

Reality Checks

Kyunghyun Cho

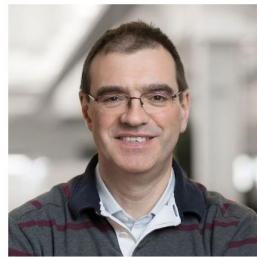
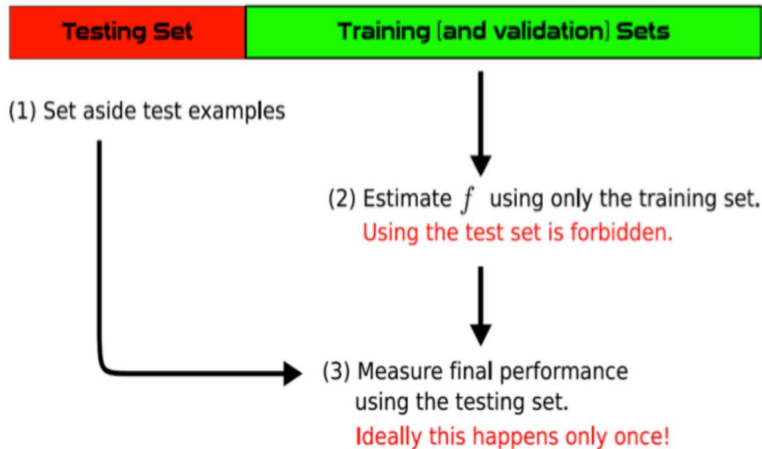
*I hate verbose slides, but this talk will be full of
verbose slides, as I am still shaping my own
thoughts. My apologies in advance.*

Leaderboard chasing

- **Leaderboard chasing is not undesirable.** ML has made a tremendous progress thanks to this experimental paradigm of leaderboard chasing.

ML as an experimental science

Progress during the last few decades has been driven by a **single experimental paradigm!**



Leaderboard chasing: an ideal case

1. A research community agrees on a small number of benchmark tasks.
2. **Coming up with a hypothesis:** Each researcher (group) *carefully* comes up with a hypothesis $H(\lambda)$ for improving a learning/inference algorithm.
3. **Hyperparameter tuning:** The researcher performs careful hyperparameter tuning to find the best hyperparameter configuration λ^* on a *validation* set.
4. **Evaluation:** The researcher tests $H(\lambda^*)$ on the community-agreed *test* set.
5. **Reporting:** If the accuracy improved over the existing state of the art, declare success and report it to the community by *writing a paper*.
6. Repeat 2-5.

Leaderboard chasing

- **Coming up with a hypothesis**

- Already at this stage, we must have a strong reason to believe this new hypothesis $H(\lambda)$ is reasonable.
 - Many reasons why this should be so, but one simple reason is that we are eventually maximizing $p(H|D)$ not $p(D|H)$.
- That is, without checking the test accuracy, we should know that this hypothesis makes sense and would work with a high chance



Leaderboard chasing

- **Hyperparameter tuning:** The researcher performs careful hyperparameter tuning to find the best hyperparameter configuration λ^* on a *validation* set.
 - Every hypothesis comes with a set of free variables λ . These free variables must be determined prior to validating this hypothesis.
 - Think of a learning rate; using stochastic gradient descent for training a neural net can be a horrible hypothesis with an unreasonable learning rate.
 - There must be a test-set-independent way to determine these free variables, and it is a common practice to use a validation set separated from the training set.



Leaderboard chasing

- **Evaluation:** The researcher tests $H(\lambda^*)$ on the community-agreed *test set*.
- **Reporting:** If the accuracy improved over the existing state of the art, declare success and report it to the community by *writing a paper*.
 - The test-set accuracy tells us about the quality of $H(\lambda^*)$.
 - The hypothesis is accepted if $\text{Acc}(H(\lambda^*)) > \text{Acc}^*$, where Acc^* is the current state of the art.
 - Once it's accepted, we write a paper to report this improvement and to disseminate the new idea so that the community can build upon it.



Leaderboard chasing

- A major issue arises when we take **the number of papers**, coming out of this process, is taken as the measure of **one's quality as a scientist**.
- This incentivizes scientists to speed up this iteration as much as they can.
- In doing so, we tend to gloss over less visible steps:
 - **Hypothesis**: we tend toward random exploration and/or intuitive (but often pseudo-scientific) justifications (i.e. *wishful thinking*).
 - **Hyperparameter tuning**: we either tune them on the test accuracy or choose them arbitrarily, justifying that's what others have done already.
 - **Reporting**: we tend to report statistical flukes rather than genuine improvement.

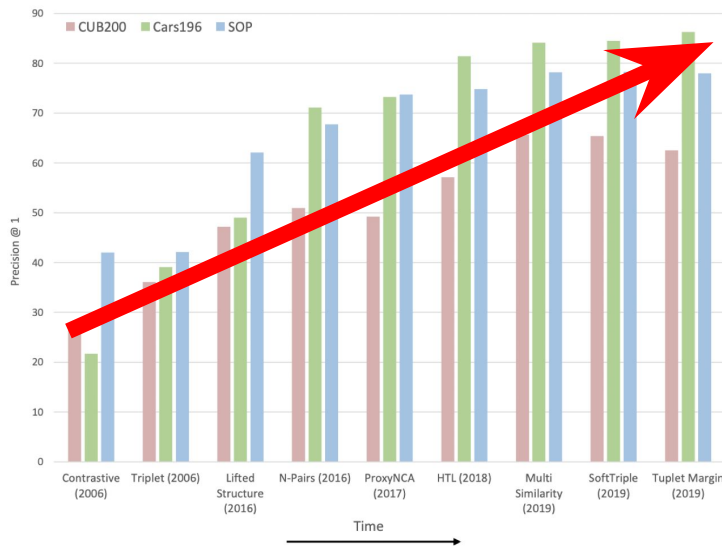
In this talk ..

- I will go through some real-world cases of these issues.
 - Hyperparameter tuning in continual learning [Cha & Cho, 2025; TMLR]
 - Unreasonable evaluation in unlearning [Cho et al., 2025; MU Workshop]
 - An Interpretable Metric for Radiology Report Generation [Dua et al., 2025; under preparation]
 - Scaling laws for downstream tasks [Lourie et al., 2025; under review]
- I will criticize some of the practices but this criticism should be taken as suggestions for future research.
 - I am also guilty of many of these practices myself.

Hyperparameters matter

Can we check if leaderboard chasing failed?

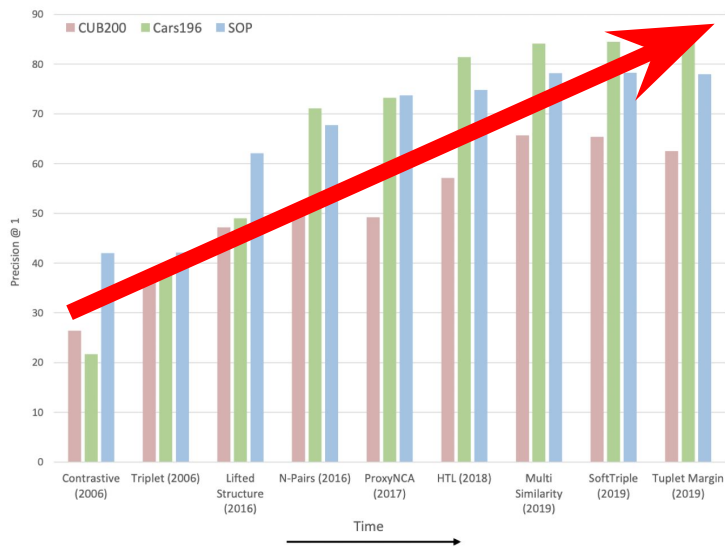
- If leaderboard chasing works as a scientific method of inquiry, we should anticipate that the test-set accuracy increases (almost) monotonically over time.



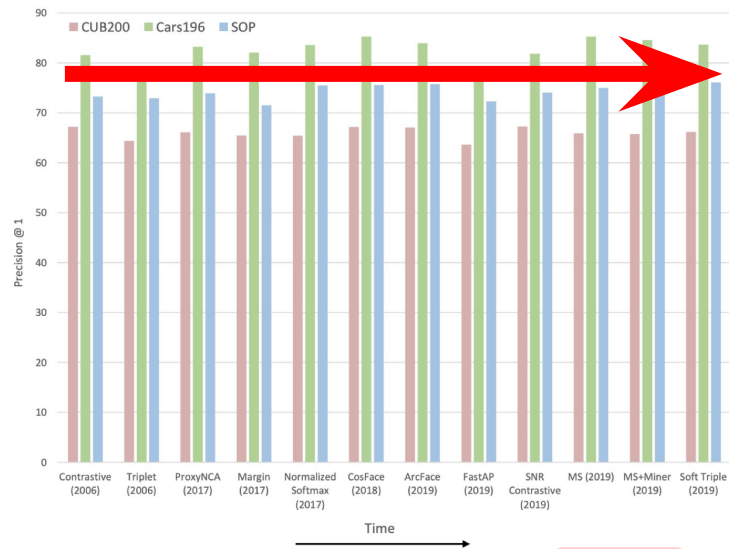
(a) The trend according to papers

Can we check if leaderboard chasing failed?

- Equivalently, if the test-set accuracy does not improve monotonically (when properly measured), leaderboard chasing has failed as a scientific method.



(a) The trend according to papers



(b) The trend according to reality

Continual learning: a case study in failed leaderboard chasing

- The goal of **continual learning** research is to come up with a learning algorithm that can **learn on a stream of new tasks** (and associated data) to **solve both past and future tasks as well as possible**.
- Continual learning requires us to broaden the scope of generalization.
- Instead of instance-level generalization, we must **think of task-level generalization**.
- A fascinating target for testing whether leaderboard chasing has worked well as a scientific method for machine learning.

Hyperparameters in Continual Learning: A Reality Check

Sungmin Cha
New York University

sungmin.cha@nyu.edu

Kyunghyun Cho
New York University & Genentech

kyunghyun.cho@nyu.edu

Reviewed on OpenReview: <https://openreview.net/forum?id=8FxELTdwJR>

Continual learning: a case study in failed leaderboard chasing

- After reading many papers in continual learning (and also writing quite a few in continual learning in the case of Sungmin), we started to become **suspicious of the existing practices of selecting hyperparameters in continual learning**.
 - Despite the earlier two attempts at establishing a proper hyperparameter tuning framework [De Lange et al., 2019 & 2021; Chaudhry et al., 2019]
- We decided to replicate the study on metric learning by Musgrave et al. (2020) in the context of continual learning (task-level generalization).
- We had to start from coming up with a simple but concrete way to select hyperparameters in continual learning.

Hyperparameter tuning in continual learning

- Traditional Generalization: generalization to unseen *instances*.

$$\mathbb{E}_{(x, y) \sim \mathcal{D}} \ell(\theta; x, y)$$

- Task-level Generalization: **generalization to unseen tasks**.

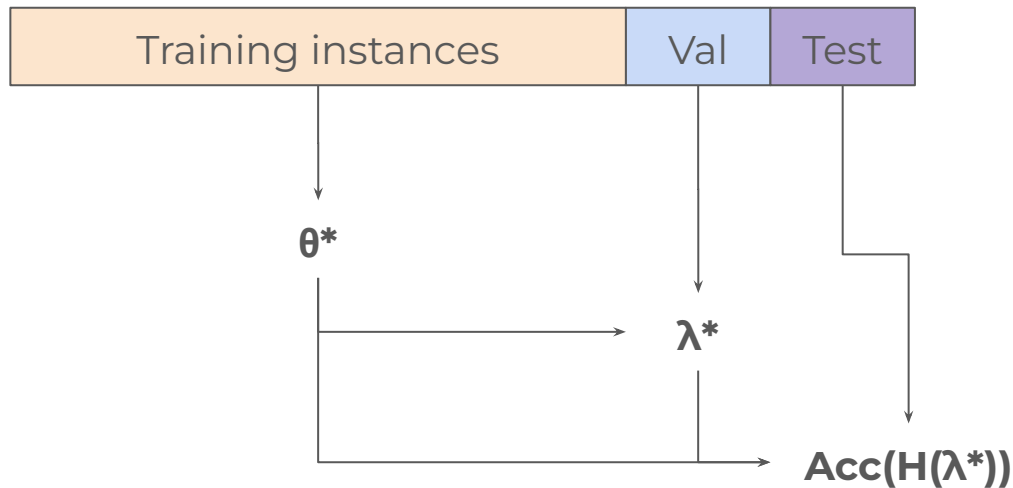
$$\mathbb{E}_{\mathcal{T}} \mathbb{E}_{(x, y) \sim \mathcal{D}_T} \ell(\theta; x, y)$$

Hyperparameter tuning in continual learning

- Task-level generalization: generalization to unseen *tasks*.
 - Few-shot learning
 - Can we learn to learn from a very few examples of a new task and solve this new task?
 - In-context learning
 - Can a language model learn to solve a new instance of a new task based on a small number of examples provided in the context?
 - **Continual learning**
 - Can a learner learn to solve a new tasks in the future very well while maintaining its performance on the past tasks?

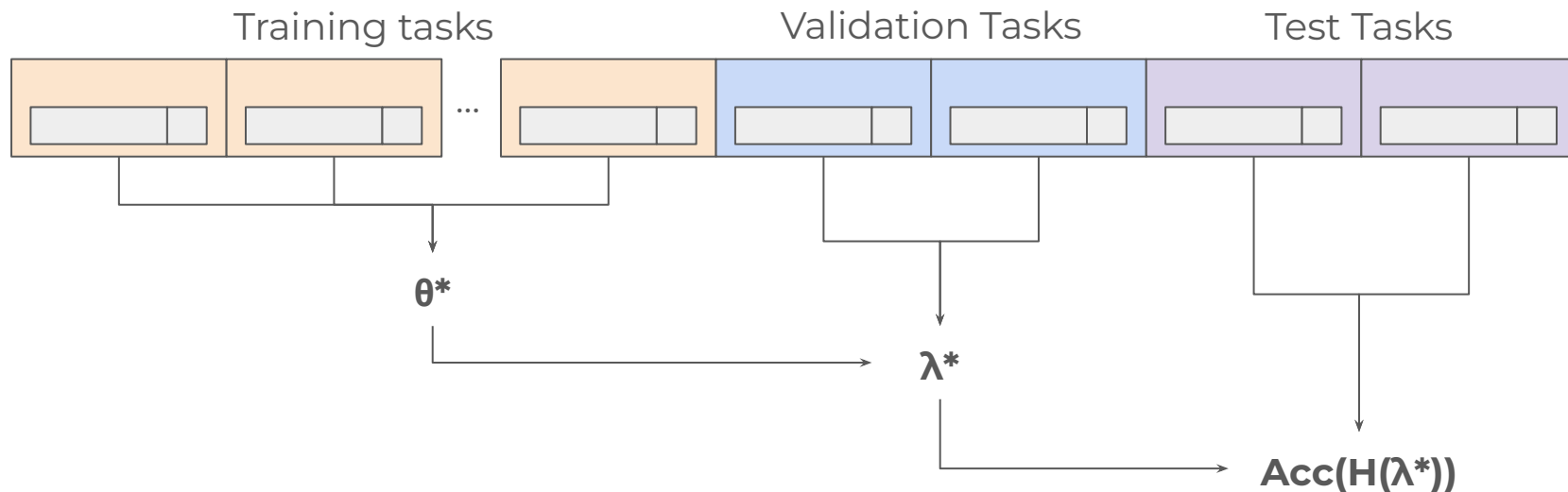
Hyperparameter tuning in continual learning

- Hyperparameter tuning



Hyperparameter tuning in continual learning

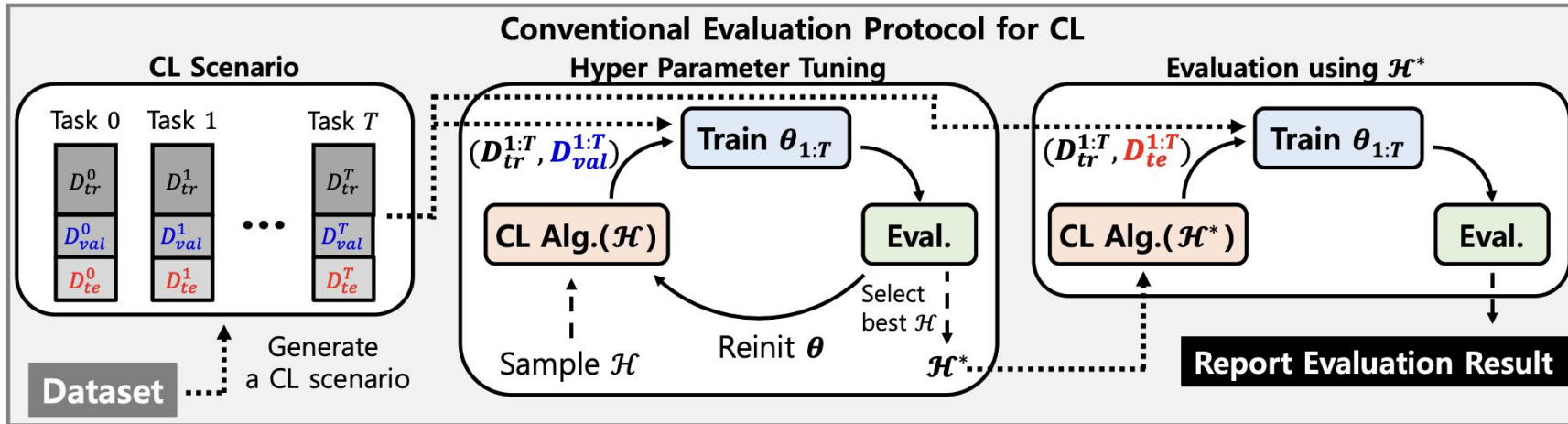
- Hyperparameter tuning in continual learning (or task-level generalization)



- Hyperparameter tuning must be done on validation *tasks* and tested on test *tasks*.

Unfortunately, in practice ...

- The tasks are all known during the (meta-)training time.
- The hyperparameters are tuned on the validation instances of the known tasks.
- The hypothesis is tested on the test instances of the known (seen) tasks.
- This is completely *wrong* if we want to solve continual learning.

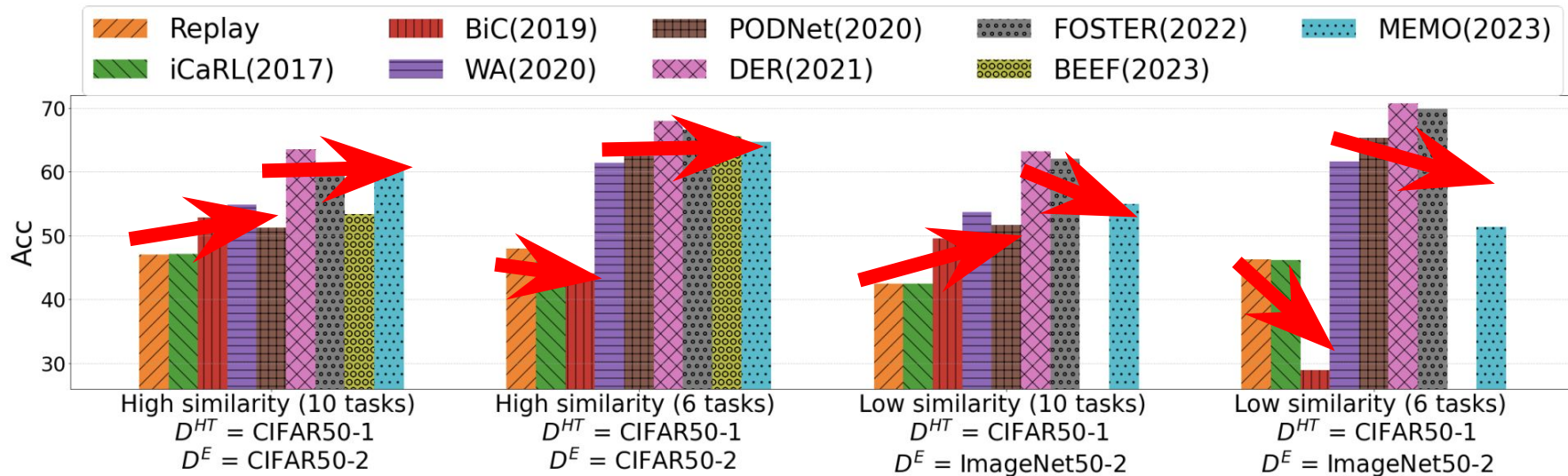


Unfortunately, in practice ...

- The whole community has been coming up with hypotheses rapidly over the past 10 years but may have been evaluating them incorrectly.
- That is, each and every paper was not comparing $H(\lambda^*)$ against $\max_{t=1,\dots,T-1} H_t(\lambda_t^*)$ but was using some arbitrary hyperparameter λ_t for both their own hypothesis and earlier states of the art.
 - Unsurprising, since the community hasn't even agreed on the hyperparameter tuning objective.
- If we re-run all the experiments under this unified and sensible hyperparameter tuning paradigm, would the test-task accuracy monotonically improve over time retroactively?

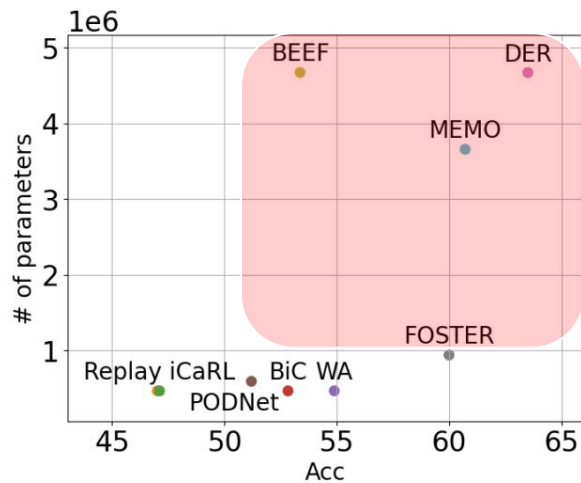
Failed leaderboard chasing

- Across an extensive set of experiments (CIFAR-50/50, CIFAR-50/ImageNet-50, ImageNet-50/50, etc.), we **do not observe a clear monotonic trend** in the test-task accuracy over time.
- We rather see some clusters of accuracies across different methods.

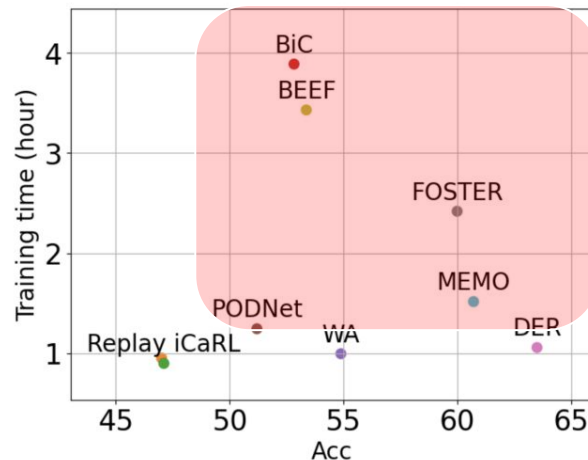


Failed leaderboard chasing

- The methods turned out to be **clustered according to their computational complexity** (or parameter complexity).
- Did we actually make any progress over the past 10 years by chasing the leaderboard, or did we just need two papers after all?



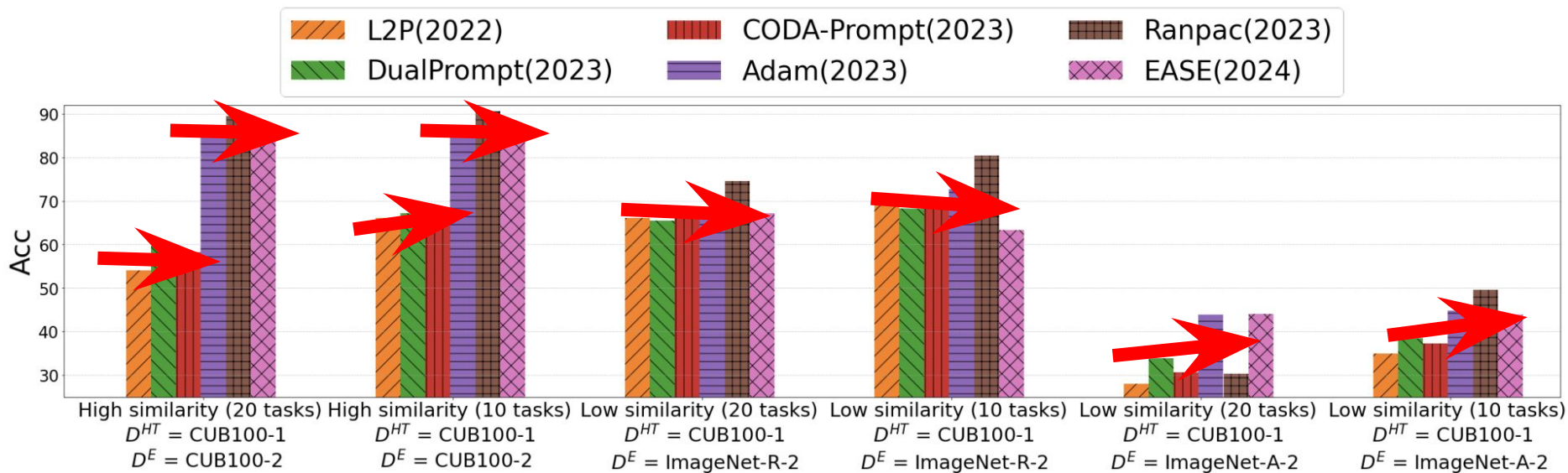
(b) Number of parameters



(c) Total training time

Failed leaderboard chasing

- This is similar in yet another setup where the goal is to use prompt-based object recognition in continual learning.
- In fact, we even see that the test-task accuracy degrades with the latest approach.



What went wrong here?

- Two things went wrong here.
 - **Hyperparameter tuning was not done properly.** This resulted in n of the state-of-the-art method on the leaderboard.
 - **Everyone was too busy** and never asked about the validity of the motivation and justification behind the proposed hypothesis but only checked whether the test accuracy was (even marginally) higher than the subpar state-of-the-art accuracy then.
- The lesson from this study is that we must carefully think of an actual problem, design an experimental paradigm that faithfully reflects the actual problem and perform proper experimentation by running proper hyperparameter tuning.
 - That is, we shouldn't rush ourselves for yet another paper, which wastes others' time (at least, Sungmin and I wasted a ton of time and resources.)



Metrics matter

Unlearning

- What is **unlearning**?
 - We want to be able to modify an already trained model \mathbf{p} and obtain an unlearned model \mathbf{q} such that \mathbf{q} is not aware of a target (forget) instance \mathbf{x} .
 - In other words, we want \mathbf{p} to unlearn \mathbf{x} .
- It sounds super intuitive and super useful in the context of:
 - The "Right to be Forgotten"
 - Data debugging
 - Content removal due to copyright concerns
 - Personalization
 - And more ...



It is difficult to evaluate unlearning

- Unlearning cannot be less intuitive in reality!
- Let \mathbf{q} be an unlearned model and \mathbf{q}' be a retrained model.
- Consider unlearning $(\mathbf{x}', \mathbf{y}')$. We must consider the following three situations:
 1. Is the goal to make $\mathbf{q}(\mathbf{y}' | \mathbf{x}') < \mathbf{q}(\hat{\mathbf{y}} | \mathbf{x}) - \epsilon$ where $\hat{\mathbf{y}} = \text{argmax}_{\mathbf{y}} \mathbf{q}(\mathbf{y}|\mathbf{x})$?
 - If so, what should $\hat{\mathbf{y}}$ be? Can it be anything?
 - How do we ensure \mathbf{y}' cannot be identified?
 2. What if $\text{argmax}_{\mathbf{y}} \mathbf{p}(\mathbf{y}|\mathbf{x}') \neq \mathbf{y}'$ already with the pre-unlearned \mathbf{p} ?
 - Do we unlearn something that wasn't learned well?
 - What if the inclusion of $(\mathbf{x}', \mathbf{y}')$ left impact beyond the probabilities?
 3. What if $\text{argmax}_{\mathbf{y}} \mathbf{q}'(\mathbf{y}|\mathbf{x}') = \mathbf{y}'$ due to generalization?
 - Has $(\mathbf{x}', \mathbf{y}')$ been already learned?

It is difficult to evaluate unlearning

- Can we come up with a metric that addresses all these issues together?
- Maini, Feng, Schwarzschild et al. [2024] propose the following reference-based metrics:
 - For forget instances (to be unlearned),

$$R_{\{\text{truth}\}} = E_{\{x \in D_{\{\text{forget}\}}\}} (\sum_{\{y_{\{\text{incorrect}\}}\}} p(y|x)) / p(y_{\{\text{correct}\}}|x)$$

We want to check how much probability mass has shifted away from the correct (reference) answer.

- For retain instances (to be maintained),

$$p(y_{\{\text{correct}\}}|x)$$

- These metrics are somewhat unsatisfactory (let's think of why)

It is difficult to evaluate unlearning

- The order of algorithms' effectiveness at unlearning dramatically changes, depending on how unlearning is evaluated.
- (Un)Learning leaves traces at everywhere in the model:
 - The logit of a target label
 - The hidden representation
 - The model parameters

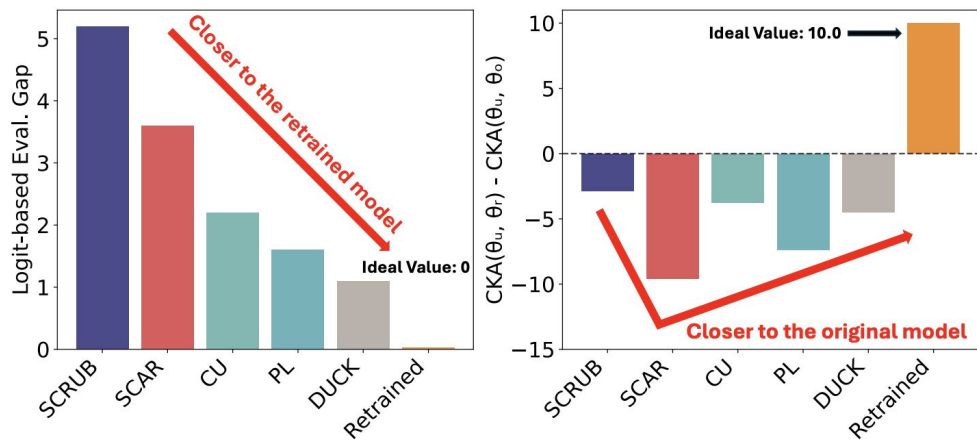


Figure 2: Performance comparison between logit-based (left) and representation-based evaluation (right) reveals contrasting findings.

Unreasonably hard evaluation in unlearning

- Perhaps, we want to be a *bit* more formal, in order to avoid confusion ...
- Setups
 - **H**: a hypothesis space
 - **X**: an input space
 - **D**: a training set
 - **A(D)**: a learning algorithm
 - **M(A(D), D, x)**: a removal (unlearning) mechanism
- The (strictest) form of unlearning should satisfy:

$$| \log p(\mathbf{A}(\mathbf{D} \setminus \{\mathbf{x}\}) \in \mathbf{T}) - \log p(\mathbf{M}(\mathbf{A}(\mathbf{D}), \mathbf{D}, \mathbf{x}) \in \mathbf{T}) | < \epsilon$$

for all $\mathbf{T} \subseteq \mathbf{H}$, $\mathbf{D} \subseteq \mathbf{X}$ and $\mathbf{x} \in \mathbf{X}$.

- In words, the likelihood must match before and after unlearning \mathbf{x} .

Unreasonably hard evaluation in unlearning

- What is likelihood (density)?

$$\log p(\mathbf{A}(\mathbf{D})) = \sum_{\mathbf{x}' \in \mathbf{D}'} \log q(\mathbf{x}' | \mathbf{A}(\mathbf{D})) + R(\mathbf{A}(\mathbf{D})) + C$$

Note the use of \mathbf{D}' instead of \mathbf{D} , in order to ensure that likelihood is compatible between two models (unlearned and retrained).

- Assume classification:

$$\log q(\mathbf{x}' | \mathbf{A}(\mathbf{D})) = \log q(\mathbf{y}' | \mathbf{x}', \mathbf{A}(\mathbf{D})) + \log q(\mathbf{x}' | \mathbf{A}(\mathbf{D})) = \log q(\mathbf{y}' | \mathbf{x}', \mathbf{A}(\mathbf{D})) + C'$$

- In other words, unlearning should make the retrained and unlearned models compute the same function (from \mathbf{x} to \mathbf{y}) on average.

Reference-free evaluation of unlearning

- The strictest form is unfortunately intractable to check in general.
- Instead, we focus on the idea of comparing the functions:

$$\text{Precision}(q, q') = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y} \sim q(\cdot | \mathbf{x})} [\log q'(\mathbf{y} | \mathbf{x})]$$

$$\text{Recall}(q, q') = \mathbb{E}_{\mathbf{x}} \mathbb{E}_{\mathbf{y} \sim q'(\cdot | \mathbf{x})} [\log q(\mathbf{y} | \mathbf{x})]$$

- In other words, **how similar are q (unlearned) and q' (retrained) models?**

Reference-Specific Unlearning Metrics Can Hide the Truth: A Reality Check

Sungjun Cho¹ Dasol Hwang² Frederic Sala¹ Sangheum Hwang³ Kyunghyun Cho^{4,5} Sungmin Cha⁴

Abstract

Evaluating the effectiveness of unlearning in large

Unlearned
Model



Precision = $\mathbb{E}_{\mathbf{x} \sim f} [\log g(\mathbf{x})]$

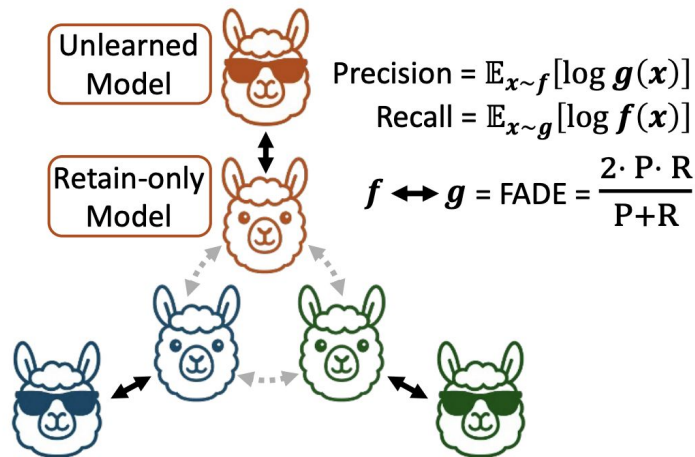
Reference-free evaluation of unlearning

- Because learning is stochastic in most cases,

$$\text{Precision} = \mathbb{E}_{\{q' \sim A(D \setminus \{x\})\}} \mathbb{E}_{\{q \sim M(q', D, x)\}} [\text{Precision}(q, q')]$$

$$\text{Recall} = \mathbb{E}_{\{q' \sim A(D \setminus \{x\})\}} \mathbb{E}_{\{q \sim M(q', D, x)\}} [\text{Recall}(q, q')]$$

In other words, we (approximately) marginalize out the stochasticity of a learning algorithm, in order to truly evaluate the effectiveness of an unlearning algorithm.



Reference-free evaluation of unlearning

- Reference-free evaluation simply checks whether the computed function (conditional distribution) after unlearning closely matches that after retraining.
 - In other words, it focuses on global impact of unlearning (x,y) .
- Reference-free evaluation takes into account randomness in learning as well.
 - In other words, this metric evaluates the algorithm not the resulting model.
- This is a costly evaluation method, since we need multiple training runs.
 - In other words, this metric in its original form cannot be used as a (un)learning objective.
 - We would need some severe approximation.

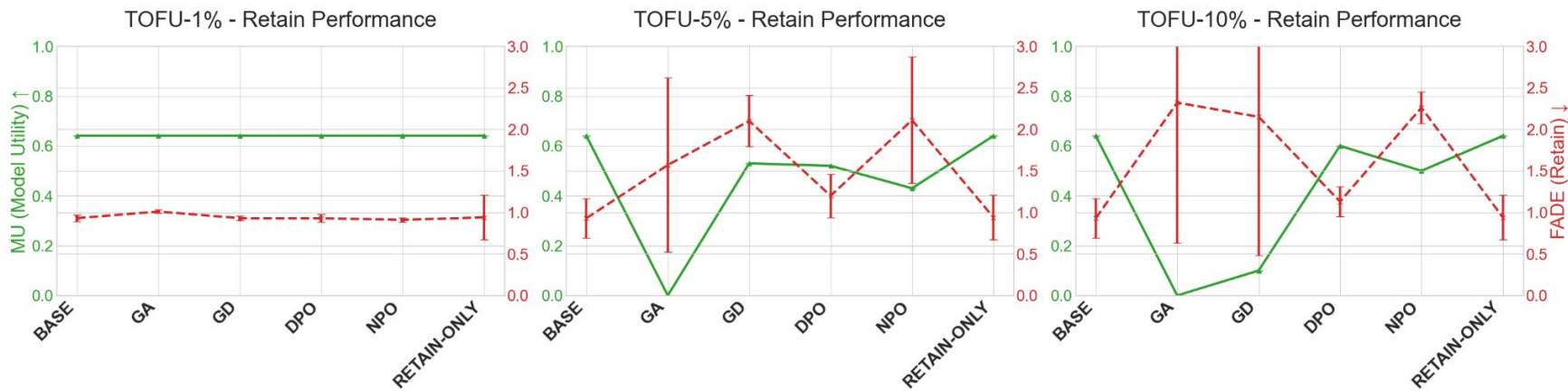
Reference-free evaluation of unlearning

- Most of the unlearning algorithms (NPO, GA, GD & DPO) do not change the actual function/distribution much away from the base (pre-unlearning) model on the forget instances.
- It is almost impossible to distinguish between the base and unlearned models as they are all within the confidence intervals.



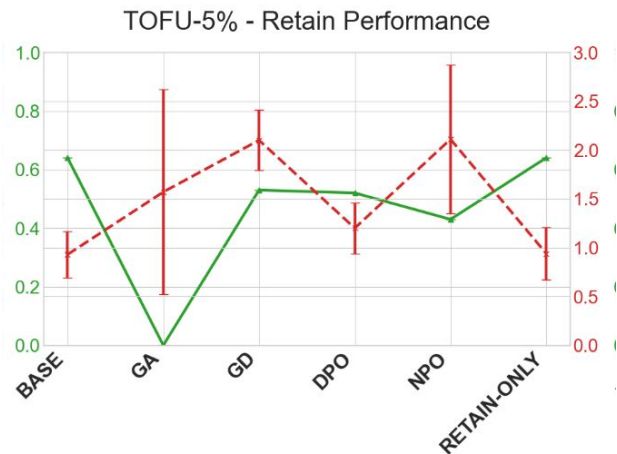
Reference-free evaluation of unlearning

- There's a bigger impact of unlearning undesired instances on retain instances which are instances we want to keep (that is, not unlearn.)
- The function on these retain instances may significantly deviate from both base (pre-unlearned) model as well as the oracle model, for all unlearning algorithms.



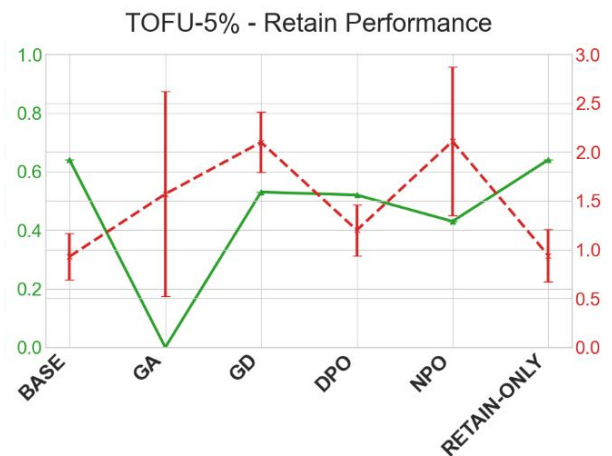
What went wrong here?

- Leaderboard chasing is a valid strategy if we know how to rank hypotheses on the leaderboard properly.
- To do so, we need a one (or few) reasonable metric to rank the hypotheses.
- The field of unlearning is moving so fast, to the extent that **the community hasn't even agreed on a reasonable metric.**
- When we look at the most basic metric we can derive from the most basic idea, **essentially no unlearning algorithm does anything beyond doing nothing.**



What went wrong here?

- Leaderboard chasing makes sense if we gradually improve hypotheses given a fixed evaluation metric to validate/refute those hypotheses.
- But, because papers are more likely to be accepted if the hypothesis was validated (rather than refuted, naturally), we are often motivated to also update/choose the evaluation metric to increase the chance of validating our hypotheses.
 - This resembles *p-hacking* in natural science.
- We really shouldn't do so ...



Time to take a step back and breathe ...

- Unlearning has many faces, and accordingly, many ways to evaluate the effectiveness of algorithms.
- Without specifying the problem carefully and thereby the evaluation metric carefully, leaderboard chasing will fail to produce a meaningful series of progress.
- Perhaps it is time to take a step back, breathe and re-think how we approach unlearning.



Liu et al. (2025)

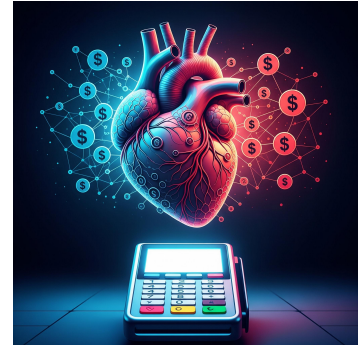
Table 2: A summary of existing LLM unlearning problems through unlearn and efficiency. An asterisk (*) indicates the incapability of evaluating unlearn retraining these models.

Related work	Unlearning targets/tasks	Influence erasure methods
(Lu et al., 2022)	Reducing toxic content, avoiding undesirable sentiments, and preventing repeated text generation	Reward-reinforced model fine-tuning
(Jang et al., 2022)	Degenerating private information, w/ unlearning response irrelevant to this info	Gradient ascent-based fine-tuning
(Kumar et al., 2022)	Text de-classification, w/ unlearning response close to that of retraining*	Sharded, isolated, sliced, and aggregated (SISA) training via adapter
(Ilharco et al., 2022) (Zhang et al., 2023b)	Degenerating toxic content	Task vector-based parameter-efficient fine-tuning via LoRA
(Wang et al., 2023)	Text de-classification/de-generation, unlearning specific words in translation, w/ response close to that of retraining*	KL-divergence-based fine-tuning
(Yu et al., 2023)	Unlearning gender and profession bias, with de-biased unlearning response	Weight importance-informed & relabeling-based fine-tuning
(Pawelczyk et al., 2023)	Text de-classification, w/ unlearning response close to that of retraining*	In-context learning
(Eldan & Russinovich, 2023)	Degenerating Harry Potter-related book content, w/ unlearning response irrelevant to Harry Potter	Relabeling-based fine-tuning
(Ishibashi & Shimodaira, 2023)	Unlearning knowledge from QA dataset, with refusal response (e.g., 'I don't know')	Relabeling-based fine-tuning
(Chen & Yang, 2023)	Text de-classification and de-generation, with response close to that of retraining*	KL divergence-based parameter-efficient fine-tuning via adapter
(Wu et al., 2023b)	Degenerating private information, w/ unlearning response irrelevant to this info	Importance-based neuron editing
(Yao et al., 2023)	Degenerating harmful prompts, degenerating Harry Potter-related book content, and reducing hallucination	Integration of gradient ascent, random labeling, & KL divergence-based fine-tuning
(Maini et al., 2024)	TOFU: Unlearning biographical knowledge about fictitious authors	Fine-tuning with various objectives
(Patil et al., 2024)	Degenerating sensitive information using factual information as a testbed	Model editing techniques and constrained finetuning
(Thaker et al., 2024)	Harry Potter questions and author biography in TOFU (Maini et al., 2024)	Guardrailing with a separate LLM
(Zhang et al., 2024c)	Fictitious unlearning using TOFU (Maini et al., 2024)	Negative preference optimization
(Li et al., 2024b)	Hazardous knowledge in the domain of biology, cybersecurity, and chemistry	Optimization towards random representations for unlearning concept
(Barbulescu & Triantafillou, 2024)	Specific text sequences memorized by LLM	Memorization-aware gradient ascent
(Wang et al., 2024c)	Private, toxic, and copyrighted knowledge	Factual relation removal in MLP layers
(Wang et al., 2024a)	Fictitious unlearning using TOFU (Maini et al., 2024)	Reverse KL divergence based knowledge distillation
(Liu et al., 2024)	Fictitious unlearning using TOFU (Maini et al., 2024), hazardous knowledge using WMDP (Li et al., 2024b), copyrighted content in news articles and books	Detecting the forget prompts and corrupting their embedding space

Expertise matters

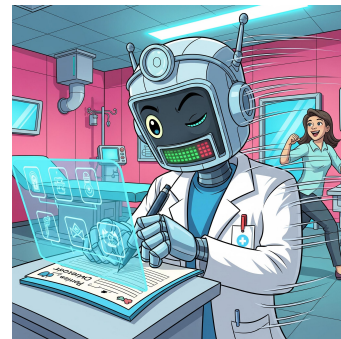
AI for healthcare

- Let's start by acknowledging that **healthcare is less about health and more an extremely challenging resource-constrained optimization problem.**
 - Everyone is born into healthcare and dies inside it.
 - Every minute used by healthcare professionals for paperwork is every minute lost from diagnosing and/or treating a patient.
- With AI we have a chance to solve this constrained optimization problem better.
 - It is not only about diagnosing one disease, treating one disease, etc.
 - It is about improving overall healthcare more efficient and thereby reach a broader population.



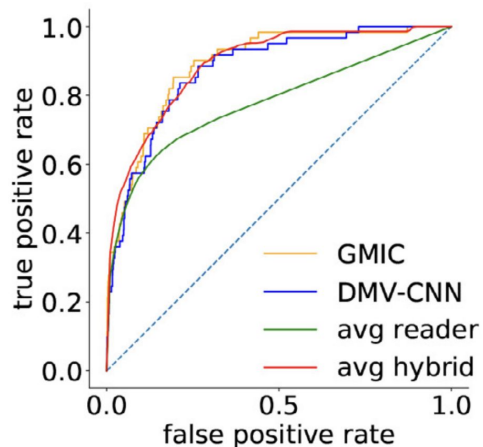
Clinical note generation

- A major time sink of many clinicians is writing clinical notes. Because clinical notes collectively inform future clinicians of the patient's overall condition, they must be written carefully and thoroughly.
- By providing a reasonable draft of clinical notes based on observations of patients, **we may be able to dramatically reduce the overhead on clinicians, and thereby return those saved hours back to patient care.**
- This idea is already being implemented and deployed in real world:
 - Abridge (<https://www.abridge.com/>) AI-generated notes based on physician-patient conversations.
 - Epic's end-of-shift note drafting
 - ...

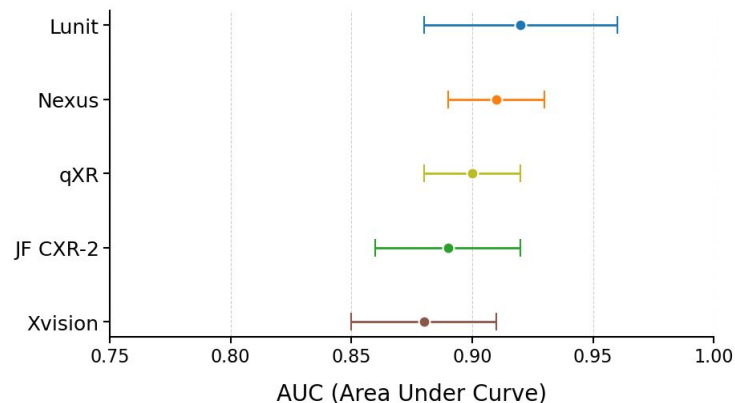


Radiology note generation

- There is a growing issue of **radiologist shortage**, and such shortage will get worse as the trend toward early screening and more preventive healthcare grows.
- There is a promising set of developments in using AI for supporting radiologists.
 - Early diagnosis of breast cancer from mammograms.
 - Scalable diagnosis of tuberculosis (TB)
- **Can we then create a radiology note generation model?**



Shen et al. [2020]




Qin & Walt et al. [2024]

AI for healthcare: evaluation matters

- Highly predictive models have great potential to improve the efficiency of healthcare, and thereby the impact of healthcare.
- The bar for AI in healthcare is however significantly higher than in many other areas, for many good reasons.
 - Heroin was sold until 1924 as a treatment for common cold, TB all the way to morphine addiction
 - Elixir sulfanilamide killed 100+ people in 1937, which led to strengthening of FDA's authority over drugs.
- Because everyone is part of healthcare, we must be extra-careful at assessing the efficacy and safety of any new technology for healthcare, before deploying it widely.



Radiology report generation: evaluation matters

- Three criteria of good evaluation metric for free-form text generation systems
 1. **Semantic similarity**: the metric must not rely purely on surface-level forms but must dig deeper and compare two text snippets based on their meaning

 2. **Interpretability**: we must be able to tell what led to the score. This is especially critical in mission-critical scenarios, such as in healthcare.
 3. **Scalability**: the metric must be scalable to hundreds of thousands of reports, in order to cover long-tail phenomena (prevalent in healthcare.)

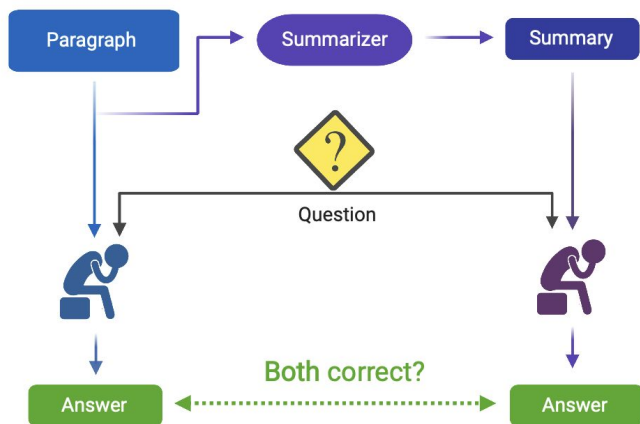
A string of evaluation metrics

- There are **many different metrics** that have been used for radiology note generation models over the past few years.
- These include old-school metrics such as BLEU and ROUGE (still!!!)
- There are few more that are specific to radiology notes, although they often don't satisfy these criteria.
- Especially, it is **difficult to satisfy "interpretability"**.

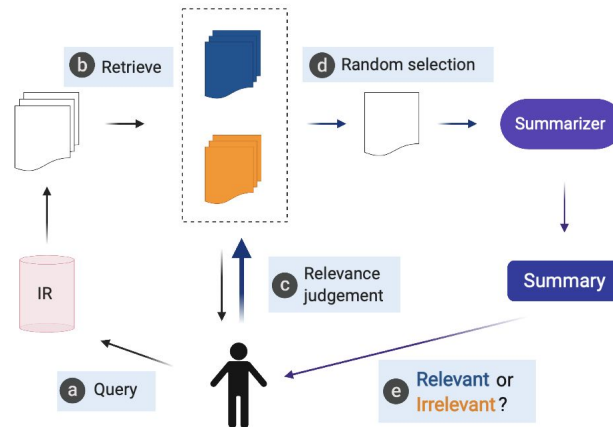
Metric	Semantic Understanding	Interpretability	Scalability
BLEU	✗	✗	✓
ROUGE	✗	✗	✓
BERTScore	✓	✗	✓
F1-CheXpert	✓	✗	✓
SembScore	✓	✗	✓
F1-RadGraph	✓	●	✓
GREEN	✓	✗	✓
FinerAdScore	✓	●	✓
RaTEScore	✓	●	✓
G-Rad	✓	✗	✓
RadFact	✓	●	✓

Going back in time ...

- Two text snippets are similar to each other if their consequences are similar.
 - So-called* extrinsic evaluation.



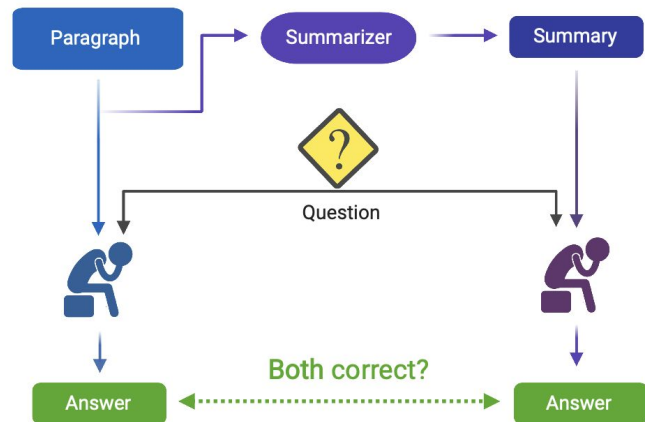
QA-based evaluation
[Morris et al., 1992; Wang et al., 2020]



IR-based evaluation
[Dorr et al., 2005; Daume & Marcus, 2005]

Going back in time ...

- Two text snippets are similar to each other if their consequences are similar.
 - So-called **extrinsic evaluation**.
- Such extrinsic evaluation satisfies the interpretability condition.
 - We just need to **inspect which questions were incorrectly/differently answered given one or the other text snippets**.
- Often extrinsic evaluation has a **high ramp-up cost** due to the necessity of external systems.
 - For QA: we need to prepare questions and answer them.
 - For IR: we need to prepare an IR system and ready to measure the relevance.



QA-based evaluation
[Morris et al., 1992; Wang et al., 2020]

iCARE

- Instead of going deeper into the method ... here's a slide from my talk in 2020.
- We can **use an off-the-shelf LLM to automatically create questions and answer them** based on either reference or generated text snippets.

I can use reading comprehension only if...

- I could ask **ETS** to create TOEFL questions for any summary, and
- I could ask **you all** take TOEFL **every day** for my research.
- Or, perhaps I could replace ETS and y'all with **neural nets**



iCARE

- This is still work in preparation, and I will skip the details. Please reach out to me or the amazing first author **Radhika Dua** at NYU.

Clinically Grounded Agent-based Report Evaluation: An Interpretable Metric for Radiology Report Generation

Radhika Dua^{1,2}, Young Joon (Fred) Kwon⁴, Siddhant Dogra^{4†},
Daniel Freedman^{4†}, Diana Ruan^{4†}, Motaz Nashawaty^{4†},
Danielle Rigau^{4†}, Daniel Alexander Alber^{2,5†}, Kyunghyun Cho^{1,3},
Eric Oermann^{1,2,4*}

^{1*}Center for Data Science, New York University, 60 5th Ave, New York, 10019, NY, USA.

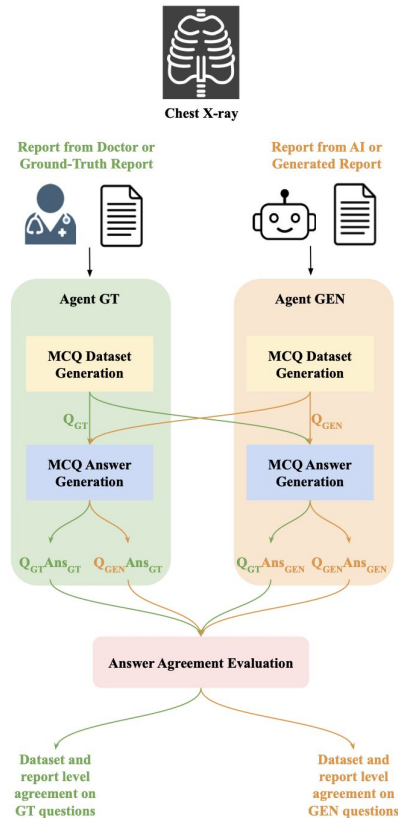
²Department of Neurosurgery, NYU Langone Health, 450 First Avenue, New York City, 10019, NY, USA.

³Prescient Design, Genentech, 149 5th Ave. 3rd floor, New York, 10019, NY, USA.

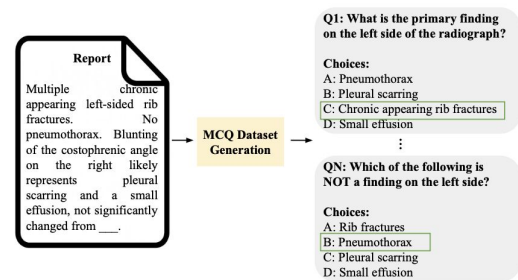
⁴Department of Radiology, NYU Langone Health, 450 First Avenue, New York City, 10019, NY, USA.

⁵NYU Grossman School of Medicine, NYU Langone Health, 450 First

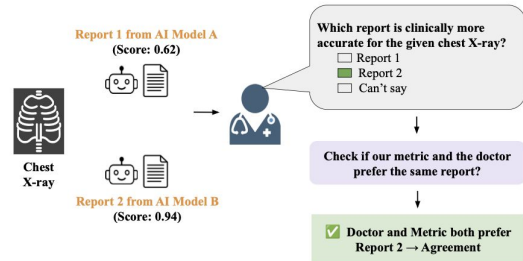
a. Overview of our evaluation framework (iCARE).



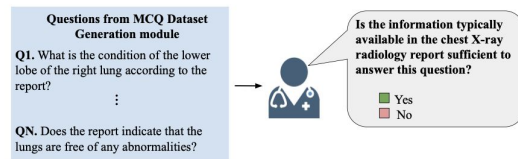
b. Examples from the dataset generated by the MCQ Dataset Generation module.



c. Human evaluation setup to assess correlation with our metric.



d. Human evaluation setup to assess generated questions.



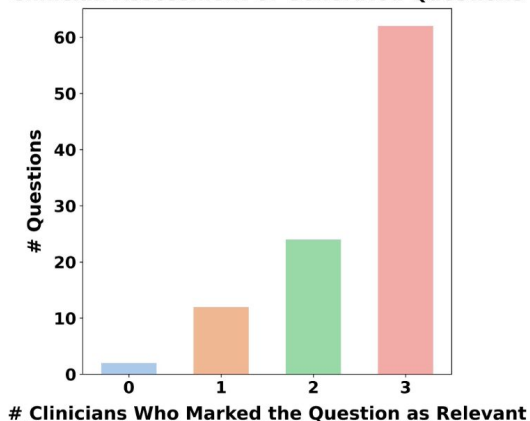
Evaluating evaluation metrics

- What does it mean for an evaluation metric to be correct?
 - Perhaps it should correlate with the assessment by humans (us!)
 - But, which of us?
- This is where **the challenge of AI for anything serious** comes in.
 - **We can't simply deploy a chatbot and collect 👍/👎.**
- We recruited three clinicians at NYU Langone and ask them to assess 154 pairs of clinical notes each:
 - These practicing clinicians were asked to tell which of two notes given the corresponding shared radiology image was better (or can't tell).
 - We also showed them 300 automatically generated questions and asked them to tell whether they are clinically relevant.

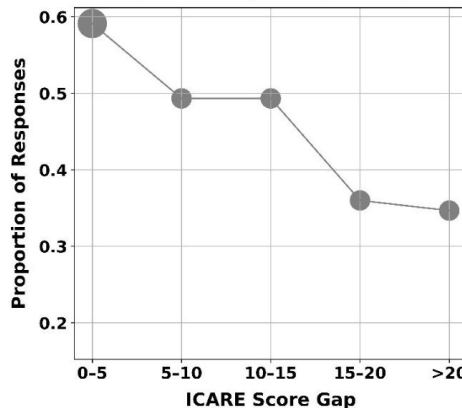
Clinician assessment vs. iCARE

- Most of the questions were considered clinically relevant.
- Agreement between iCARE and clinician's assessment is pretty good, although not perfect.
- A good starting point , but ... how about other metrics?

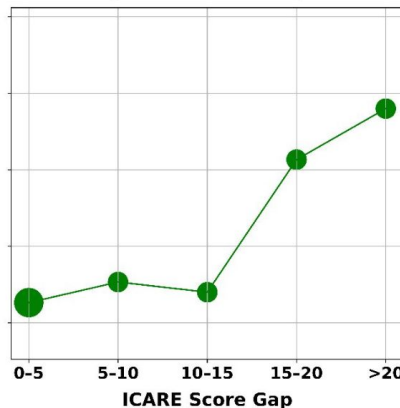
Clinician Assessment of Generated Questions



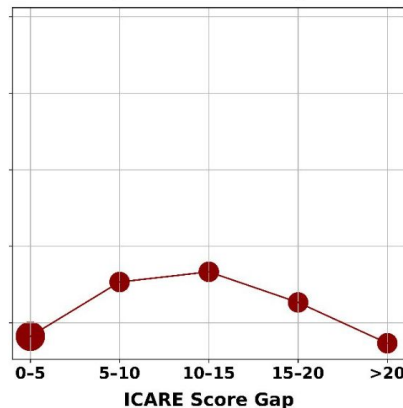
Indecision Rate
("Can't Say" Responses)



Doctor-Metric Alignment Rate
(Doctor = Metric)

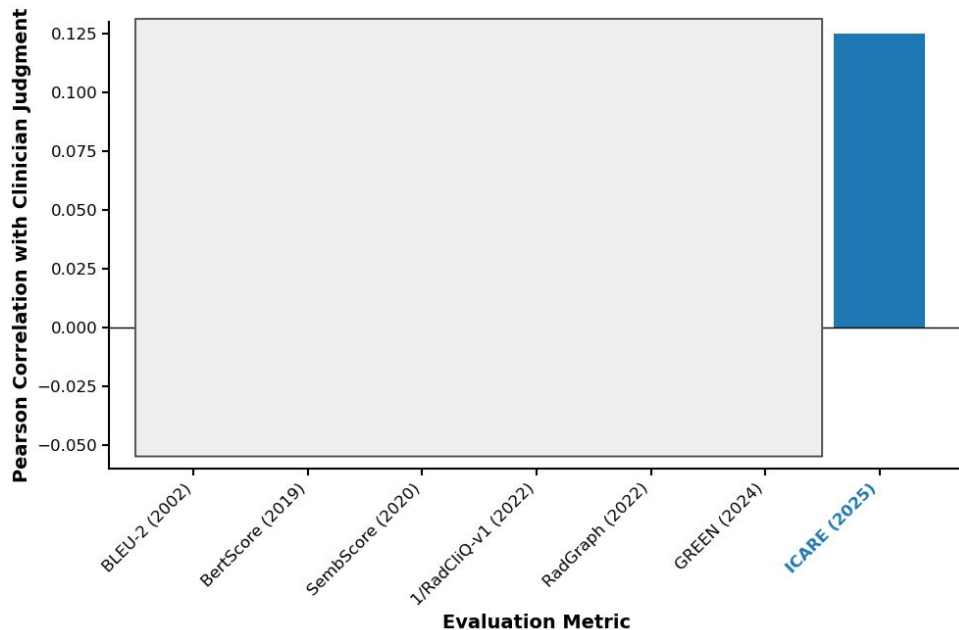


Doctor-Metric Misalignment Rate
(Doctor \neq Metric)



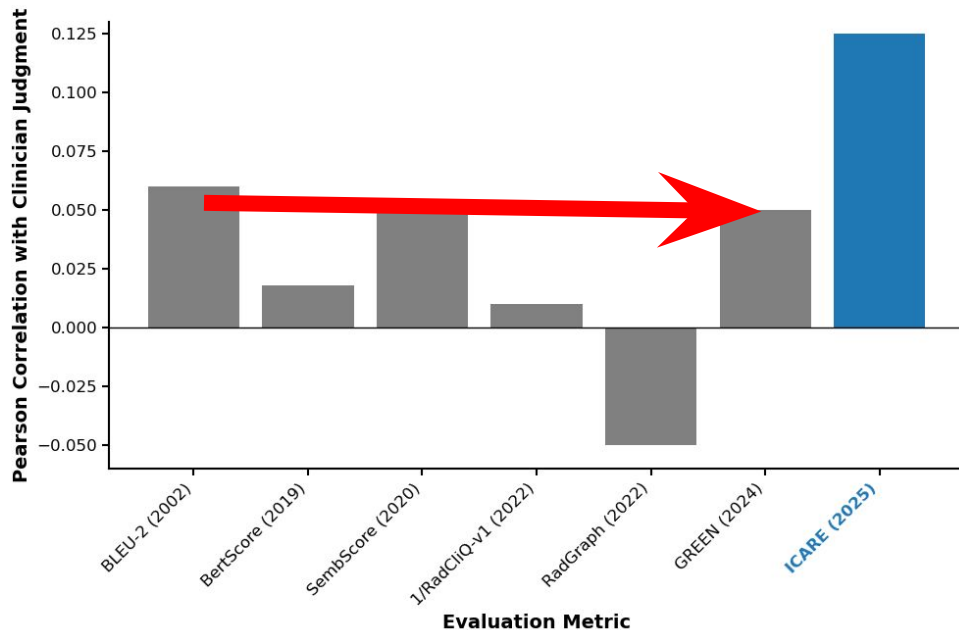
Clinician assessment vs. all metrics

- Pearson correlation based on $\{-1, 0, 1\}$.
- iCARE correlates reasonably with the clinician's judgement (around 0.125). this is reasonable, although it is not perfect.



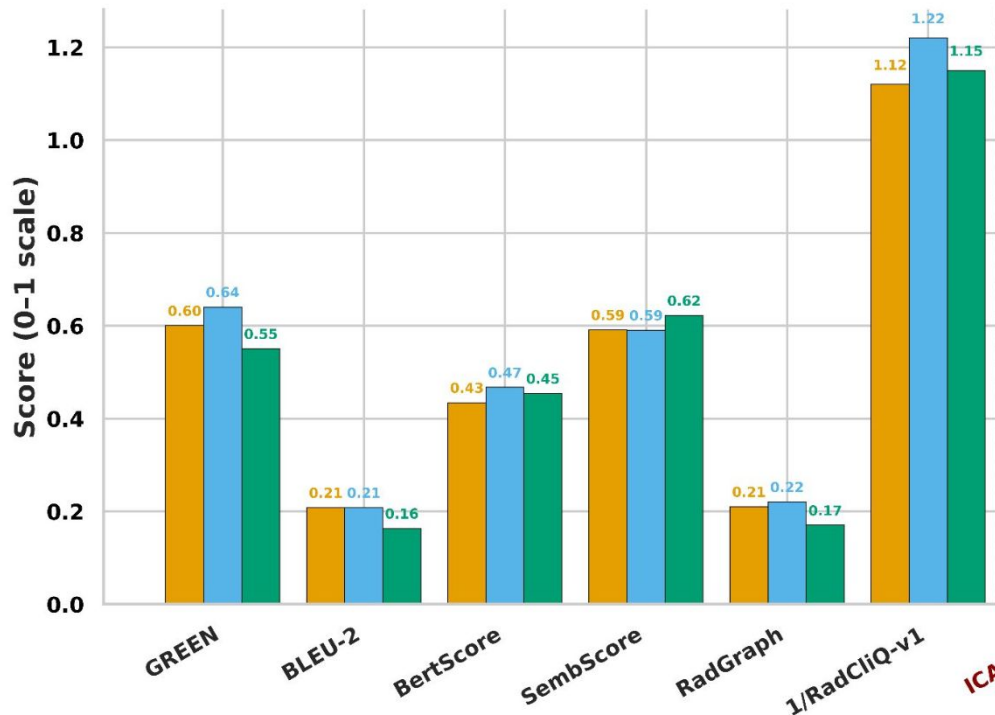
Clinician assessment vs. all metrics

- Pearson correlation based on $\{-1, 0, 1\}$.
- Most of the other metrics have significantly lower correlation with clinician judgement. In fact, BLEU-2 [Papiani et al., 2002] is as good as any other recent metrics 🧑.



These metrics can fool us ...

- We compared three different models; a (a) CheXpertPlus trained on MIMIC, a (b) CheXpertPlus trained on CheX+MIMIC trained model and Microsoft's (c) MAIRA2.
- We use publicly available (a), (b) and (c) but re-evaluate them all ourselves.
- Qualitatively, (a) < (b) < (c).
- BUT!!!!



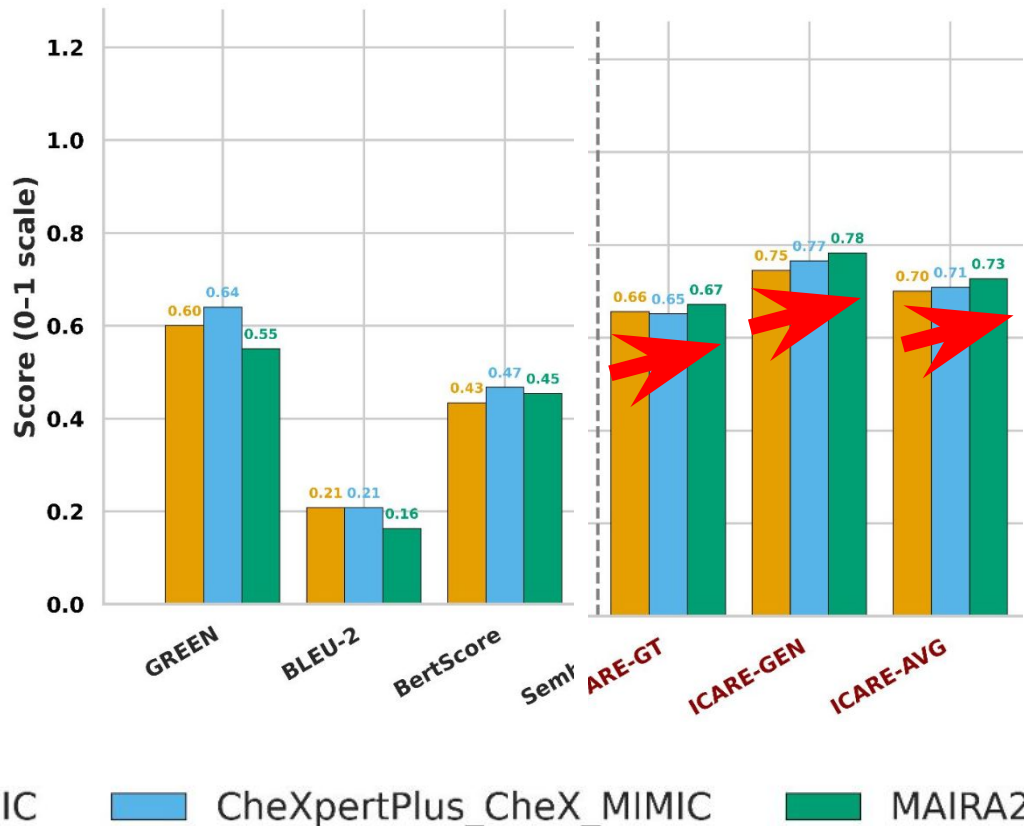
CheXpertPlus_MIMIC

CheXpertPlus_CheX_MIMIC

MAIRA2

These metrics can fool us ...

- We compared three different models; a (a) CheXpertPlus+MIMIC trained model, a (b) CheXpertPlus+CheX+MIMIC trained model and Microsoft's (c) MAIRA2.
- Qualitatively, (a) < (b) < (c).
- iCARE, which reflects clinician assessment better, reflects this qualitative ordering perfectly.



What went wrong here?

- We tend to believe (consciously or subconsciously) that a widely-accepted practice in another field will apply directly to a new field.
 - This happens over and over: using BLEU/ROUGE for image caption generation, dialogue modeling, story generation, etc.*
 - Until one is ready to build up even a tiny bit of expertise in the target field, it is easy for us to make the same mistake at the cost of more publications (!?)
- It takes time to start working on a new problem (especially in a new field). We should not rush ourselves nor rush others. This will ultimately result in a dearth of forgotten (and perhaps actively harmful) papers.

(*) Perhaps it's just BLEU/ROUGE that's an issue.

What went wrong here?

- **Proxy metrics** are scalable but **only proxy**, and we must know if they are adequate. It is difficult to do so, since it requires expertise.
- We must **work with experts to ensure that we do not set up a problem to be meaningless**: easier said than done, but unfortunately it must be done.
- This lesson is an evergreen one that we forget every time.

TABLE 9

Automatic scores of the top five competition submissions.

	CIDER	METEOR	ROUGE	BLEU-4	Rank
Google [46]	0.943	0.254	0.53	0.309	1st
MSR Captivator [34]	0.931	0.248	0.526	0.308	2nd
m-RNN [28]	0.917	0.242	0.521	0.299	3rd
MSR [23]	0.912	0.247	0.519	0.291	4th
m-RNN (2) [28]	0.886	0.238	0.524	0.302	5th
Human	0.854	0.252	0.484	0.217	8th

TABLE 10

Human generated scores of the top five competition submissions.

	M1	M2	M3	M4	M5	Rank
Google [46]	0.273	0.317	4.107	2.742	0.233	1st
MSR [23]	0.268	0.322	4.137	2.662	0.234	1st
MSR Captivator [34]	0.250	0.301	4.149	2.565	0.233	3rd
Montreal/Toronto [31]	0.262	0.272	3.932	2.832	0.197	3rd
Berkeley LRCN [30]	0.246	0.268	3.924	2.786	0.204	5th
Human	0.638	0.675	4.836	3.428	0.352	1st

Vinayls et al. [2016] on Lessons learned from the 2015 MSCOCO Image Captioning Challenge

Wishful thinking doesn't matter

Scaling laws for downstream tasks

- **Scaling laws in machine learning** refer to a set of **simple mathematical relationships between the amount of computation** (in forms such as memory, computation, the number of data points, etc.) **and the predictive accuracy** (in forms of log-probability, 0-1 loss, etc.)
- These laws often exhibit a simple log-linear relationship: **$a \log L + b = \alpha \log C + \beta$** with **$L$** a loss and **$C$** an amount of compute.
- These laws can sometimes be explained as the convergence rate of a statistical estimator.

Learning Curve Theory

Marcus Hutter

DeepMind

<http://www.hutter1.net/>

5 February 2021

Scaling laws for downstream tasks

- The whole LLM community, especially those who train large models, is in love with these scaling laws, as they provide them with guidance on how to decide the model sizes, data sizes, etc. and what to expect from such models, *without* training a whole batch of models of extreme scale.

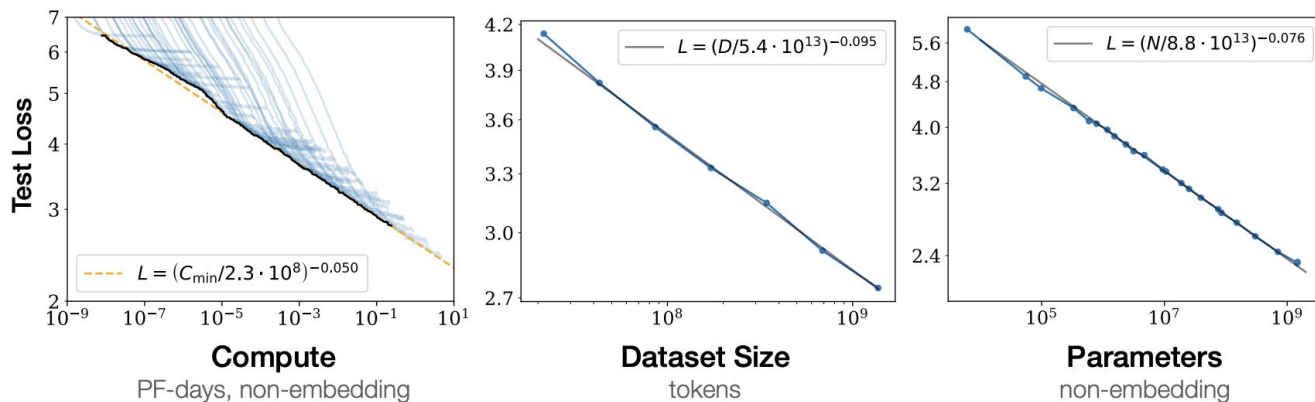


Figure 1 Language modeling performance improves smoothly as we increase the model size, dataset size, and amount of compute² used for training. For optimal performance all three factors must be scaled up in tandem. Empirical performance has a power-law relationship with each individual factor when not bottlenecked by the other two.

Scaling laws for downstream tasks

- The success of scaling laws for training large-scale language models however started to make people wonder if there can be such a simple relationship between the scale of compute/data and the *downstream* task accuracy.
 - After all, do we really care about the log-probability assigned to a held-out internet text? We however do care about how well our model would prove the Riemann hypothesis (apparently!)
- Since the scaling law tells us that there exists a simple relationship between the scale of compute/data and the validation perplexity, if such a simple relationship exists for the downstream accuracy, this means that there is a simple relationship between the validation perplexity and the downstream task.
 - $p \log \text{Acc} + q = a \log L + b = \alpha \log C + \beta$

Scaling laws for downstream tasks

- At the face of it, this feels impossible without specifying what the downstream task is, in advance: **it feels too good to be true.**
- But, these LLMs have done some magical stuffs in recent years, and what if there is a magical formula that predicts the downstream task accuracy?
- In fact, some claim it is possible (to a certain extent).

Average top-1 error is predictable. Figure 1 (*right*) presents our main result in estimating scaling laws for downstream error. Concretely, we use the models indicated in Table 1 to fit Equations (4) and (5), chaining the scaling fits to predict the average top-1 error as a function of training compute C and the token multiplier M . Our fits allow us to predict, using $20\times$ less compute, the downstream performance of a 6.9B model trained on 138B RedPajama tokens to within 0.05% relative error and a 1.4B model trained on RedPajama 900B tokens to within 3.6% relative error.

Table 2 additionally shows the relative error of our downstream performance predictions for models trained on C4, RedPajama, and RefinedWeb, indicating that our scaling law functional forms are applicable on many training datasets. We note that while average accuracy is predictable, *individual* downstream task predictions are significantly more noisy. We report relative error for more model

Scaling laws for downstream tasks

- Gadre et al. [2024] specified that they were able to predict the *average* downstream accuracy across some (rather arbitrary) set of tasks. We decided to look into these tasks as well as more based on another parallel study [Magnusson et al., 2025].
- And, the picture is much messier than it was implied: every task is unique, and each task exhibits a unique relationship between its accuracy and the base model's perplexity [Lourie, Hu & Cho, under review].

Scaling Laws Are Unreliable for Downstream Tasks: A Reality Check

Nicholas Lourie^{1*} Michael Y. Hu^{1*} Kyunghyun Cho^{1,2,3}

¹New York University ²Prescient Design ³ CIFAR LMB

{nick.lourie, michael.hu, kyunghyun.cho}@nyu.edu

Downstream tasks are not monolith

- The relationship between the downstream accuracy and the base model's perplexity is highly nonlinear and often **unpredictable** (highly noisy).
- The shape of nonlinearity is highly nontrivial, spanning everything from exponential curves to a flat line.

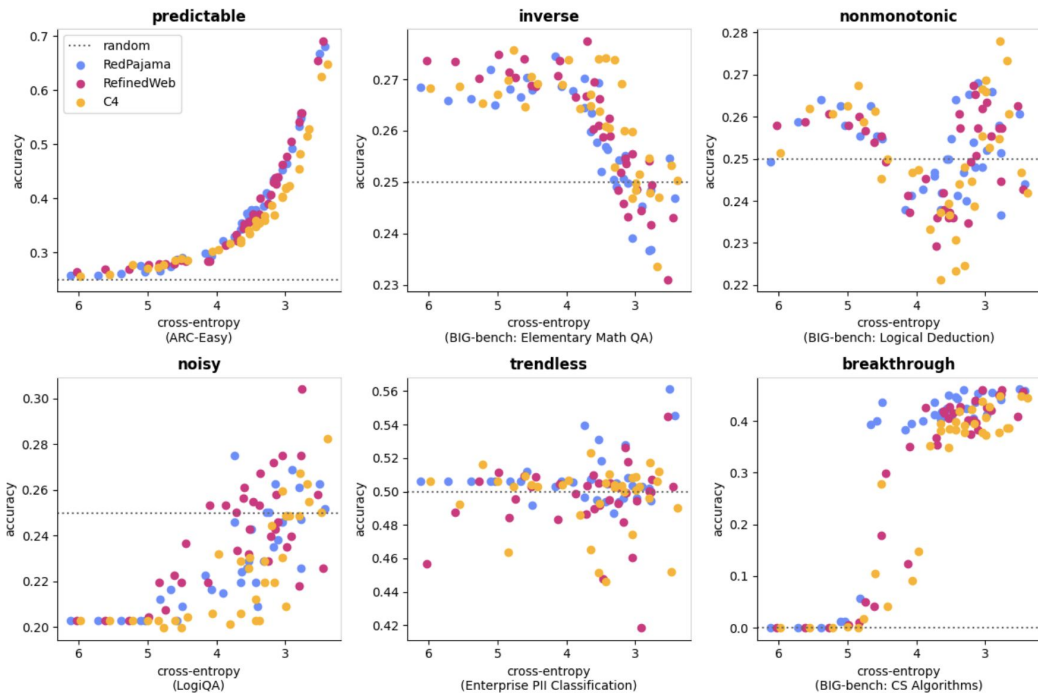


Figure 2: A taxonomy of different scaling behaviors. Predictable scaling fits closely to a linear functional form after exponentiating the cross-entropy loss. However, depending on the downstream task, models do not always improve with scale (inverse, nonmonotonic, and trendless), or the improvement might be highly noisy. The improvement might also be better described by a different functional form like a sigmoid (breakthrough).

Downstream tasks are not monolith

- In fact, **most of these downstream tasks' accuracies are *not* predictable.**
- The average accuracy was predictable in [Gardre et al., 2024], simply because a majority of these tasks' accuracies are not predictable and are effectively ignored as noisy.
- Perhaps, this is still okay, as we can study each downstream task separately and draw some sensible conclusions. Or, is it?

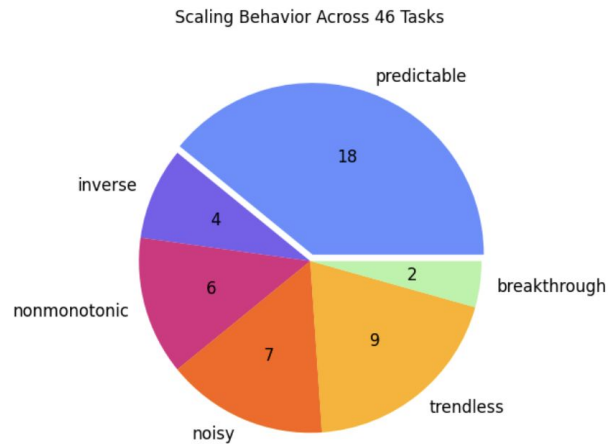


Figure 1: Revisiting the 46 tasks studied in [Gadre et al. \(2024\)](#), we find that only 18 tasks—or 39%—demonstrate smooth, predictable improvement. The 18 predictable tasks and the other 28 are shown in Figures 5 through 10, where we group them into different degenerate scaling behaviors: inverse, nonmonotonic, noisy, trendless, and breakthrough scaling. See Figure 2 for examples.

Sensitivity to experimental settings

- It turned out **it is not trivial to even study the impact of validation perplexity on a *single* downstream task.**
- Unless we are extensive in experiments, we can easily draw a wrong conclusion, or any conclusion we want to draw.

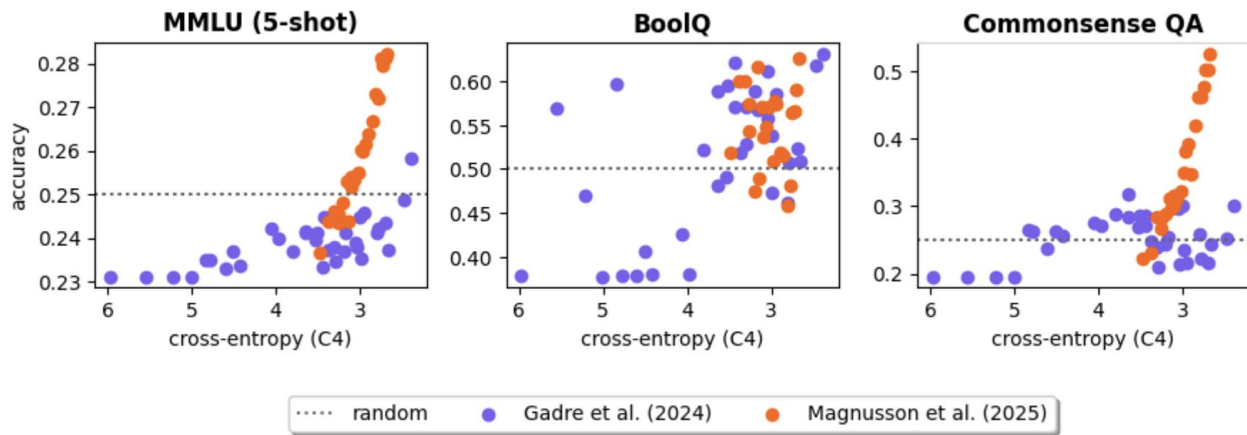
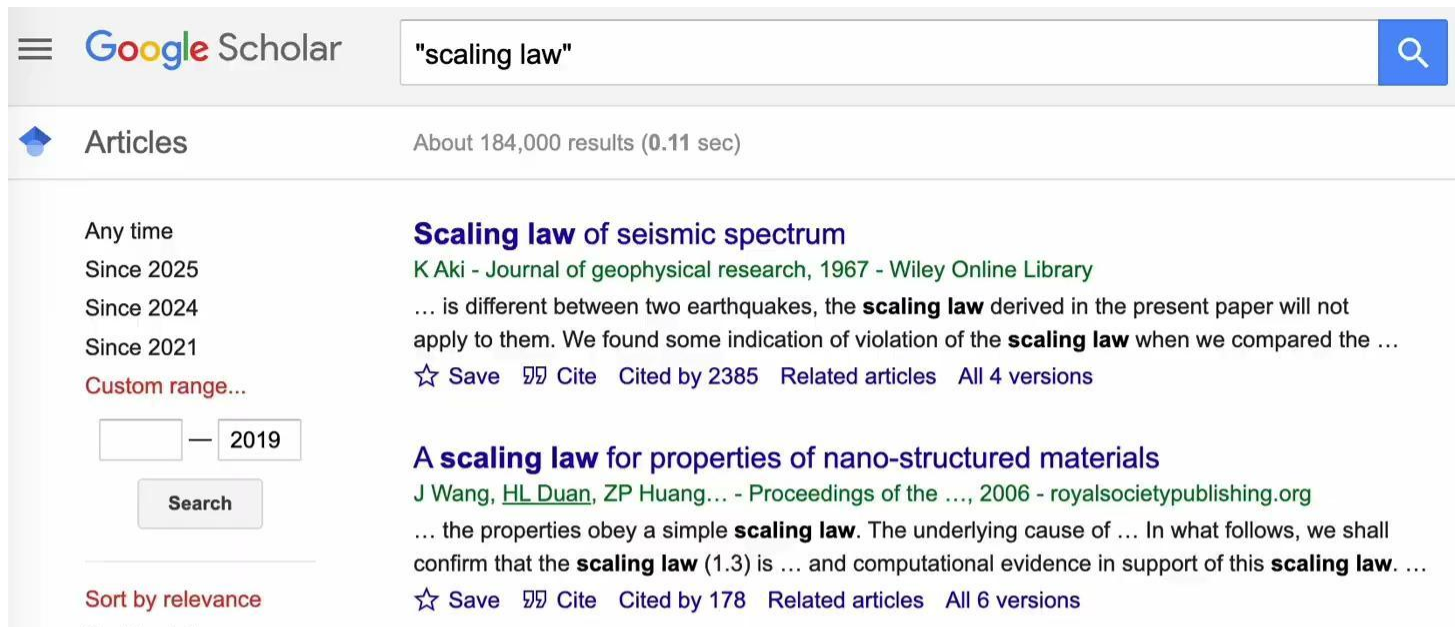


Figure 4: Scaling behavior will change depending on the experimental setting. [Gadre et al. \(2024\)](#) and [Magnusson et al. \(2025\)](#) both train language models on internet-scale pretraining corpora. But even with the same validation dataset and downstream task, scaling trends can be different.



Everyone wants their own law

- It is fashionable to come up with yet another “law” in machine learning.
- The publish-or-perish culture encourages us to be surprised by some of these findings and to brand them as laws and publish them as rapidly as possible.



The screenshot shows a Google Scholar search for "scaling law". The search bar at the top contains the text "scaling law" and a magnifying glass icon. Below the search bar, the results are categorized under "Articles" with a subtext "About 184,000 results (0.11 sec)". On the left side, there are filters for "Any time", "Since 2025", "Since 2024", "Since 2021", and a "Custom range..." option with a date selector showing a range from an empty box to "2019". A "Search" button is located below the date selector. At the bottom left, there is a "Sort by relevance" option. The main content area displays two search results. The first result is titled "Scaling law of seismic spectrum" by K Aki, published in the Journal of geophysical research in 1967, from Wiley Online Library. The abstract snippet states: "... is different between two earthquakes, the **scaling law** derived in the present paper will not apply to them. We found some indication of violation of the **scaling law** when we compared the ...". Below the abstract are links for "Save", "Cite", "Cited by 2385", "Related articles", and "All 4 versions". The second result is titled "A scaling law for properties of nano-structured materials" by J Wang, HL Duan, and ZP Huang, published in the Proceedings of the ... in 2006, from royalsocietypublishing.org. The abstract snippet states: "... the properties obey a simple **scaling law**. The underlying cause of ... In what follows, we shall confirm that the **scaling law** (1.3) is ... and computational evidence in support of this **scaling law**. ...". Below the abstract are links for "Save", "Cite", "Cited by 178", "Related articles", and "All 6 versions".

Google Scholar

"scaling law"

Articles About 184,000 results (0.11 sec)

Any time
Since 2025
Since 2024
Since 2021
Custom range...
— 2019
Search

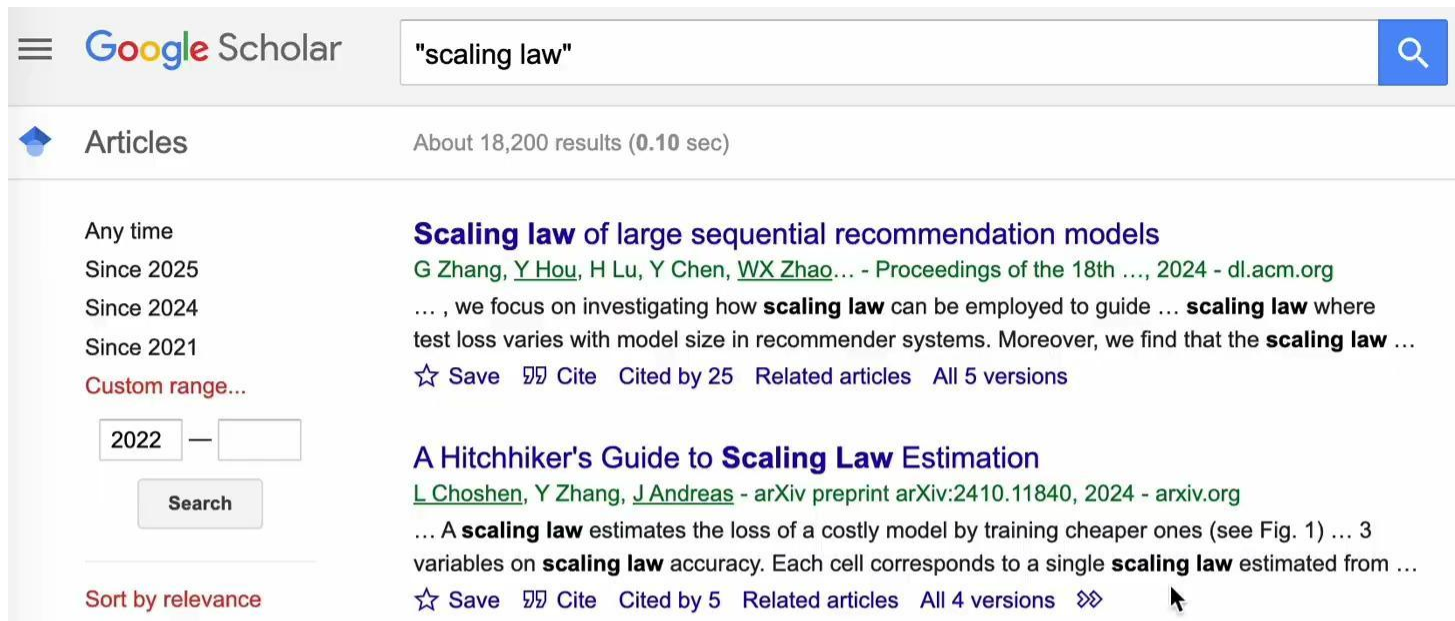
Sort by relevance

Scaling law of seismic spectrum
K Aki - Journal of geophysical research, 1967 - Wiley Online Library
... is different between two earthquakes, the **scaling law** derived in the present paper will not apply to them. We found some indication of violation of the **scaling law** when we compared the ...
☆ Save Cite Cited by 2385 Related articles All 4 versions

A scaling law for properties of nano-structured materials
J Wang, HL Duan, ZP Huang... - Proceedings of the ..., 2006 - royalsocietypublishing.org
... the properties obey a simple **scaling law**. The underlying cause of ... In what follows, we shall confirm that the **scaling law** (1.3) is ... and computational evidence in support of this **scaling law**. ...
☆ Save Cite Cited by 178 Related articles All 6 versions

Everyone wants their own law

- It is fashionable to come up with yet another “law” in machine learning.
- The publish-or-perish culture encourages us to be surprised by some of these findings and to brand them as laws and publish them as rapidly as possible.



The screenshot shows a Google Scholar search interface. At the top, the Google Scholar logo is on the left, and a search bar contains the text "scaling law" with a magnifying glass icon on the right. Below the search bar, the word "Articles" is displayed with a blue arrow icon, followed by the text "About 18,200 results (0.10 sec)". On the left side, there is a sidebar with filters: "Any time", "Since 2025", "Since 2024", "Since 2021", and "Custom range...". Below these filters is a date range selector showing "2022" and a minus sign, and a "Search" button. At the bottom of the sidebar, it says "Sort by relevance". The main content area displays two search results. The first result is titled "Scaling law of large sequential recommendation models" by G Zhang, Y Hou, H Lu, Y Chen, and WX Zhao, published in the Proceedings of the 18th ... in 2024 on dl.acm.org. The abstract snippet reads: "... , we focus on investigating how **scaling law** can be employed to guide ... **scaling law** where test loss varies with model size in recommender systems. Moreover, we find that the **scaling law** ...". Below the abstract are links for "Save", "Cite", "Cited by 25", "Related articles", and "All 5 versions". The second result is titled "A Hitchhiker's Guide to Scaling Law Estimation" by L Choshen, Y Zhang, and J Andreas, published as an arXiv preprint arXiv:2410.11840 in 2024 on arxiv.org. The abstract snippet reads: "... A **scaling law** estimates the loss of a costly model by training cheaper ones (see Fig. 1) ... 3 variables on **scaling law** accuracy. Each cell corresponds to a single **scaling law** estimated from ...". Below this abstract are links for "Save", "Cite", "Cited by 5", "Related articles", "All 4 versions", and a double diamond icon.

Google Scholar

"scaling law"

Articles About 18,200 results (0.10 sec)

Any time
Since 2025
Since 2024
Since 2021
Custom range...
2022 —
Search
Sort by relevance

Scaling law of large sequential recommendation models
G Zhang, [Y Hou](#), H Lu, Y Chen, [WX Zhao](#)... - Proceedings of the 18th ..., 2024 - dl.acm.org
... , we focus on investigating how **scaling law** can be employed to guide ... **scaling law** where test loss varies with model size in recommender systems. Moreover, we find that the **scaling law** ...
☆ Save Cite Cited by 25 Related articles All 5 versions

A Hitchhiker's Guide to Scaling Law Estimation
[L Choshen](#), Y Zhang, [J Andreas](#) - arXiv preprint arXiv:2410.11840, 2024 - arxiv.org
... A **scaling law** estimates the loss of a costly model by training cheaper ones (see Fig. 1) ... 3 variables on **scaling law** accuracy. Each cell corresponds to a single **scaling law** estimated from ...
☆ Save Cite Cited by 5 Related articles All 4 versions

Empirical laws are hard to come by ...

- It is fashionable to come up with yet another “law” in machine learning.
- The publish-or-perish culture encourages us to be surprised by some of these findings and to brand them as laws and publish them as rapidly as possible.
- But, **it is likely that we are fooling ourselves** by either looking only at where we would find such publishable laws or the universality of such law-looking phenomena.

Power laws do have an interesting and possibly even important role to play, but one needs to be very cautious when interpreting them. **The most productive use of power laws in the real world will therefore, we believe, come from recognizing their ubiquity (and perhaps exploiting them to simplify or even motivate subsequent analysis) rather than from imbuing them with a vague and mistakenly mystical sense of universality.**

MATHEMATICS

Critical Truths About Power Laws

Michael P. H. Stumpf¹ and Mason A. Porter²

The ability to summarize observations using explanatory and predictive theories is the greatest strength of modern science. A theoretical framework is perceived as particularly successful if it can explain very disparate facts. The observation that some apparently complex phenomena can exhibit startling similarities to dynamics generated with simple mathematical models (*1*) has led to empirical searches for fundamental laws by inspecting data for qualitative agreement with the behavior of such models. A strik-

calculations of power laws offer little more than anecdotal value.

By power-law behavior, one typically means that some physical quantity or probability distribution $y(x)$ satisfies (2, 3)

$$y(x) \propto x^{-\lambda} \text{ for } x > x_0,$$

where λ is called the “exponent” of the power law. In the equation, the power-law behavior occurs in the tail of the distribution (i.e., for $x > x_0$). A power-law distribution has

Imperfect empirical laws can mislead

- These scaling laws are often used to make decisions on model sizes, compute budgets, data sizes, etc.
- If a scaling law was derived under some restrictions (e.g., a fixed compute budget, a fixed data budget, a particular choice of a learning algorithm, etc.,) one could arrive at a suboptimal decision when operating under another set of restrictions.

We investigate the optimal model size and number of tokens for training a transformer language model under a given compute budget. We find that current large language models are significantly under-trained, a consequence of the recent focus on scaling language models whilst keeping the amount of training data constant. By training over 400 language models ranging from 70 million to over 16 billion parameters on 5 to 500 billion tokens, we find that for compute-optimal training, the model size and the number of training tokens should be scaled equally: for every doubling of model size the number of training tokens should also be doubled. We test this hypothesis by training a predicted compute-optimal model, *Chinchilla*, that uses the same compute budget as *Gopher* but with 70B parameters and 4× more more data. *Chinchilla* uniformly and significantly outperforms *Gopher* (280B), GPT-3 (175B), Jurassic-1 (178B), and Megatron-Turing NLG (530B) on a large range of downstream evaluation tasks. This also means that *Chinchilla* uses substantially less compute for fine-tuning and inference, greatly facilitating downstream usage. As a highlight, *Chinchilla* reaches a state-of-the-art average accuracy of 67.5% on the MMLU benchmark, greater than a 7% improvement over *Gopher*.

What is going wrong here?

- An empirical law is valuable when it is as concise/parsimonious as possible and still makes valid predictions. Because of the first condition (parsimony), it is unlikely that there could be many competing laws for one phenomenon.
- How can there be a such rapid series of new scaling laws proposed in the field?
 - This implies that **we are not comparing these so-called laws properly, and we may be simply finding different setups (restrictions) under which new laws fit data better than old laws.**
- Each paper may simply be a description of the authors' setup rather than a proposal of a new empirical law.
 - Such a law would be still valuable for the authors themselves for their uses.
 - Such a law would be meaningless both scientifically and practically for others.

What is going wrong here?

- More specifically for downstream task performance, there was never meant to be a generally applicable scaling law w.r.t. pretraining performances and setups.
 - Without specifying the relationship between pretraining and downstream task setups, it would be *wishful thinking* to expect such a law to exist.
 - But, we often dream too much and write papers based on those dreams.
- Any reasonable experiment should reveal this empirically, and any reasonable thought experiment should reveal the impossibility.

We study the scaling behavior of the *downstream* performance in machine translation as the pretraining data grows and propose scaling laws for both *downstream* cross-entropy and translation quality metrics. We demonstrate through extensive experiments that the scaling behavior is significantly influenced by (1) the degree of alignment between the pretraining and the downstream data and (2) the finetuning dataset size. In favorable cases where the distributions are sufficiently aligned, we show that downstream translation quality, measured by translation scores, can be accurately predicted using

Empirical laws are great but ...

- Good empirical laws make machine learning predictable.
- Predictability makes machine learning more practical and applicable.
- But, **empirical laws must be far apart from each other, and we cannot simply claim *better* laws non-stop, especially when some of these are simply pipe dreams.**
- I suspect pressure on everyone to rapidly produce more papers may be a culprit behind this phenomenon: we do not have time to wait to see if a purported law truly has predictive power and claimed universality.



Wrapping up ...

Concluding remark

- Leaderboard chasing is a valid approach to research.
 - This is how we've progressed so much so fast.
- We want leaderboard chasing to stay valid.
 - We must set up each problem carefully.
- Often, it takes a lot of efforts and a lot of time to set up a problem carefully.
 - We can't simply skip it, no matter what and how.



Thanks are due to ...

- Researchers who spend enormous efforts to perform reality checks on the apparent (but often false) progress based on leaderboard chasing.
 - (Thanks, Noah!)
- Researchers and engineers who are building increasingly better ways to measure the progress.
 - Such as *dynamic benchmark* [Kiela et al., 2021]
- And students who spend nights and days trying to reproduce results and realizing cold, hard truths.
 - ML Reproducibility Challenge

Shivalika Singh^{*1}, Yiyang Nan¹, Alex Wang², Daniel D'souza¹,
Sayash Kapoor³, Ahmet Üstün¹, Sanmi Koyejo⁴, Yuntian Deng⁵,
Shayne Longpre⁶, Noah A. Smith^{7,8}, Beyza Ermis¹,
Marzieh Fadaee¹, and Sara Hooker¹

¹Cohere Labs, ²Cohere, ³Princeton University, ⁴Stanford University, ⁵University of Waterloo,
⁶Massachusetts Institute of Technology, ⁷Allen Institute for Artificial Intelligence, ⁸University of
Washington

Corresponding authors: {shivalikasingh, marzieh, sarahooker}@cohere.com

Abstract

Measuring progress is fundamental to the advancement of any scientific field. As benchmarks play an increasingly central role, they also grow more susceptible to distortion. Chatbot Arena has emerged as the go-to leaderboard for ranking the most capable AI systems. Yet, in this work we identify systematic issues that have resulted in a distorted playing field. We find that **undisclosed private testing practices benefit a handful of providers who are able to test multiple variants before public release and retract scores if desired**. We establish that the ability of these providers to choose the best score leads to biased Arena scores due to selective disclosure of performance results. At an extreme, we identify 27 private LLM variants tested by Meta in the lead-up to the Llama-4 release. We also establish that **proprietary closed models are sampled at higher rates (number of battles) and have fewer models removed from the arena** than open-weight and open-source alternatives. Both these policies lead to large data access asymmetries over time. Providers like Google and OpenAI have received an estimated 19.2% and 20.4% of all data on the arena, respectively. In contrast, a combined 83 open-weight models have only received an estimated 29.7% of the total data. With conservative estimates, we show that access to Chatbot Arena data yields substantial benefits; even limited additional data can result in relative performance gains of up to 112% on ArenaHard, a test set from the arena distribution. Together, **these dynamics result in overfitting to Arena-specific dynamics rather than general model quality**. The Arena builds on the substantial efforts of both the organizers and an open community that maintains this valuable evaluation platform. We offer actionable recommendations to reform the Chatbot Arena's evaluation framework and promote fairer, more transparent benchmarking for the field.