

# The Easy, the Hard and the Ugly

## ✓ Easy

- ✓ Inference (*fit*)
- ✓ Fine-Tuning (*predict*)

## ✓ Hard

- ✓ Pre-training

## ➤ Ugly (Responsible AI)

- Bias
- Toxicity
- Misinformation
- Hallucinations
- Plagiarism



# Ugly: Outline

- **Benchmarking** (Omar)
- Smooth-Talking Machines/Trust (Ken)

# Evaluating Evaluations: A Perspective on Benchmarks

# The importance of evaluation

- More and more benchmarks, data sets and evaluations available
- Industry ships a new product/feature ...
  - ... users test it, high expectations, harsh reaction ...
  - ... back to the lab to fix things, iterate; very expensive process
- As a community:
  - Are we really evaluating the right stuff?
  - Are we using solid principles to construct and maintain benchmarks?
  - What are we learning?

# Validity and Reliability

Krippendorff (2018) *Content analysis: An introduction to its methodology*

- Reliability is about data, and validity is about truth:
  - Reliability:
    - The attribute of Data on which researchers can rely in answering their Research questions – Krippendorff, p. 411
  - Validity:
    - The quality of a claim to be as stated, true, or correct. – Krippendorff, p. 413
- Validity assumes a hypothesis/claim.
  - Hard to test validity without a clearly stated hypothesis
- There are more papers on reliability than validity in our field(s).
  - 18k matches for “inter-annotator agreement” in ACL Anthology
  - <https://aclanthology.org/search/?q=inter-annotator+agreement>

# Reliability of BLEU

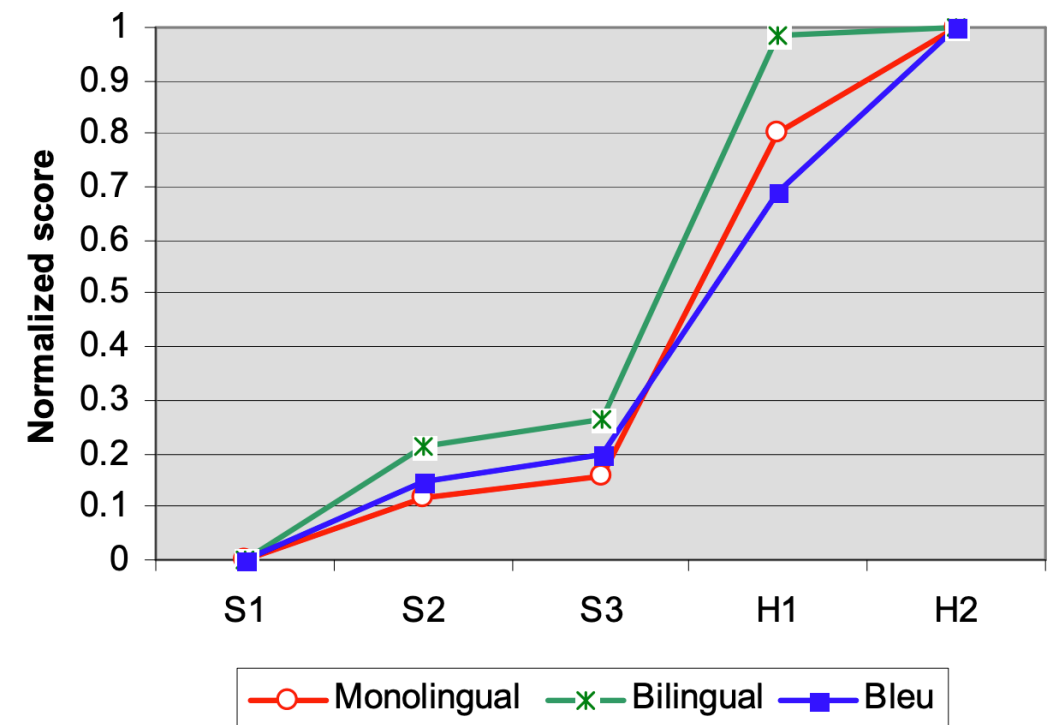
<https://aclanthology.org/P02-1040/>

- The original paper asks 3 excellent questions about reliability
  - How reliable is the difference in BLEU metric?
  - What is the variance of BLEU score?
  - If we were to pick another random set of 500 sentences,
    - would we still judge S3 to be better than S2?

# Reliability of BLEU

- In their Figure 7,
  - they use BLEU to compare three machine translation systems (S1, S2 and S3) and
  - two human (non-professional) translators (H1 and H2).
- They report that humans score better than machines.
- They conclude:
  - “BLEU’s strength is that it correlates highly with human judgments
    - by averaging out individual sentence judgment errors over a test corpus
    - rather than attempting to divine the exact human judgment for every sentence:
    - *quantity leads to quality.*”

Figure 7: BLEU vs Bilingual and Monolingual Judgments



<https://aclanthology.org/P02-1040/>



# Validity of BLEU

- Less discussion of validity in original paper
- Some papers raise serious questions about validity
  - of BLEU and other metrics
  - at least, for some use cases that go beyond the original BLEU paper
- It is hard to talk about validity without a clearly stated use case
  - <https://aclanthology.org/D17-1238/>
    - *This paper shows that state-of-the-art automatic evaluation metrics for NLG systems*
      - *do not sufficiently reflect human ratings,*
      - *which stresses the need for human evaluations.*
  - <https://aclanthology.org/J18-3002>
    - *Overall, the evidence supports using BLEU for diagnostic evaluation of MT systems*
      - *(which is what it was originally proposed for),*
    - *but does not support using BLEU outside of MT, for evaluation of individual texts, or for scientific hypothesis testing.*



# Use Cases

- BLEU was originally proposed
  - to compare a small number of systems for DARPA competitions.
- But soon after BLEU was introduced,
  - Och suggested using BLEU for a very different use case.
- It had been standard practice to to use
  - different metrics for testing and training.
- Och found that if one is going to use BLEU to evaluate systems,
  - then his system would do better in the evaluation if he also used BLEU for training.
- Och's suggestion did well in competitions,
  - but raises questions about reliability and validity

# Consequence of Evaluation: Proposed Scale

- Minor: e.g., SOTA-chasing, leaderboards
- Moderate: e.g., Multitask learning
  - Generalizing results over workloads and tasks.
- Major: e.g., Och training
  - Significant consequences for system performance
- Mission Critical: e.g., SPEC
  - What should I buy? And what performance should I expect on my workloads?
  - Go/No-go decisions

# Averaging: Arithmetic vs. Geometric

- There are many benchmarks in our field:
  - e.g., GLUE, SciRepEval, MS MARCO and Big Bench
- Many designed for SOTA-chasing,
  - but hopefully, results will generalize to more important use cases.
- More likely to generalize to more important use cases
  - if they were designed to do so in the first place
- SPEC was designed to report performance relative to baseline
  - How much better is the candidate CPU relative to VAX 11/780?
  - On the user's (unspecified) workload?
- Mashey argues that
  - geometric means generalize better
  - over workloads than arithmetic means.
- Mashey suggests that results on our benchmarks would
  - generalize beyond less important SOTA-chasing use cases
  - if we replaced arithmetic means with geometric means in GLUE
  - (and many of our other benchmarks).

# SPEC: A benchmark for evaluating CPUs

Code	App Area	Lines	Remarks
gcc	Compiler	87,800	CNU C Compiler V1.35, compiles 76 soruces, 10% I/O
Espresso	Logic De- sign	14,800	PALs generation tool, heuristic minimization, little paging
Li	Interpreter	7700	Lisp interpreter (XLIST 16), solves 8-queens problem using recursive backtracking, many jumps/loops
Eqntott	Logic design	3500	Creates truth tables; >95% of time in qsort
Compress	Data com- pression	1500	Compress/decompress 1MB file 20 times using adaptive Lempel-Ziv coding
Sc	Spreadsheet	8500	Spreadsheet app based on the Unix "curses"

**Table 1: SPEC CINT92 suite (from Table 2 in [8])**

Code	App Area	Lines	Remarks
Spice2g6	Circuit De- sign	18,900	Analog circuit simulation tool, unvectorizable, unparallelizable, uses cache
Doduc	Physics, simula- tion	5300	Monte Carlo simulation of thermohydraulic neclear reactor, unvectorizable, many jumps/loops
Fpppp	Quantum chemistry	2700	Electron integral, unvectorizable, no jumps (good for pipelining)
Tomcatv	Geometry	200	Mesh generation, 90-98% vectorizable, exercises data cache
Nasa7	Kernels	1300	Some kernels are vectorizable
Mdljdp2	Chemistry	4500	Motion equations for 500-atom model
Wave5	Physics	7600	Particle and Maxwell's equations
Ora	Optics	500	Ray Tracing
Alvinn	Robotics	300	Neural network propagation training
Ear	Medicine	5200	Ear simulation using FFT
Mdljsp2	Chemistry	3900	Single precision of Mdljdp2
Swm256	Simulation	500	Shallow water equations system
Su2cor	Quantum physics	2500	Masses of elementary particles, 98.5% vectorizable
Hydro2d	Astrophysics	4500	Calactic jets, 99.5% vectorizable

**Table 2: SPEC CFP92 suite (from Table 3 in [8])**

# Challenges for Validity and Reliability

- Unrepresentative Samples
- Test/Train Splits: Interpolation vs. Extrapolation
- Leakage
- Labeling and Inter-annotator Agreement
- Too many (irrelevant) tasks

# Unrepresentative Samples

- Ideally, a benchmark should be
  - a representative sample
  - of a larger population of interest.
- **Balanced Corpora:**
  - 1960s: Brown Corpus
  - 1990s: British National Corpus
- **Current view: catch as catch can**
  - *there is no data like more data* – Mercer
- **Representative samples → More credible generalizations**

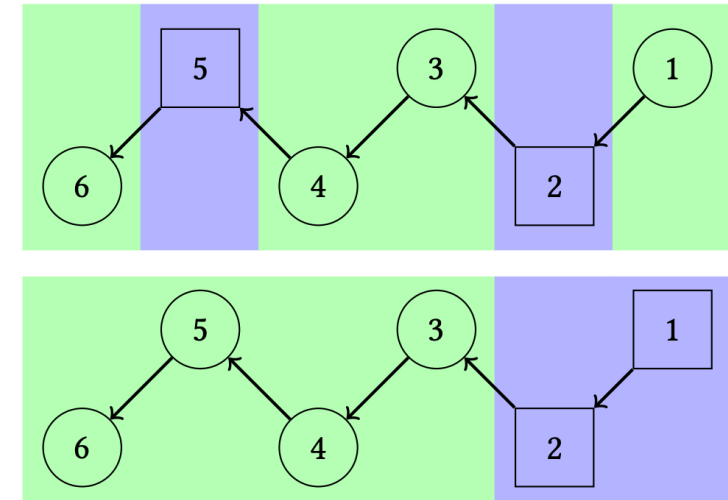
	Semantic Scholar		Citation Task	
	%	N	%	N
$ A $	47%	99M	82%	21,885
$ L $	53%	111M	97%	25,850
$ A \cup L $	70%	145M	99%	26,378
$ A \cap L $	31%	65M	80%	21,357
totals	100%	208M	100%	26,657

**Table 3: Comparison of Semantic Scholar with a benchmark (SciRepEval Cite task). There are large mismatches in  $|A|$  (papers with abstracts) and  $|L|$  (papers with links in  $G$ ).**

# Test/Train Splits: Interpolation $\neq$ Extrapolation

## *It's Difficult to Make Predictions, Especially About the Future*

- It is common for graphs to evolve over time.
  - For example, the academic literature is growing very quickly,
  - doubling every nine years
- Benchmarks such as OGB focus on static snapshots from a few years ago,
  - missing opportunities to encourage the community to study growth and timeliness.
- Random splits are common in graph learning benchmarks, e.g., WN18RR,
  - a popular Knowledge Graph Completion (KGC) task based on WordNet



**Figure 1: Random Splits (top) vs Causal Split (bottom).**

Paper	Year	Title
1	2018	[...] Photogrammetric imaging
2	2016	Convenient probe of S(1D2)[...]
3	2005	Megapixel ion imaging [...]
4	2003	Direct current slide imaging [...]
5	1995	profiles of CI(2Pj) photofragments [...]
6	1988	Adiabatic dissociation of [...]

**Table 4: 1 cites 2, 2 cites 3,..., 5 cites 6**



# Leakage

- Leakage is common in many benchmarks
  - There is considerable discussion of leakage in SciDocs
    - (see 4.2 of <https://aclanthology.org/2022.emnlp-main.802/>)
  - WN18 → WN18RR (WordNet benchmarks)
    - If  $x$  is-a  $y$  (a car is a vehicle),
      - then there will be two links between  $x$  and  $y$ : hypernym and hyponym
    - Since WN18 randomly splits links into test and train,
      - one of these links is likely to be in test and the other in train
    - Unfortunately, WN18RR corrects some (but not all) of the leakage
      - See table 4 of <https://aclanthology.org/2021.emnlp-main.501/>
    - Despite this leakage, there are many papers on WN18RR
      - <https://paperswithcode.com/dataset/wn18rr>

# Labeling and Inter-annotator Agreement

- The documentation on SciRepEval makes it clear that some labels are “silver” (less reliable) [underlining added]:
  - ... a new large-scale field of study (FoS) multi-label training set of more than 500K papers with silver FoS labels based on publication venue
- We compared FoS labels in SciRepEval with FoS labels in MAG and found large differences.
  - More agreement in some fields (Computer Science)
    - Less agreement in History, Sociology and Art.
  - It is possible that the annotators
    - are more familiar with Computer Science
    - than History, Sociology and Art.

# The bad news first

- Labeling is hard
  - Facebook
  - Points-Of-Interests (Foursquare, etc.)
- Labeling is going to get more difficult
  - Enterprise
  - Personalization
  - Healthcare
  - New data sets
- Some context
  - We assume supervised or semi-supervised learning
  - Large scale
  - Continuous

# What is a label?



● **M. Choi Young Deuk .** <info@undptours.com>  
To: oralonso@yahoo.com



Wed, Mar 31 at 8:16 AM

--

Hello Omar Alonso,

I am a banker working with CIMB bank Cambodia. I contacted you for a reason, one of my late customer have the same family name as yours. He died 6 years ago and left 10.7 million United States dollars in his account. Since then no relative have come to claim his money .. I think we can work things out.

Best regards,

M. Choi Young Deuk .



● **Chase**  
From: contact@melendezlillian.com  
To: oralonso@yahoo.com

**CHASE** 

**Crucial Message**

Message ID: [TR3D71452024](#)



Spam email?  
Label: yes, no

Dear Valued Client :  
oralonso@yahoo.com

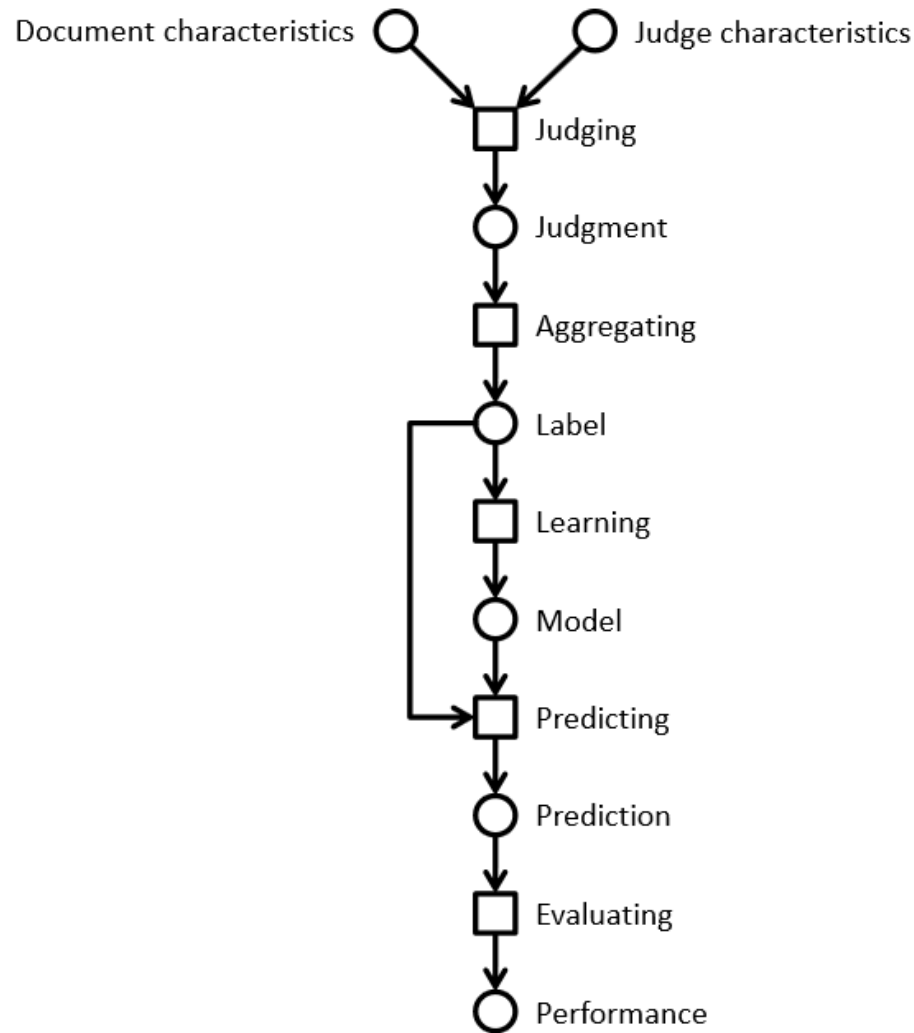
An imperative communication from **Chase** necessitates your immediate consideration. Neglecting to reply promptly could lead to restrictions on your account.

# Why we care?

- Provenance
- Reproducibility & debugging
- Explainability & interpretability
  - How a training set was created
- Bias and fairness
- Data management
  - ML/AI models live & die by the quality of input data
  - Metadata about labels
  - Maintenance

# Lifecycle of a label

- Information retrieval example



Using a crowd to label a data set

Using ML to process the complete data set

# Relevance labels

- Indicate whether a search result is valuable to a searcher
- Key in evaluation and optimization IR systems
- Editors or experts
  - TREC-style
- Crowdsourcing
- LLMs



# Careful with that ~~axe~~ data, Eugene

- In the era of big data and machine learning
  - labels -> features -> predictive model -> optimization
- Labeling perceived as boring
- Tendency to rush labeling
- Quality is key
  - Garbage in, garbage out
- Own the entire stack
  - Labeling, modeling, infrastructure, deployment

# The state of the field

- Human-labeled data is more important than ever
- Requirements
  - Throughput -> ASAP; I need the labels for yesterday
  - Cost -> cheap; if possible free
  - Quality -> top
- Performed as a one-off by 3<sup>rd</sup> party (crowd or editors)
  - Human Intelligence Task (HIT)
  - Micro-tasks
- Needs development work to get good results
- Very limited functionality in current platforms
  - Mechanical Turk, SageMaker (Amazon)
  - Figure Eight (Appen)
  - Toloka (Yandex)
  - Start-ups
- LLMs

# The need for humans

- Many examples where humans are involved
- Adult content and moderation
- Baby sitting algorithms

## The Laborers Who Keep Dick Pics and Beheadings Out of Your Facebook Feed

BY ADRIAN CHEN 10.23.14 | 6:30 AM | PERMALINK  
Share 60.5k Tweet 7,274 718 674 Pin it



<https://www.wired.com/2014/10/content-moderation/>

## 'A Permanent Nightmare': Pinterest Moderators Fight to Keep Horrifying Content Off the Platform

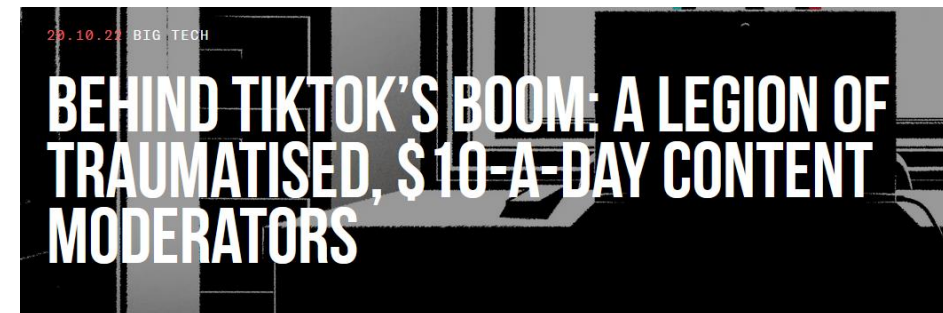
Moderators reported seeing child pornography content 'every couple hours'



Sarah Emerson · Follow

Published in OneZero · 11 min read · Jul 28, 2020

<https://onezero.medium.com/a-permanent-nightmare-pinterest-moderators-fight-to-keep-horrifying-content-off-the-platform-4d8e7ec822fe>



<https://www.thebureauinvestigates.com/stories/2022-10-20/behind-tiktoks-boom-a-legion-of-traumatized-10-a-day-content-moderators>

# Problems

- Monolithic HITs
  - The structure of a HIT mirrors the structure of the task the developer is working on
  - Similar to Conway's law in software engineering
- Task complexity
- Lengthy instructions
  - RTFM doesn't work
- We don't think of HC/crowdsourcing as programming
- How to improve
  - Use established programming practices
  - Careful, we are dealing with humans and not machines

# A spectrum of labeling tasks

Nature of task	Aggregation approach	Evaluation technique
Objective question has a correct answer (objective)	Reliable judge assigns appropriate label for an item	Evaluate workers by comparing individual results with gold set
Judgment question has a best answer (partially objective)	Inter-rater agreement determines label for an item	Evaluate workers by comparing individual results with consensus
Subjective question has consistent answer (subjective)	Repeatable polling determines probability of a label for an item	Evaluate workers by computing the consistency of results between groups

# Prepare the environment

- Homework before you label
  - Assess the lay of the land
  - Identify your use cases
  - Understand your product's data
  - Design your HITs
  - Determine your guidelines
  - Communicate your task
  - Maintain high quality
- Ongoing vs. one-offs HITs
- Labels for the machine != labels for humans

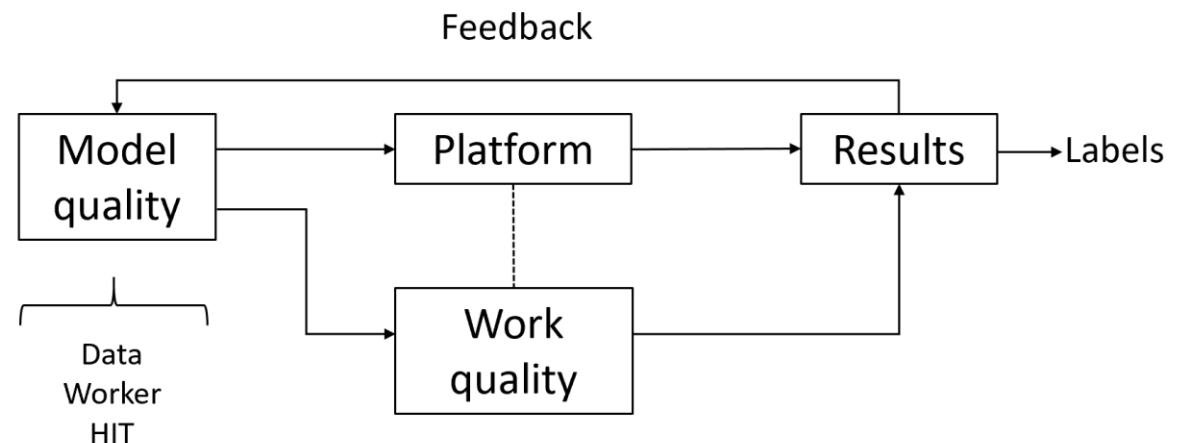
# HIT design principles

- Self-contained, short, and simple
- Document presentation
  - Text alignment & legibility; reading level; multi-cultural and multilingual
- Cognitive biases
  - Implications on the final output: anchor effect, mere exposure, picture superiority
- Task complexity
  - High cognitive load; low usability, specific expertise



# Quality control in general

- Extremely important part of the task
- Approach as “overall” quality; not just for workers
- Bi-directional channel
  - You may think the worker is doing a bad job.
  - The same worker may think you are a lousy requester.
- Quality framework
  - Module quality
  - Work quality
- Measuring agreement



# Algorithms used in practice

- Voting
  - Majority vote, Borda, tiers
  - Strong baseline
- Honey pots and programmatic gold
- Expectation-Maximization
- Get another label
- Adaptivity
  - Quality-cost tradeoff
  - How many workers?
  - When to stop?
  - Stopping rules
  - Automatic honey pots creation

# Behavioral features

- Focus on the way workers work instead of what they produce
- Task fingerprinting
- High correlation with work quality
- Wernicke
  - Information Extraction scenario
  - Weighted majority voting
  - Behavioral features outperform performance-based methods

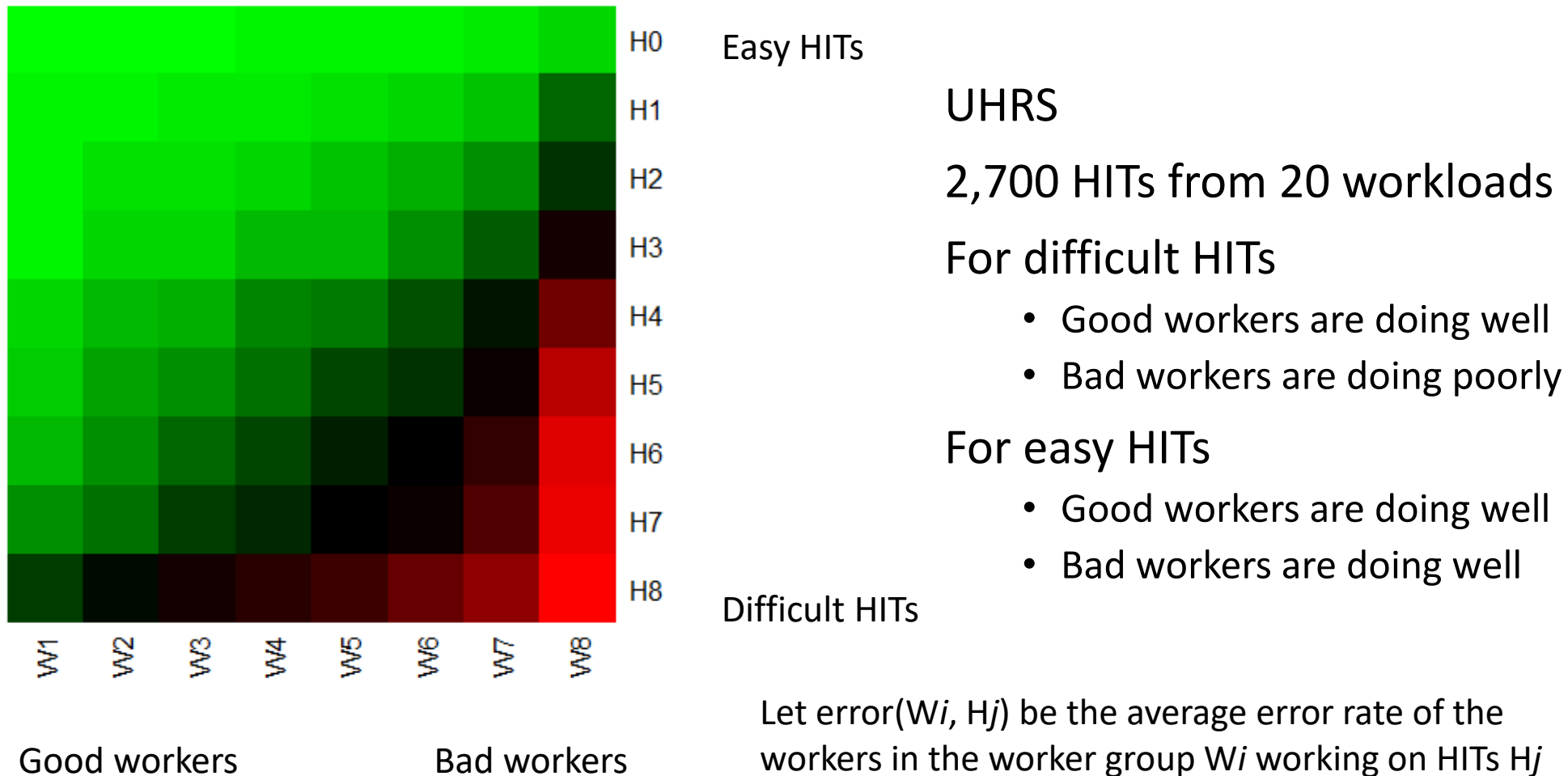
J. Rzeszotarski and A. Kittur. “Instrumenting the Crowd: Using Implicit Behavioral Measures to Predict Task Performance”. UIST 2011.

S. Han, P. Dai, P. Paritosh, D. Huynh. “Crowdsourcing Human Annotation on Web Page Structure: Infrastructure Design and Behavior-Based Quality Control”. ACM TIST 2016

# Active learning

- Accuracy
  - Limited budget for annotating a small % of the unlabeled data
- Speed
  - Model more accurate more quickly
- Diversity
- Uncertainty sampling
  - Least confidence, margin of confidence, ratio of confidence
- Diversity sampling
  - Clustering to partition the data, real-world diversity

# Error rates for different worker/HIT groups



# Snorkel approach

- Formalizing programmatic labeling
- Models are commodities
  - `pip install <what-you-want>`
- Training data is the interface for software 2.0
- Labeling functions as black boxes that predict a label
- Learn from agreements/disagreements between labeling functions

# LLMs

- Human labels are expensive
  - Expert > Crowd-based worker > LLM
  - Automatic label is not a new idea
- How about using LLMs to label documents?
- Potential advantages
  - Cost and performance
  - Allocate humans where are needed the most
- Potential issues
  - Reliability
  - Quality control



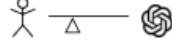
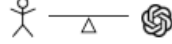
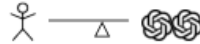
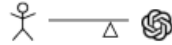
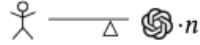

P. Thomas et al. "Large language models can accurately predict searcher preferences" [arxiv.org/abs/2309.10621](https://arxiv.org/abs/2309.10621)

G. Faggioli et al. "Perspectives on Large Language Models for Relevance Judgment" ICTIR 2023



# Spectrum of human-machine collaboration

- LLMs judgement quality
- LLMs cost
- Multiple LLMs as judges
- Truthfulness
- Bias
- Explanations/justifications

Collaboration Integration	Task Organization
Human Judgment	
	The human will do all judgments manually without any kind of support.
	Humans have full control of judging but are supported by text highlighting, document clustering, etc.
AI Assistance	
	The human assessor judges an LLM-generated summary of the document.
	Balanced competence partitioning. Humans and LLMs focus on tasks they are good at.
Human Verification	
	Two LLMs each generate a judgment, and a human selects the better one.
	An LLM produces a judgment (and an explanation) that a human can accept/reject.
	LLMs are considered crowdworkers, varied by specific characteristics, and controlled by a human.
Fully Automated	
	Fully automatic assessment.

# Prompting

- In-context learning
- New capabilities can be unlocked in LLMs
- LLM is prompted with a few in-context demonstrations
- Learns to perform a certain task
- Task performance is very sensitive to prompts

# Setup

- Similar to crowdsourcing work
- Take TREC judgement guidelines
  - HIT in Mturk
  - Prompt for GPT or similar LLM
- Compute agreement using Cohen's kappa
- Two main approaches
  - Prompt “as is”
  - Prompt engineering

# Prompt structure

- Relevance evaluation task
- Task instructions
  - You are a search quality rater evaluating relevance of web pages
- Query-document pair to be labelled
  - Query {query}
  - Document {document}
  - Relevant?
- Re-state the task
- Output format

# Prompts compared to HITs

role      You are a search quality rater evaluating the relevance of web pages. Given a query and a web page, you must provide a score on an integer scale of 0 to 2 with the following meanings:

- 2 = highly relevant, very helpful for this query
- 1 = relevant, may be partly helpful but might contain other irrelevant content
- 0 = not relevant, should never be shown for this query

Assume that you are writing a report on the subject of the topic. If you would use any of the information contained in the web page in such a report, mark it 1. If the web page is primarily about the topic, or contains vital information about the topic, mark it 2. Otherwise, mark it 0.

## Query

description,  
narrative

A person has typed [query] into a search engine.

They were looking for: *description narrative*

## Result

Consider the following web page.

—BEGIN WEB PAGE CONTENT—

*page text*

—END WEB PAGE CONTENT—

## Instructions

Split this problem into steps:

Consider the underlying intent of the search.

aspects

Measure how well the content matches a likely intent of the query (M).

aspects

Measure how trustworthy the web page is (T).

Consider the aspects above and the relative importance of each, and decide on a final score (O).

multiple

We asked five search engine raters to evaluate the relevance of the web page for the query. Each rater used their own independent judgement.

Produce a JSON array of scores without providing any reasoning. Example: [{"M": 2, "T": 1, "O": 1}, {"M": 1 . . .

## Results

[{

## Document Relevance Evaluation

Please evaluate the **relevance** of a document to the given topic. A document is relevant if it directly discusses the topic. Each document should be judged on its own merits. That is, a document is still relevant even if it is the thirtieth document you have seen with the same information.

### Tips

- Payment based on quality of the work completed. Please follow the instructions and be consistent in your judgments.
- **Bonus** payment if you provide a good justification
- Please justify your answer, otherwise you may not get paid.
- A document should not be judged as relevant or irrelevant based only on the title of the document. You **must** read the document.

### Task

Please evaluate the relevance of the following document about **art, stolen, forged**.

Description: What incidents have there been of stolen or forged art?

More information: Instances of stolen or forged art in any media are relevant. Stolen mass-produced things, even though they might be decorative, are not relevant (unless they are mass-produced art reproductions). Pirated software, music, movies, etc. are not relevant.

## CHASE ENDS IN ARREST OF 3 AFTER LATEST JEWEL HEIST;

### CRIME: SANTA ANA ROBBERY FITS A PATTERN OF NEARLY 100 SIMILAR THEFTS IN THE WEST SINCE 1989. THE SUSPECTS MAY BE PART OF A LOS ANGELES COUNTY RING.

By WENDY PAULSON, TIMES STAFF WRITER

#### SANTA ANA

A freeway chase from Huntington Beach to Compton ended with the arrests of three men who allegedly robbed a department store jewelry counter at gunpoint, the latest in a series of Southland jewel heists, police said Thursday.

And although Orange County police were tight-lipped about investigations of two similar robberies in the last month, Los Angeles police said the incidents fit a pattern of nearly 100 similar thefts in the western United States since 1989 that may stem from a criminal network in southwest Los Angeles County.

Please rate the above document according to its relevance to **art, stolen, forged** as follows. Note that the task is about how relevant to the topic the document is.

- ☐ **Relevant.** A relevant document for the topic.
- ☐ **Not relevant.** The document is not good because it doesn't contain any relevant information.

Does the topic look difficult? Please rate the difficulty from 1 to 5 (1=easy, 5=very difficult):

- ☐ **1** Easy
- ☐ **2** Somewhat easy
- ☐ **3** Neither easy nor difficult
- ☐ **4** Somewhat difficult
- ☐ **5** Very difficult

Please justify your answer or comment on your selection. Please use your own words. You may get a bonus payment if your comment is useful.

Submit

# Preliminary results

- With no prompt engineering
- With prompt features
  - R (role), D (description), A (aspects), M (multiple judges)
  - Performance varies per feature
  - Cohen's  $\kappa$  (0.20 to 0.64)

		Model	
		0	1 or 2
TREC assessor	0	866	95
	1 or 2	405	1585

Table 3: Overview of TREC-8 relevance judgment agreement between TREC assessors and each of the LLMs. Based on a sample of 1000 topic-document pairs.

LLM	Prediction	TREC-8 Assessors		Cohen's $\kappa$
		Relevant	Not relevant	
GPT-3.5	Relevant	237	48	0.38
	Not relevant	263	452	
YouChat	Relevant	33	26	0.07
	Not relevant	67	74	

Table 4: Overview of TREC-DL 2021 relevance judgment agreement between TREC assessors and each of the LLMs based on a sample of 400 question-passage pairs. TREC assessments were made on a graded scale from 3 (highly relevant) to 0 (not relevant).

LLM	Prediction	TREC-DL 2021 Assessors				Cohen's $\kappa$
		3	2	1	0	
GPT-3.5	Relevant	89	65	48	16	0.40
	Not relevant	11	35	52	84	
YouChat	Relevant	96	93	79	42	0.49
	Not relevant	4	7	21	58	

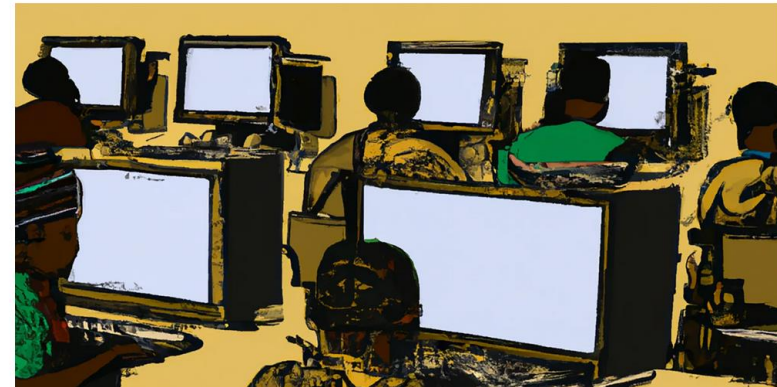
# Discussion

- In favor
  - LLMs are able to produce an explanation
  - This could be used to assist humans in relevance judgements
- Against
  - LLMs are not users
  - IR is about relevance to an information need
  - No proof that evaluation by LLM has any relationship to reality
- Things to consider
  - Reliability over time
  - Cost in prompt engineering
- Caveat
  - LLMs are systems
  - Query intent and answer construction

# LLMs and human computation

- For a LLM like ChatGPT,  $p(w_i | w_1, \dots, w_{i-1})$  is defined by a transformer
- LLM-based system do require editorial work
- Not different from any major property on the Internet

Exclusive: OpenAI Used Kenyan Workers on Less Than \$2 Per Hour to Make ChatGPT Less Toxic



<https://time.com/6247678/openai-chatgpt-kenya-workers/>

ARTIFICIAL INTELLIGENCE

## ChatGPT is powered by these contractors making \$15 an hour

Two OpenAI contractors spoke to NBC News about their work training the system behind ChatGPT.

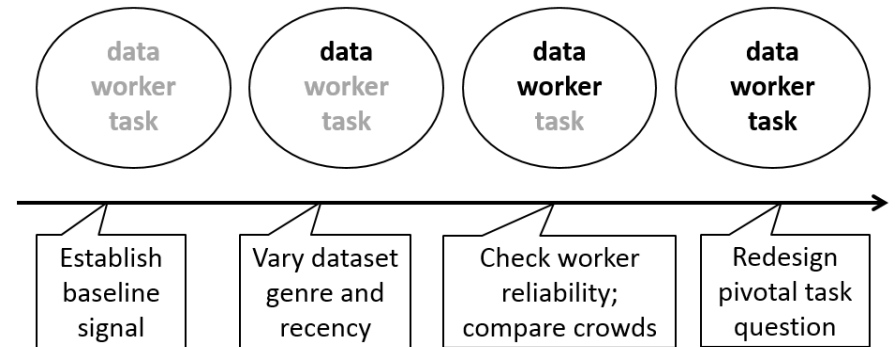
<https://www.nbcnews.com/tech/innovation/openai-chatgpt-ai-jobs-contractors-talk-shadow-workforce-powers-rcna81892>



# The main process is unchanged

- Regardless if labeling is done by machines or humans
- Three main components
  - Task design
    - HIT or prompt engineering
  - Data
  - Crowd
    - Human-based crowd or LLM-based crowd
- Quality control
- Debugging

	Machine computation	Human computation
Design	Throw away	Reluctant to throw away
Testing	Systematic	Ad-hoc
Debugging	Programmer's fault	Worker's fault



# Too many (irrelevant) tasks

- SciRepEval example
- There are so tasks that
  - some will be more relevant to our use cases,
  - and others will be less relevant
- The FoS task, for example, classifies documents into 23 fields of study.
  - The FoS task is probably not relevant to recommendation use cases

# Principles

- Characterize use case & audience
- Validity:
  - Relevance of task to use case
- Reliability:
  - Inter-rater agreement
- Realistic workloads
- Labeling and annotation
  - Documentation:
    - How was it done?
    - Availability of instructions
- Maintenance
  - Include a feedback loop mechanism to maximize adoption
  - Workload Evolution
  - Lessons learned and addendum(s)
  - Lifecycle and deprecation
- High standards
  - (for high-stakes use cases)
    - Use of established software engineering and data management techniques
      - (e.g., code review, versioning, configurations, dependencies, and testing).
    - How was the data sourced?
      - Provenance?
      - Can we data set be generated again easily?
    - Clean and well-documented data model

# Ugly: Outline

- ✓ Benchmarking (Omar)
- **Smooth-Talking Machines/Trust** (Ken)

## EMERGING TRENDS

# Emerging trends: Smooth-talking machines

Kenneth Ward Church  and Richard Yue 

Institute for Experiential AI, Northeastern University, San Jose, CA, USA

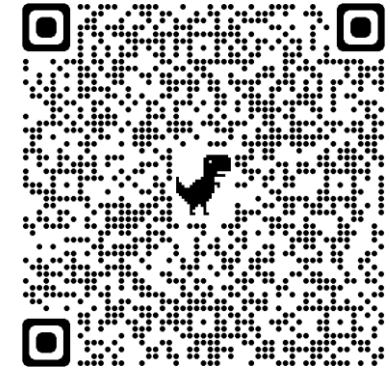
**Corresponding author:** Kenneth Ward Church; Email: [k.church@northeastern.edu](mailto:k.church@northeastern.edu)

(Received 15 August 2023; accepted 17 August 2023)

### Abstract

Large language models (LLMs) have achieved amazing successes. They have done well on standardized tests in medicine and the law. That said, the bar has been raised so high that it could take decades to make good on expectations. To buy time for this long-term research program, the field needs to identify some good short-term applications for smooth-talking machines that are more fluent than trustworthy.

**Keywords:** Large language models; Hallucinations; ChatGPT; Responsible AI



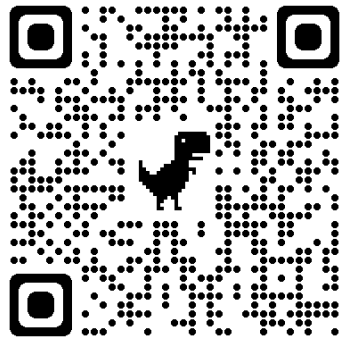
# Too Good to be True

- Standardized tests in medicine, law, etc.
- Press
  - *ChatGPT is, quite simply, the best artificial intelligence chatbot ever released to the general public.* -- New York Times
- Back-peddling
  - *You shouldn't expect a computer to hang a shingle... anytime soon, but...*
  - *It's best to think of ChatGPT as autocomplete on steroids...*
  - *anyone who uses the internet knows that*
    - *the internet is, well, not always accurate.*
  - *What's more, we don't know precisely what information ChatGPT is being fed.*

# Fluency $\neq$ Intelligence

- People want to believe in chatbots
- What is the difference between a hallucination and a con?

<https://didapelled.bandcamp.com/track/smooth-talkin-con-man>



*Smooth talking, soft spoken, con man*  
*Smooth talking, soft spoken, con man*  
*You stole all of my love*  
*Then you washed me off your hands*

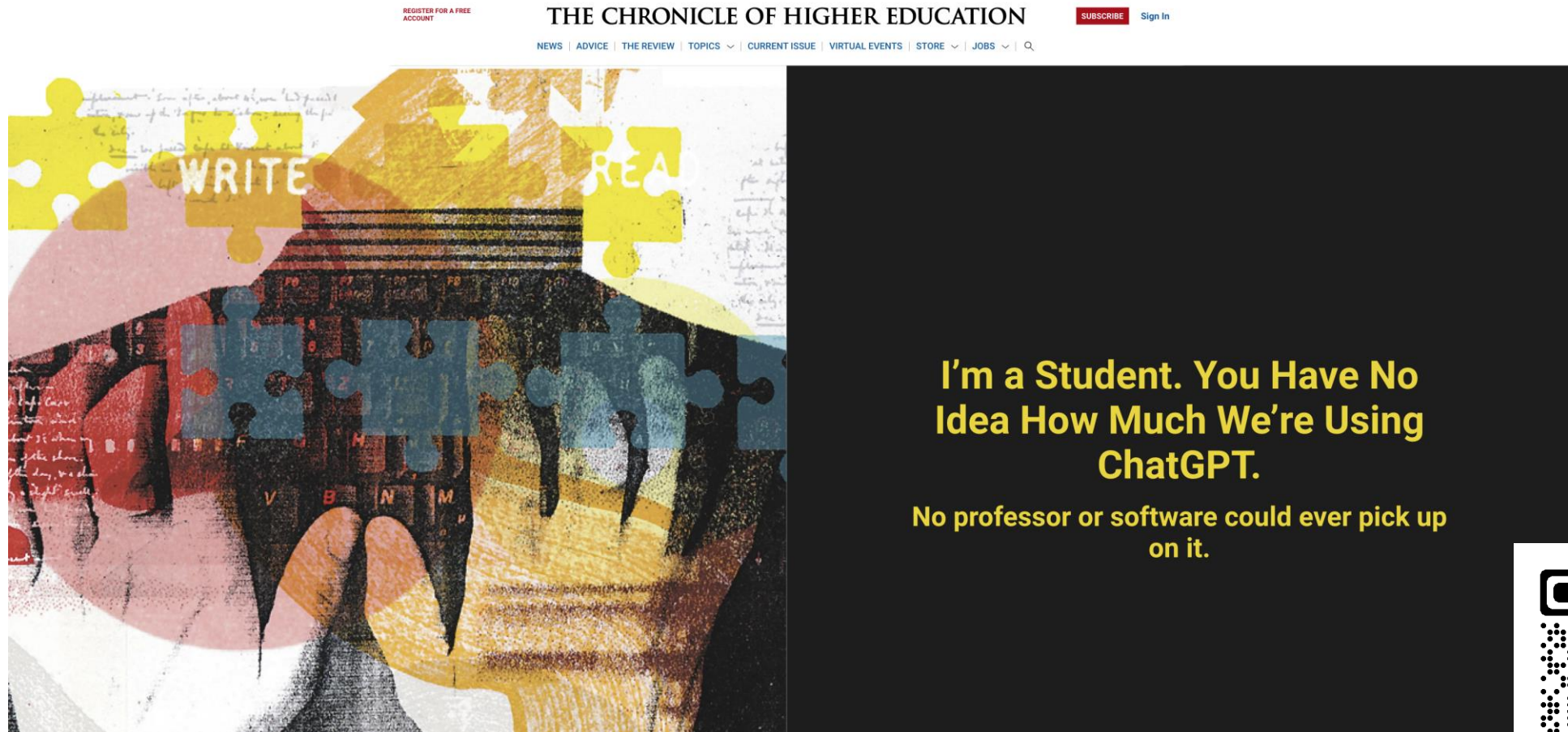
*Your words were well reversed lies*  
*You make me change my mind*  
...  
*Now I'm smooth talking swinging too*  
...  
*Smooth talkin', smooth walkin'*  
*soft spoken, slick workin'*  
*hip swinging, fast talkin' too*



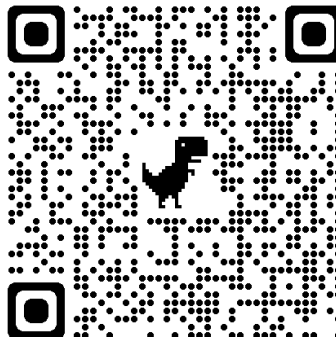
# Good Applications for Smooth-Talking Machines

- There have been booms and busts in AI
  - Busts → AI Winters
- *Good Applications for Crummy Machine Translation* – Church & Hovy (1993)
  - Even though machine translation did not work very well at the time,
    - we argued that it would help advance the field in the long-term
    - to look for promising short-term use cases.
  - We needed a few quick successes to support the field
    - to buy time for longer-term investments in more fundamental improvements.
  - It was clear at the time that
    - it would take decades to make good on expectations.
- Similar comments apply to LLMs.
  - The bar has been raised so high that it could take decades to make good on expectations.
  - In the meantime, we should be on the lookout for short-term quick hits
  - to buy time for more fundamental improvements.

<https://www.chronicle.com/article/im-a-student-you-have-no-idea-how-much-were-using-chatgpt>



[https://github.com/kwchurch/WSDM\\_2024\\_tutorial](https://github.com/kwchurch/WSDM_2024_tutorial)



# Essay Writing Subtasks:

## Human-machine Collaboration

### **Chatbots are good at**

- Thesis statements
- Outlines

### **Chatbots are not good at**

- Capturing student's voice
- Quotes (makes them up)

# Is it cheating to use a chatbot?

## Yes

- According to Owen Terry, author of
  - *You have no idea how much we're using ChatGPT*

## No

- A professor, Inara Scott, on NPR
  - <https://www.wbur.org/hereandnow/2023/05/22/chatgpt-academia>
  - Creative way to learn creative writing

# Good applications: Take advantage of strengths and avoid weaknesses

## Strengths

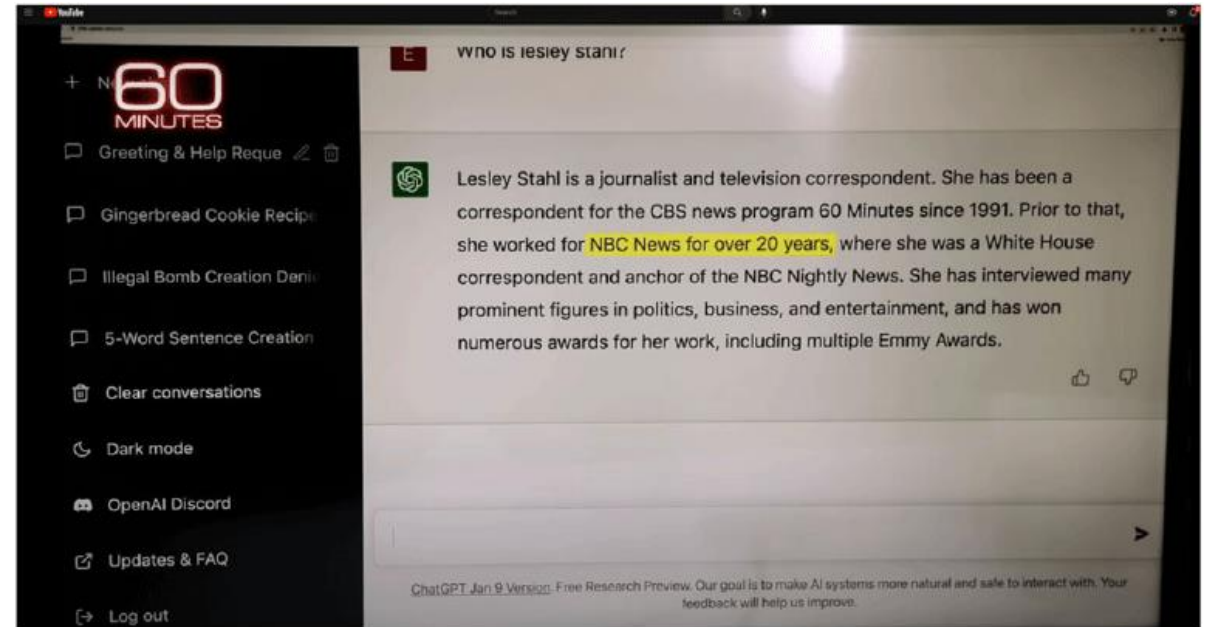
- Fluency

Thesaurus

## Weaknesses

- Hallucinations

Complement with  
fact-checking



# Smooth-Talking Conclusions

- Low Road: Give up; hallucinations are too hard
- **Middle Road:** Use search to verify assertions (fact-checking)
- High Road: Revive Rationalism
  - 1950s: Empiricism (Firth, Harris, Skinner)
  - 1970s: Rationalism (Minsky, Chomsky)
  - 1990s: Empiricism (EMNLP)

## EMERGING TRENDS

# Emerging trends: When can users trust GPT, and when should they intervene?

Kenneth Church 

Northeastern University, Boston, MA, USA

Email: [k.church@northeastern.edu](mailto:k.church@northeastern.edu)

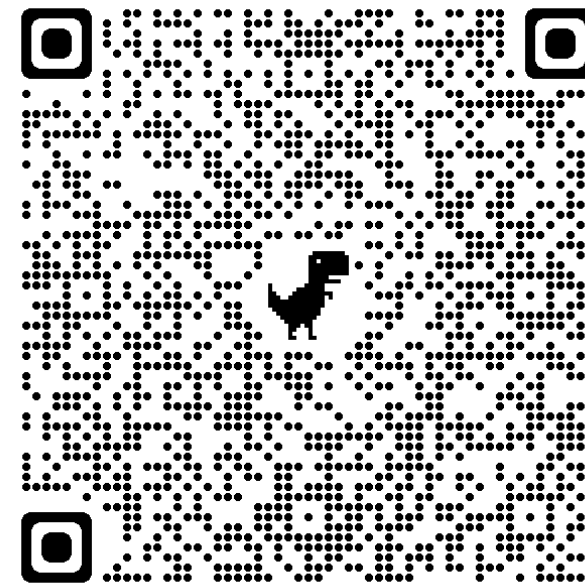
(Received 19 December 2023; accepted 19 December 2023)

### Abstract

Usage of large language models and chat bots will almost surely continue to grow, since they are so easy to use, and so (incredibly) credible. I would be more comfortable with this reality if we encouraged more evaluations with humans-in-the-loop to come up with a better characterization of when the machine can be trusted and when humans should intervene. This article will describe a homework assignment, where I asked my students to use tools such as chat bots and web search to write a number of essays. Even after considerable discussion in class on hallucinations, many of the essays were full of misinformation that should have been fact-checked. Apparently, it is easier to believe ChatGPT than to be skeptical. Fact-checking and web search are too much trouble.

**Keywords:** ChatGPT; fluency; trustworthiness; human-in-the-loop; evaluation in situ; fact-checking

[https://github.com/kwchurch/WSDM\\_2024\\_tutorial](https://github.com/kwchurch/WSDM_2024_tutorial)



# Evaluation of Human-in-the-Loop

## Homework: use tools to help them write essays

**Table 1.** LLMs have amazing strengths and amazing weaknesses

Task	Traditional NLP	Web search	LLMs
Metaphor	“AI Complete”	Very good	Amazingly good
Documentation	NA	Useful	Amazingly good
Outlines	NA	NA	Useful
Directions	NA	Useful	Poor
Quotes	NA	Useful	Amazingly bad
References	NA	Useful	Amazingly bad
Perspective	NA	Useful	Amazingly bad



# Use Cases:

## Some work well and some do not

### Amazingly Good

- *cover all the bases*
- *drop the ball*
- *dunk*
- *fumble*
- *get on base*
- *hit a home run*
- *out in left field*
- *ragging the puck*
- *sticky wicket*

### Amazingly Bad

- Made-up Reference
  - Title: “A Survey of Nearest Neighbor Search Algorithms”
  - Authors: Yufei Tao, Dongxiang Zhang
  - Link: Survey Paper (link to <https://arxiv.org/abs/1904.06188>)

- *During the First Opium War (1839–1842),*
  - *the British government was led by the Conservative Party under Prime Minister Sir Robert Peel.*
- *The opposition, primarily the Whigs, had varying views on the war.*
  - *Some opposed it on moral grounds, criticizing the ethics of trading in opium,*
  - *while others were concerned about the potential impact on international relations and trade.*
- Factual Errors
- Lack of Depth/Perspective
- Question for students
  - *What were the Opium Wars?*
  - *Where did the name come from?*
  - *Summarize the conflict from multiple perspectives, including:*
    - *England*
    - *China*
    - *India*
    - *United States*
    - *France*

# In fact, the Opium War had little to do with opium

- Neither the government (Whigs) nor the opposition (Conservatives)
  - supported drugs.
- The Whigs had just abolished slavery
  - and viewed drugs to be a form of slavery.
  - The conservatives viewed drugs as bad for business (in textiles and tea).
- The name of the conflict, Opium Wars, comes from an editorial on March 23, 1840, in the conservative newspaper: *The Times*, which argued that
  - *The British would be saddled with the massive expense of an unnecessary foreign campaign*
  - *that would cost far more than the entire value of the lost opium.* Platt (2019), p. 393.
- The government was put in an awkward corner because, Charles Elliot, their representative in China mishandled the situation.
  - He convinced the smugglers to give him their drugs in return for British IOUs,
  - and then he handed over the drugs to the Chinese authorities for destruction.
- When Parliament did not want to make good on the IOUs,
  - they thought they could force
  - the Chinese to pay for the lost opium.

# Lack of Perspective → Danger

## Chatbots ≠ Historian (Platt)

- Most of the essays from the students repeated output from ChatGPT more or less as is.
- These essays contained factual errors,
  - but more seriously,
  - the essays lack depth and perspective.
- In Platt (2019), p. 444, Platt argued that Napoleon understood that
  - it would be foolish for Britain to use its short-term advantage in technology to humiliate the Chinese.
  - Eventually, the Chinese would do what they have done (become stronger).
- Since the 1920s, these events are referred to as
  - the “century of humiliation”
  - by the authorities in China.
- Platt makes it clear that
  - the current Chinese government is using this terminology
  - to motivate its efforts to compete with the West in technologies such as AI
- When we discussed these essays in class, I tried to argue
  - that over-simplifying the truth, and
  - taking the Western side of the conflict,
  - could be dangerous and could lead to a trade war, if not a shooting war