Framing Issues of Measurement in AI

Luca Mari
Università Cattolica

Milan Network of Logic and Philosophy of Science
Università degli Studi di Milano, 15 January 2026

The context

The New York Times

A.I. Has a Measurement Problem

Which A.I. system writes the best computer code or generates the most realistic image? Right now, there's no easy way to answer those questions.

(15 April 2024, https://www.nytimes.com/2024/04/15/technology/ai-models-measurement.html)

Framing the analysis

Today we say *Artificial Intelligence* (AI) and mean *Machine Learning*

The relation between AI and *Measurement Science* (MS) is twofold:

- **AI for MS**: how can AI help improving measurement?
  ("smart" meters, automated test generation and evaluation, …)
- **MS for AI**: how can MS help improving *Learning Machines* (LMs)
  and our knowledge about them?

We focus here on the latter, and specifically about the evaluation of LM behavior:

- can we measure the quality of the behavior of a LM? how?
- for a given kind of task, what are the best LMs?

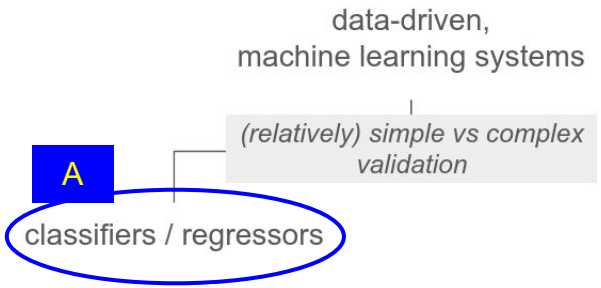
Framing the analysis

About the evaluation of LM behavior two main issues need to be considered:

- do we know **what** we want to measure?
  → is the measurand (“quality of behavior”) well defined?
- do we know **how** we want to measure?
  → is the measuring system well designed?

In an evaluation-oriented framework three classes of LMs can be identified

Class A



Traditional ML systems,
for classifications or regressions

- the measurand is well defined
- labels / true values for training / calibration are available

Tools: k-nearest neighbors, logistic regression, decision trees, neural networks, …
Examples: handwritten character recognition, antispam filtering, recommendation systems, sentiment analysis, time series forecast, …

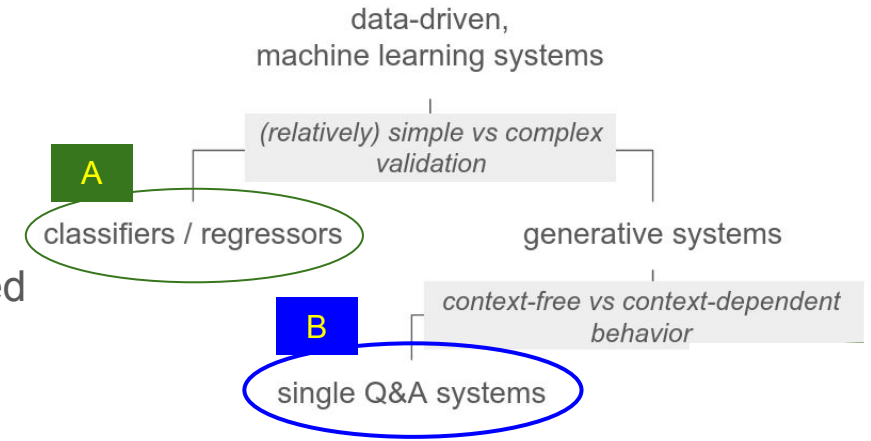
Well-known statistical / data mining techniques: distinction between features and targets; training vs test set split; bias vs variance (underfitting vs overfitting); …

Well-known statistical / data mining quality parameters: precision, recall, accuracy, …

Two main **measurement-related problems**:

- **instability**, if the relation between the features and the targets changes in time (“independent and identically distributed variables” (IID) condition not fulfilled)
- **bias**, if the training set is not sufficiently representative of the population (incorrect choice of measurement standards used for calibration)

Class B



Single Q&A GenAI systems,
for context-free tasks

- the measurand is not so well defined
- labels used in training may be controversial in inference

Tools: RNNs, Transformers
Examples: translation, summarization, …

All common benchmarks for language models assume single Q&A

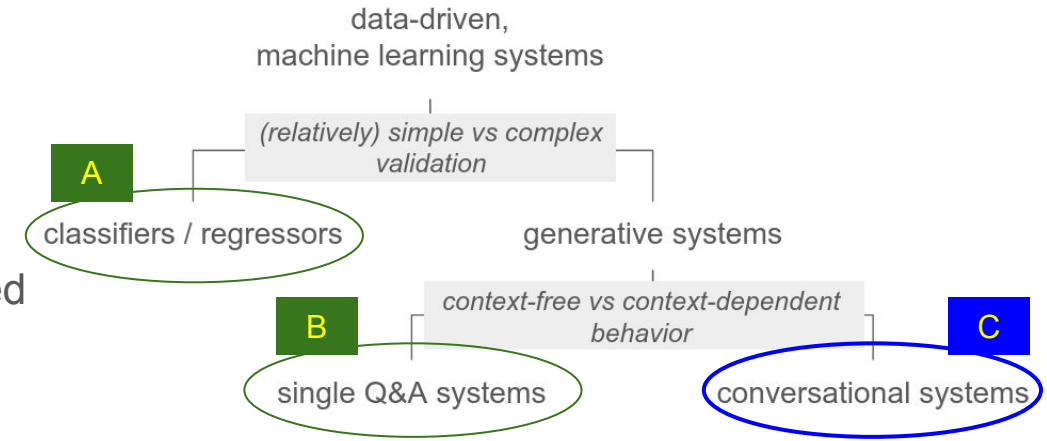
– MMLU (Massive Multitask Language Understanding): performance on a wide range of tasks (“the SAT for chatbots”)
– HellaSwag: commonsense reasoning
– PIQA (Physical Interaction Question Answering): comprehension of physical interactions
– WinoGrande: common sense reasoning, complex pronoun disambiguation

Together with what was mentioned for Class A systems, the key **measurement-related problem**: there could be no intersubjective criteria to assess the quality of inference results
(see the case of BLEU: “the closer a machine translation is to a professional human translation, the better it is”)

Class C

Conversational GenAI systems,
for context-sensitive tasks

- the measurand is not so well defined
- labels used in training may be controversial in inference

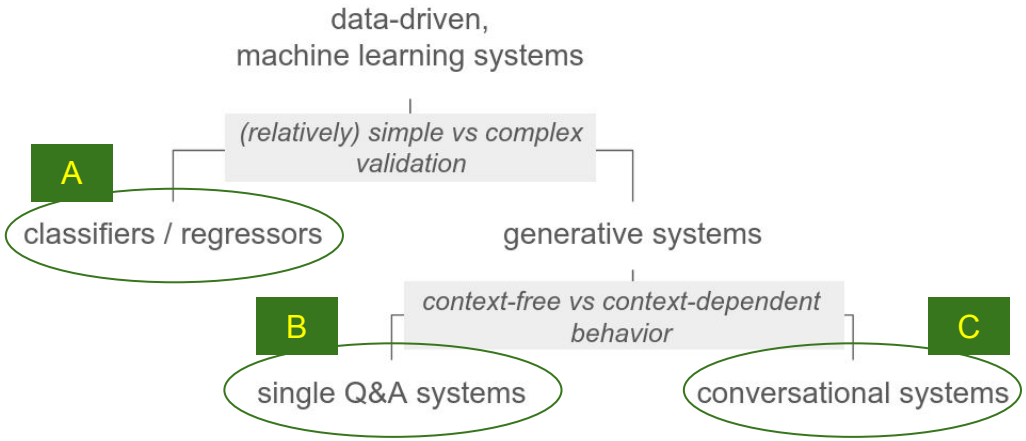


Tools: Transformers
Examples: like for Class B, plus conversations

We are not aware of any benchmark / metric specifically devoted to context-sensitive tasks
(LMArena Leaderboard (https://lmarena.ai/leaderboard) is based on direct comparison and using the Elo rating system)

Together with what was mentioned for Class A & B systems, the key **measurement-related problem**: there could be no intersubjective criteria to assess the quality of inference results

In summary



Classes of systems

A: traditional ML systems

B: single Q&A GenAI systems

C: conversational GenAI systems

Measurement-related problems

solved or well-known

partially solved, hard

unsolved, very hard

Open issues / Why this is important

The evaluation of the quality of behavior of Class C systems (chatbots…),
as a context-sensitive task, is still an open issue
(in particular: how to avoid the “learn to test” effect?)

Moreover:

- chatbots can be trained for function / tool calling and therefore as hybrid systems
- chatbots can be enabled to interact with each other in agent-based architectures

How to evaluate the quality of behavior of these systems is still an open issue

Finally, the training process of chatbots includes today not only self-supervised learning and supervised fine tuning (both token oriented), but also reinforcement learning (task oriented): particularly the version with human feedback (RLHF) is unavoidably ideological,
and therefore very hard to evaluate objectively
(see the case of the OpenAI Model Spec: https://model-spec.openai.com)

Thanks for your attention

Luca Mari
https://lmari.github.io
luca.mari@unicatt.it