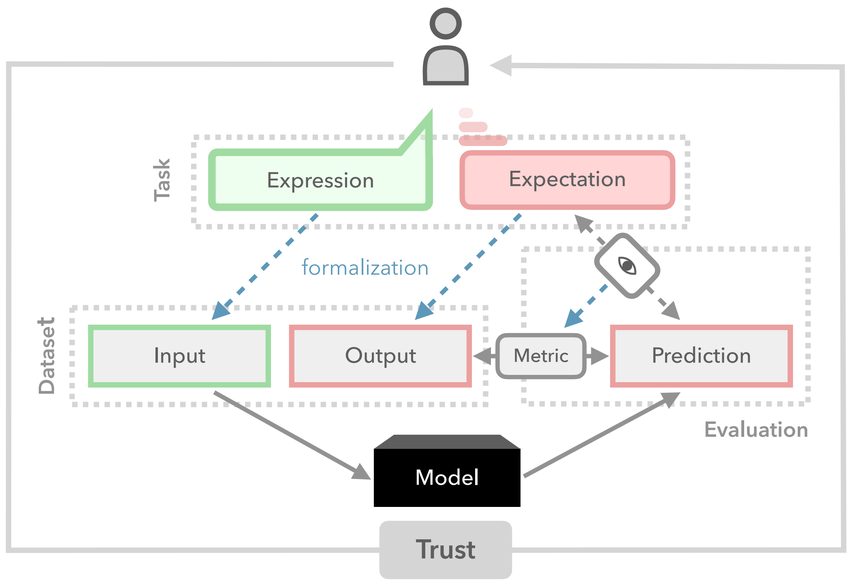
Establishing Trustworthiness: Rethinking Tasks and Model Evaluation

Robert Litschko*, Max Müller-Eberstein*, Rob van der Goot, Leon Weber, Barbara Plank

Working Definition

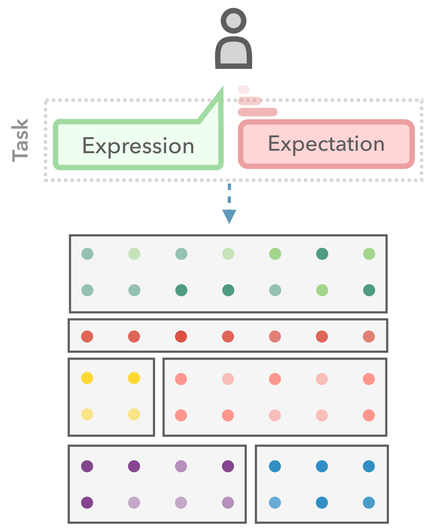


Trust arises from knowledge of origin as well as from knowledge of functional capacity.

Trustworthiness - Working Definition by David G. Hays, 1979

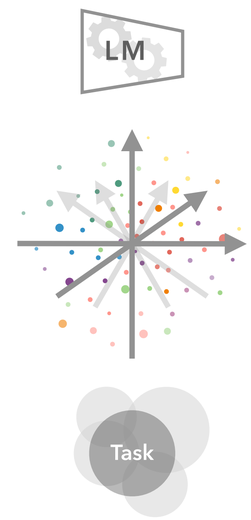
Paradigm Shift

Old: Compartmentalized NLP



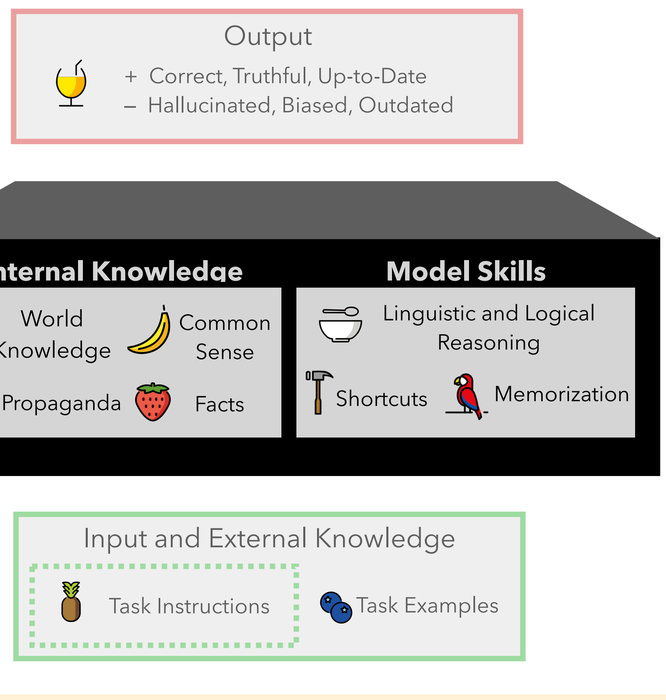
Task-specific Models and Evaluation Protocols

New: Large Language Models

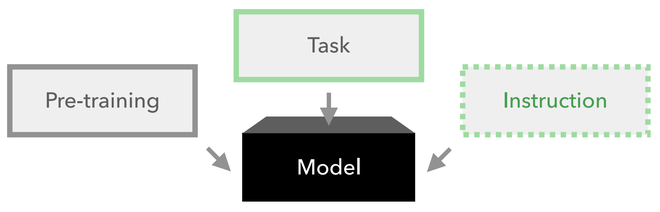


Massive Multi-task and Multilingual Learning

Trustworthiness Desiderata



Openmoji, CC BY-SA 4.0

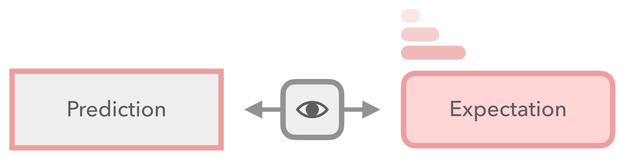


D1 Knowledge about Model Input

- Over time, we have lost the ability to inspect intermediate outputs and control model inputs.
- What "ingredients" (, , ,) are used to generate output?

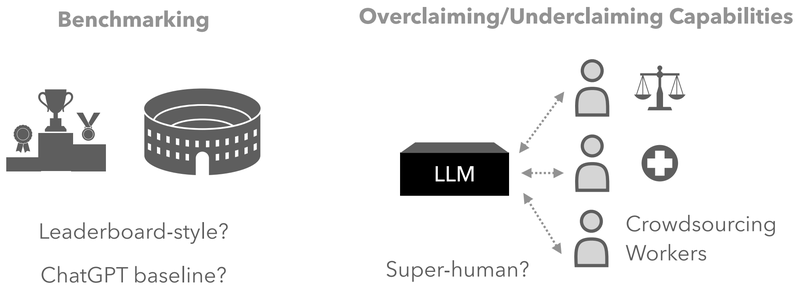
D2 Knowledge about Model Behaviour

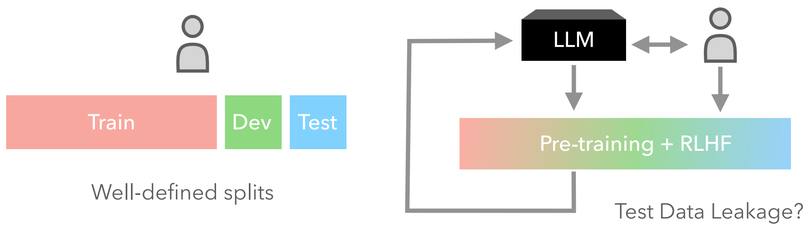
- Over time, we lost the ability to interpret decision boundaries and model behaviour.
- Which skills (, ,) are employed to process ingredients into outputs?

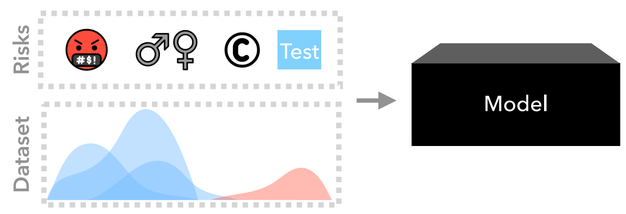


D3 Knowledge of Evaluation Protocols

- Today, LLMs are used to solve tasks outside of the benchmark: user-formulated tasks via instructions.
- What do human-level comparisons and leader-boards tell us about model capabilities?







D4 Knowledge of Data Origin

- Is model performance reflective of its capabilities or overestimated due to data leakage?
- Additional risks from unknown provenance include hate speech, biases, copyright violations, …

What Can We Do to Gain Trust?

Explain skills required ( ) vs. skills employed (,)

- **Linguistically-motivated:**
  - Probing tasks, Checklists
  - Linguistic Profiling
- **Model-based:** Attribute skills to parameter regions.
- **Interpretability** methods.

Facilitate Representative and Comparable Qualitative Analysis

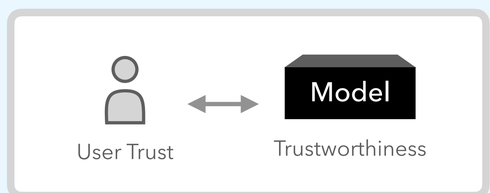
- **Faceted** quantitative analysis.
- **Standardized qualitative** evaluation protocols.
- Expert-curated **diagnostics annotations** w.r.t cognitive abilities required to solve task.

Be explicit about data provenance.

- Opt for cross-X evaluation.
- **Closed-source models** (e.g ChatGPT) typically **evolve over time** and have unknown data provenance.

→ Untrustworthy baselines

User Trust



- Trustworthiness: Knowledge about LLM's functional capacity and origin.
- User Trust: What do users do with the model output across multiple interactions (e.g., verify, fact check, revise, accept)?

IT-UNIVERSITETET I KØBENHAVN
north
LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN
nlp
CIS
mcml Munich Center for Machine Learning
EMNLP 2023