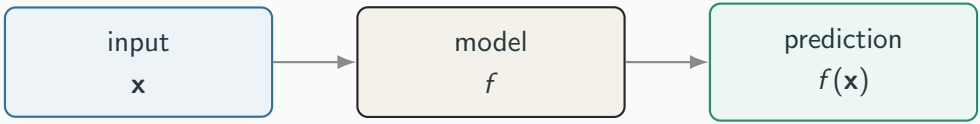
Generalization and Evaluation

What does it mean for a model to "work"?

CSCI 567 — Machine Learning • Lecture 01 • Aug 28, 2026

Setup: input $\rightarrow$ model $\rightarrow$ prediction



Linear regression, SVMs, neural networks, transformers — every method this semester is an instance of this diagram. Before we learn how to build f, we need to agree on what makes f *good*.

This lecture: defining “does it work”

“Does this model work?” sounds like a yes/no question. It is not, and this lecture — which trains **zero** models — builds the vocabulary every later lecture leans on when it claims a method “generalizes” or a benchmark “measures” something.

The field’s answer to “does it work” has gotten **more demanding** over time — four times:

a clean statistical assumption $\longrightarrow$ a check against distribution shift $\longrightarrow$ a check against memorizing the test $\longrightarrow$ a check on what the number even means

Outline

1. The classical answer: IID and the generalization gap
2. Complication 1: even a careful IID benchmark has hidden gaps
3. Complication 2: out-of-distribution and external validity
4. Complication 3: the LLM-benchmark era — contamination
5. Complication 4: benchmarks as proxies — construct validity
6. Synthesis: two different questions, not one list
7. Where this course goes next

1. The classical answer

IID and the generalization gap

The formal setup: a distribution and a loss

a fixed, unknown distribution $\mathcal{D}$ over input-output pairs $(\mathbf{x}, \mathbf{y})$ — every example we ever see (training, testing, deployment) is a draw from $\mathcal{D}$; we never see $\mathcal{D}$ itself, only samples

A model f is scored by a **loss function** $\ell(f(\mathbf{x}), \mathbf{y})$:

$\ell(f(\mathbf{x}), \mathbf{y})$ **small**

$f(\mathbf{x})$ is a good prediction for $\mathbf{y}$

$\ell(f(\mathbf{x}), \mathbf{y})$ **large**

$f(\mathbf{x})$ is a bad prediction for $\mathbf{y}$

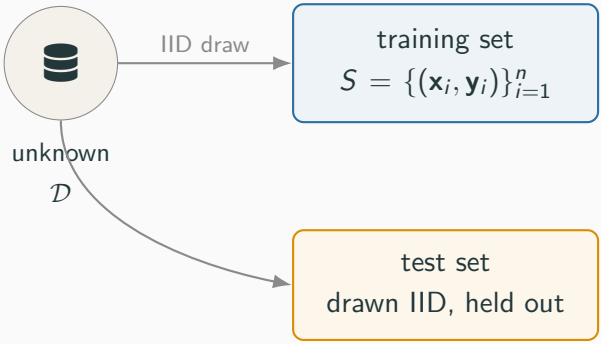
True risk: the quantity we actually care about

$$\mathcal{R}(f) \;=\; \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim\mathcal{D}}\left[\ell(f(\mathbf{x}),\mathbf{y})\right]$$

the average loss f would incur over the *entire population* — including examples we will never see.

✗ we cannot compute $\mathcal{R}(f)$ directly — we do not have access to $\mathcal{D}$, only to a finite sample

Training and test sets: IID draws from $\mathcal{D}$



IID: independently and identically distributed — every draw is independent, and every draw comes from the same $\mathcal{D}$.

The generalization gap

$$\text{generalization gap} \;=\; \mathcal{R}(f) \;-\; \underbrace{\frac{1}{n}\sum_{i=1}^{n} \ell(f(\mathbf{x}_i), \mathbf{y}_i)}_{\text{training risk}}$$

a model that fits training data perfectly but has a **large** generalization gap has not learned anything useful — it has **memorized** S

Held-out test error: an unbiased, noisy estimate of $\mathcal{R}(f)$

The point of a held-out test set is to catch a large generalization gap before deployment.

✅ if the test set is truly IID with training data, and no modeling decision was made by looking at it, test error is an **unbiased estimate** of $\mathcal{R}(f)$

But unbiased is not the same as precise: an estimate can be correct *on average* and still swing wildly on any single draw. What makes a large IID test set *trustworthy* is a **concentration bound**.

Hoeffding's inequality

$Z_1, \ldots, Z_n$ independent, $Z_i \in [0,1]$, $\mathbb{E}[Z_i] = \mu$, $\bar{Z} = \frac{1}{n}\sum_i Z_i$.

$$\mathbb{P}\big[\,|\bar{Z} - \mu| \geq \epsilon\,\big] \;\leq\; 2\exp(-2n\epsilon^2)$$

$Z_i = \ell(f(\mathbf{x}_i), \mathbf{y}_i)$, $\mu = \mathcal{R}(f)$: test error concentrates around true risk — **provided f was fixed before the test set was drawn**.

Proof, step 1: Markov's inequality on the moment generating function

Target: the one-sided tail $\mathbb{P}[\bar{Z} - \mu \geq \epsilon]$ — the other tail and the factor of 2 come in step 3. **Markov's inequality** ($V \geq 0$, $c > 0$: $\mathbb{P}[V \geq c] \leq \mathbb{E}[V]/c$) is the only probabilistic input.

The trick: apply it to $e^{t(\bar{Z}-\mu)}$, not to $\bar{Z} - \mu$ itself, with $t > 0$ a free parameter. Since $x \mapsto e^{tx}$ is strictly increasing, $\{\bar{Z} - \mu \geq \epsilon\}$ and $\{e^{t(\bar{Z}-\mu)} \geq e^{t\epsilon}\}$ are the *same event*:

$$\mathbb{P}[\bar{Z} - \mu \geq \epsilon] = \mathbb{P}\Big[e^{t(\bar{Z}-\mu)} \geq e^{t\epsilon}\Big] \;\leq\; e^{-t\epsilon}\,\mathbb{E}\Big[e^{t(\bar{Z}-\mu)}\Big]$$

Independence enters exactly once: $e^{t(\bar{Z}-\mu)} = \prod_i e^{t(Z_i-\mu)/n}$, and for *independent* factors the expectation of the product is the product of the expectations:

$$\mathbb{P}[\bar{Z} - \mu \geq \epsilon] \;\leq\; e^{-t\epsilon} \prod_{i=1}^{n} \mathbb{E}\Big[e^{t(Z_i-\mu)/n}\Big] \qquad \text{(holds for } \textit{every } t > 0 \text{ at once)}$$

Proof, step 2: Hoeffding's lemma bounds each factor

Hoeffding's lemma: W mean-zero, supported on $[a, b] \implies \mathbb{E}[e^{sW}] \leq \exp\big(s^2(b-a)^2/8\big)$ for all s

Why the lemma holds, move 1 — convexity: on $[a, b]$, e^{sw} lies below its chord, so $e^{sW} \leq \frac{b-W}{b-a} e^{sa} + \frac{W-a}{b-a} e^{sb}$; taking expectations and using $\mathbb{E}[W] = 0$ leaves a bound depending only on a, b, s.

Move 2 — calculus: the log of that bound, as a function ψ of $s(b-a)$, has $\psi(0) = \psi'(0) = 0$ and $\psi'' \leq \frac{1}{4}$ everywhere (a variance term, maximized by a fair coin), so Taylor gives $\psi \leq s^2(b-a)^2/8$.

Apply it to each factor: $W = Z_i - \mu$ is mean-zero on the width-1 interval $[-\mu, 1-\mu]$ $(b - a = 1)$, with $s = t/n$, so each factor is at most $e^{t^2/(8n^2)}$; multiplying n of them into step 1:

$$\mathbb{P}[\bar{Z} - \mu \geq \epsilon] \leq e^{-t\epsilon} \cdot \left(e^{t^2/(8n^2)}\right)^n = \exp\left(-t\epsilon + \frac{t^2}{8n}\right)$$

Proof, step 3: optimize t, then union-bound the two tails

Step 2's bound holds for every $t > 0$, so use the best one: minimize $g(t) = -t\epsilon + \frac{t^2}{8n}$.

$$g'(t) = -\epsilon + \frac{t}{4n} = 0 \implies t^\star = 4n\epsilon > 0 \qquad \left(g'' = \tfrac{1}{4n} > 0\text{: a minimum}\right)$$

$$g(t^\star) = -4n\epsilon^2 + \frac{(4n\epsilon)^2}{8n} = -4n\epsilon^2 + 2n\epsilon^2 = -2n\epsilon^2 \quad \implies \quad \mathbb{P}[\bar{Z} - \mu \geq \epsilon] \leq \exp(-2n\epsilon^2)$$

The other tail: run the identical argument on $Z_i' = 1 - Z_i$ (same width-1 support, mean $1 - \mu$) to get $\mathbb{P}[\mu - \bar{Z} \geq \epsilon] \leq \exp(-2n\epsilon^2)$; a union bound over the two disjoint bad events gives the factor of 2:

$$\mathbb{P}\left[\,|\bar{Z} - \mu| \geq \epsilon\,\right] \;\leq\; 2\exp(-2n\epsilon^2) \qquad \blacksquare$$

The recipe — Markov on e^{tX}, factor by independence, bound each moment generating function, optimize t — is the generic **Chernoff method** behind essentially every exponential tail bound in this field.

A concrete number: $n = 10{,}000$

A test set of $n = 10{,}000$, solved at 95% confidence:

$$\epsilon = \sqrt{\frac{\ln(2/0.05)}{2 \cdot 10{,}000}} \approx \sqrt{1.84 \times 10^{-4}} \approx 1.4\%$$

Sampling noise alone moves test accuracy by at most ± 1.4 points.

everything above leans on one clause — *"IID, and untouched by modeling decisions"* — and the rest of this lecture is what happens when it quietly breaks

2. Complication 1

Even a careful IID benchmark has hidden gaps

Recht et al. (2019): build a second test set and check

A test set built with **exactly** the same collection methodology as the training data satisfies IID on paper — does that settle the question?

> 🔍 *Do ImageNet Classifiers Generalize to ImageNet?*
> — Recht, Roelofs, Schmidt, Shankar, ICML 2019

original CIFAR-10 / ImageNet test sets

a decade of use, thousands of papers

new test set

same sourcing pipeline, same labeling protocol, same class balance

Then re-evaluate a broad range of **already-published** models on the new test set.

How the new test sets were built

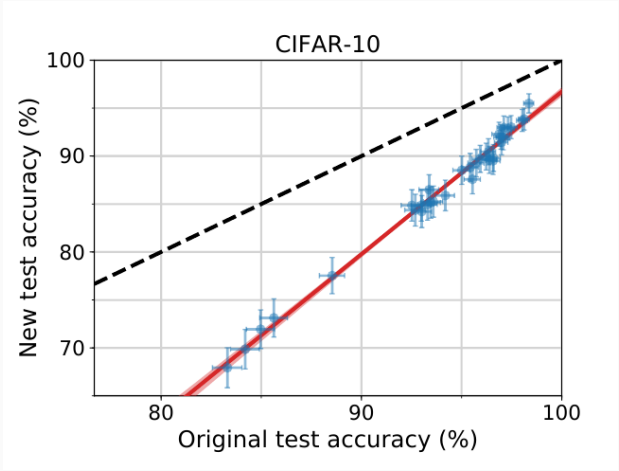
1. Gather from the original source: CIFAR-10 candidates from the larger Tiny Images repository, matching each class's keyword composition; ImageNet candidates from new Flickr queries matching each class's original query structure and upload window.

2. Annotate with the original pipeline: CIFAR-10 labeled by two graduate-student authors under the original instructions; ImageNet via rebuilt MTurk tasks (48-image grids, class definition + Wikipedia link, same selection rules, $\geq$10 workers/image). Each image gets a **selection frequency** — the fraction of workers selecting it; $\geq$6 embedded original-validation images per task calibrate the scale.

3. Sample the final sets: 2,000 / 10,000 images ($\approx \pm 1\%$ confidence intervals), small next to the source pools ($\approx$3% / $<$1%) to limit "harder leftovers" bias; the headline ImageNet set (*MatchedFrequency*) matches the original validation set's per-class selection-frequency distribution.

design choices move the answer: sampling by top selection frequency instead (avg. 0.93 vs. 0.73) swings accuracy by $\sim$14 points — and still: *"it is astonishingly hard to replicate the test set distributions of CIFAR-10 and ImageNet"*

Result: every model's accuracy dropped



Recht et al. (2019), Figure 1 (CIFAR-10 panel). Dashed: $y = x$. Every point sits below it.

3–15
percentage points
CIFAR-10

11–14
percentage points
ImageNet

The drop follows a precise linear fit

$$\begin{aligned} \text{acc}_{\text{new}} &= 1.69 \cdot \text{acc}_{\text{orig}} - 72.7\% \quad \text{(CIFAR-10)} \\ \text{acc}_{\text{new}} &= 1.11 \cdot \text{acc}_{\text{orig}} - 20.2\% \quad \text{(ImageNet)} \end{aligned}$$

slope 95% CI: $[1.63, 1.76]$ / $[1.07, 1.19]$ (100,000 bootstrap resamples) — reliably > 1.

Higher original accuracy $\Rightarrow$ *smaller* drop.

Ruling out sampling noise: a p-value from the Hoeffding bound

Null hypothesis H_0: both test sets are IID draws from the same $\mathcal{D}$, and every model was fixed before either was drawn — then the drop is pure sampling noise. **Test it with the bound we just proved**: the same Chernoff recipe over all $n_1 + n_2$ independent terms gives a two-sample bound; plug in ImageNet's $n_1 = 50{,}000$ (original validation), $n_2 = 10{,}000$ (new), smallest drop $\epsilon = 0.11$:

$$\mathbb{P}\left[\,|\mathrm{acc}_{\mathrm{orig}} - \mathrm{acc}_{\mathrm{new}}| \geq \epsilon\,\right] \;\leq\; 2\exp\!\left(-2\epsilon^2/(1/n_1 + 1/n_2)\right) \quad \text{under } H_0$$

$$p \;\leq\; 2\exp\!\left(-2\,(0.11)^2/(1/50{,}000 + 1/10{,}000)\right) \;=\; 2\exp(-202) \;\approx\; 5 \times 10^{-88}$$

reject H_0 at any significance level anyone has ever used — what explains the drop is adaptivity or a genuine distribution change, not noise

Hoeffding is loose (it assumes worst-case variance): on CIFAR-10's *smallest* drops ($\epsilon = 0.03$, $n_1 = 10{,}000$, $n_2 = 2{,}000$) it gives only $p \leq 0.10$; the paper's exact Clopper-Pearson intervals ($\approx \pm 1\%$) still rule noise out.

Candidate explanation: adaptive overfitting

> 💡 **adaptive overfitting**: a decade of researchers tuning architectures and hyperparameters against the same fixed test set, implicitly fitting it — the same failure mode as touching the test set during model selection

Sounds right. Let's test it rather than just argue about it.

The mechanism, made precise: p-hacking

m groups each test a *false* hypothesis against the same fixed data, at significance level α:

$$\mathbb{P}[\text{at least one false positive}] = 1 - (1-\alpha)^m \qquad m = 20,\ \alpha = 0.05 : \approx \mathbf{64.2\%}$$

a decade of researchers iterating configurations against
one fixed test set S, publishing only what scores well,
is this mechanism — no individual needs to cheat

If it dominates, it predicts a checkable fingerprint: later, more-selected models should show a *growing* gap between test accuracy and true performance.

Testing the fingerprint: no diminishing returns

If adaptive overfitting drove the drop, later, more-tuned models should show **diminishing returns**: population accuracy plateaus while test accuracy keeps climbing.

Observed instead: later models see *smaller* drops (slope > 1). No diminishing returns anywhere in a decade of models — the predicted fingerprint is absent.

Evidence against *substantial* adaptive overfitting here — not proof it never happens elsewhere.

Decomposing the drop: three terms, two ruled out

$$\underbrace{L_S - L_{S'}}_{\text{observed drop}} = \underbrace{(L_S - L_{\mathcal{D}})}_{\text{adaptivity gap}} + \underbrace{(L_{\mathcal{D}} - L_{\mathcal{D}'})}_{\text{distribution gap}} + \underbrace{(L_{\mathcal{D}'} - L_{S'})}_{\text{generalization gap}}$$

- **adaptivity gap**: ruled out — the predicted fingerprint is absent (previous slide).
- **generalization gap**: already ruled out — the two-sample Hoeffding test rejects pure sampling noise at $p \approx 5 \times 10^{-88}$.

the **distribution gap** is what's left, by elimination

The distribution gap: genuinely harder images

two test sets built to the same specification, both nominally IID with the original training distribution — and the new one contained **genuinely slightly harder images**

"harder," measured: images MTurk crowdworkers agreed on *less often* (lower **selection frequency**) are exactly the ones models get wrong more often

Honest limit: selection frequency is a measurable *proxy* for difficulty, not an explanation — candidate causes (object size, filters, pose) remain unresolved.

Human control: not harder for human vision

$\sim$5% hardest CIFAR-10 images, nine human graduate-student annotators:

-0.8%

average human accuracy difference, original vs. new

Statistically indistinguishable from zero — **harder for the models, not for human vision**.

IID is a modeling assumption we choose to make, not a property we can verify by inspection.

Getting the collection process to match is **necessary**, not **sufficient**. Localizing a real gap took a formal decomposition, a decade of models, and a human baseline.

3. Complication 2

Out-of-distribution and external validity

Out-of-distribution data: deployment $\neq$ training

Once a model leaves the benchmark and reaches the real world, the IID assumption gets even more strained:

spam classifier trained on 2020 email

meets 2027 spam tactics

model trained on one hospital's scanner

meets a *different* scanner, different hospital

neither is IID with training: this is **out-of-distribution (OOD)** data — drawn from $\mathcal{D}' \neq \mathcal{D}$

External validity: the same question in research methodology

external validity: whether findings from a specific study sample hold for the target population, and for other outcomes/contexts beyond the one studied

ML's OOD generalization literature is answering the *same* question under a different name: does strong performance on this benchmark's distribution predict strong performance on the distribution you actually care about?

The failure mode, and the stricter bar it forces

even a *mild* input-distribution shift at deployment can make a model far worse — a **near-zero IID generalization gap** can coexist with being **useless in production**

IID story
(Bar 1)

+

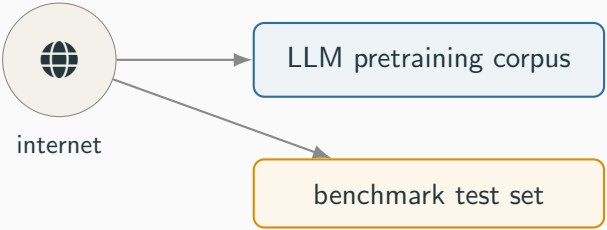
evidence the benchmark distribution resembles deployment

Passing OOD generalization requires *both*.

4. Complication 3

The LLM-benchmark era: contamination

Data contamination



if a benchmark's test examples — or close paraphrases — end up inside the pretraining corpus, the model can score well by **recalling** the answer rather than generalizing to it

Not a distribution-shift problem — a **data-leakage** problem, and it inflates a number that otherwise looks like a clean, IID-style evaluation.

Contamination vs. OOD: not the same failure

	OOD	Contamination
mechanism	distribution shift	data leakage
direction of harm	**under**states ability	**over**states ability
when it bites	deployment	benchmark time

Measuring it, take 1: the GPT-3 overlap audit

Brown et al. (2020), Appendix C — flag a benchmark example as potentially leaked if it shares a **13-gram** with anything in the pretraining set; compare the confirmed-clean subset vs. the full benchmark.

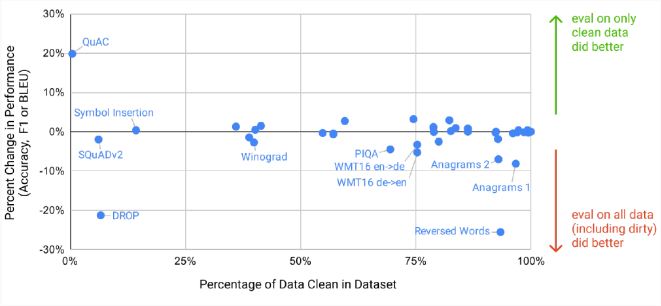


Figure 4.2. x-axis: % of a benchmark confirmed clean. y-axis: performance change, clean-only vs. full.

Reading the GPT-3 result: cluster, false alarms, real hits

most benchmarks: cluster near 0% — clean and full evaluation agree

QuAC (+20%), DROP (−22%): large swings, investigated and found to be *false alarms* — overlap was background passages, not the question/answer pairs tested

PIQA, Winograd: real contamination (132 Winograd schemas confirmed in training data) — marked with an asterisk in the paper, not waved away

Self-reported caveat: a filtering-pipeline bug undercounted some true overlaps; retraining to fix it wasn't feasible given GPT-3's training cost.

Measuring it, take 2: Sainz et al. (2023)

NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for Each Benchmark — Findings of EMNLP 2023

Push back precisely on generalizing Brown et al.'s finding:

- Field-wide contamination is largely **unmeasured**, not shown to be small.
- Most published LLM results report *no* contamination analysis at all.
- Calls for benchmark-specific contamination measurement as standard practice.

Measuring it, take 3: a recipe you can run yourself

Golchin & Surdeanu (2024), *Time Travel in LLMs* — "guided instruction"

1. Prompt the model with the benchmark's name, its partition (e.g. "test split"), and a random-length initial fragment of a real instance.
2. Ask it to complete the rest.
3. If the completion matches the true continuation *closely* — more closely than chance, or than an uncontaminated model would manage — flag that instance.

Aggregating instance-level flags gives a partition-level contamination estimate.

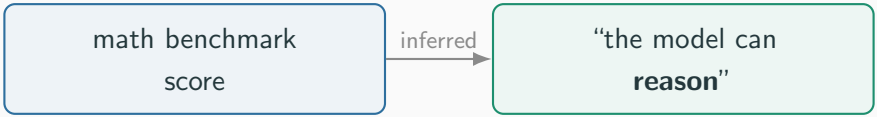
In the LLM era, a high benchmark score answers a different question than it used to.

"Did this generalize" now requires ruling out "did this just memorize the benchmark" — a check the classical IID story has **no mechanism** to catch.

5. Complication 4

Benchmarks as proxies: construct validity

Benchmark scores are used as proxies for capabilities



A high coding-benchmark score is read as evidence the model can propose **genuinely novel** solutions — not just solve benchmark-shaped problems it has practiced.

Bars 1–3 all ask: is this **number** trustworthy? Here the number can be entirely trustworthy — the question is whether it means what it is being used to mean

Construct validity (Cronbach & Meehl, 1955)

construct validity: does a measurement actually reflect the theoretical construct it is assumed to measure — not just the literal, narrow task performance it directly records?

content	does the test cover the right material?
criterion	does the score predict an external outcome?
construct	does it reflect the intended latent trait?

Construct validity is the hardest of the three: the construct itself — "reasoning" — is not directly observable.

A latent-factor model of benchmark scores

Let $\theta \in \mathbb{R}^d$ be a model's (unobserved) vector of latent cognitive constructs; benchmark i's expected score:

$$\mathbb{E}[X_i] \;=\; \mu_i \;+\; \Lambda_i \theta, \qquad \Lambda_i \in \mathbb{R}^{1\times d} \text{ the } \textbf{loading vector}$$

benchmark i has good construct validity for target construct j if Λ_i **concentrates** on coordinate j: Λ_{ij} large, $\Lambda_{ik} \approx 0$ for $k \neq j$

Bars 1–3 say nothing about Λ_i: a benchmark can be IID-valid, deployment-representative, *and* uncontaminated, while measuring something other than what it is credited with.

$d = 1$ makes a falsifiable prediction

"High performance here means high general reasoning" implicitly assumes $d = 1$.

one scalar θ explains performance across domains $\implies$ any two such benchmarks i, j satisfy $\mathrm{Corr}(X_i, X_j) \to 1$ up to noise

Not a rhetorical claim — a precise, falsifiable prediction about a correlation matrix.

Salaudeen et al. (2025): validity-centered evaluation

> *"Grand claims, such as models achieving general reasoning capabilities, are supported with model performance on narrow benchmarks . . . which provide a limited and potentially misleading assessment."*

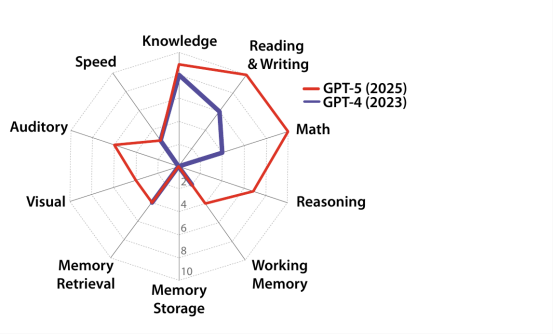
Their fix is not a new benchmark — it is a discipline: ask explicitly whether the evidence supports the claim being made.

Testing $d = 1$ directly: Hendrycks et al. (2025)

A Definition of AGI — 33 authors, incl. Bengio, Salaudeen. Adapts a human psychometric battery (Cattell-Horn-Carroll theory) across **ten** cognitive domains, weighted **equally** (10% each) — breadth over any one specialty

Knowledge	Reading/Writing	Mathematics
Reasoning	Working Memory	Memory Storage
Memory Retrieval	Visual	Auditory
	Speed	

The result: a jagged capability profile



Hendrycks et al. (2025), Figure 1. A single shared factor would inflate the polygon evenly. This is not that.

27%

AGI-proficiency, GPT-4

57%

AGI-proficiency, GPT-5

jumps driven by *uneven* gains: sharp peaks (Math, Reading/Writing, Knowledge), **Memory Storage near 0%** — the single most severe bottleneck

Jaggedness rejects the one-factor story

if $d = 1$

ten domain scores tightly correlated, one dominant direction

observed

jagged: strong here, weak there, no single dominant direction

Either the true $d > 1$, or Λ_i doesn't concentrate on the construct it's credited with. Either way: one narrow benchmark score cannot license a one-dimensional capability claim, because the model's own measured structure is not one-dimensional.

Beyond item-level leakage: class-level exposure

Bar 3 asked: did the model see this **exact** test item?

> the harder question: did exposure to a broad *class* of similar problems — not the literal items — already produce a high score via pattern-matching, with no general capability behind it at all?

Unlike 13-gram overlap, there is no clean test for this.

A benchmark number can be perfectly trustworthy and still not mean what it is being used to mean.

Unlike Bars 1–3, no amount of care about sampling, deployment matching, or decontamination fixes a construct-validity problem — it is a question about what the number **means**, not whether it is **accurate**.

6. Synthesis

Two different questions, not one list

The four complications separate into two kinds of question

is the number trustworthy? — three progressively harder bars, about measurement integrity

does the number mean what you think it means? — a fourth question, about inference

The four bars, side by side

	bar	the question it asks
1	IID generalization	does test performance reflect $\mathcal{R}(f)$, with nothing leaking from the test set into modeling decisions?
2	OOD / external validity	does performance on the benchmark's distribution predict performance on the distribution you deploy to?
3	freedom from contamination	is the score the model's ability, or partly what it memorized during pretraining?
4	construct validity	does this specific, narrow score license the broader capability claim being made?

Bars 1–3: is the number **trustworthy**. Bar 4: is it **meaningful**. Lecture 03 builds the formal tools for Bar 1 (bias-variance, overfitting).

Four independent checks

A model can pass any subset of these four while failing the rest.

Every method we build this semester will be judged against them — worth being explicit *before* there is a model on the board to judge.

7. Roadmap

Where this course goes next

Course roadmap

L02–L06: classical supervised & unsupervised learning — optimization, linear/logistic regression, perceptron, regularization, kernels, double descent, SVM, PCA, clustering

Exam 1, then **L07–L10**: the deep-learning & LLM arc — neural networks, ensembling, self-supervised learning, transformers, scaling laws

L11–L12: reinforcement learning; safety & interpretability — this *same* evaluation question, asked once more, about whether we can **trust** a model at all

Next time: Lecture 02

optimization algorithms (GD, SGD) + the three foundational models they fit: linear regression, logistic regression, the perceptron

Pre-recorded video, fully self-contained; discussion section meets in person as usual.

References I

- Recht, Roelofs, Schmidt, Shankar. *Do ImageNet Classifiers Generalize to ImageNet?* ICML 2019.
- Brown et al. *Language Models are Few-Shot Learners.* NeurIPS 2020. (GPT-3; see Appendix C on data contamination.)
- Sainz, Campos, García-Ferrero, Etxaniz, Lopez de Lacalle, Agirre. *NLP Evaluation in Trouble: On the Need to Measure LLM Data Contamination for Each Benchmark.* Findings of EMNLP 2023.
- Golchin, Surdeanu. *Time Travel in LLMs: Tracing Data Contamination in Large Language Models.* ICLR 2024.

References II

- Cronbach, Meehl. *Construct Validity in Psychological Tests.* Psychological Bulletin, 52(4), 281–302, 1955.
- Salaudeen et al. *Measurement to Meaning: A Validity-Centered Framework for AI Evaluation.* arXiv 2505.10573, 2025.
- Hendrycks et al. (33 authors, incl. Bengio, Salaudeen). *A Definition of AGI.* arXiv 2510.18212, 2025.

Full citation details:

`references/sources/lecture01-generalization-evaluation-citations.md`

the question this course keeps asking

does it generalize — IID, out-of-distribution, free of contamination, and measuring the right construct?

Questions?