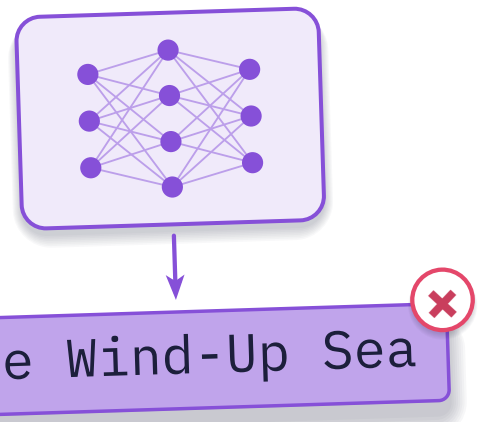
Prof. Aviv Zohar

Semester A, 2026–7

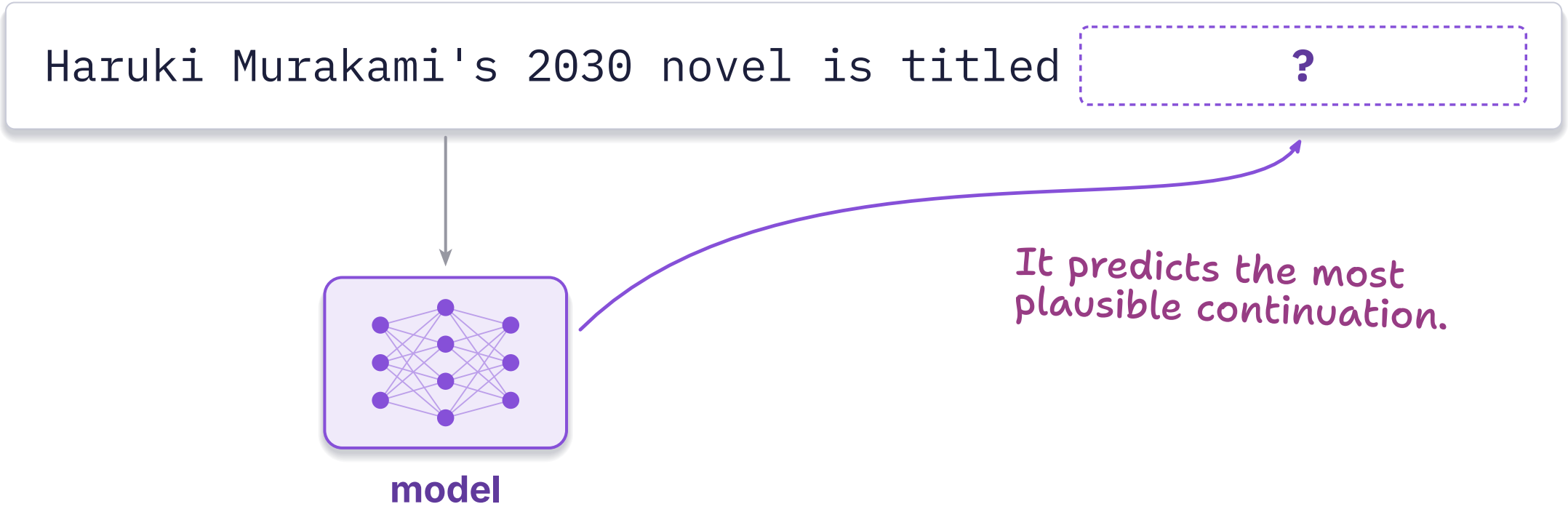
What the Model Knows, and Doesn't

Hallucinations, memory and context, the knowledge cutoff



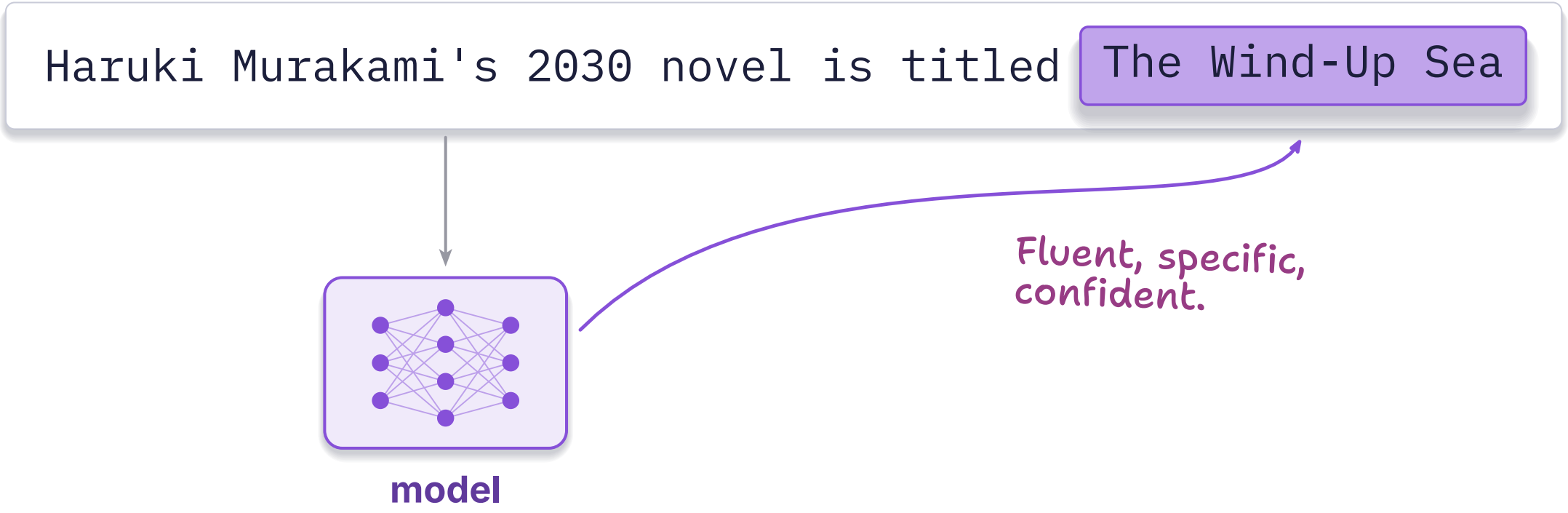
Fluent is not the same as true

the text so far



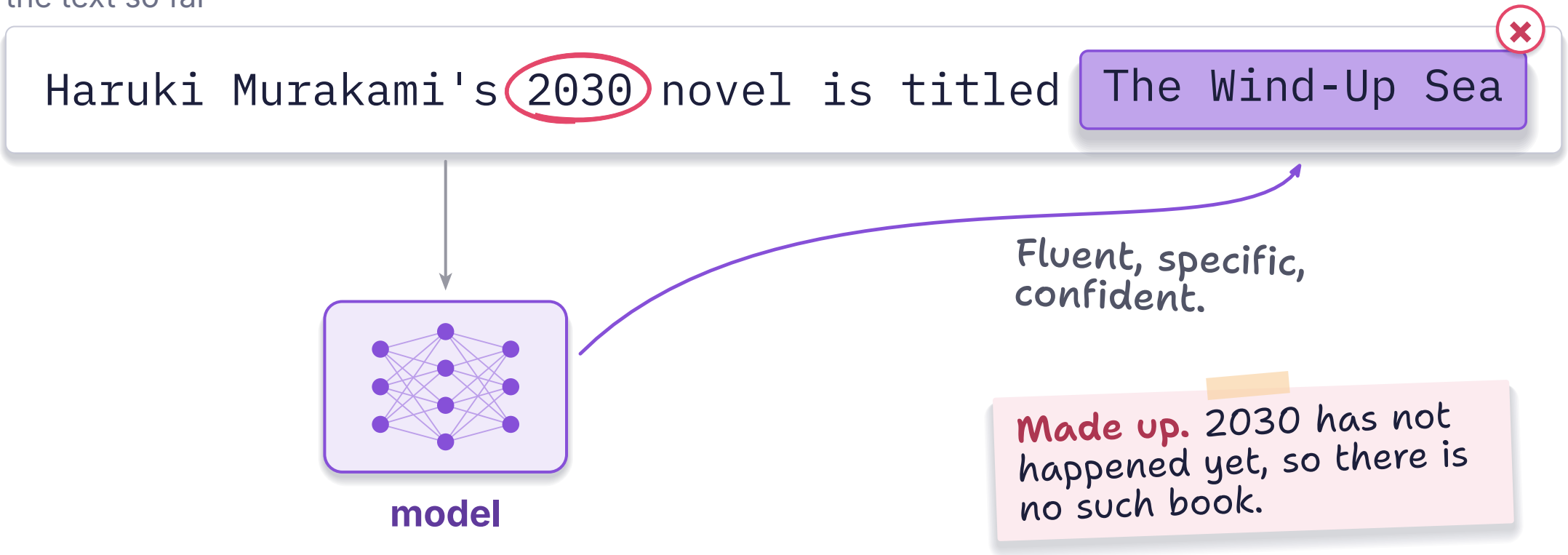
Fluent is not the same as true

the text so far

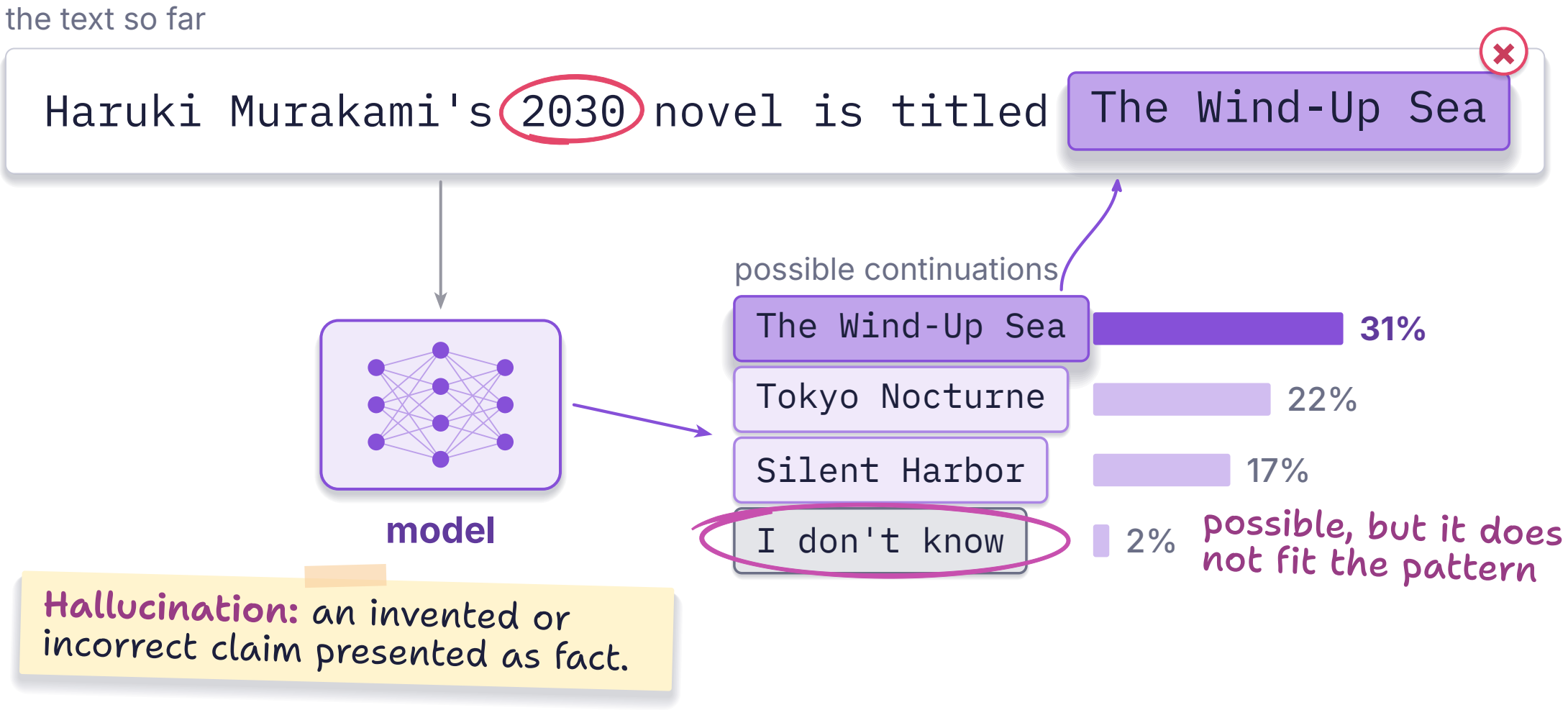


Fluent is not the same as true

the text so far



Fluent is not the same as true



probabilities are illustrative

Later training helps, but does not cure it

completion engine

a made-up title



“I don’t know”



→

after later training

a made-up title



“I don’t know”



bar lengths are schematic

How a benchmark scores an answer

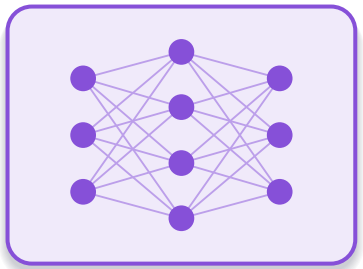
a lucky guess	**1 point**
a wrong guess	0 points
“I don’t know”	0 points

Easier to say “I don’t know.” Hallucinations still happen.

Guessing never scores worse, so high scores can come with confident errors.

Two sources: the weights and the context

learned during training



model

supplied in this request

Two sources: the weights and the context

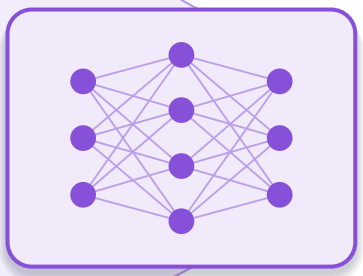
Not the instruction files the harness reads.

memory

in the weights, set by training

```
 0.21  -1.30   0.07   0.88
-0.42   0.15   1.12  -0.03
 0.93  -0.27  -0.61   0.40
-0.08   0.74   0.33  -1.05
```

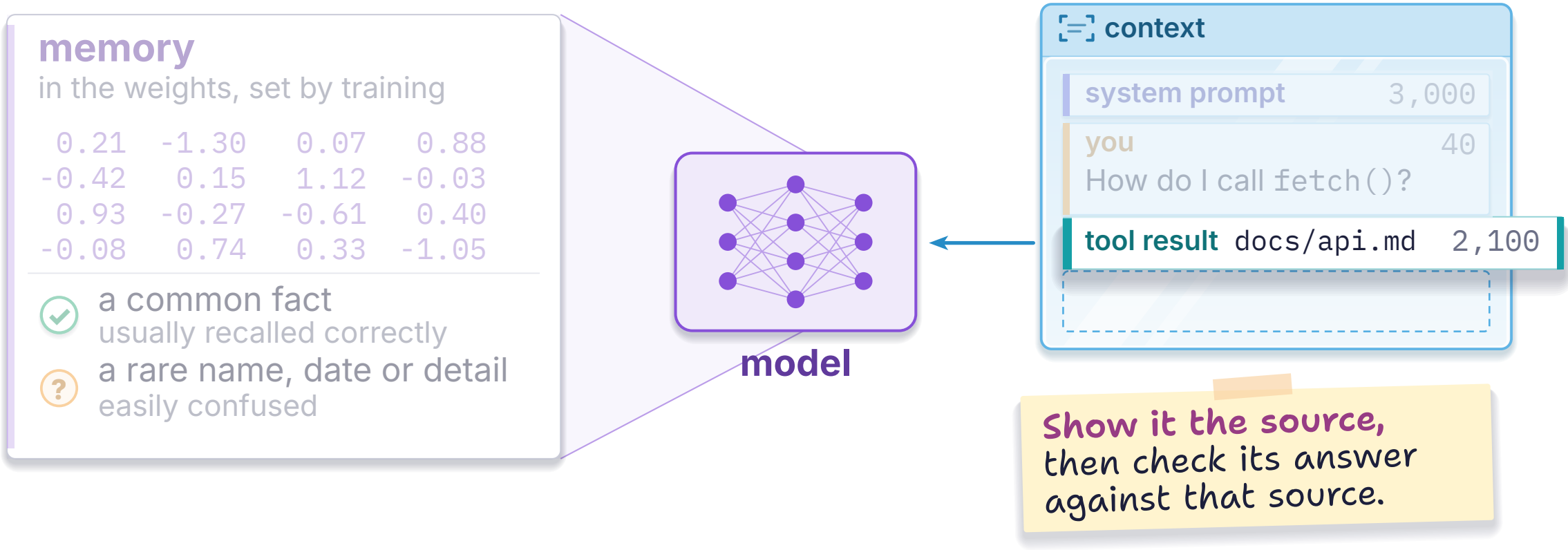
- a common fact
  usually recalled correctly
- a rare name, date or detail
  easily confused



model

Not a database of exact facts.

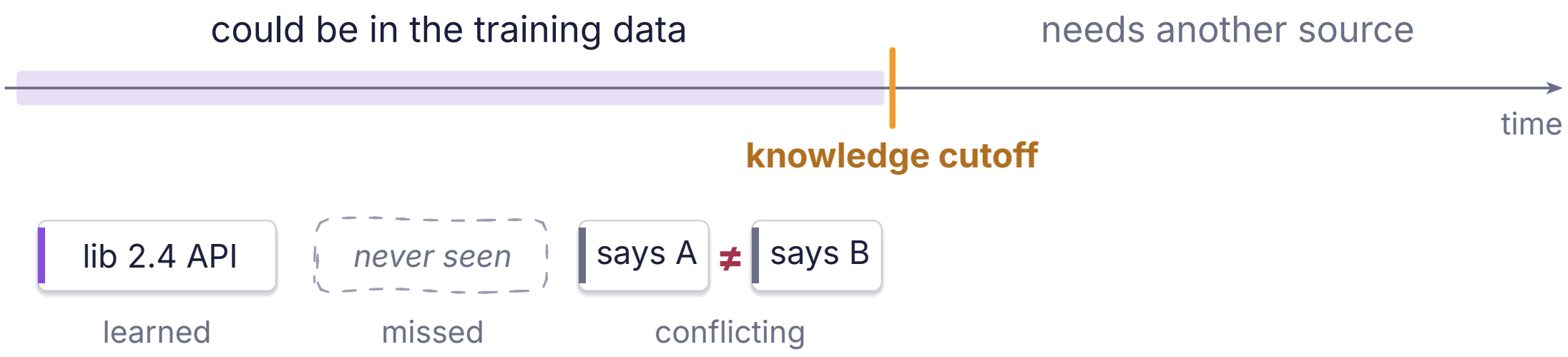
Two sources: the weights and the context



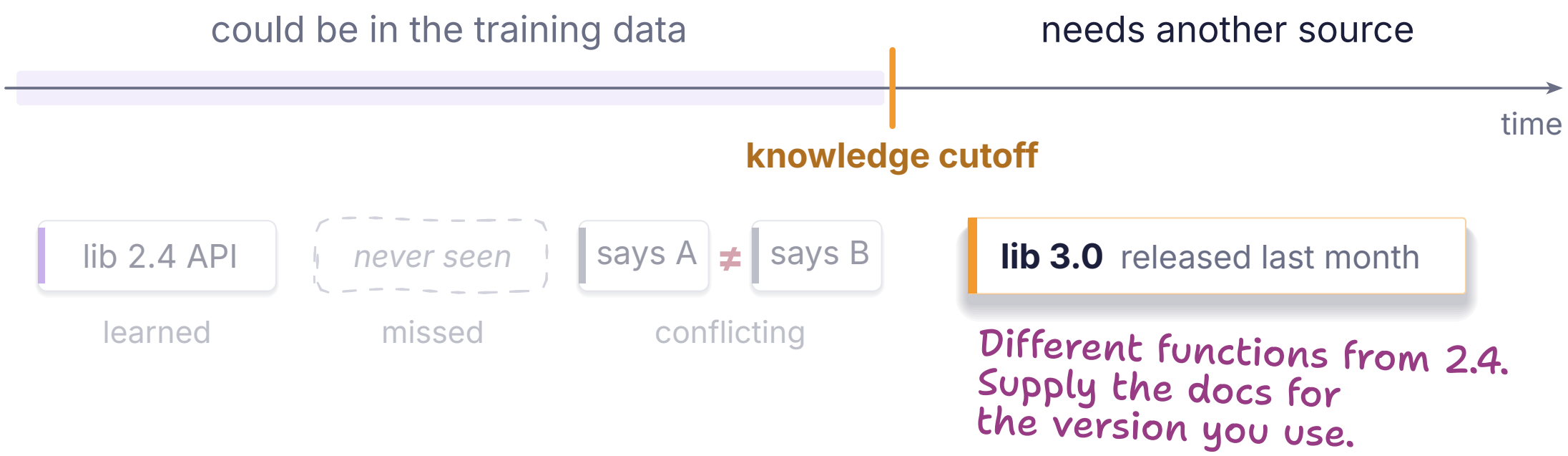
Training knowledge stops at a cutoff



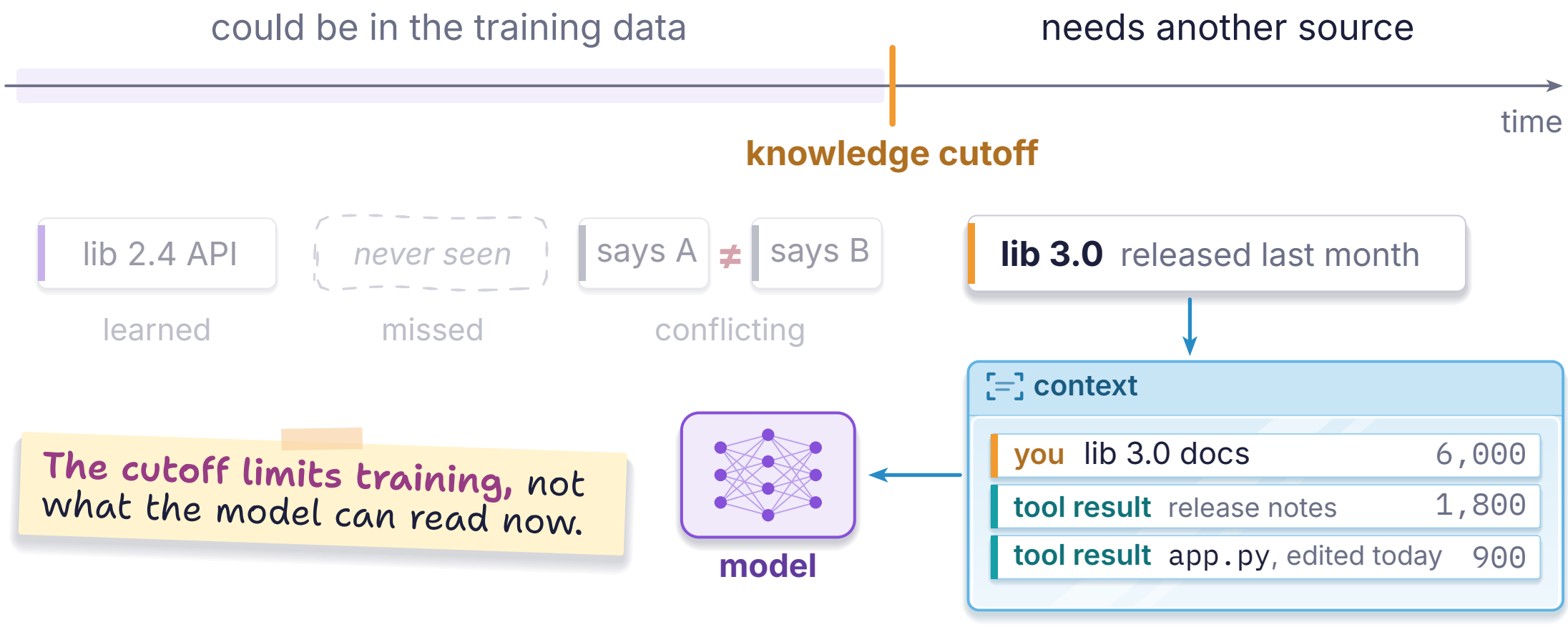
Training knowledge stops at a cutoff



Training knowledge stops at a cutoff



Training knowledge stops at a cutoff



In code: invented methods and packages

agent session `~/sales`

you typed by you

Reshape my pandas DataFrame from wide to long, and suggest a CSV helper.

In code: invented methods and packages

agent session `~/sales`

you typed by you

Reshape my pandas DataFrame from wide to long, and suggest a CSV helper.

agent written by the model

Reshape it with:

```
df.unpivot(index="region")
```

For reading CSV files, try:

```
pip install kewl-utils
```

In code: invented methods and packages

agent session `~/sales`

you typed by you

Reshape my pandas DataFrame from wide to long, and suggest a CSV helper.

agent written by the model

Reshape it with:
`df.unpivot(index="region")` ⓐ
For reading CSV files, try:
`pip install kewl-utils` ⓑ

shell

```
~/sales $ pip install kewl-utils
ERROR: No matching distribution
found for kewl-utils

~/sales $ python reshape.py
AttributeError: 'DataFrame'
object has no attribute 'unpivot'
```

ⓐ pandas calls this `melt`
ⓑ an invented package name

A clean install alone does not prove it is the package you meant.

In code: invented methods and packages

agent session `~/sales`

you typed by you

Reshape my pandas DataFrame from wide to long, and suggest a CSV helper.

agent written by the model

Reshape it with:

`df.unpivot(index="region")` ⓐ

For reading CSV files, try:

`pip install kewl-utils` ⓑ

Give it the docs and types for the version you use. Then check the code with your tools.

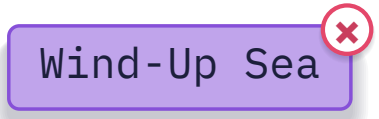
package suggestions that did not exist

about 1 in 5

across the models and tasks tested

Spracklen et al., USENIX Security 2025

Four limits, one way to work with them



Hallucination

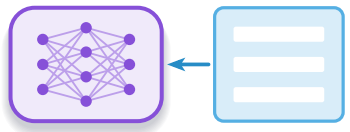
A plausible continuation can be false. Fluent wording is not evidence.

Four limits, one way to work with them



Hallucination

A plausible continuation can be false. Fluent wording is not evidence.



Memory vs. context

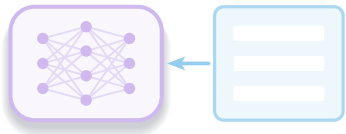
Provide the text instead of relying on recall. Check how it is used.

Four limits, one way to work with them



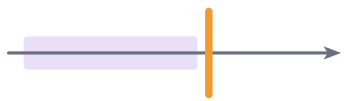
Hallucination

A plausible continuation can be false. Fluent wording is not evidence.



Memory vs. context

Provide the text instead of relying on recall. Check how it is used.



Knowledge cutoff

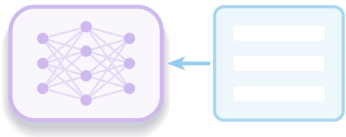
Bring new events and library releases into the context.

Four limits, one way to work with them



Hallucination

A plausible continuation can be false. Fluent wording is not evidence.

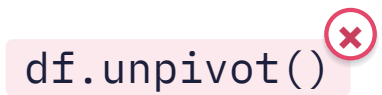


Memory vs. context

Provide the text instead of relying on recall. Check how it is used.

Knowledge cutoff

Bring new events and library releases into the context.



Invented APIs

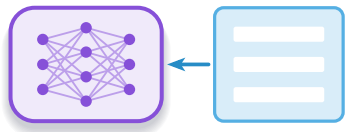
Methods and packages that don't exist. Your tools often catch them.

Four limits, one way to work with them



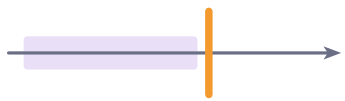
Hallucination

A plausible continuation can be false. Fluent wording is not evidence.



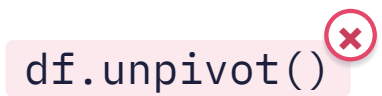
Memory vs. context

Provide the text instead of relying on recall. Check how it is used.



Knowledge cutoff

Bring new events and library releases into the context.



Invented APIs

Methods and packages that don't exist. Your tools often catch them.

Put what matters into the context. **Verify the output afterward.**