CS 489: Introduction to Natural Language Processing

Lecture 14: Evaluation and Benchmarking

Instructor: Freda Shi

November 10, 2026



UNIVERSITY OF WATERLOO

Today: Three Questions

1. **Target:** What behavior and population do we want to measure?
2. **Scoring:** How do we score probabilities, answers, and generated text?
3. **Evidence:** What makes a comparison trustworthy?

What Does “Good” Mean?

Goal: measure language model **behavior on a specific task**.

Claim	Evidence
Predicts held-out text well	Probability assigned to that text
Answers test questions correctly	Correct answers on sampled questions
Preserves a reference’s content	Matches to reference answers
Produces useful responses	Judgments under a quality rubric

What Does “Good” Mean?

Goal: measure language model **behavior on a specific task**.

Claim	Evidence
Predicts held-out text well	Probability assigned to that text
Answers test questions correctly	Correct answers on sampled questions
Preserves a reference’s content	Matches to reference answers
Produces useful responses	Judgments under a quality rubric

“Good” is a relative term. Define "good" before any evaluation.

A Benchmark Is More Than a Dataset

Suppose we want to compare systems for **scientific question answering**.

1. **Target and data:** what questions and which subjects?
2. **System and protocol:** which model, prompts, tools, and computation budget?
3. **Scoring:** how are answers extracted, graded, and aggregated?

MMLU: Knowledge Across 57 Subjects

Hendrycks et al. (2021): four-choice questions across STEM, humanities, social sciences, and professional domains. A common protocol uses 5-shot prompting; random choice gives 25% accuracy.

Illustrative physics question:

A $10\,\text{kg}$ mass slides on a frictionless incline at 30°. What is its acceleration down the incline, using $g = 9.8\,\text{m/s}^2$?

(A) 2.5 (B) 4.9 (C) 7.5 (D) 9.8 m/s^2

MMLU: Knowledge Across 57 Subjects

Hendrycks et al. (2021): four-choice questions across STEM, humanities, social sciences, and professional domains. A common protocol uses 5-shot prompting; random choice gives 25% accuracy.

Illustrative physics question:

A 10 kg mass slides on a frictionless incline at 30°. What is its acceleration down the incline, using $g = 9.8\,\mathrm{m/s^2}$?

(A) 2.5 (B) 4.9 (C) 7.5 (D) 9.8 $\mathrm{m/s^2}$

$$ma = mg\sin 30^\circ \quad \Longrightarrow \quad a = 9.8 \times \frac{1}{2} = 4.9\,\mathrm{m/s^2}.$$

A correct choice alone does not reveal whether the model used this reasoning. Report subject-level scores alongside the aggregate.

Metric: Perplexity

Score the probability assigned to held-out text.
For N scored tokens, the average log-loss in bits is

$$H = -\frac{1}{N}\sum_{i=1}^{N} \log_2 P(w_i \mid w_{<i}), \qquad \text{PPL} = 2^H.$$

Metric: Perplexity

Score the probability assigned to held-out text.
For N scored tokens, the average log-loss in bits is

$$H = -\frac{1}{N}\sum_{i=1}^{N} \log_2 P(w_i \mid w_{<i}), \qquad \text{PPL} = 2^H.$$

- Lower PPL means higher probability assigned to the observed text.
- Uniform prediction over V tokens gives PPL $= V$.
  PPL is an effective number of choices.
- Useful for monitoring prediction during training; task accuracy and usefulness need their own measurements.

Perplexity Depends on the Counting Unit

Toy example: two lossless tokenizations of the same 12-character text. Suppose each model assigns its token sequence a total negative log-probability of 12 bits.

	Scored tokens	**Bits/token**	**PPL**
Model A	6	$12/6 = 2$	$2^2 = 4$
Model B	3	$12/3 = 4$	$2^4 = 16$

Perplexity Depends on the Counting Unit

Toy example: two lossless tokenizations of the same 12-character text. Suppose each model assigns its token sequence a total negative log-probability of 12 bits.

	Scored tokens	Bits/token	PPL
Model A	6	$12/6 = 2$	$2^2 = 4$
Model B	3	$12/3 = 4$	$2^4 = 16$

$$\mathrm{BPC} = \frac{N}{C} \log_2 \mathrm{PPL} \qquad \Longrightarrow \qquad \mathrm{BPC}_A = \mathrm{BPC}_B = \frac{12}{12} = 1.$$

C counts characters in the original text; bits per byte uses bytes instead.

Compare on identical text with consistent preprocessing and context limits. Vocabulary size alone does not determine per-token perplexity.

Closed-Answer Metrics

When the output is an answer or a set of extracted items, score its correctness.

- **Accuracy:** fraction of examples assigned the correct label.
- **Exact match (EM):** the normalized output matches an accepted answer.
- **Precision and recall:** count correct predictions or extracted items.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad F_1 = \frac{2PR}{P + R}.$$

Closed-Answer Metrics

When the output is an answer or a set of extracted items, score its correctness.

- **Accuracy:** fraction of examples assigned the correct label.
- **Exact match (EM):** the normalized output matches an accepted answer.
- **Precision and recall:** count correct predictions or extracted items.

$$P = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FP}}, \qquad R = \frac{\mathrm{TP}}{\mathrm{TP} + \mathrm{FN}}, \qquad F_1 = \frac{2PR}{P + R}.$$

Example: extract 4 entities, of which 3 match the 5 gold entities.

$$\mathrm{TP} = 3, \quad \mathrm{FP} = 1, \quad \mathrm{FN} = 2 \quad \Longrightarrow \quad P = \frac{3}{4},\ R = \frac{3}{5},\ F_1 = \frac{2}{3}.$$

A Discrete Metric Can Create a Jump

The emergent ability hypothesis (Wei et al., 2022): without changing model architecture, a model's performance on a task can improve sharply at a certain scale.

A Discrete Metric Can Create a Jump

The emergent ability hypothesis (Wei et al., 2022): without changing model architecture, a model's performance on a task can improve sharply at a certain scale.

Toy binary task: A is correct; prediction uses $\arg\max$.

Model size	1	2	3	4
P(A)	0.40	0.49	0.51	0.60
P(B)	0.60	0.51	0.49	0.40
Accuracy on this item	0	0	1	1

A Discrete Metric Can Create a Jump

The emergent ability hypothesis (Wei et al., 2022): without changing model architecture, a model's performance on a task can improve sharply at a certain scale.

Toy binary task: A is correct; prediction uses arg max.

Model size	1	2	3	4
P(A)	0.40	0.49	0.51	0.60
P(B)	0.60	0.51	0.49	0.40
Accuracy on this item	0	0	1	1

Accuracy changes when the top-ranked answer changes. Log-loss records improvement on both sides of that boundary.

Check the metric before interpreting a scaling curve as an emergent ability.

Can we use confidence to decide when to trust an answer?

Confidence: the probability a model assigns to its chosen answer.

Can we use confidence to decide when to trust an answer?

Confidence: the probability a model assigns to its chosen answer.

Consider many multiple-choice questions where that probability is **90%**.

Can we use confidence to decide when to trust an answer?

Confidence: the probability a model assigns to its chosen answer.

Consider many multiple-choice questions where that probability is **90%**.

Actual accuracy in this group	Interpretation
About 90%	Confidence matches accuracy
60%	Overconfident
99%	Underconfident

Can we use confidence to decide when to trust an answer?

Confidence: the probability a model assigns to its chosen answer.

Consider many multiple-choice questions where that probability is **90%**.

Actual accuracy in this group	Interpretation
About 90%	Confidence matches accuracy
60%	Overconfident
99%	Underconfident

Calibration asks whether this match holds at every confidence level p:

$$P(\text{correct} \mid \text{confidence} = p) = p.$$

Expected Calibration Error: Example

Group N predictions into confidence bins S_b. Each bin has empirical accuracy $\text{acc}(S_b)$ and mean confidence $\text{conf}(S_b)$.

$$\text{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left|\text{acc}(S_b) - \text{conf}(S_b)\right|.$$

Bin	$\lvert S_b \rvert$	**Mean confidence**	**Correct**	**Accuracy**
Lower confidence	4	0.60	3	$3/4 = 0.75$
Higher confidence	6	0.90	3	$3/6 = 0.50$

Expected Calibration Error: Example

Group N predictions into confidence bins S_b. Each bin has empirical accuracy $\mathrm{acc}(S_b)$ and mean confidence $\mathrm{conf}(S_b)$.

$$\mathrm{ECE} = \sum_{b=1}^{B} \frac{|S_b|}{N} \left|\mathrm{acc}(S_b) - \mathrm{conf}(S_b)\right|.$$

Bin	$\lvert S_b \rvert$	Mean confidence	Correct	Accuracy
Lower confidence	4	0.60	3	$3/4 = 0.75$
Higher confidence	6	0.90	3	$3/6 = 0.50$

$$\mathrm{ECE} = \frac{4}{10}|0.75 - 0.60| + \frac{6}{10}|0.50 - 0.90| = 0.30.$$

The higher-confidence bin is overconfident. ECE depends on binning and sample size; report both with the score.

Robustness: Adding an Irrelevant Detail

Oliver picks 44 kiwis on Friday and 58 on Saturday. On Sunday, he picks twice as many as on Friday. How many kiwis does he have in total?

Robustness: Adding an Irrelevant Detail

Oliver picks 44 kiwis on Friday and 58 on Saturday. On Sunday, he picks twice as many as on Friday. How many kiwis does he have in total?

$$44 + 58 + 2 \times 44 = 190.$$

Robustness: Adding an Irrelevant Detail

Oliver picks 44 kiwis on Friday and 58 on Saturday. On Sunday, he picks twice as many as on Friday. How many kiwis does he have in total?

$$44 + 58 + 2 \times 44 = 190.$$

Added detail: Five of Sunday's kiwis are smaller than average.

Robustness: Adding an Irrelevant Detail

Oliver picks 44 kiwis on Friday and 58 on Saturday. On Sunday, he picks twice as many as on Friday. How many kiwis does he have in total?

$$44 + 58 + 2 \times 44 = 190.$$

Added detail: Five of Sunday's kiwis are smaller than average.

The correct total stays **190**. Reported o1-mini error: subtract the smaller kiwis, giving $190 - 5 = 185$.

A robustness test checks whether irrelevant information changes the answer.

HELM: Evaluate More Than Accuracy

Liang et al. (2022): the original HELM organizes evaluation across 42 scenarios and seven metric dimensions.

Dimension	Question
Accuracy	How often is the answer correct?
Calibration	Does confidence match correctness?
Robustness	Does performance survive input perturbations?
Fairness	How does performance vary across groups?
Bias	What stereotypes or associations appear?
Toxicity	How often are harmful outputs produced?
Efficiency	What resources does prediction require?

HELM: Evaluate More Than Accuracy

Liang et al. (2022): the original HELM organizes evaluation across 42 scenarios and seven metric dimensions.

Dimension	Question
Accuracy	How often is the answer correct?
Calibration	Does confidence match correctness?
Robustness	Does performance survive input perturbations?
Fairness	How does performance vary across groups?
Bias	What stereotypes or associations appear?
Toxicity	How often are harmful outputs produced?
Efficiency	What resources does prediction require?

Use standardized prompts and settings, publish outputs, and examine trade-offs. Calibration and robustness give two concrete ways to go beyond accuracy.

One Reference, Many Valid Outputs?

Translation, summarization, and free-form QA often allow multiple answers. Exact match gives little credit to a valid paraphrase.

Reference	the cat sat on the mat
Candidate A	a cat was sitting on the mat
Candidate B	the mat sat on the cat
Candidate C	the feline rested on the rug

One Reference, Many Valid Outputs?

Translation, summarization, and free-form QA often allow multiple answers. Exact match gives little credit to a valid paraphrase.

Reference	the cat sat on the mat
Candidate A	a cat was sitting on the mat
Candidate B	the mat sat on the cat
Candidate C	the feline rested on the rug

A preserves the event with different wording. B uses exactly the reference words but reverses the roles. C uses related words, with some change in meaning.

Compare *n*-gram matches, subsequence matches, and embedding similarity.

BLEU

Papineni et al. (2002): use **clipped n-gram precision** for translation evaluation.

Reference R: the cat sat on the mat
Candidate C: the the the the

$$p_n = \frac{\sum_{g \in \mathcal{G}_n(C)} \min\big(\text{count}_C(g), \text{count}_R(g)\big)}{\sum_{g \in \mathcal{G}_n(C)} \text{count}_C(g)}.$$

$\mathcal{G}_n(C)$ contains distinct candidate n-gram types.

BLEU: Combine Precision and a Length Penalty

For nonempty candidate length c and reference length r:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right), \qquad \text{BP} = \begin{cases} 1 & c \geq r, \\ \exp(1 - r/c) & c < r. \end{cases}$$

Standard BLEU-4 uses $N = 4$ and $w_n = 1/4$.
The brevity penalty discourages short outputs that omit content.

BLEU: Combine Precision and a Length Penalty

For nonempty candidate length c and reference length r:

$$\text{BLEU} = \text{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right), \qquad \text{BP} = \begin{cases} 1 & c \geq r, \\ \exp(1 - r/c) & c < r. \end{cases}$$

Standard BLEU-4 uses $N = 4$ and $w_n = 1/4$.
The brevity penalty discourages short outputs that omit content.

With multiple references, clip each type using its maximum reference count and use the closest reference length for each candidate (shorter in a tie).

BLEU: Trace the n-gram Counts

Keep the reference and score a different candidate:

Reference: the cat sat on the mat
Candidate: the cat is on the mat

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right).$$

BLEU: Trace the n-gram Counts

Keep the reference and score a different candidate:

Reference: the cat sat on the mat
Candidate: the cat is on the mat

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right).$$

n	**Clipped matches**	**Candidate total**	p_n
1	the (2), cat, on, mat	6	$5/6$
2	the cat; on the; the mat	5	$3/5$
3	on the mat	4	$1/4$
4	none	3	$0/3$

BLEU: Trace the n-gram Counts

Keep the reference and score a different candidate:

Reference: the cat sat on the mat
Candidate: the cat is on the mat

$$\mathrm{BLEU} = \mathrm{BP} \cdot \exp\left(\sum_{n=1}^{N} w_n \ln p_n\right).$$

n	**Clipped matches**	**Candidate total**	p_n
1	the (2), cat, on, mat	6	$5/6$
2	the cat; on the; the mat	5	$3/5$
3	on the mat	4	$1/4$
4	none	3	$0/3$

$$c = r = 6 \implies \mathrm{BP} = 1, \qquad \text{BLEU-4} = 0.$$

A zero precision makes the unsmoothed geometric mean zero. Sentence-level variants often add smoothing and must specify the method.

ROUGE-N: Measure Reference Coverage

Lin (2004): ROUGE was developed for summarization evaluation. With one reference, ROUGE-N recall is

$$R_n = \frac{\sum_{g \in \mathcal{G}_n(R)} \min\big(\text{count}_C(g), \text{count}_R(g)\big)}{\sum_{g \in \mathcal{G}_n(R)} \text{count}_R(g)}.$$

$\mathcal{G}_n(R)$ contains distinct reference n-gram types.

ROUGE-N: Measure Reference Coverage

Lin (2004): ROUGE was developed for summarization evaluation. With one reference, ROUGE-N recall is

$$R_n = \frac{\sum_{g \in \mathcal{G}_n(R)} \min\big(\text{count}_C(g), \text{count}_R(g)\big)}{\sum_{g \in \mathcal{G}_n(R)} \text{count}_R(g)}.$$

$\mathcal{G}_n(R)$ contains distinct reference n-gram types.

BLEU's p_n divides by the **candidate** total. ROUGE-N recall divides by the **reference** total.

Precision and Recall on the Same Output

Extend the reference while keeping the candidate fixed:

Reference:	the cat sat on the mat by the door
Candidate:	the cat is on the mat

Precision and Recall on the Same Output

Extend the reference while keeping the candidate fixed:

Reference: the cat sat on the mat by the door
Candidate: the cat is on the mat

	Matches	**Precision**	**ROUGE recall**
Unigrams	5	$p_1 = 5/6$	$R_1 = 5/9$
Bigrams	3	$p_2 = 3/5$	$R_2 = 3/8$

Precision and Recall on the Same Output

Extend the reference while keeping the candidate fixed:

Reference:	the cat sat on the mat by the door
Candidate:	the cat is on the mat

	Matches	**Precision**	**ROUGE recall**
Unigrams	5	$p_1 = 5/6$	$R_1 = 5/9$
Bigrams	3	$p_2 = 3/5$	$R_2 = 3/8$

These precisions are **components** of BLEU, not complete BLEU scores.

$$\mathrm{BP} = e^{1-9/6} = e^{-0.5} \approx 0.607,$$

$$\text{BLEU-1} = \mathrm{BP}\, p_1 \approx 0.505, \qquad \text{BLEU-2} = \mathrm{BP}\sqrt{p_1 p_2} \approx 0.429.$$

BERTScore: Match Contextual Token Embeddings

Zhang et al. (2020): compare contextual embeddings for candidate tokens $\mathbf{c}_1, \ldots, \mathbf{c}_m$ and reference tokens $\mathbf{r}_1, \ldots, \mathbf{r}_k$. Let $s_{ij} = \cos(\mathbf{c}_i, \mathbf{r}_j)$.

$$P_{\text{BERT}} = \frac{1}{m} \sum_{i=1}^{m} \max_j s_{ij}, \qquad R_{\text{BERT}} = \frac{1}{k} \sum_{j=1}^{k} \max_i s_{ij}.$$

Precision averages **row maxima**; recall averages **column maxima**.

BERTScore: Match Contextual Token Embeddings

Zhang et al. (2020): compare contextual embeddings for candidate tokens $\mathbf{c}_1, \ldots, \mathbf{c}_m$ and reference tokens $\mathbf{r}_1, \ldots, \mathbf{r}_k$. Let $s_{ij} = \cos(\mathbf{c}_i, \mathbf{r}_j)$.

$$P_{\text{BERT}} = \frac{1}{m} \sum_{i=1}^{m} \max_j s_{ij}, \qquad R_{\text{BERT}} = \frac{1}{k} \sum_{j=1}^{k} \max_i s_{ij}.$$

Precision averages **row maxima**; recall averages **column maxima**.

$$F_{\text{BERT}} = \frac{2 P_{\text{BERT}} R_{\text{BERT}}}{P_{\text{BERT}} + R_{\text{BERT}}}.$$

Each token independently selects its best match. Matches can be many-to-one, with no requirement that they preserve order.

BERTScore Example: Row Maxima and Column Maxima

Candidate: “feline rested”; reference: “cat sat down”.

	cat	**sat**	**down**
feline	**0.90**	0.20	0.10
rested	0.15	**0.80**	0.60

BERTScore Example: Row Maxima and Column Maxima

Candidate: “feline rested”; reference: “cat sat down”.

	cat	**sat**	**down**
feline	**0.90**	0.20	0.10
rested	0.15	**0.80**	0.60

$$P_{\text{BERT}} = \frac{0.90 + 0.80}{2} = 0.85, \qquad R_{\text{BERT}} = \frac{0.90 + 0.80 + 0.60}{3} \approx 0.767.$$

BERTScore Example: Row Maxima and Column Maxima

Candidate: “feline rested”; reference: “cat sat down”.

	cat	**sat**	**down**
feline	**0.90**	0.20	0.10
rested	0.15	**0.80**	0.60

$$P_{\text{BERT}} = \frac{0.90 + 0.80}{2} = 0.85, \qquad R_{\text{BERT}} = \frac{0.90 + 0.80 + 0.60}{3} \approx 0.767.$$

$$F_{\text{BERT}} \approx \frac{2(0.85)(0.767)}{0.85 + 0.767} \approx 0.806.$$

Related words receive partial credit, but similarity does not guarantee the same meaning.

What Reference Similarity Might Miss

- **Roles and order:** “the cat sat on the mat” versus “the mat sat on the cat” has identical unigram counts. Equal length gives BLEU-1 $= 1$, despite the role reversal.
- **Valid alternatives:** a paraphrase can express the same content with few exact matches. Extra references help cover this variation.
- **Factuality:** replacing “Paris” with “Lyon” in “Paris is the capital of France” makes a false statement while preserving most wording. Embedding similarity can also remain high.

Human Evaluation: Starts with a Rubric

- **Absolute rating:** score each response on a defined scale. Provide examples regarding what each score means.
- **Pairwise comparison:** choose A, B, or a tie under the rubric. Blind model identities and randomize display order.
- **Ranking:** order several responses under the same criteria. The annotation burden grows with their number and length.

Human Evaluation: Starts with a Rubric

- **Absolute rating:** score each response on a defined scale. Provide examples regarding what each score means.
- **Pairwise comparison:** choose A, B, or a tie under the rubric. Blind model identities and randomize display order.
- **Ranking:** order several responses under the same criteria. The annotation burden grows with their number and length.

Report annotator expertise, target population, and agreement.

Human Evaluation: Starts with a Rubric

- **Absolute rating:** score each response on a defined scale. Provide examples regarding what each score means.
- **Pairwise comparison:** choose A, B, or a tie under the rubric. Blind model identities and randomize display order.
- **Ranking:** order several responses under the same criteria. The annotation burden grows with their number and length.

Report annotator expertise, target population, and agreement.

Caveat: Disagreement can reflect genuine differences in preferences or interpretation.

Chatbot Arena: From Votes to Ratings

Zheng et al. (2023): users compare anonymous model responses to their own questions.
Elo updates relative ratings from pairwise wins, ties, and losses.

- R_A, R_B: current ratings; E_A: A's expected score.
- Actual score S_A: 1 for a win, $1/2$ for a tie, 0 for a loss.
- $K > 0$ controls the update size; R'_A is A's new rating.

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}, \qquad R'_A = R_A + K(S_A - E_A).$$

Chatbot Arena: From Votes to Ratings

Zheng et al. (2023): users compare anonymous model responses to their own questions.
Elo updates relative ratings from pairwise wins, ties, and losses.

- R_A, R_B: current ratings; E_A: A's expected score.
- Actual score S_A: 1 for a win, $1/2$ for a tie, 0 for a loss.
- $K > 0$ controls the update size; R'_A is A's new rating.

$$E_A = \frac{1}{1 + 10^{(R_B - R_A)/400}}, \qquad R'_A = R_A + K(S_A - E_A).$$

Example: $R_A = R_B = 1500$ gives $E_A = 0.5$. If A wins, $S_A = 1$; with $K = 32$:

$$R'_A = 1500 + 32(1 - 0.5) = 1516.$$

Beating a stronger opponent earns a larger rating increase.

MT-Bench and LM-as-Judge

Zheng et al. (2023): MT-Bench has 80 two-turn questions across eight categories.

Use a strong LM as a judge:

1. Provide the conversation, response(s), and an explicit rubric.
2. Request a single-response score or a pairwise preference.
3. Compare judgments against independent human annotations.

MT-Bench and LM-as-Judge

Zheng et al. (2023): MT-Bench has 80 two-turn questions across eight categories.

Use a strong LM as a judge:

1. Provide the conversation, response(s), and an explicit rubric.
2. Request a single-response score or a pairwise preference.
3. Compare judgments against independent human annotations.

First-turn agreement, S2 protocol excluding ties	Rate
Human–Human	81%
GPT-4 pairwise–Human	85%

Agreement is specific to the task, rubric, and tie handling. It does not establish that the judge can verify arbitrary factual or reasoning claims.

LM Judges: Validate the Failure Modes

For the explanation task, compare controlled versions of the same response.

Risk	Control or check
Position bias	Swap response order; flag inconsistent judgments.
Verbosity bias	Compare concise and verbose explanations with the same content.
Self-enhancement	Check preferences for the judge's own style or model family.
Factual or reasoning errors	Have an expert check the explanation's claims.

LM Judges: Validate the Failure Modes

For the explanation task, compare controlled versions of the same response.

Risk	Control or check
Position bias	Swap response order; flag inconsistent judgments.
Verbosity bias	Compare concise and verbose explanations with the same content.
Self-enhancement	Check preferences for the judge's own style or model family.
Factual or reasoning errors	Have an expert check the explanation's claims.

Validate important conclusions with independent human or task-specific checks. Report judge version, prompt, rubric, decoding settings, and disagreement handling.

With a scoring rule in place, we can examine the reliability of comparisons.

Data Contamination: What Did Training Expose?

Contamination: benchmark content enters the model's training data.
Consider: "What is Freda's made-up word in CS 489 Lecture 14?"
(Answer: "Catito")

Data Contamination: What Did Training Expose?

Contamination: benchmark content enters the model's training data.
Consider: "What is Freda's made-up word in CS 489 Lecture 14?"
(Answer: "Catito")

System A: unseen content

Can search the web.
Finds this slide and extracts the word.

System B: seen in training

Can recall the tutorial's content.
Retrieves the word from memory.

Data Contamination: What Did Training Expose?

Contamination: benchmark content enters the model's training data.
Consider: "What is Freda's made-up word in CS 489 Lecture 14?"
(Answer: "Catito")

System A: unseen content

Can search the web.
Finds this slide and extracts the word.

System B: seen in training

Can recall the tutorial's content.
Retrieves the word from memory.

Both can answer: `Catito`.

The same correct answer can reflect different solution and generalization paths.

Detecting and Reducing Contamination

Lecture 10 introduced training-data overlap checks.
Evaluation also requires keeping test information out of development.

- **Search training data:** exact and *n*-gram matches find copies but can miss paraphrases.
- **Probe exposure:** reproduction of distinctive examples or canary strings can provide evidence of memorization.
- **Use fresh data:** private tests or data created after a verified training cutoff reduce exposure risk.
- **Separate development and testing:** select prompts and models on development data; reserve the test set for final evaluation.

Detecting and Reducing Contamination

Lecture 10 introduced training-data overlap checks.
Evaluation also requires keeping test information out of development.

- **Search training data:** exact and *n*-gram matches find copies but can miss paraphrases.
- **Probe exposure:** reproduction of distinctive examples or canary strings can provide evidence of memorization.
- **Use fresh data:** private tests or data created after a verified training cutoff reduce exposure risk.
- **Separate development and testing:** select prompts and models on development data; reserve the test set for final evaluation.

Caveat: Exposure alone does not quantify score inflation. Clean and overlapping subsets can also differ in difficulty.

Data Curation Can Change the Ranking

The GSM8K-Platinum authors removed 110 questions and corrected 10 answers. Reported errors below come from zero-shot evaluation; lower is better.

	GSM8K	**GSM8K-Platinum**
Questions	1,319	1,209
Claude 3.7 Sonnet	38	
DeepSeek-R1	56	

Data Curation Can Change the Ranking

The GSM8K-Platinum authors removed 110 questions and corrected 10 answers. Reported errors below come from zero-shot evaluation; lower is better.

	GSM8K	**GSM8K-Platinum**
Questions	1,319	1,209
Claude 3.7 Sonnet	38	9
DeepSeek-R1	56	4

The reported ordering reverses after curation.

Data Curation Can Change the Ranking

The GSM8K-Platinum authors removed 110 questions and corrected 10 answers. Reported errors below come from zero-shot evaluation; lower is better.

	GSM8K	GSM8K-Platinum
Questions	1,319	1,209
Claude 3.7 Sonnet	38	9
DeepSeek-R1	56	4

The reported ordering reverses after curation.

Question removal and answer correction both change the measurement.
Takeaway: Inspect the benchmark's errors as well as the model's errors.

Prompt Sensitivity Is Protocol Sensitivity

The same multiple-choice question can be evaluated in different ways:

Protocol	What is scored
Score A/B/C/D after a fixed prefix	Highest-probability answer label
Generate an answer directly	Label extracted from generated text
Generate a rationale, then answer	Extracted label after extra computation

Prompt Sensitivity Is Protocol Sensitivity

The same multiple-choice question can be evaluated in different ways:

Protocol	What is scored
Score A/B/C/D after a fixed prefix	Highest-probability answer label
Generate an answer directly	Label extracted from generated text
Generate a rationale, then answer	Extracted label after extra computation

- Demonstration selection, order, system prompts, and formatting can affect results.
- Fix or report the model version, tools, token budget, decoding settings, and answer parser.
- Tune prompts on development data and report variation across settings.

Is a Two-Point Accuracy Gain Meaningful?

Consider an 100-example multiple-choice test with two models, A and B.

Outcome	Questions
Both correct	70
Only A correct	4
Only B correct	6
Both wrong	20

Is a Two-Point Accuracy Gain Meaningful?

Consider an 100-example multiple-choice test with two models, A and B.

Outcome	Questions
Both correct	70
Only A correct	4
Only B correct	6
Both wrong	20

A scores 74%; B scores 76%. The difference is **2 percentage points**, driven by 10 disagreements.

Are they meaningfully different, or could this difference arise from sampling variation?

Paired Bootstrap: Resample the Comparison

On the same 100 independently sampled questions, A scores 74% and B scores 76%. How uncertain is the accuracy gap $\hat{\Delta} = 0.76 - 0.74 = 0.02$?

Paired Bootstrap: Resample the Comparison

On the same 100 independently sampled questions, A scores 74% and B scores 76%. How uncertain is the accuracy gap $\hat{\Delta} = 0.76 - 0.74 = 0.02$?

1. Store each question's paired results (a_i, b_i): correct $= 1$, incorrect $= 0$.
2. Draw 100 question indices **with replacement**.
3. Use **the same indices for both models**; recompute the accuracy gap $\hat{\Delta}^*$.
4. Repeat many times; use the 2.5th and 97.5th percentiles of these gaps for an approximate 95% bootstrap confidence interval.

Paired Bootstrap: Resample the Comparison

On the same 100 independently sampled questions, A scores 74% and B scores 76%. How uncertain is the accuracy gap $\hat{\Delta} = 0.76 - 0.74 = 0.02$?

1. Store each question's paired results (a_i, b_i): correct $= 1$, incorrect $= 0$.
2. Draw 100 question indices **with replacement**.
3. Use **the same indices for both models**; recompute the accuracy gap $\hat{\Delta}^*$.
4. Repeat many times; use the 2.5th and 97.5th percentiles of these gaps for an approximate 95% bootstrap confidence interval.

Reuse stored predictions; no retraining or new model runs. If the interval includes 0, the data do not clearly favor either model.

Paired Bootstrap: One Sampling Example

Toy example: 5 questions, with both models' results stored in each row.

Original results

i	a_i	b_i	$b_i - a_i$
1	1	1	0
2	1	0	-1
3	0	1	1
4	0	1	1
5	0	0	0

$$\hat{\Delta} = \frac{0 - 1 + 1 + 1 + 0}{5} = 0.20.$$

Paired Bootstrap: One Sampling Example

Toy example: 5 questions, with both models' results stored in each row.

Original results

i	a_i	b_i	$b_i - a_i$
1	1	1	0
2	1	0	-1
3	0	1	1
4	0	1	1
5	0	0	0

Sampled indices: $(2, 2, 3, 5, 5)$

i	a_i	b_i	$b_i - a_i$
2	1	0	-1
2	1	0	-1
3	0	1	1
5	0	0	0
5	0	0	0

$$\hat{\Delta} = \frac{0 - 1 + 1 + 1 + 0}{5} = 0.20.$$

Paired Bootstrap: One Sampling Example

Toy example: 5 questions, with both models' results stored in each row.

Original results

i	a_i	b_i	$b_i - a_i$
1	1	1	0
2	1	0	-1
3	0	1	1
4	0	1	1
5	0	0	0

$$\hat{\Delta} = \frac{0 - 1 + 1 + 1 + 0}{5} = 0.20.$$

Sampled indices: $(2, 2, 3, 5, 5)$

i	a_i	b_i	$b_i - a_i$
2	1	0	-1
2	1	0	-1
3	0	1	1
5	0	0	0
5	0	0	0

$$\hat{\Delta}^* = \frac{-1 - 1 + 1 + 0 + 0}{5} = -0.20.$$

A Score Is an Experimental Result

Report the experiment

- Target population and test data.
- System, protocol, and trial budget.
- Scoring rules and uncertainty.

A Score Is an Experimental Result

Report the experiment

- Target population and test data.
- System, protocol, and trial budget.
- Scoring rules and uncertainty.

Inspect marked-wrong answers

- Missing knowledge or failed reasoning?
- Correct answer in an unparsed format?
- Judge error or incorrect gold label?

An arithmetic error, a parser failure, and a wrong gold label need different fixes.

A Score Is an Experimental Result

Report the experiment

- Target population and test data.
- System, protocol, and trial budget.
- Scoring rules and uncertainty.

Inspect marked-wrong answers

- Missing knowledge or failed reasoning?
- Correct answer in an unparsed format?
- Judge error or incorrect gold label?

An arithmetic error, a parser failure, and a wrong gold label need different fixes.

Next: decoding algorithms that change the outputs we evaluate.