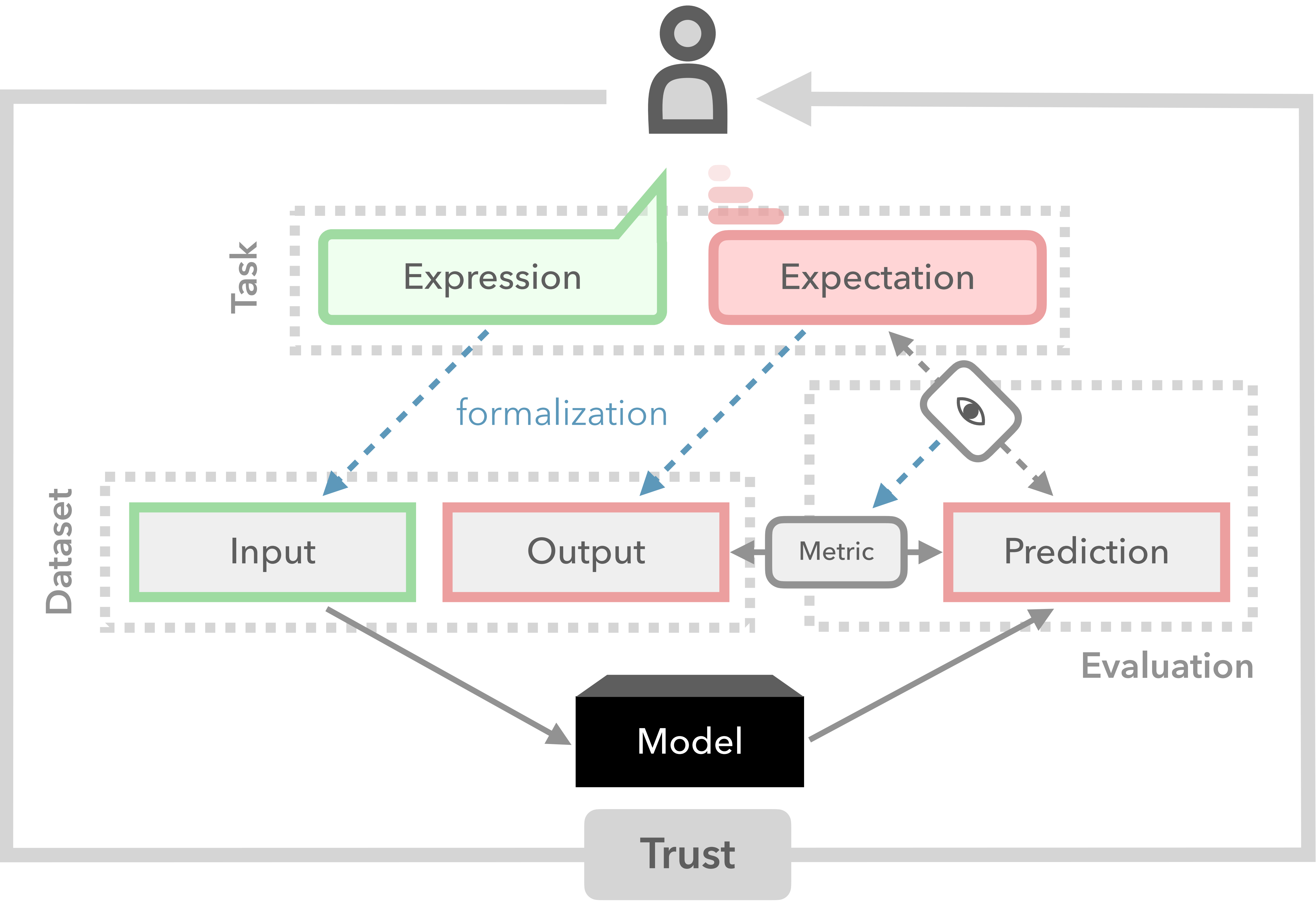
Establishing Trustworthiness: Rethinking Tasks and Model Evaluation

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

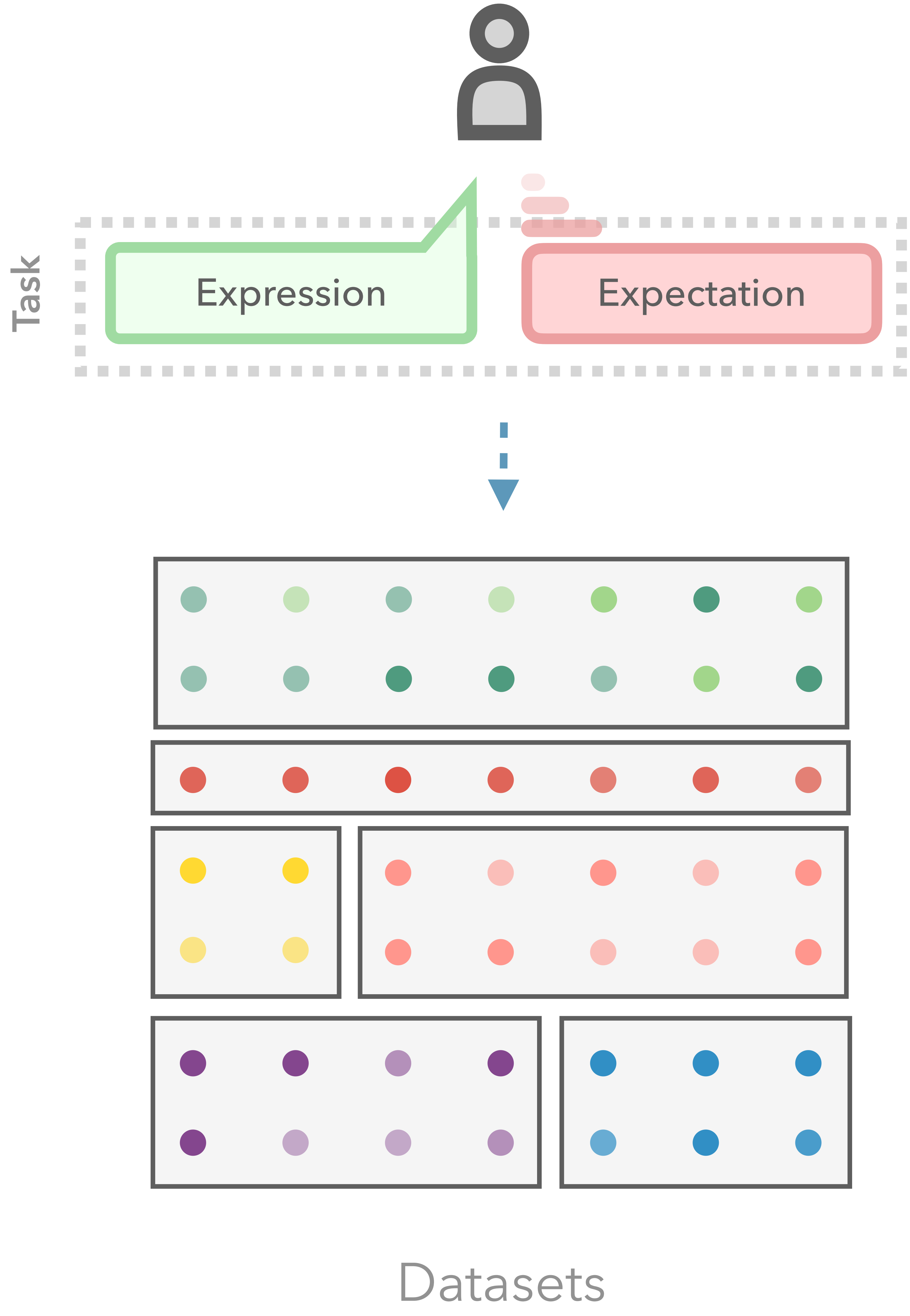
EMNLP 2023

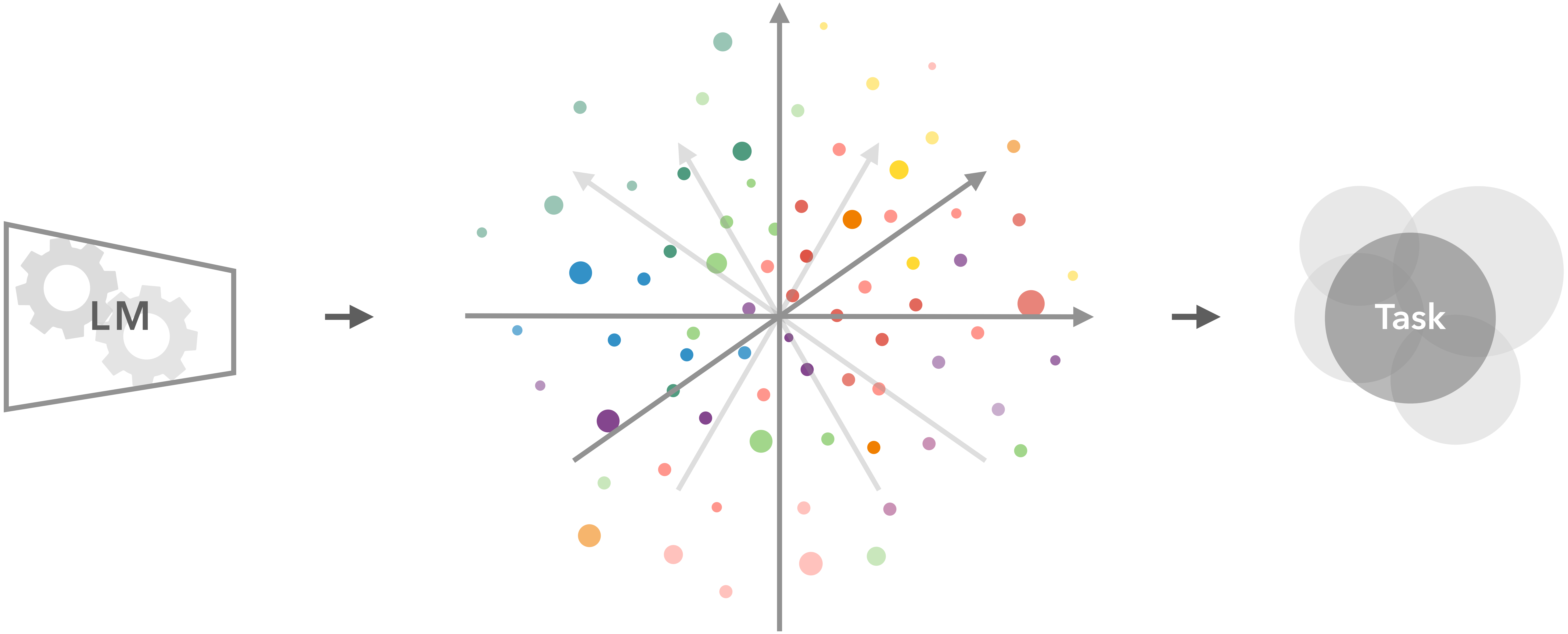
IT-UNIVERSITETET I KØBENHAVN / north × LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN / nlp

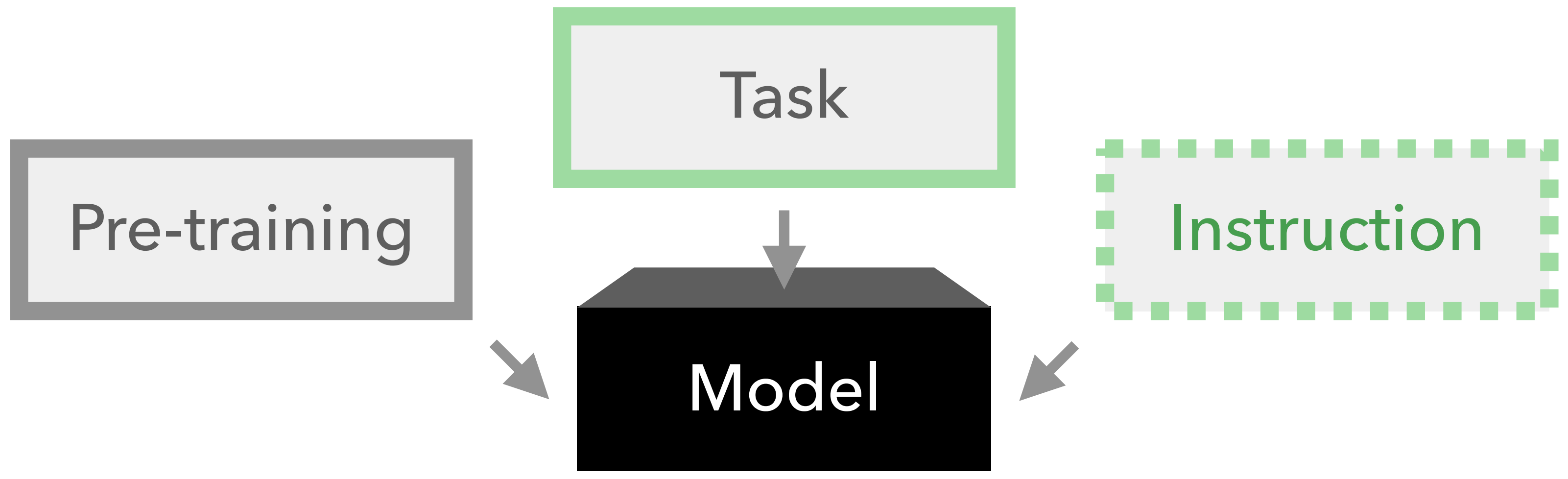


Trust arises from **knowledge of origin** as well as from **knowledge of functional capacity.**

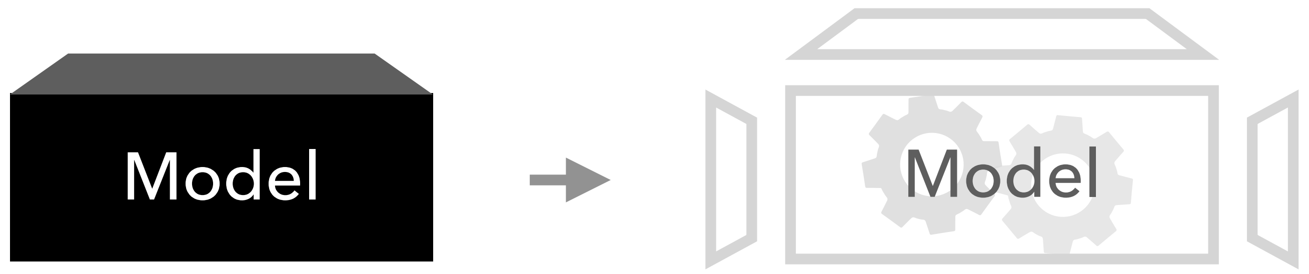
Trustworthiness - Working Definition by David G. Hays, 1979


Task
Expression
Expectation
Datasets

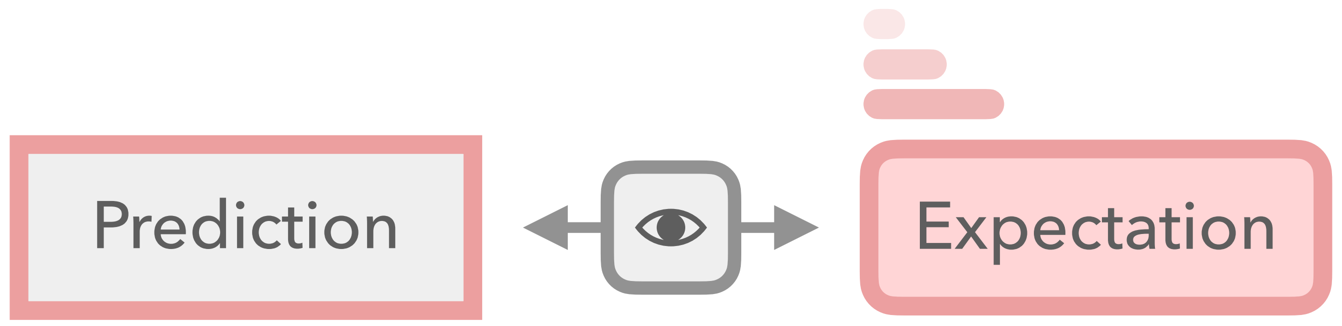

LM
Task



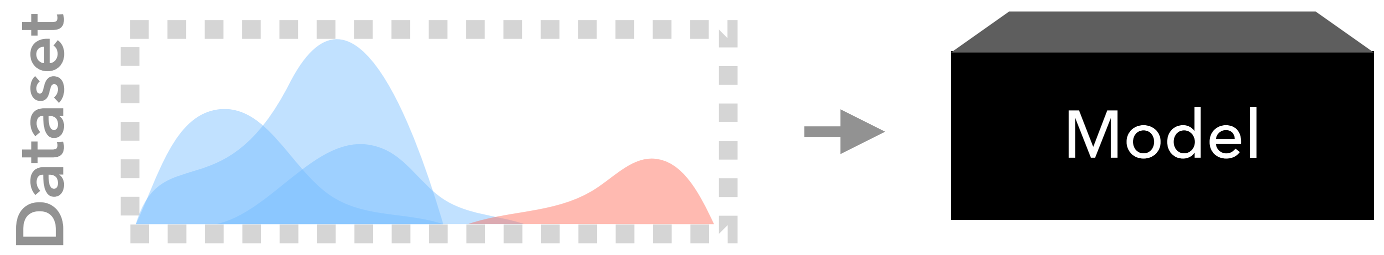
D1 Knowledge about Model Input



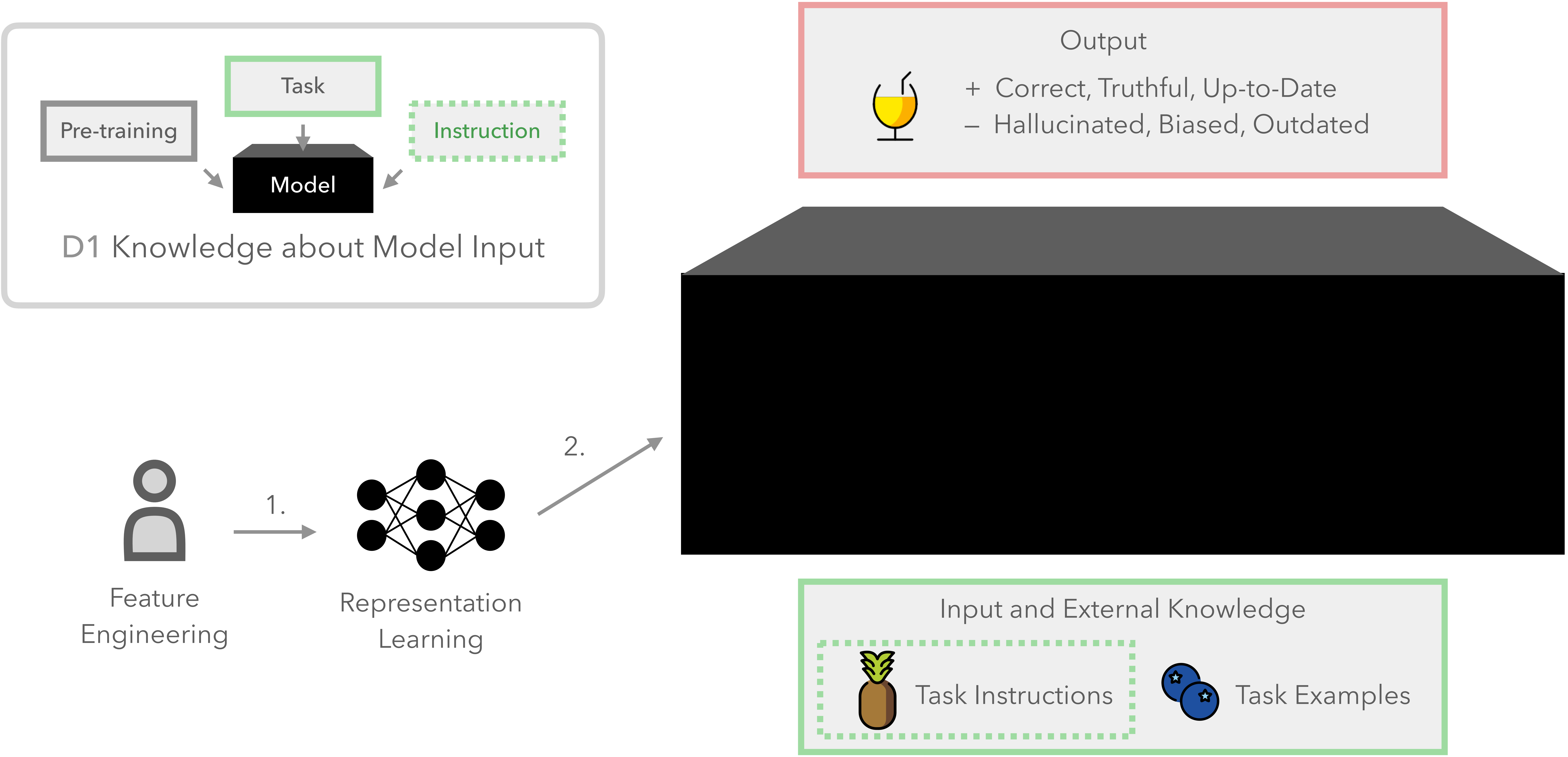
D2 Knowledge about Model Behaviour



D3 Knowledge of Evaluation Protocols



D4 Knowledge of Data Origin



Task

Pre-training

Instruction

Model

D1 Knowledge about Model Input

Output

+ Correct, Truthful, Up-to-Date

– Hallucinated, Biased, Outdated

2.

1.

Feature Engineering

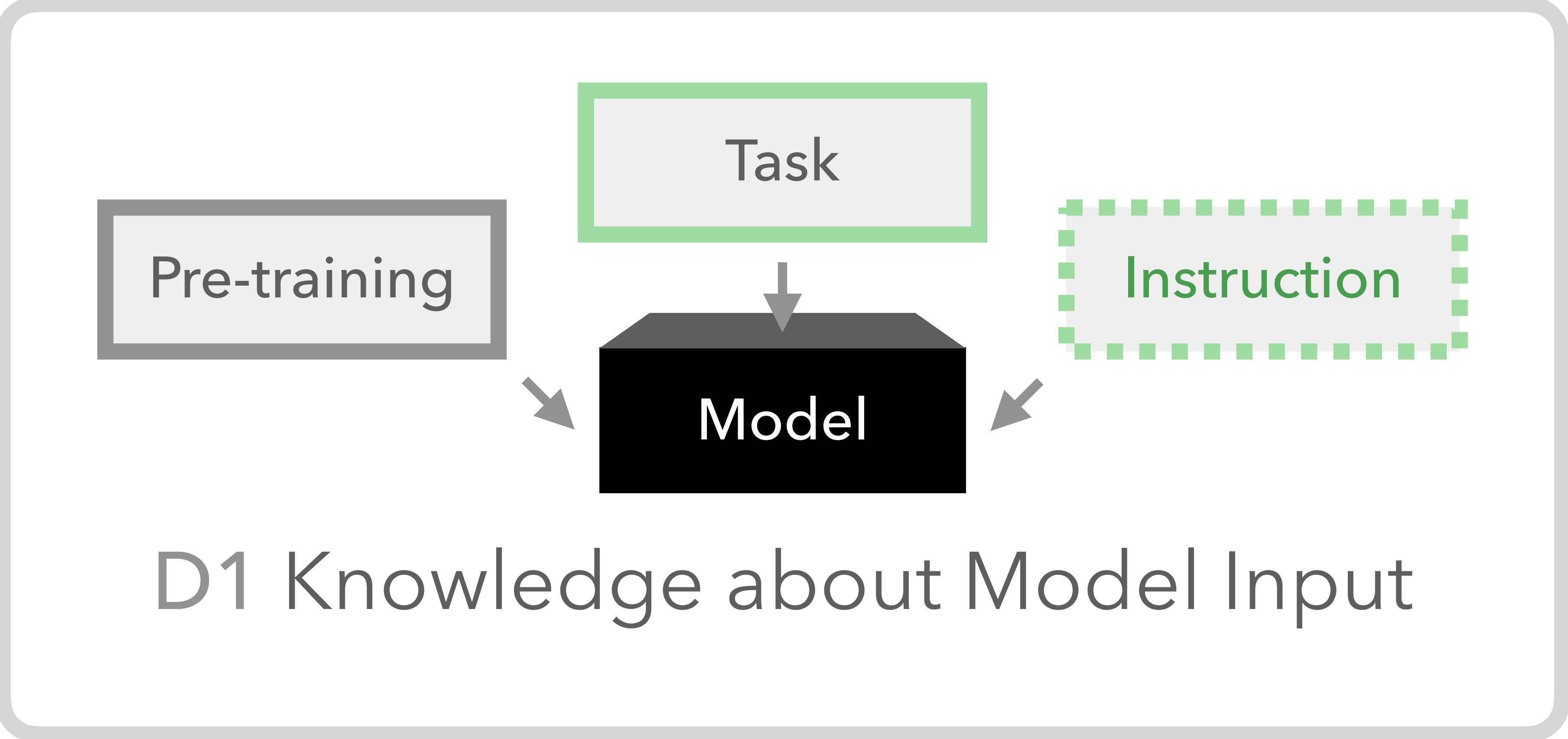
Representation Learning

Input and External Knowledge

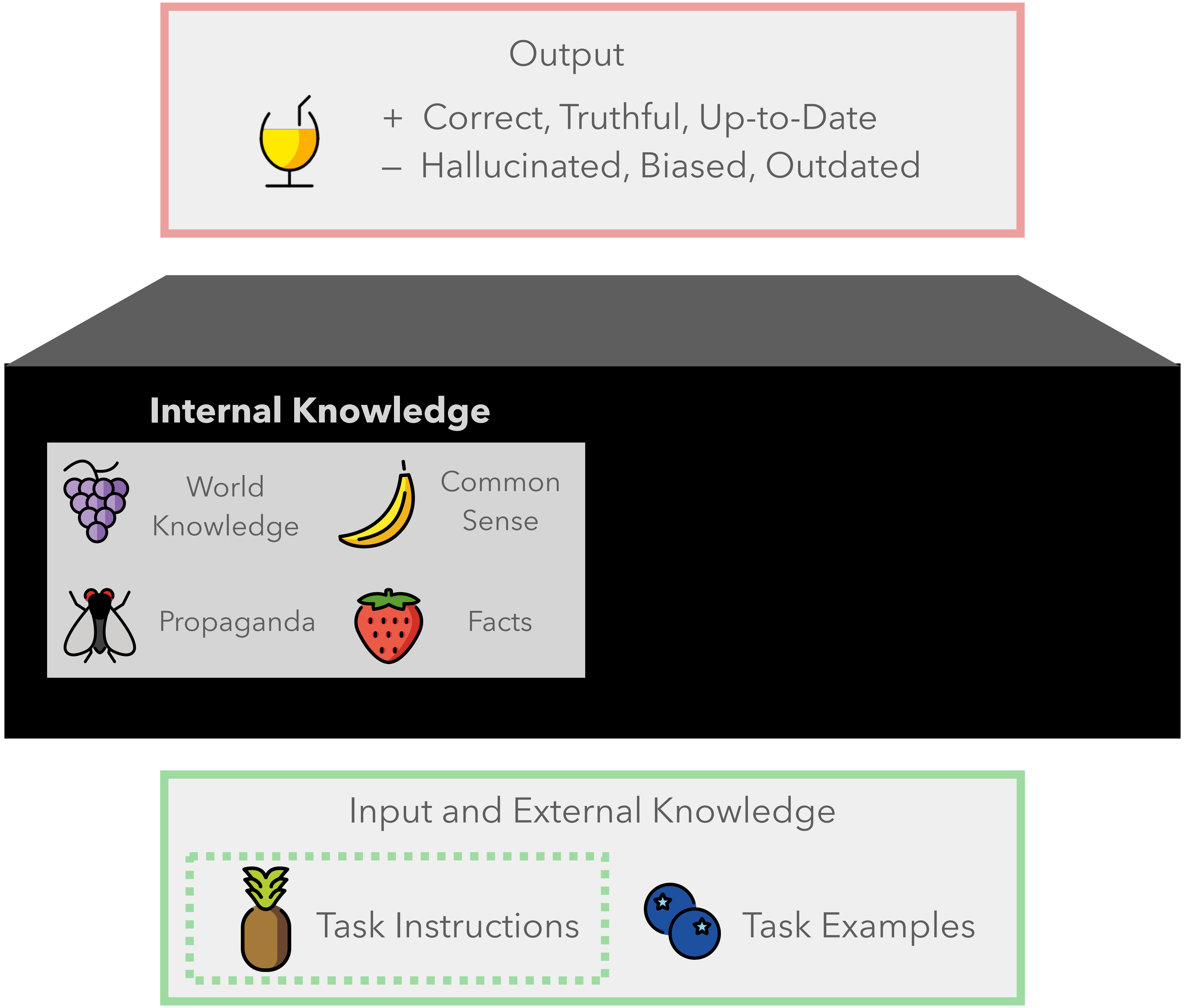
Task Instructions

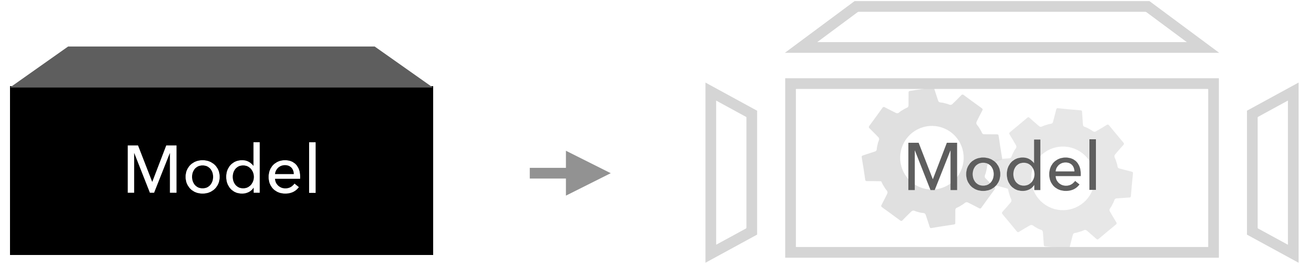
Task Examples

Openmoji, CC BY-SA 4.0



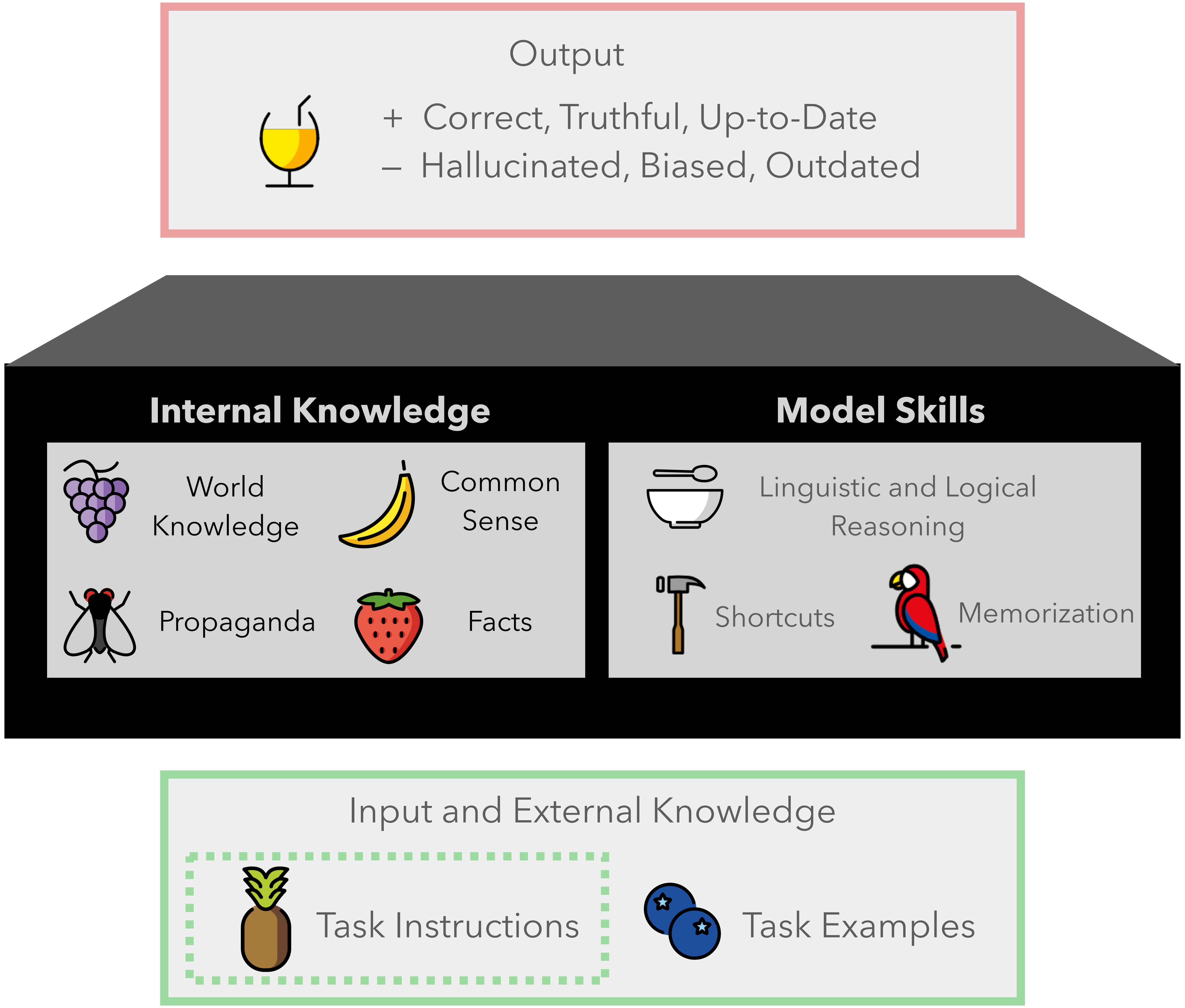
Which internal and external knowledge (ingredients) is used to generate output?





D2 Knowledge about Model Behaviour

What skills are employed to process ingredients into model output?



Output

+ Correct, Truthful, Up-to-Date

– Hallucinated, Biased, Outdated

Internal Knowledge

World Knowledge

Common Sense

Propaganda

Facts

Model Skills

Linguistic and Logical Reasoning

Shortcuts

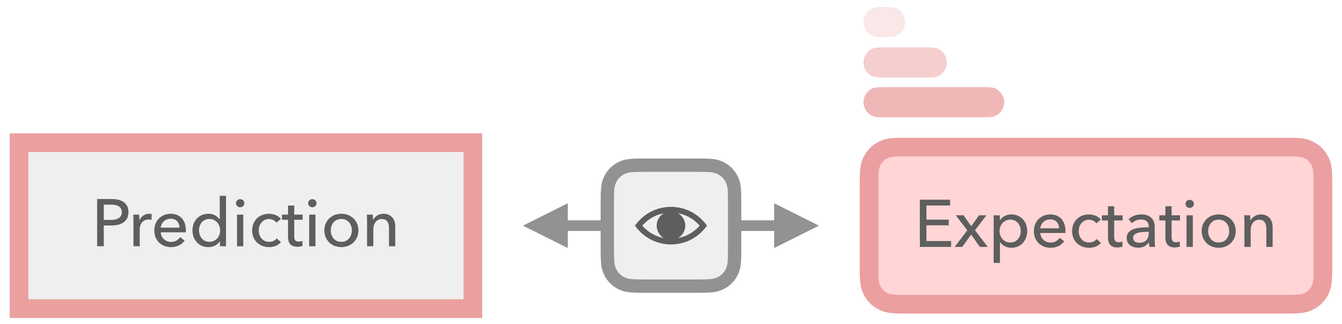
Memorization

Input and External Knowledge

Task Instructions

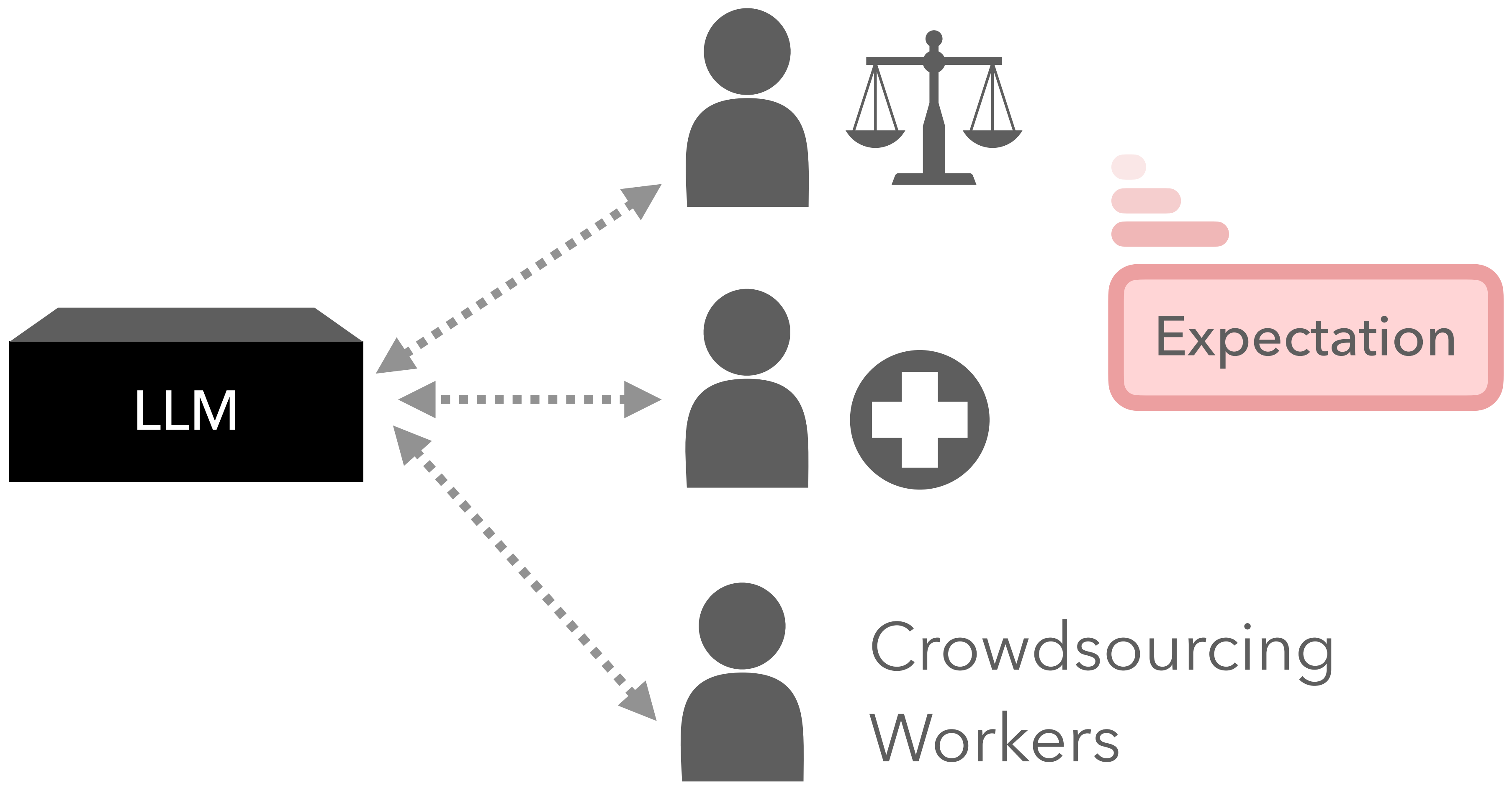
Task Examples

Openmoji, CC BY-SA 4.0

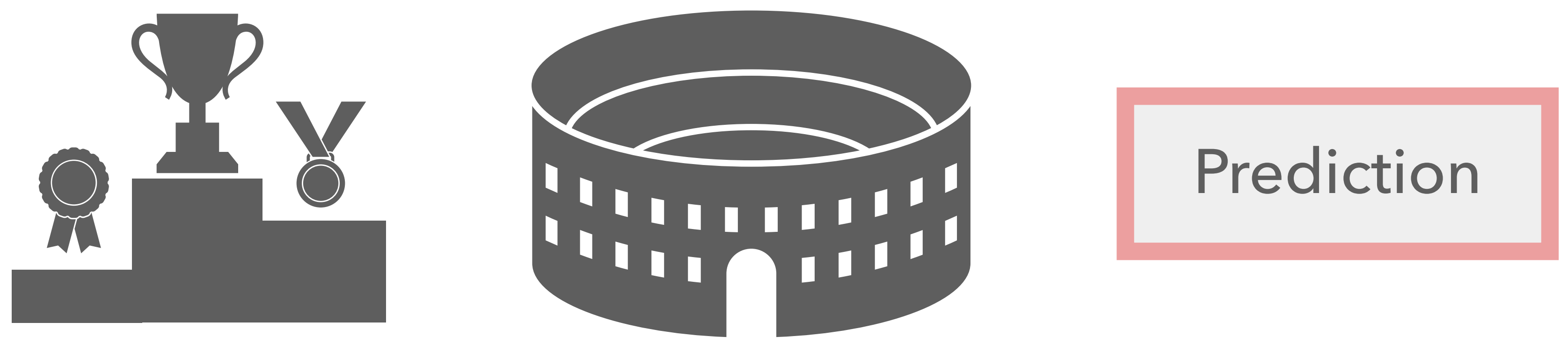


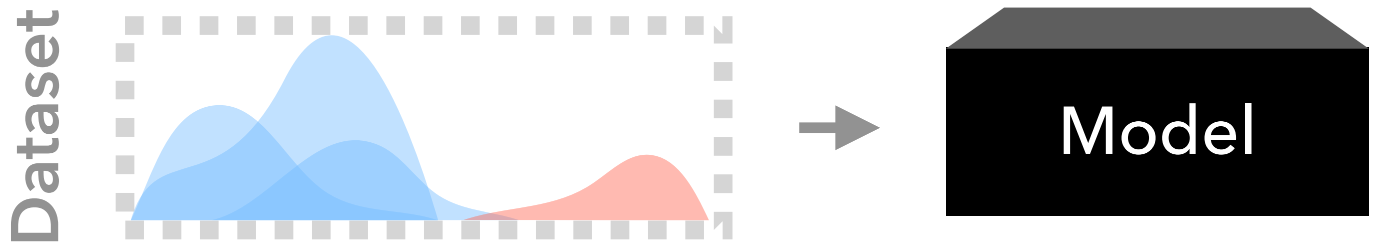
D3 Knowledge of Evaluation Protocols

Baselines: Super-human Performance?



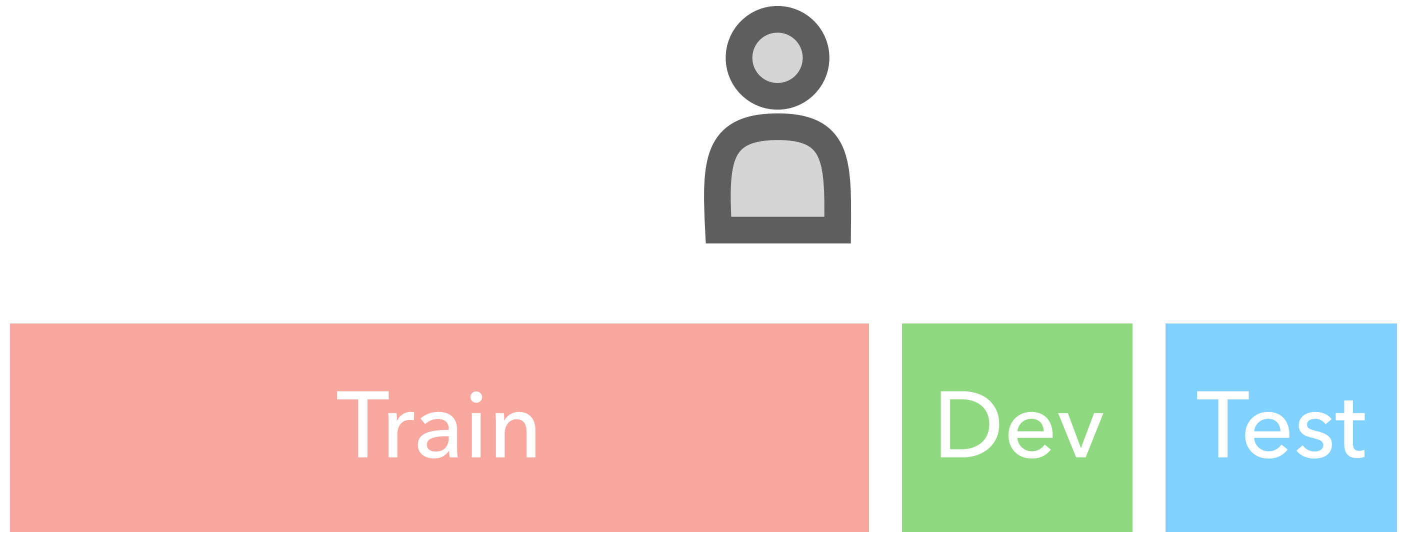
Benchmarking: Leaderboard-style?



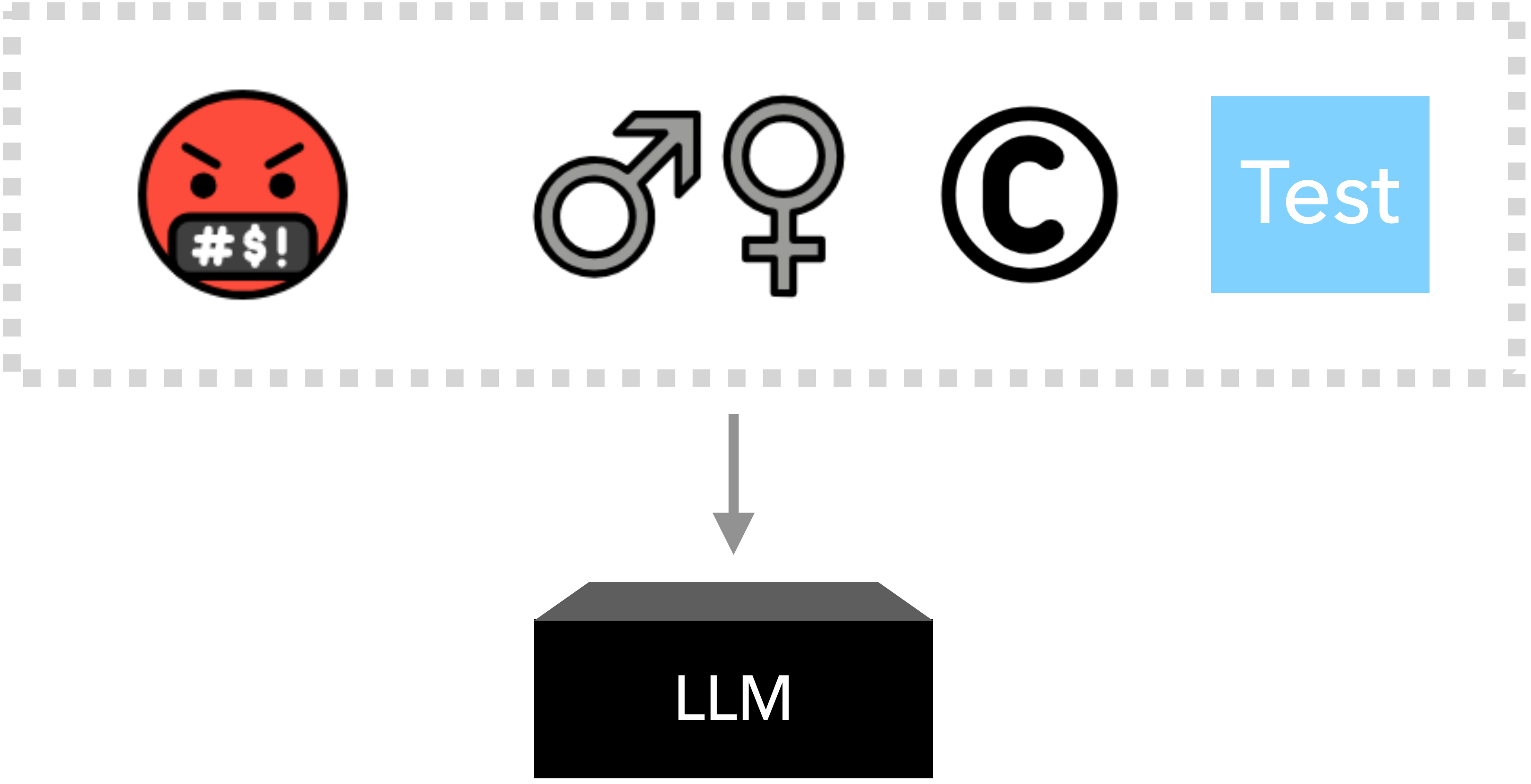


D4 Knowledge of Data Origin

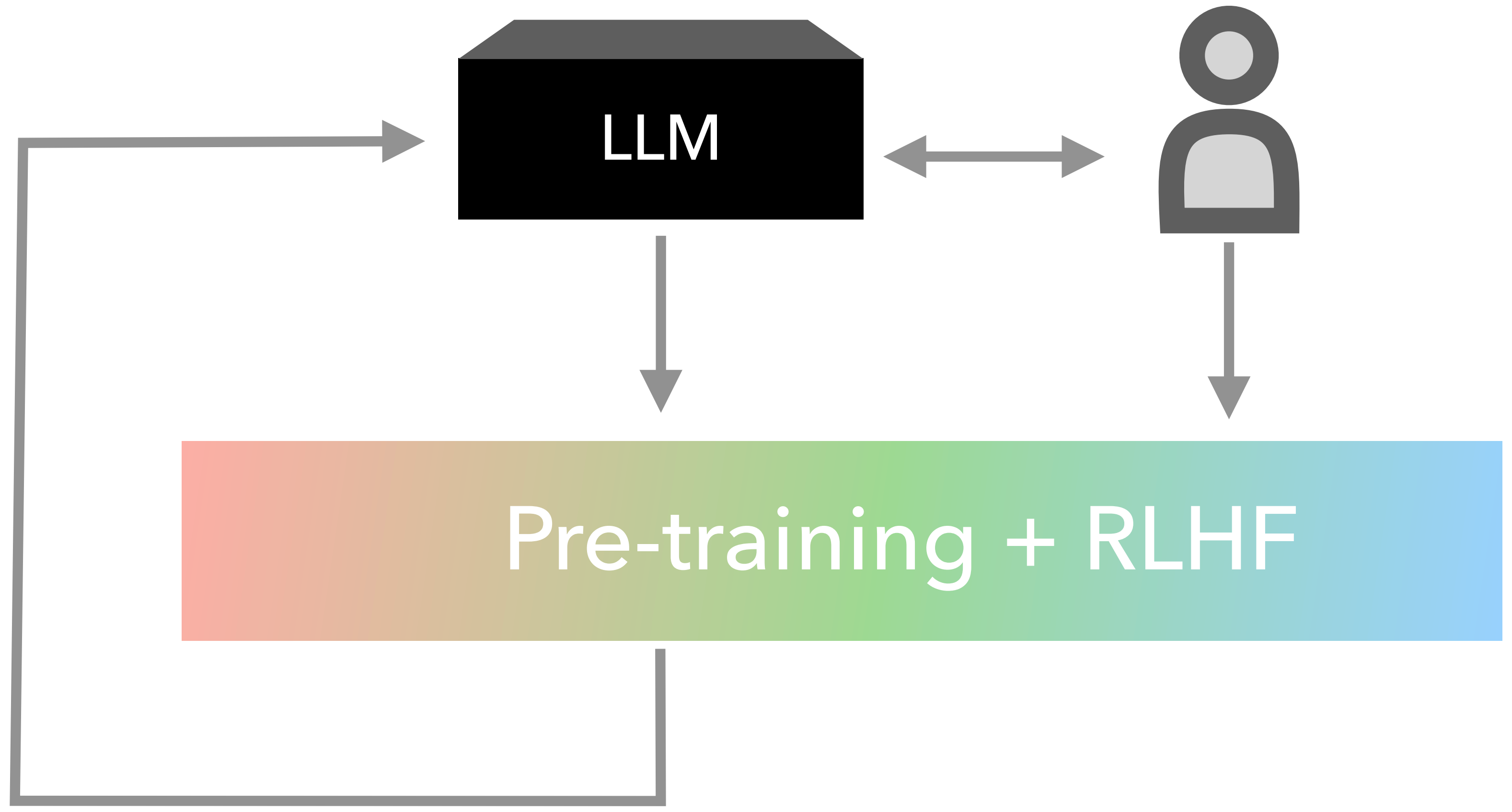
Compartmentalized NLP



Hate Speech, Biases, Copyright?



Current Trend in NLP



Openmoji, CC BY-SA 4.0

Conclusion

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

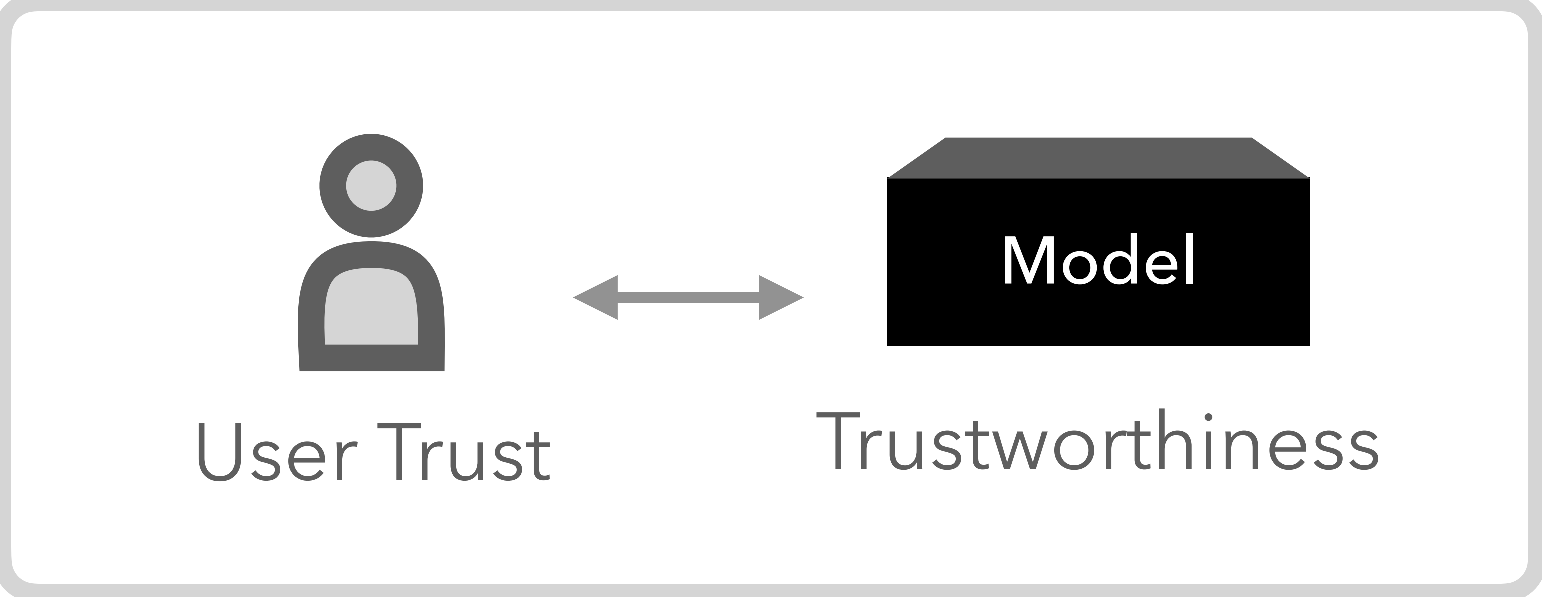
Trustworthiness - Working Definition by David G. Hays, 1979

D1 Knowledge about Model Input

D2 Knowledge about Model Behaviour

D3 Knowledge of Evaluation Protocols

D4 Knowledge of Data Origin



What Can We Do to Gain Trust Now and in Future? (See paper)

- Explain Skills Required versus Skills Employed.
- Facilitate Representative and Comparable Qualitative Analysis.
- Be Explicit about Data Provenance.

Thanks

See you in Singapore!

IT-UNIVERSITETET I KØBENHAVN / north × LMU LUDWIG-MAXIMILIANS-UNIVERSITÄT MÜNCHEN / nlp