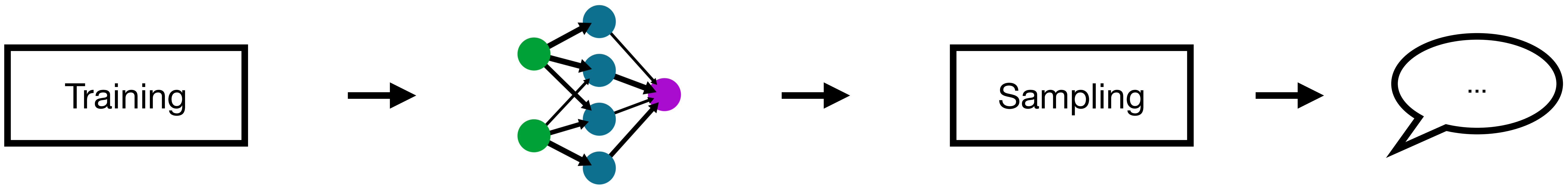
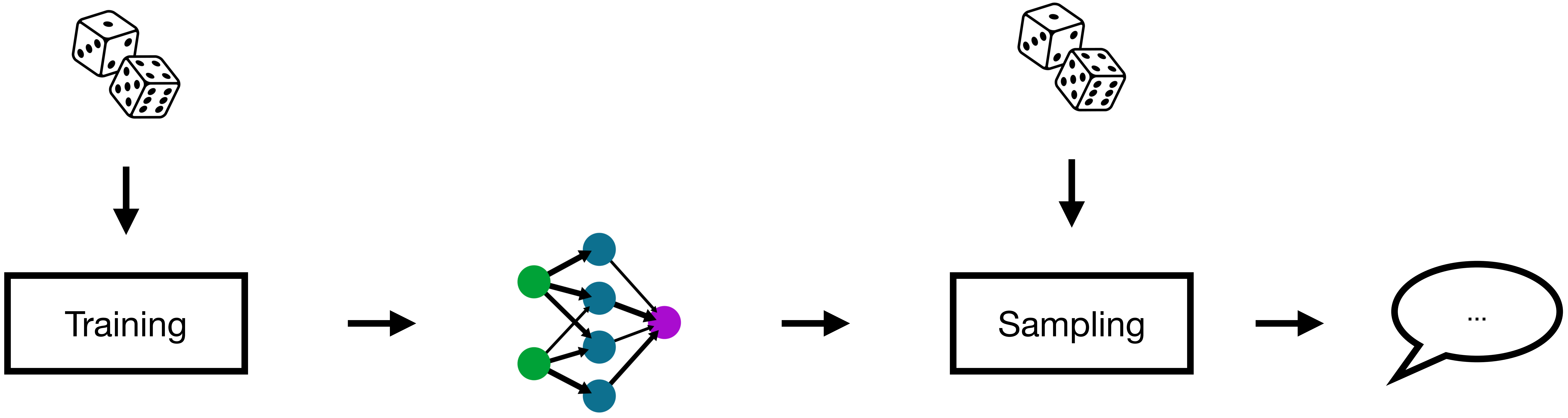
Distortion-free mechanisms for language model provenance

based on joint work with Sally Zhu, Ahmed Ahmed, John Thickstun, Tatsu Hashimoto, and Percy Liang

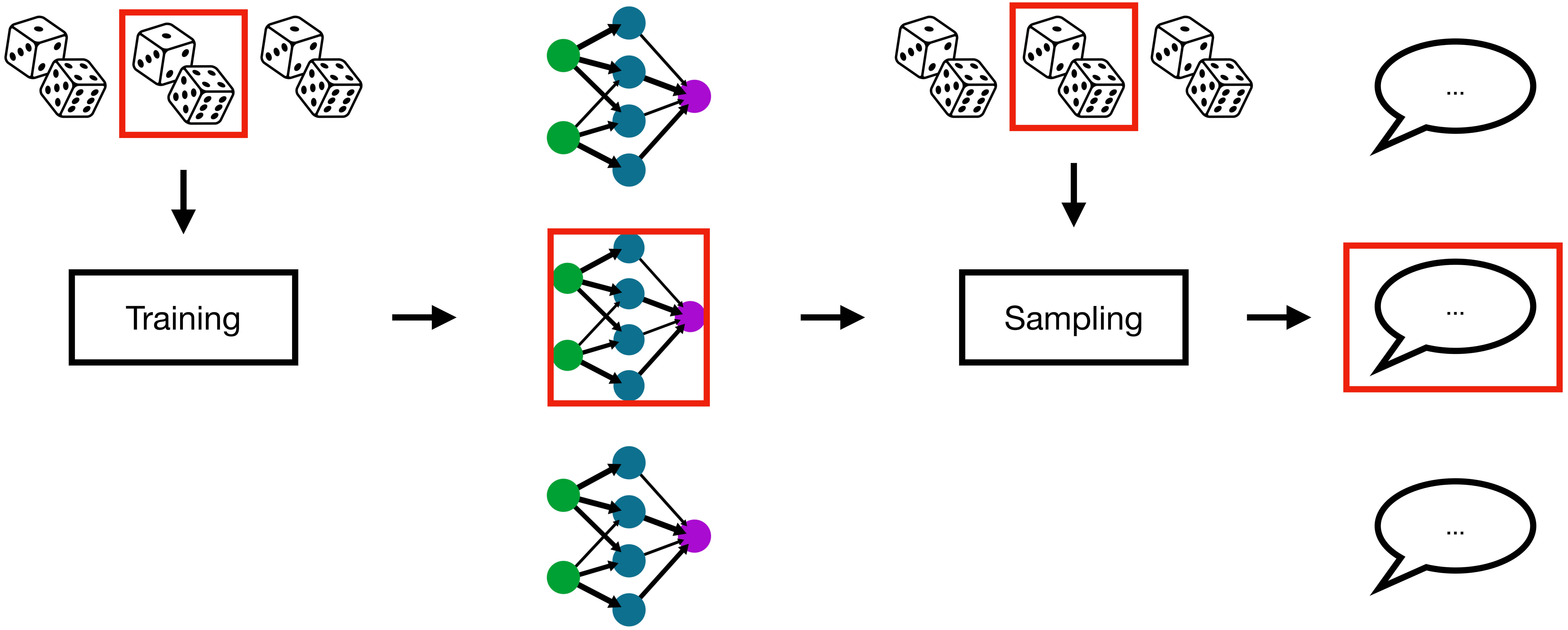
The lifecycle of a language model



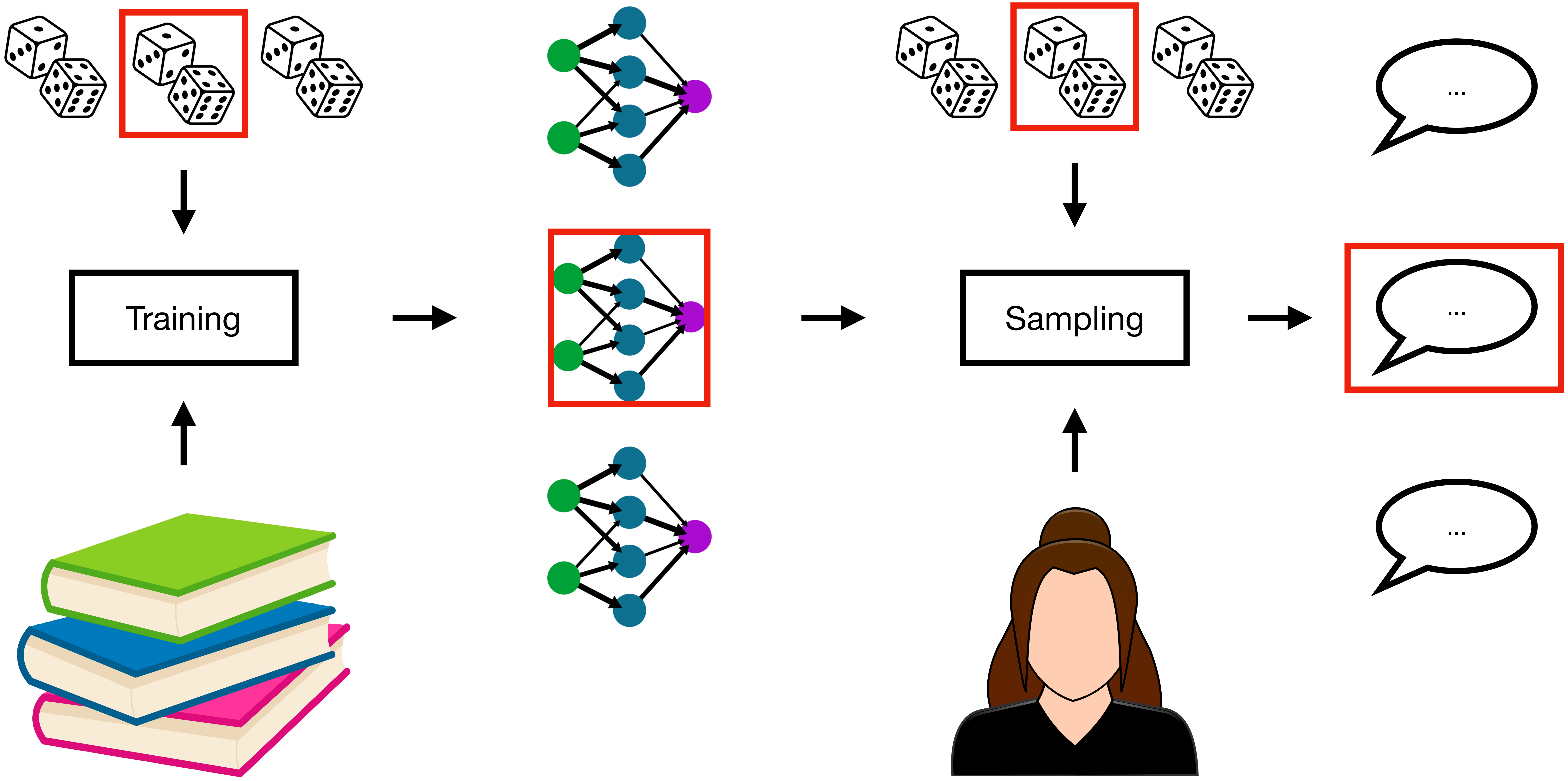
Provenance via independence testing



Provenance via independence testing



Provenance via independence testing



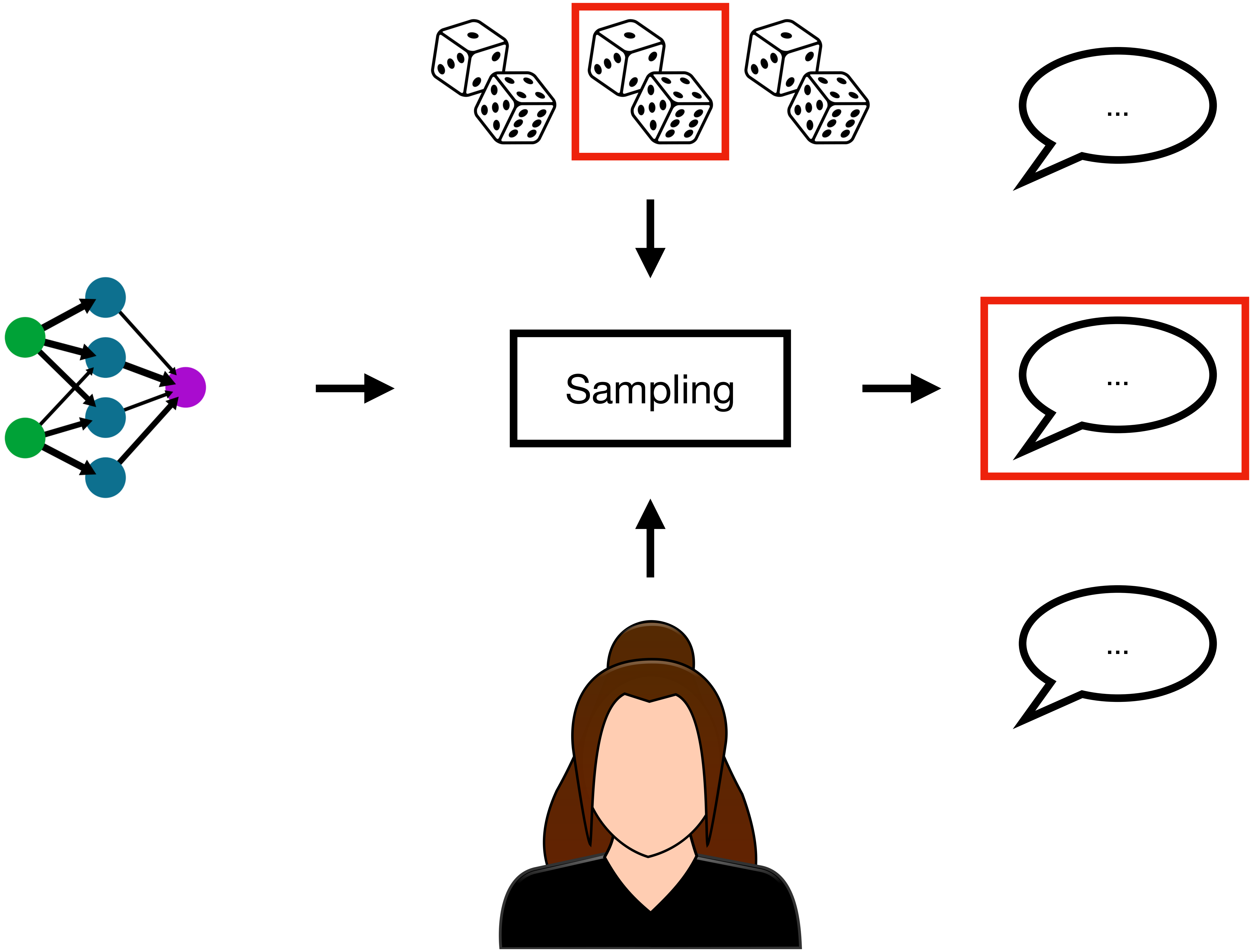
Part 1: Text

John Thickstun

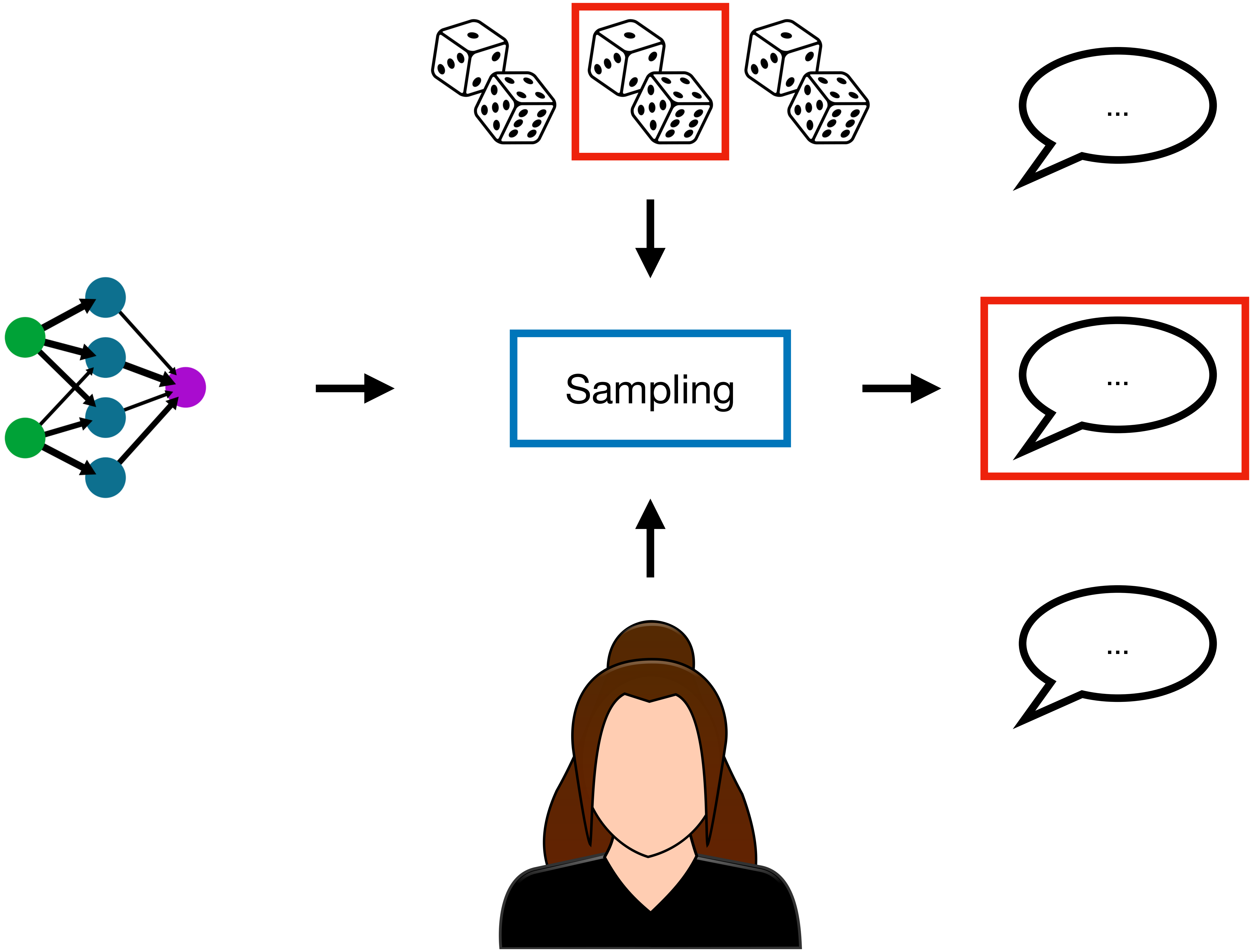
Tatsu
Hashimoto

Percy Liang

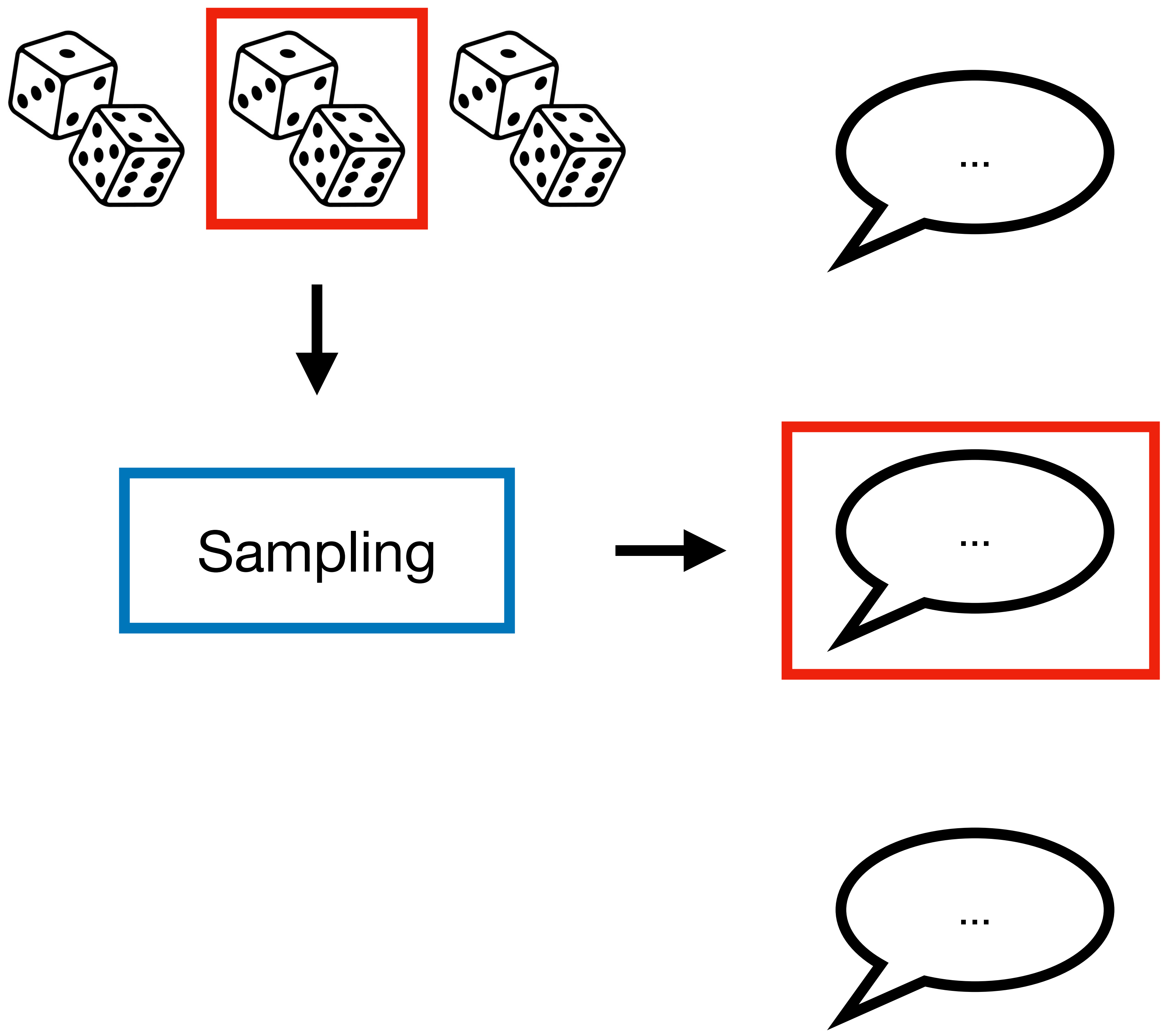
Provenance via independence testing



Provenance via independence testing



Provenance via independence testing

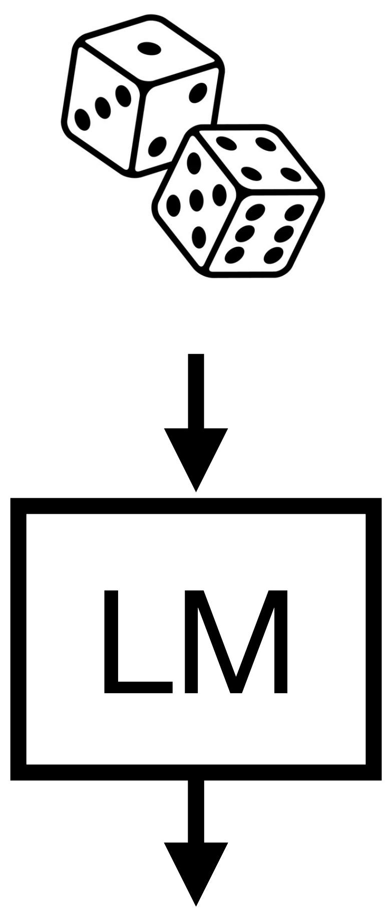


Sampling from a language model

LM

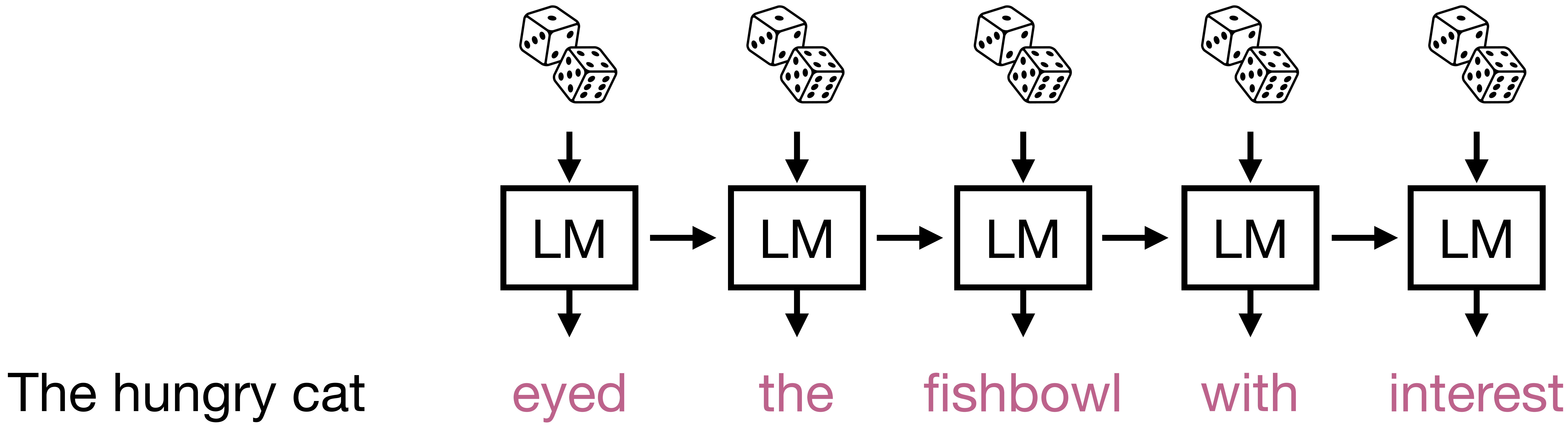
The hungry cat

Sampling from a language model

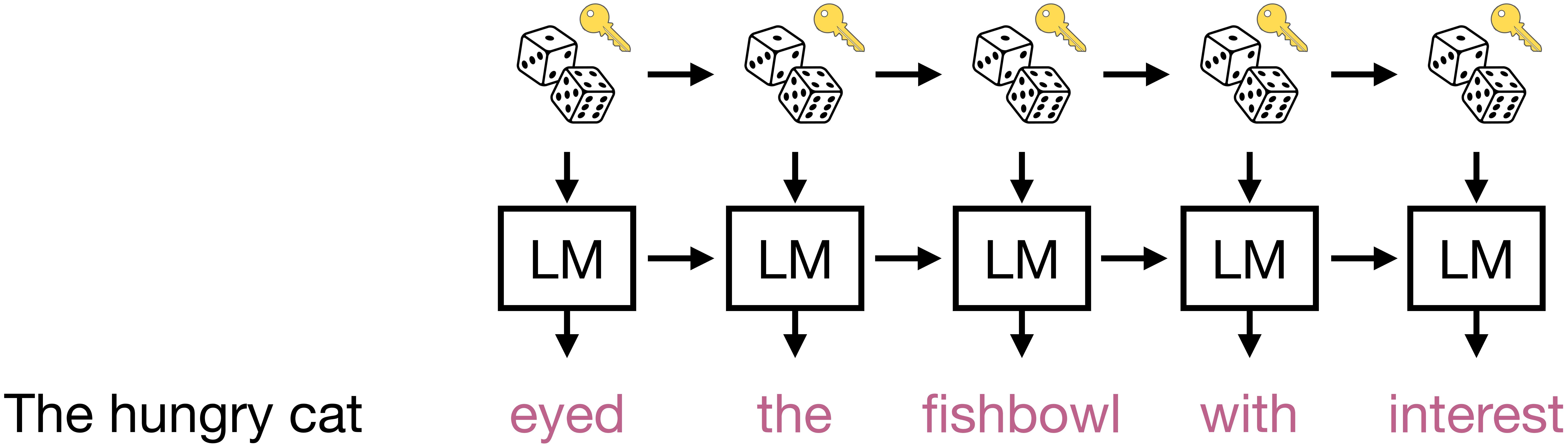


The hungry cat eyed

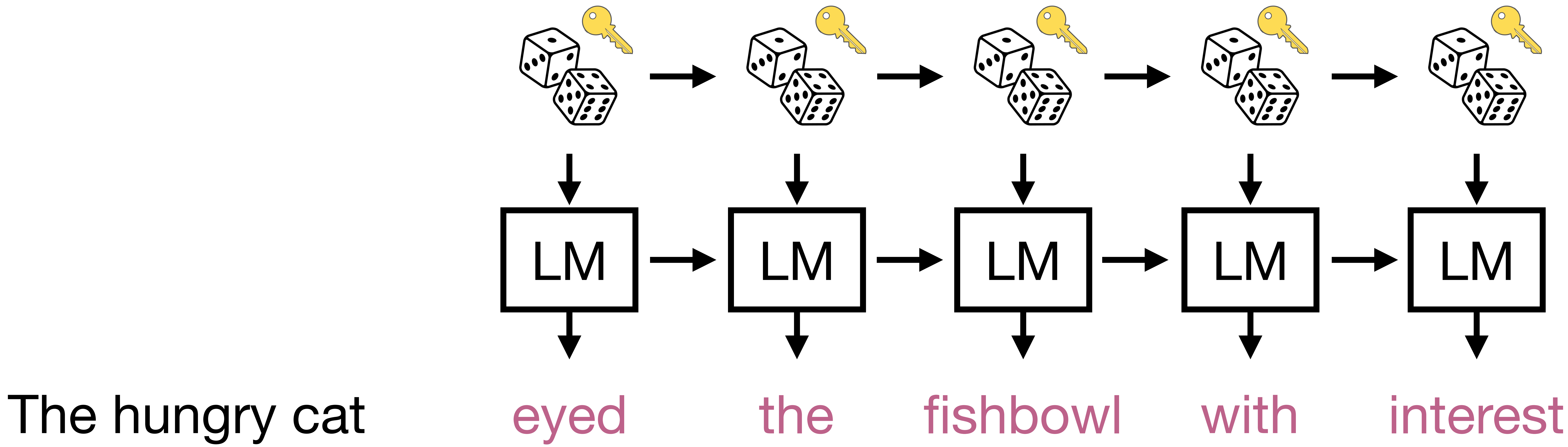
Sampling from a language model



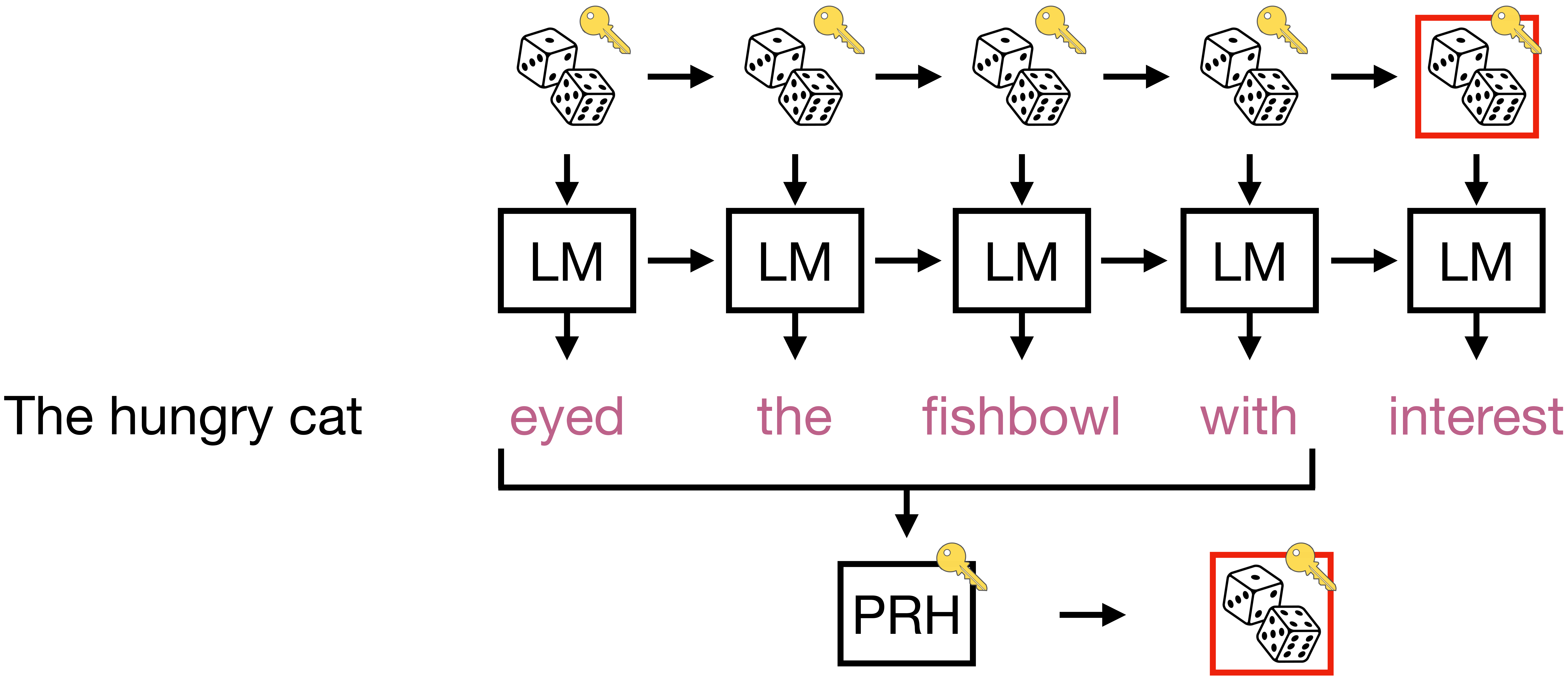
Watermarking a language model



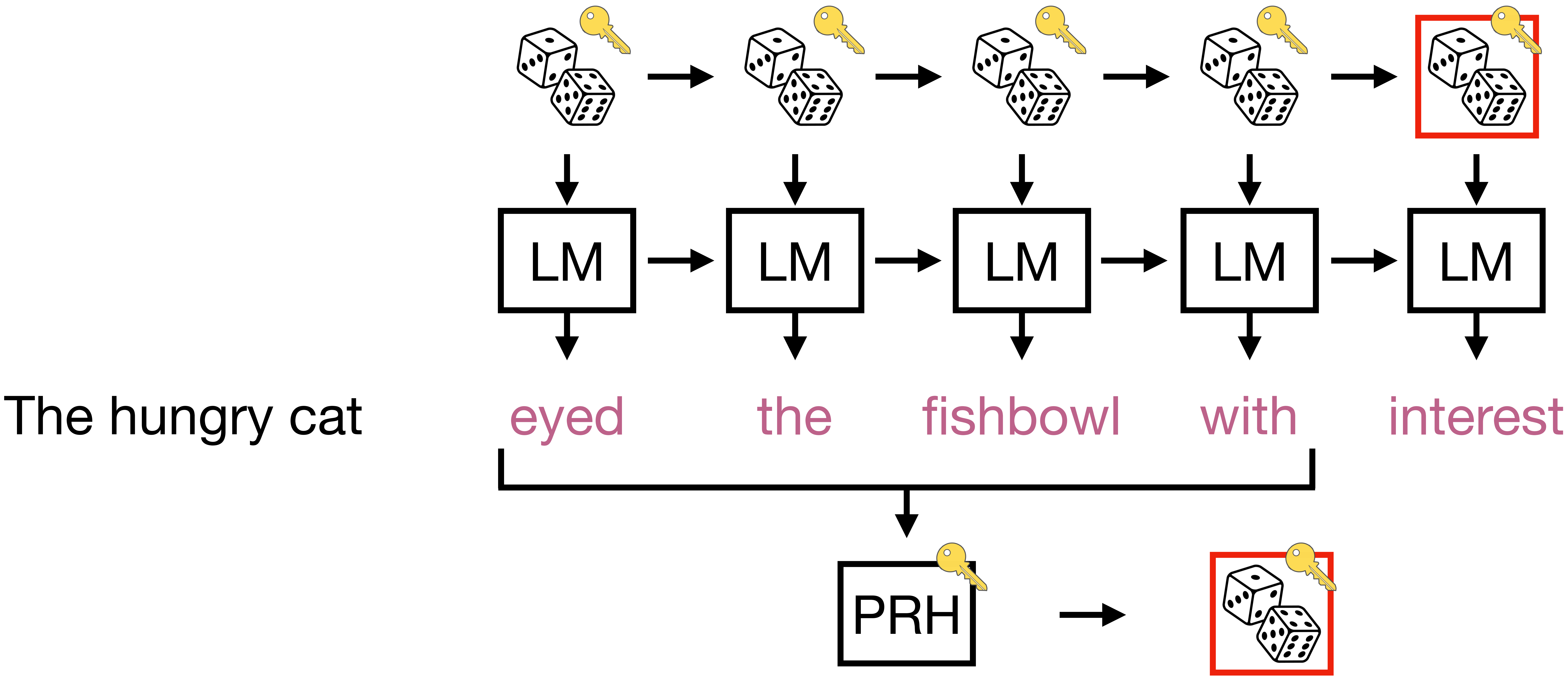
Hash-based watermarks [KGW+'23; AK'23; CGZ'23]



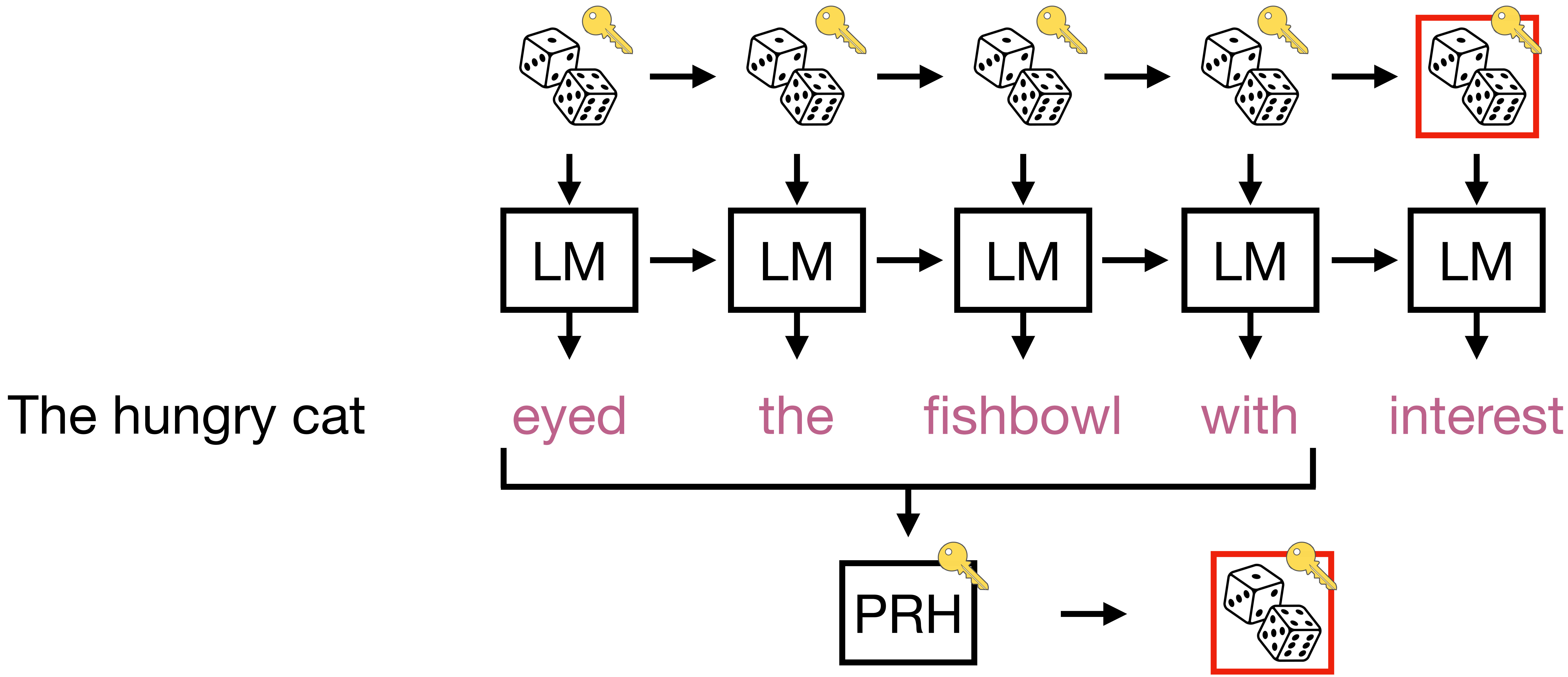
Hash-based watermarks [KGW+'23; AK'23; CGZ'23]



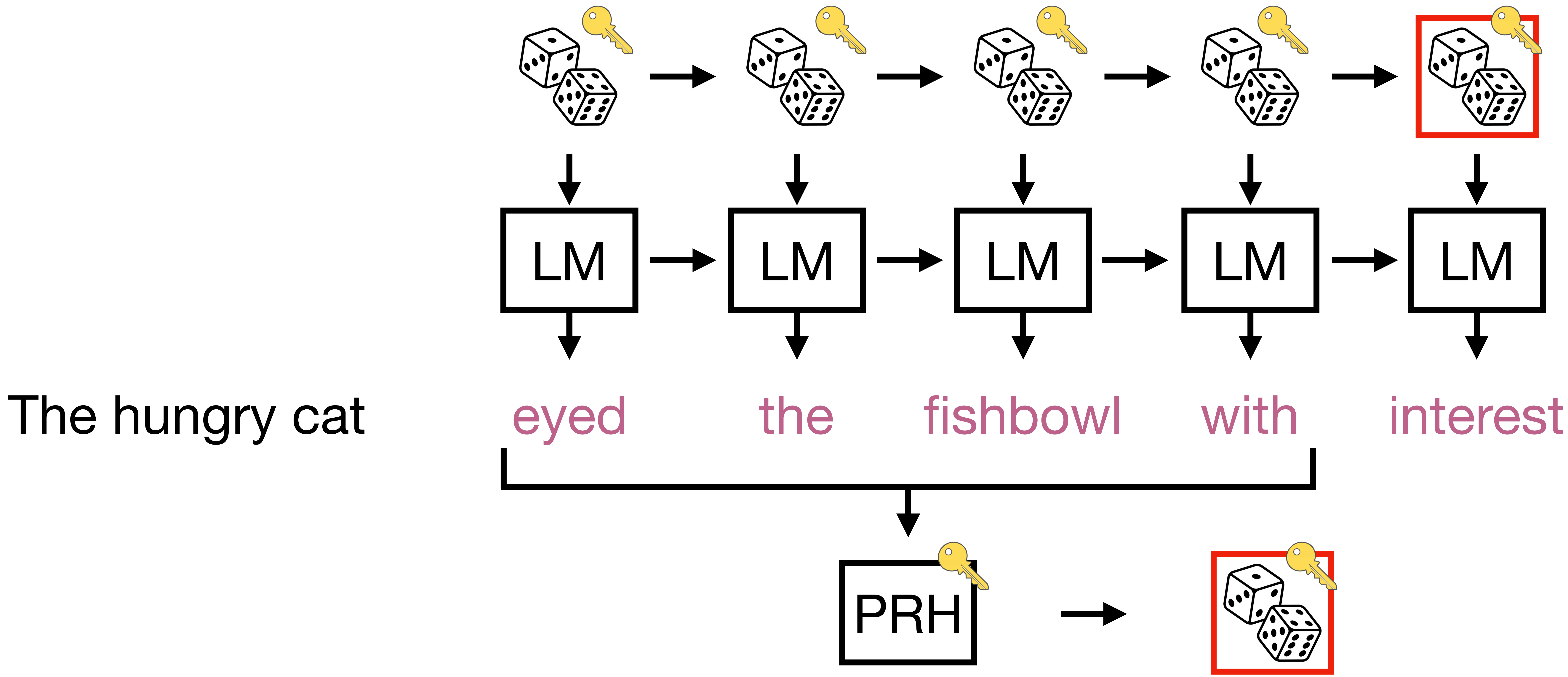
Hash-based watermarks [**KGW+'23**; AK'23; CGZ'23]



Hash-based watermarks [KGW+'23; **AK'23; CGZ'23**]



Hash-based watermarks [KGW+'23; AK'23; CGZ'23]



Prompt: **Give me a list of 20 movies.**

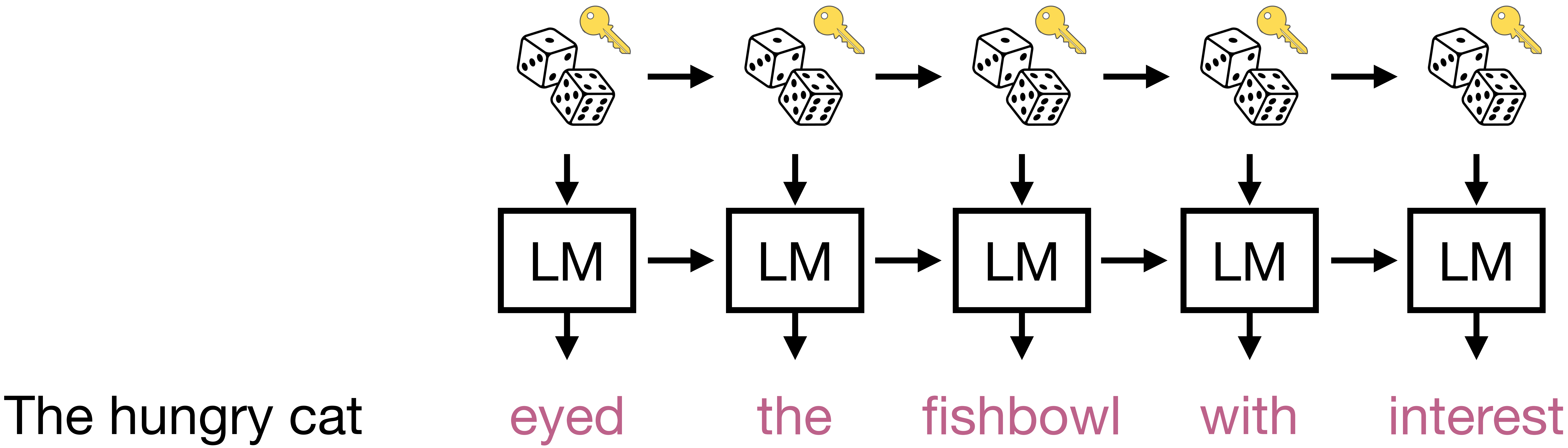
...

5. The Lord of the Rings: The Fellowship of the Ring
6. The Lord of the Rings: The Two Towers
7. The Lord of the Rings: The Return of the King
8. The Imitation Game
9. The Matrix
10. The Matrix Reloaded
11. The Matrix Revolutions
12. The Lord of the Rings: The Animated Version
13. The Lord of the Rings: The Angmar Wars
14. The Lord of the Rings: The Angmar Wars II
15. The Lord of the Rings: The Angmar Wars III

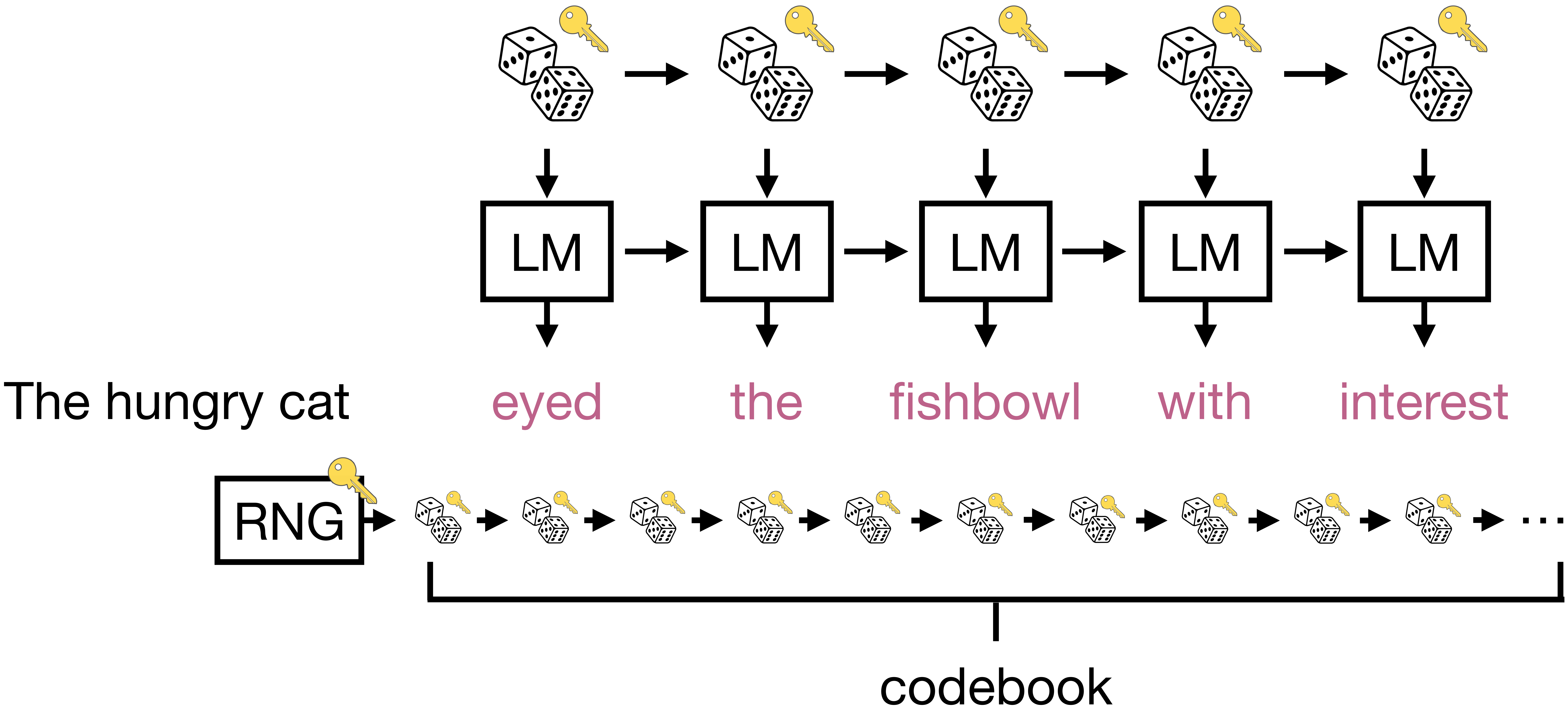
...

Model: Alpaca 7B
(hash-based watermark)

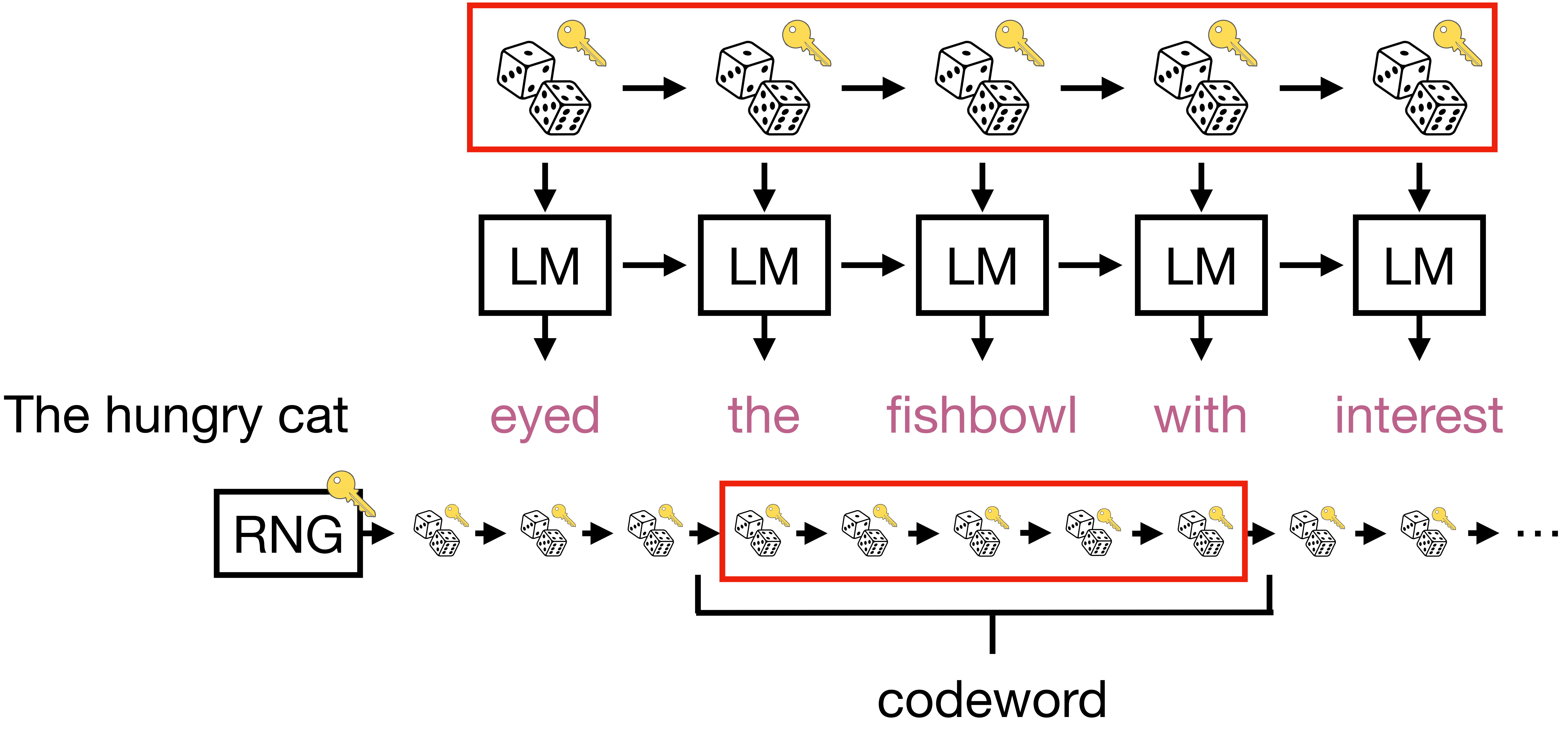
Distortion-free watermarks [**K**THL'23]



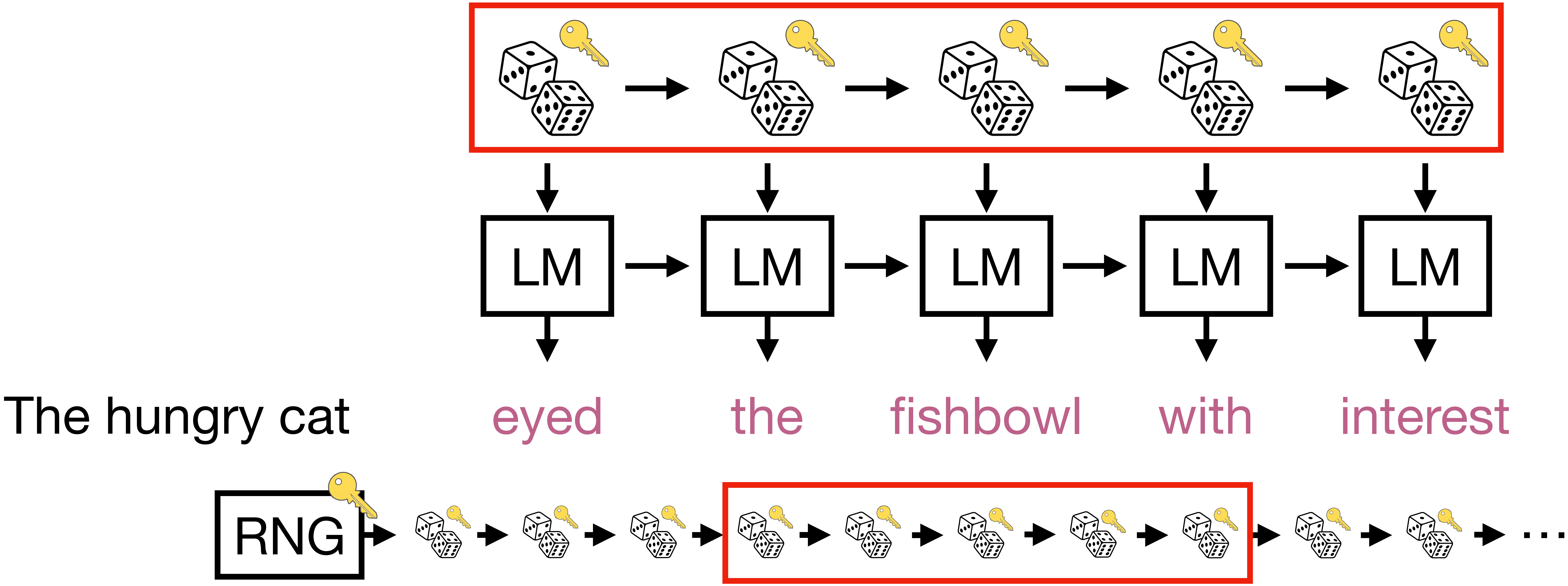
Distortion-free watermarks [KTHL'23]



Distortion-free watermarks [KTHL'23]

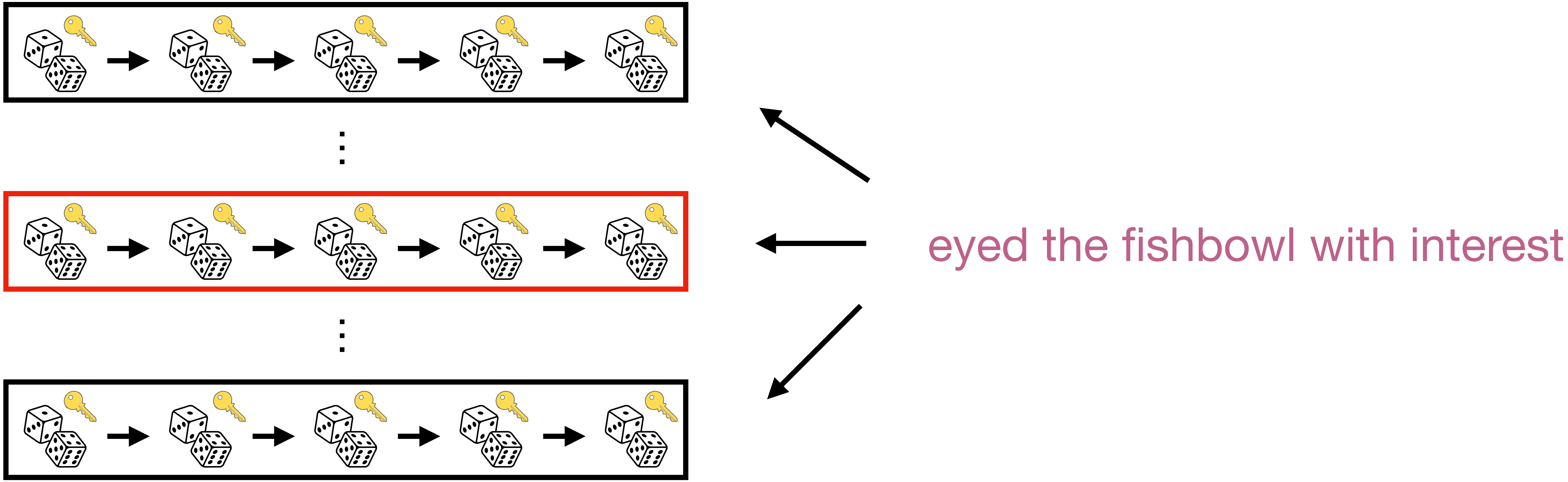


Distortion-free watermarks [KTHL'23]

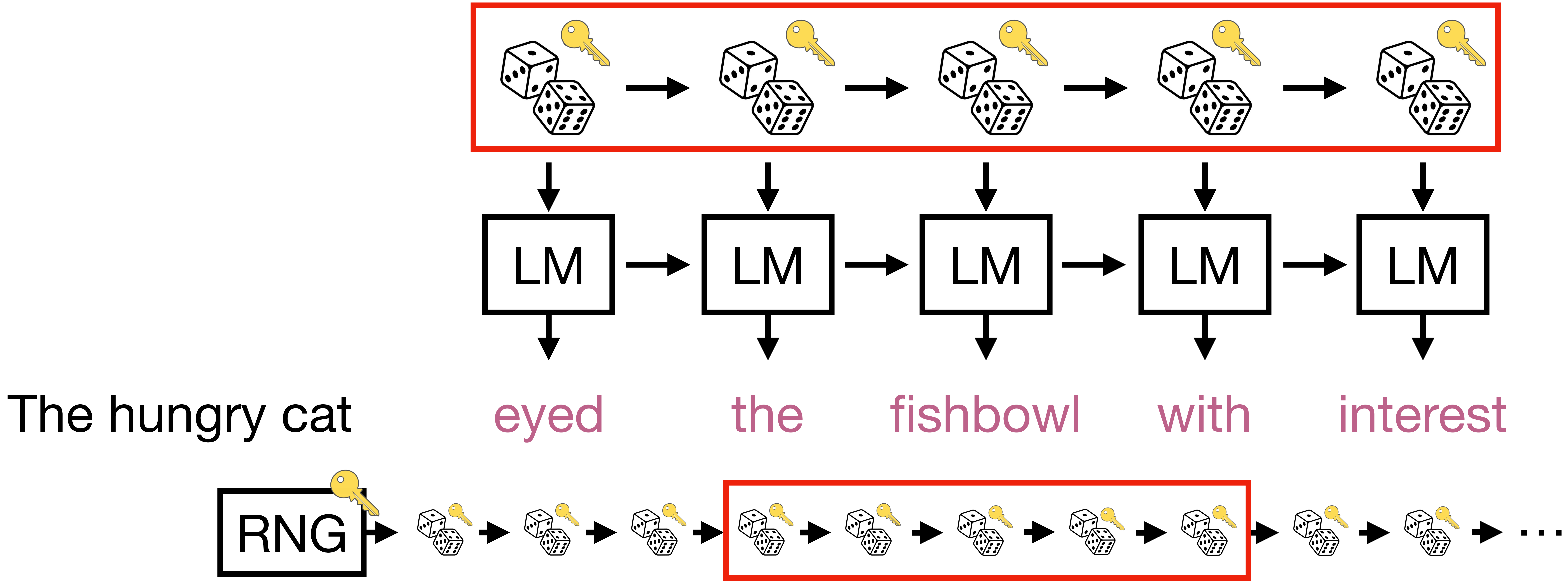


Generation is *distortion-free* until you re-use the dice.

Distortion-free watermarks [**K**THL'23]

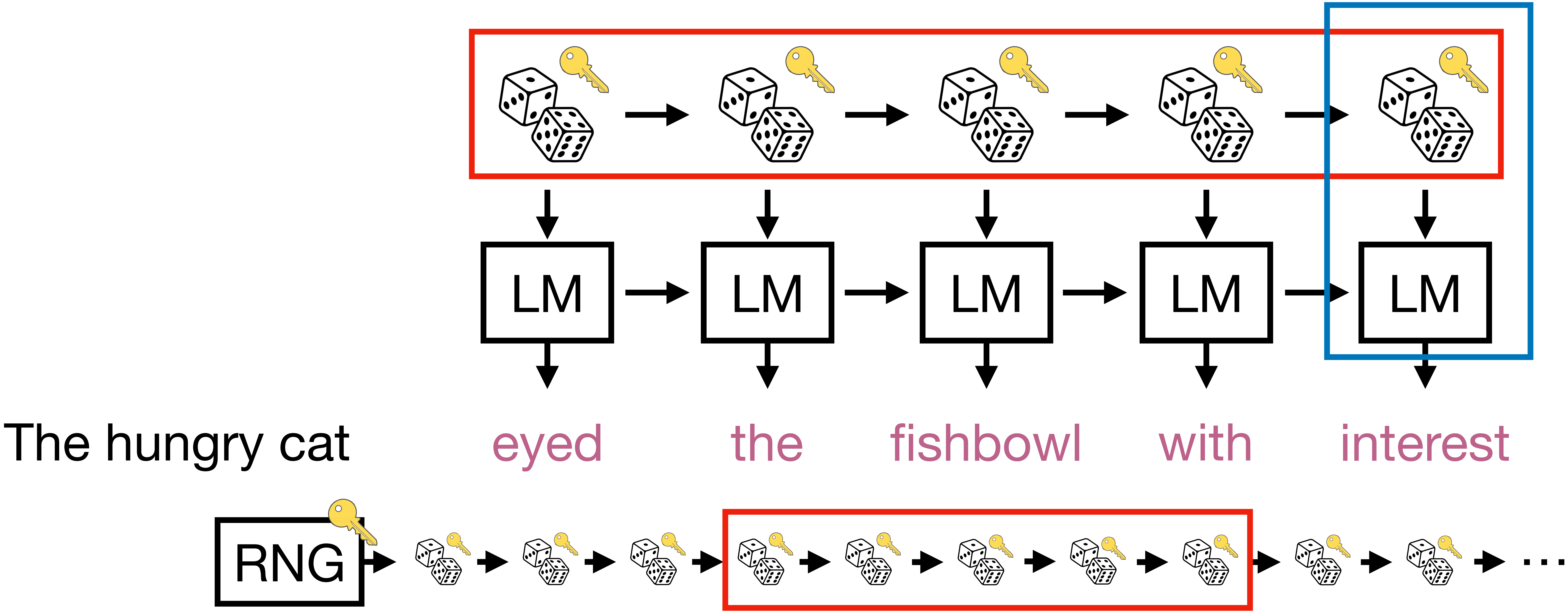


Distortion-free watermarks [KTHL'23]



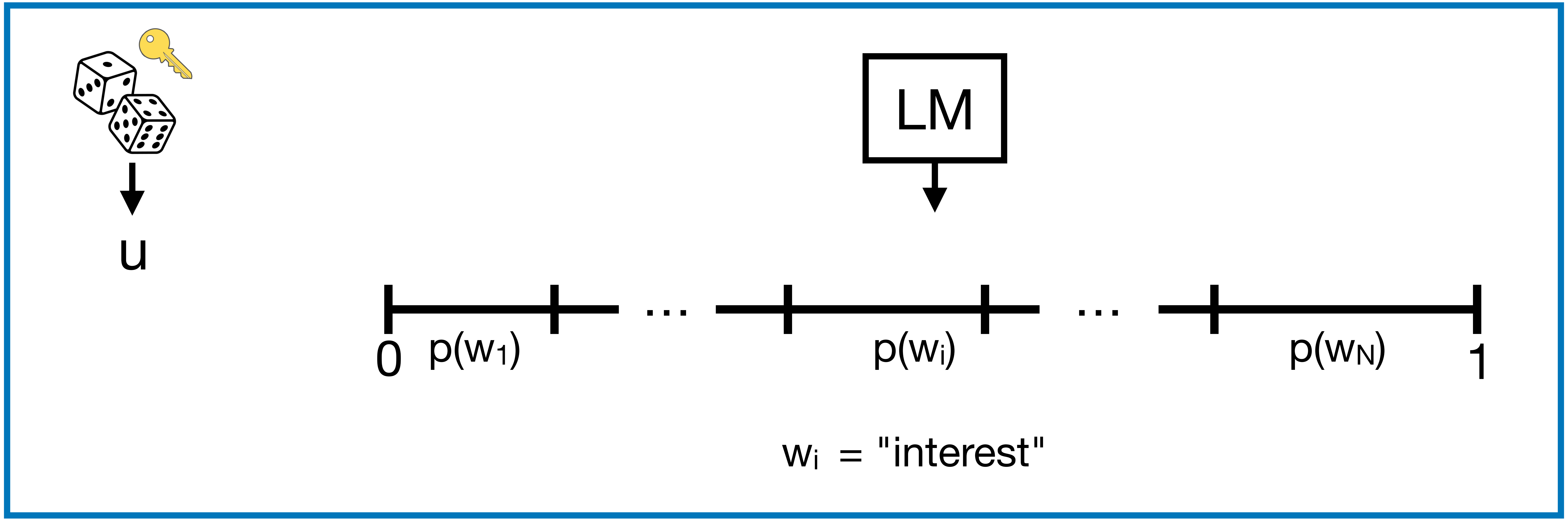
Generation is *distortion-free* until you re-use the dice.

Distortion-free watermarks [KTHL'23]

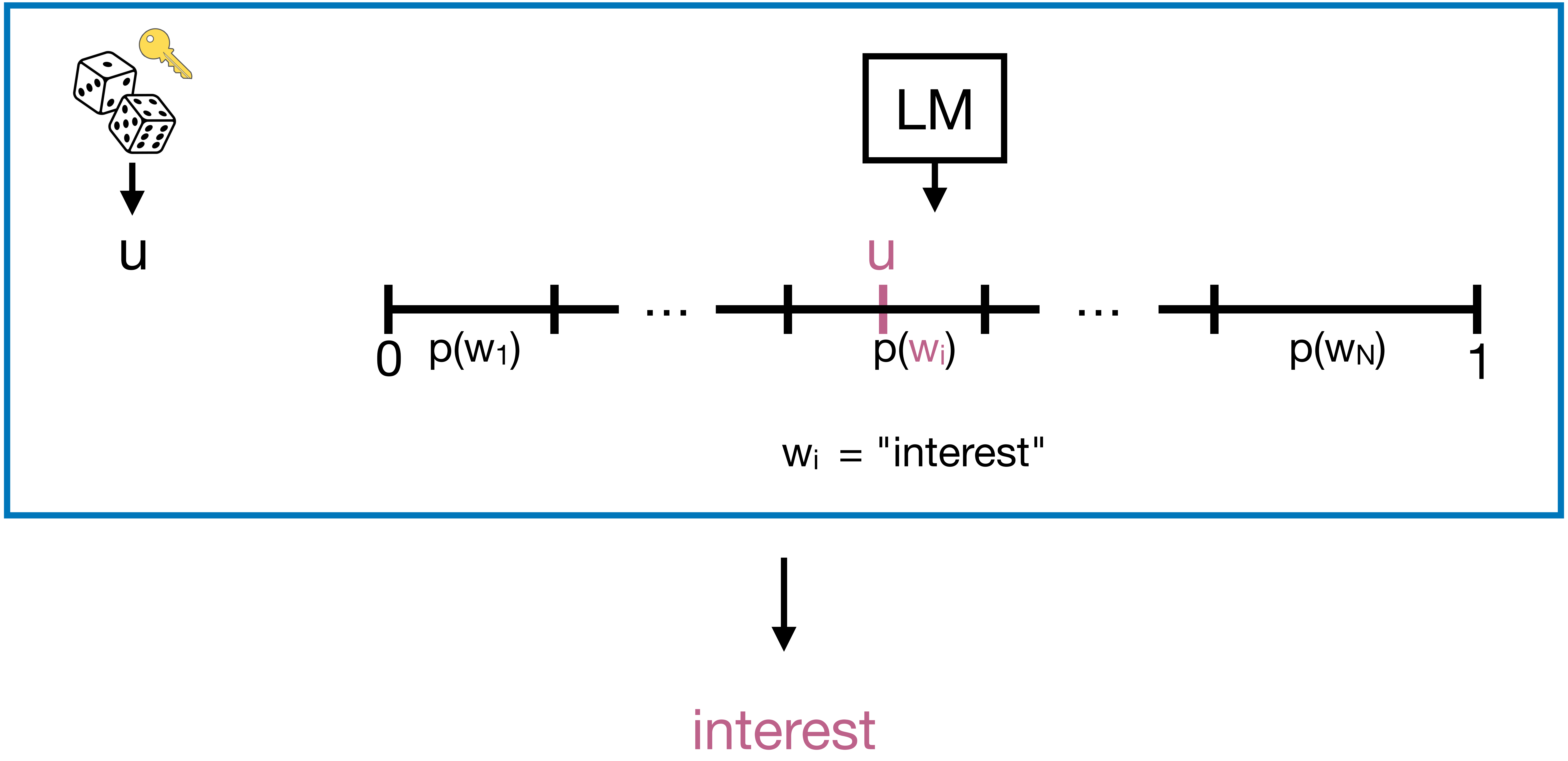


Generation is *distortion-free* until you re-use the dice.

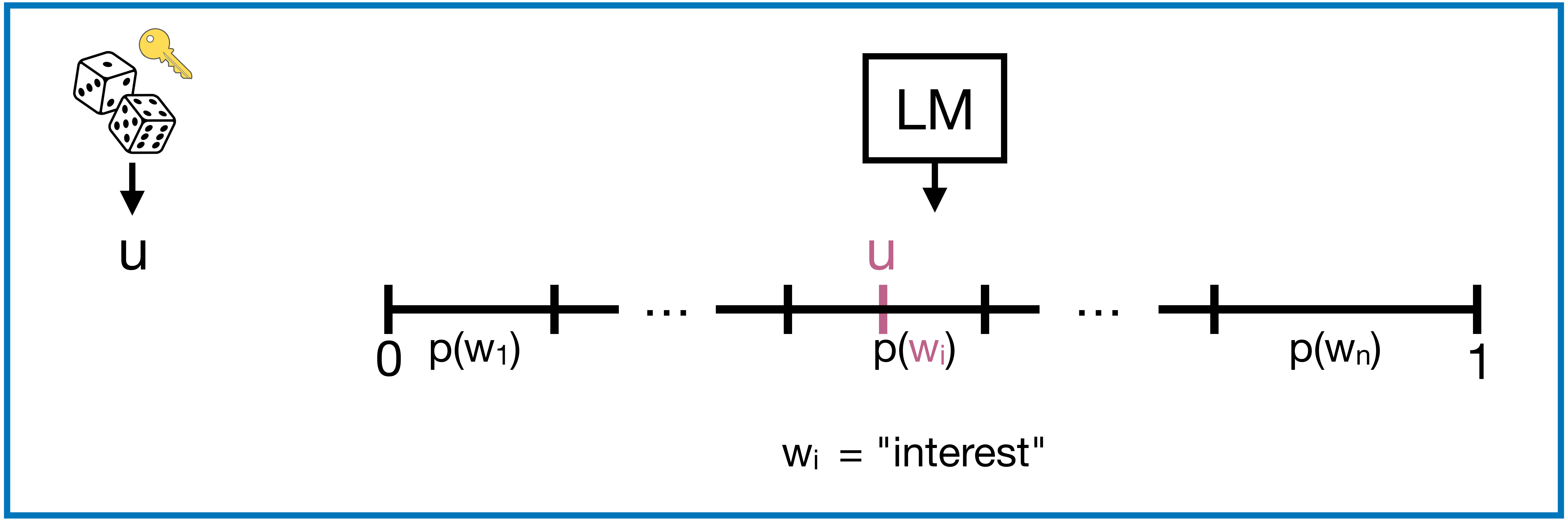
Generating watermarked text



Generating watermarked text

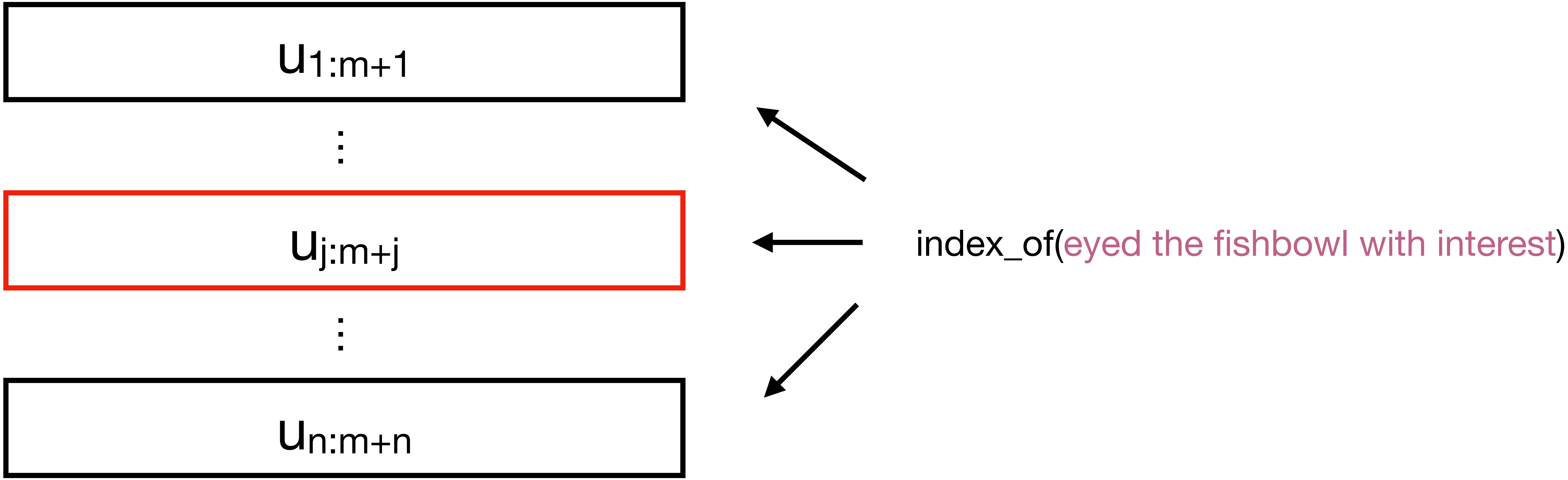


Detecting watermarked text

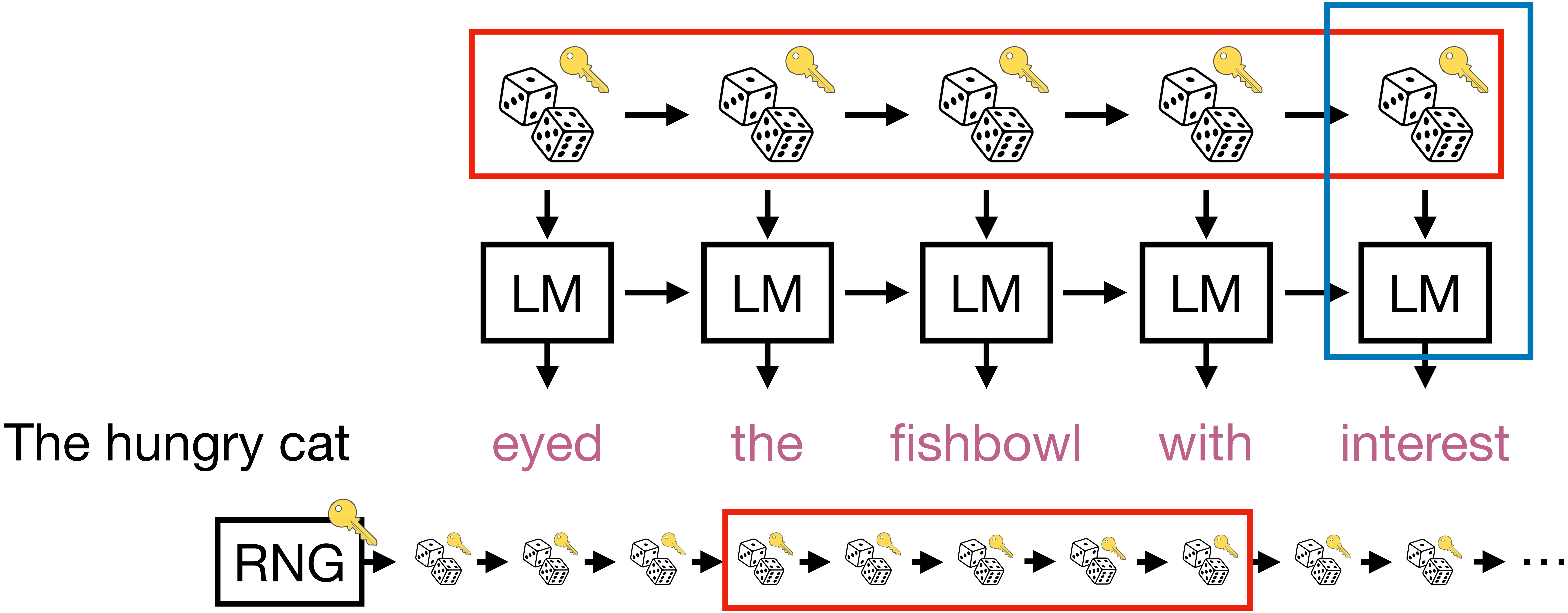


index_of("interest") correlates with u

Detecting watermarked text



Distortion-free watermarks [KTHL'23]



Generation is *distortion-free* until you re-use the dice.

Watermarking: what works and what doesn't

Our watermark is distortion-free *and* robust...

Watermarking: what works and what doesn't

Our watermark is distortion-free *and* robust...

...but detection is expensive.

Watermarking: what works and what doesn't

Our watermark is distortion-free *and* robust...

...but detection is expensive.

Can we have all three?

Watermarking: what works and what doesn't

Our watermark is distortion-free *and* robust...

...but detection is expensive.

Can we have all three? [CG'24; GM'24; GG'24]

Watermarking: what works and what doesn't

Our watermark is distortion-free *and* robust...

...but detection is expensive.

Can we have all three? [**CG'24**; GM'24; GG'24]

Watermarking: what works and what doesn't

Our watermark is distortion-free *and* robust...

...but detection is expensive.

Can we have all three? [CG'24; GM'24; GG'24]

Part 2: Weights

Sally Zhu

Ahmed Ahmed

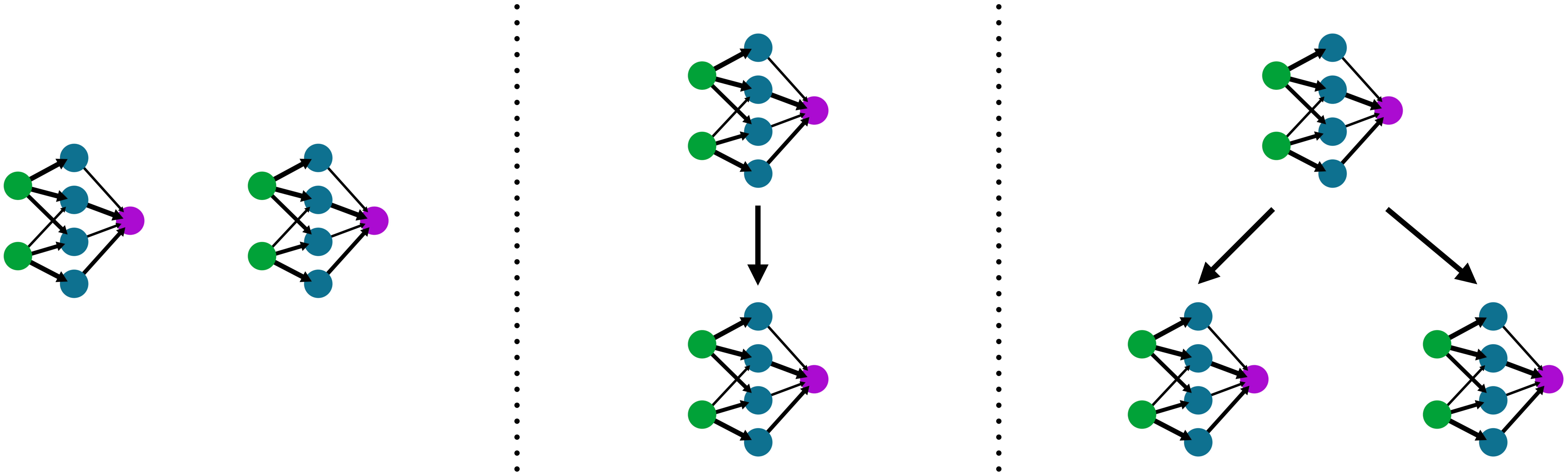
Percy Liang

Model independence testing

Can a third party infer the relationship between two models from their weights?

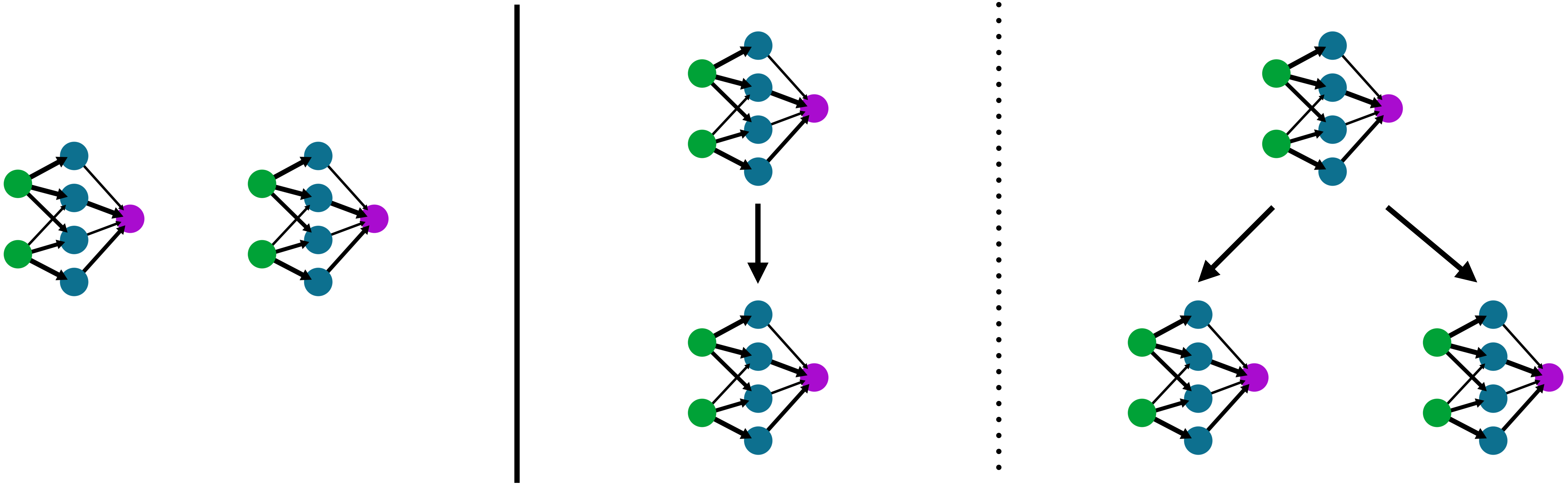
Model independence testing

Can a third party infer the relationship between two models from their weights?



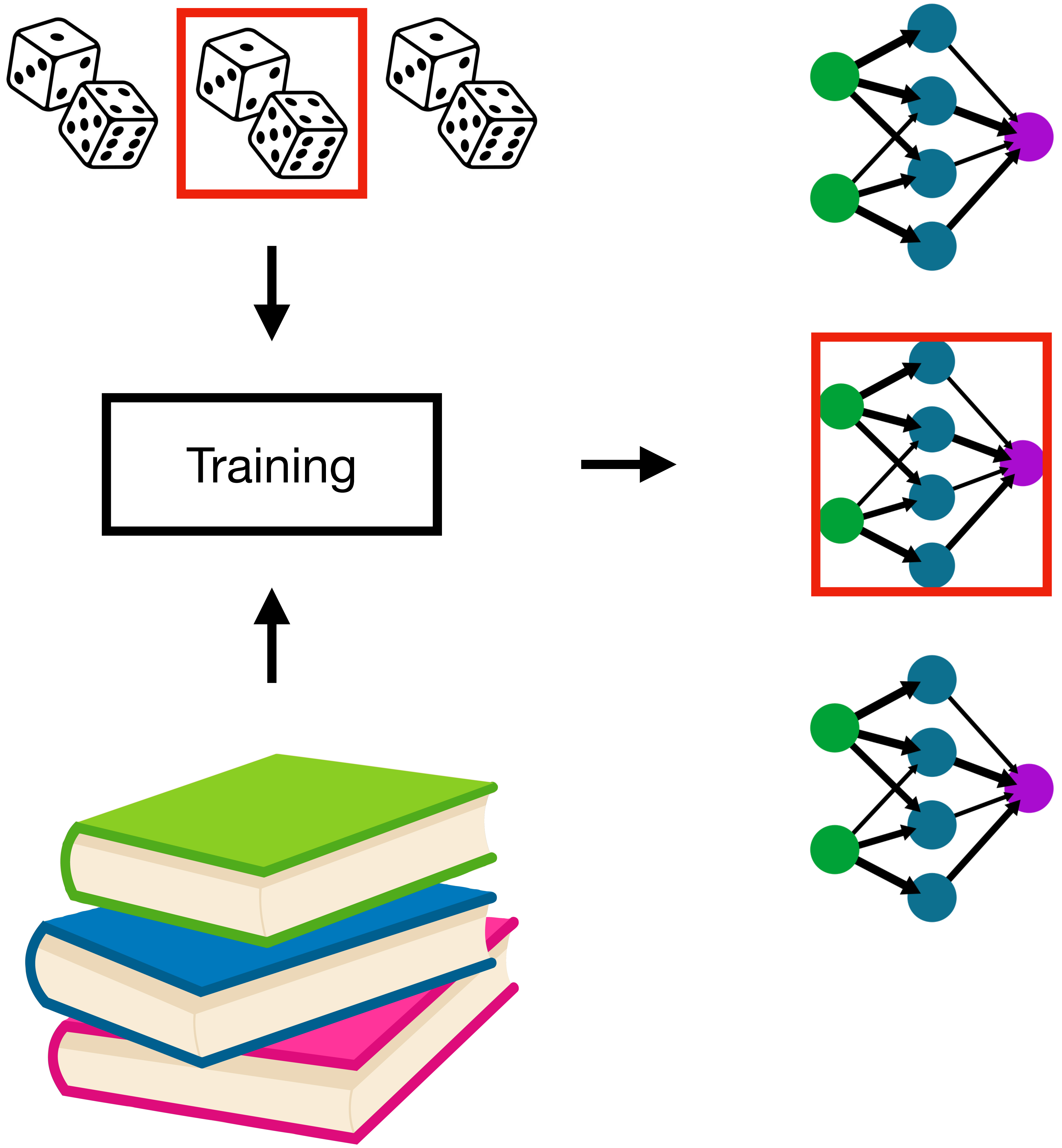
Model independence testing

Can a third party infer the relationship between two models from their weights?



Independent **Dependent**

Provenance via independence testing



Assumptions on training

$A : \Theta \to \Theta$ is Π-equivariant if $\pi(A(\theta_0)) = A(\pi(\theta_0))$ for any $\theta_0 \in \Theta$ and $\pi \in \Pi$.

Assumptions on training

$A : \Theta \to \Theta$ is Π-equivariant if $\pi(A(\theta_0)) = A(\pi(\theta_0))$ for any $\theta_0 \in \Theta$ and $\pi \in \Pi$.

$\mu \in \mathscr{P}(\Theta)$ is Π-invariant if we have $\mu(\theta_0) = \mu(\pi(\theta_0))$ for any $\theta_0 \in \Theta$ and $\pi \in \Pi$.

Assumptions on training

$A : \Theta \to \Theta$ is Π-equivariant if $\pi(A(\theta_0)) = A(\pi(\theta_0))$ for any $\theta_0 \in \Theta$ and $\pi \in \Pi$.

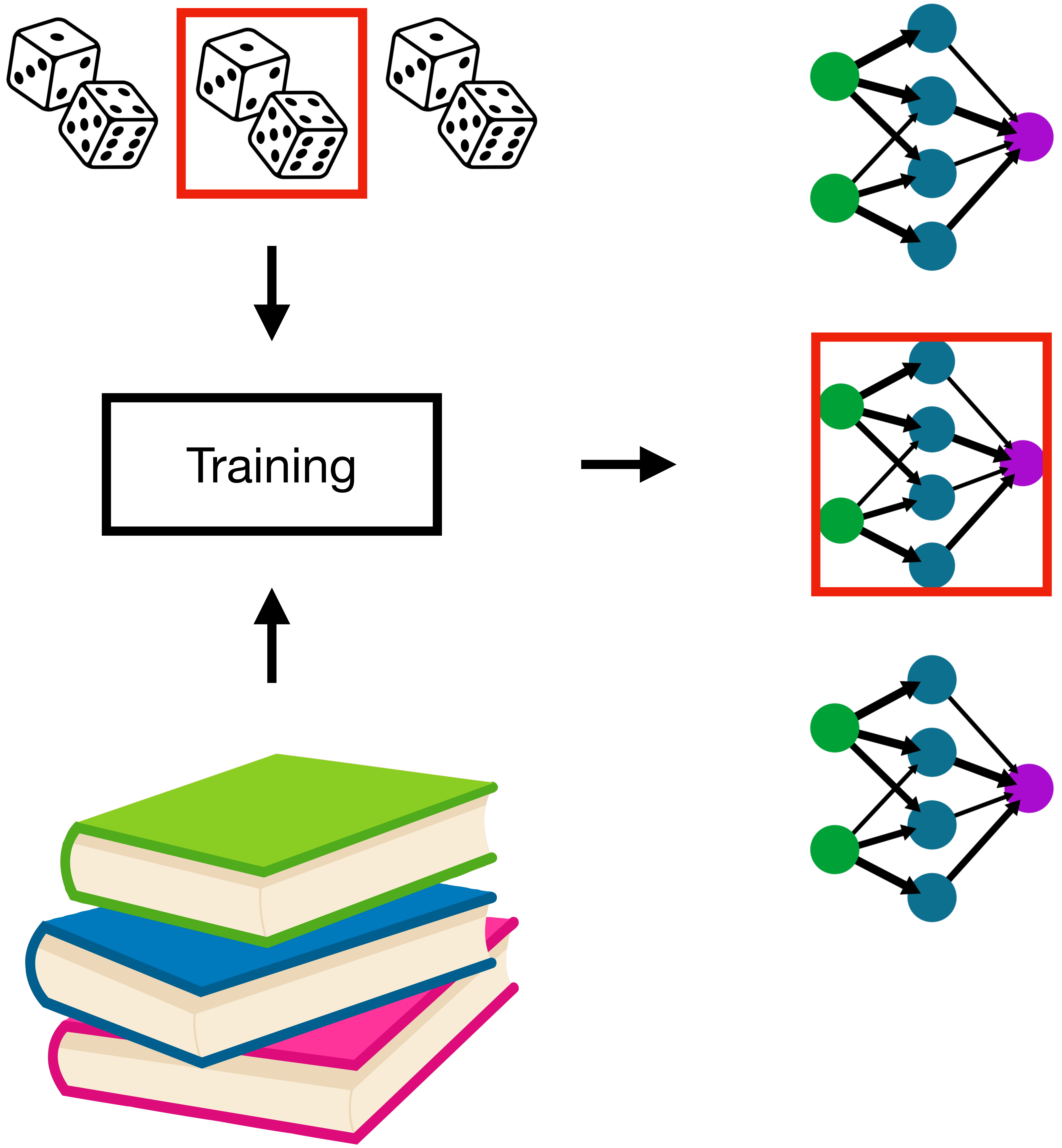
$\mu \in \mathscr{P}(\Theta)$ is Π-invariant if we have $\mu(\theta_0) = \mu(\pi(\theta_0))$ for any $\theta_0 \in \Theta$ and $\pi \in \Pi$.

Example (2-layer MLP):

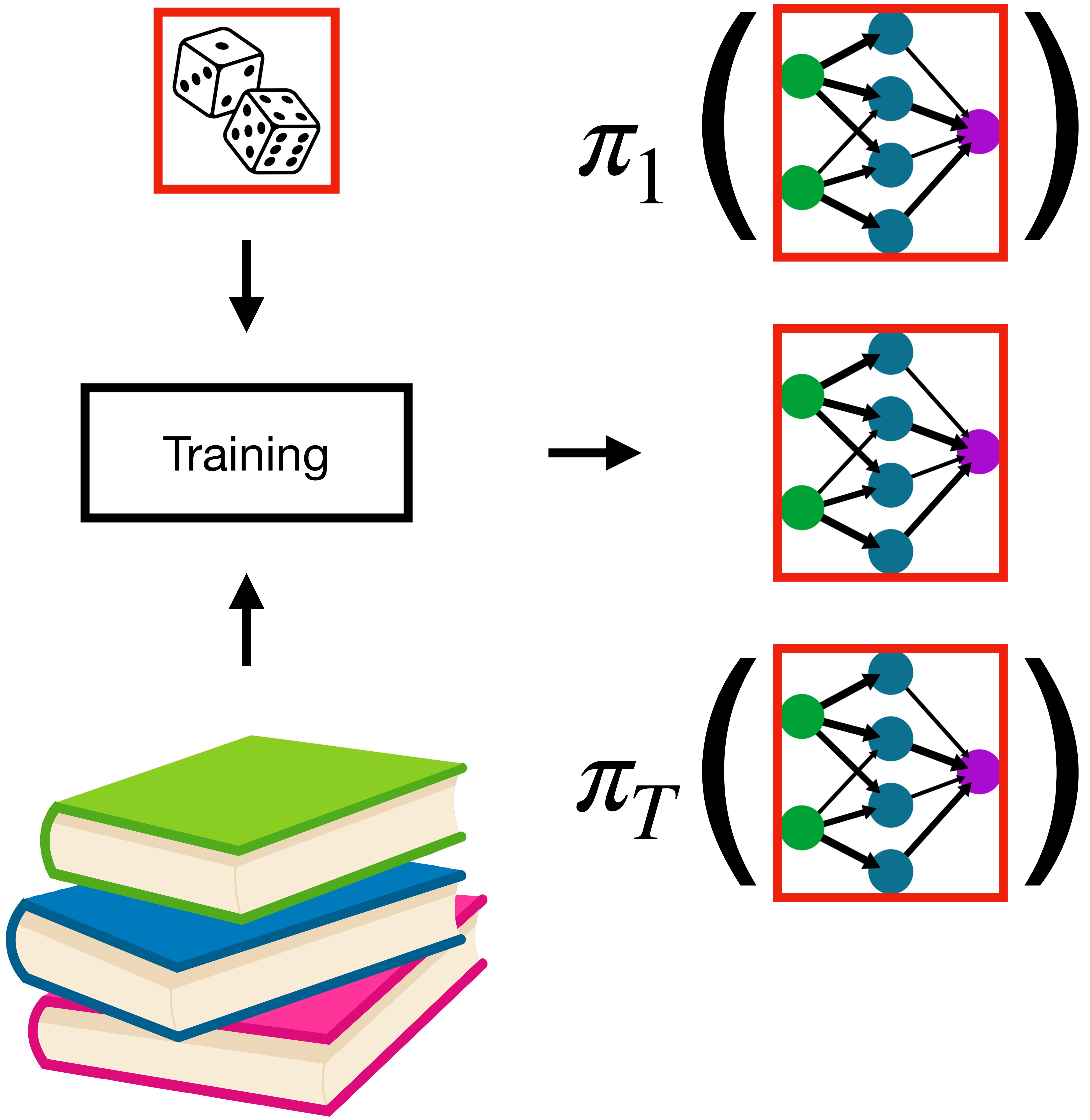
$$\theta = (W_1, W_2), \pi(\theta) = (W_2\pi^T, \pi W_1)$$

$$f(x; \theta) = W_2\sigma(W_1x) = W_2\pi^T\sigma(\pi W_1x) = f(x; \pi(\theta))$$

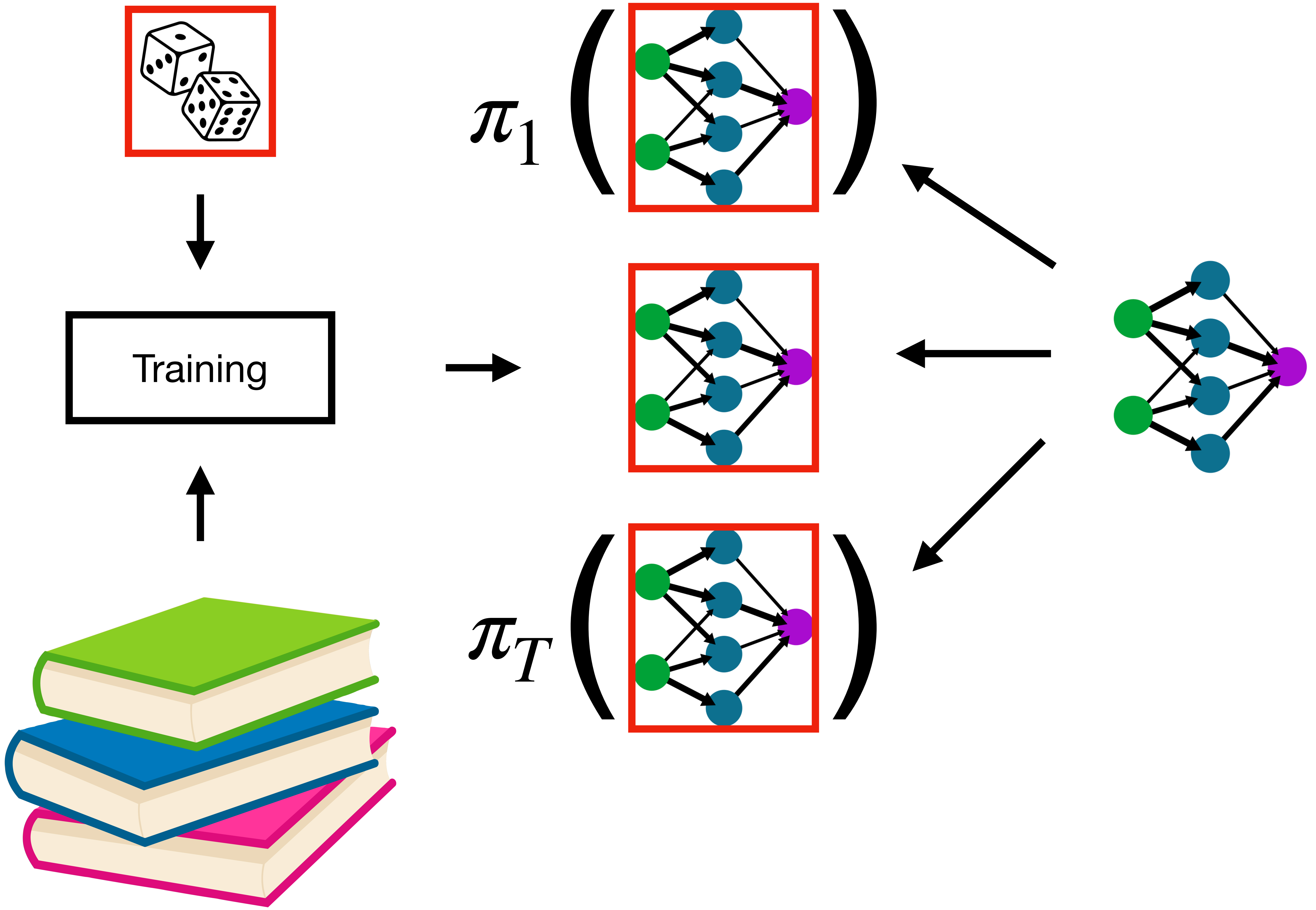
Provenance via independence testing



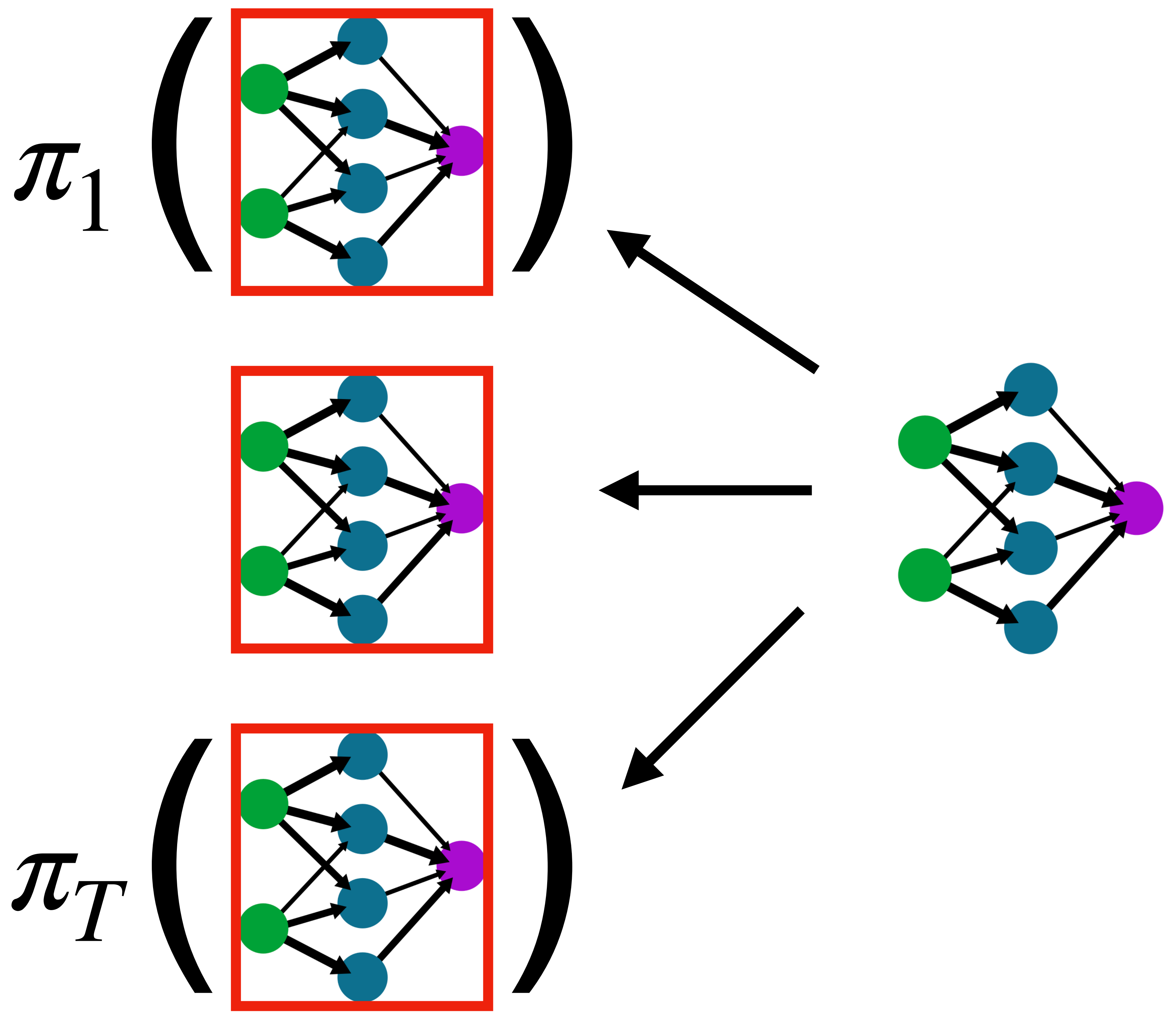
Provenance via independence testing



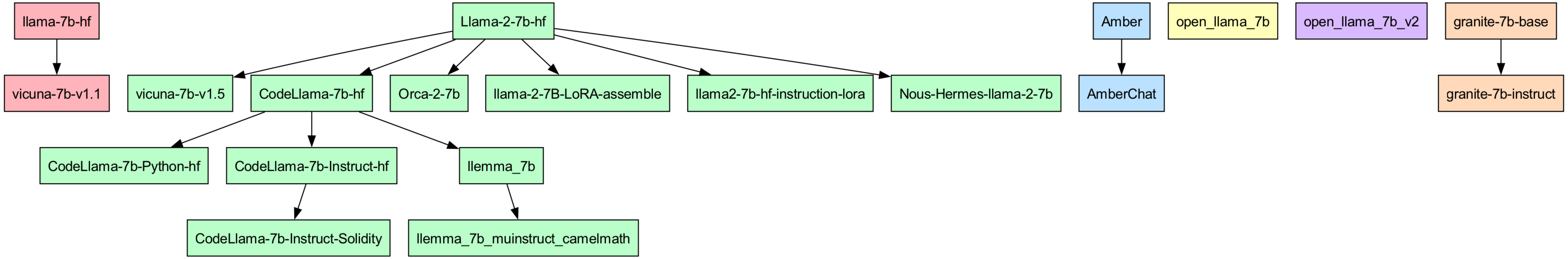
Provenance via independence testing



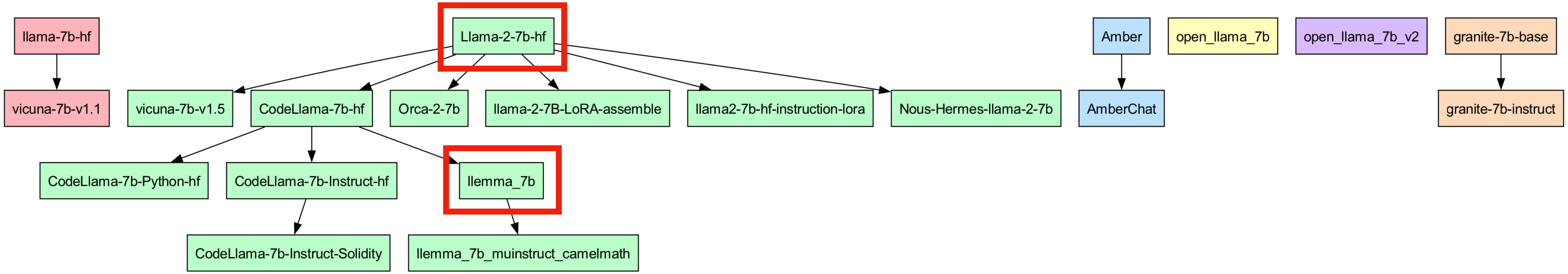
Provenance via independence testing



Empirical validation



Empirical validation



What about robustness to adversaries?

What about robustness to adversaries?

Easy to break our tests by permuting hidden units.

What about robustness to adversaries?

Easy to break our tests by permuting hidden units.

Can we design a test with non-trivial robustness?

MLPs with gated linear units (GLUs) [DFAG'17; Sha'20]

MLPs with gated linear units (GLUs) [DFAG'17; Sha'20]

Standard: $\theta = (W_1, W_2)$ $f(x;\theta) = W_2\sigma(W_1 x)$

MLPs with gated linear units (GLUs) [DFAG'17; Sha'20]

Standard: $\theta = (W_1, W_2)$ $f(x;\theta) = W_2\sigma(W_1 x)$

GLU: $\theta = (W_u, W_g, W_d)$ $f(x;\theta) = W_d(\sigma(W_g x) \odot W_u x)$

Matching activations between models

$$\theta = (W_u, W_g, W_d) \qquad f(x;\theta) = W_d(\sigma(W_g x) \odot W_u x)$$

$$\theta' = (W'_u, W'_g, W'_d)$$

''

Matching activations between models

$$\theta = (W_u, W_g, W_d) \qquad f(x;\theta) = W_d(\sigma(W_g x) \odot W_u x)$$

$$\theta' = (W'_u, W'_g, W'_d) \qquad \text{''}$$

$$\phi(\theta, \theta') = \underbrace{\text{match}(W_g X, W'_g X)}_{\pi_g}, \underbrace{\text{match}(W_u X, W'_u X)}_{\pi_u}$$

Matching activations between models

$$\theta = (W_u, W_g, W_d) \qquad f(x;\theta) = W_d(\sigma(W_g x) \odot W_u x)$$

$$\theta' = (W'_u, W'_g, W'_d) \qquad ''$$

$$\phi(\theta, \theta') = \text{spearman-pval}(\underbrace{\text{match}(W_g X, W'_g X)}_{\pi_g}, \underbrace{\text{match}(W_u X, W'_u X)}_{\pi_u})$$

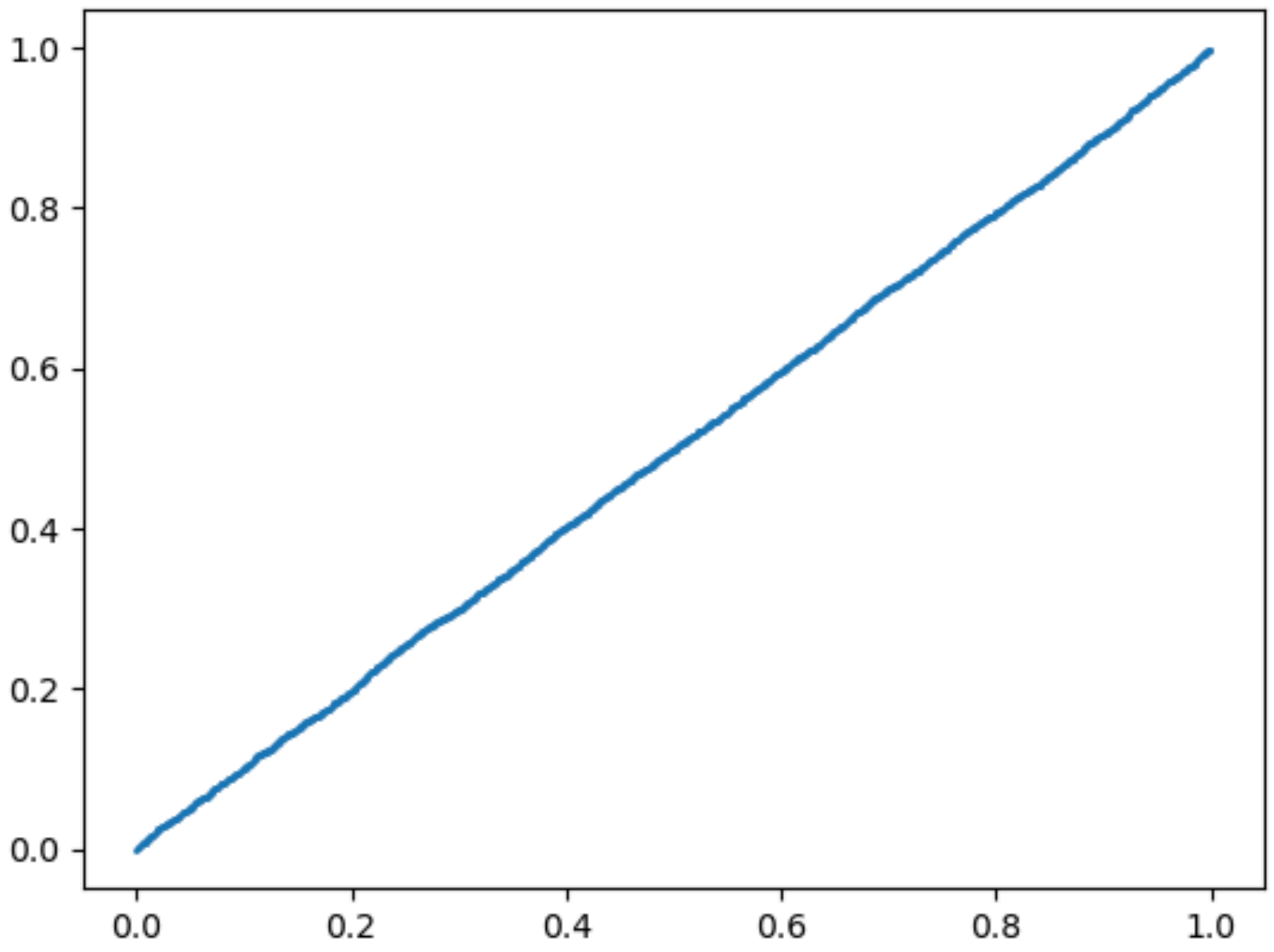
Empirical validation: precision

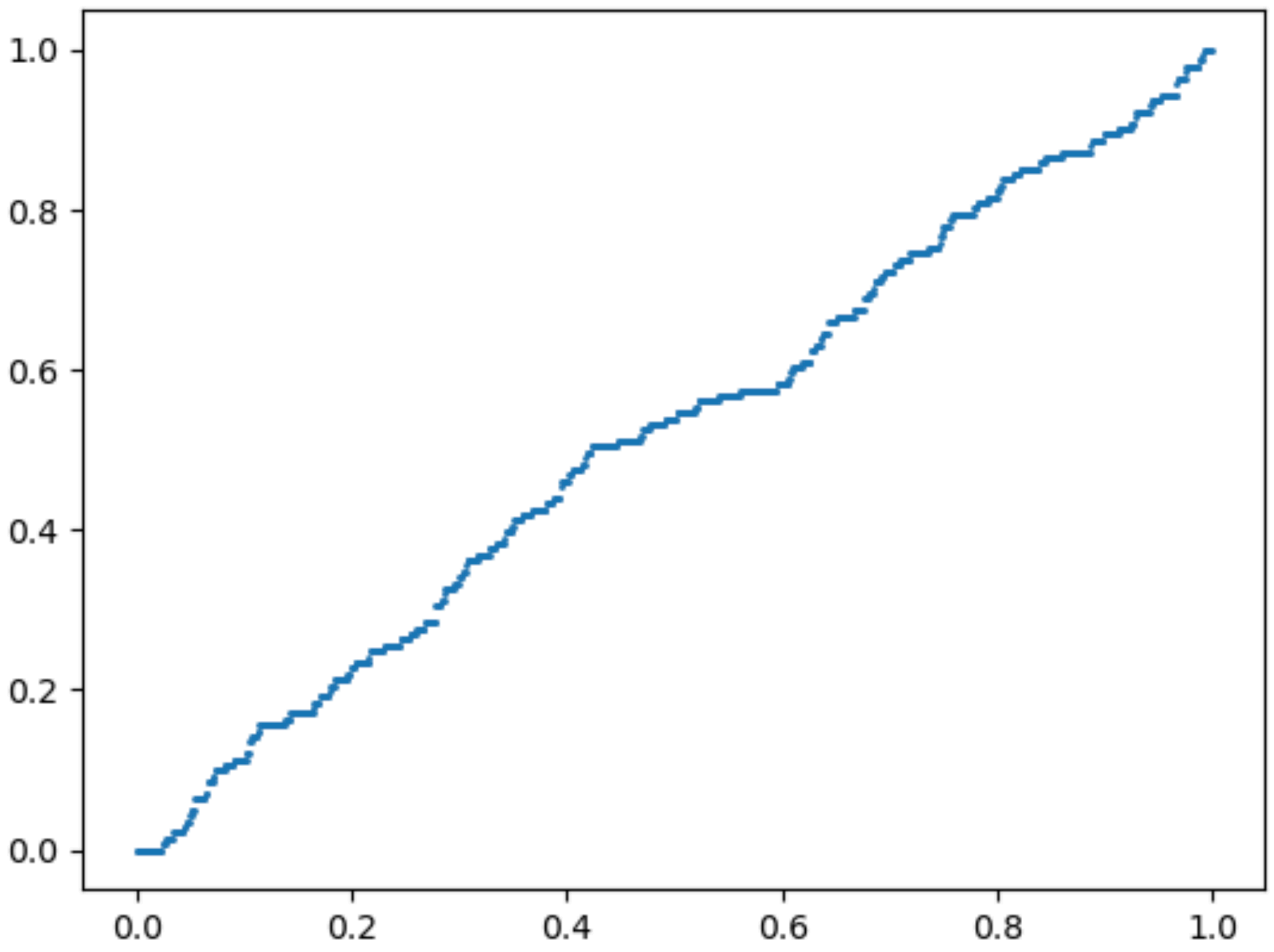
Empirical validation: precision

blockwise

aggregated

$\hat{P}(\phi < x)$





x

Empirical validation: recall

Goal: robustness to output-preserving transformations.

Empirical validation: recall

Goal: robustness to output-preserving transformations.

But how do we exhaustively enumerate these transformations? [ZZWL'24]

Empirical validation: recall

Crazy idea: let's retrain each MLP from scratch (by distilling activations)

Empirical validation: recall

Crazy idea: let's retrain each MLP from scratch (by distilling activations)

We found ϕ remains (very) small after doing this...

Empirical validation: recall

Crazy idea: let's retrain each MLP from scratch (by distilling activations)

We found ϕ remains (very) small after doing this...

...but not after retraining the entire Transformer model.

Empirical validation: recall

Crazy idea: let's retrain each MLP from scratch (by distilling activations)

We found ϕ remains (very) small after doing this...

...but not after retraining the entire Transformer model.

???

Sally Zhu

Ahmed Ahmed

John Thickstun

Tatsu
Hashimoto

Percy Liang

References

[KGW+'23] John Kirchenbauer, Jonas Geiping, Yuxin Wen, Jonathan Katz, Ian Miers, and Tom Goldstein. (2023) A Watermark for Large Language Models.

[AK'23] Scott Aaronson and Hendrik Kirchner. (2023) Watermarking GPT Outputs.

[CGZ'24] Miranda Christ, Sam Gunn, and Or Zamir. (2023) Undetectable Watermarks for Language Models.

[CG'24] Miranda Christ and Sam Gunn. (2024) Pseudorandom Error-Correcting Codes.

[GM'24] Noah Golowich and Ankur Moitra. (2024) Edit Distance Robust Watermarks for Language Models.

[GG'24] Surendra Ghentiyala and Venkatesan Guruswami. (2024) New Constructions of Pseudorandom Codes.

[DFAG'17] Yann Dauphin, Angela Fan, Michael Auli, and David Grangier. (2017) Language Modeling for Gated Convolutional Networks.

[Sha'20] Noam Shazeer. (2020) GLU Variants Improve Transformer.

[ZZWL'24] Boyi Zheng, Chenghu Zhou, Xinbing Wang, and Zhouhan Lin. (2024) Human-readable Fingerprint for Large Language Models.