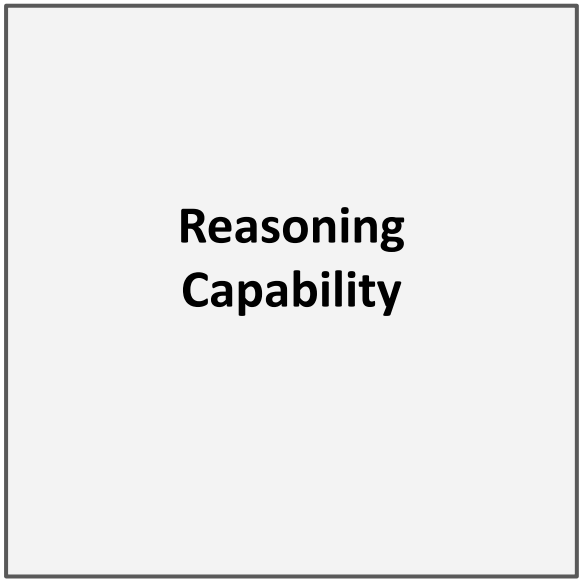
Validity Assessment



Wellington, J., & Szczerbinski, M. (2007). Research methods for the social sciences. A&C Black.

How does **measurement theory**
connect to
language technology evaluation?

NLP Evaluations as Measurement

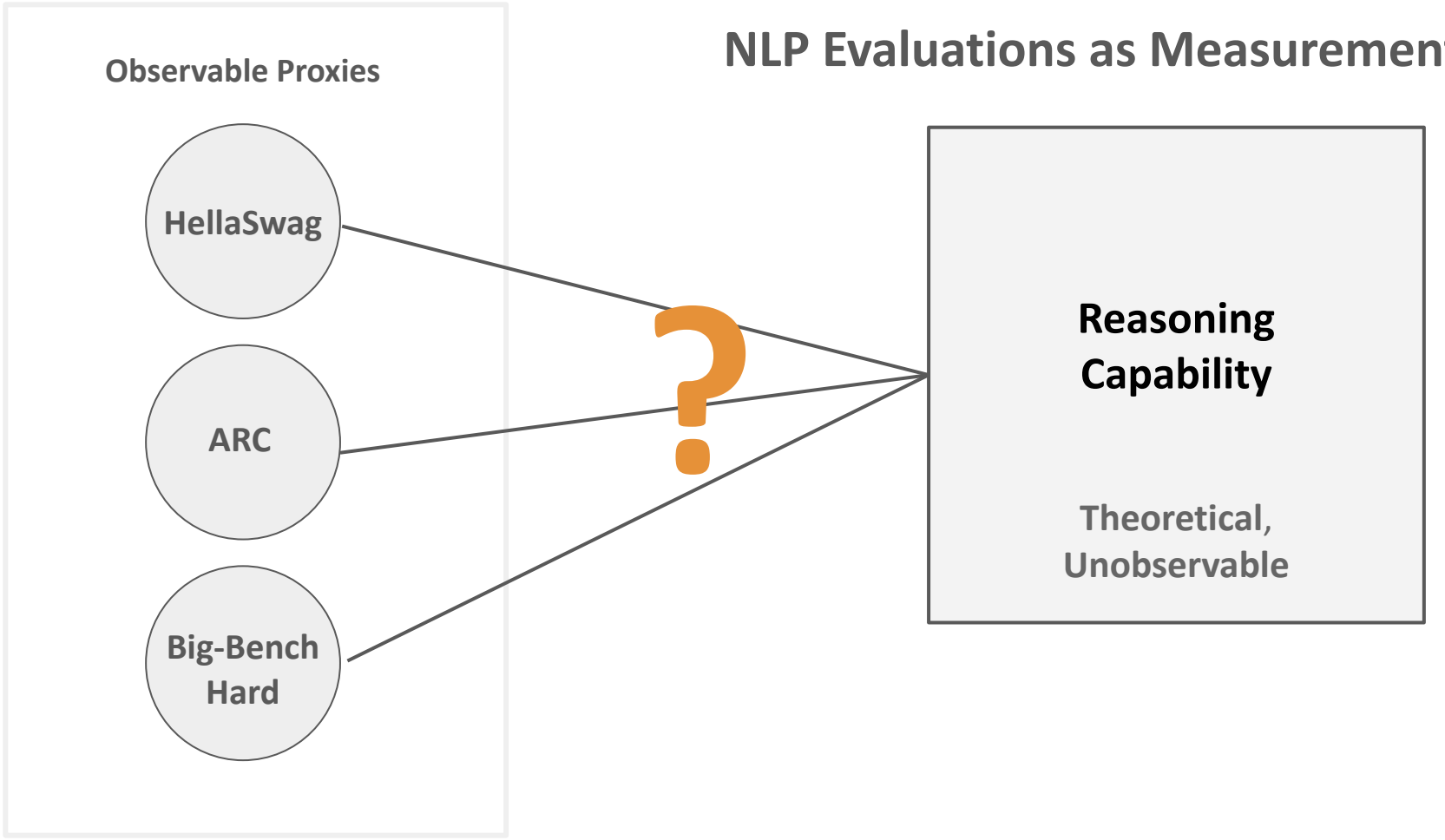


NLP Evaluations as Measurement

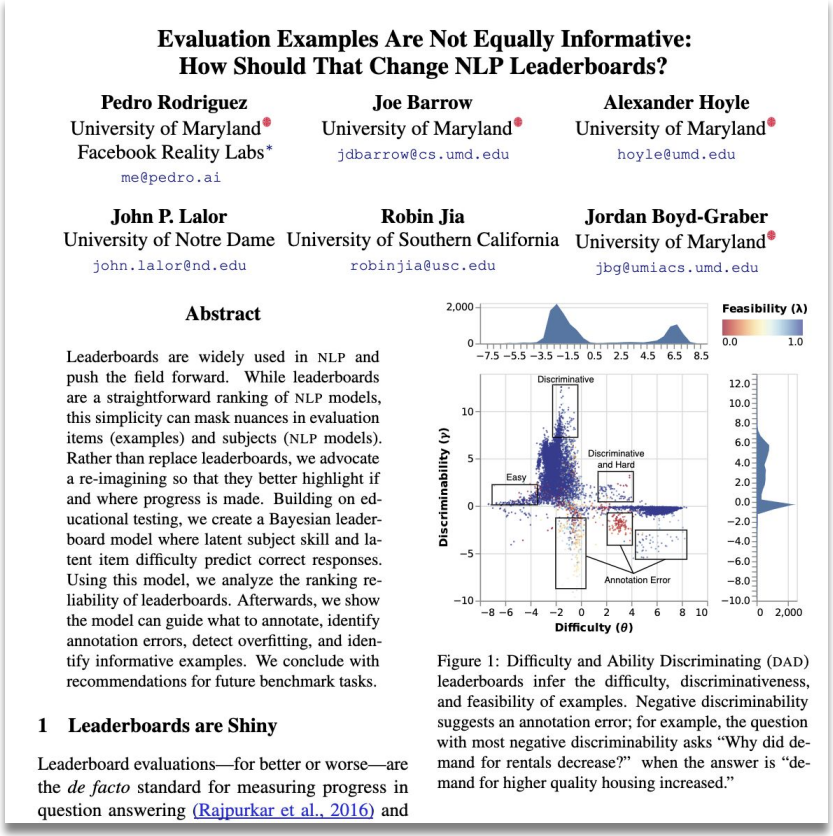
Reasoning Capability

Theoretical, Unobservable

NLP Evaluations as Measurement



Connecting measurement theory to language technology eval



Evaluation Examples Are Not Equally Informative: How Should That Change NLP Leaderboards?

Pedro Rodriguez
University of Maryland
Facebook Reality Labs*
me@pedro.ai

Joe Barrow
University of Maryland
jdbarrow@cs.umd.edu

Alexander Hoyle
University of Maryland
hoyle@umd.edu

John P. Lalor
University of Notre Dame
john.lalor@nd.edu

Robin Jia
University of Southern California
robinjia@usc.edu

Jordan Boyd-Graber
University of Maryland
jbg@umiacs.umd.edu

Abstract

Leaderboards are widely used in NLP and push the field forward. While leaderboards are a straightforward ranking of NLP models, this simplicity can mask nuances in evaluation items (examples) and subjects (NLP models). Rather than replace leaderboards, we advocate a re-imagining so that they better highlight if and where progress is made. Building on educational testing, we create a Bayesian leaderboard model where latent subject skill and latent item difficulty predict correct responses. Using this model, we analyze the ranking reliability of leaderboards. Afterwards, we show the model can guide what to annotate, identify annotation errors, detect overfitting, and identify informative examples. We conclude with recommendations for future benchmark tasks.

1 Leaderboards are Shiny

Leaderboard evaluations—for better or worse—are the *de facto* standard for measuring progress in question answering (Rajpurkar et al., 2016) and

Figure 1: Difficulty and Ability Discriminating (DAD) leaderboards infer the difficulty, discriminativeness, and feasibility of examples. Negative discriminability suggests an annotation error; for example, the question with most negative discriminability asks "Why did demand for rentals decrease?" when the answer is "demand for higher quality housing increased."

Item-response theory (IRT) for benchmark construction and results interpretation.

Rodriguez, P., Barrow, J., Hoyle, A. M., Lalor, J. P., Jia, R., & Boyd-Graber, J. (2021, August). Evaluation examples are not equally informative: How should that change NLP leaderboards?. In ACL-IJCNLP 2021

Connecting measurement theory to language technology eval

ECBD: Evidence-Centered Benchmark Design for NLP

Yu Lu Liu[1,2] **Su Lin Blodgett**[3] **Jackie Chi Kit Cheung**[1,2,4]
Q. Vera Liao[3] **Alexandra Olteanu**[3] **Ziang Xiao**[3,5]
[1]Mila – Quebec Artificial Intelligence Institute [2]McGill University
[3]Microsoft Research, Montréal, Canada [4]Canada CIFAR AI Chair [5]Johns Hopkins University
yu.l.liu@mail.mcgill.ca jackie.cheung@mcgill.ca
{sulin.blodgett,veraliao,alexandra.olteanu}@microsoft.com
ziang.xiao@jhu.edu

Abstract

Benchmarking is seen as critical to assessing progress in NLP. However, creating a benchmark involves many design decisions (e.g., which datasets to include, which metrics to use) that often rely on tacit, untested assumptions about what the benchmark is intended to measure or is actually measuring. There is currently no principled way of analyzing these decisions and how they impact the validity of the benchmark's measurements. To address this gap, we draw on evidence-centered design in educational assessments and propose Evidence-Centered Benchmark Design (ECBD), a framework which formalizes the benchmark design process into five modules. ECBD specifies the role each module plays in helping practitioners collect evidence about capabilities of interest. Specifically, each module requires benchmark designers to describe, justify, and support benchmark design choices—e.g., clearly specifying the capabilities the benchmark aims to measure or how evidence about those ca-

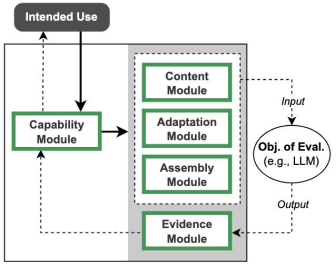


Figure 1: Simplified schema of the Evidence-Centered Benchmark Design (ECBD) framework. Solid line arrows indicate the process of designing a benchmark (e.g., designers should decide on the intended uses of the benchmark before deciding what capabilities are of interest). The dotted line arrows indicate the process wherein the benchmark gathers necessary evidence.

At the same time, as NLP models are increasingly believed to be more performant and to exhibit a wider range of capabilities, evaluation in NLP

Evidence-Centered Design framework for analyzing NLP benchmark design decisions and their impact on resulting measurements' validity.

Liu, Y. L., Blodgett, S. L., Cheung, J. C. K., Liao, Q. V., Olteanu, A., & Xiao, Z. (2024). ECBD: Evidence-Centered Benchmark Design for NLP. *ACL 2024*.

Evaluating evaluation is

a **process** of gathering multiple evidence to support the claim that a measurement accurately measures what it is intended to measure

- Messick (1994)

Messick, S. (1994). The interplay of evidence and consequences in the validation of performance assessments. *Educational researcher*, *23*(2), 13-23.

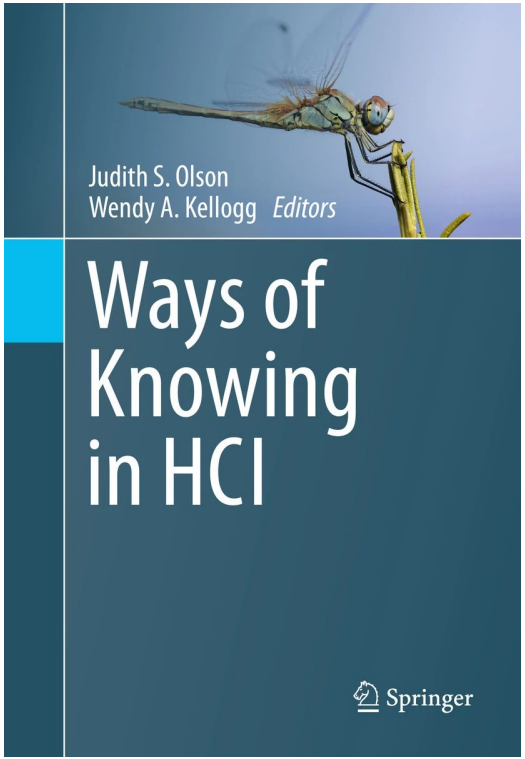
Evaluation Practices in Human-Computer Interaction (HCI)

Q. Vera Liao

Why HCI?

- A field that concerns itself with design and *evaluation* of technologies
  - Human-centered: evaluation of “human interaction”
- Interdisciplinary roots: inherits evaluation methods and desiderata from the social sciences
  - E.g., reliability and validity when designing quantitative measurements
- Embraces diverse methods beyond “human annotation/rating” used in NLP to get to: ***what*** (to evaluate), ***how well***, and ***why***
  - Often utilizes mixed-methods approaches (i.e., multiple methods in one study)

Many Ways of Knowing in HCI



Evaluation Methods in HCI

	Qualitative	Quantitative
Empirical	e.g., interview-based, ethnographic studies or think aloud	e.g., lab studies measuring completion time, error rate or surveys
Analytical	e.g., cognitive walk-through, heuristic evaluation	e.g., analysis of logs and cognitive models

Barkhuus & Rode. From mice to men-24 years of evaluation in CHI. In CHI 2007 EC

How to Choose? (more later)

Quantitative v.s. Qualitative?

- Research question: how well v.s. what or why
- Ecological validity
- Pragmatic costs

Empirical v.s. Analytical?

- Ecological validity
- Pragmatic costs

Empirical & Quantitative

- Lab studies with quantitative measurements
  - Task outcome measures
  - Behavioral measures
  - Subjective measures (e.g. with questionnaire)

- Survey studies (e.g. with close ended questions)

Crash Course on Quantitative Experimental Design

- What alternatives to compare? → **Experimental conditions**
  - E.g., with the new technique v.s. baseline without
- What effect(s) is the research question interested in?→ **Measurement(s)**
- Who are the target users?→ **Participant recruitment**
- What is the prototypical usage and the context?→ **Experimental task and procedure**
- What other factors might make a difference? → **Control variables or controlling in the experiment**

Will illustrate with examples in the next section

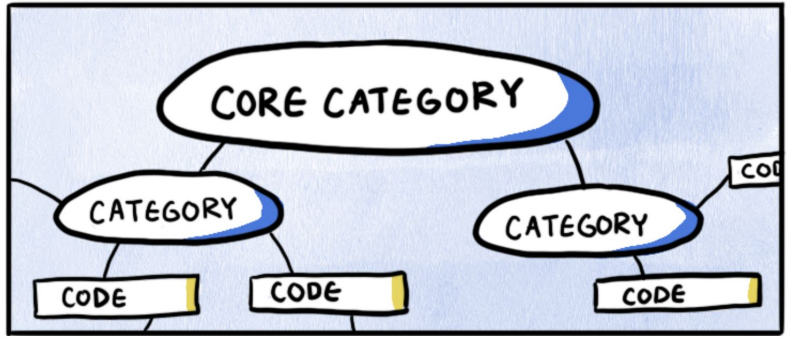
Empirical & Qualitative

- Interview
- Observational study

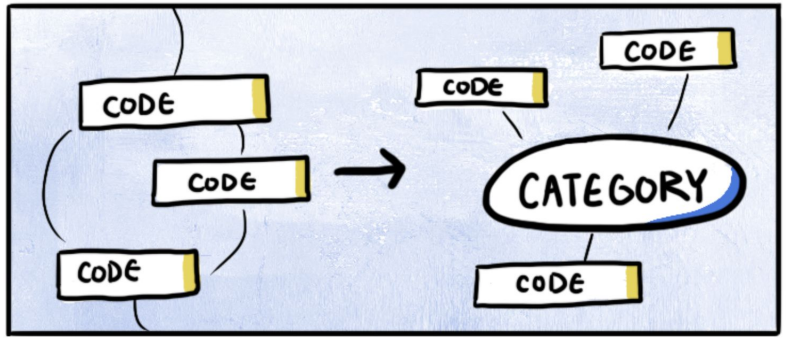
Crash Course on Interview Study

- Formative study (*what* and *why*) v.s. summative study (*how well*)
- Structured v.s. semi-structured v.s. Non-structured
- Data analysis using **grounded theory method**: iterative development of interpretation and theorizing

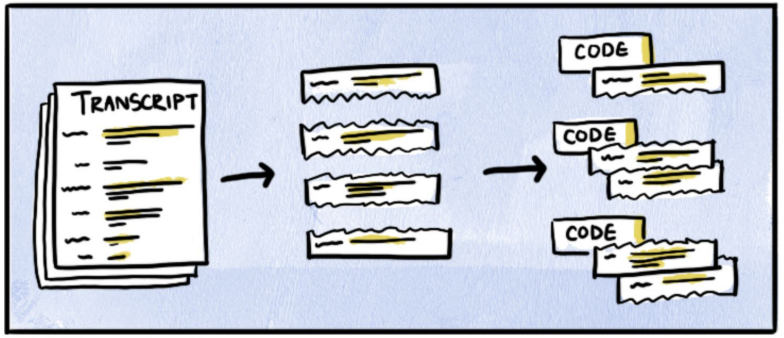
Step 3: Selective coding



Step 2: Axial coding



Step 1: Open coding



Analytical & Quantitative

- User modeling/simulation
  - Cognitive models to simulate how users would operate/click/browse
  - Agent-based modeling to anticipate outcomes of multi-user systems (e.g. social media platforms）

Analytical & Quantitative

- User modeling/simulation
  - Cognitive models to simulate how users would operate/click/browse
  - Agent-based modeling to anticipate outcomes of multi-user systems (e.g. social media platforms)

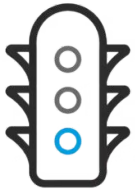
Lessons for LLM simulated evaluation?

- Theoretical grounding of how people would behave
- Rigorous validation with empirical human data

Analytical & Qualitative

- Cognitive walkthrough: domain/design experts simulate user interactions (e.g. to identify possible breakdowns)
- Heuristic evaluation: design experts rate interfaces based on usability heuristics

10 Usability Heuristics



Visibility of System Status

Match Between System & the Real World

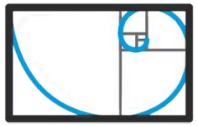
User Control & Freedom

Consistency & Standards

Error Prevention

Recognition Rather than Recall

Flexibility & Efficiency of Use



Aesthetic & Minimalist Design

Help Users Recognize, Diagnose & Recover from Errors

Help & Documentation

Interaction Design Foundation
interaction-design.org

Nielsen & Molich. Heuristic evaluation of user interfaces. *CHI 1990*

Analytical & Qualitative

- Cognitive walkthrough: domain/design experts simulate user interactions
- Heuristic evaluation: design experts rate interfaces based on usability heuristics

Lessons for human (experts) rating evaluation?

- Rigorously developed evaluation criteria and rating protocol
- Contextualize the rating: help the rater think through the criteria and think like the user

How to Choose Evaluation Method?

Quantitative v.s. Qualitative?

- ***Research question***: how well v.s. what or why
- ***Ecological validity/realism***: qualitative methods often engage more deeply with individual experience in the natural context
- ***Cost***: quantitative methods can (but not always) be less costly of researcher time and effort (e.g., when recruiting from crowdsourcing platform)

Empirical v.s. Analytical?

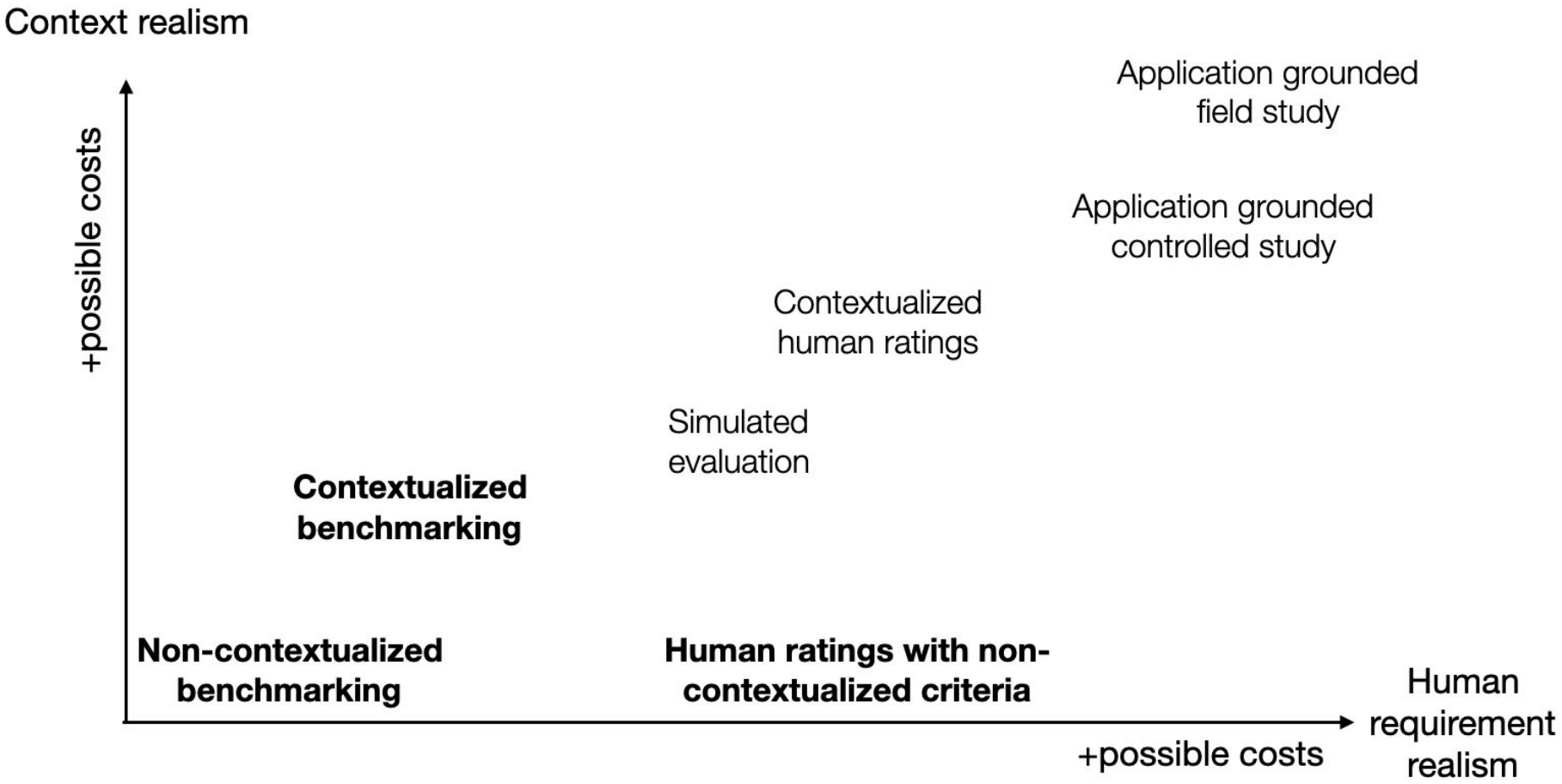
- ***Ecological validity/realism***: empirical methods are naturally more valid/realistic
- ***Cost***: analytical methods are less costly in researcher time, effort; also less or zero costs for users
- Analytical methods are often only used in the early stage of technical development or sensitive contexts

More on Realism/Ecological Validity

Ecological validity: whether one can generalize from the conclusions of a *laboratory study* to the real world (Schmuckler, 2001)

- **Context:** how close is the task or test environment to the real-world context?
- **Human response**: how well does the measurement represent people's actual response and is appropriate to the constructs that matter?
- **Stimuli**: how close is the stimuli (i.e. system behavior) used in the test to those encountered in real-world?

Realism: the situation or context within which the evidence is gathered, in relation to the contexts to which you want your evidence to apply (McGrath 1995)



Liao & Xiao, Rethinking model evaluation as narrowing the socio-technical gap. *ICML2023 Workshop on AI & HCI*

Take-Away

- Inform evaluation by understanding downstream use cases: contexts, user/stakeholder needs and behaviors, system behaviors
  - Start with “what”, utilize qualitative approaches
- Acknowledge “easy” approaches (e.g. automatic metrics, crowd ratings) are often compromising realism/validity for lower cost. We can improve by:
  - Better contextualization: reflect the usage contexts and user behavior in the test; articulate in what contexts the results can or cannot apply
  - Formalization and validation based on the “more realistic” approaches
- Embrace diverse evaluation approaches and justify your choices
  - E.g., Lower-cost, non-empirical approaches are often useful in early stage of technology development, but insufficient for systems that are impacting people’s lives

Example HCI Evaluation of Language Technologies

Q. Vera Liao & Ziang Xiao

Quantitative Empirical Evaluation with Human-Subjects

	Qualitative	Quantitative
Empirical	e.g., interview-based, ethnographic studies or think aloud	e.g., lab studies measuring completion time, error rate or surveys
Analytical	e.g., cognitive walk-through, heuristic evaluation	e.g., analysis of logs and cognitive models

Barkhuus & Rode. From mice to men-24 years of evaluation in CHI. In CHI 2007 EC

Crash Course on Quantitative Experimental Design

- What alternatives to compare? → **Experimental conditions**
  - E.g., with the new technique v.s. baseline without
- What effect(s) is the research question interested in?→ **Measurement(s)**
- Who are the target users?→ **Participant recruitment**
- What is the prototypical usage and the context?→ **Experimental task and procedure**
- What other factors might make a difference? → **Control variables or controlling in the experiment**

Use Case: Writing Support

Study 1: System Supporting Metaphor Creation for Science Writing

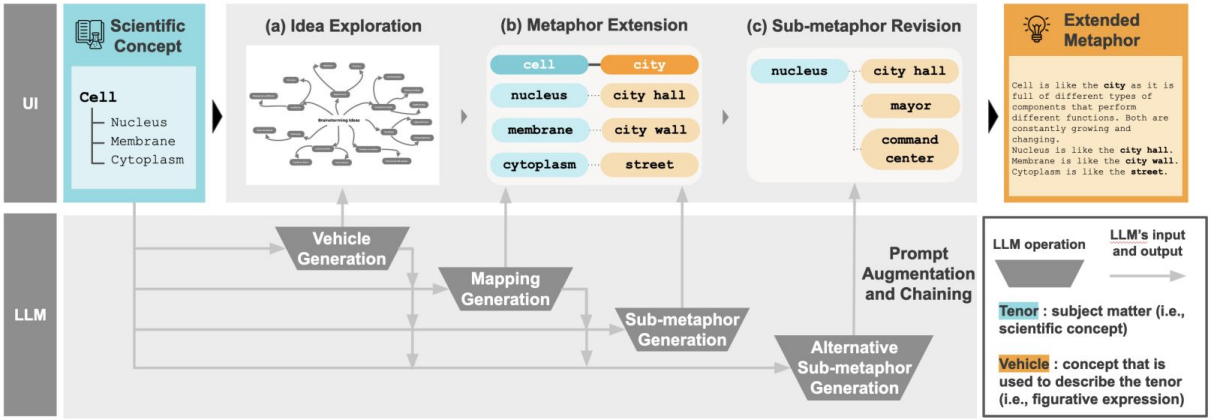
Metaphorian: Leveraging Large Language Models to Support Extended Metaphor Creation for Science Writing

Jeongyeon Kim
University of California, San Diego
San Diego, California, United States

Sangho Suh
University of California, San Diego
San Diego, California, United States

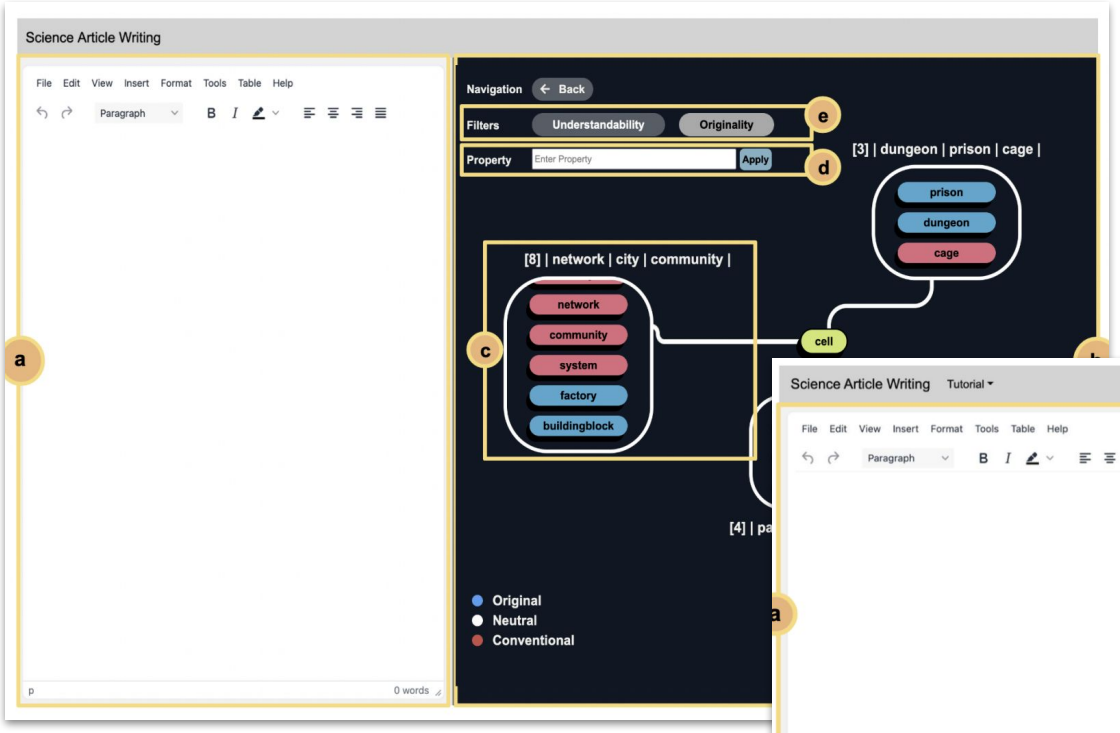
Lydia B Chilton
Columbia University
New York, New York, United States

Haijun Xia
University of California, San Diego
San Diego, California, United States

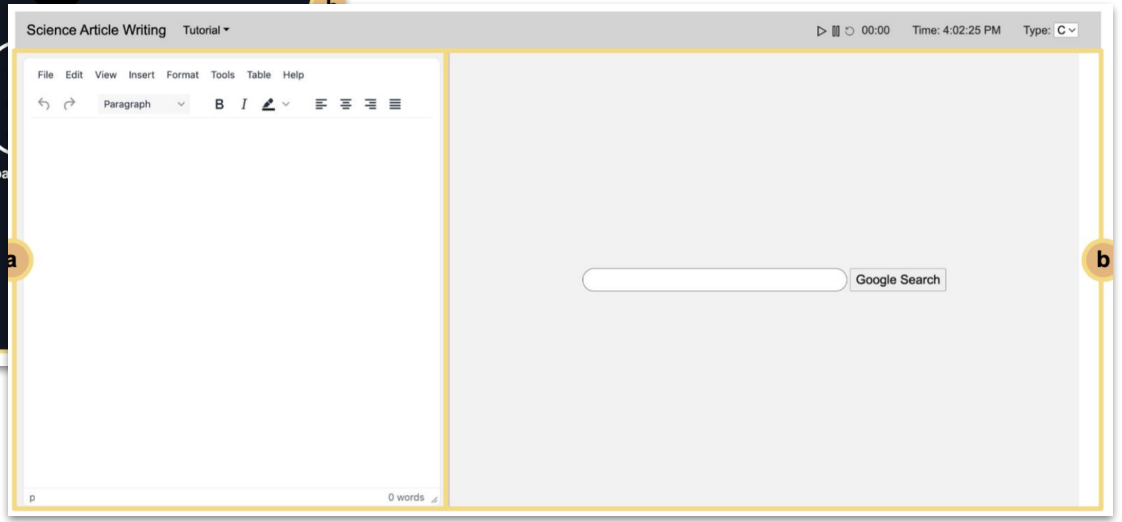


Experimental Design

- What alternatives to compare? Writing with Metaphorian v.s. Baseline interface without Metaphorian



Metaphorian



Baseline

Experimental Design

- What effects is the research question interested in?
  - Writing outcome quality: expert writers rated understandability, originality, scientific accuracy and overall quality
  - Writer experience: post-task survey on user satisfaction; subjective workload using NASA-LTX questionnaire;

Experimental Design

- Who are the target users? Experienced science writer, recruited from Upwork with publishing experience
- What is the prototypical usage and the context? Write a short article to explain a given scientific concept to the general public, with no strict time limit
- What other factors might make a difference? Participants were asked to write on one given topic and one topic of their own choosing

Study 2: Evaluating Influence of Opinionated LLM for Writing Support

Co-Writing with Opinionated Language Models Affects Users' Views

Maurice Jakesch
Cornell University
Ithaca, New York, USA
mpj32@cornell.edu

Advait Bhat
Microsoft Research
Bengaluru, India

Daniel Buschek
University of Bayreuth
Bayreuth, Germany

Lior Zalmanson
Tel Aviv University
Tel Aviv, Israel

Mor Naaman
Cornell Tech
New York, New York, USA

ABSTRACT

If large language models like GPT-3 preferably produce a particular point of view, they may influence people's opinions on an unknown scale. This study investigates whether a language-model-powered writing assistant that generates some opinions more often than others impacts what users write – and what they think. In an online experiment, we asked participants (N=1,506) to write a post discussing whether social media is good for society. Treatment group participants used a language-model-powered writing assistant configured to argue that social media is good or bad for society. Participants then completed a social media attitude survey, and independent judges (N=500) evaluated the opinions expressed in their writing. Using the opinionated language model affected the opinions expressed in participants' writing and shifted their opinions in the subsequent attitude survey. We discuss the wider implications of our results and argue that the opinions built into AI language technologies need to be monitored and engineered more carefully.

computer hardware and software architecture [97], large language models produce human-like language [56] by iteratively predicting likely next words based on the sequence of preceding words. Applications like writing assistants [38], grammar support [66], and machine translation [45] inject the models' output into what people write and read [51].

Using large language models in our daily communication may change how we form opinions and influence each other. In conventional forms of persuasion, a persuader crafts a compelling message and delivers it to recipients – either face-to-face or mediated through contemporary technology [94]. More recently, user researchers and behavioral economists have shown that technical choice architectures, such as the order of options presented affect people's behavior as well [42, 72]. With the emergence of large language models that produce human-like language [25, 56], interactions with technology may influence not only behavior but also opinions: when language models produce some views more often than others, they may persuade their users. We call this new paradigm of influence *latent persuasion* by language models, illustrated

Crash Course on Quantitative Experimental Design

- What alternatives to compare? → Experimental conditions
  - E.g., with the new technique v.s. baseline without
- What effect(s) is the research question interested in?→ Measurement(s)
- Who are the target users?→ Participant recruitment
- What is the prototypical usage and the context?→ Experimental task and procedure
- What other factors might make a difference? → Control variables or controlling in the experiment

Experimental Design

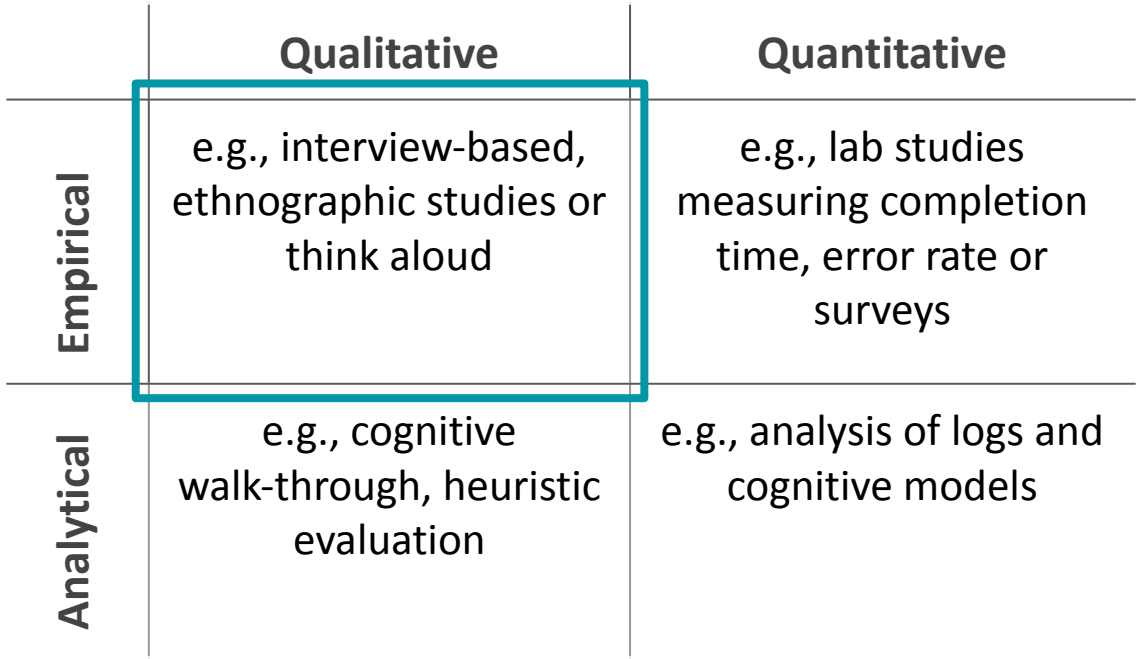
- What effects is the research question interested in? **Risk of LLM influencing writer's views**
  - Outcome measure of LLM's influence
    - Opinion expressed in writing, by crowd-worker rating position of each sentence, then calculate percentages of pro versus anti positions
    - Attitude change on topic, measured by the difference between self-reported attitude post- and pre-writing-task
  - Writing behaviors: how many suggestions accepted; how long paused to consider suggestions

Experimental Design

- What conditions/alternatives to compare?
- What other factors might make a difference? **Writer's original position**

(1) *Control group:* participants wrote their answers without a writing assistant.
(2) *Techno-optimist language model treatment:* participants were shown suggestions from a language model configured to argue that social media is good for society.
(3) *Techno-pessimist language model treatment:* participants received suggestions from a language model configured to argue that social media is bad for society.

Qualitative Empirical Evaluation with Human-Subjects



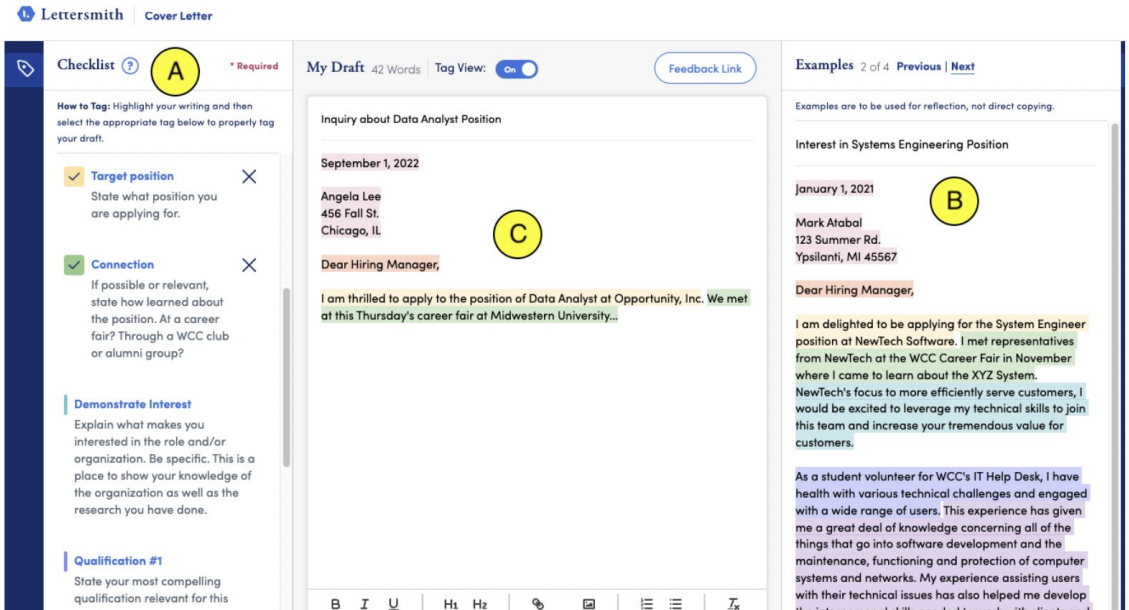
Barkhuus & Rode. From mice to men-24 years of evaluation in CHI. In CHI 2007 EC

Study 3: Evaluating Professional Communication Support

Lettersmith: Scaffolding Written Professional Communication Among College Students

Julie Hui
juliehui@umich.edu
School of Information, University of Michigan
Ann Arbor, Michigan, USA

Michelle Sprouse
sprouse@umich.edu
English and Education, University of Michigan
Ann Arbor, Michigan, USA



Interview Method

- **Situated experience**: recruited instructors to use the system in 7 communication/writing classes
- Interviewed 11 instructors and 19 students: their experience using the system, how it impacted them, whether or not they found it useful

Findings: Lettersmith Is Useful and *Why*

Students found Lettersmith useful for:

- Learning structure and content in a new genre
- Identifying language to express appropriate professional tone,
- Reflecting on their own writing

Instructors found that using Lettersmith:

- Helped them articulate writing task expectations
- Pinpoint where students had gaps in their understanding
- Scale instructional support for early-stage drafting

Use Case: Conversational AI

Analytical Methods

	Qualitative	Quantitative
Empirical	e.g., interview-based, ethnographic studies or think aloud	e.g., lab studies measuring completion time, error rate or surveys
Analytical	e.g., cognitive walk-through, heuristic evaluation	e.g., analysis of logs and cognitive models

Barkhuus & Rode. From mice to men-24 years of evaluation in CHI. In CHI 2007 EC

Heuristic Evaluation of Conversational Agents

Raina Langevin
rlangevi@uw.edu
Human Centered Design and Engineering, University of Washington
Seattle, WA

Ross Lordon
rolordon@microsoft.com
Microsoft
Redmond, WA

Thi Avrahami
thi@rul.ai
Rulai
Mountain View, CA

Benjamin Cowan
benjamin.cowan@ucd.ie
School of Information and Communication Studies, University College Dublin
Dublin, Ireland

Tad Hirsch
tad.hirsch@northeastern.edu
Department of Art + Design, Northeastern University
Boston, MA

Gary Hsieh
garyhs@uw.edu
Human Centered Design and Engineering, University of Washington
Seattle, WA

ABSTRACT

Conversational interfaces have risen in popularity as businesses and users adopt a range of conversational agents, including chatbots and voice assistants. Although guidelines have been proposed, there is not yet an established set of usability heuristics to guide and evaluate conversational agent design. In this paper, we propose a set of heuristics for conversational agents adapted from Nielsen's heuristics and based on expert feedback. We then validate the heuristics through two rounds of evaluations conducted by participants on two conversational agents, one chatbot and one voice-based personal assistant. We find that, when using our heuristics to evaluate both interfaces, evaluators were able to identify more usability issues than when using Nielsen's heuristics. We propose that our heuristics successfully identify issues related to dialogue content, interaction design, help and guidance, human-like characteristics, and data privacy.

CCS CONCEPTS

• **Human-centered computing** → **Heuristic evaluations**; **User interface design**.

KEYWORDS

heuristic evaluation, conversational agents, user interface design

ACM Reference Format:
Raina Langevin, Ross Lordon, Thi Avrahami, Benjamin Cowan, Tad Hirsch,

1 INTRODUCTION

Conversational agents are growing in popularity, through the uptake of text based and voice based conversational systems such as chatbots and Intelligent Personal Assistants (IPAs) respectively. Unlike other forms of human-computer interfaces, there is little consensus as to best practice for the design of conversational agents [5]. Recently there have been strides towards consolidating and validating guidance in related areas, such as human-AI interaction [1], and human-like chatbot experiences [24]. Our work looks to build upon recent efforts [20][26], to develop a comprehensive set of heuristics for conversational agent based interactions. The use of heuristics to guide design and evaluation is a widely used practice for interface design. Our research takes the approach of using Nielsen's heuristics [22] as a foundation upon which to build, adapting these for conversational agent based interaction.

We sought to expand on Nielsen's heuristics using a four phased design process. We first developed a set of heuristics for the design of conversational agent interfaces using prior research findings as well as our own experiences in developing these interfaces. Second, we presented these heuristics to nine experts in conversational agent design and heuristic evaluation, and incorporated their feedback. In the third phase, we evaluated our heuristics on two interfaces, a voice assistant on the Amazon Echo and an online chatbot. We compared our heuristics with Nielsen's heuristics to observe their effectiveness in identifying usability issues with conversational agents. After finding that the conversational agent heuristics

Heuristic Evaluation for Conversational Agent

Langevin, R., Lordon, R. J., Avrahami, T., Cowan, B. R., Hirsch, T., & Hsieh, G. (2021, May). Heuristic evaluation of conversational agents. CHI 2021

Match between system and the real world	**Consistency and standards**	**Error Prevention**	**Context preservation**	**Trustworthiness**
The system should **understand and speak the users' language**—with words, phrases and concepts familiar to the user and an appropriate voice…... Include dialogue elements that create a **smooth conversation** through openings, mid-conversation guidance, and graceful exits.	Users should not have to wonder whether **different words,options, or actions mean the same thing**…. Users should also be able to receive **consistent responses** even if they communicate the same function in multiple ways (and modalities). Within the interaction, the system should have a **consistent voice, style of language, and personality**.	Even better than good error messages is a careful design of the conversation and interface to reduce the likelihood of a problem from occurring in the first place. **Be prepared for pauses, conversation fillers, and interruptions, as well as dialogue failures, dead ends or sidetracks.** Proactively prevent or eliminate potential error-prone conditions, and **check and confirm with users before they commit an action**.	Maintain **context preservation** regarding the conversation topic intra-session, and if possible inter-session. Allow the user to reference past messages for further interactions to **support implicit user expectations of conversations**.	The system should convey trustworthiness by ensuring **privacy of user data**, and by being **transparent and truthful** with the user. The system should not falsely claim to be human.

Analytical Methods

	Qualitative	Quantitative
Empirical	e.g., interview-based, ethnographic studies or think aloud	e.g., lab studies measuring completion time, error rate or surveys
Analytical	e.g., cognitive walk-through, heuristic evaluation	e.g., analysis of logs and cognitive models

Barkhuus & Rode. From mice to men-24 years of evaluation in CHI. In CHI 2007 EC

A Conversation Analysis of Non-Progress and Coping Strategies with a Banking Task-Oriented Chatbot

Chi-Hsun Li, Su-Fang Yeh, Tang-Jie Chang, Meng-Hsuan Tsai, Ken Chen, Yung-Ju Chang
National Chiao Tung University, Hsinchu, Taiwan
{armuro}@cs.nctu.edu.tw, {funing314.iem05g}@g2.nctu.edu.tw,
{sfy.iem07g, tjchang.cs08g, kent31.iem05g}@nctu.edu.tw, {clairetsai818}@gmail.com

ABSTRACT
Task-oriented chatbots are becoming popular alternatives for fulfilling users' needs, but few studies have investigated how users cope with conversational 'non-progress' (NP) in their daily lives. Accordingly, we analyzed a three-month conversation log between 1,685 users and a task-oriented banking chatbot. In this data, we observed 12 types of conversational NP; five types of content that was unexpected and challenging for the chatbot to recognize; and 10 types of coping strategies. Moreover, we identified specific relationships between NP types and strategies, as well as signs that users were about to abandon the chatbot, including 1) three consecutive incidences of NP, 2) consecutive use of message reformulation or switching subjects, and 3) using message reformulation as the final strategy. Based on these findings, we provide design recommendations for task-oriented chatbots, aimed at reducing NP, guiding users through such NP, and improving user experiences to reduce the cessation of chatbot use.

Author Keywords
chatbot; conversation analysis; breakdowns; non-progress; coping strategies

CSS CONCEPTS
•Human-centered computing~Human computer interaction (HCI)~Interaction paradigms~Natural language interfaces

with their rapid growth in popularity; and we argue that this problem can be ascribed chiefly to lack of understanding of how users use chatbots in their daily lives. Various researchers have sought to develop better natural-language processing techniques, or to reduce recognition errors [22,26], since conversation breakdowns can be caused by difficulties with the complexities of natural-language [25].

Researchers have also started to develop guidelines for the chatbot interaction design. For instance, Jain et al. [12] explored how first-time users communicated with several kinds of chatbots and generated a set of guidelines based on the findings, and Ashktorab et al. [4] studied which strategies users prefer chatbots to adopt to repair conversation breakdowns. However, the resulting guidelines have thus far been based on studies in which the participants were given specific interaction instructions or scenarios. Therefore, their uses of chatbots were not driven by their own day-to-day needs, and the realism of the obstacles to human-chatbot interaction reported in these studies remains uncertain. Likewise, unknown are the frequency of these obstacles, how users deal with them, and which of them are most likely to cause users to break off communication with a chatbot. We argue that obstacles to conversation, or the non-progress (NP) of a conversation, between a human and a task-oriented chatbot are just as important to address as improving the usability of a website or mobile app. Moreover, it might be possible to anticipate NP and prioritize it for repair if we have

A Conversation analysis of a **three-month** conversation log between **1,685 users** and a task-oriented banking chatbot

Li, C. H., Yeh, S. F., Chang, T. J., Tsai, M. H., Chen, K., & Chang, Y. J. (2020, April). A conversation analysis of non-progress and coping strategies with a banking task-oriented chatbot. CHI 2020

Findings

Recognition Error	Mis-recognition	Non-recognition
Expected content	43.0%	45.2%
Unexpected content/Intention gaps		
Extra explanation	1.6%	2.5%
Restart the subject	0.4%	0.4%
Stay in the previous topic	0.4%	1.3%
Unfinished message	0.8%	2.5%
Finishing an unfinished message	0.7%	1.1%

Table 1. Non-progress types, by frequency

Message reformulation		
C1	add words	6.68%
C2	remove words	4.76%
C3	rephrase	8.82%
C4	repeat	5.75%
C5	ask new topic	5.48%
C6	others	3.56%
Quitting		
C7	quit subject temporarily	27.16%
C8	quit conversation temporarily	6.74%
C9	switch subject	13.47%
C10	abandon chatbot service	17.58%

Table 3. Users' strategies for dealing with non-progress

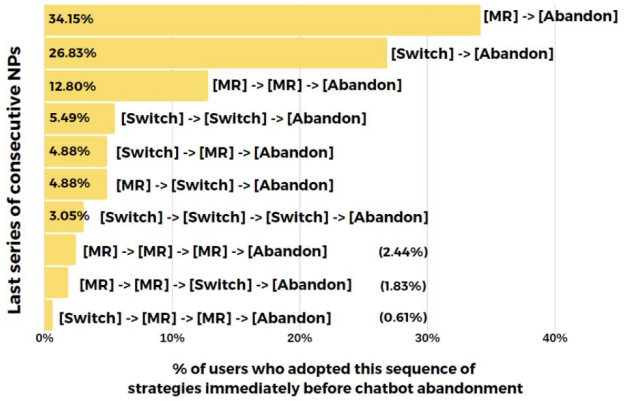


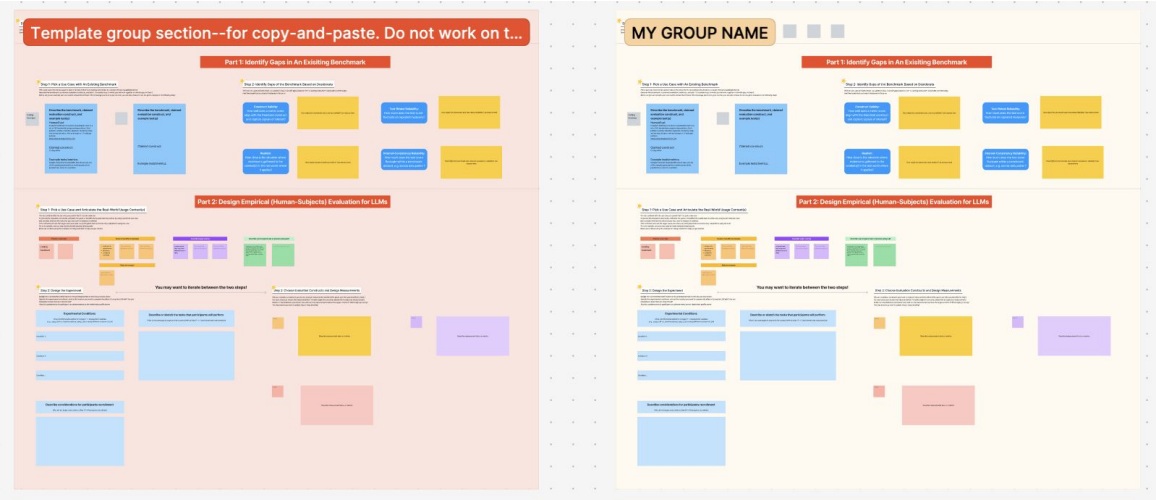
Figure 8. Relative use of message reformulation ("MR") vs. switching subjects ("Switch") as the user's final strategy before chatbot abandonment.

Reflection and Open Questions

Hands-on Exercise

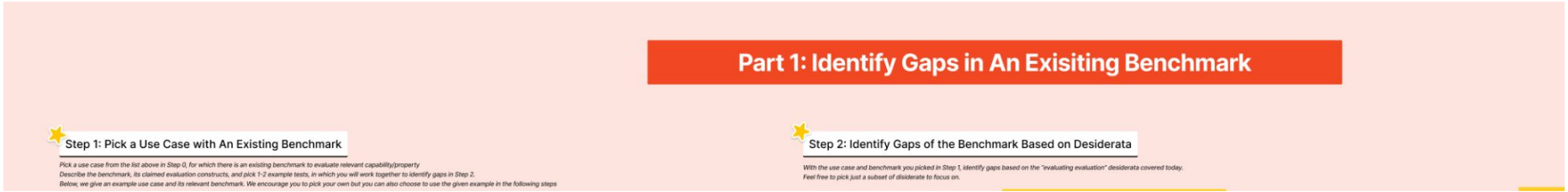
Getting Started

- Form teams of 4-6 people
- Create a working area for your team by copy-and-pasting from the template. Give your team a name!



- Or, write things down on sticky notes/notebook/papers

Exercise Overview



We encourage you to follow the steps and complete both parts
It is also ok if your team decide to focus on one task
If you get stuck, feel free to ask for help from the instructors

Step 0 (ice-breaker): List Common Downstream Use Cases of LLMs

Introduce yourself and brainstorm common use cases of LLMs.
Spend no more than 5-10 minutes on this part. You will ony need to pick one from these use cases to complete the exercise.

Exercise 1: Identify Gaps in an Existing Benchmark

Step 1: Pick a Use Case with An Existing Benchmark

Pick a use case from the list above in Step 0, for which there is an existing benchmark to evaluate relevant capability/property
Describe the benchmark, its claimed evaluation constructs, and pick 1-2 example tests, in which you will work together to identify gaps in Step 2.
Below, we give an example use case and its relevant benchmark. We encourage you to pick your own but you can also choose to use the given example in the following steps

Coding Assistant

Describe the benchmark, claimed evaluation construct, and example test(s)

HumanEval:

Evaluate functional correctness of generated code on a set of 164 handwritten programming problems. Each problem includes a function signature, docstring, body, and several unit tests, with an average of 7.7 tests per problem.
https://arxiv.org/abs/2107.03374

Claimed construct:

Coding ability

Example tests/metrics:

Pass@k measures the probability that at least one out of kkk samples generated by a model passes all the provided test cases for a problem.

Describe the benchmark, claimed evaluation construct, and example test(s)

Claimed construct:

Example tests/metrics:

Step 2: Identify Gaps of the Benchmark Based on Desiderata

With the use case and benchmark you picked in Step 1, identify gaps based on the "evaluating evaluation" desiderata covered today. Feel free to pick just a subset of disiderate to focus on.

Criterion-related Validity: How well does the benchmark results align with the theorized construct and capture signals of interest?

How might the benchmark lack construct validity? List reasons here

Test-Retest Reliability: How much does the test score fluctuate on repeated measures?

How might the benchmark lack test-retest reliability? List reasons here

Ecological Validity: How well does the result generalize to the real-world contexts where the technology will be used?
(think about the usage context, human behaviors and system behaviors)

How might the benchmark lack ecological validity? List reasons here

Internal-Consistency Reliability: How much does the test score fluctuate within a benchmark dataset, e.g. across data points ?

How might the benchmark lack internal-consistency reliability? List reasons here

Exercise 2: Design An Empirical Evaluation Study (for an LLM use case)

Part 2: Design An Empirical (

Step 1: Pick a Use Case and Articulate the Real-World Usage Context(s)

You can continue with the use case you used in Part 1 or pick a new one
To ground the evaluation constructs, articulate the goals or benefits that people want to achieve by using LLM in this use case, also consider what are the risks that you may want to measure in addition.
Then articulate who are the target users and what is a prototypical task or process they will perform using the LLM.
Try to be specific, and you may want to iterate between these points.
Below we continue using the example of coding assistant to help you get started.

Choose a use case

Coding Assistant

Goals or benefits to measure

- coding task performance
- efficiency
- subjective workload

Type anything, @mention anyone

Type anything, @mention anyone

Risks to measure

Type anything,

Describe target user(s)

programmers, who may have different level of skills

Type anything, @mention anyone

Type anything, @mention anyone

Describe a prototypical task or process using LLM

A programmer has a coding task to complete. They prompt the LLM with a description of the task. Then they will check and edit the output until satisfied. They may also choose to revise the prompt to get an updated output

Type anything, @mention anyone

Step 2: Design the Experiment

You may want to iterate between the two steps!

Design the experimental task(s) based on the prototypical task or process you have above.
Specify the experimental conditions: what is the baseline you want to compare the effect of using the LLM with? Are you interested in more than one kind of LLM?
Also list considerations for participant recruitment based on the stakeholder profile above

Experimental Conditions

What alternatives to compare?
(e.g., using LLM v.s. baseline without using LLM; or using different versions of LLM)...

Condition 1:

Condition 2:

Condition ...

Describe or sketch the tasks that participants will perform

What is the prototypical usage and the context (refer to Step 1)?

Describe considerations for participants recruitment

Who are the target users (refer to Step 1)?

Step 3: Choose Evaluation Constructs and Design Measurements

Choose evaluation constructs you want to measure based on the LLM benefits, goals and risks you identified in Step 1.
For each construct, design the measurement(s). Consider objective outcome, behavioral or subjective measurements
Reflect on the limitations and iterate: how well can the measurement capture the target construct? What might go wrong?
You may need to go back to update Step 2. Keep iterating!

Construct 1

Describe measurement tests or metrics

Questions?

EMNLP 2024

Human-Centered Evaluation of Language Technologies

Su Lin Blodgett Microsoft Research Montréal
Jackie C.K. Cheung McGill University
Q. Vera Liao Microsoft Research Montréal
Ziang Xiao Johns Hopkins University

Blodgett, S. L., Cheung, J. C. K., Liao, V., & Xiao, Z. (2024, November). Human-Centered Evaluation of Language Technologies. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Tutorial Abstracts* (pp. 39-43).

Microsoft

McGill

Johns Hopkins University