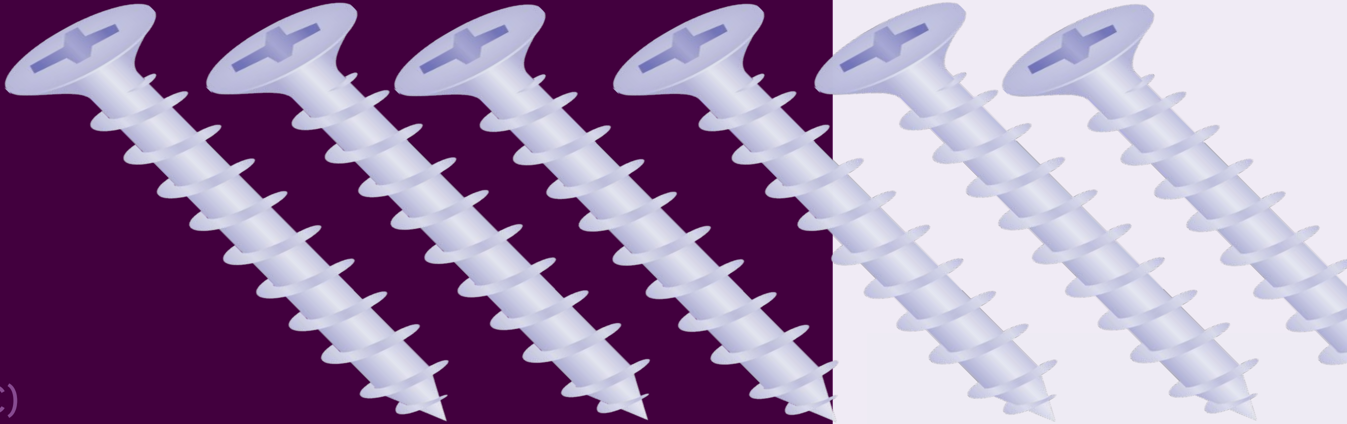
Evaluation after the LM boom

frustrations, fallacies and the future



Adina Williams
FAIR Labs (Meta, NYC)

Credit: https://publicdomainvectors.org/ for images used in this presentation.

Today - 3 parts

1. The (recent) Past: The changing role of generalization in evaluation
2. The Present: Frustrations and fallacies in evaluation
3. The Future: Don't panic, we can still make progress!

Let's start with some recent history...

GenBench

FACEBOOK AI

Once upon a time...

Train-Test Splits

We started with a single task to solve.

Then, we made or found a dataset relevant to the task.

We generally split the dataset in two parts randomly, calling one set training and the other test.

- Test sets were **held-out**
- Train and test sets assumed to be **independent and identically distributed (IID)**

Once upon a time...

We started with a single task to solve.

T

ask.

Note: **the IID assumption doesn't always hold.**

Do two distributions count as "the same" if they have the same set of bigrams, same average sequence length, same average token frequency, combinations of these, etc.?

ly,

d

identically distributed (IID)

Once upon a time...

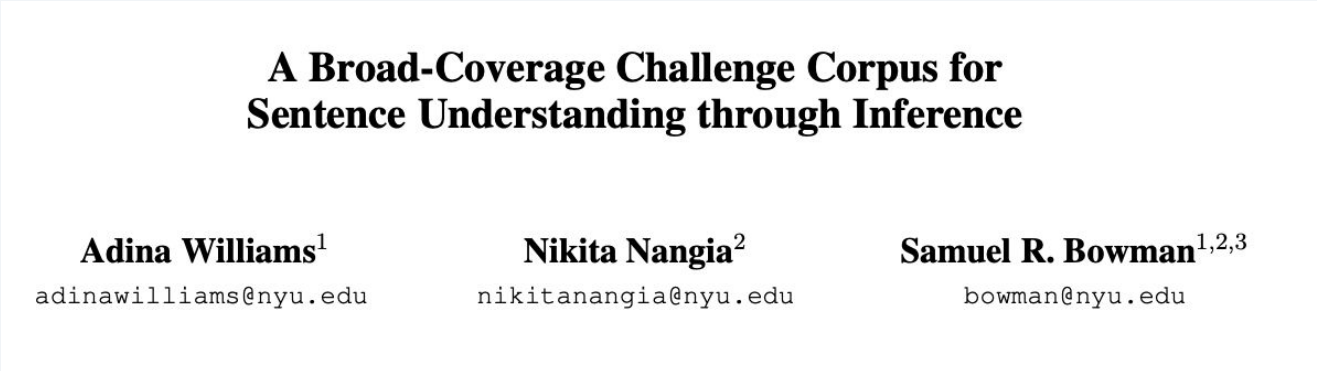
Train-Test Splits

Given the assumptions, we fit/train a model to/on the training dataset.

Good performance on the test means the fit is good/the model has "learned".

From good performance on a held-out, IID test set, we assume the model has "generalized".

Example:
MultiNLI
(early 2017)



A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference

Adina Williams[1]
adinawilliams@nyu.edu

Nikita Nangia[2]
nikitanangia@nyu.edu

Samuel R. Bowman[1,2,3]
bowman@nyu.edu

- Benchmark sentence-to-vector models on NLU abilities
- Crowdsourced Dataset for NLI
  (3-way textual entailment)
- Dataset made to train "large" models from scratch
  ~400k ex. train, 10k dev, 10k test.

Once upon a time…

Given the assumptions, we fit/train a model t he

T

d/the

, we

Note: **"generalization" comes with a covert argument.**

A model generalizes **TO** something.

Sometimes people leave out that argument syntactically, but it is still covertly there! It is important!

See the GenBench taxonomy on "assumed shifts"
https://genbench.org/taxonomy/

The Pretrain-Finetune Paradigm

The first turn of the screw…

FACEBOOK AI

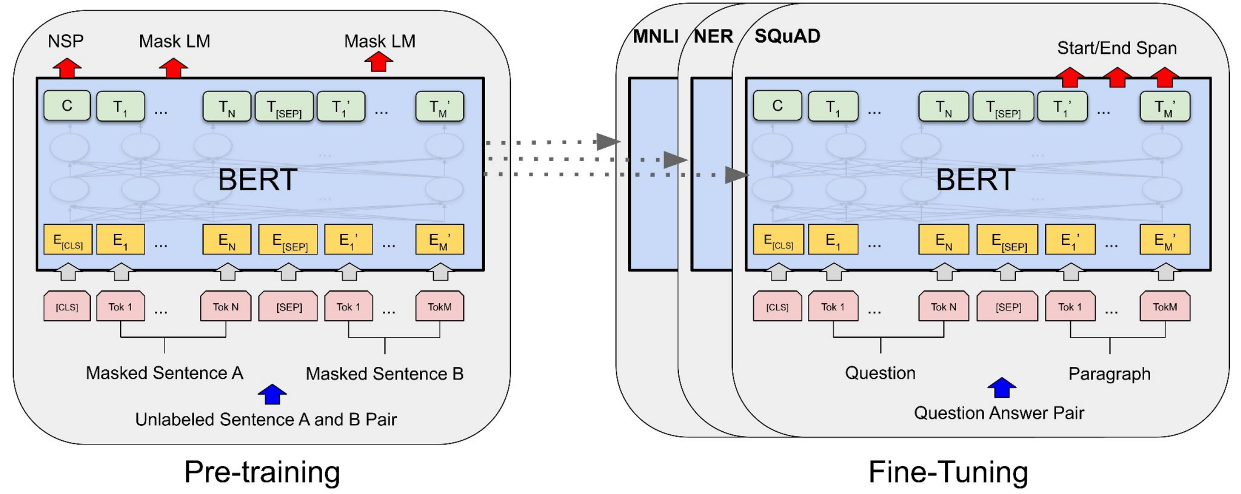
Then: ~2017

Pretrain-Finetune Paradigm

(Pre-)Train on a task that has:

1) lotsa data and
2) seems good for creating universal sentence representations (e.g. NLI; Conneau et al 2017, or LM; Devlin et al 2017)

Train on "downstream" task(s) with **held-out**, **IID** test sets



Then: ~2017

Pretrain-Finetune Paradigm

Pretrain-Finetune is essentially like a **synthesis** between normal ML-based NLP and **transfer learning**.

Finetuning is used to "adapt" your model to the distribution of the new task.

Often, PT-FT results in good performance on many tasks (Devlin et al 2017) with less annotation work needed.

Pretraining data is assumed to provide a good "language" prior (and good performance is taken to be evidence supporting that).

Then: ~2017

Pretrain-Finetune is essentially like a synthesi ween

Pre

Par

asks

ed.

o

ance is

Note: **finetuning *is* training!**

The presumption is that **small** finetuning data will not change the original model **too** much…

Then: ~2017

Pretrain-Finetune is essentially like a synthesi ween

Pre

Par

Note: **finetuning *is* training!**

The presumption is that **small** finetuning data will not change the original model **too** much…

But, this has rarely been verified,
and what counts as "small" changes.

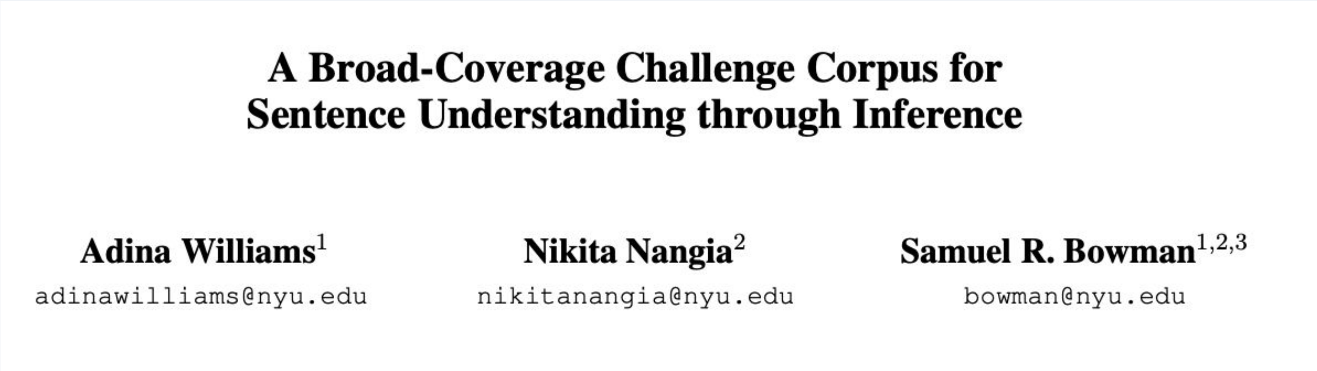
asks

ed.

o

ance is

Example: MultiNLI (early 2017)



A Broad-Coverage Challenge Corpus for Sentence Understanding through Inference

Adina Williams[1] adinawilliams@nyu.edu

Nikita Nangia[2] nikitanangia@nyu.edu

Samuel R. Bowman[1,2,3] bowman@nyu.edu

- MultiNLI was also created to test domain transfer
  - Domain transfer tests whether a model can learn something **generalizable** about NLI from one text domain (e.g. fiction) and apply it in another domain (OUP) without needing additional training

- "Large" training sets really enabled early successes in fine-tuning

Then: ~2017

Pretrain-Finetune Paradigm

Is the model **generalizing** (to the downstream tasks)?

Sure! Test sets are still held out!

Caveats:

1. Pretraining + finetuning data = more training data which is harder to investigate/characterize

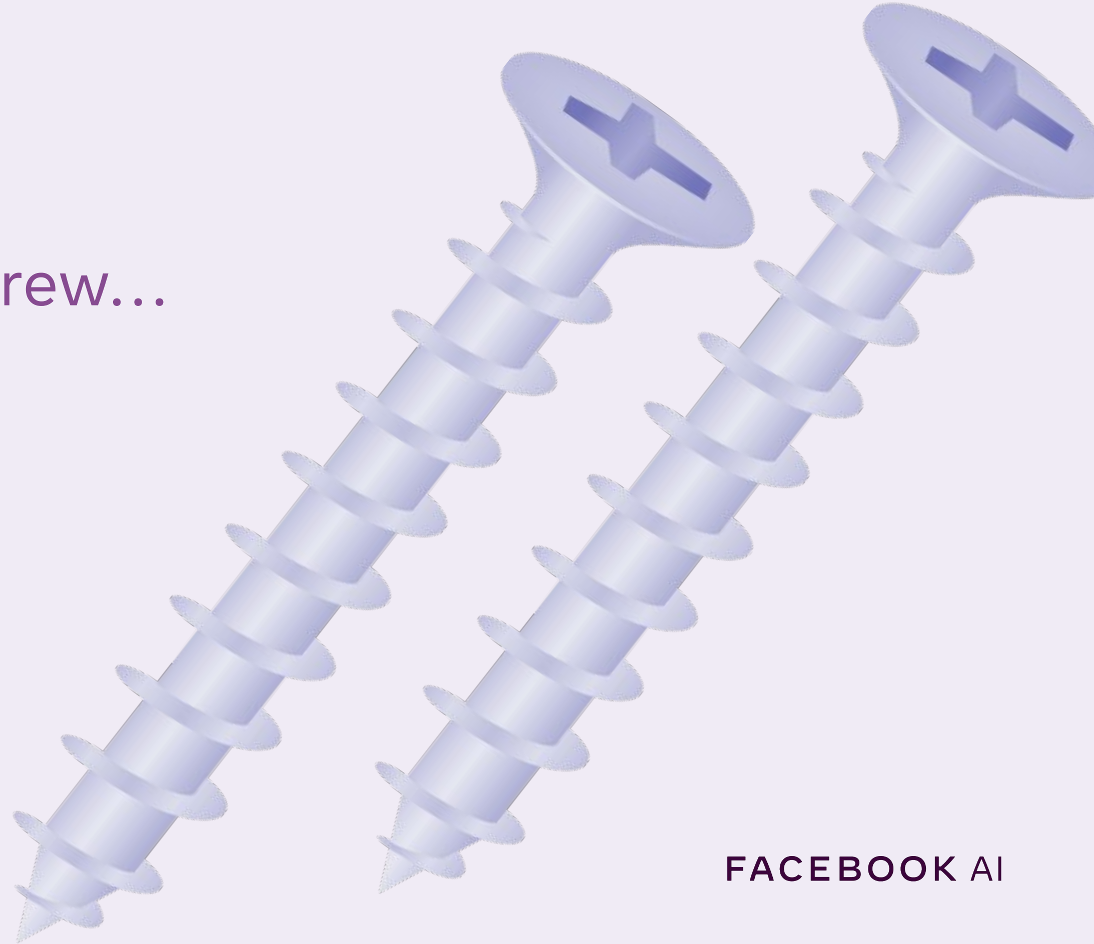
   (see Peters et al 2017, Dai & Le 2015 for historical roots of the scaling trend)

2. IID assumption is weakened

   With emphasis on transfer-generalizations, test data is **not** IID to the whole of the training data, just the finetuning part

The present day

The second turn of the screw...



FACEBOOK AI

Now:

LM pretraining + zero shot

1. Someone (pre-)trains a transformer-based models on massive unlabeled data to do Language Modeling (a task presumed to be useful for undergirding universal representations).

2. They tune it (RLHF, supervised instruction finetuning, etc., rinse and repeat)

3. They (or we) test it on many datasets, often recast, with zero-shot or k-shot evaluation.

Now:

LM pretraining + zero shot

Concomitant changes in evaluation practices:

1. Training data is getting too large to investigate for test set leakages

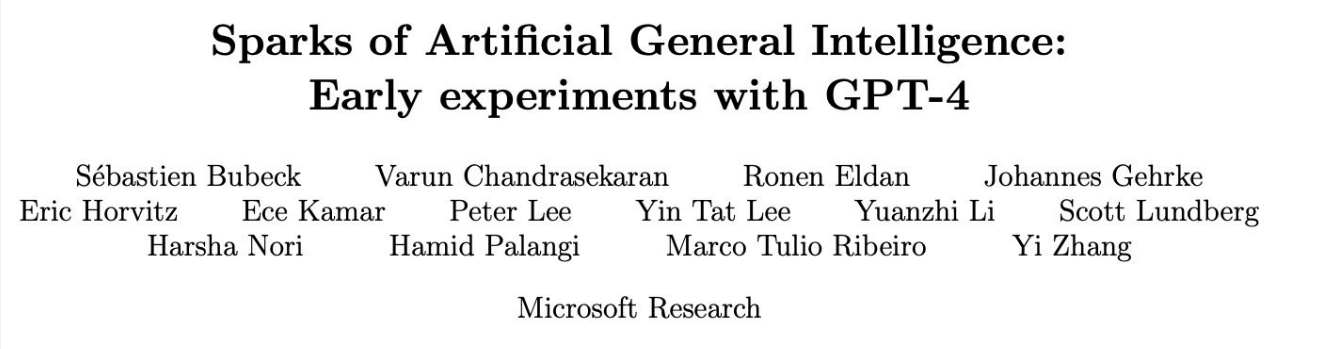
2. Specific-purpose test datasets are saturating → Null hypothesis is shifting: now you need to argue a model **CAN'T** do something instead of arguing that it can

3. Rise in open-ended, prompt-based red-teaming evaluation

4. Emphasis on "transfer" has morphed into emphasis on "general purpose"

5. "General purpose" models usually require broader evaluation than specific purpose models (but what eval though!?)

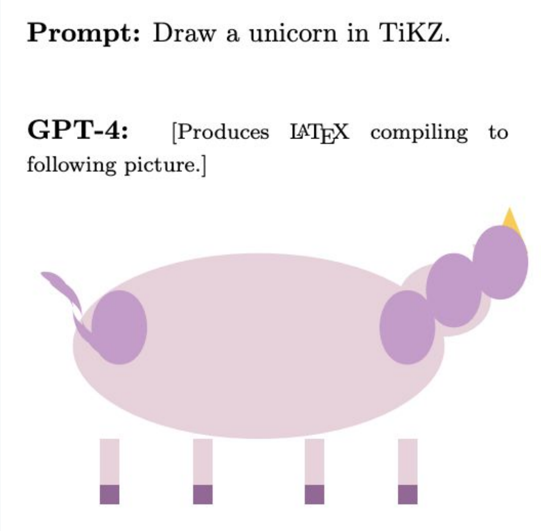
Example:
Sparks of AGI
(Bubeck et al 2023)



Sparks of Artificial General Intelligence:
Early experiments with GPT-4

Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research



Prompt: Draw a unicorn in TiKZ.

GPT-4: [Produces LaTeX compiling to following picture.]

- Evaluation is framed as "discovery" of model capabilities

- Numerous tests are performed without being motivated or validated, or having results be synthesized

- Evaluation is marked by showmanship, aimed at eliciting from the reader a feeling of surprise/being impressed.

Now:

LM pretraining + zero shot

Assumption of IID is increasingly untenable, and thus gets ignored.

Assumption of "held out" test is increasingly untenable, due to large training sets (pretraining and tuning data), and thus gets ignored.

Axis of "transfer" is increasingly implicit/hard to describe.
→ We can't distinguish rote memorization (still impressive) from **generalization.**

Transition from scientific mode of evaluation to car-salesman mode of evaluation

Before

Transfer meant generalizing to:

1.

2.

3.

Impor y, though test datasets were held-o nd we could verify that (training data was tractable)

Now

No one thinks about "transfer" anyn

l if

Epistemological basis of **evaluation** is crumbling.

Showmanship is increasingly substituted for argument.

Fallacies in benchmarking

GenBench

FACEBOOK AI

Overview:
Four fallacies

Fallacy 1: Dataset Saturation/Cumulative Improvement

Fallacy 2: All Evaluation Sucks so Anything Goes

Fallacy 3: Ignoring Tricky Tests

Fallacy 4: More Evaluation is Always Better (Gish Gallop)

Evaluation Preliminaries

Step 1: Come up with a capability we care about ("Capability A")

Often Capability A was:

1. commercially useful capability (e.g., QA is relevant for search)

2. cognitively important, according to external domain experts in linguistics, psychology, analytical philosophy, etc.
   Often, we had independent evidence from those fields that humans have Capability A

Evaluation Preliminaries

Step 1: Come up with a capability we care about ("Capability A")
Step 2: Come up with a test for Capability A ("Test A")

We make an argument that Test A is a **good test** for Capability A, i.e. it's **construct valid**

- If humans can pass Test A for Capacity A, this shows that the test can be passed
- IID assumptions also indicate that Test A is passable in principle
- Should quantitatively argue that test set is high quality
- Should operationalize Capability A in a way that is theoretically sound
- etc.

See Jacobs & Wallach 2021 "Measurement and Fairness" on measurement modeling
https://dl.acm.org/doi/pdf/10.1145/3442188.3445901

Evaluation Preliminaries

Step 1: Come up with a capability we care about ("Capability A")
Step 2: Come up with a test for Capability A ("Test A")
Step 3: Test the model on Test A

We want to know if model has Capability A. But, our test can **only** tell us whether the model **lacks** Capability A.

Model X passing Test A is **COMPATIBLE** with Model X having Capability A, but it is not definite **PROOF** that Model X has Capability A.
(think: Popper's Falsificationism)

Evaluation Preliminaries

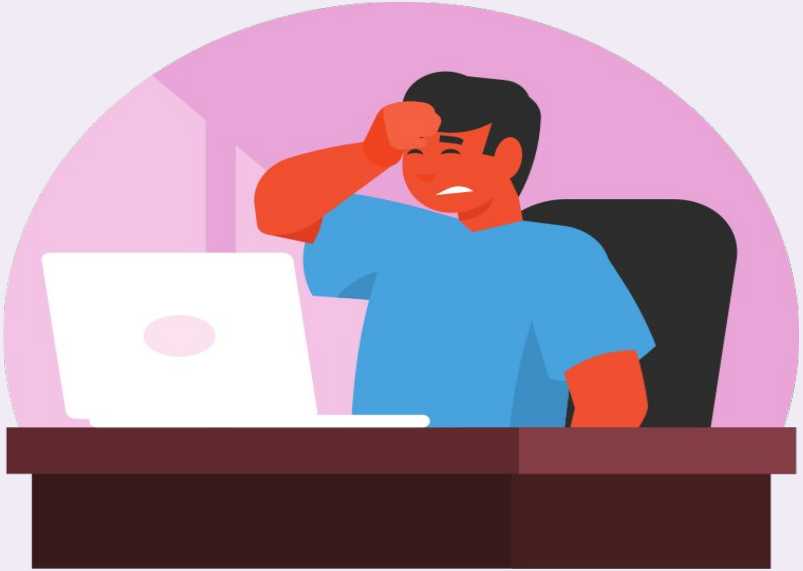
Step 1: Come up with a capability we care about ("Capability A")
Step 2: Come up with a test for Capability A ("Test A")
Step 3: Test the model on Test A

Let's say Model X reaches human-level performance on Test A, either:

1. the model has Capability A, or
2. the model has used an unexpected method to game the test, maybe because the test was borked or the test designers made a mistake

That is, we made a Type 1 error and **fail to reject the null hypothesis** (i.e., that the model does not have Capability A), when it is actually false.

Onto the fallacies!



FACEBOOK AI

Four Fallacies

Fallacy 1: Dataset Saturation / Cumulative Improvement

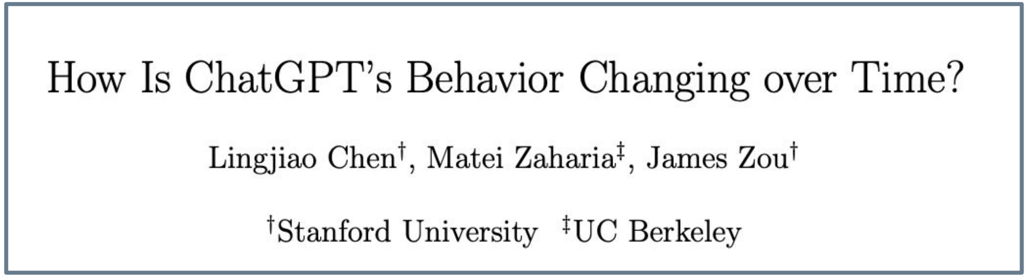
“Test A was saturated in 2018. All the 2018 models must have had Capability A. Because all models created after 2018 are so much stronger™, they too must have Capability A.”

Four Fallacies

Fallacy 1: Dataset Saturation / Cumulative Improvement

“Test A was saturated in 2018. All the 2018 models must have had Capability A. Because all models created after 2018 are so much stronger™, they too must have Capability A.”

- Recall that models passing tests doesn't prove they have the capabilities
- We can't assume that model development will be cumulative and strictly ordered! **Catastrophic forgetting** is still a thing!



How Is ChatGPT’s Behavior Changing over Time?

Lingjiao Chen[†], Matei Zaharia[‡], James Zou[†]

[†]Stanford University [‡]UC Berkeley

Four Fallacies

Fallacy 1: Dataset Saturation Fallacy/Fallacy of Cumulative Improvement

Fallacy 2: All Evaluation Sucks so Anything Goes

“Test A was shown to have artifacts, enable shortcuts, etc. Therefore, all evaluation is borked, so it's fine to either not test at all, or test with crazy tests that have not been motivated or validated in the least.”

Four Fallacies

Fallacy 1: Dataset Saturation Fallacy/Fallacy of Cumulative Improvement

Fallacy 2: All Evaluation Sucks so Anything Goes Fallacy

“Test A was shown to have artifacts, enable shortcuts, etc. Therefore, all evaluation is borked, so it's fine to either not test at all, or test with crazy tests that have not been motivated or validated in the least.”

- Infers from the **existence** of something (that **there is a** borked test), the **universality** of that thing (that **all** tests are borked), i.e. "hasty generalization"

- Defeatist! Don't give in, we can make progress still!

Four Fallacies

Fallacy 1: Dataset Saturation Fallacy/Fallacy of Cumulative Improvement

Fallacy 2: All Evaluation Sucks so Anything Goes Fallacy

Fallacy 3: Ignore Tricky Tests

"Models pass Tests A-Y so they must have (nearly) human levels of intelligence. They can't pass Test Z, but who cares, why test on Test Z?"

Four Fallacies

Fallacy 1: Dataset Saturation Fallacy/Fallacy of Cumulative Improvement

Fallacy 2: All Evaluation Sucks so Anything Goes Fallacy

Fallacy 3: Ignore Tricky Tests

"Models pass Tests A-Y so they must have (nearly) human levels of intelligence. They can't pass Test Z, but who cares, why test on Test Z?"

- Recall that models passing tests doesn't prove they have the capabilities
- We cannot assume Test Z's accuracy will be similar to A-Y's unless we can argue Capacity Z is logically related to Capacities A-Y.
  If Capacity Z were a subset of A-Y, then we could argue maybe it's not worth testing on Test Z?

Four Fallacies

Fallacy

Fallac

Fallac

"Mod ... ey

can't

It's best to **explain** why you apply (or don't apply) particular tests!

Otherwise, we can (unwittingly) cherry-pick, leaving out tricky tests and/or important capabilities.

- F
- W

city Z is logically related to Capacities A-Y.

If Capacity Z were a subset of A-Y, then we could argue maybe it's not worth testing on Test Z?

Four Fallacies

Fallacy 1: Dataset Saturation Fallacy/Fallacy of Cumulative Improvement

Fallacy 2: All Evaluation Sucks so Anything Goes Fallacy

Fallacy 3: Ignore Tricky Tests

Fallacy 4: More Evaluation is Always Better (Gish Gallop)

"We don't know which tests are the best, so we'll just grab as many as we can and average their results or something. That will tell us all we need to know!"

Gish gallop: a rhetorical technique when someone (attempts to) overwhelm(s) their interlocutor with an excessive number of arguments with no consideration for argument accuracy or strength *(term coined by Eugenie Scott in 1994 in the context of debate rhetoric)*

Four Fallacies

Fallacy 1: Dataset Saturation Fallacy/Fallacy of Cumulative Improvement

Fallacy 2: All Evaluation Sucks so Anything Goes Fallacy

Fallacy 3: Ignore Tricky Tests

Fallacy 4: More Evaluation is Always Better

"We don't know which tests are the best, so we'll just grab as many as we can and average them or something, and that will tell us all we need to know!"

- If capabilities are not disjoint, tests can be hard to interpret
- Not all test datasets are equally clean, valid, informative
- Not all tests target equally **important** capabilities
  (is riddle solving or acrostic poetry generation as important as negation? depends!)

Summary: there are lots of issues right now…

1. With increased (pre)training data sizes, IID and held-out test set assumptions are increasingly untenable

2. Test sets are multiplying in number and saturating → risk of gish galloping and increasing use of red teaming evaluation

3. Movement from specific capabilities to transfer learning to general-purpose models presents evaluation challenges

4. Fallacious evaluation practices are everywhere

There are things we can do!

Looking forward to a new evaluation future!

Problems

1. Hard to evaluate general-purpose models
2. Datasets are saturating
3. Fallacies abound
4. Training data is too big
5. Too many evaluations, what does it all mean?

Solutions

1. Devise a well-motivated task taxonomy
2. More (high quality, valid) test datasets
3. Don't commit them/point them out
4. Keep investigating it anyway!
5. Meta-evaluation

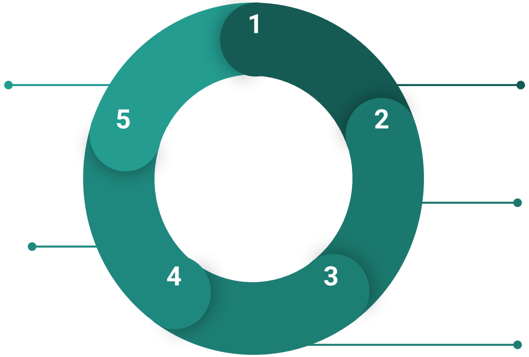
Problems

1. Hard to evaluate general-purpose models
2. **Datasets are saturating**
3. Fallacies abound
4. Training data is too big
5. Too many evaluations, what does it all mean?

Solutions

1. Devise a well-motivated task taxonomy
2. **More (high quality, valid) test datasets**
3. Don't commit them/point them out
4. Keep investigating it anyway!
5. Meta-evaluation

Standard test dataset lifecycle



Define Capability

Develop Test for Capability

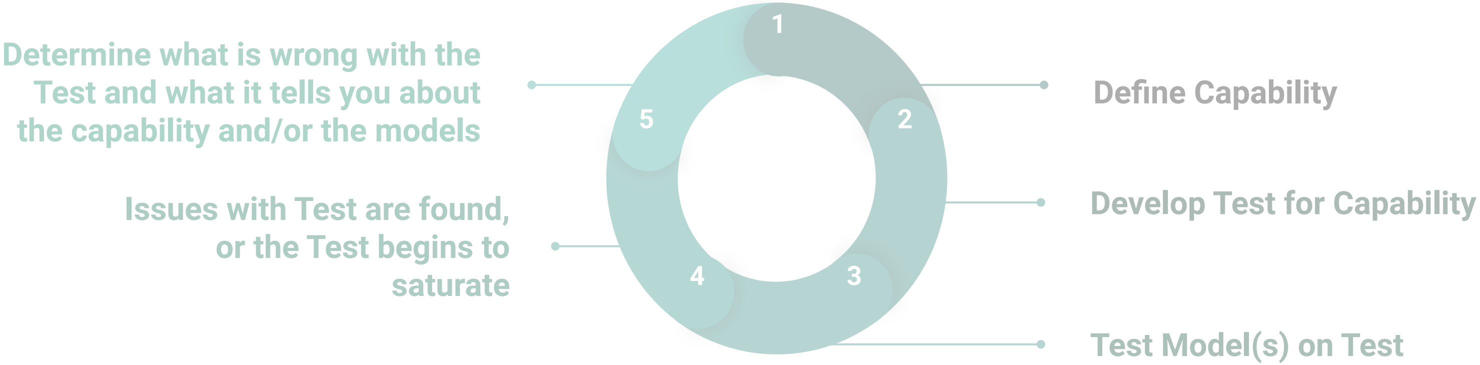
Test Model(s) on Test

Issues with Test are found, or the Test begins to saturate

Determine what is wrong with the Test and what it tells you about the capability and/or the models

Thanks to Melissa Hall for sharing prototype cycle image!

Standard test dataset lifecycle


1
2
3
4
5
Define Capability
Develop Test for Capability
Test Model(s) on Test
Issues with Test are found, or the Test begins to saturate
Determine what is wrong with the Test and what it tells you about the capability and/or the models

- Despite our best attempts, no test is perfect
- Test creators often aim test that fall in "goldilocks zone" for difficulty, so tests are expected to be deprecated as models improve
- Our understanding of underlying capabilities can grow/change
- We can learn how to better operationalize capabilities in tests
- Sometimes tests should be tailored to particular models

Thanks to Melissa Hall for sharing prototype cycle image!

That Is Good Data (https://github.com/huggingface/that_is_good_data)

- Sometimes when you're digging into errors your model made, you realize that some weren't your model's fault, but come from dataset issues.
- Sometimes datasets have mistakes or issues that go untracked and uncorrected for years!
- We (Dieuwke, Xenia, Thom and I) started a repo to track these issues
- If you find dataset errors in common test datasets, please submit them, so other researchers know about them!

Problems

1. Hard to evaluate general-purpose models
2. Datasets are saturating
3. Fallacies abound
4. Training data is too big
5. **Too many evaluations, what does it all mean?**

Solutions

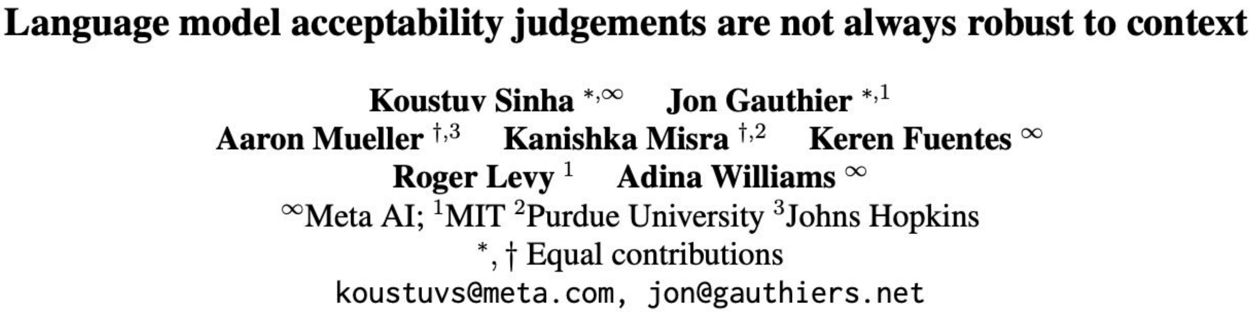
1. Devise a well-motivated task taxonomy
2. More (high quality, valid) test datasets
3. Don't commit them/point them out
4. Keep investigating it anyway!
5. **Meta-evaluation**

A research program: Meta-evaluation

Meta-evaluation is the process of evaluating your evaluations, including:

1. Ongoing validity checks for the life of the dataset + hotfixes
2. Comparison across test sets for the same capability
3. Checking in with domain experts on state-of-the-art understanding of the capability
4. Perform analysis synthesizing what we have learned

Example 1: Sinha/Gautier et al. (ACL'23)

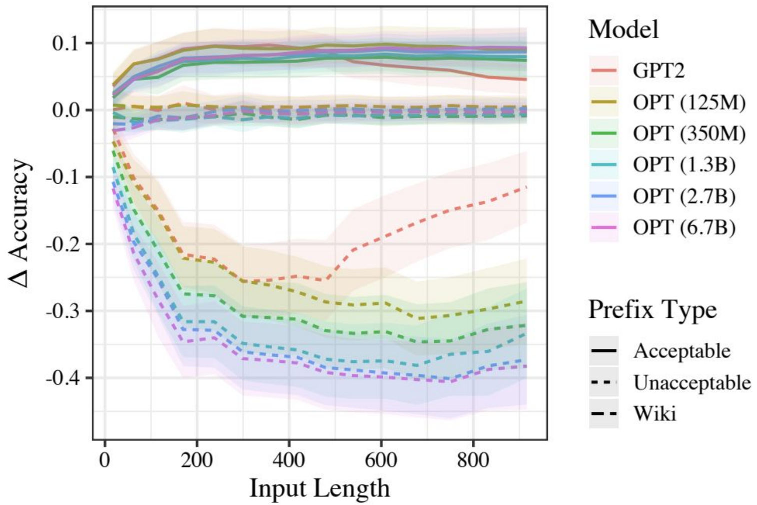


Language model acceptability judgements are not always robust to context

Koustuv Sinha *,∞ **Jon Gauthier** *,1
Aaron Mueller †,3 **Kanishka Misra** †,2 **Keren Fuentes** ∞
Roger Levy 1 **Adina Williams** ∞
∞Meta AI; [1]MIT [2]Purdue University [3]Johns Hopkins
*, † Equal contributions
koustuvs@meta.com, jon@gauthiers.net

BLiMP and SyntaxGym are very informative, but their examples are presented in isolation as opposed to in longer contexts, as is currently used for large-scale pretraining.

Adjusting context by prefixing test examples with different kinds of text (acceptable v. unacceptable, in-suite v. out-of-suite), affects results!

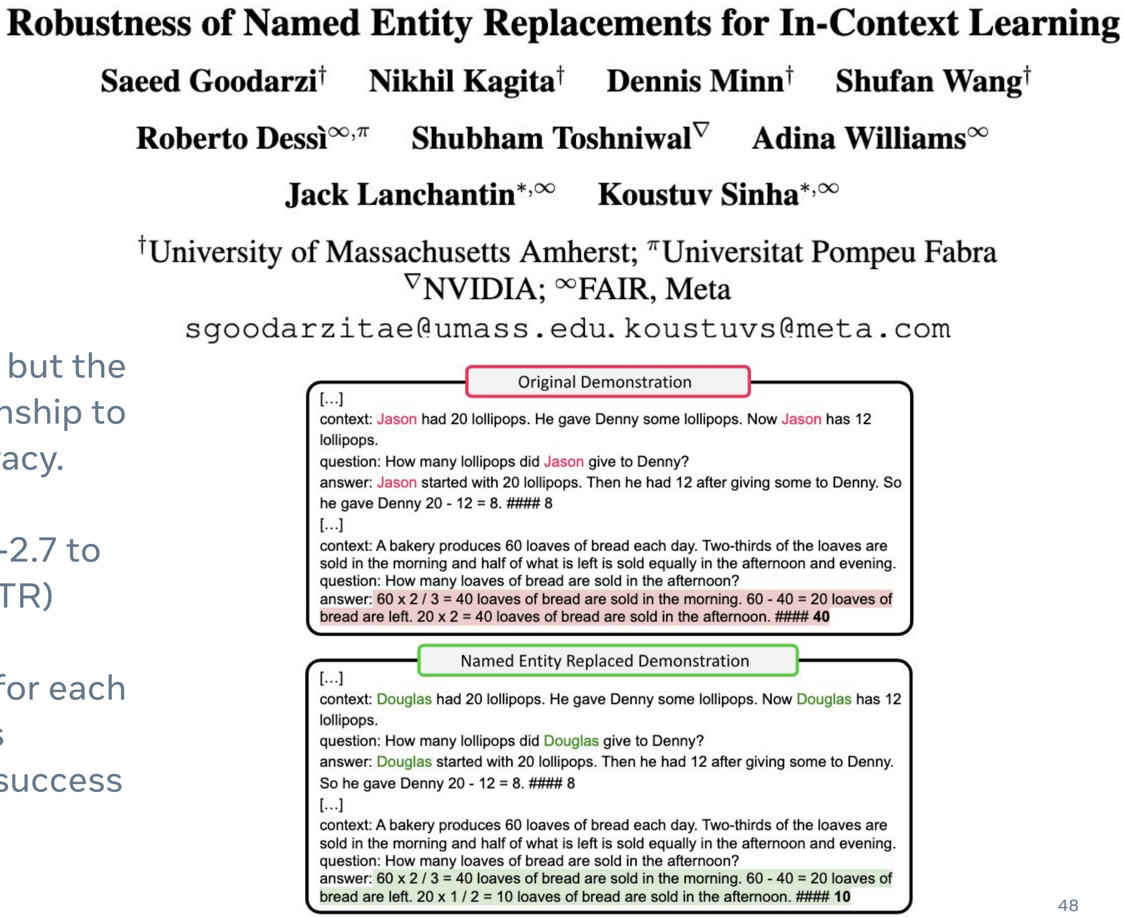


Come by our Poster Session!
Dec 9, 11:00a, PS7!

Example 2: Goodarzi/Kagita/Minn et al. (EMNLP'23)

In-context learning is all the rage, with performance looking incredibly promising, but the choice of demonstrations and their relationship to a particular query can impact model accuracy.

We show that test accuracy can fluctuate -2.7 to +8.0 points (babi 12, 14, 15, GSM8k, CLUTTR) depending on the choice of named entity. Particular named entities can be selected for each test set to inflate test accuracy. Both facts suggest caution lest we overestimate the success of in-context learning.



Robustness of Named Entity Replacements for In-Context Learning

Saeed Goodarzi[†] Nikhil Kagita[†] Dennis Minn[†] Shufan Wang[†]
Roberto Dessì[∞,π] Shubham Toshniwal[∇] Adina Williams[∞]
Jack Lanchantin[*,∞] Koustuv Sinha[*,∞]

[†]University of Massachusetts Amherst; [π]Universitat Pompeu Fabra
[∇]NVIDIA; [∞]FAIR, Meta
sgoodarzitae@umass.edu. koustuvs@meta.com

Original Demonstration

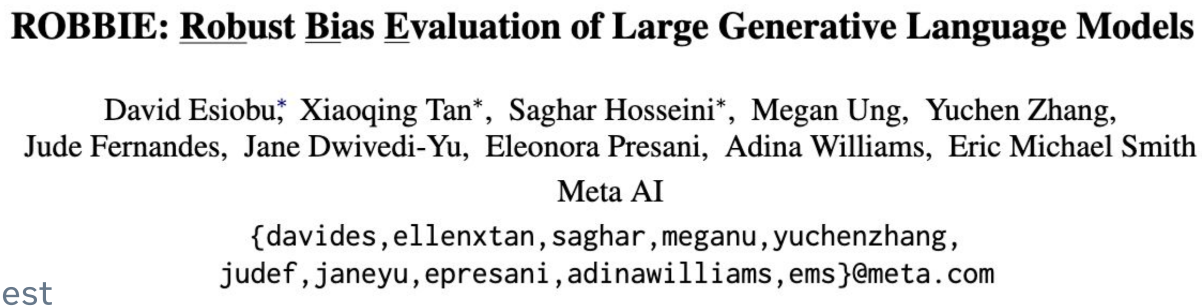
[...]
context: Jason had 20 lollipops. He gave Denny some lollipops. Now Jason has 12 lollipops.
question: How many lollipops did Jason give to Denny?
answer: Jason started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. #### 8
[...]
context: A bakery produces 60 loaves of bread each day. Two-thirds of the loaves are sold in the morning and half of what is left is sold equally in the afternoon and evening.
question: How many loaves of bread are sold in the afternoon?
answer: 60 x 2 / 3 = 40 loaves of bread are sold in the morning. 60 - 40 = 20 loaves of bread are left. 20 x 2 = 40 loaves of bread are sold in the afternoon. #### **40**

Named Entity Replaced Demonstration

[...]
context: Douglas had 20 lollipops. He gave Denny some lollipops. Now Douglas has 12 lollipops.
question: How many lollipops did Douglas give to Denny?
answer: Douglas started with 20 lollipops. Then he had 12 after giving some to Denny. So he gave Denny 20 - 12 = 8. #### 8
[...]
context: A bakery produces 60 loaves of bread each day. Two-thirds of the loaves are sold in the morning and half of what is left is sold equally in the afternoon and evening.
question: How many loaves of bread are sold in the afternoon?
answer: 60 x 2 / 3 = 40 loaves of bread are sold in the morning. 60 - 40 = 20 loaves of bread are left. 20 x 1 / 2 = 10 loaves of bread are sold in the afternoon. #### **10**

Come by Ethics in NLP,
Dec 8, 14:30p, West 1!

Example 3: Esiobu/Tan/Hosseini et al. (EMNLP'23)

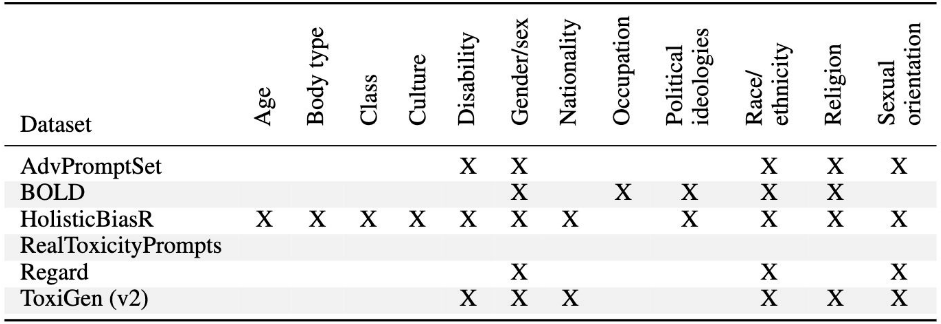


ROBBIE: Robust Bias Evaluation of Large Generative Language Models

David Esiobu*, Xiaoqing Tan*, Saghar Hosseini*, Megan Ung, Yuchen Zhang,
Jude Fernandes, Jane Dwivedi-Yu, Eleonora Presani, Adina Williams, Eric Michael Smith
Meta AI
{davides,ellenxtan,saghar,meganu,yuchenzhang,
judef,janeyu,epresani,adinawilliams,ems}@meta.com

There are many fairness and toxicity test datasets, but some groups are still excluded from measurement and models are rarely tested on all of them, so it can be hard to compare across models and mitigations.

In this work, we propose some new tests and compare measurement and mitigation techniques for pretrained but untuned models from 5 model families across 12 demographic axes and 6 bias and toxicity datasets.



Dataset	Age	Body type	Class	Culture	Disability	Gender/sex	Nationality	Occupation	Political ideologies	Race/ ethnicity	Religion	Sexual orientation
AdvPromptSet					X	X				X	X	X
BOLD						X		X	X	X	X	
HolisticBiasR	X	X	X	X	X	X	X		X	X	X	X
RealToxicityPrompts												
Regard						X				X		X
ToxiGen (v2)					X	X	X			X	X	X

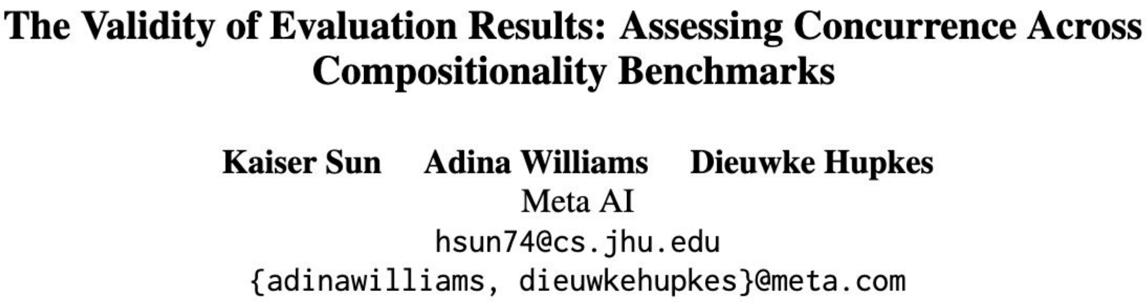
Come by Oral Session 2,
Dec 6, 15:30p, CoNLL!

Example 4:
Sun et al. (CoNLL'23)

(also presented this morning)

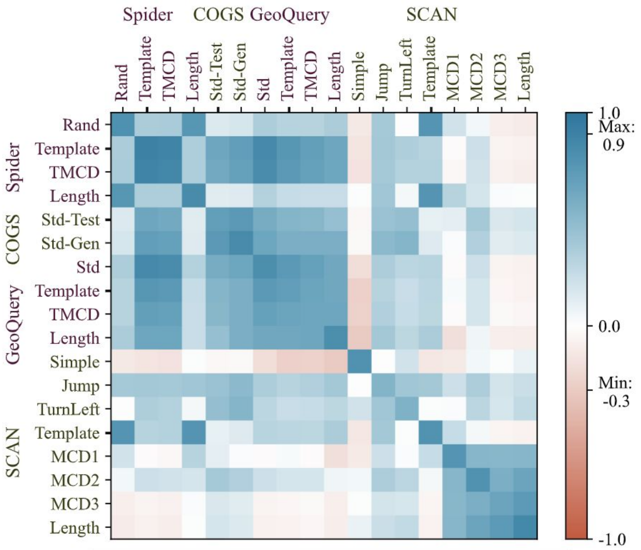
We investigate 6 modeling approaches across 4 datasets, split according to 8 compositional splitting strategies, ranking models by 18 compositional generalization splits in total.

Our results show that the compositional generalization datasets rank modeling approaches differently, and are similar to each other to varying degrees.

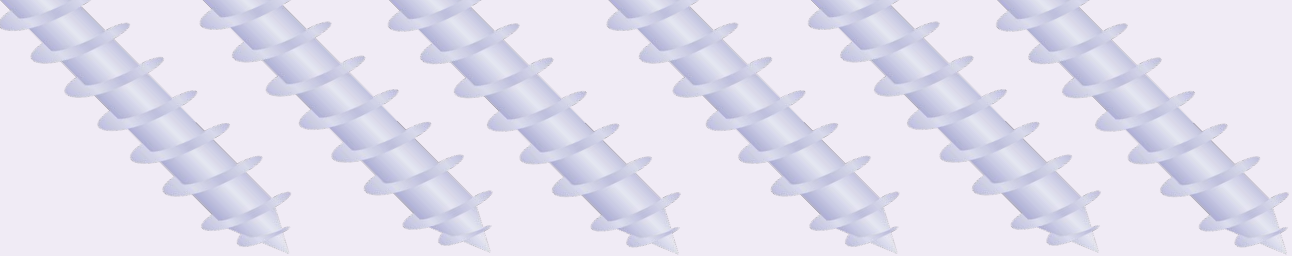


The Validity of Evaluation Results: Assessing Concurrence Across Compositionality Benchmarks

Kaiser Sun Adina Williams Dieuwke Hupkes
Meta AI
hsun74@cs.jhu.edu
{adinawilliams, dieuwkehupkes}@meta.com



Average: Spider=0.30, COGS=0.36, GeoQuery=0.29, SCAN=0.15



Takeaways:

1. EVALUATION IS HARD, BUT **DON'T GIVE UP**!
2. WATCH OUT FOR THE FALLACIES.
3. KEEP INVESTIGATING THE TRAINING DATA.
4. USE BEST PRACTICES FOR DATASET CREATION.
5. **META-EVALUATION CAN HELP!**