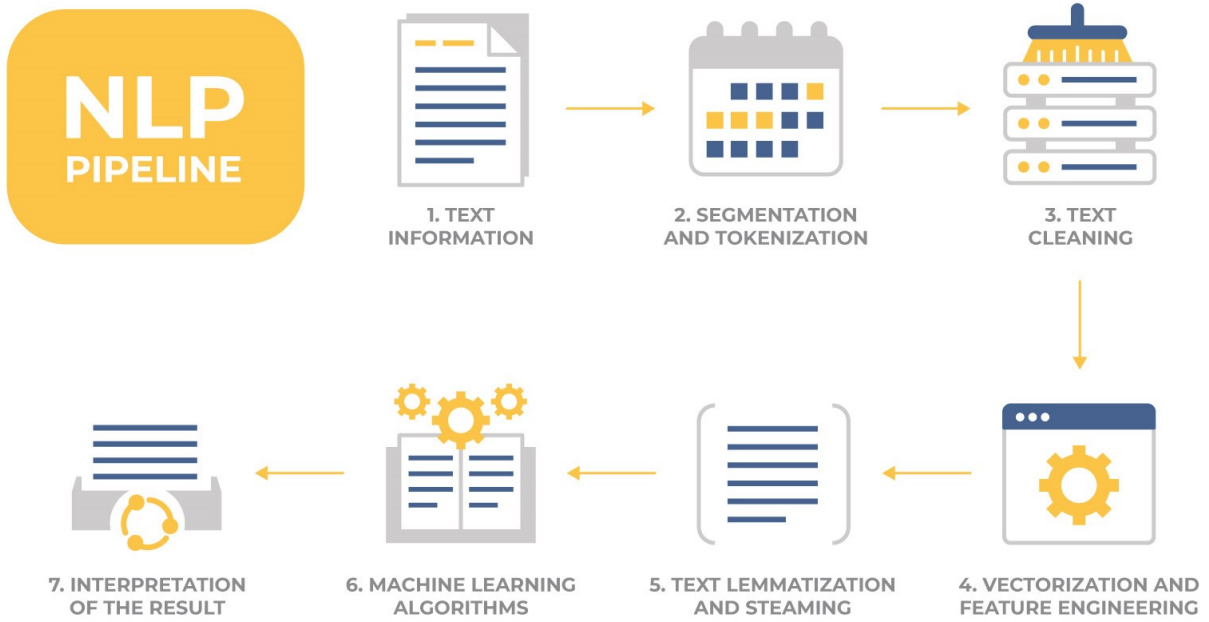
Learning How to Ask: Querying LMs with Mixtures of Soft Prompts

Guanghui Qin and Jason Eisner

Johns Hopkins University

LMs Already Did Your Job



LMs Already Did Your Job

Prompt: _____ x died in _____ y .

Input

BERT

GPT-3

Task: What year did Mary Cassatt die?

LMs Already Did Your Job

Prompt: Mary Cassatt _____ x died in _____ 1926 y .

Input

Output

BERT

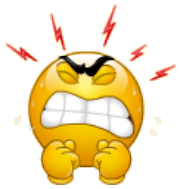
GPT-3

Task: What year did Mary Cassatt die?

The Form of the Question Matters

Opinion Poll

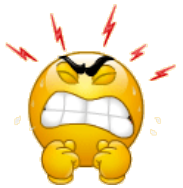
- Should employers be **forced** to negotiate with unions?



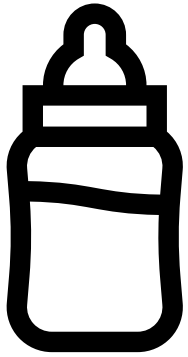
- Should unions have the **right** to negotiate with employers?

The Form of the Question Matters

Opinion Poll



- Should employers be **forced** to negotiate with unions?
- Should unions have the **right** to negotiate with employers?



Children's Mind

- Why **can't we** see the moon now?
- Where does **the moon** go during the day?

The Form of the Question Matters

Opinion Poll

Children's Mind

We're prompting LMs!

BERT

GPT-3

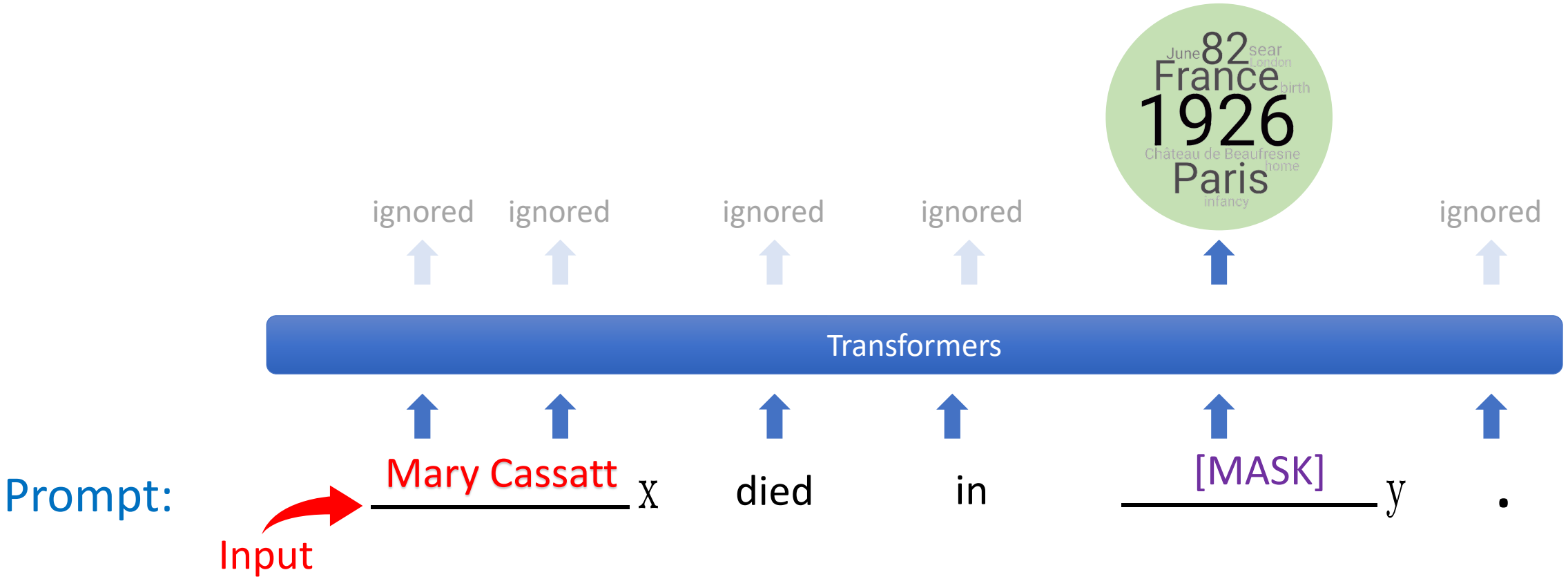
Prompt LMs

Prompt: Mary Cassatt _____ x died in _____ [MASK] y .

Input

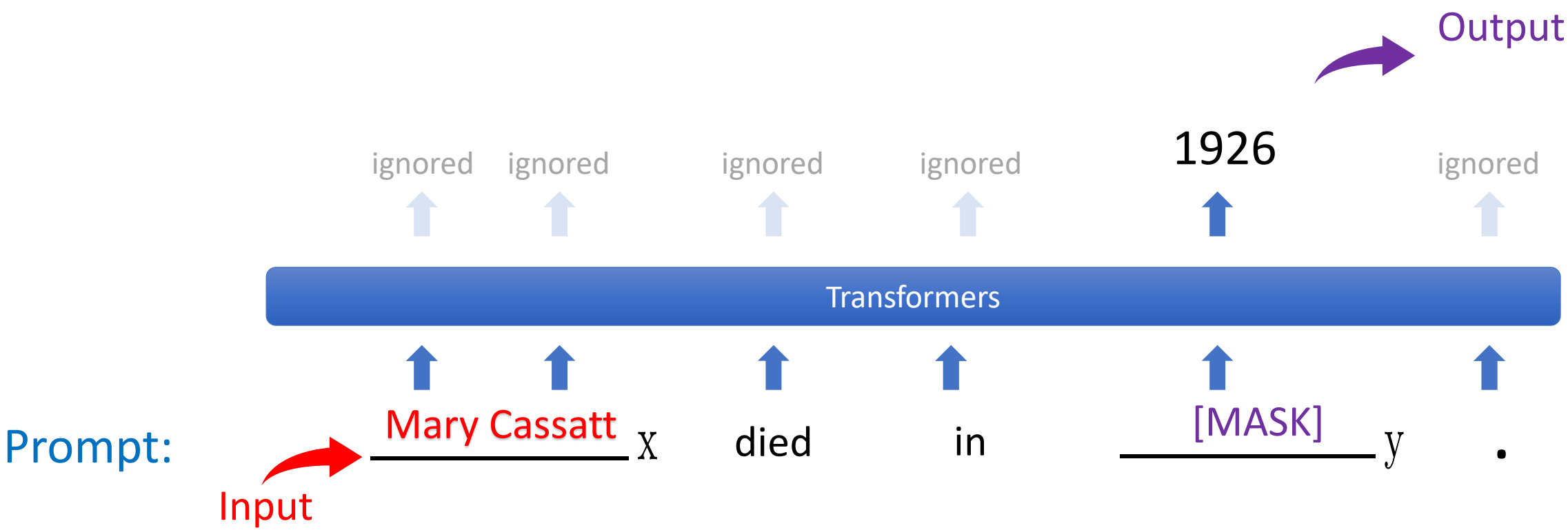
Task: Fact Queries: Year-of-Death

Prompt LMs



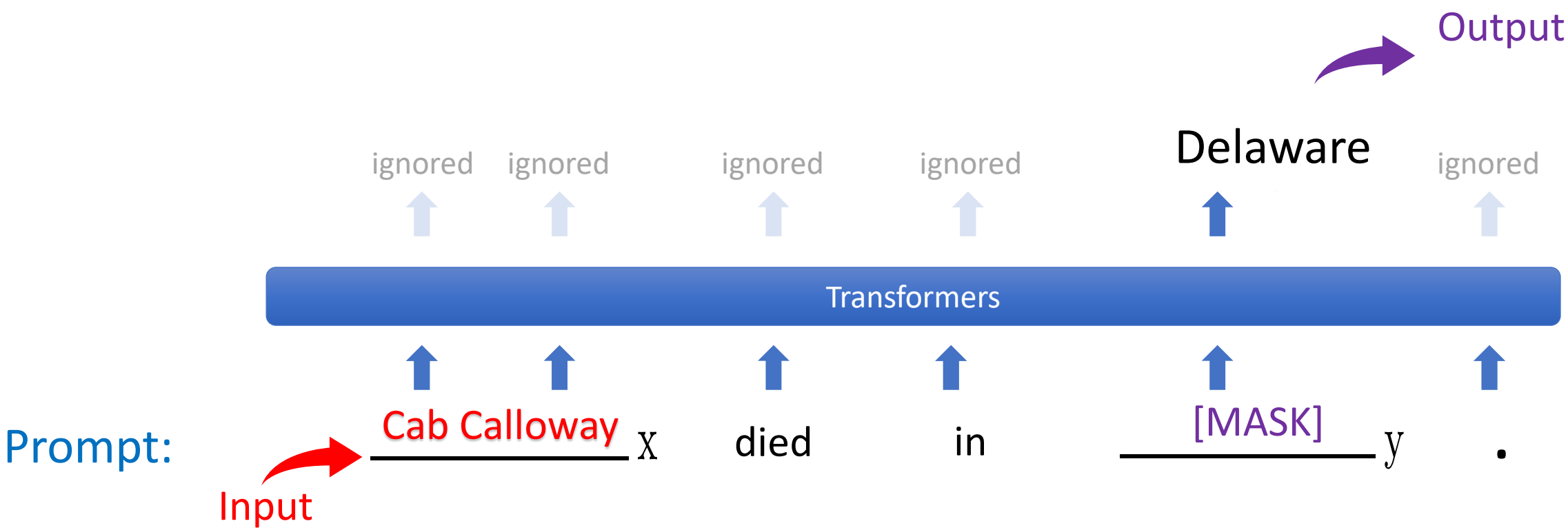
Task: Fact Queries: Year-of-Death

Prompt LMs



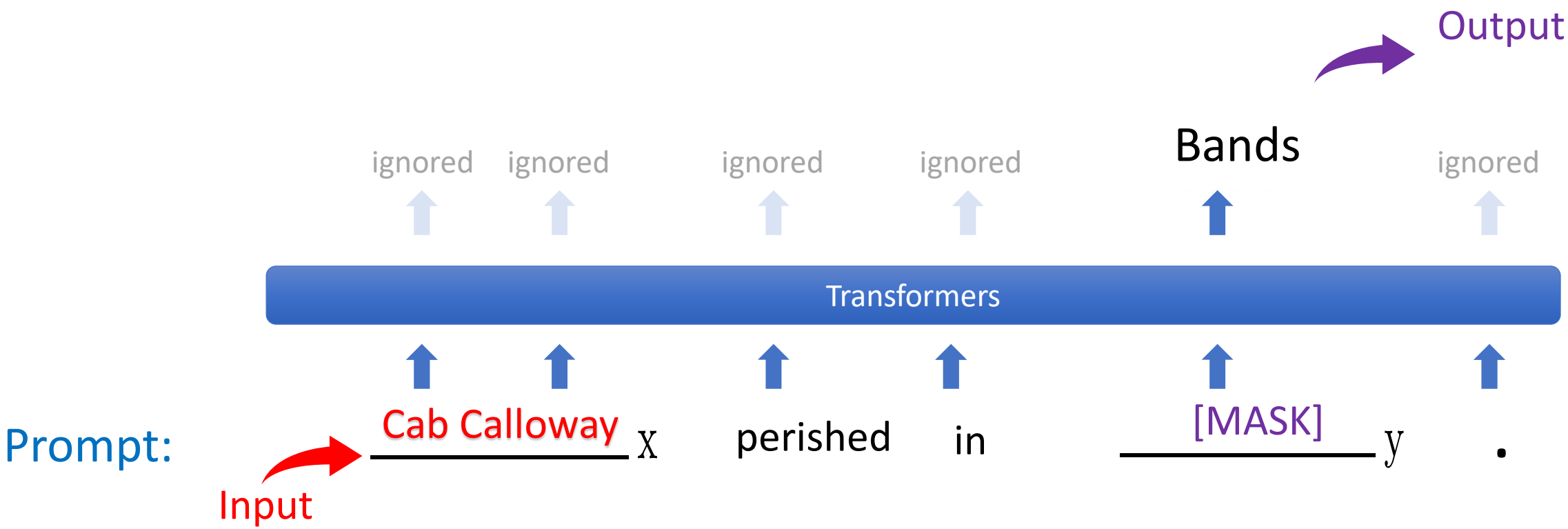
Task: Fact Queries: Year-of-Death

Prompt LMs



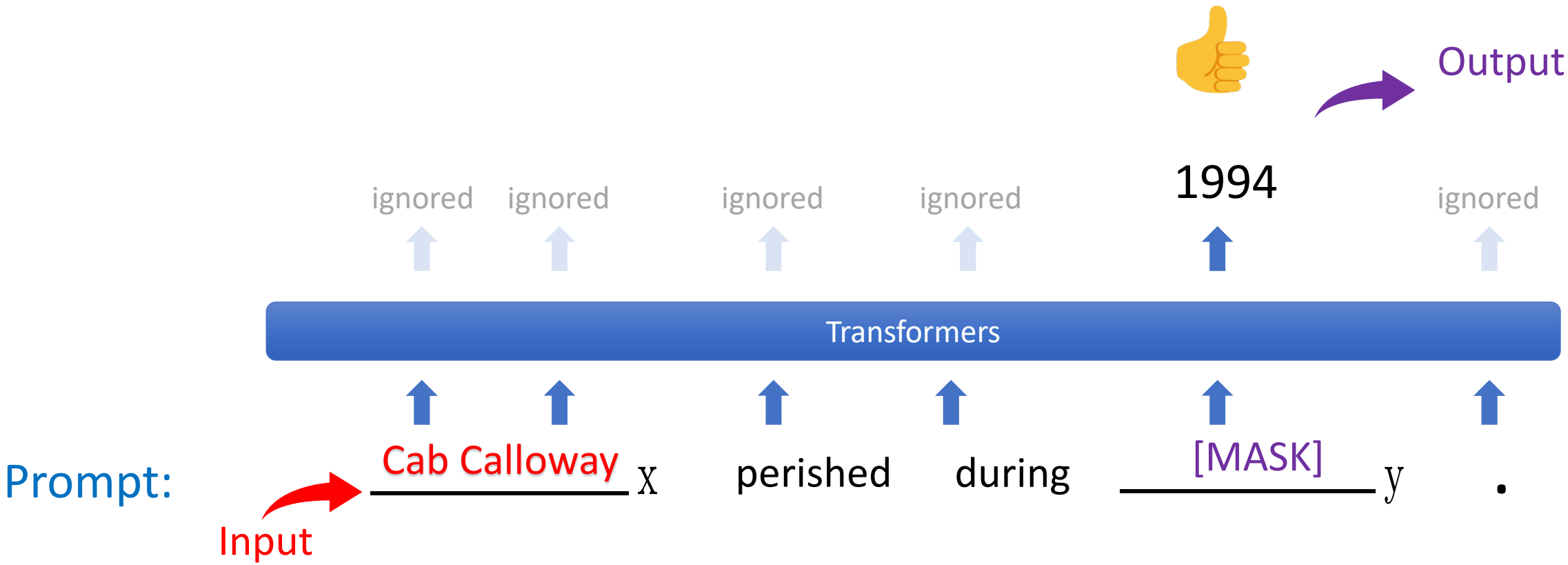
Task: Fact Queries: Year-of-Death

Prompt LMs



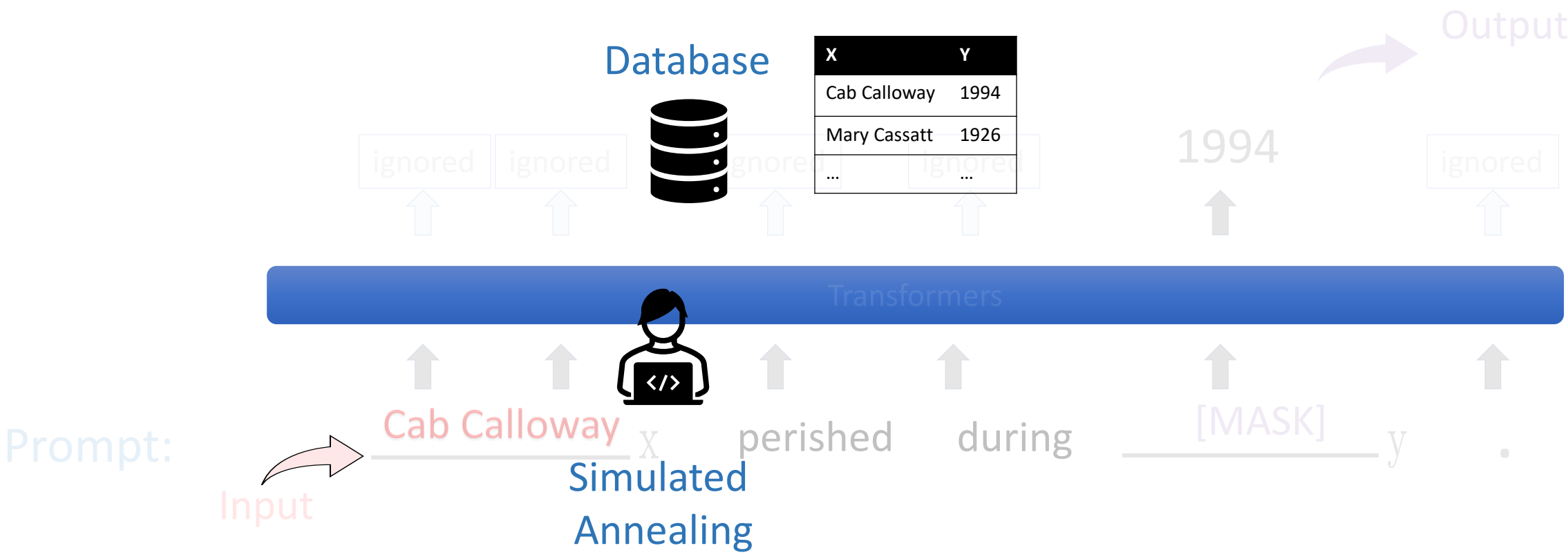
Task: Fact Queries: Year-of-Death

Prompt LMs



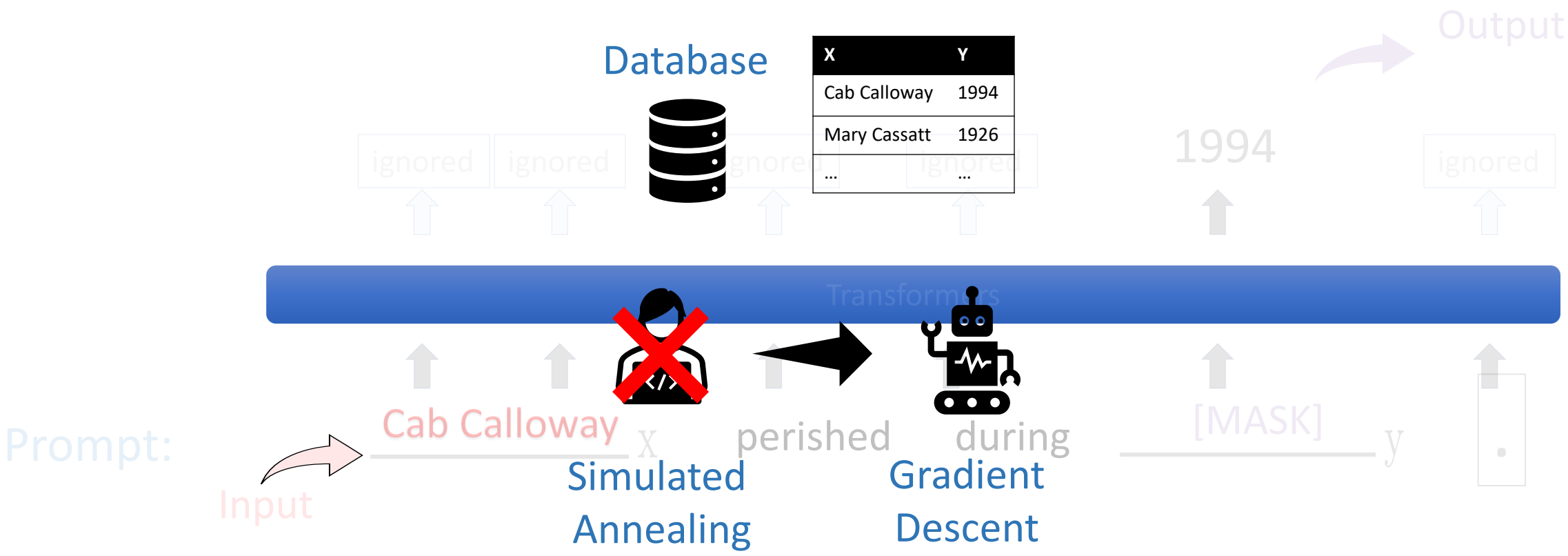
Task: Fact Queries: Year-of-Death

Prompt LMs



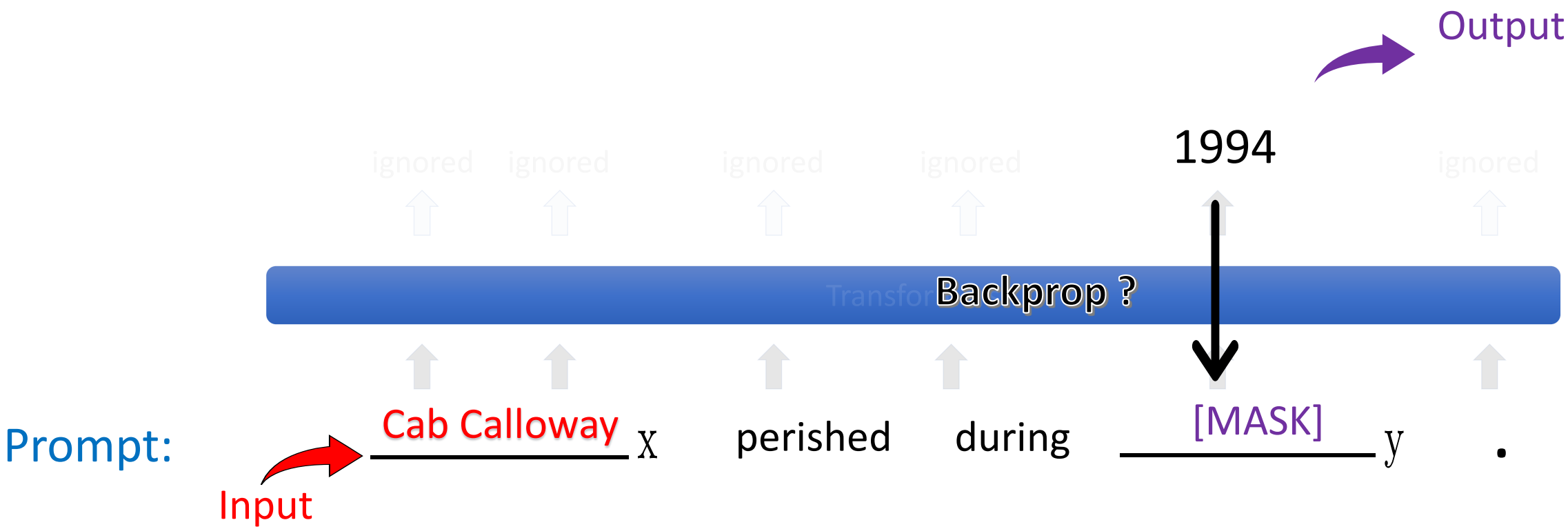
Task: Fact Queries: Year-of-Death

Prompt LMs



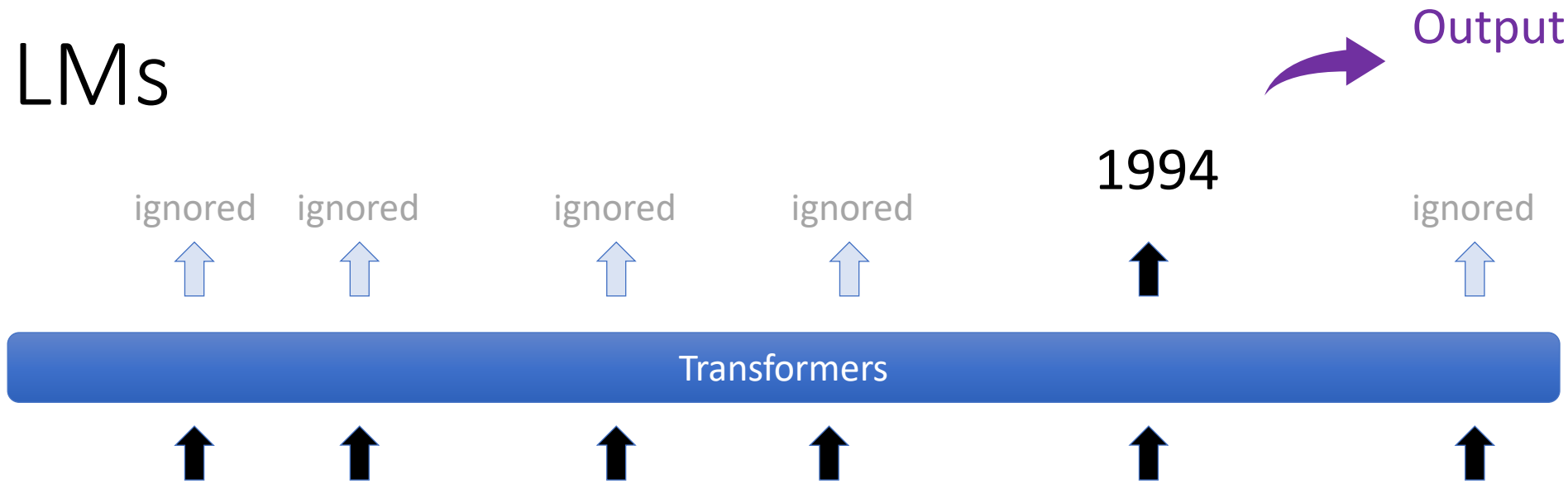
Task: Fact Queries: Year-of-Death

Prompt LMs



Task: Fact Queries: Year-of-Death

Prompt LMs



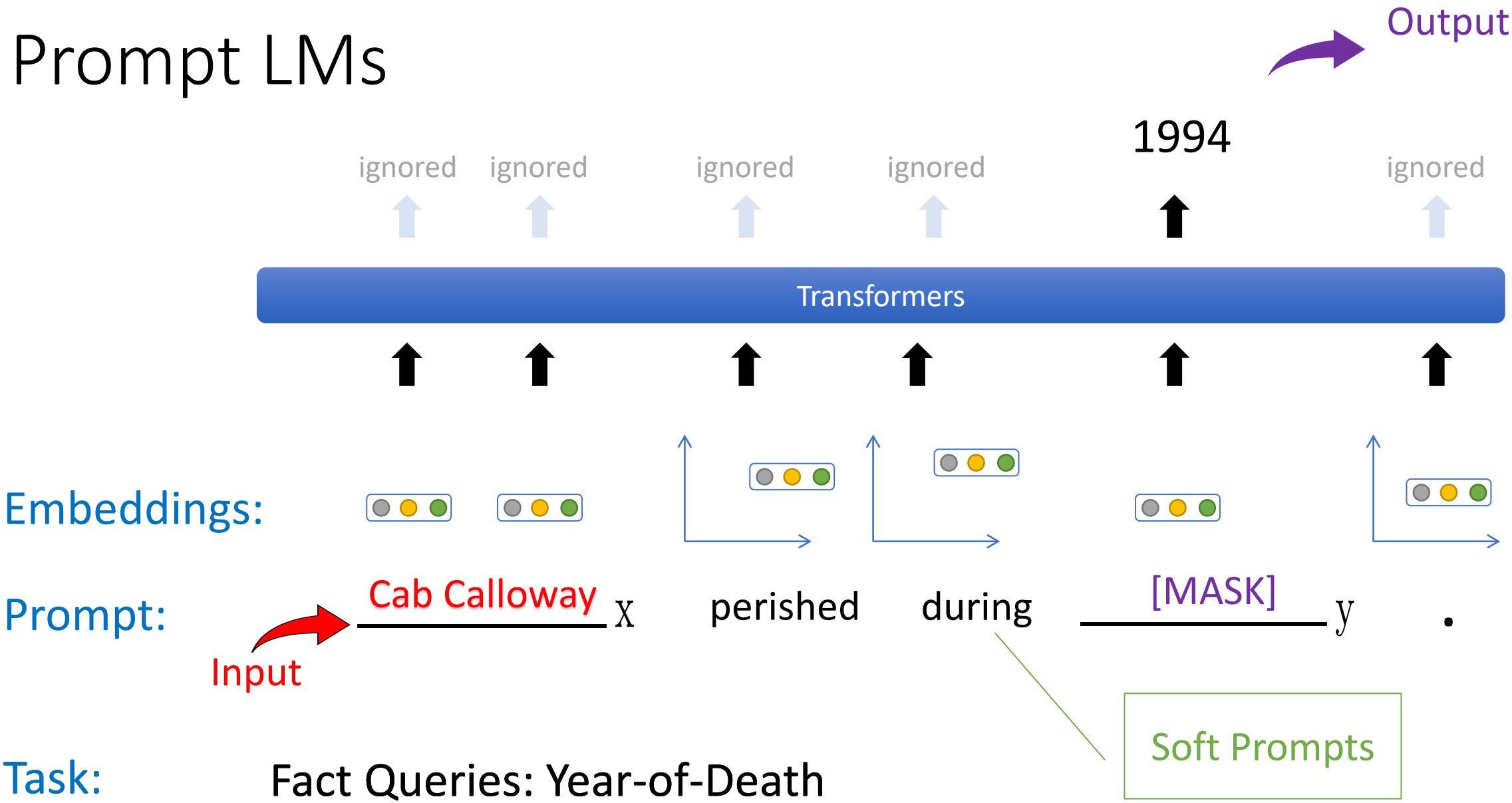
Embeddings:

Prompt: Cab Calloway _____ x perished during _____ [MASK] y .

Input

Task: Fact Queries: Year-of-Death

Prompt LMs


Output
1994
ignored
ignored
ignored
ignored
ignored
Transformers
Embeddings:
Prompt:
Cab Calloway
x
perished
during
[MASK]
y
.
Input
Soft Prompts
Task:
Fact Queries: Year-of-Death

Why Soft Prompts?

Cab Calloway x played until his death in [MASK] y .

Why Soft Prompts?

- Easy to search with backprop.

Cab Calloway x played until his death in [MASK] y .

Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.

Cab Calloway x played until his death in [MASK] y .

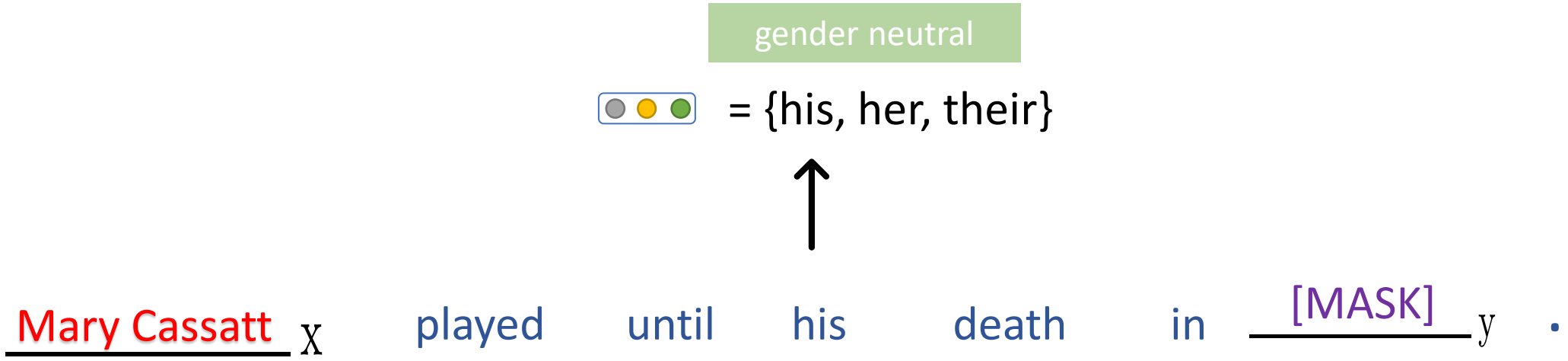
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.

Mary Cassatt x played until his death in [MASK] y .

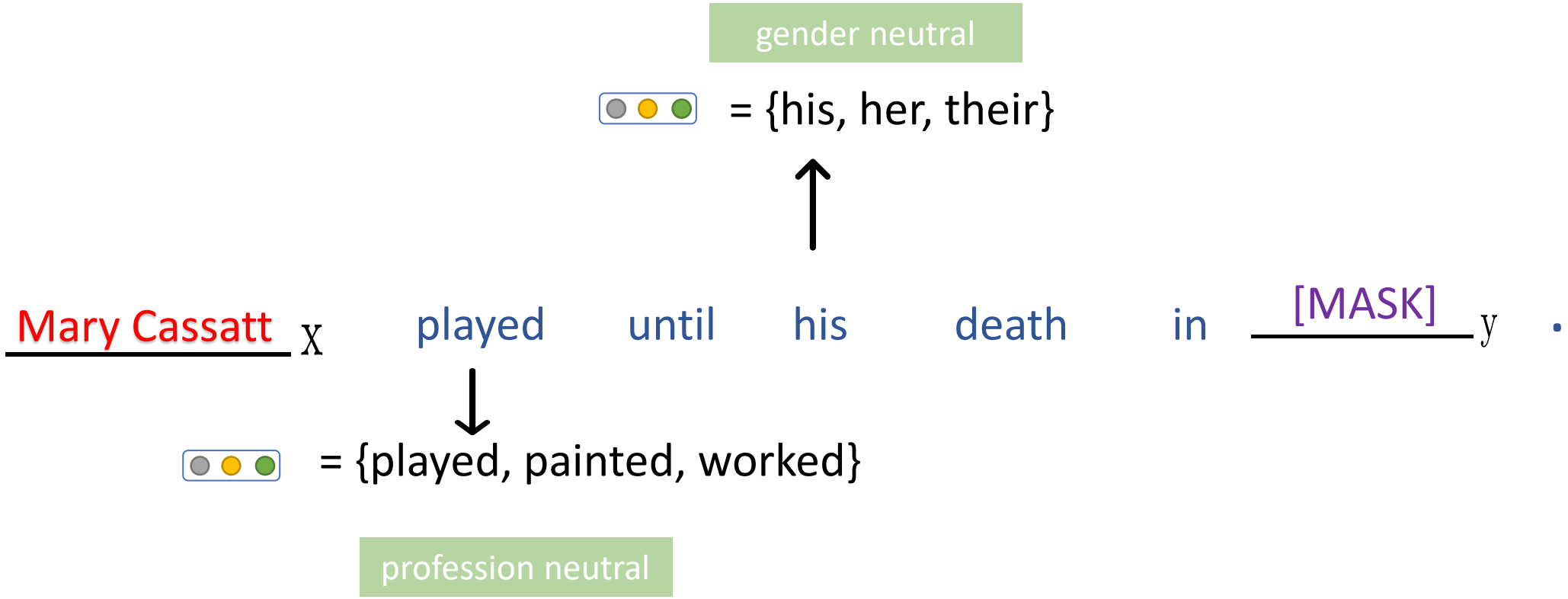
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.



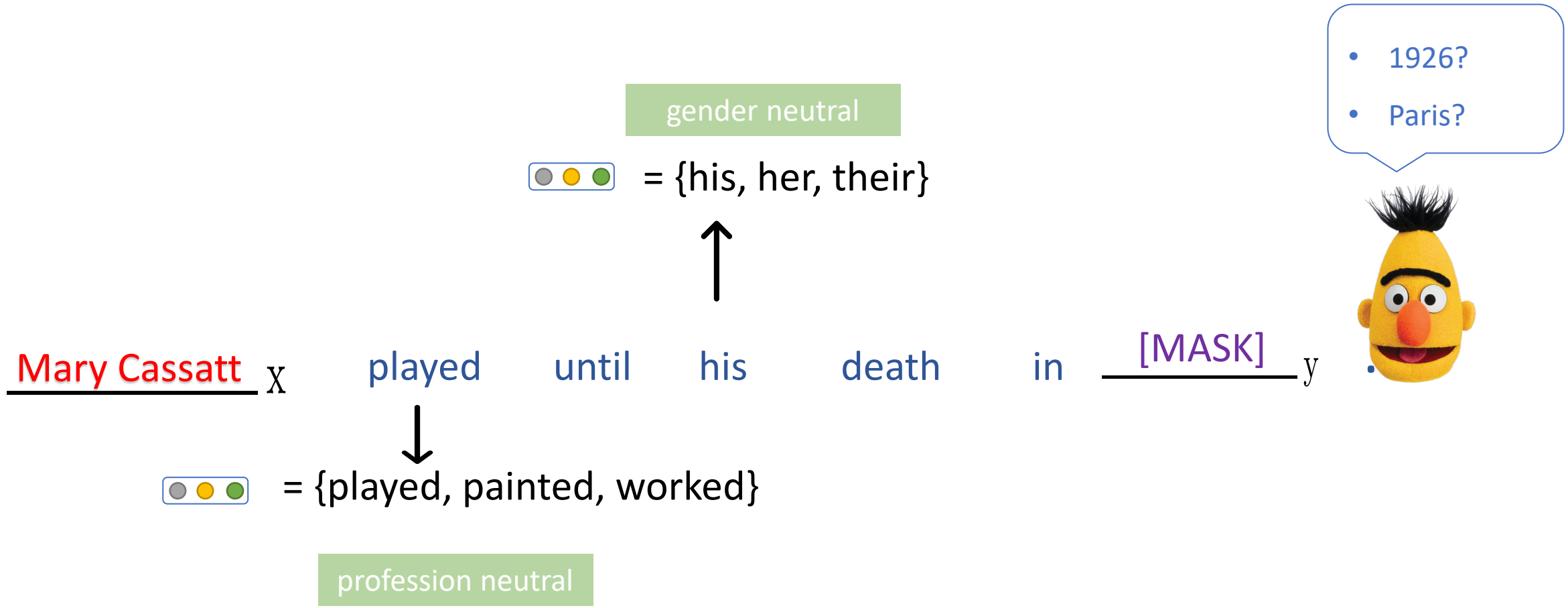
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.



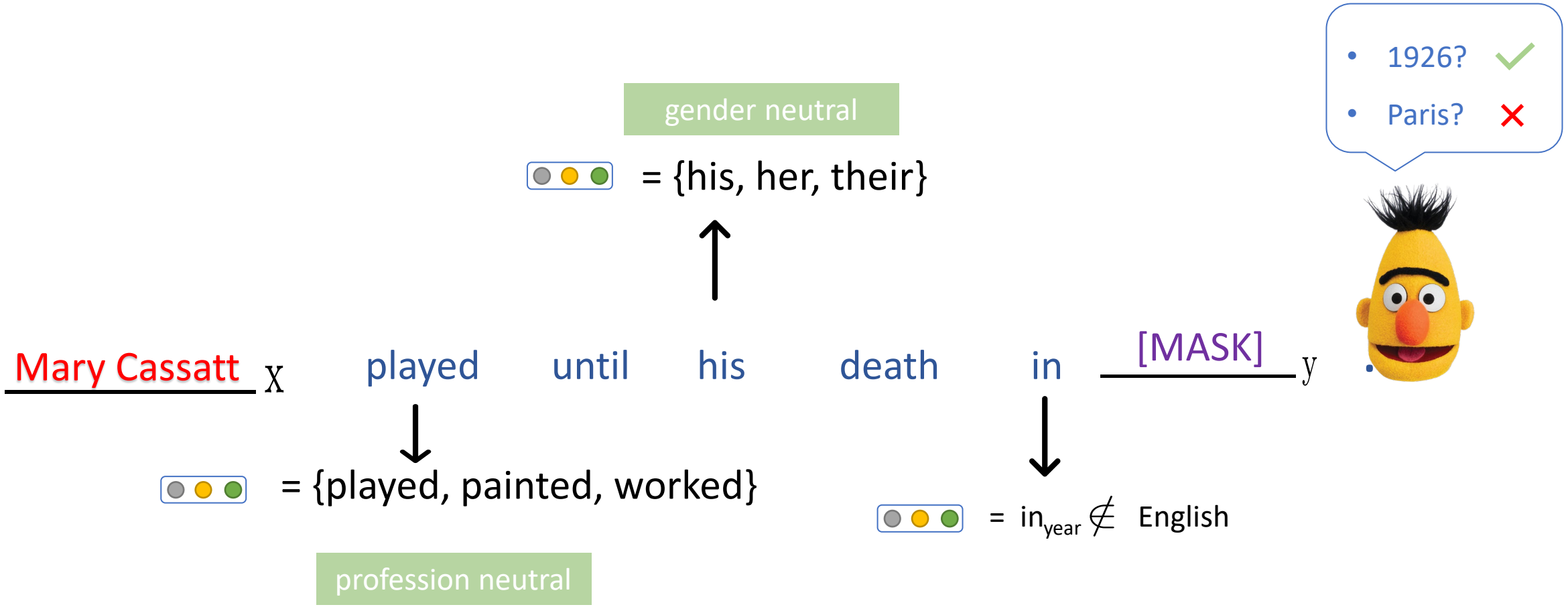
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.



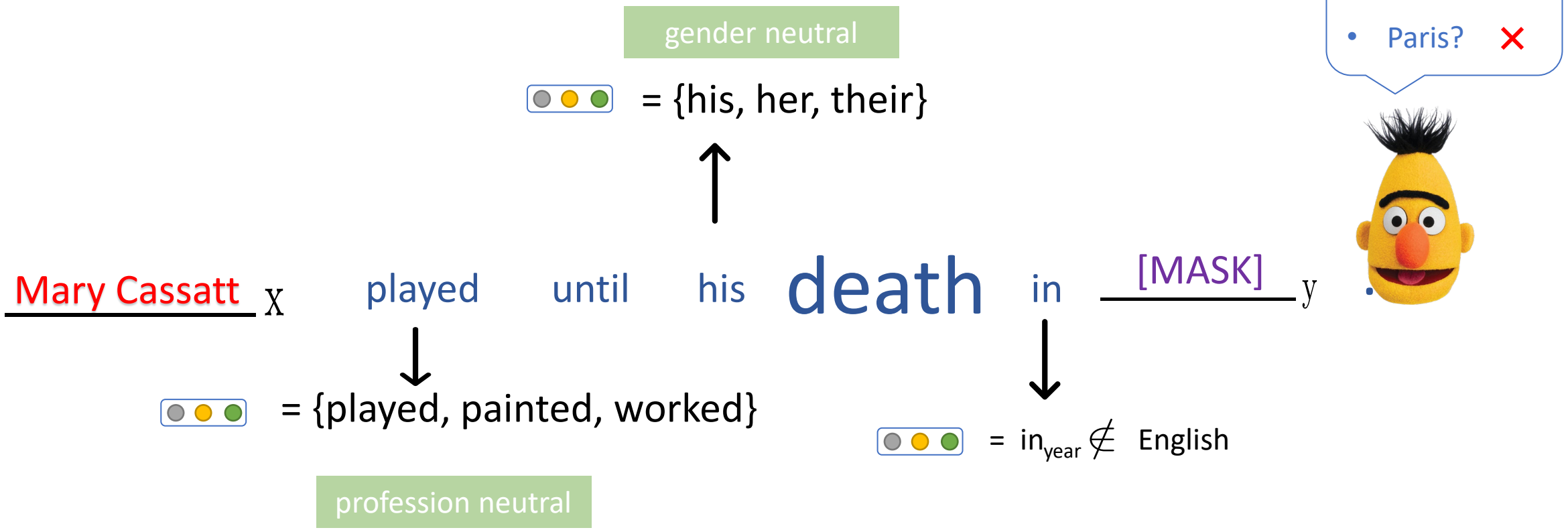
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.



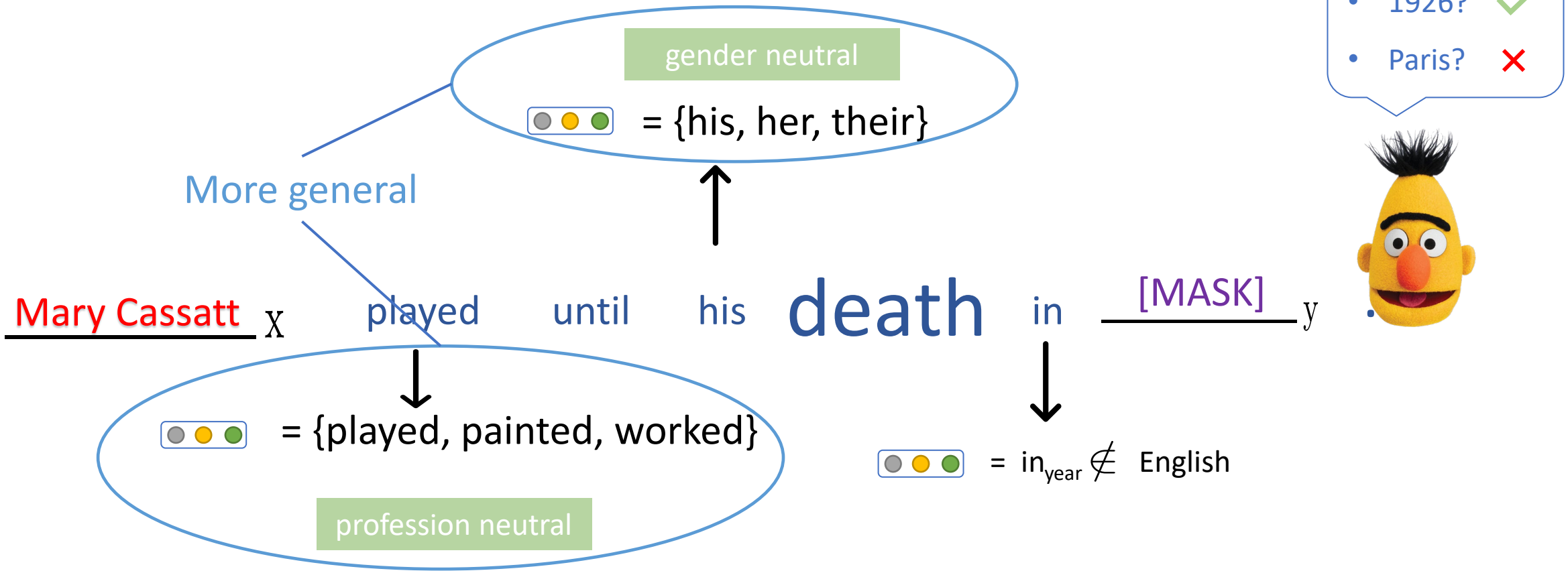
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.
- They can emphasize certain keywords, even particular dimensions.



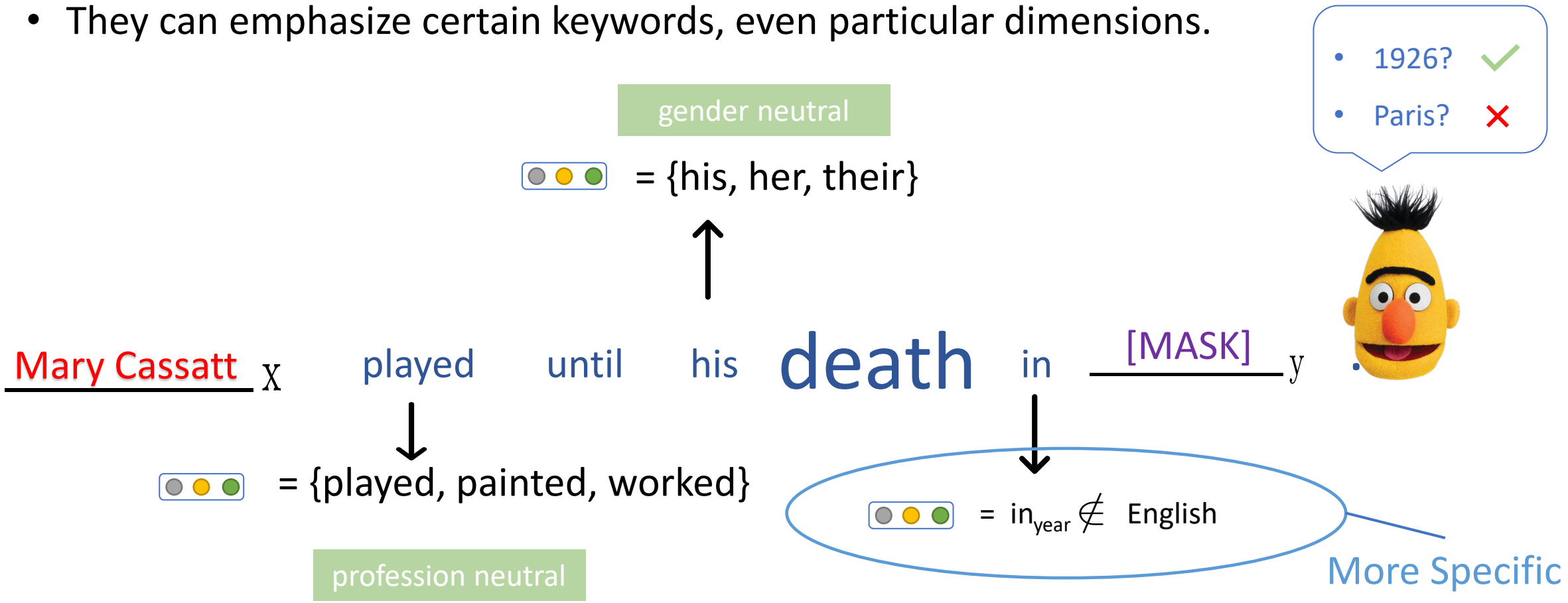
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.
- They can emphasize certain keywords, even particular dimensions.



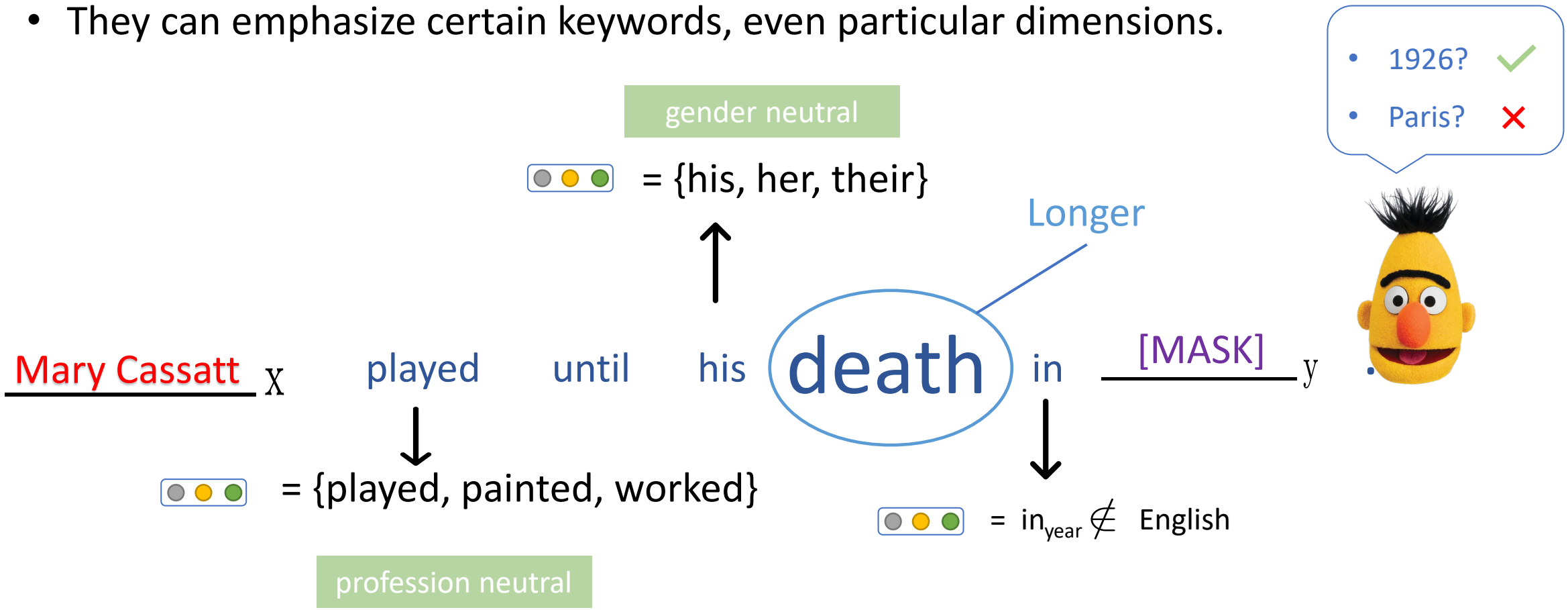
Why Soft Prompts?

- Easy to search with backprop.
- We have a larger space of prompts.
- They can emphasize certain keywords, even particular dimensions.

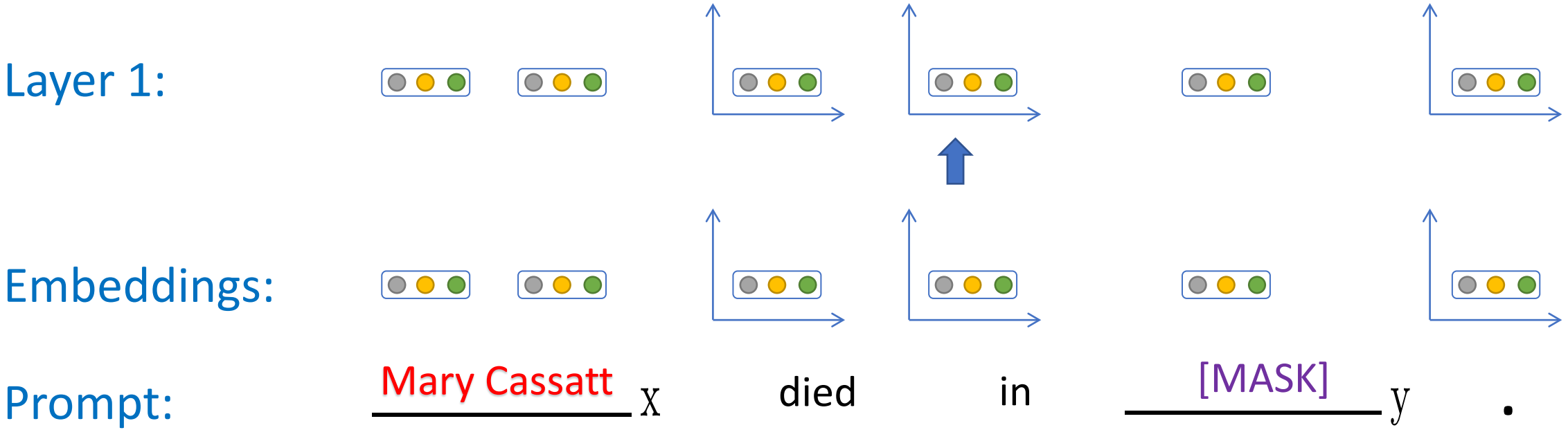


Why Soft Prompts?

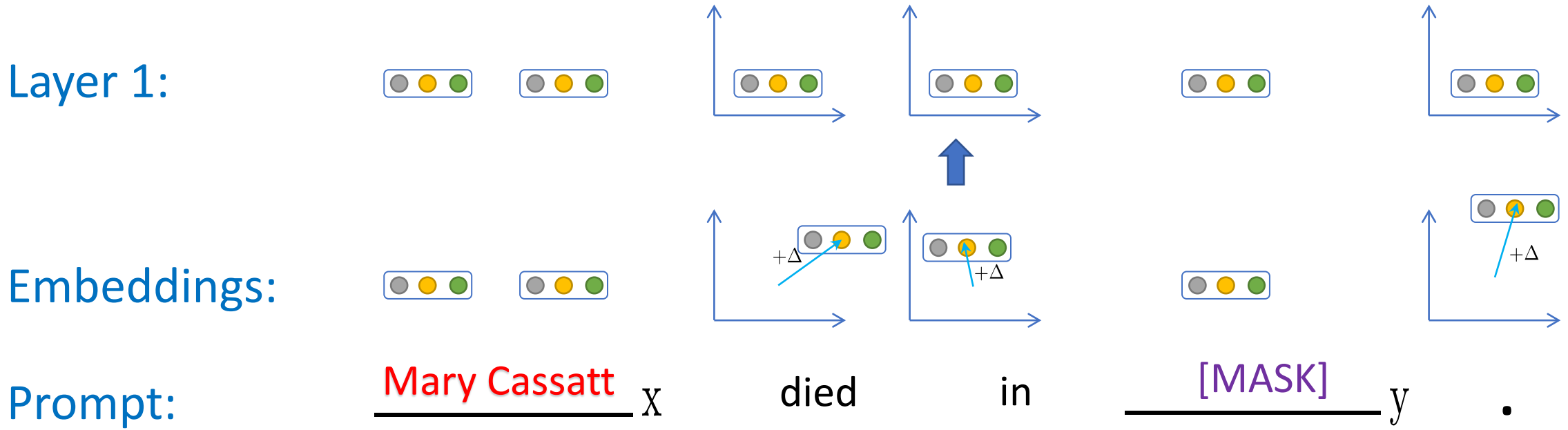
- Easy to search with backprop.
- We have a larger space of prompts.
- They can emphasize certain keywords, even particular dimensions.



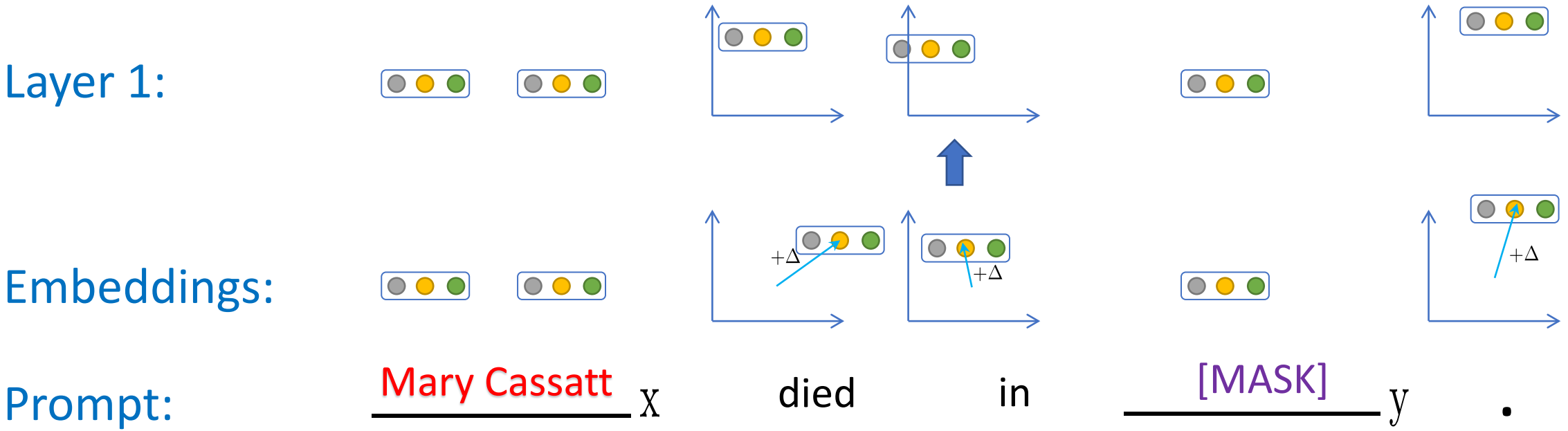
Deep Perturbation of Prompts



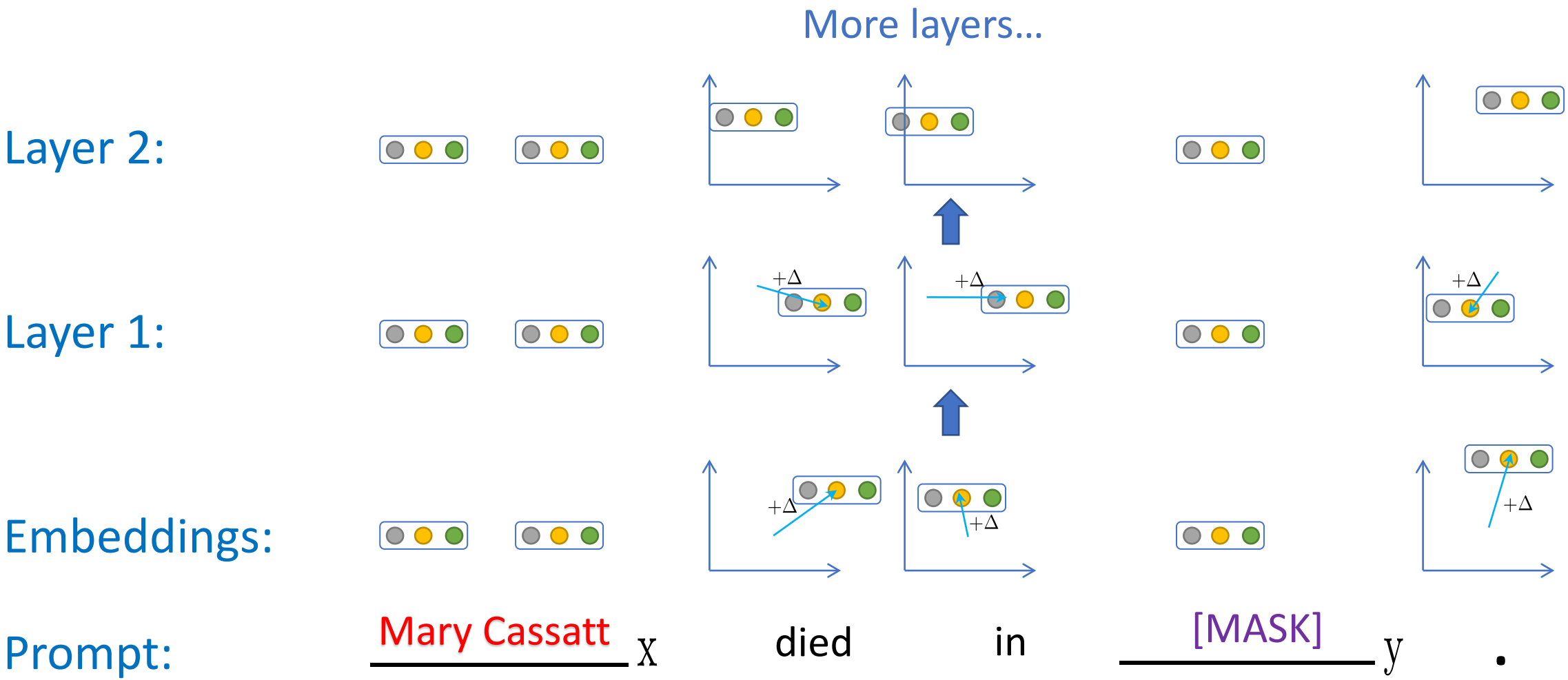
Deep Perturbation of Prompts



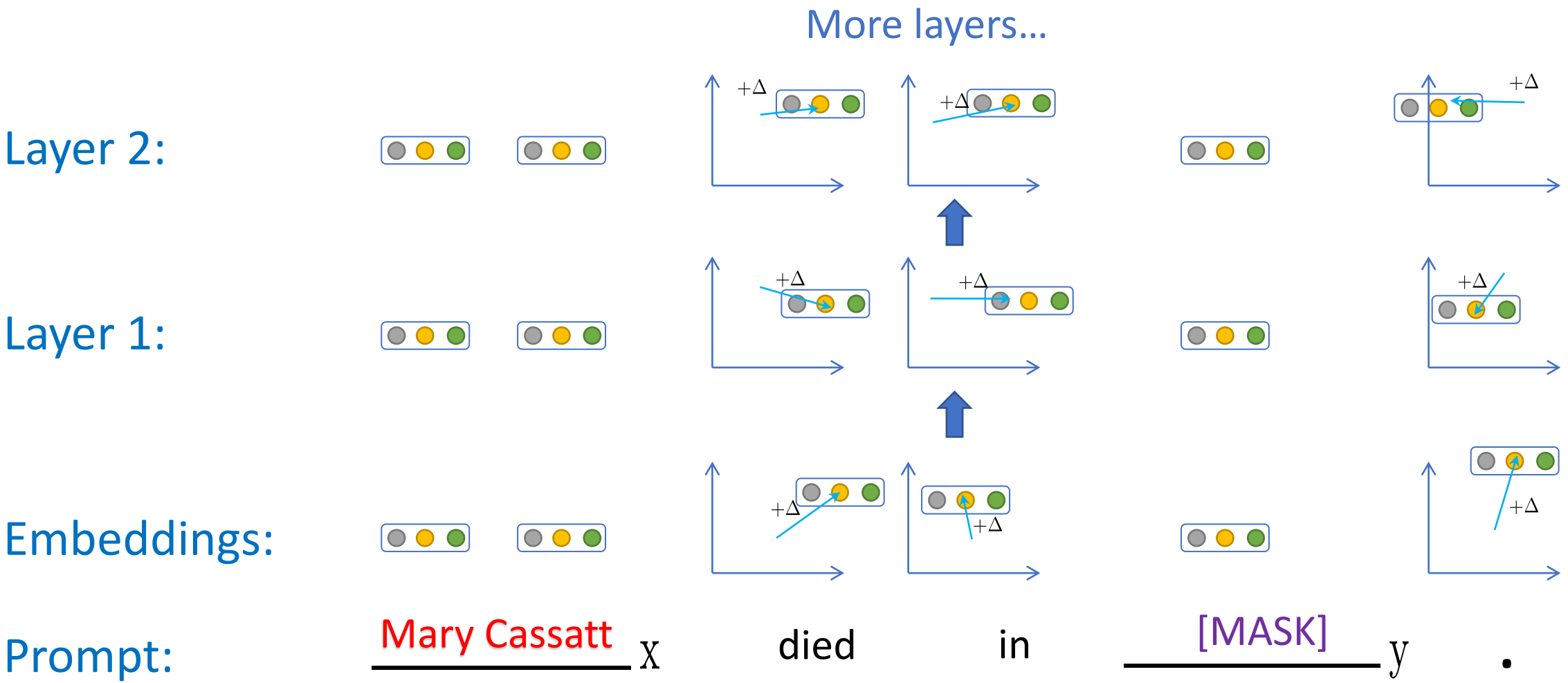
Deep Perturbation of Prompts



Deep Perturbation of Prompts

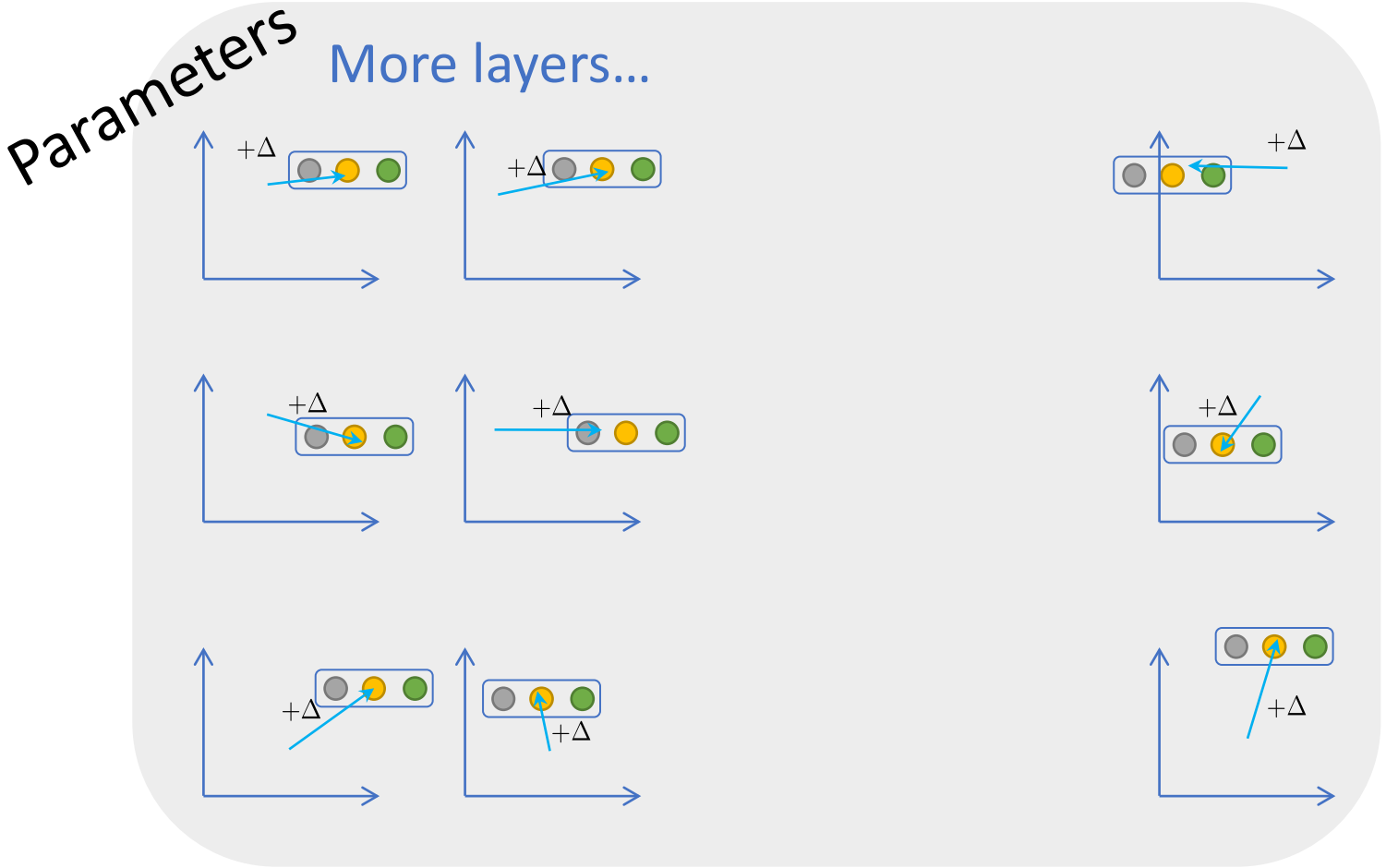


Deep Perturbation of Prompts

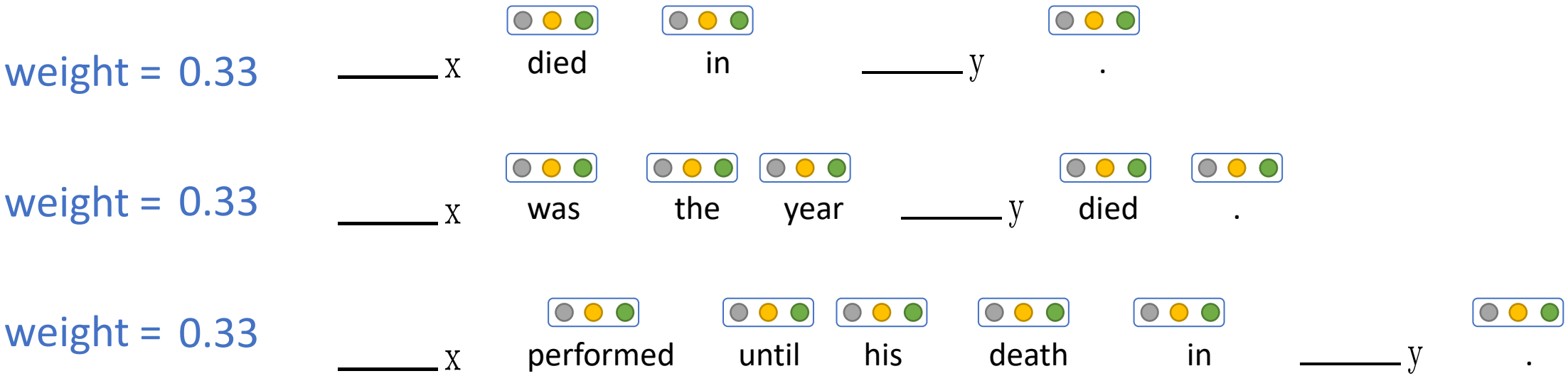


Deep Perturbation of Prompts

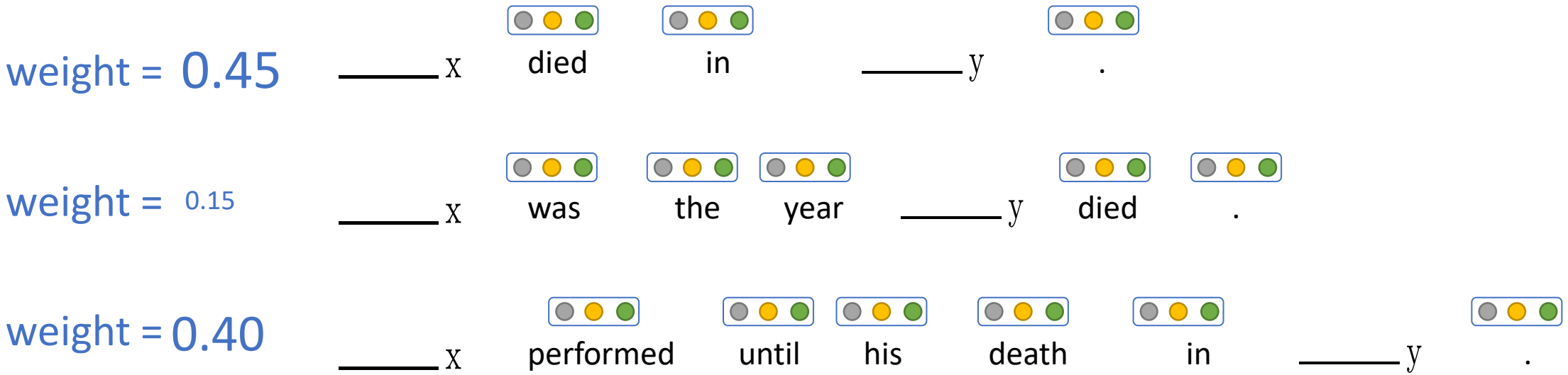
#param ratio	
soft-prompt	1
deep-perturbation	25
fully-fine-tuning	1e5



Ensembling (Mixture of Prompts)



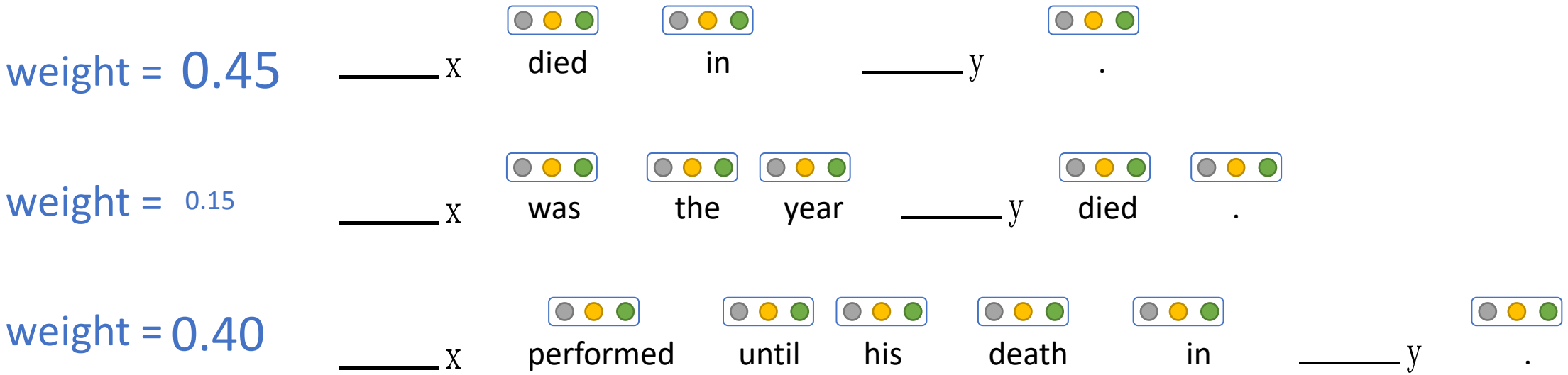
Ensembling (Mixture of Prompts)



$$p(y \mid x) = \sum_{\text{prompt}} p(y \mid \text{prompt}(x)) \cdot p(\text{prompt})$$

Ensembling (Mixture of Prompts)

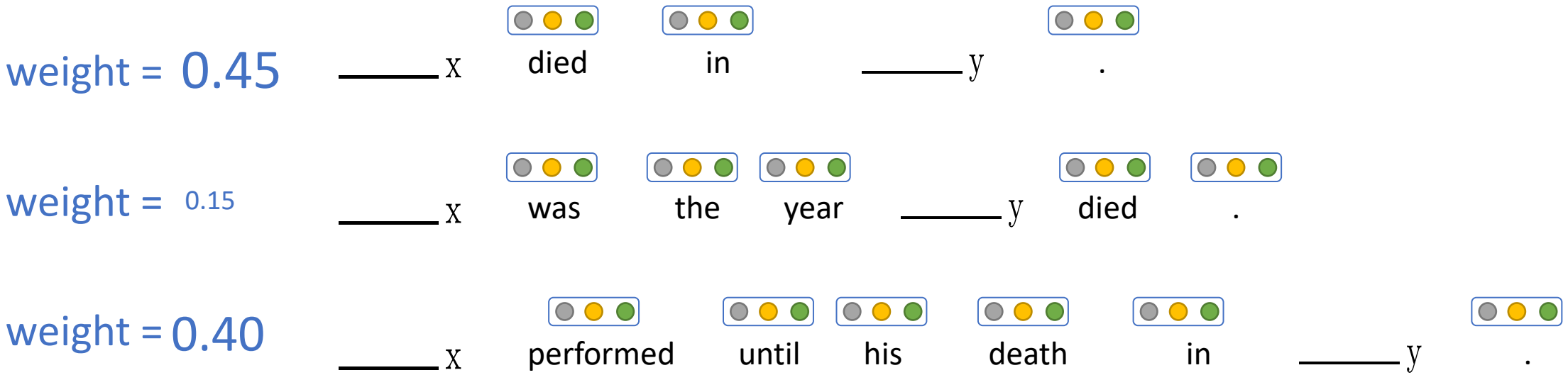
- Bigger model: When one prompt is unsure, others can help



$$p(y \mid x) = \sum_{\text{prompt}} p(y \mid \text{prompt}(x)) \cdot p(\text{prompt})$$

Ensembling (Mixture of Prompts)

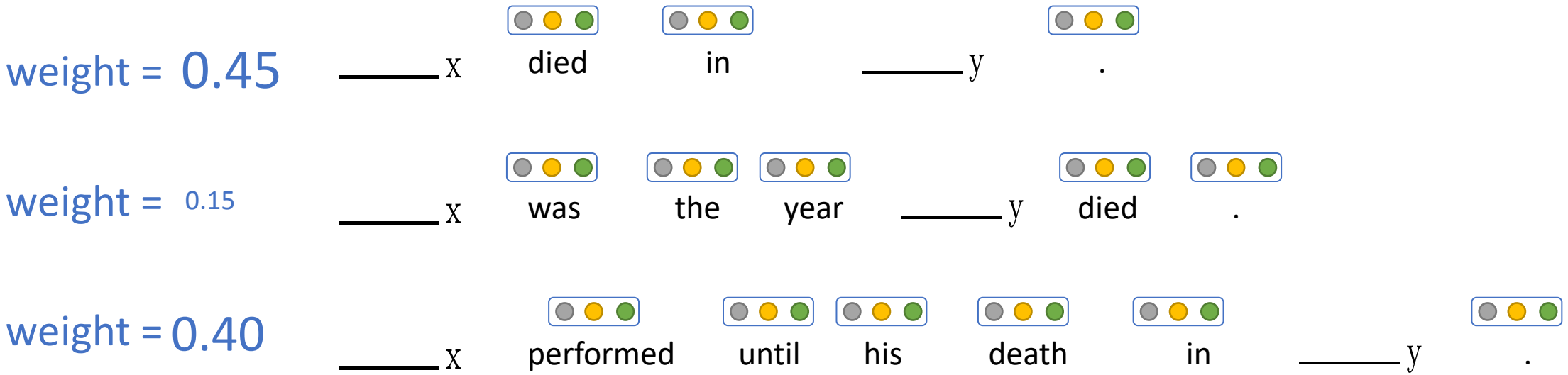
- Bigger model: When one prompt is unsure, others can help
- Better predictions: Ensembling reduces variance



$$p(y \mid x) = \sum_{\text{prompt}} p(y \mid \text{prompt}(x)) \cdot p(\text{prompt})$$

Ensembling (Mixture of Prompts)

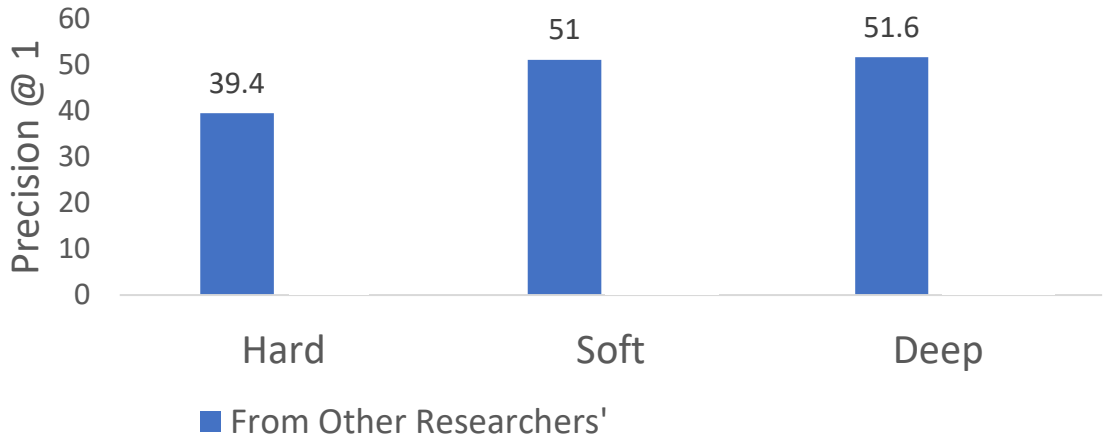
- Bigger model: When one prompt is unsure, others can help
- Better predictions: Ensembling reduces variance
- Better optimization: Explore multiple starting points in parallel



$$p(y \mid x) = \sum_{\text{prompt}} p(y \mid \text{prompt}(x)) \cdot p(\text{prompt})$$

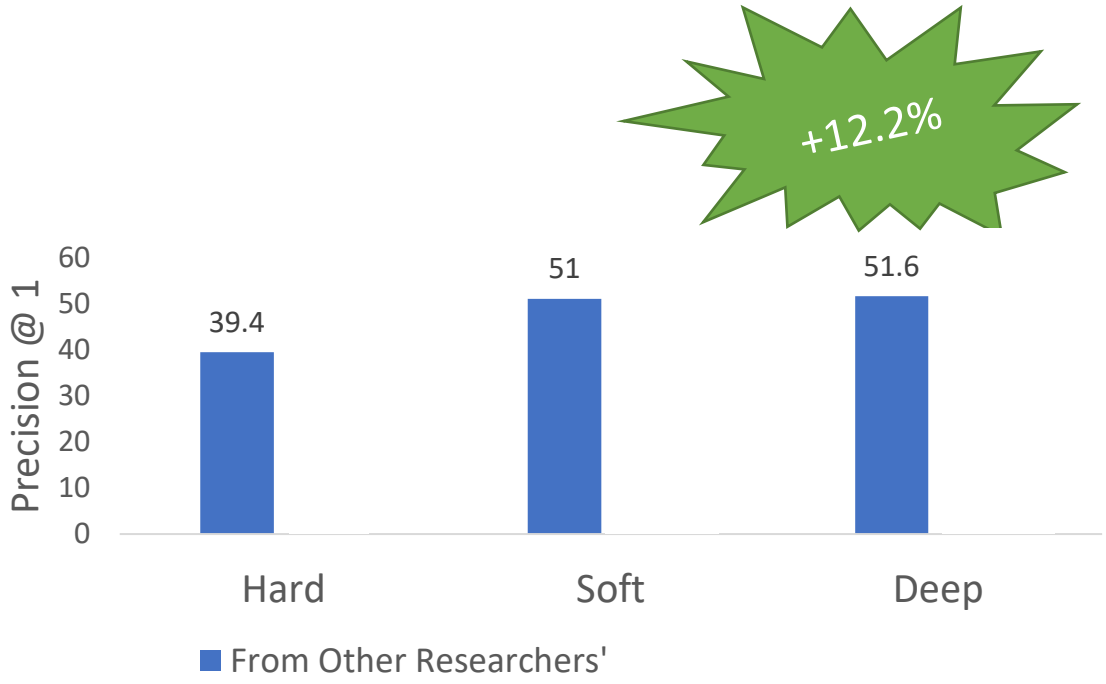
Main Experiments

- Predict factual relations from T-REx dataset by prompting BERT-large
  - About 1000 training examples per relation



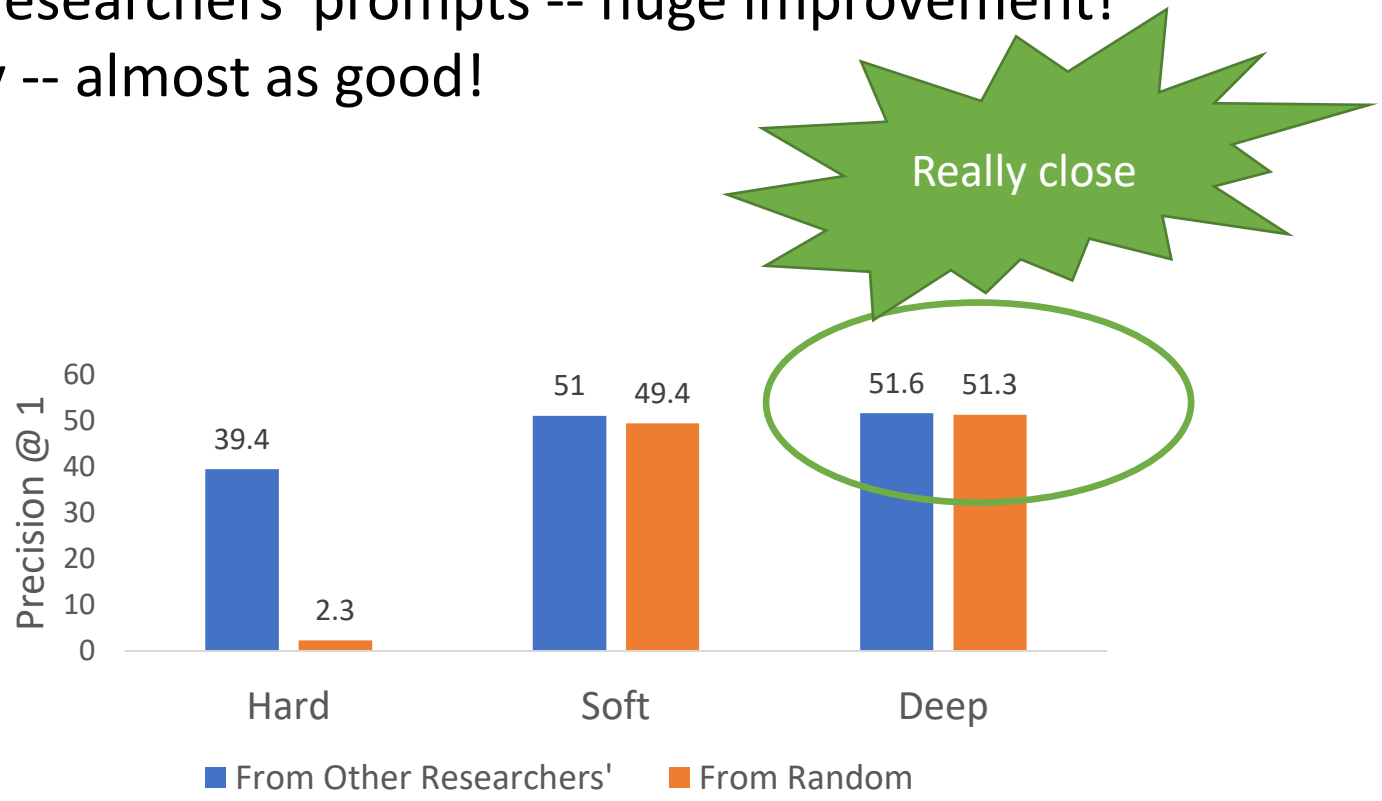
Main Experiments

- Predict factual relations from T-REx dataset by prompting BERT-large
  - About 1000 training examples per relation
- Initialize at other researchers' prompts -- huge improvement!



Main Experiments

- Predict factual relations from T-REx dataset by prompting BERT-large
  - About 1000 training examples per relation
- Initialize at other researchers' prompts -- huge improvement!
- Initialize randomly -- almost as good!



Lots of Experiments

LM	Method	Precision@1	Precision@10	MRR
		init → soft → deep	init → soft → deep	init → soft → deep
BEb	LAMA	31.1	59.5	40.3
	LPAQA	34.1	62.0	43.6
	Soft (sin.)	31.1 $\xrightarrow{+14.6^?}$ 45.7 $\xrightarrow{+2.0}$ 47.7	59.5 $\xrightarrow{+16.3^?}$ 75.8 $\xrightarrow{+3.2}$ 79.0	40.3 $\xrightarrow{+15.9^?}$ 56.2 $\xrightarrow{+2.2}$ 58.4
	Soft (min.)	34.1 $\xrightarrow{+14.7^?}$ 48.8 $\xrightarrow{+1.9}$ **50.7**$^?$	62.0 $\xrightarrow{+15.6^?}$ 79.6 $\xrightarrow{+1.1}$ **80.7**$^?$	43.6 $\xrightarrow{+15.8^?}$ 59.4 $\xrightarrow{+1.7}$ **61.1**$^?$
	Soft (par.)	34.1 $\xrightarrow{+12.8^?}$ 46.9 $\xrightarrow{+1.5}$ 48.4	62.0 $\xrightarrow{+16.8^?}$ 78.8 $\xrightarrow{+0.8}$ 79.6	43.6 $\xrightarrow{+14.2^?}$ 57.8 $\xrightarrow{+1.3}$ 59.1
	Soft (ran.)	0.7 $\xrightarrow{+46.6}$ 47.3 $\xrightarrow{+0.8}$ 48.1	4.6 $\xrightarrow{+74.0}$ 79.1 $\xrightarrow{+0.0}$ 79.1	2.3 $\xrightarrow{+56.1}$ 58.4 $\xrightarrow{+0.5}$ 58.9
BEl	LAMA	28.9†	57.7†	38.7†
	LPAQA	39.4†	67.4†	49.1†
	Soft (sin.)	28.9 $\xrightarrow{+16.9}$ 45.8 $\xrightarrow{+5.3}$ 51.1	57.7 $\xrightarrow{+19.0}$ 76.7 $\xrightarrow{+4.4}$ 81.1	38.7 $\xrightarrow{+17.8}$ 56.5 $\xrightarrow{+5.0}$ 61.5
	Soft (min.)	39.4 $\xrightarrow{+11.6}$ 51.0 $\xrightarrow{+0.6}$ **51.6**	67.4 $\xrightarrow{+14.0}$ 81.4 $\xrightarrow{+0.5}$ **81.9**	49.1 $\xrightarrow{+12.5}$ 61.6 $\xrightarrow{+0.5}$ **62.1**
	Soft (par.)	39.4 $\xrightarrow{+9.2}$ 48.6 $\xrightarrow{+2.5}$ 51.1	67.4 $\xrightarrow{+12.6}$ 80.0 $\xrightarrow{+1.7}$ 81.7	49.1 $\xrightarrow{+10.5}$ 59.6 $\xrightarrow{+2.1}$ 61.7
	Soft (ran.)	2.3 $\xrightarrow{+47.1}$ 49.4 $\xrightarrow{+1.9}$ 51.3	8.0 $\xrightarrow{+73.0}$ 81.0 $\xrightarrow{+0.7}$ 81.7	4.5 $\xrightarrow{+55.9}$ 60.4 $\xrightarrow{+1.5}$ 61.9
Rob	LPAQA	1.2†	9.1†	4.2†
	AutoPrompt	40.0	68.3	49.9
	Soft (min.)	1.2 $\xrightarrow{+39.4}$ **40.6** $\xrightarrow{-7.3}$ 33.2	9.1 $\xrightarrow{+66.3}$ **75.4** $\xrightarrow{-22.3}$ 53.0	4.2 $\xrightarrow{+48.8}$ **53.0** $\xrightarrow{-12.1}$ 40.8
BAb	LPAQA	0.8†	5.7†	2.9†
	Soft (min.)	0.8 $\xrightarrow{+39.1}$ **39.9**	5.7 $\xrightarrow{+69.7}$ **75.4**	2.9 $\xrightarrow{+49.2}$ **52.1**
BAl	LPAQA	3.5†	5.6†	4.8†
	Soft (min.)	3.5 $\xrightarrow{+22.3}$ **25.8**	5.6 $\xrightarrow{+62.4}$ **68.0**	4.8 $\xrightarrow{+36.2}$ **41.0**

LM	Method	Precision@1	Precision@10	MRR
		init → soft → deep	init → soft → deep	init → soft → deep
BEb	LAMA	26.4	54.3	35.8
	LPAQA	31.2	57.3	39.9
	Soft (sin.)	26.4 $\xrightarrow{+22.2^?}$ 48.6 $\xrightarrow{+1.0}$ 49.6	54.3 $\xrightarrow{+23.3^?}$ 77.6 $\xrightarrow{+0.3}$ 77.9	35.8 $\xrightarrow{+22.9^?}$ 58.7 $\xrightarrow{+0.6}$ 59.3
	Soft (min.)	31.2 $\xrightarrow{+19.0^?}$ 50.2 $\xrightarrow{+0.3}$ **50.5**$^?$	57.3 $\xrightarrow{+21.9^?}$ 79.2 $\xrightarrow{+0.5}$ **79.7**$^?$	39.9 $\xrightarrow{+20.2^?}$ 60.1 $\xrightarrow{+0.4}$ **60.5**$^?$
	Soft (par.)	31.2 $\xrightarrow{+18.5^?}$ 49.7 $\xrightarrow{+0.0}$ 49.7	57.3 $\xrightarrow{+21.3^?}$ 78.6 $\xrightarrow{+0.6}$ 79.2	39.9 $\xrightarrow{+19.6^?}$ 59.5 $\xrightarrow{+0.3}$ 59.8
	Soft (ran.)	0.8 $\xrightarrow{+46.3}$ 47.1 $\xrightarrow{+3.5}$ 50.6	4.0 $\xrightarrow{+70.4}$ 74.4 $\xrightarrow{+4.9}$ 79.3	2.2 $\xrightarrow{+54.3}$ 56.5 $\xrightarrow{+3.9}$ 60.4
BEl	LAMA	24.0†	53.7†	34.1†
	LPAQA	37.8†	64.4†	44.0†
	Soft (sin.)	24.0 $\xrightarrow{+26.2}$ 50.2 $\xrightarrow{+1.2}$ 51.4	53.7 $\xrightarrow{+24.9}$ 78.6 $\xrightarrow{+0.9}$ 79.5	34.1 $\xrightarrow{+25.9}$ 60.0 $\xrightarrow{+1.2}$ 61.2
	Soft (min.)	37.8 $\xrightarrow{+13.4}$ 51.2 $\xrightarrow{+1.3}$ **52.5**	64.4 $\xrightarrow{+15.1}$ 79.5 $\xrightarrow{+1.6}$ **81.1**	44.0 $\xrightarrow{+17.0}$ 61.0 $\xrightarrow{+1.4}$ **62.4**
	Soft (par.)	37.8 $\xrightarrow{+12.5}$ 50.3 $\xrightarrow{+1.4}$ 51.7	64.4 $\xrightarrow{+14.3}$ 78.7 $\xrightarrow{+2.1}$ 80.8	44.0 $\xrightarrow{+16.1}$ 60.1 $\xrightarrow{+1.6}$ 61.7
	Soft (ran.)	1.4 $\xrightarrow{+46.1}$ 47.5 $\xrightarrow{+4.4}$ 51.9	5.4 $\xrightarrow{+68.9}$ 74.3 $\xrightarrow{+6.3}$ 80.6	5.7 $\xrightarrow{+51.2}$ 56.9 $\xrightarrow{+5.0}$ 61.9

Model	P@1	P@10	MRR
LAMA (BEb)	0.1†	2.6†	1.5†
LAMA (BEl)	0.1†	5.0†	1.9†
Soft (min.,BEb)	11.3(**+11.2**)	36.4(**+33.8**)	19.3(**+17.8**)
Soft (ran.,BEb)	**11.8**(**+11.8**)	**34.8**(**+31.9**)	**19.8**(**+19.6**)
Soft (min.,BEl)	**12.8**(**+12.7**)	**37.0**(**+32.0**)	**20.9**(**+19.0**)
Soft (ran.,BEl)	**14.5**(**+14.5**)	**38.6**(**+34.2**)	**22.1**(**+21.9**)

Model	P@1	P@10	MRR
LAMA	9.7†	27.0†	15.6†
LPAQA	10.6†	23.7†	15.3†
Soft (sin.)	**11.2** (**+1.5**)	**33.5** (**+ 6.5**)	**18.9** (**+3.3**)
Soft (min.)	**12.9** (**+2.3**)	**34.7** (**+11.0**)	**20.3** (**+5.0**)
Soft (par.)	**11.5** (**+0.9**)	**31.4** (**+ 7.7**)	**18.3** (**+3.0**)

Model	P@1	P@10	MRR
LPAQA (BEb)	18.9	40.4	26.6
Soft (BEb)	**23.0** (**+4.1**)	**45.2** (**+4.8**)	**30.5** (**+3.9**)
LPAQA (BEl)	23.8	47.7	32.2
Soft (BEl)	**27.0** (**+3.2**)	**51.7** (**+4.0**)	**35.4** (**+3.2**)

Model	P@1	P@10	MRR
baseline	39.4	67.4	49.1
adjust mixture weights	40.0	69.1	53.3
adjust token vectors	50.7	80.7	61.1
adjust both	**51.0**	**81.4**	**61.6**

Related Work

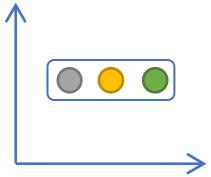
- Jiang, Zhengbao, et al. "How can we know what language models know?." *TACL* (2020).
- Shin, Taylor, et al. "AutoPrompt: Eliciting knowledge from language models with automatically generated prompts." *EMNLP* (2020).
- Haviv, Adi, Jonathan Berant, and Amir Globerson. "BERTese: Learning to speak to BERT." *EACL* (2021).
- Li, Xiang Lisa, and Percy Liang. "Prefix-Tuning: Optimizing Continuous Prompts for Generation." *ACL* (2021).
- Xiao Liu, Yanan Zheng, Zhengxiao Du, Ming Ding, Yujie Qian, Zhilin Yang, and Jie Tang. GPT understands, too. *arXiv* (2021).

Takeaways

- LMs know more facts than we thought.
  You just have to learn how to ask.

Takeaways

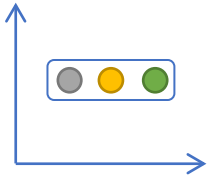
- LMs know more facts than we thought.
  You just have to learn how to ask.



- Prompts are made of vectors, not words!
  So you can tune them with backprop.

Takeaways

- LMs know more facts than we thought.
  You just have to learn how to ask.

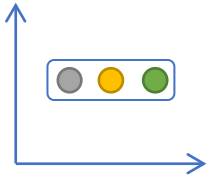


- Prompts are made of vectors, not words!
  So you can tune them with backprop.

- Random initialization works fine.
  No grad student required.

Takeaways

- LMs know more facts than we thought.
  You just have to learn how to ask.



- Prompts are made of vectors, not words!
  So you can tune them with backprop.

- Random initialization works fine.
  No grad student required.

- Prompt tuning is lightweight,
  and could also be applied to few-shot learning.

Thanks!