Implicit Representations of Meaning in Neural Language Models

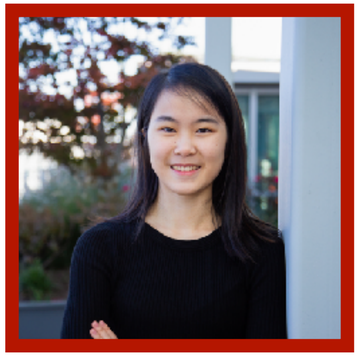
Jacob Andreas

MIT CSAIL

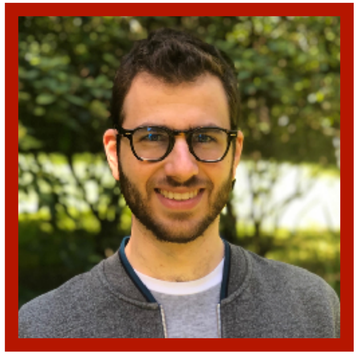
LINGO.csail.mit.edu

Implicit Representations of Meaning in Neural Language Models

[ACL 2021]



Belinda Li

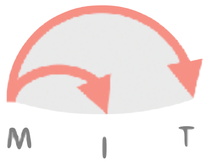


Max Nye

Language comprehension

Janet and Penny went to the store to get presents for Jack. Janet said, "I will buy Jack a top." "Don't get Jack a top," said Penny. "He has a top. He will

[Brown et al. 2020; example from Marcus & Davis 2020 / Charniak 1972]

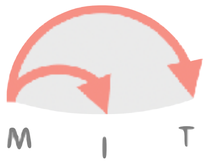


Language comprehension

Janet and Penny went to the store to get presents for Jack. Janet said, "I will buy Jack a top." "Don't get Jack a top," said Penny. "He has a top. He will

`get a top." "I will get Jack a top," said Janet.`

[Brown et al. 2020; example from Marcus & Davis 2020 / Charniak 1972]

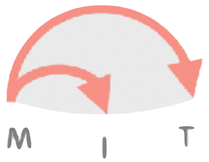


Language comprehension

Janet and Penny went to the store to get presents for Jack. Janet said, "I will buy Jack a top." "Don't get Jack a top," said Penny. "He has a top. He will

get a top." "I will get Jack a top," said Janet.

[Brown et al. 2020; example from Marcus & Davis 2020 / Charniak 1972]



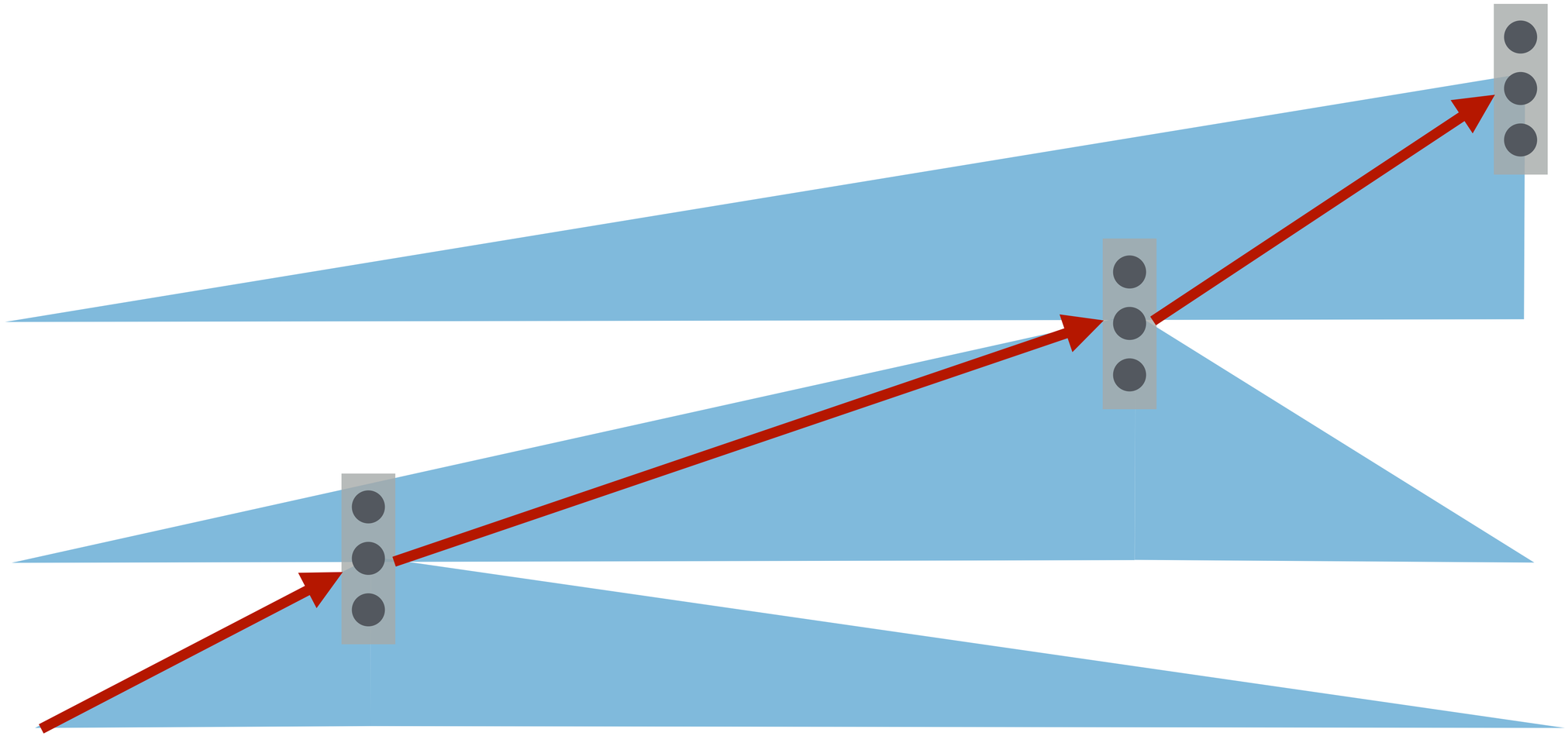
Language comprehension???

Janet and Penny went to the store to get presents for Jack. Janet said, "I will buy Jack a top." "Don't get Jack a top," said Penny. "He has a top. He will

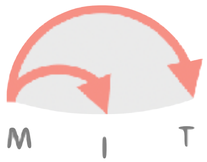
get a top." "I will get Jack a top," said Janet.

[Brown et al. 2020; example from Marcus & Davis 2020 / Charniak 1972]

Neural sequence models



John has a book. Mary has an apple. He gave her his

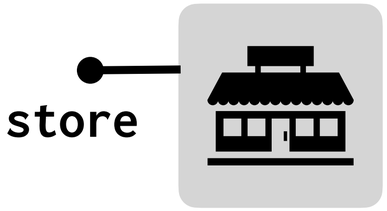


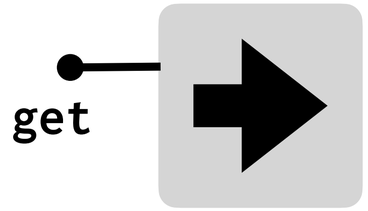
Modeling the world described by language

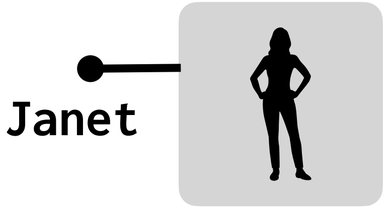
Janet went to the store to get Jack a top.

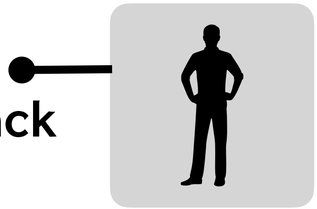
Modeling the world described by language

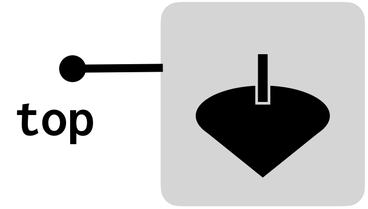
Janet went to the store to get Jack a top.





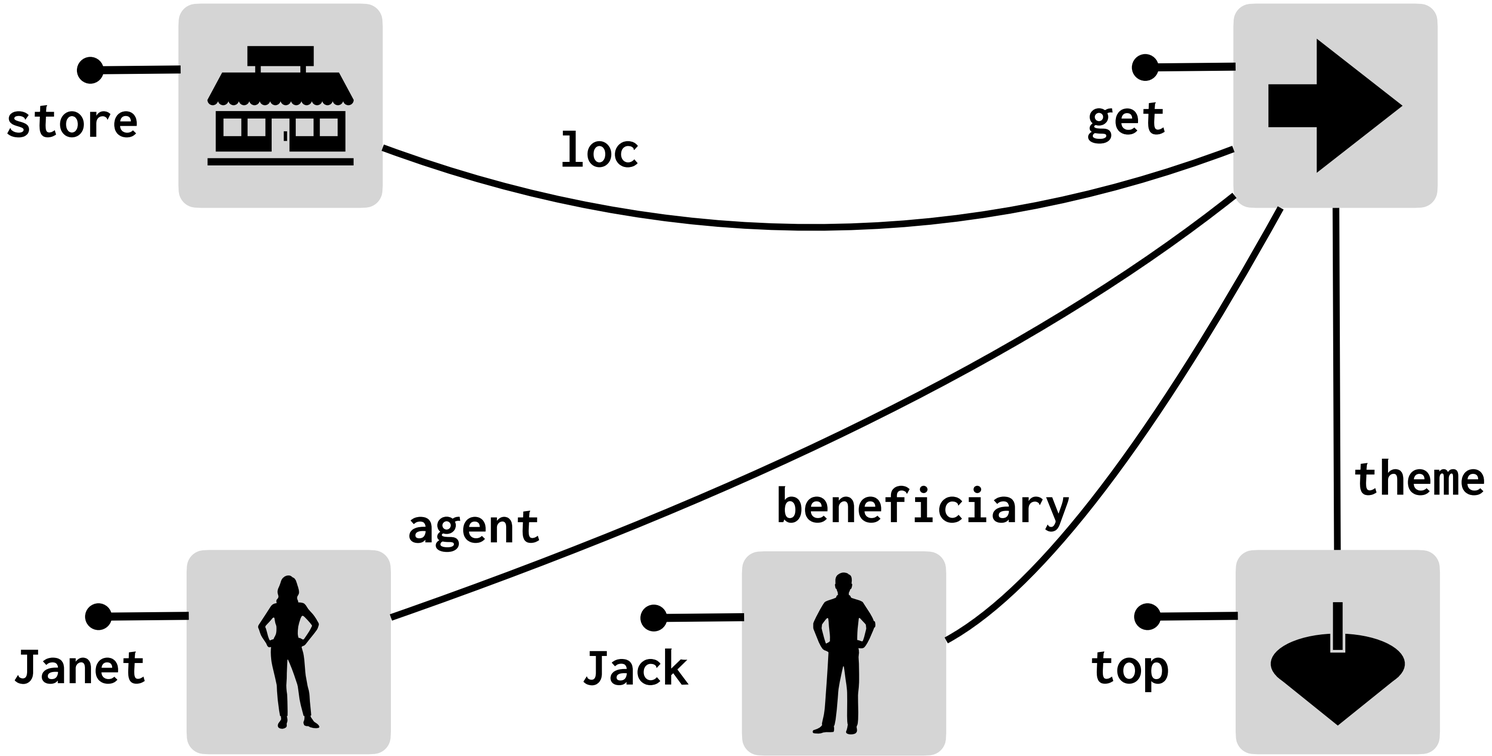






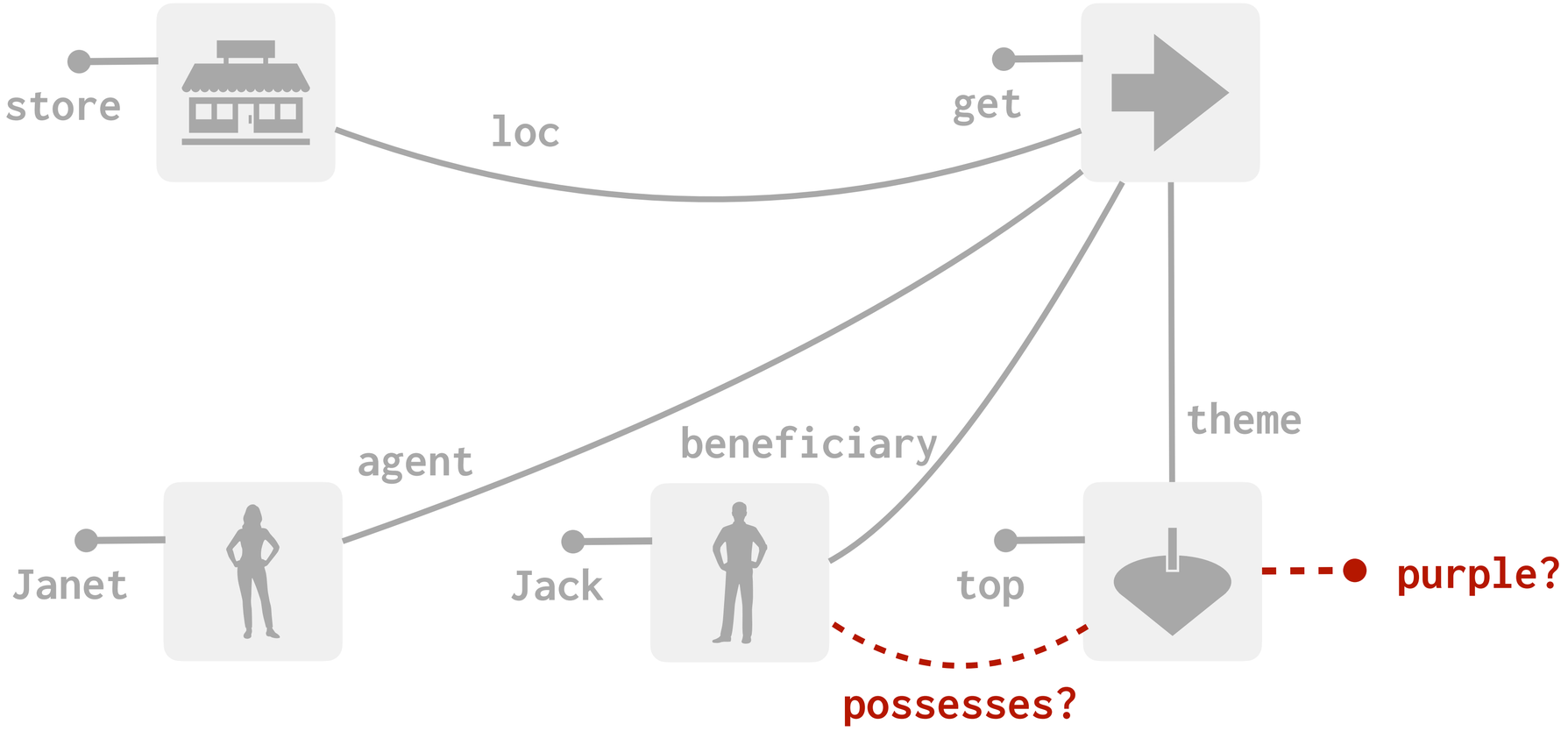
Modeling the world described by language

Janet went to the store to get Jack a top.



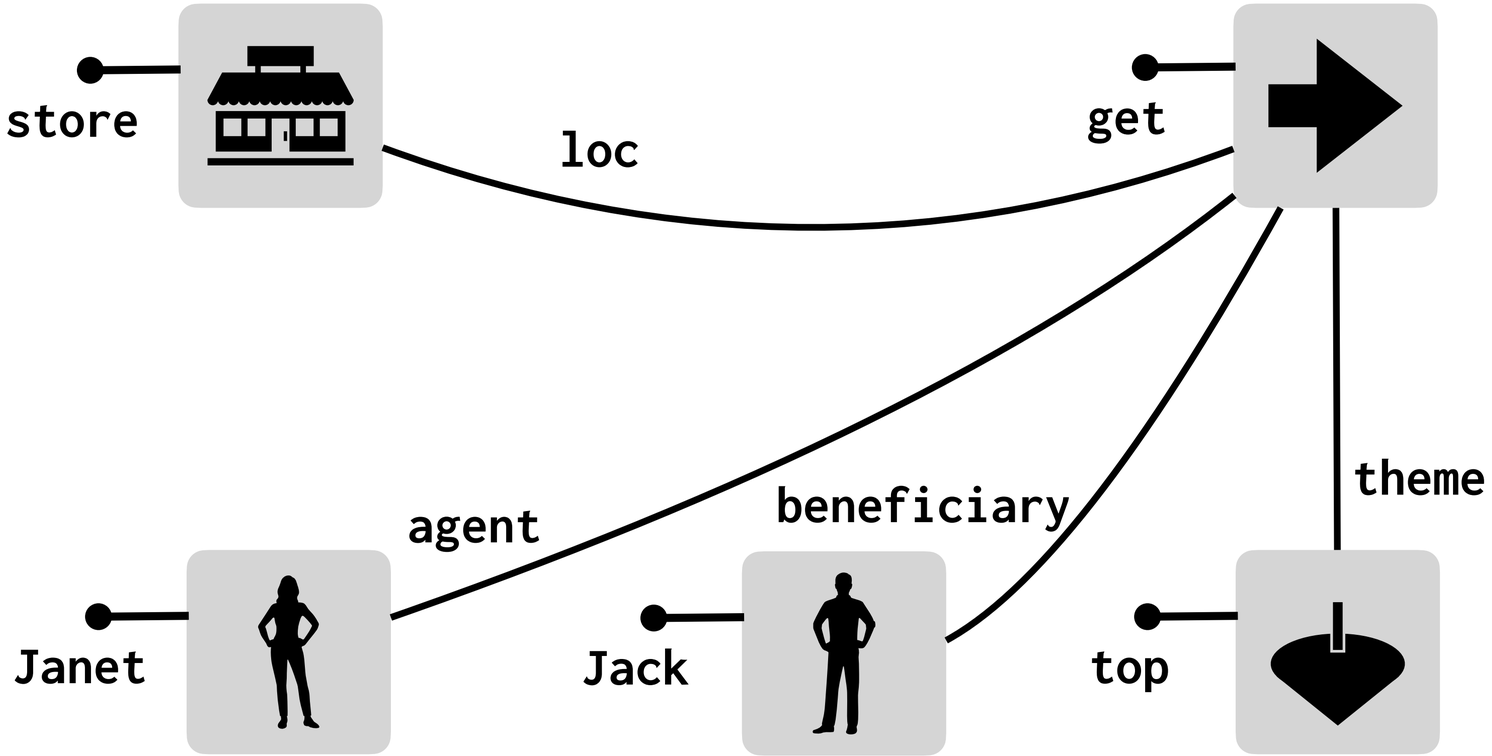
Modeling the world described by language

Janet went to the store to get Jack a top.



Modeling the world described by language

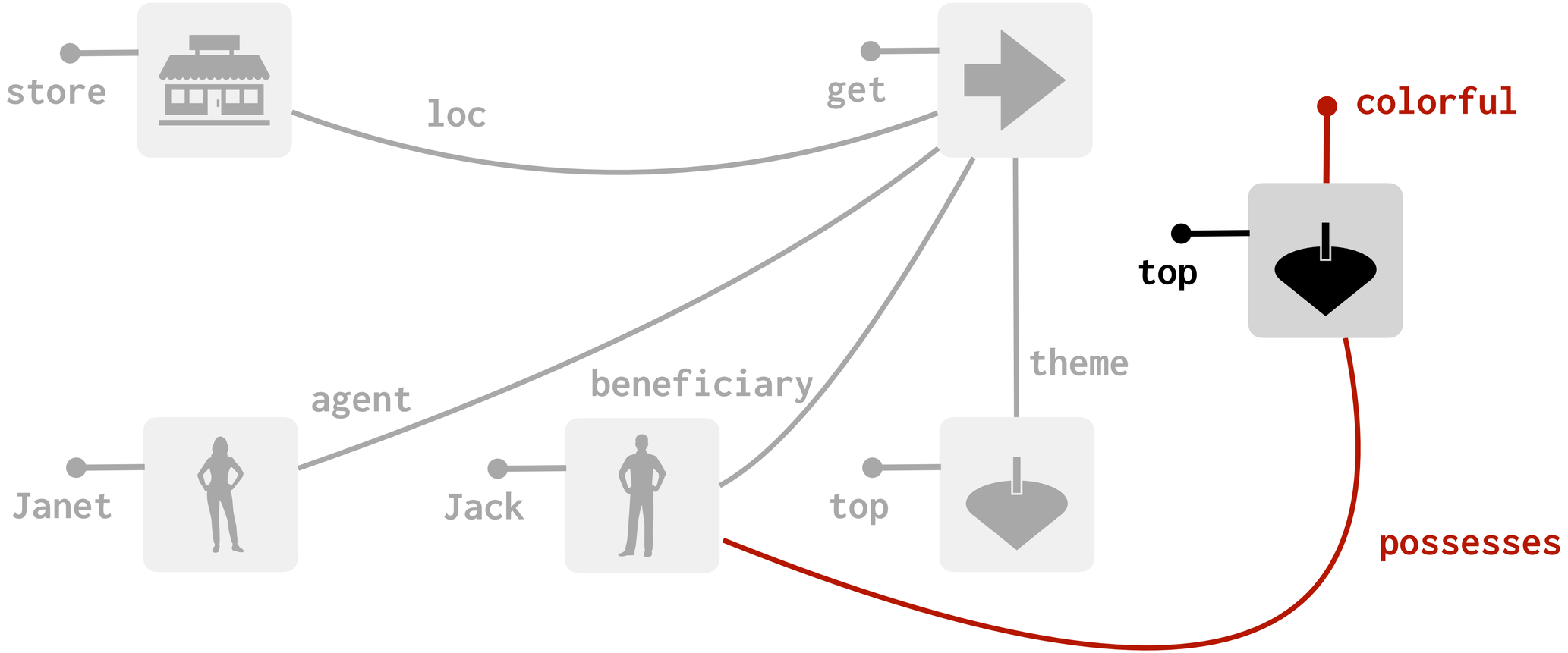
Janet went to the store to get Jack a top.



Modeling the world described by language

Janet went to the store to get Jack a top.

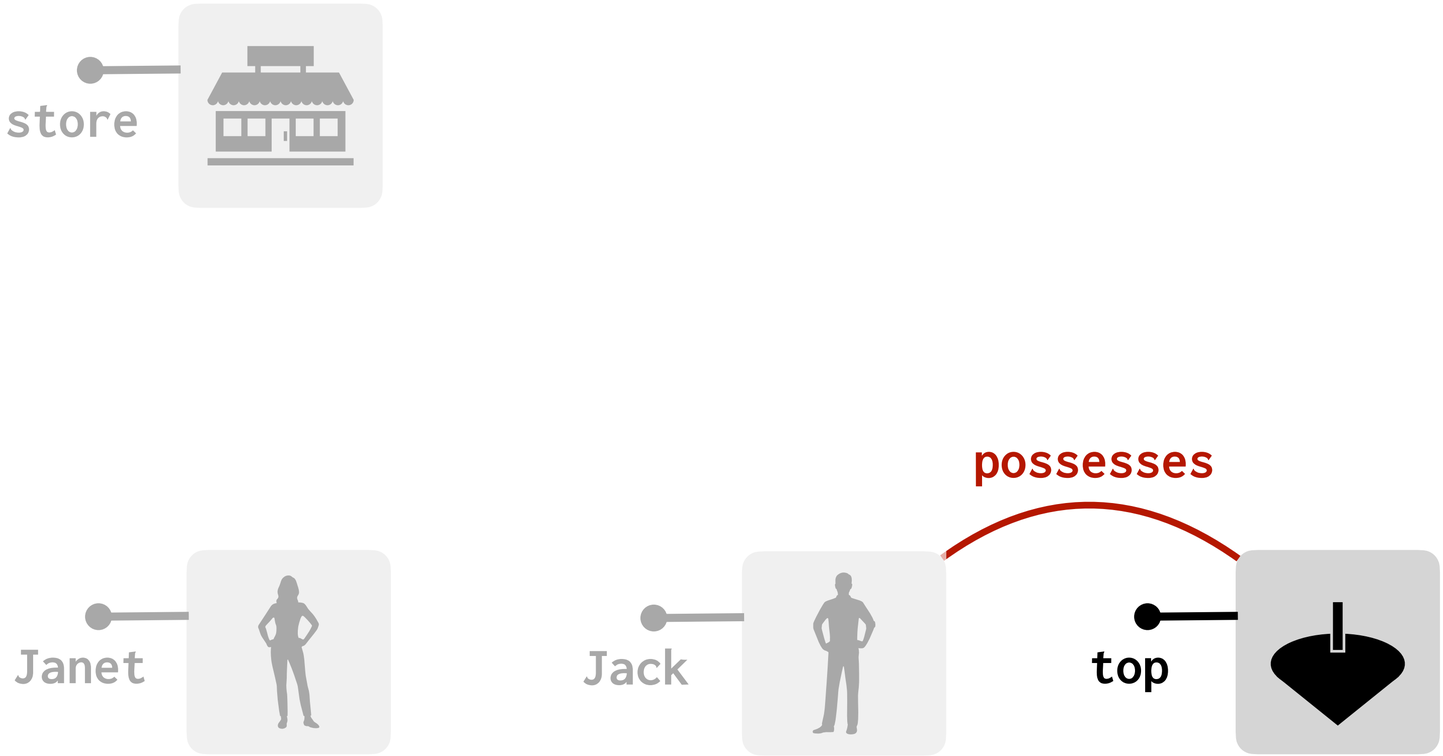
But Jack already has a colorful top.



Modeling the world described by language

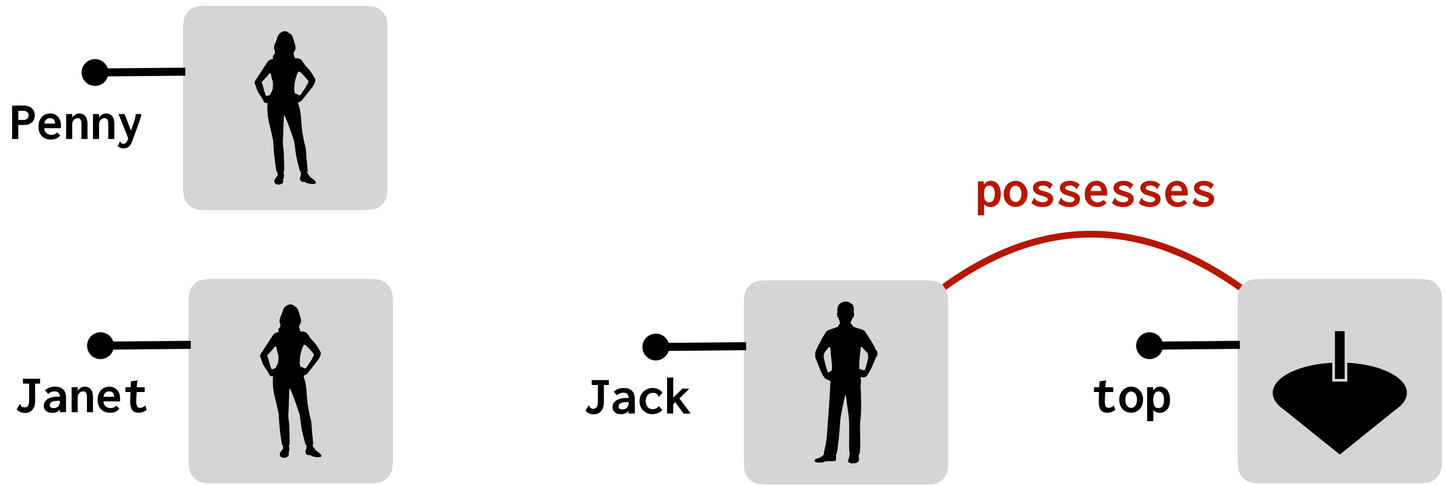
Janet went to the store to get Jack a top.

She gave it to him.



Dynamic Semantics

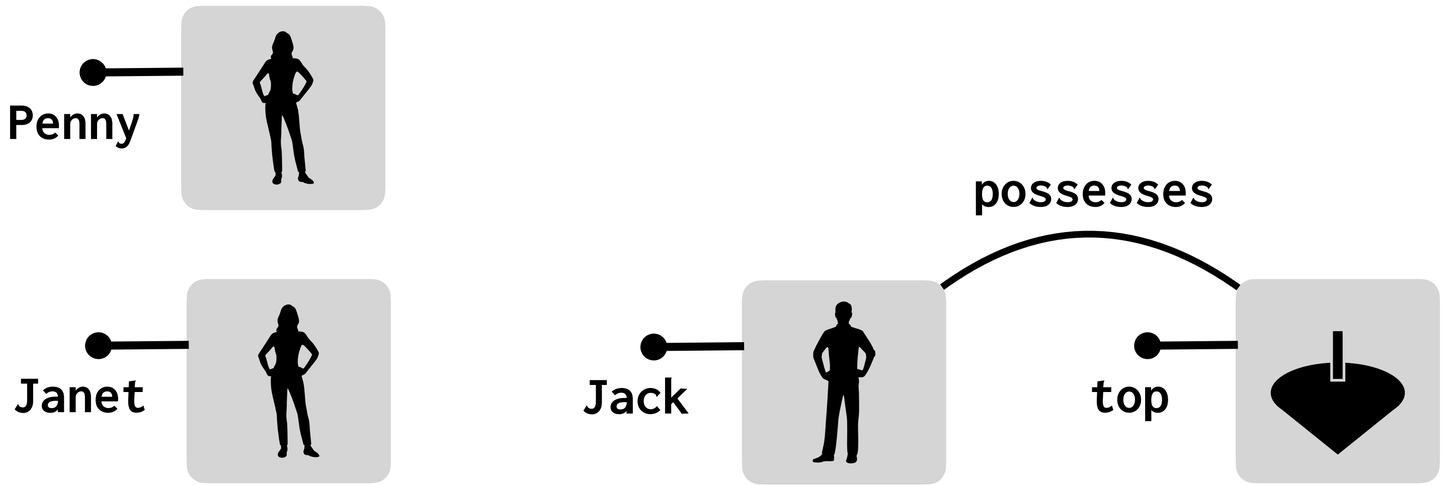
[Heim 83, “File Change Semantics”;
Kamp 81, “Discourse Representation Theory”;
Groenendijk & Stokhof 91, “Dynamic Predicate Logic”]



World models & language models

Janet and Penny went to the store to get Jack a top.
But Jack already has a colorful top.

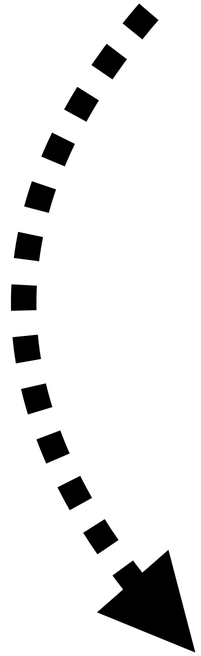
p(“Jack will get a top” | …)

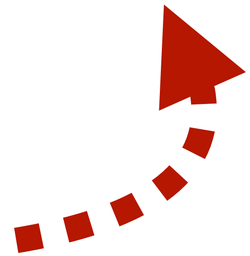


Representations in language models

Janet and Penny went to the store to get Jack a top.
But Jack already has a colorful top.

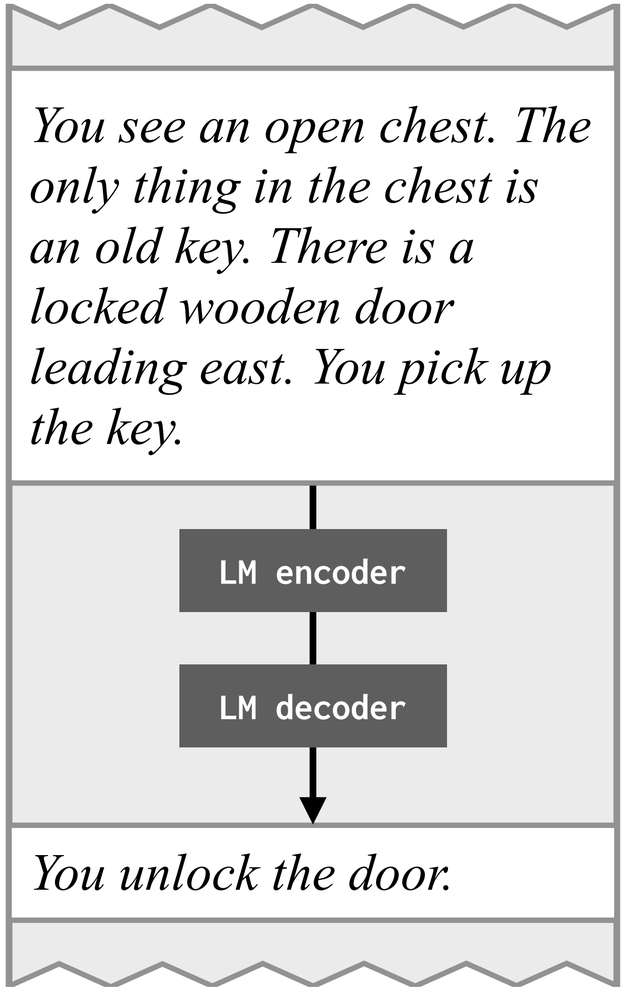
p(“Jack will get a top” | …)



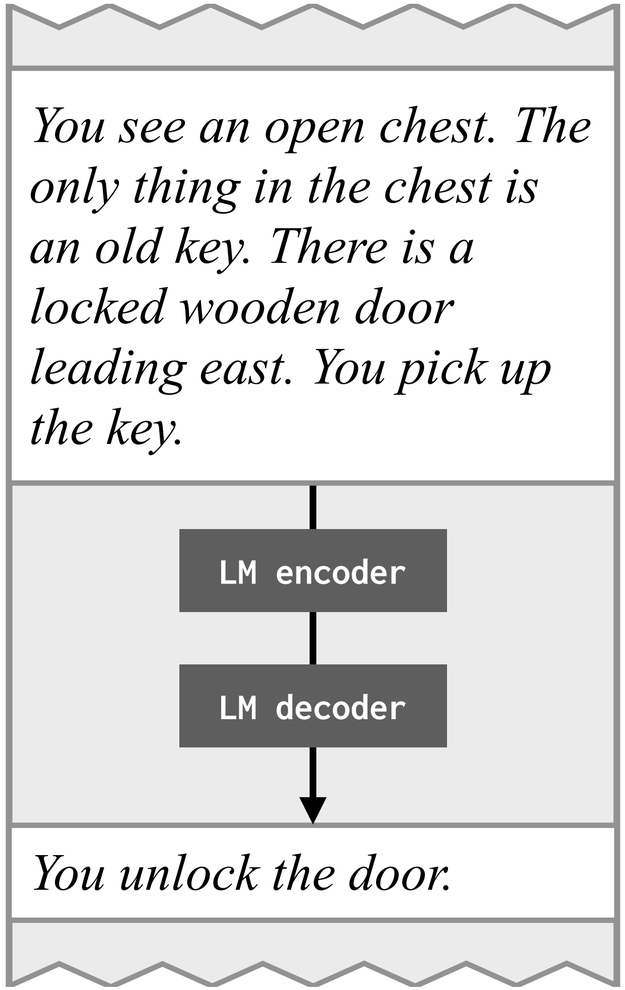


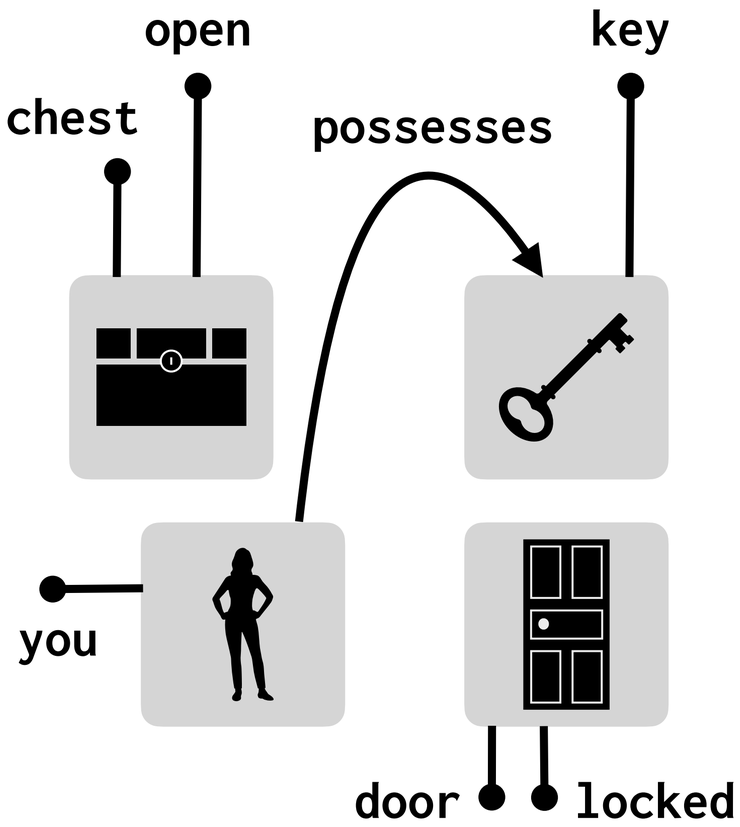
vector representations

Implicit representations of semantic state

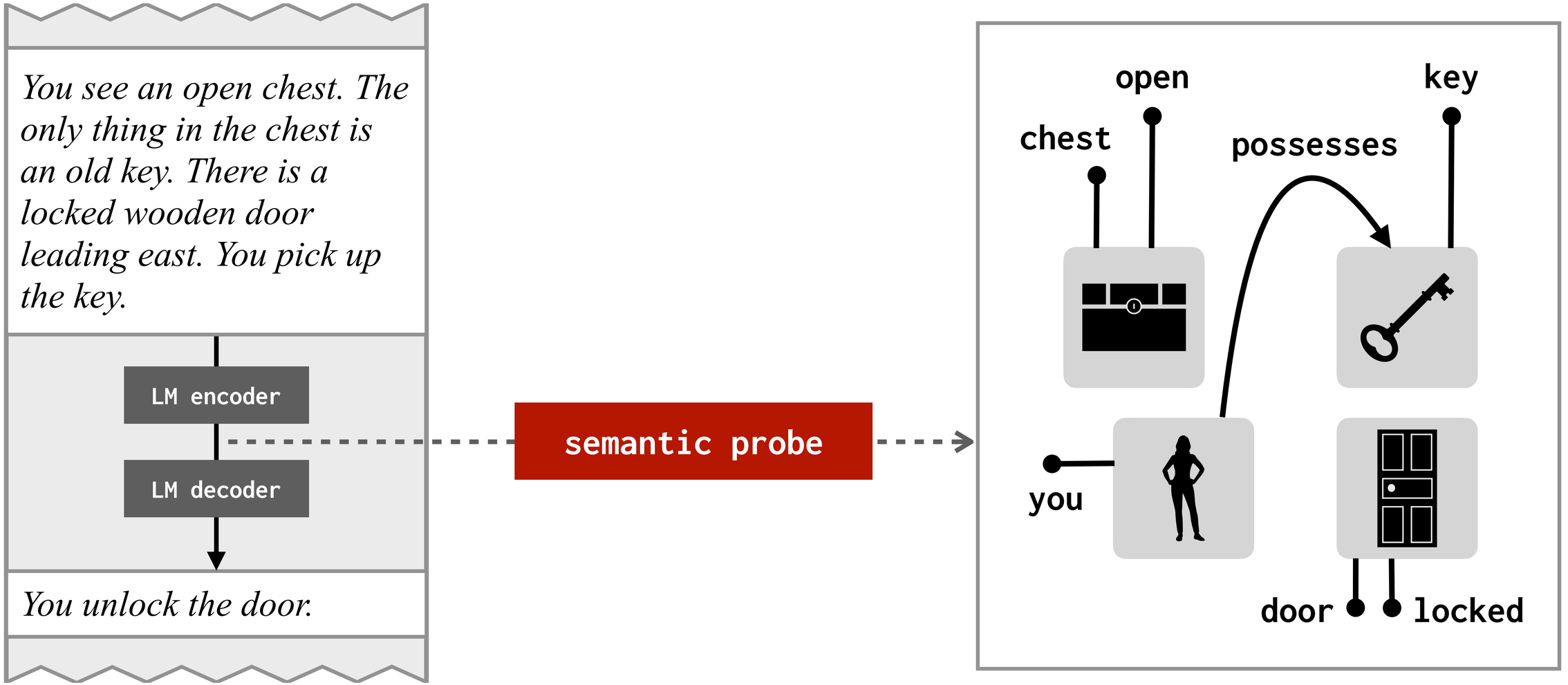


Implicit representations of semantic state

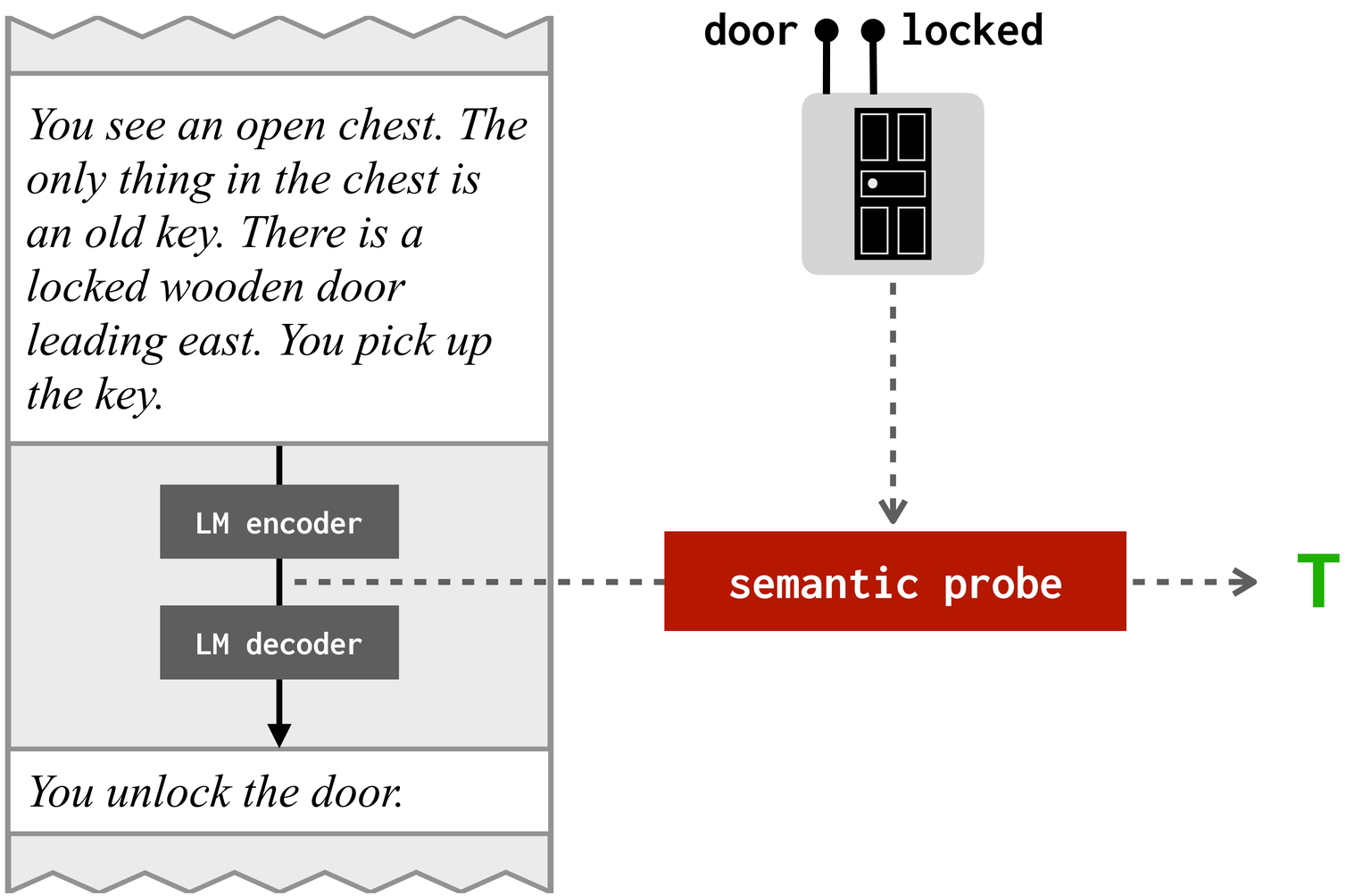




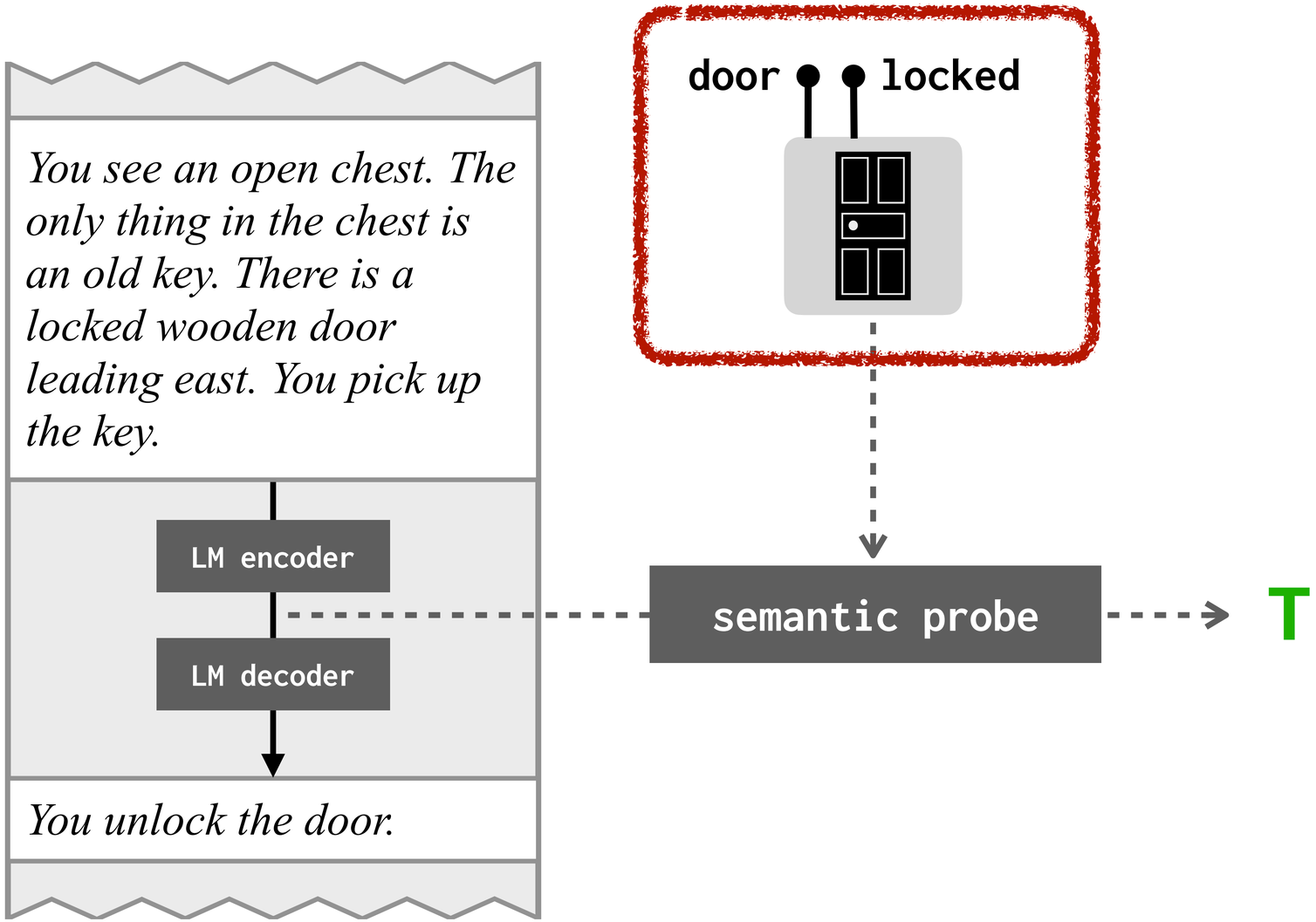
Implicit representations of semantic state



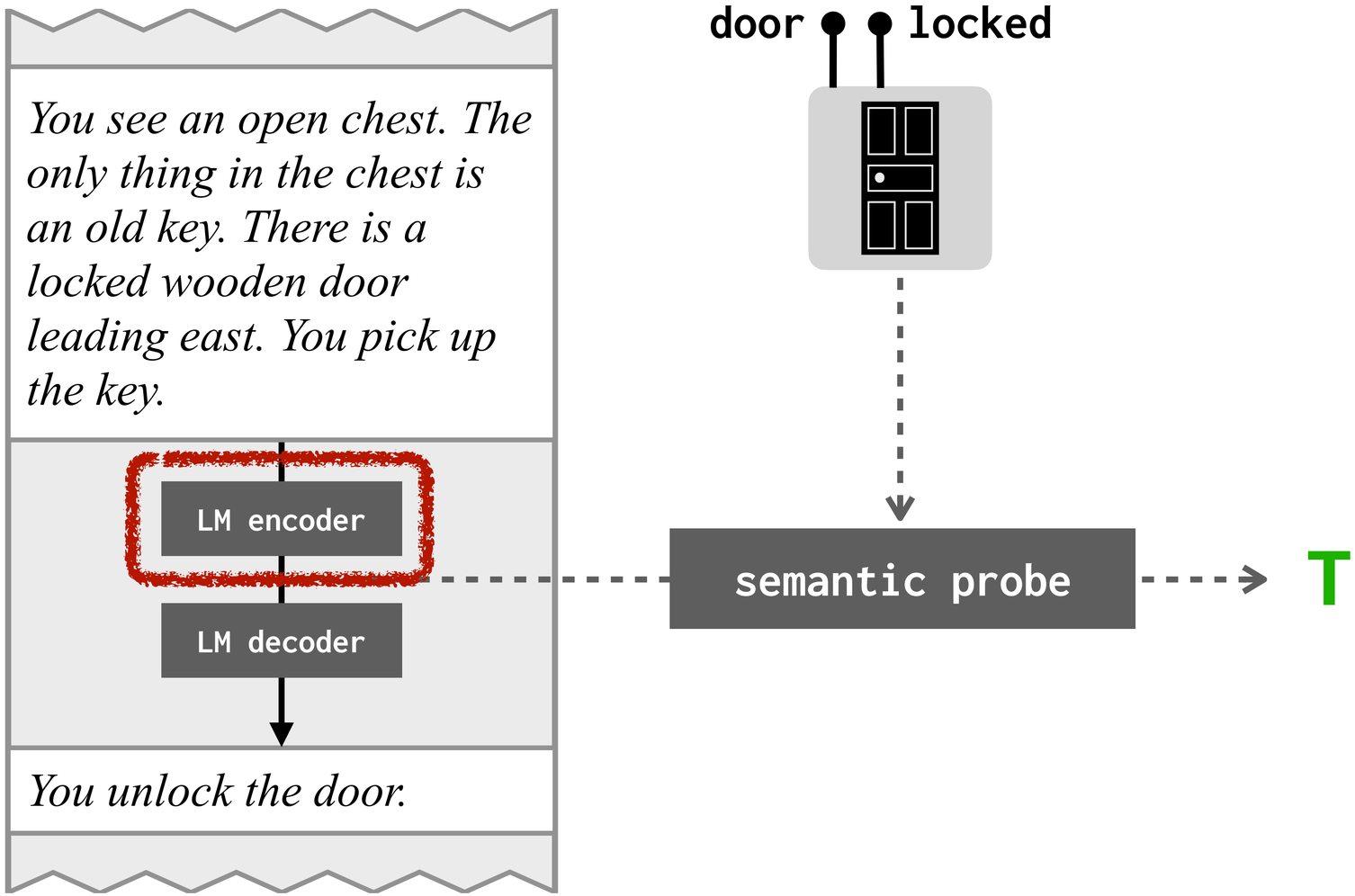
Implicit representations of semantic state



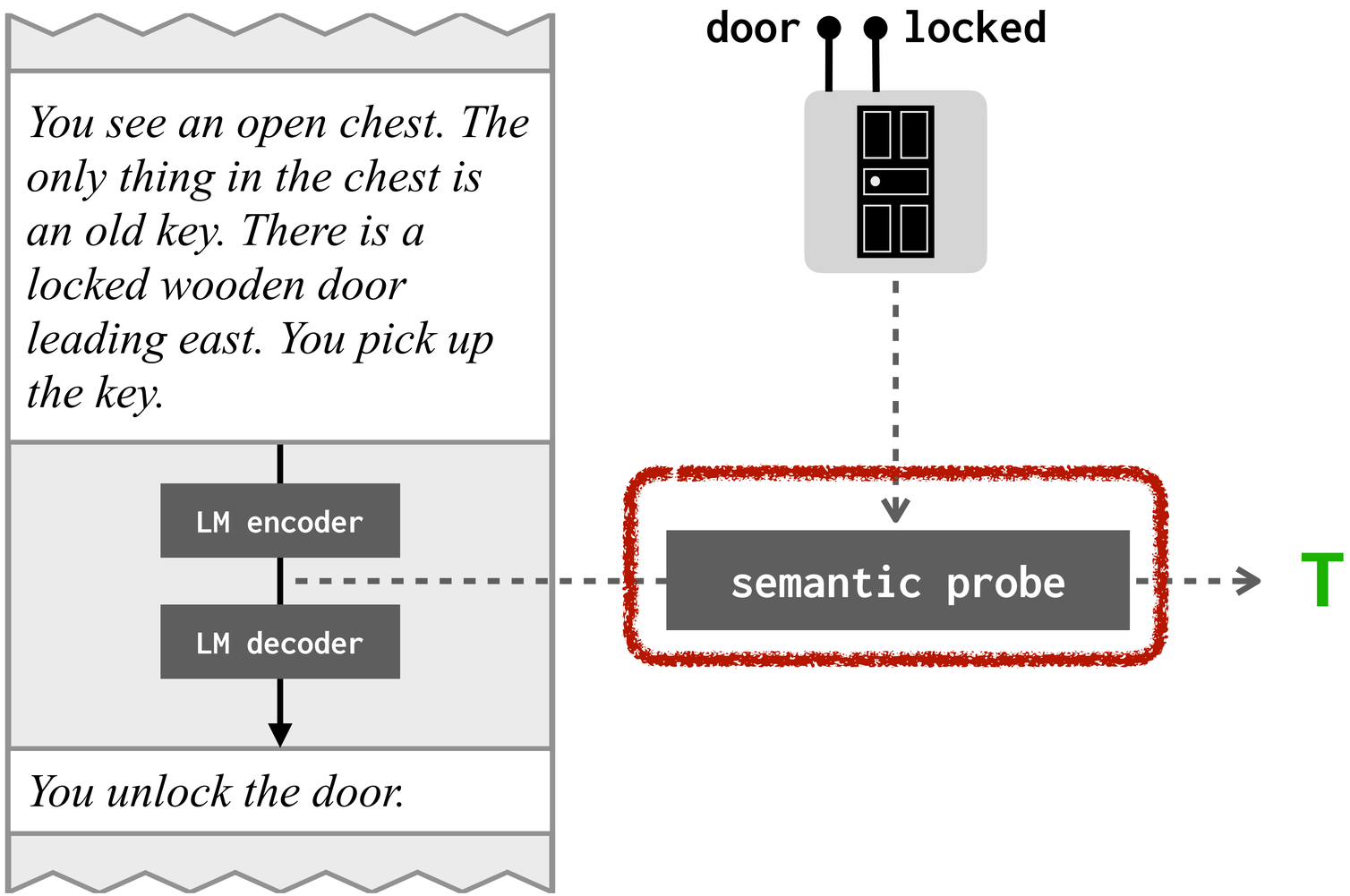
Implicit representations of semantic state



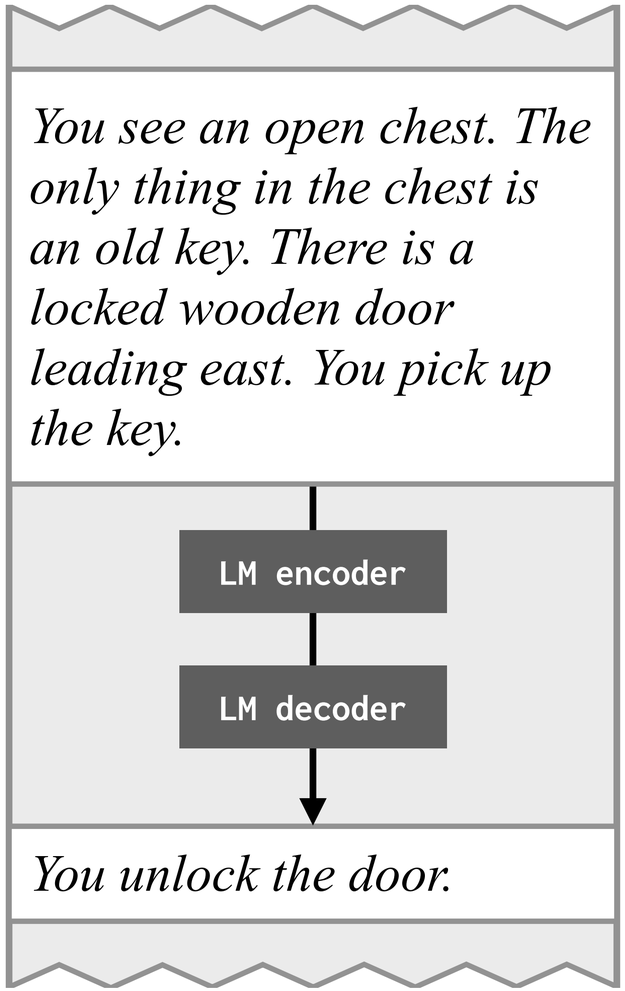
Implicit representations of semantic state

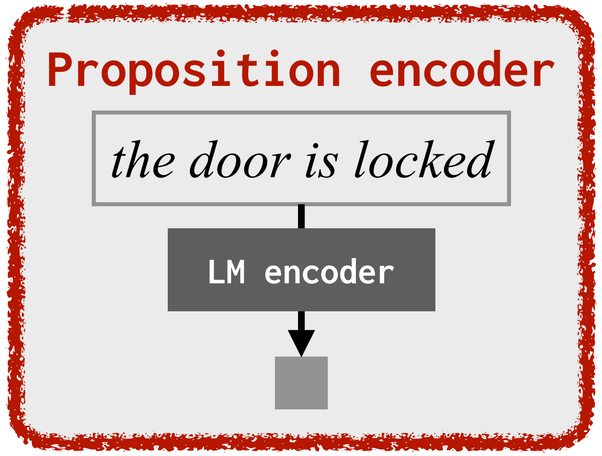


Implicit representations of semantic state

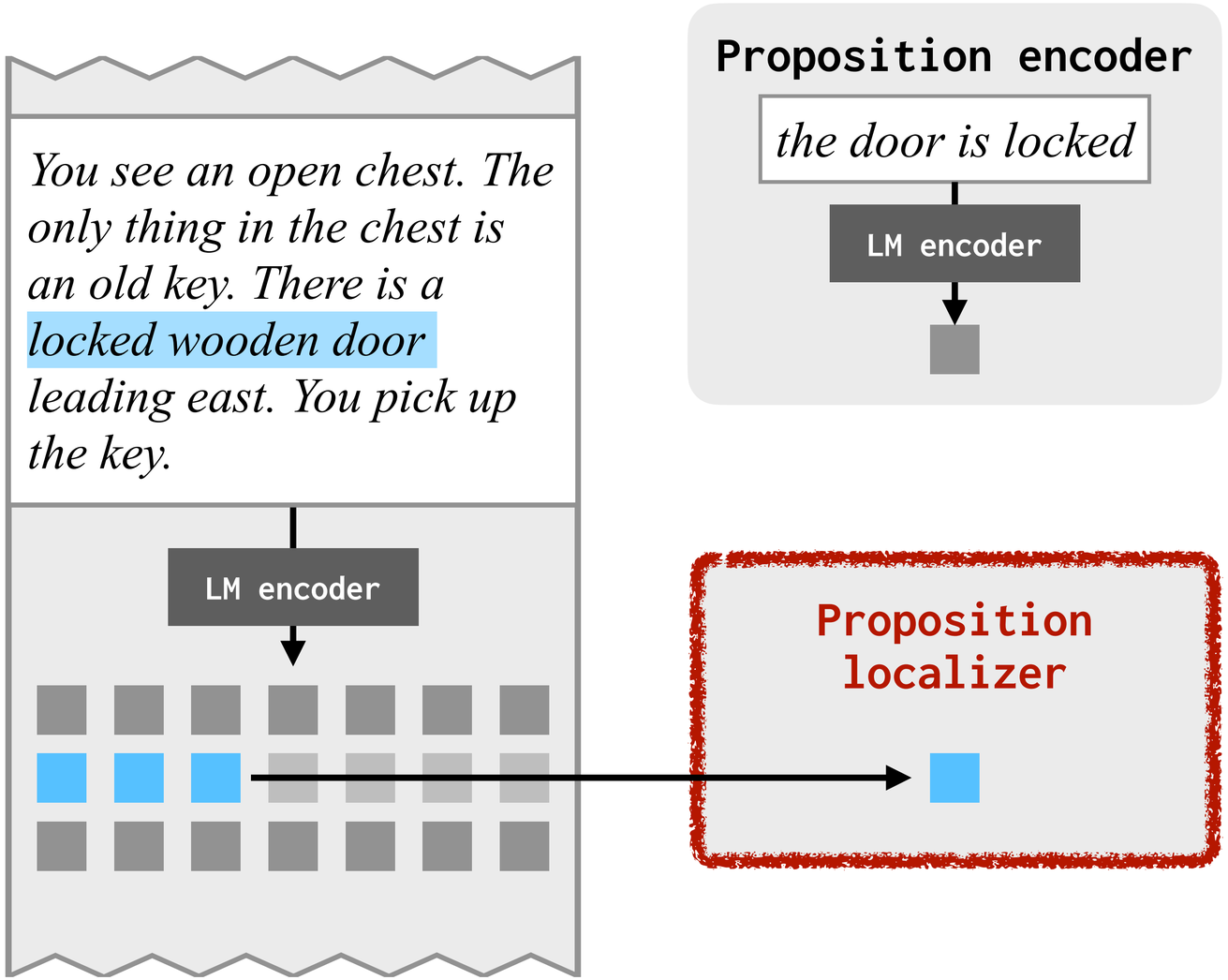


Building the probe



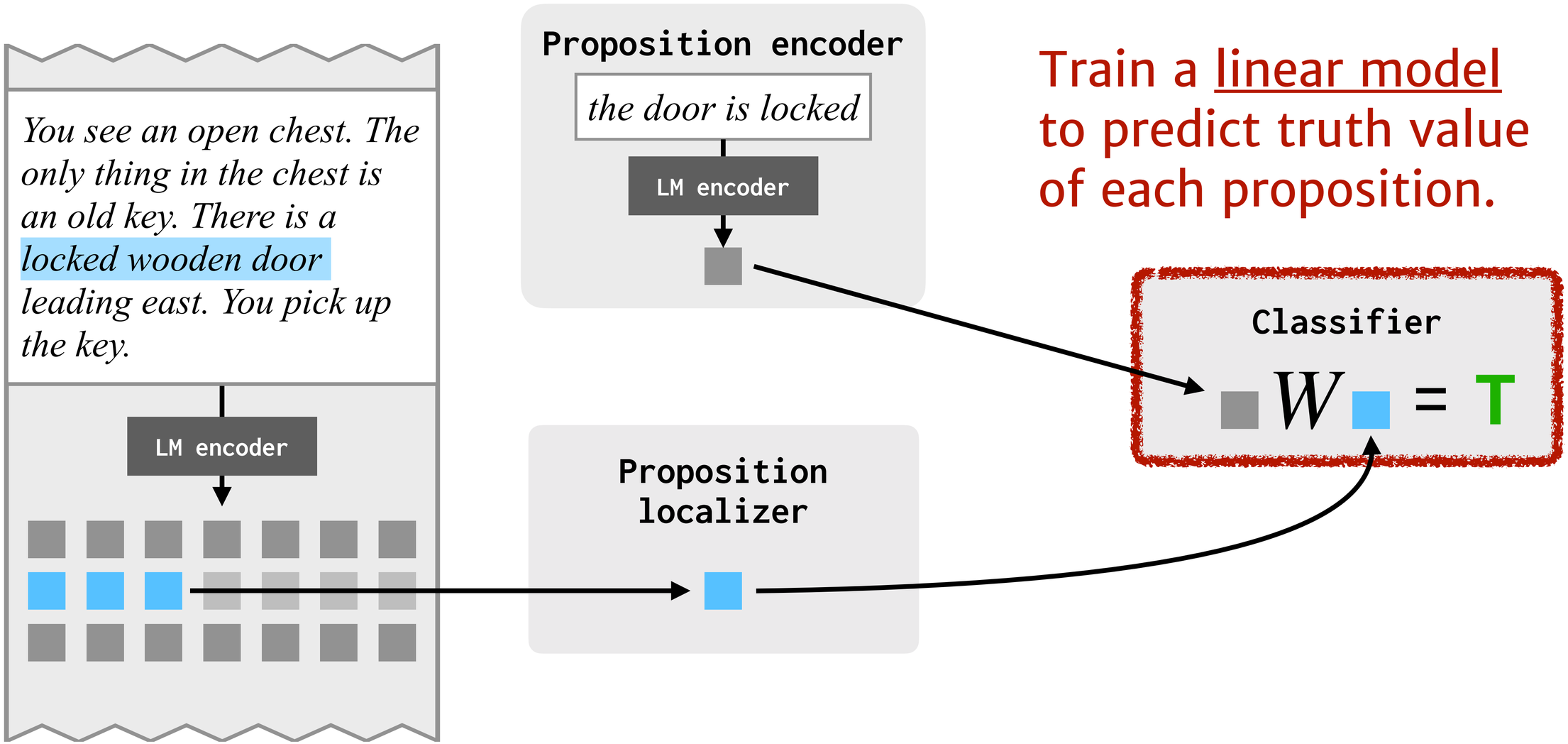


Building the probe



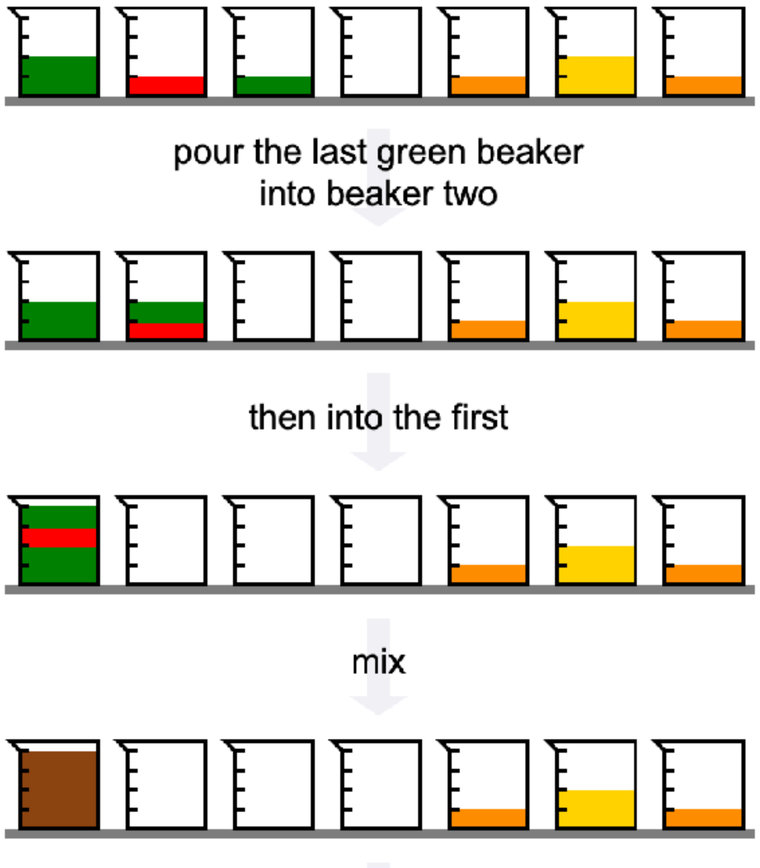
Decode facts about an entity from the encoding of its first mention.

Building the probe



Evaluation

Alchemy

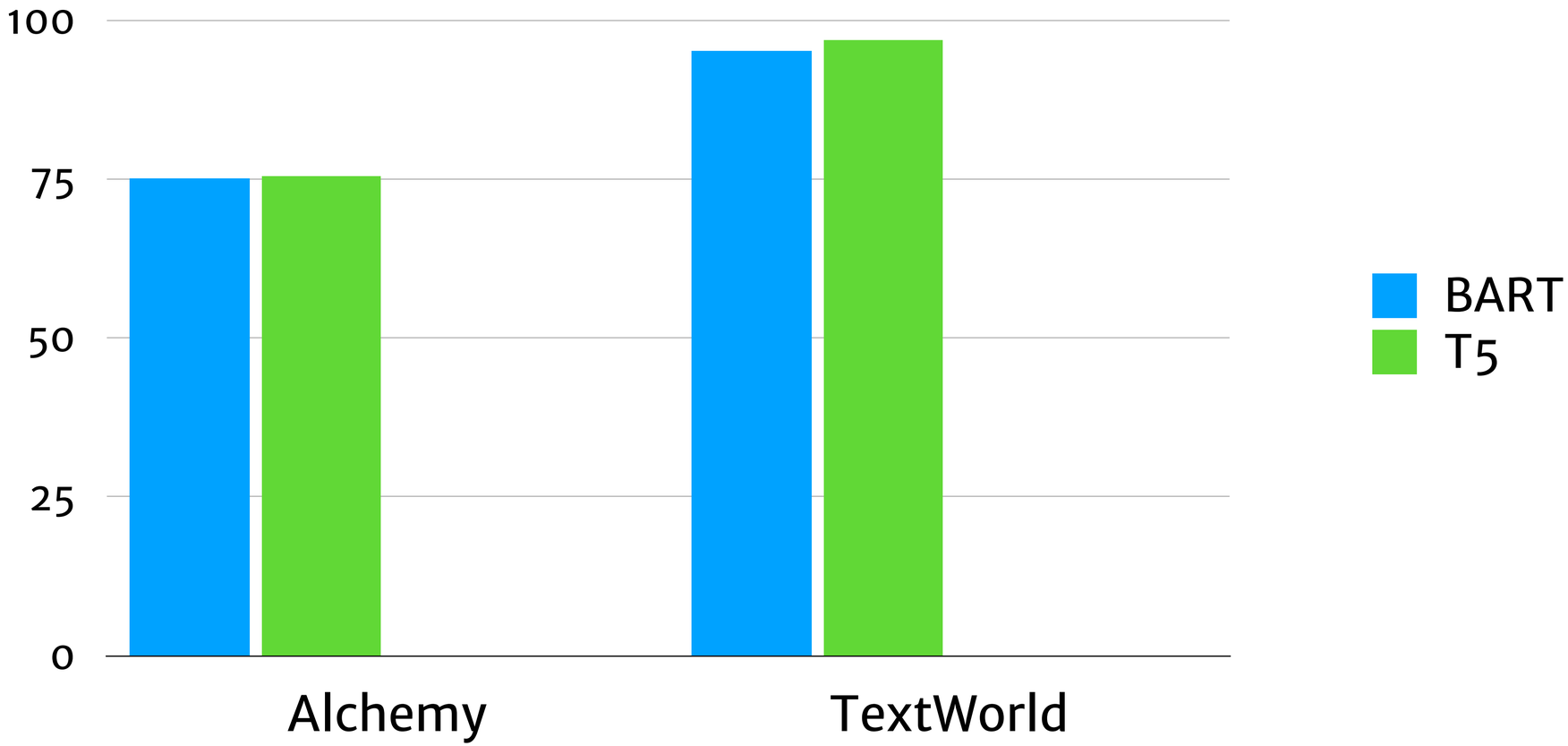


TextWorld

```
You are navigating through a
house. You've just entered a
serious study. There is a
gross looking mantle in the
room. It has nothing on it.
You see a closed rusty
toolbox. Now why would
someone leave that there?
Looks like there is a locked
door. Find the key to unlock
the door. You should try
going east.
```

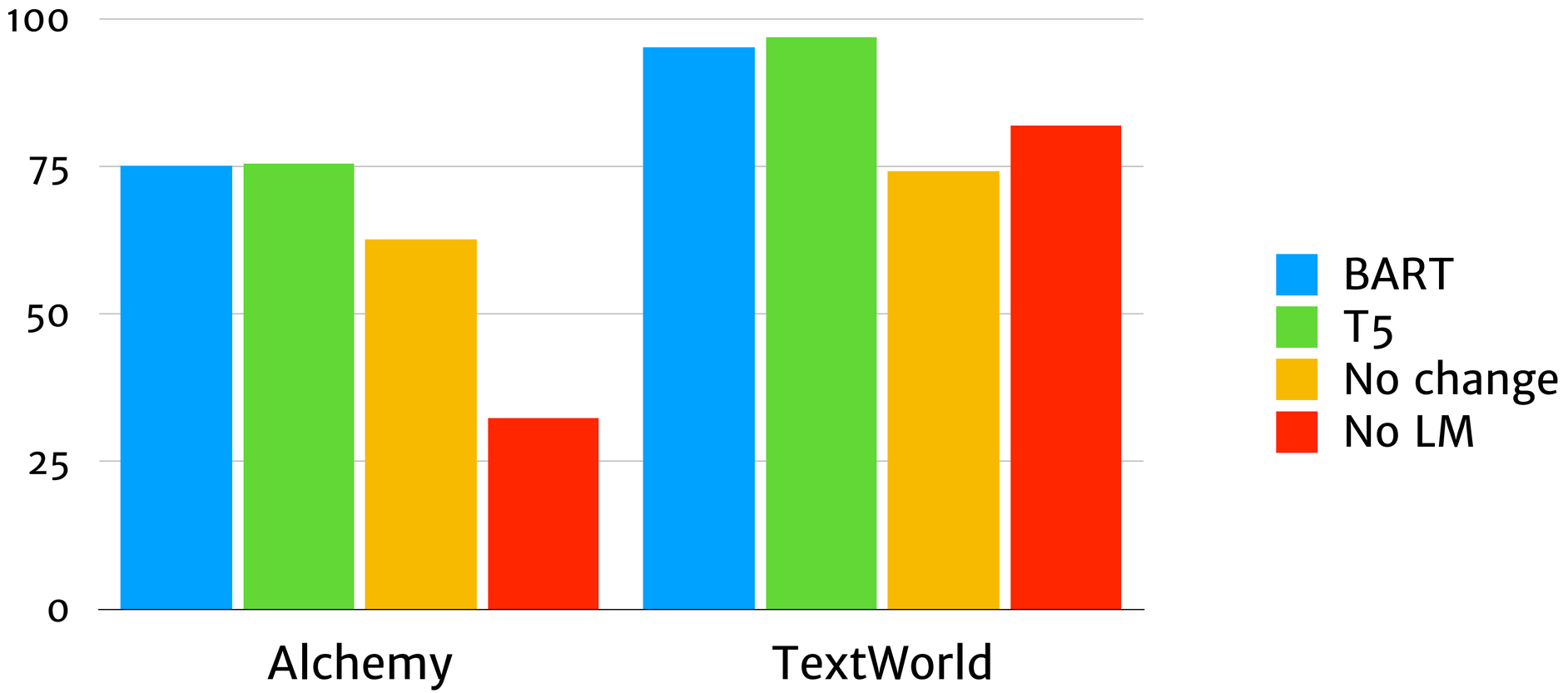
Evaluation: does it work?

What fraction of entities are exactly reconstructed?



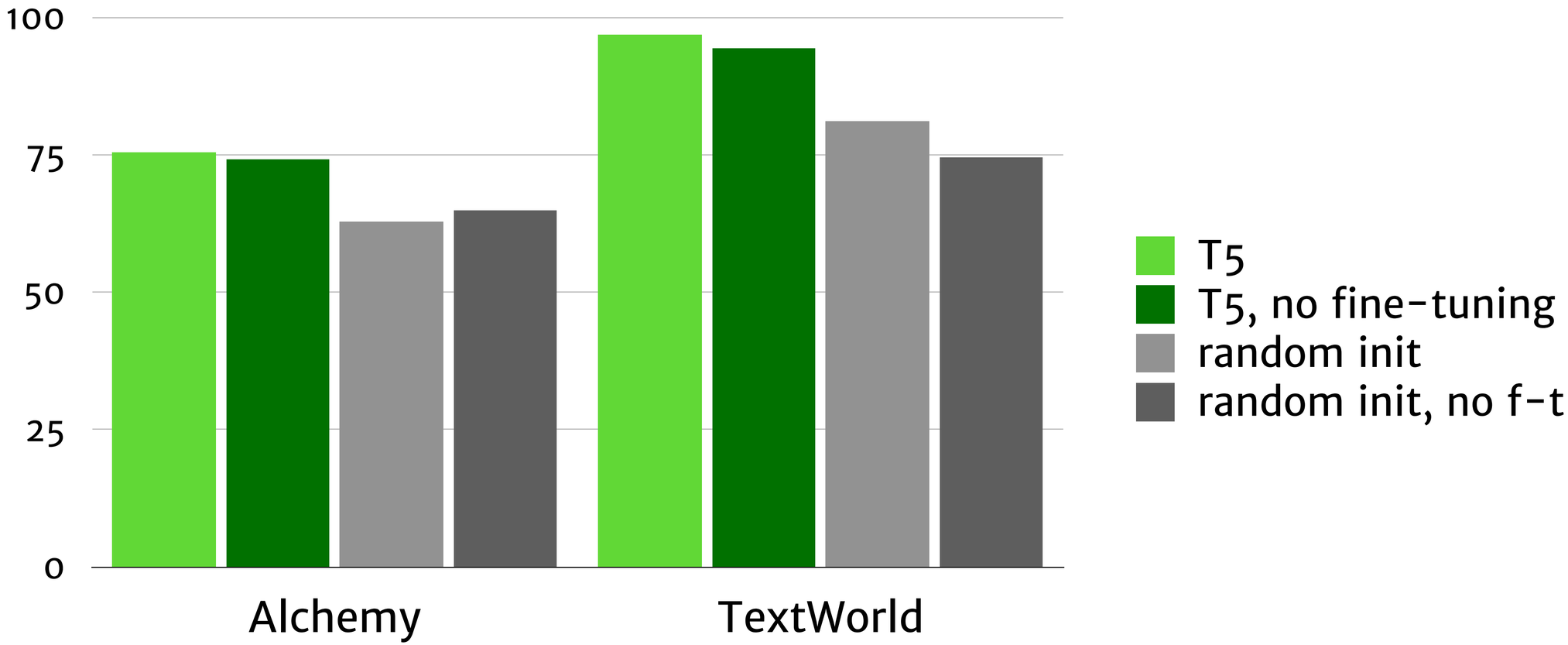
Evaluation: does it work?

What fraction of entities are exactly reconstructed?



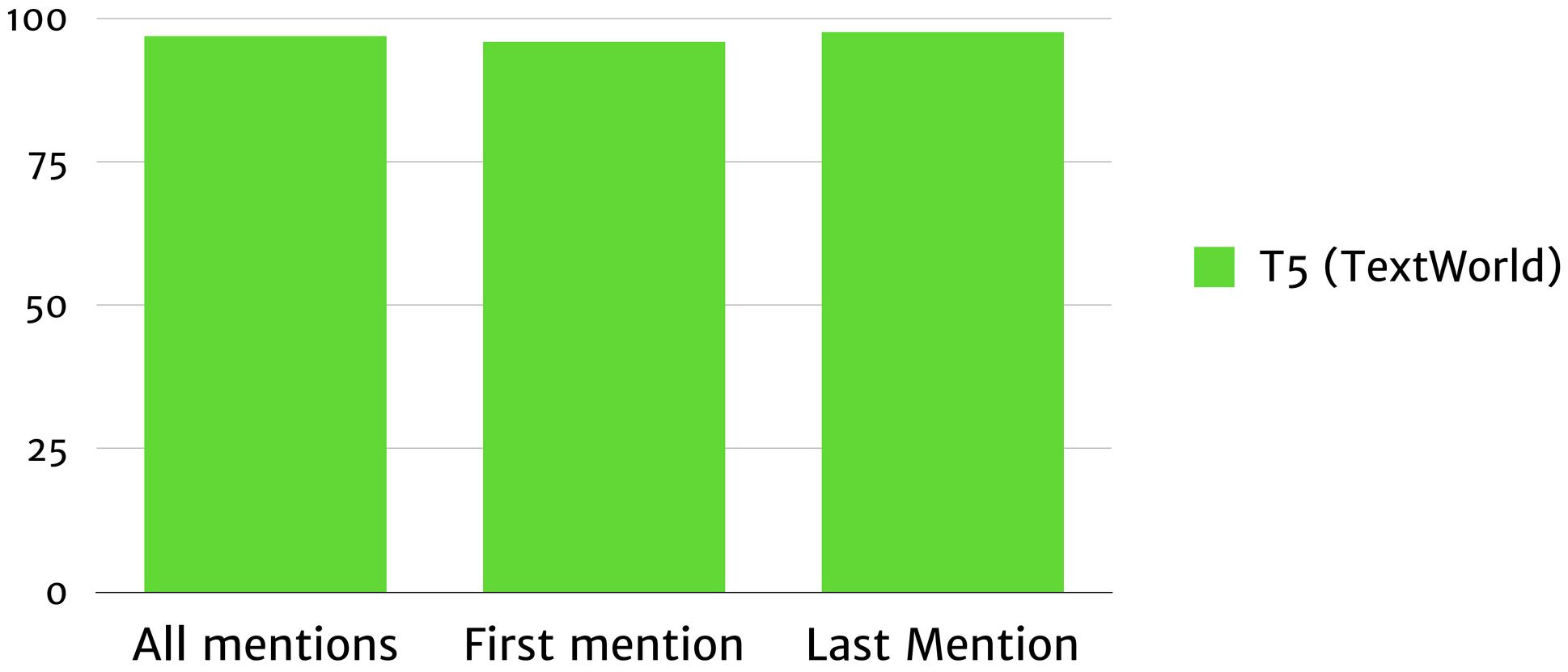
Evaluation: does it work?

What kind of training matters?



Evaluation: locality

What kind of training matters?



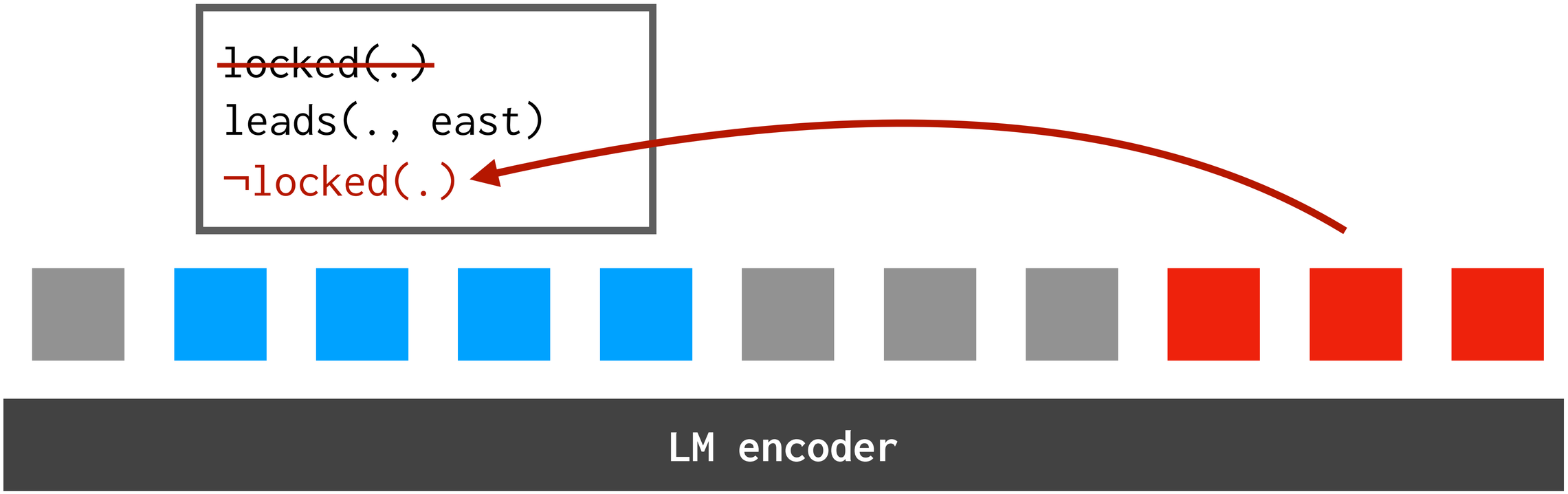
Evaluation: locality

	the	*[pos]*	*be*	*aker*	*has*	*[amount]*	*[color]*	*,*	*the*	*[pos]+1*	*[amount]*	*[color]*
(B)	39.4	41.4	42.1	41.5	58.5	68.6	74.3	64.8	45.1	35.0	35.7	34.5
(T5)	42.1	58.1	42.4	40.7	64.8	66.7	67.9	63.5	42.1	41.9	32.4	32.9
(a)	***the***	***third***	***be***	***aker***	***has***	***4***	***blue***	***,***	***the***	***fourth***	***2***	***red***

…Drain 2 from beaker 3.

58.5% / 64.8% accuracy

Language models as world models

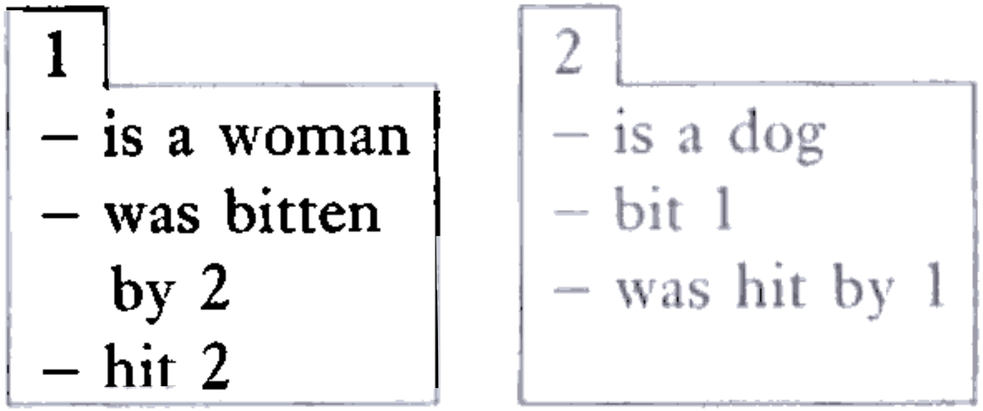


There's a locked wooden door leading east [...] you open the door.

Language models as file cards

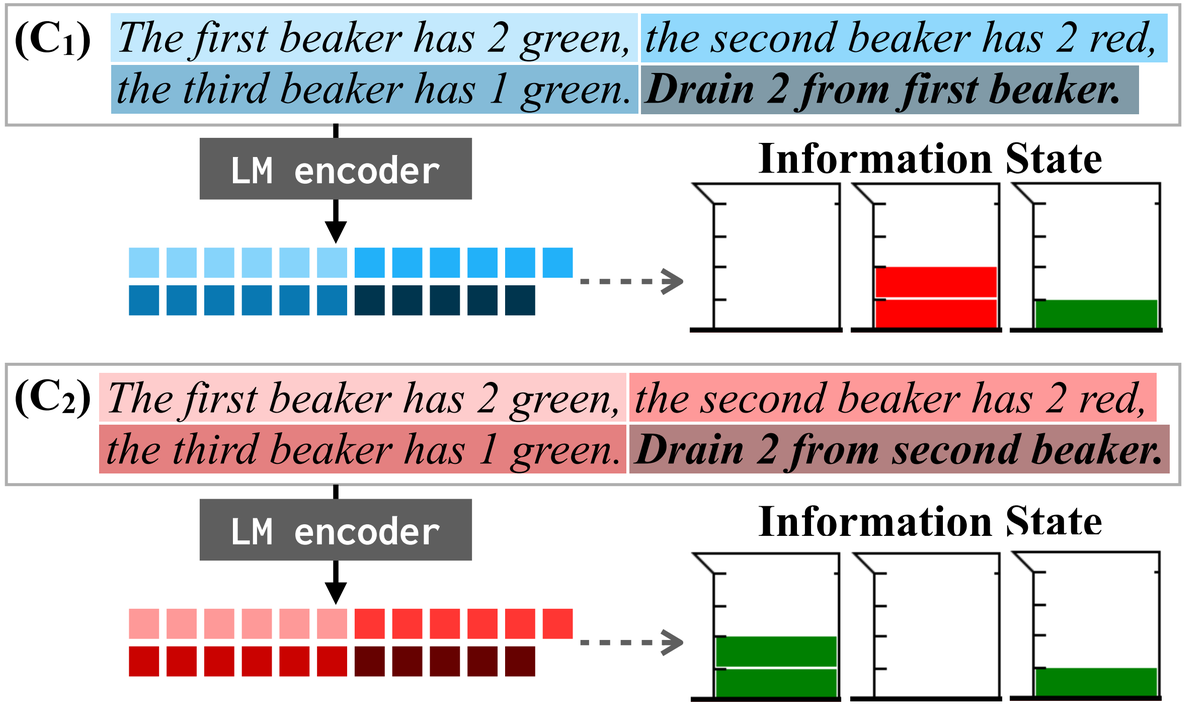
Next, (6b) gets uttered, which prompts the listener to update card 1 by adding the entry "hit 2", and to update card 2 by adding "was hit by 1". He now has F_2, still a two card file, but a different one:

F_2:

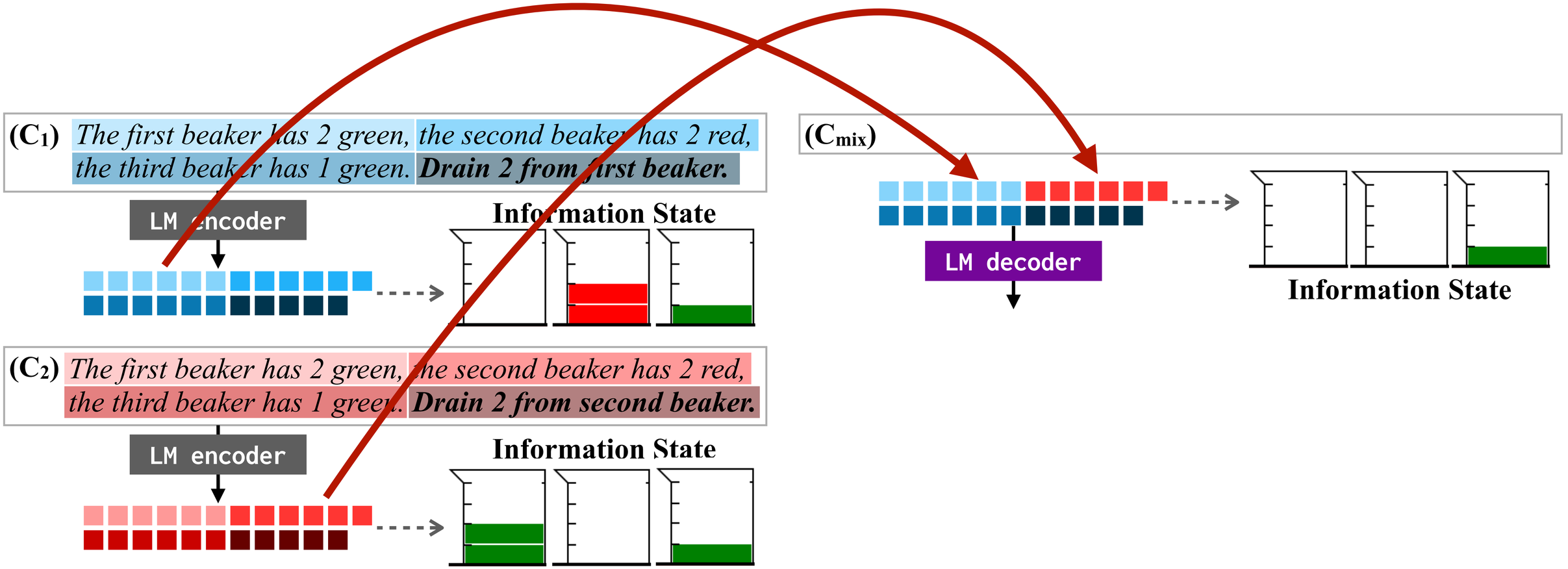


[Heim 1983!]

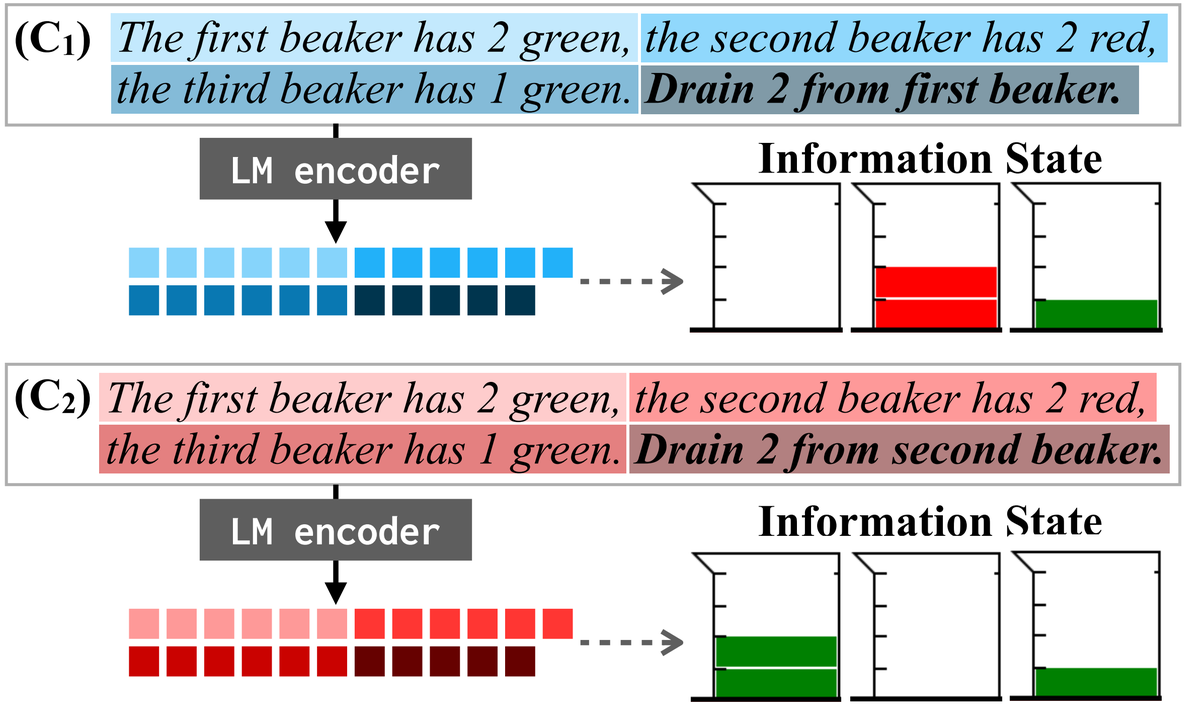
Building states from scratch

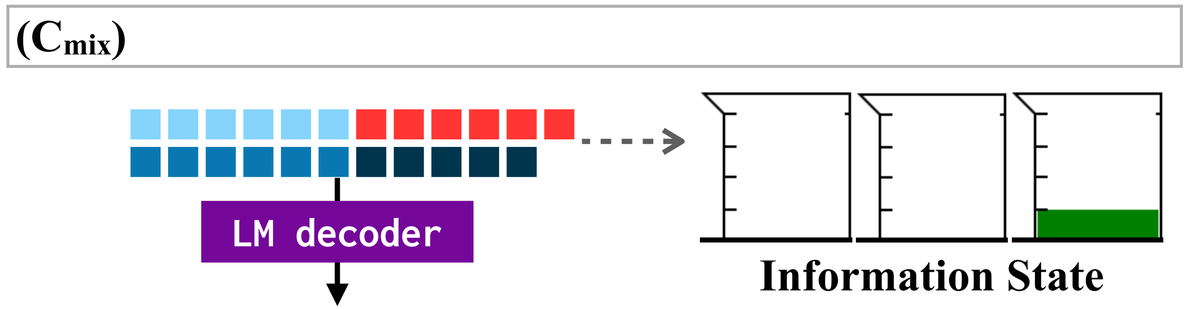


Building states from scratch



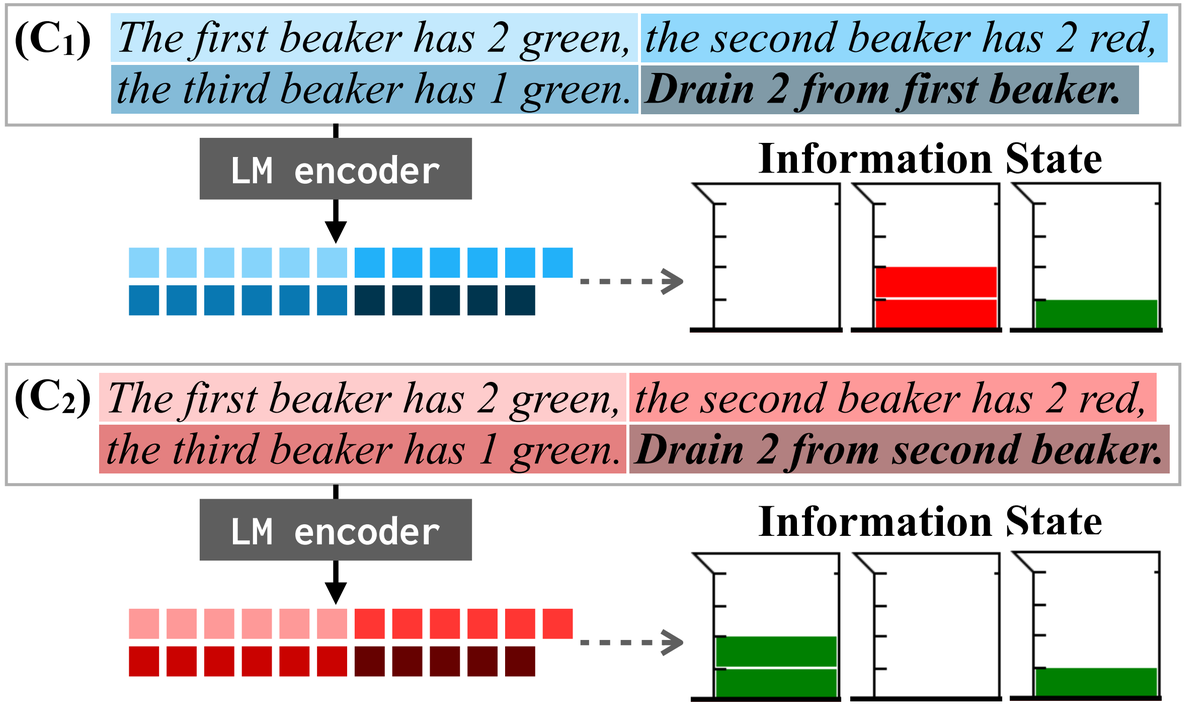
Building states from scratch

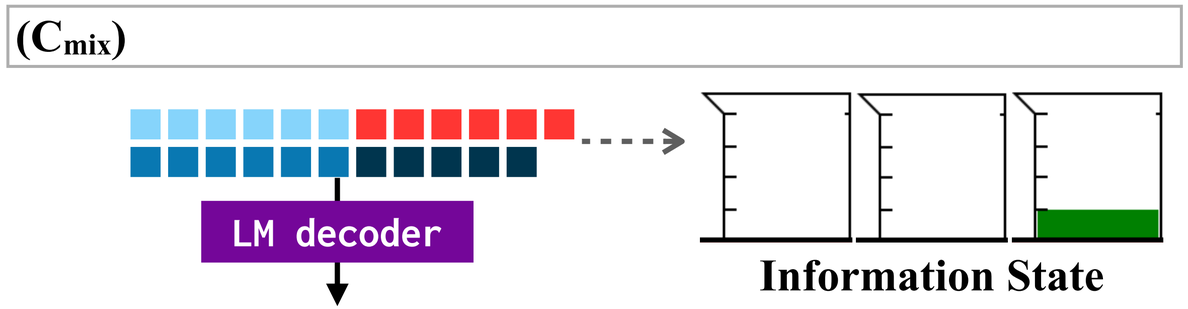




empty the third beaker

Building states from scratch

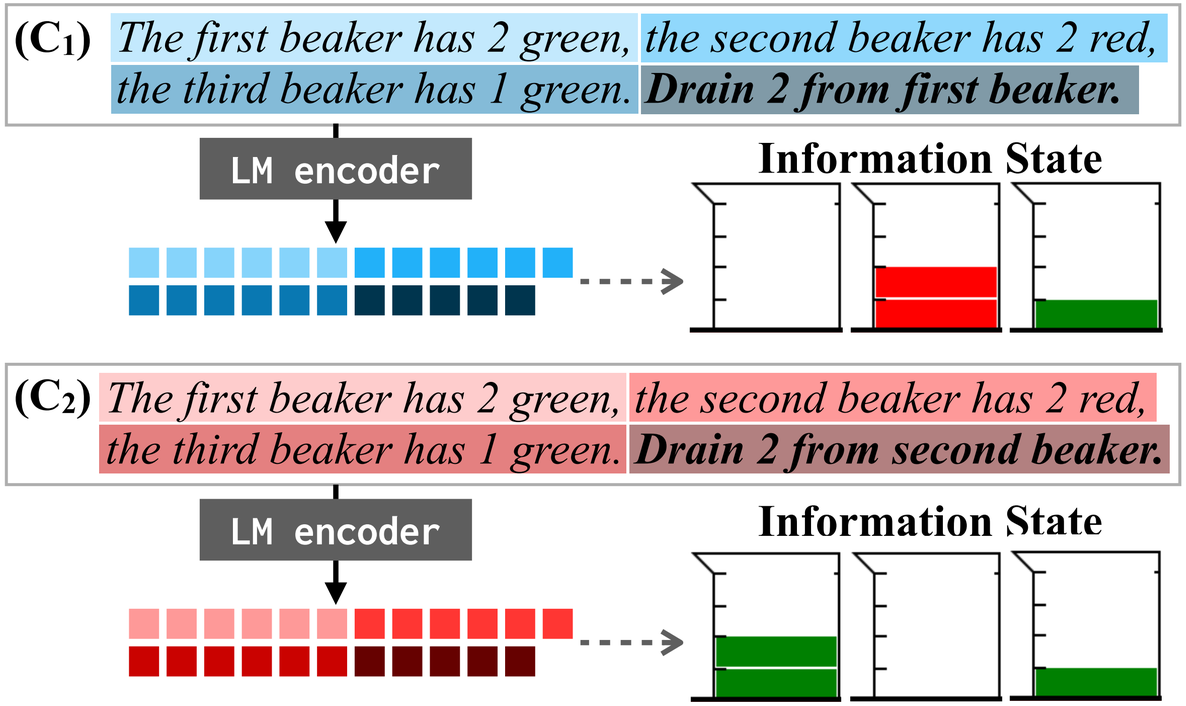


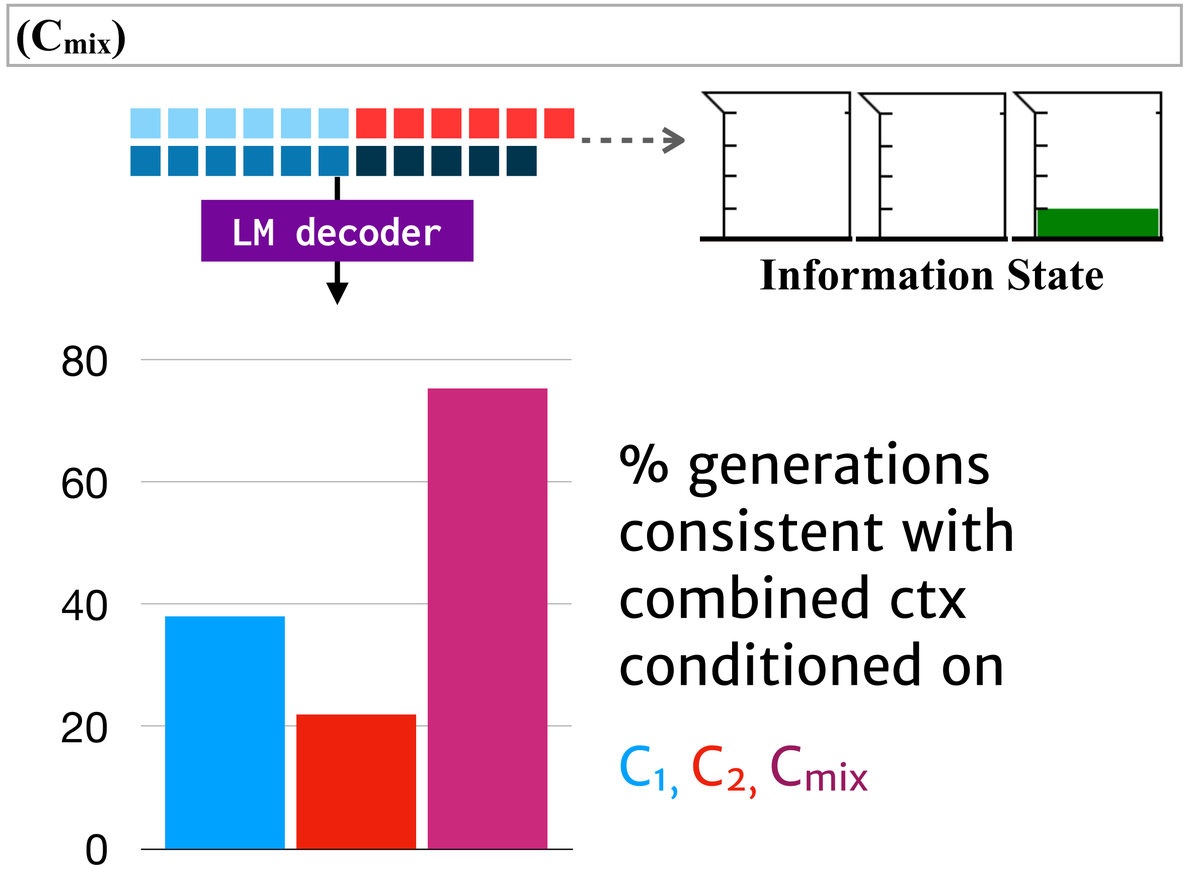


empty the third beaker

~~*stir the red beaker*~~

Building states from scratch





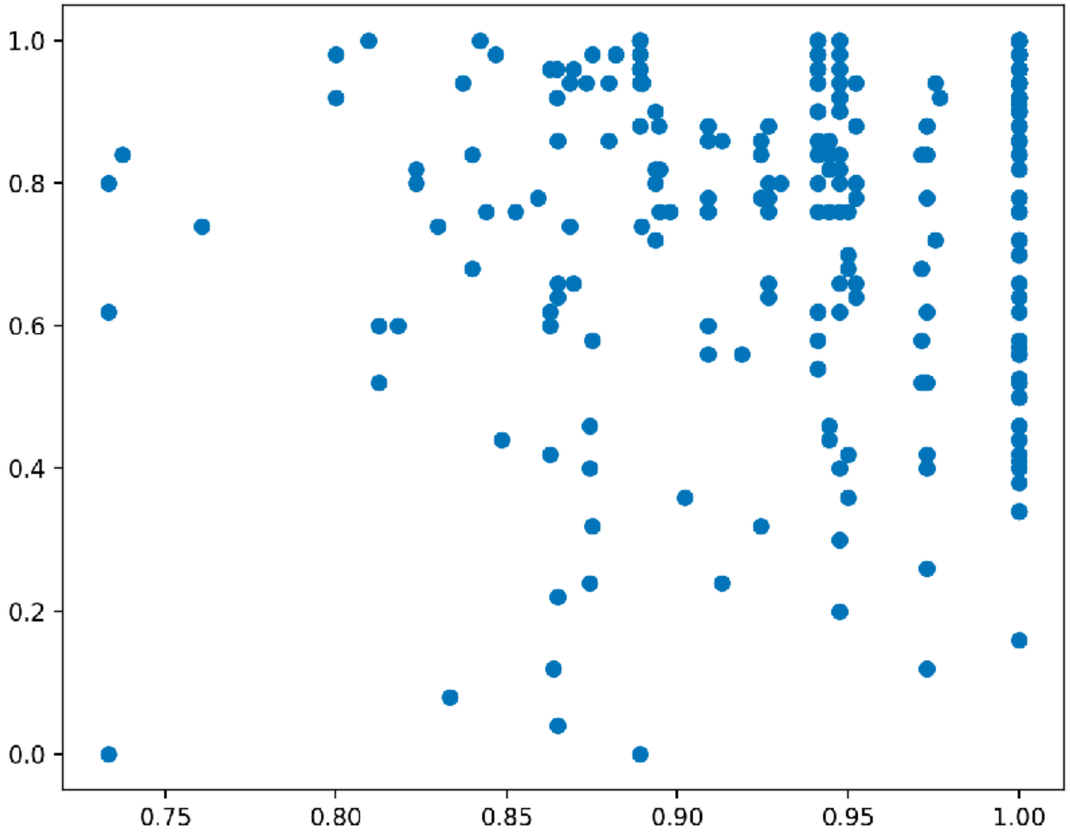
% generations consistent with combined ctx conditioned on

C_1, C_2, C_{mix}

What's still missing

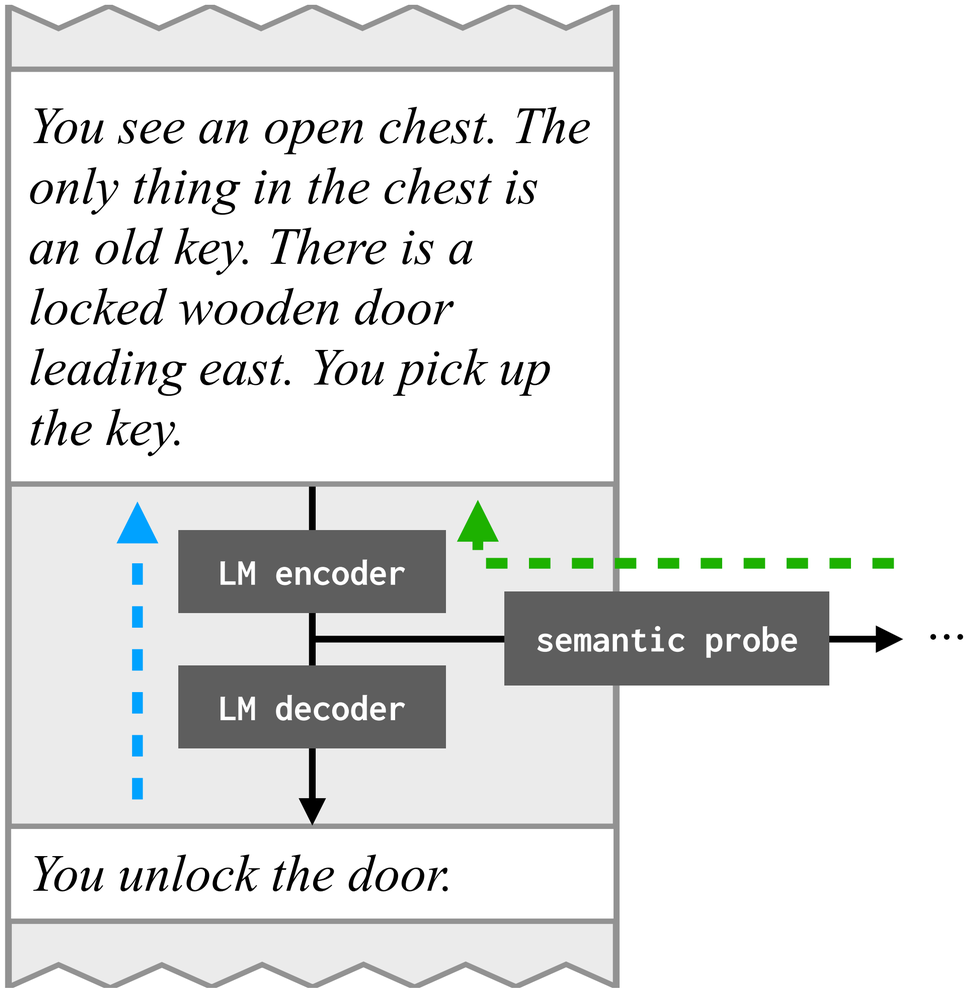
Attribution of model errors:

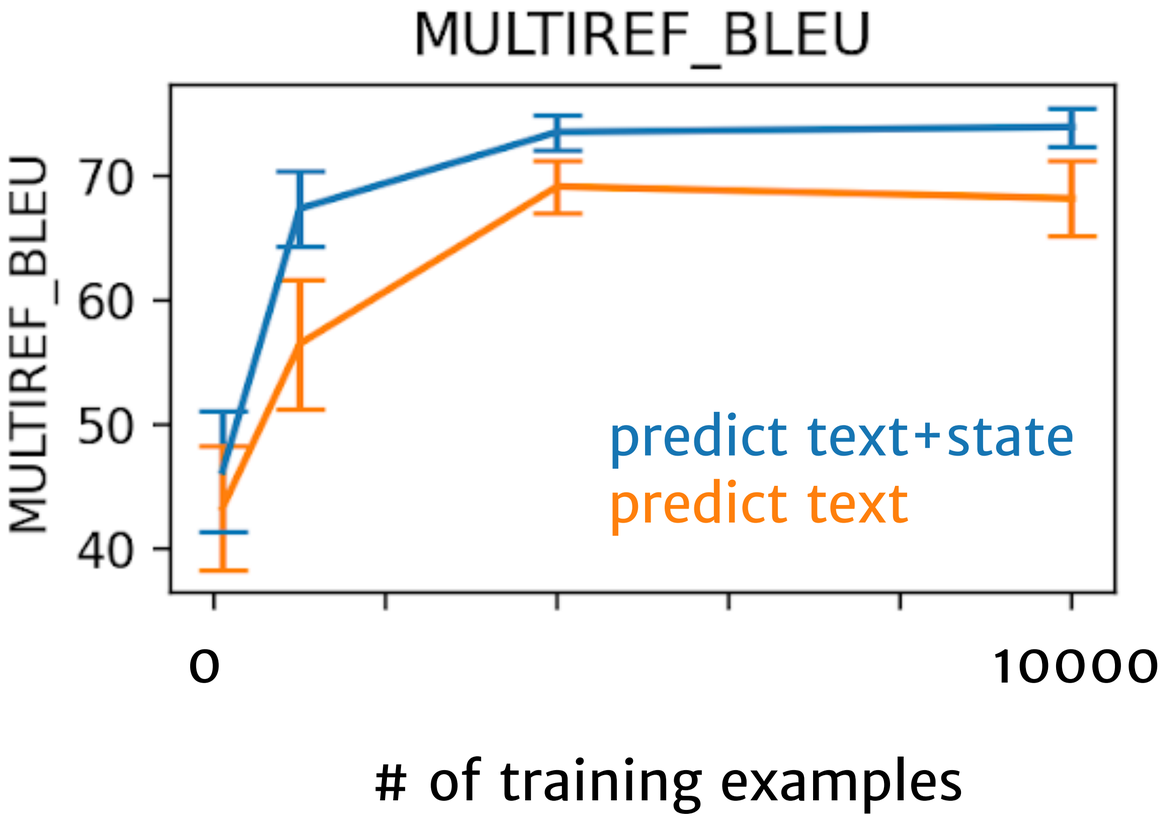
p(generation is semantically acceptable)



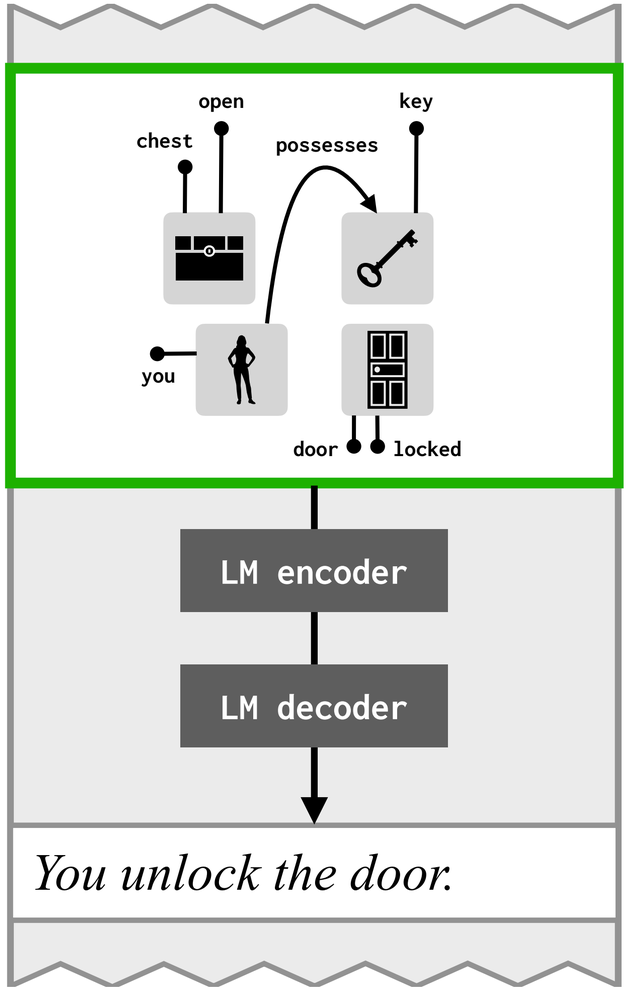
p(probe is accurate)

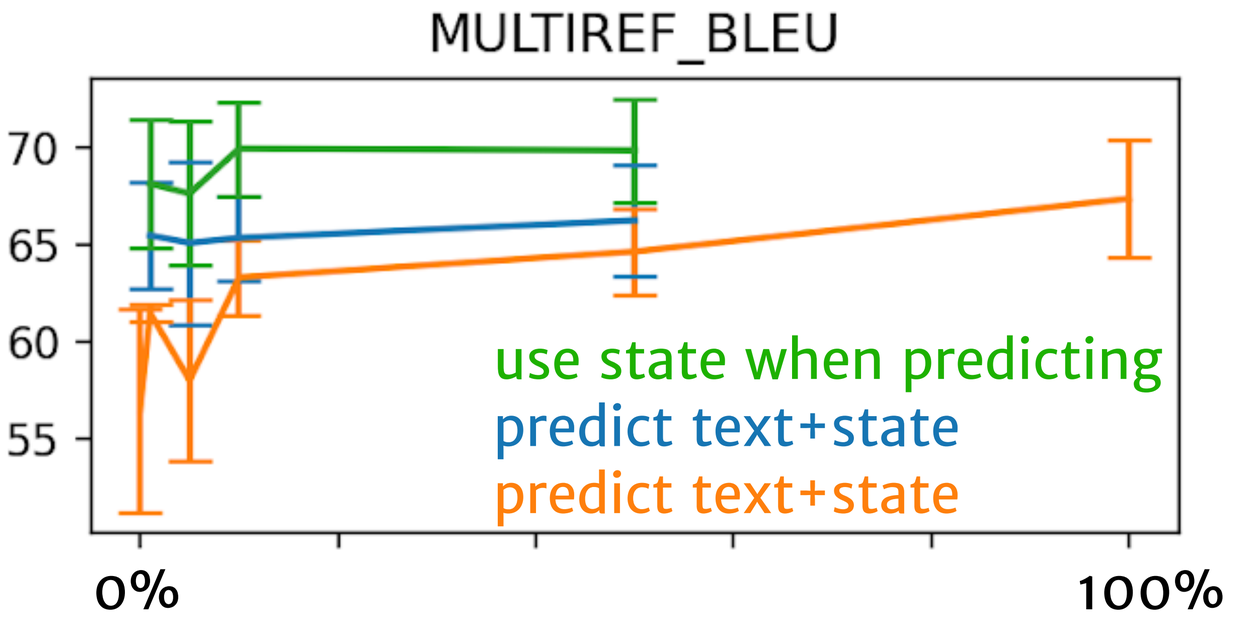
Does grounded training improve accuracy?





Would ground-truth states improve accuracy?





% of training examples with state labels

What's still missing

Quantification: *There are twenty-three reindeer; most of them have red noses.*

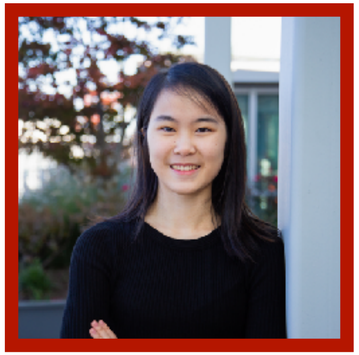
Implication and counterfactuals: *If Pat goes to the party, so will Jan. If Pat had gone to the last one, Mo would have gone too. Pat will go to the party this time.*

Summary

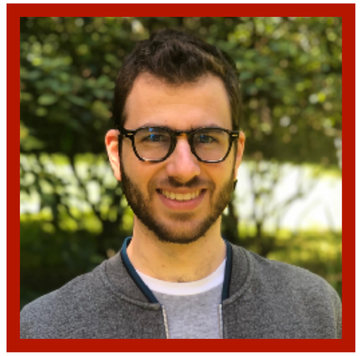
Language produce (rudimentary) representations of world states, and these states can be manipulated with predictable effects on model output.

But far from 100% reliable; lots of open questions about what these representations capture and how to improve them.

Researchers



Belinda Li



Max Nye

Sponsors

NVIDIA

SONY

MachineLearningApplications@CSAIL

Thank you!

Summary

Language produce (rudimentary) representations of world states, and these states can be manipulated with predictable effects on model output.

But far from 100% reliable; lots of open questions about what these representations capture and how to improve them.