Targeting the Benchmark
On Methodology in Current NLP Research

David Schlangen - Dept. Linguistics - University of Potsdam, Germany - david.schlangen@uni-potsdam.de

Application A

Cognitive Capability C

Cognitive Capability C'

...

Set C_L of capabilites of competent language user

Task T
- task description

exemplifies

Dataset D
- collection instructions

Model M
- architecture

Probe P
- probing task

train

test

metric

?

?

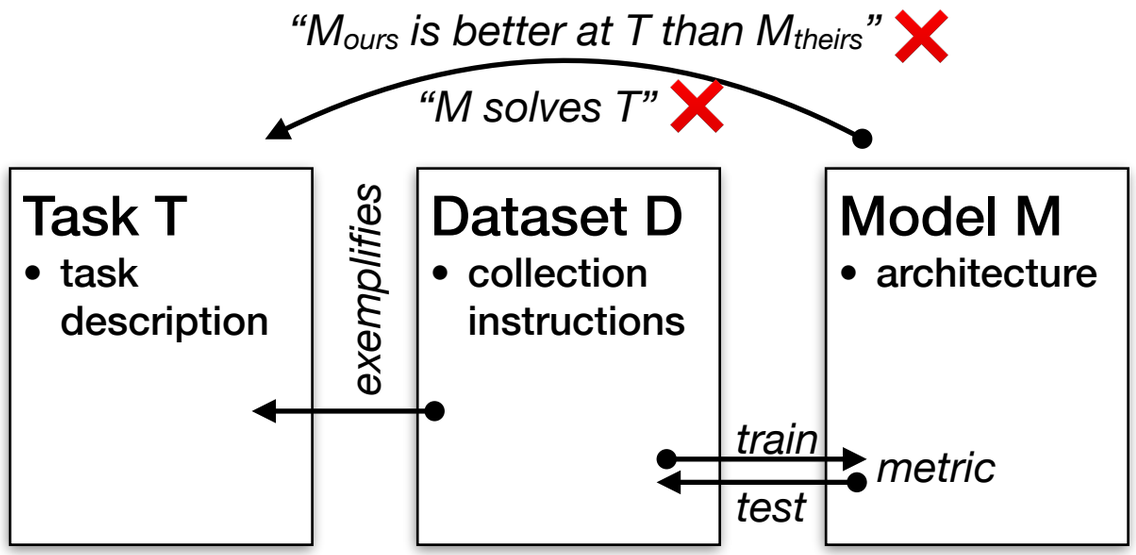
Task T'
- task description

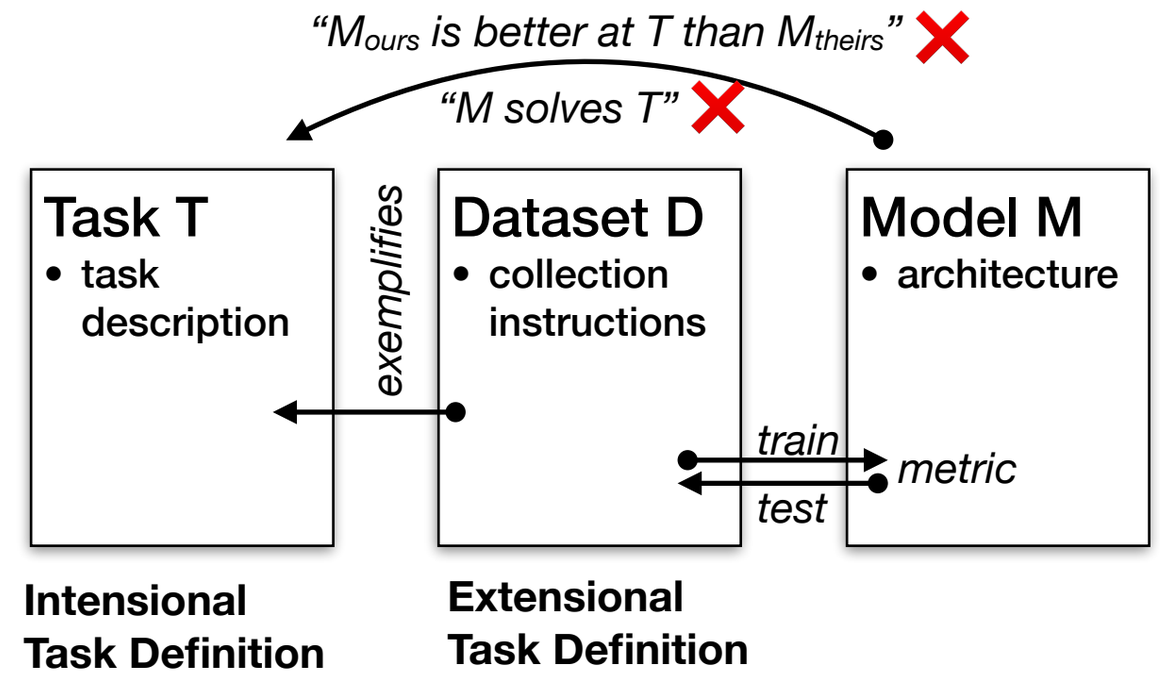
exemplifies

Dataset D'
- collection instructions

optimized for

Model M'
- architecture


“M_{ours} is better at T than M_{theirs}”
“M solves T”
Task T
• task description
exemplifies
Dataset D
• collection instructions
Model M
• architecture
train
test
metric

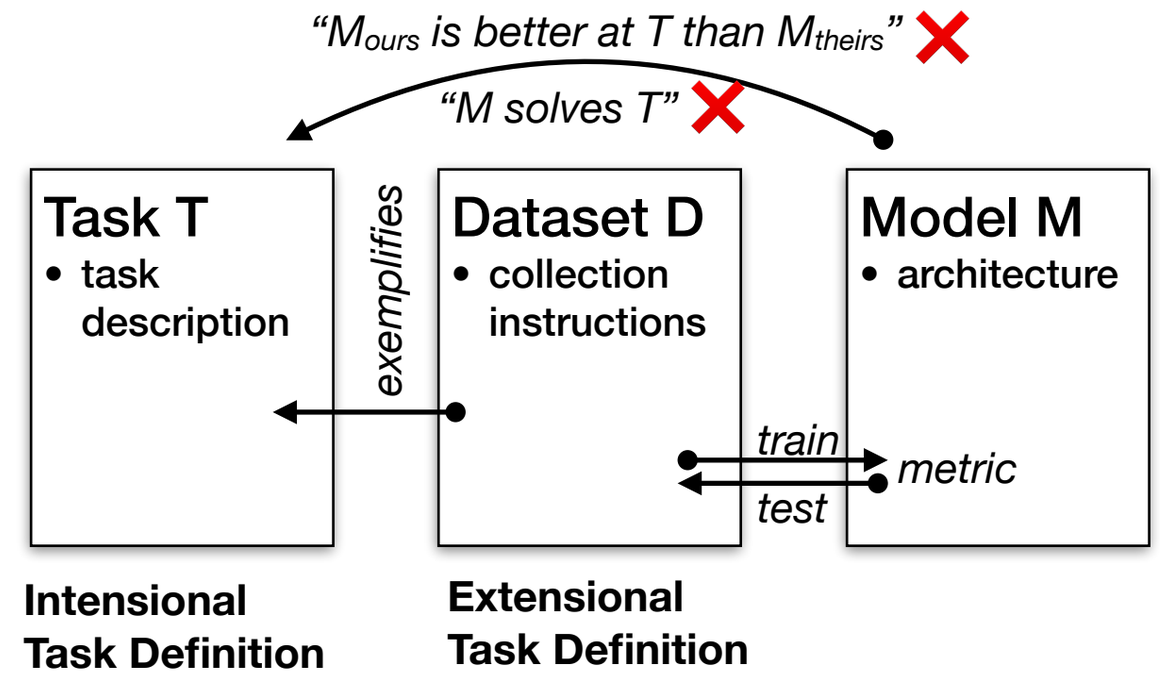


Intensional Task Definition

Image captioning: given image (of any kind), return textual descr. of its content.

Extensional Task Definition

(x, y)



Intensional Task Definition

Image captioning: given image (of any kind), return textual descr. of its content.

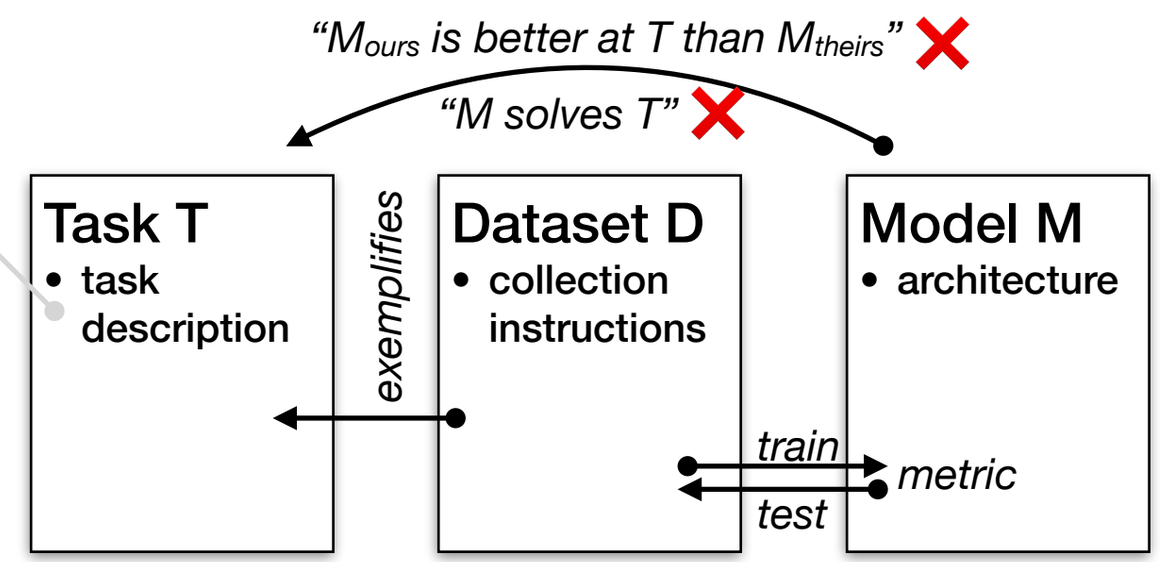
Extensional Task Definition

(x, y)

Quality Control

- *internal verification*
- *validation (how general? exploitable?)*

Application A



Intensional Task Definition

Image captioning: given image (of any kind), return textual descr. of its content.

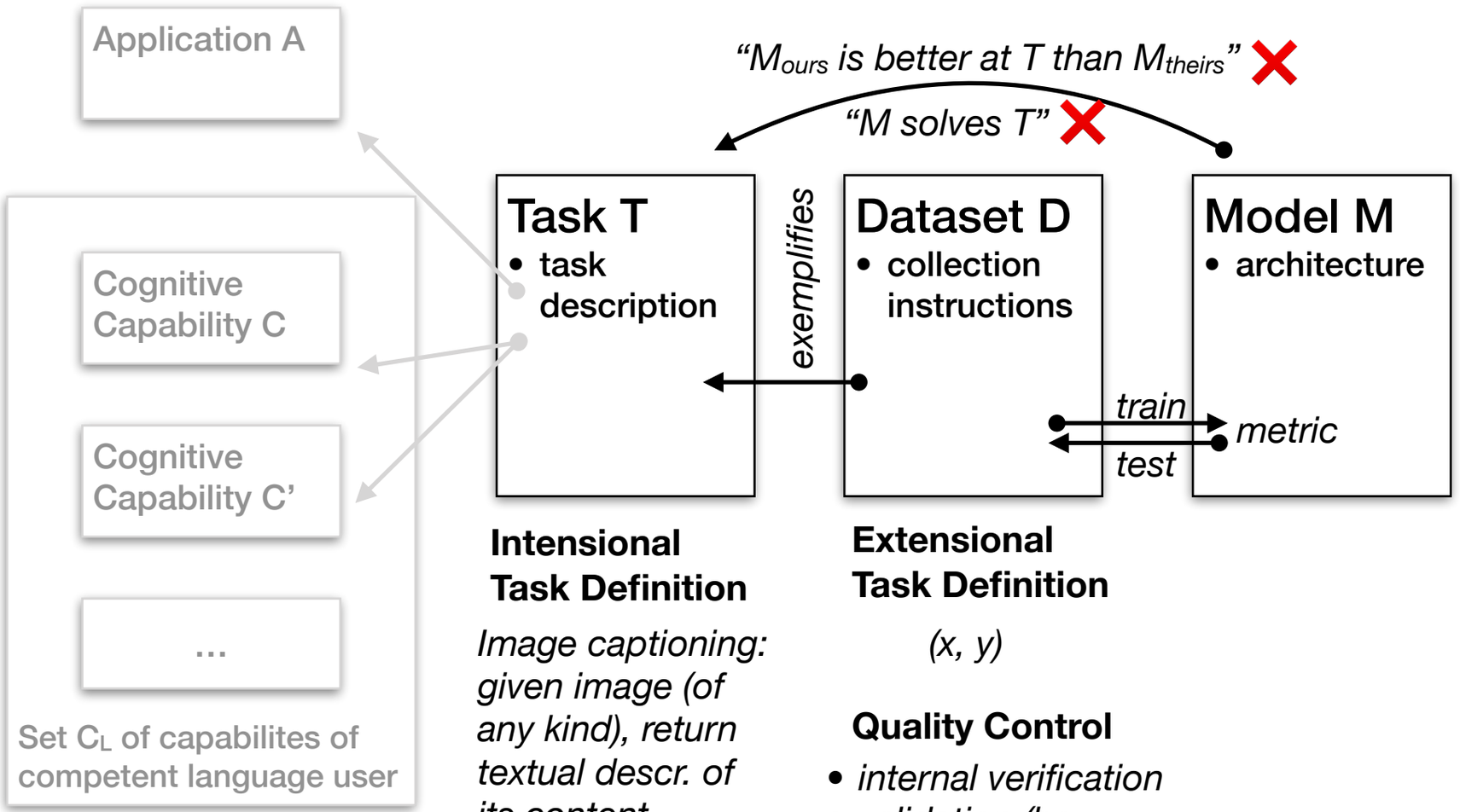
Task Justification

Extensional Task Definition

(x, y)

Quality Control

- *internal verification*
- *validation (how general? exploitable?)*



Observation: Justifications use theoretical language (from psych., ling.), w/o theory.

Intensional Task Definition

Image captioning: given image (of any kind), return textual descr. of its content.

Extensional Task Definition

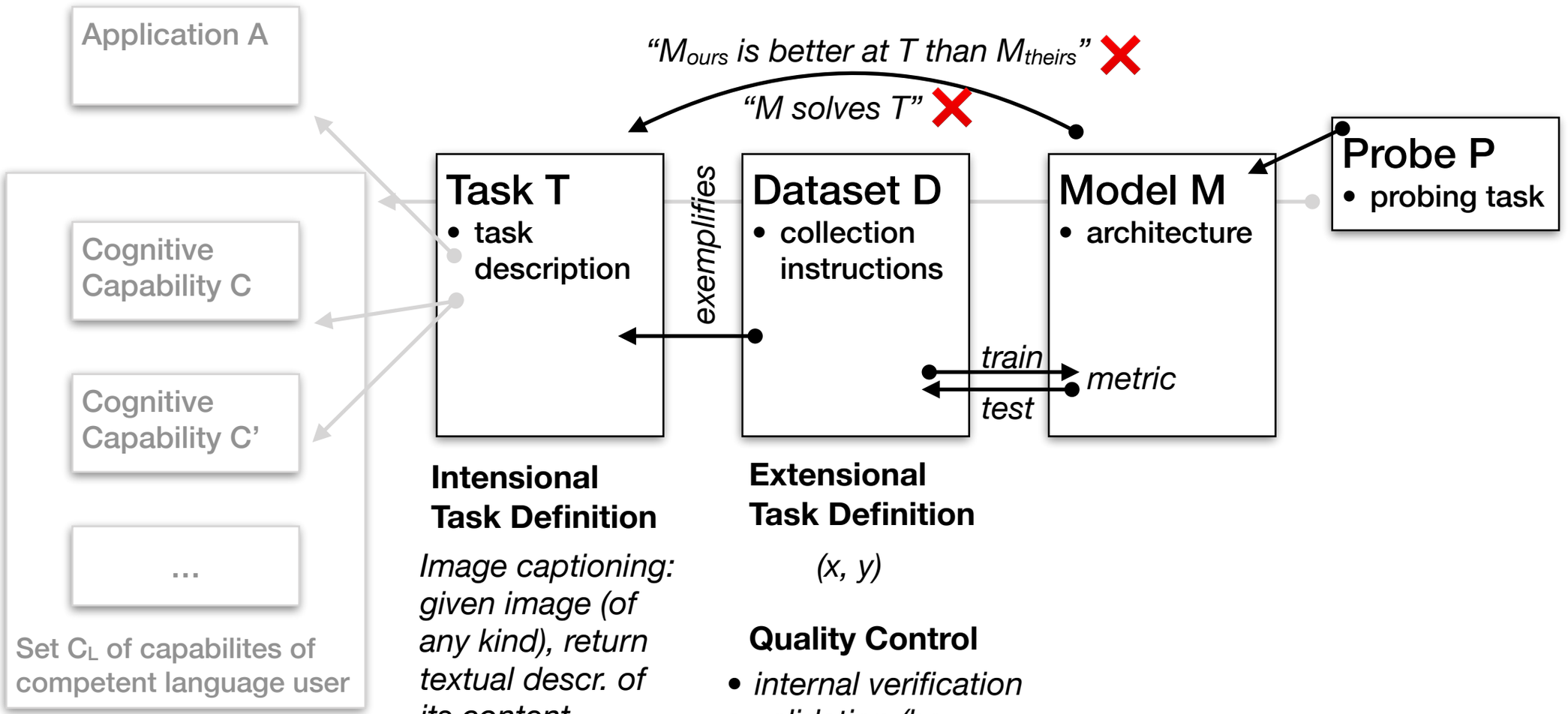
(x, y)

Quality Control

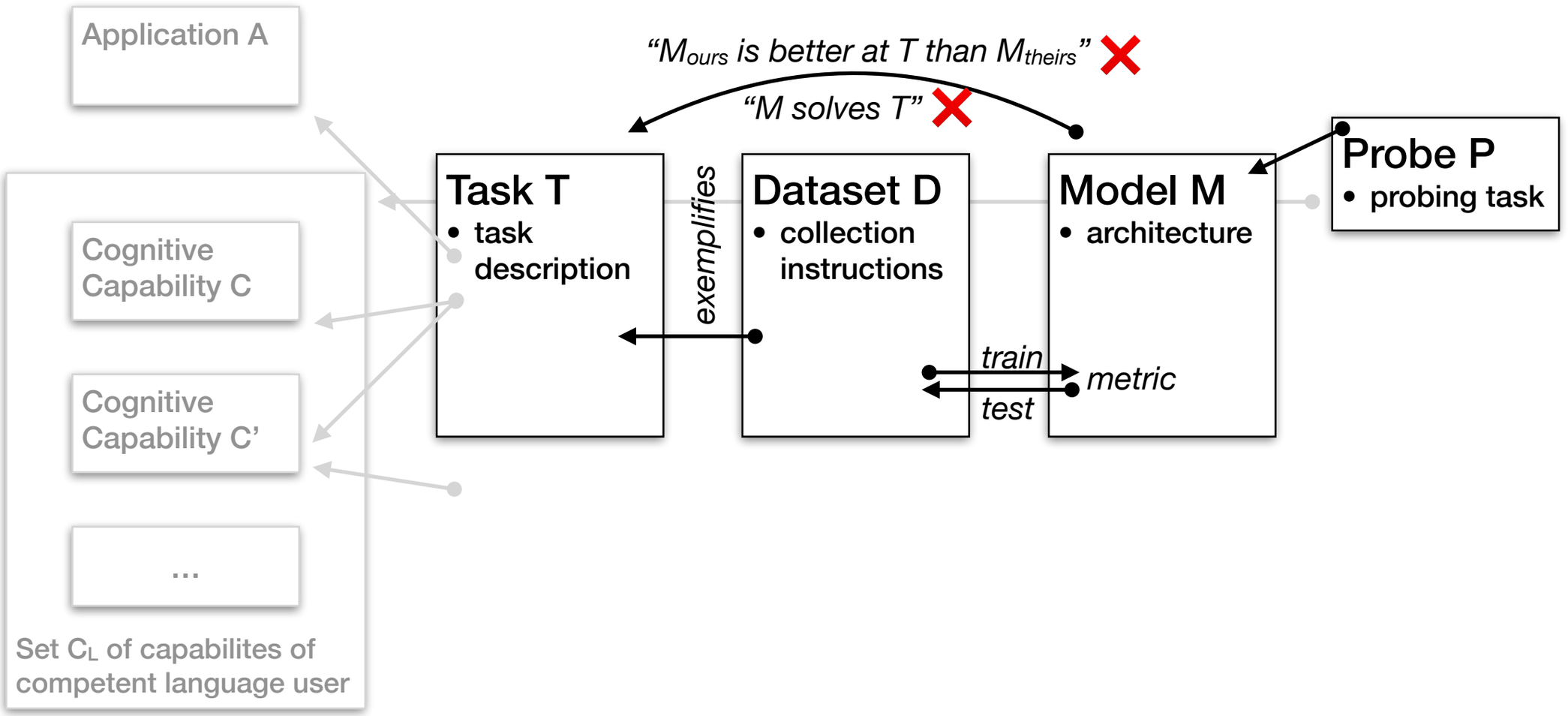
- *internal verification*
- *validation (how general? exploitable?)*

Task Justification

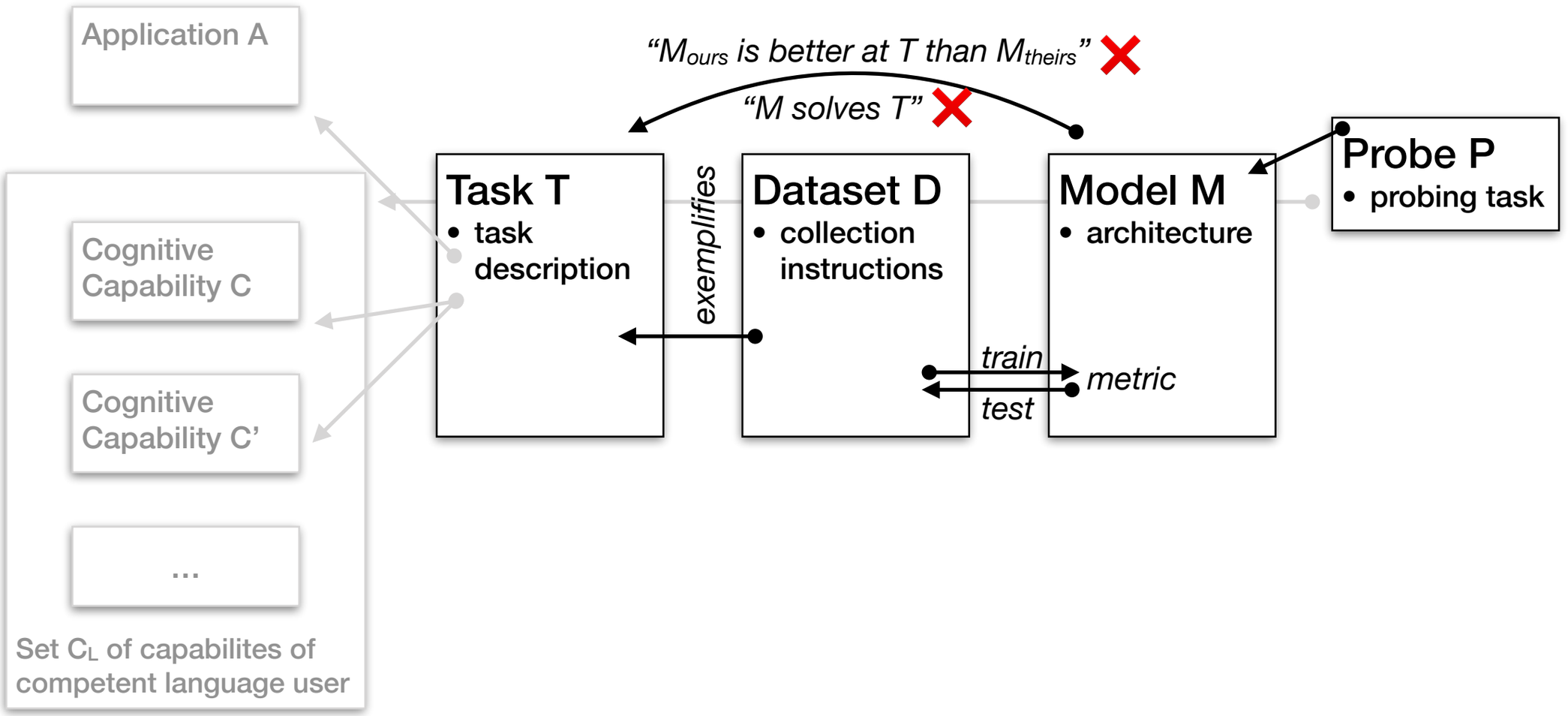
Interesting because it involves *capability C.*



Observation: Justifications use theoretical language (from psych., ling.), w/o theory.

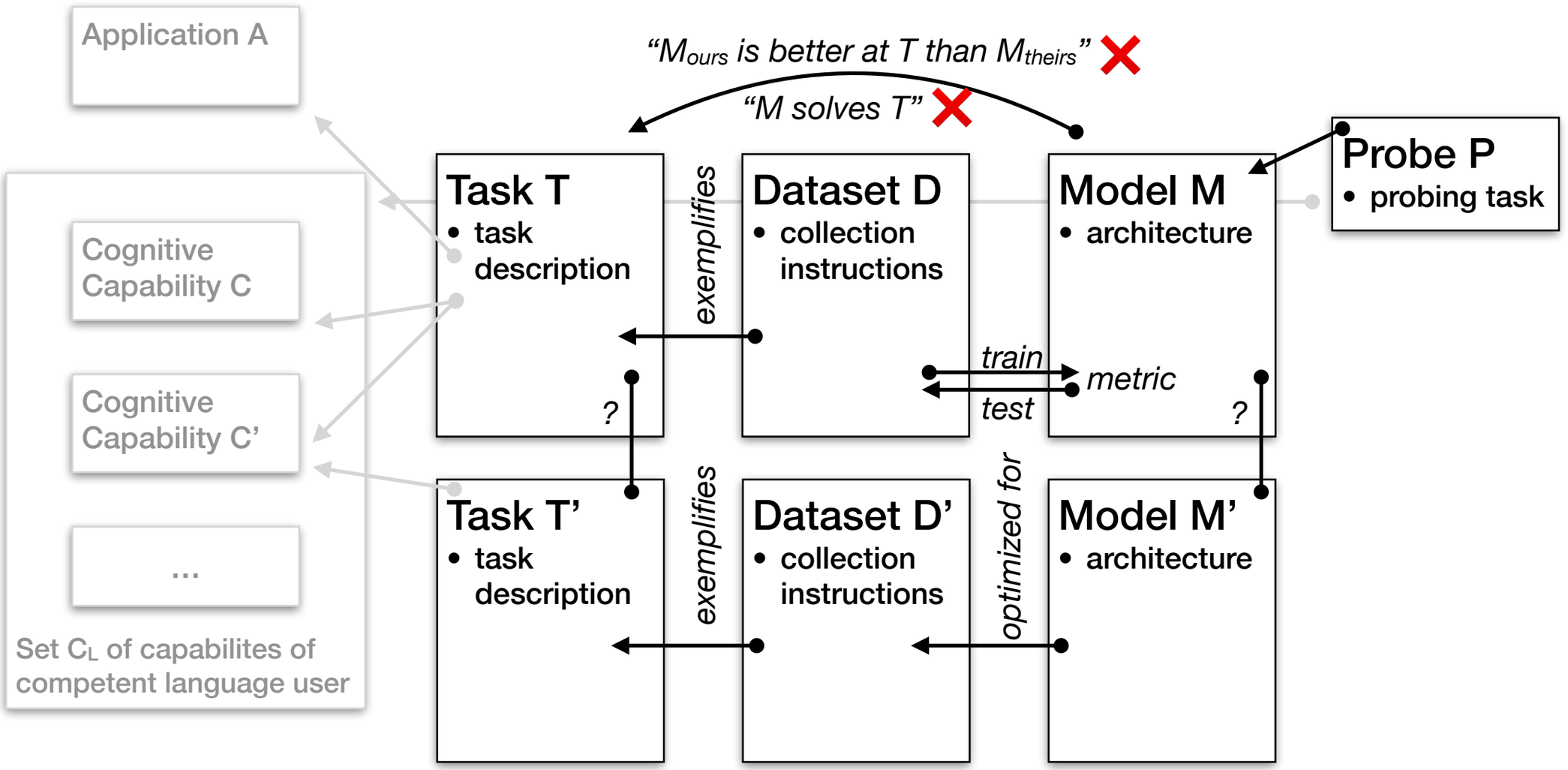


The Task / Dataset pair drives our research, perhaps more so than the Models. It should get appropriate attention.



The Task / Dataset pair drives our research, perhaps more so than the Models. It should get appropriate attention.

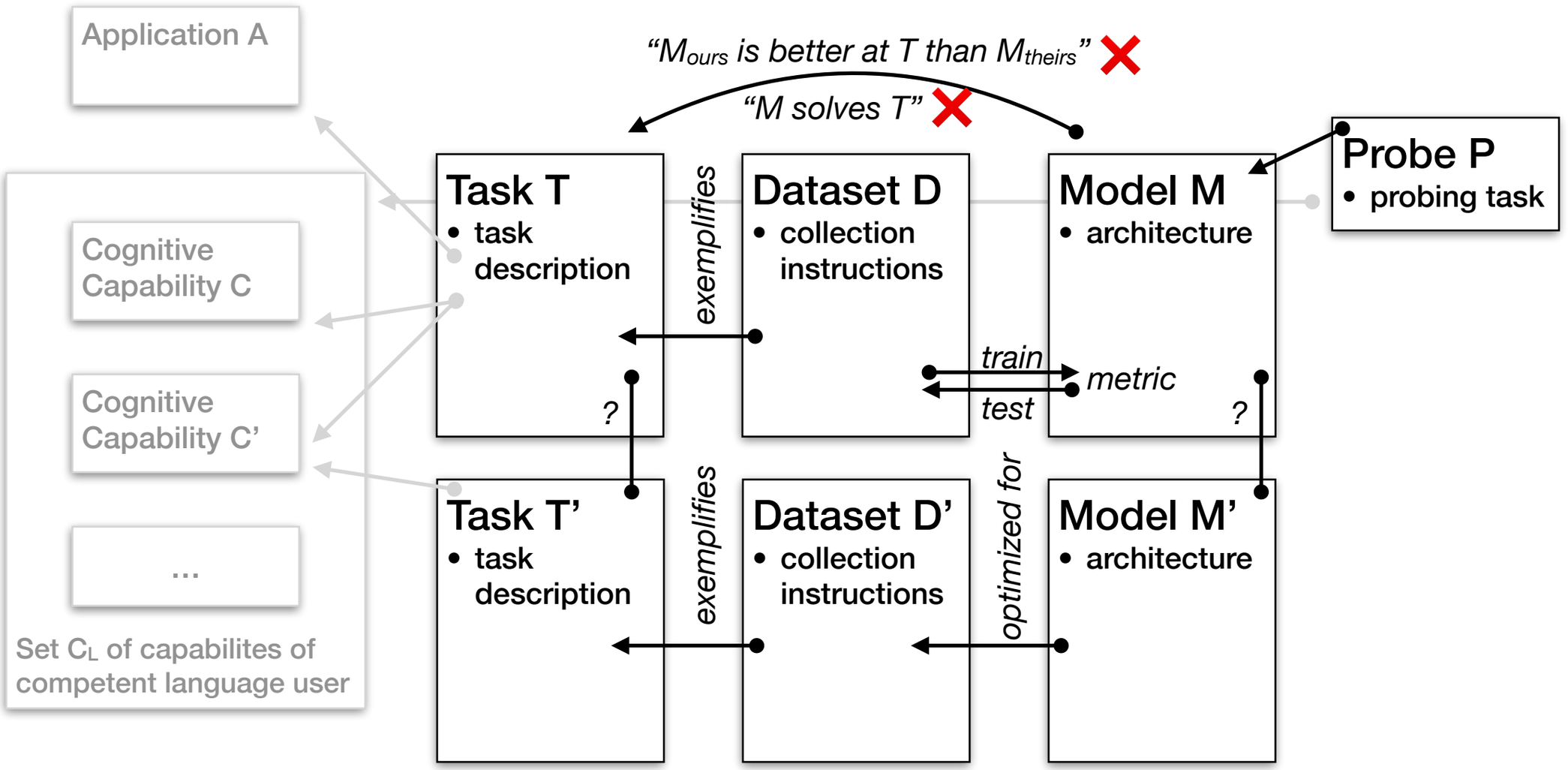
Can we find better ways to ground tasks in considerations about cognitive capabilities and the composition of the language faculty?
(Or a principled argument for why we needn’t care?)



The Task / Dataset pair drives our research, perhaps more so than the Models. It should get appropriate attention.

Can we find better ways to ground tasks in considerations about cognitive capabilities and the composition of the language faculty?
(Or a principled argument for why we needn't care?)

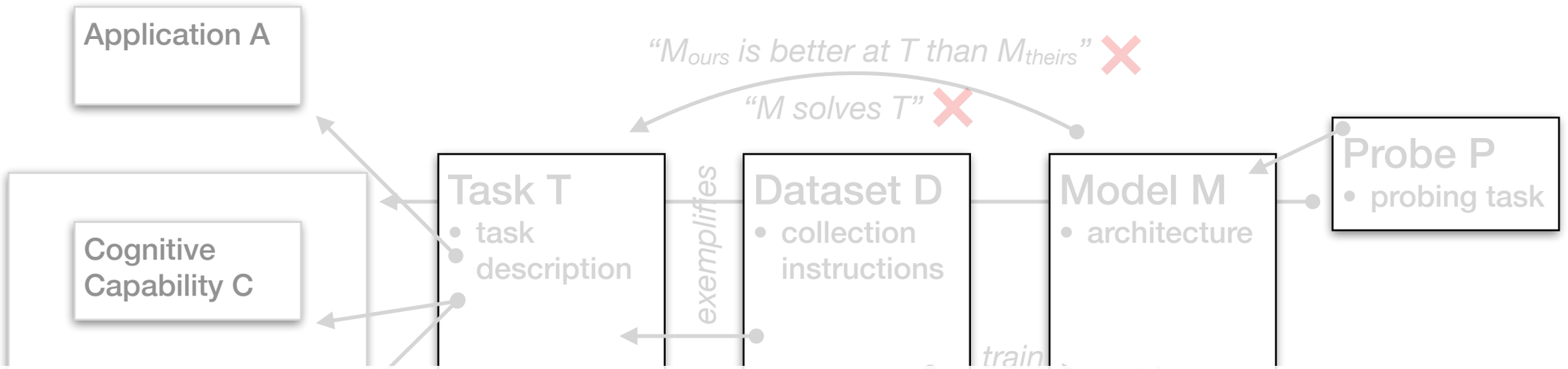
Can we make tasks *predictive*, so that performance of M at T tells us something about performance at T'?



The Task / Dataset pair drives our research, perhaps more so than the Models. It should get appropriate attention.

Can we find better ways to ground tasks in considerations about cognitive capabilities and the composition of the language faculty?
(Or a principled argument for why we needn't care?)

Can we make tasks *predictive*, so that performance of M at T tells us something about performance at T'?



Targeting the Benchmark

On Methodology in Current NLP Research

David Schlangen - Dept. Linguistics - University of Potsdam, Germany - david.schlangen@uni-potsdam.de

The Task / Dataset pair drives our research, perhaps more so than the Models. It should get appropriate attention.

Can we find better ways to ground tasks in considerations about cognitive capabilities and the composition of the language faculty?
(Or a principled argument for why we needn’t care?)

Can we make tasks *predictive*, so that performance of M at T tells us something about performance at T’?