Evaluating Language Models

Perplexity · BLEU · ROUGE · Benchmarks

Why evaluation is hard

Prompt: "Summarize the key findings of the study on climate change."

Output A:
"The study found that global temperatures rose 1.2°C since 1900, with accelerating ice loss in the Arctic."

Output B:
"The research demonstrates significant atmospheric changes leading to enhanced precipitation patterns globally."

Multiple valid outputs
Many correct ways to say the same thing

Meaning vs. form
Surface similarity $\neq$ semantic similarity

Task-dependent
Translation, summary, chat need different metrics

Intrinsic vs. extrinsic evaluation

Intrinsic

Measures the model *itself*

Perplexity
Cross-entropy loss
Bits-per-character

"How well does the model predict?"

Low PP $\neq$ good output

Extrinsic

Measures down-
stream *task* quality

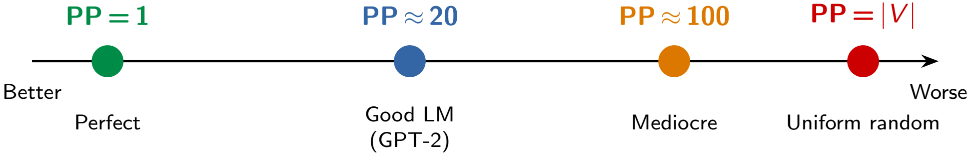
BLEU (translation)
ROUGE (summarization)
Accuracy / F1

"How useful is the output?"

Perplexity

$$\mathrm{PP}(W) \;=\; \exp\!\left(-\frac{1}{N}\sum_{i=1}^{N}\log P(w_i \mid w_{<i})\right)$$

Intuition: On average, the model is as uncertain as choosing uniformly among **PP** options at each step.



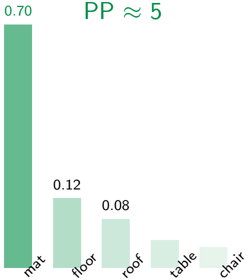
Equivalently: $\mathrm{PP} \;=\; 2^{H}$
where H is the cross-entropy

Perplexity — visual intuition

Predicting the next token after: “The cat sat on the ______”

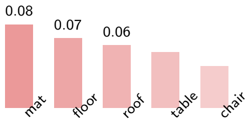
Model A

PP $\approx$ 5



Model B

PP $\approx$ 50



Lower perplexity = model concentrates probability on the right tokens

Perplexity — caveats

1. Tokenizer-dependent

Different BPE vocabularies produce different N — PP scores across models with different tokenizers are *not* comparable.

2. Not a quality metric

Low PP means good prediction, not good *content*. A model can confidently produce fluent nonsense.

3. Memorization

A model that memorizes training data has very low PP on that data but fails on new text. Check on *held-out* data.

4. LM-only

Perplexity only applies to generative (autoregressive) models. For classification, QA, etc. use task-specific metrics.

Takeaway: PP is great for comparing LMs *on the same data and tokenizer*, but don't over-interpret it.

BLEU — Bilingual Evaluation Understudy

Papineni et al., 2002 — designed for machine translation

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

match

no match

$$p_n = \frac{\text{matched } n\text{-grams in candidate}}{\text{total } n\text{-grams in candidate}}$$

Unigram precision: $p_1 = 5/6 \approx 0.83$ Bigram precision: $p_2 = 3/5 = 0.60$

Core idea: count how many n-grams in the candidate also appear in the reference

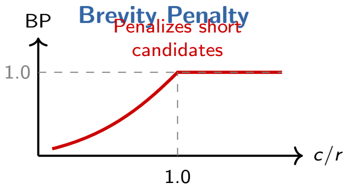
BLEU — modified precision & brevity penalty

Problem: "the the the the the" has 100% unigram precision against any sentence containing "the"!

$$\mathrm{BLEU} = \underbrace{\mathrm{BP}}_{\text{brevity penalty}} \cdot \exp\left(\sum_{n=1}^{N} w_n \log p_n\right)$$

$$\mathrm{BP} = \min\Big(1, \exp\big(1 - r/c\big)\Big)$$

r = reference length, c = candidate length, $N = 4$, $w_n = \frac{1}{4}$



BLEU-4 is standard:
geometric mean of p_1, p_2, p_3, p_4
Higher $n \Rightarrow$ stricter fluency check

BLEU — worked example

Reference: “The cat is sitting on the mat” $r = 7$

Candidate: “The the cat mat” $c = 4$

Clipped precision by n-gram order

n	**Candidate n-grams**	**Clipped matches**	p_n
1	4	3	$3/4$
2	3	1	$1/3$
3	2	0	$0/2$
4	1	0	$0/1$

$$\mathrm{BP} = \exp(1 - 7/4) = \exp(-0.75) \approx 0.47$$

$$\log\text{-avg} = \tfrac{1}{4}(\log 0.75 + \log 0.33 + \log \varepsilon + \log \varepsilon) \qquad (\varepsilon = \text{smoothed zero})$$

BLEU $\approx$ **0.47** $\times$ (very small) $\approx$ **very low** — short, incomplete candidate

ROUGE — Recall-Oriented Understudy for Gisting Evaluation

BLEU asks: "How many candidate n-grams appear in the reference?" (**precision**)

ROUGE asks: "How many *reference* n-grams appear in the candidate?" (**recall**)

$$\text{ROUGE-}N = \frac{\text{matched } n\text{-grams}}{\text{total } n\text{-grams in } \textbf{reference}}$$

ROUGE-1

Unigram recall

Content coverage

ROUGE-2

Bigram recall

Fluency + content

ROUGE-L

Longest Common Subsequence

Word-order aware

In practice, ROUGE-1, ROUGE-2, and ROUGE-L are all reported.
F1 variants (harmonic mean of precision & recall) are common.

BLEU vs. ROUGE

BLEU

Precision-based

"What fraction of candidate n-grams are correct?"

Best for: **Translation**

ROUGE

Recall-based

"What fraction of reference content is captured?"

Best for: **Summarization**

	BLEU	**ROUGE**
Focus	Precision	Recall (+ F1)
Penalizes short output	Yes (BP)	No
Penalizes missing content	No	Yes
Typical n	1–4 (geometric mean)	1, 2, or L
Range	0–1 (often $\times$100)	0–1

Limitations of *n*-gram metrics

Reference: "The movie was really excellent"

Candidate A:
"The film was truly great"
Good paraphrase, low BLEU

Candidate B:
"The movie was really bad"
Wrong meaning, high BLEU

Synonyms ignored
"great" $\neq$ "excellent" even though they mean the same

Semantics missed
"really bad" vs. "really excellent" share most *n*-grams

Multiple references needed
One reference can't capture all valid translations / summaries

Open-ended generation
No reference exists for creative writing, chat, etc.

Modern alternatives: **BERTScore** (embedding similarity), **METEOR** (synonyms + stemming), **COMET** (learned metric), **LLM-as-Judge**

Benchmarks — evaluating holistically

Instead of one metric, aggregate many tasks to measure general capability

GLUE 2018	SuperGLUE 2019	MMLU 2021	Big-Bench 2022	Chatbot Arena 2023
9 NLU tasks: sentiment, NLI, paraphrase, QA	8 harder NLU tasks: coreference, causal reasoning	57 subjects: STEM, humanities, social sciences	200+ diverse tasks: reasoning, math, knowledge	Human pairwise preference (Elo ratings)

Key idea: single metrics are fragile — benchmarks aggregate many sub-tasks for a holistic view

Benchmark saturation: GLUE was "solved" within a year — SuperGLUE soon after. The bar keeps rising.

Key benchmarks at a glance

Benchmark	What it tests	Format	Notable for
MMLU	Knowledge (57 subjects)	Multiple choice	Breadth of knowledge
HumanEval	Code generation	Function completion	Coding ability
HellaSwag	Commonsense reasoning	Sentence completion	Physical understanding
TruthfulQA	Factual accuracy	QA	Hallucination resistance
GSM8K	Math reasoning	Word problems	Step-by-step reasoning
ARC	Science questions	Multiple choice	Grade-school science
Winogrande	Coreference	Pronoun resolution	Linguistic understanding

Aggregate scores (e.g., **Open LLM Leaderboard**) let you compare models at a glance—but no single number tells the full story.

The evaluation landscape

What are you evaluating?

Language model
Perplexity
Cross-entropy

Translation
BLEU
COMET

Summarization
ROUGE
BERTScore

General capability
MMLU, HumanEval
Chatbot Arena

Human evaluation
The gold standard — expensive, slow, but captures fluency, coherence, factuality, and preference

No single metric captures everything — use the right tool for the right task

Further reading

Classic Metrics

- Papineni et al. (2002), "BLEU: A Method for Automatic Evaluation of Machine Translation"
- Lin (2004), "ROUGE: A Package for Automatic Evaluation of Summaries"
- Banerjee & Lavie (2005), "METEOR: An Automatic Metric for MT Evaluation"

Modern Evaluation

- Zhang et al. (2020), "BERTScore: Evaluating Text Generation with BERT"
- Zheng et al. (2024), "Judging LLM-as-a-Judge with MT-Bench and Chatbot Arena"

Benchmarks & Leaderboards

- Srivastava et al. (2023), "Beyond the Imitation Game" (BIG-Bench)
- Hendrycks et al. (2021), "Measuring Massive Multitask Language Understanding" (MMLU)

Questions?

Next: Early Notable Models — GPT, BERT, T5