LLM evaluation:

↳ Benchmarking:

↳ measuring LLM capabilities

↳ benchmark generally consists of

1. dataset

↳ static: fixed set of prompts &

↳ dynamic; evolving / updated inputs

↳ single task

↳ e.g. translation, summarization, etc

↳ broad task distributions

↳ e.g. random sample of user prompts to ChatGPT

2. metric

↳ measure of performance:

↳ given (x, y), how good is y?

↳ reward model!

↳ Goodhart's Law:

"When a measure becomes a target, it ceases to be a good measure"

3. generation setup

↳ decoding hyperparams

↳ tool use

↳ multiple gens or just one

↳ system prompt
↳ reasoning
etc.

Static benchmark w/ simple metric
↳ machine translation
↳ input: German sentence x
output: English sentence y
↳ metric:
↳ access to reference translation y_{ref}
↳ compute n-gram precision

BLEU computes geometric ave of n-gram precision from $n=1 \ldots 4$

$$prec_n = \frac{\text{count (matching n-grams in } y, y_{ref})}{\text{count (n-grams in } y)}$$

$$BLEU = \exp\left(\sum_{n=1}^{4} \frac{1}{4} \log prec_n\right) \cdot BP$$

BP ↓

brevity penalty

1 if $len(y) > len(y_{ref})$

decays w/ $len(y_{ref}) - len(y)$

x: Das ist ein kleines Haus.

y: This is a little house.

y_{ref}: This is a small house.

unigram prec: $\frac{4}{5}$

bigram prec: $\frac{2}{4}$

BP = 1

$BLEU_2 = 0.63$

y_2 = This house is small

unigram prec = 5/5

$BLEU_2 = 0$ (no bigram matches!)

↳ BLEU, ROUGE, etc ⇒ string matching (weak metrics, but cheap)

↳ what about multiple choice?

↳ MMLU

↳ wide task distribution

57 subjects

↳ example of knowledge exam

↳ dynamic / interaction / mix

↳ Chatbot Arena

↳ METR

↳ EQ Bench ⇒ static / easy eval

LLM judge:

- give strong LLM a rubric and candidate answers, pick better one or grade
  - who evals the evaluator?