When all you have are Logits...

Towards (Closed-Source) LLM Accountability via Logit Signatures

Swabha Swayamdipta
Assistant Professor, USC Viterbi CS
NSF - OSGAI Workshop
March 26, 2024

2

olmo

LLM360

MISTRAL AI_

2

Gemini

GPT - 4

Claude AI

OLMo

LLM360

MISTRAL AI_

GP 4

Gemini

Claude AI

olmo

LLM360

MISTRAL AI_

2

Logits of API-Protected LLMs Leak Proprietary Information

Matthew Finlayson **Xiang Ren** **Swabha Swayamdipta**
Thomas Lord Department of Computer Science
University of Southern California
{mfinlays, xiangren, swabhas}@usc.edu

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Logits of API-Protected LLMs Leak Proprietary Information

Matthew Finlayson Xiang Ren Swabha Swayamdipta
Thomas Lord Department of Computer Science
University of Southern California
{mfinlays, xiangren, swabhas}@usc.edu

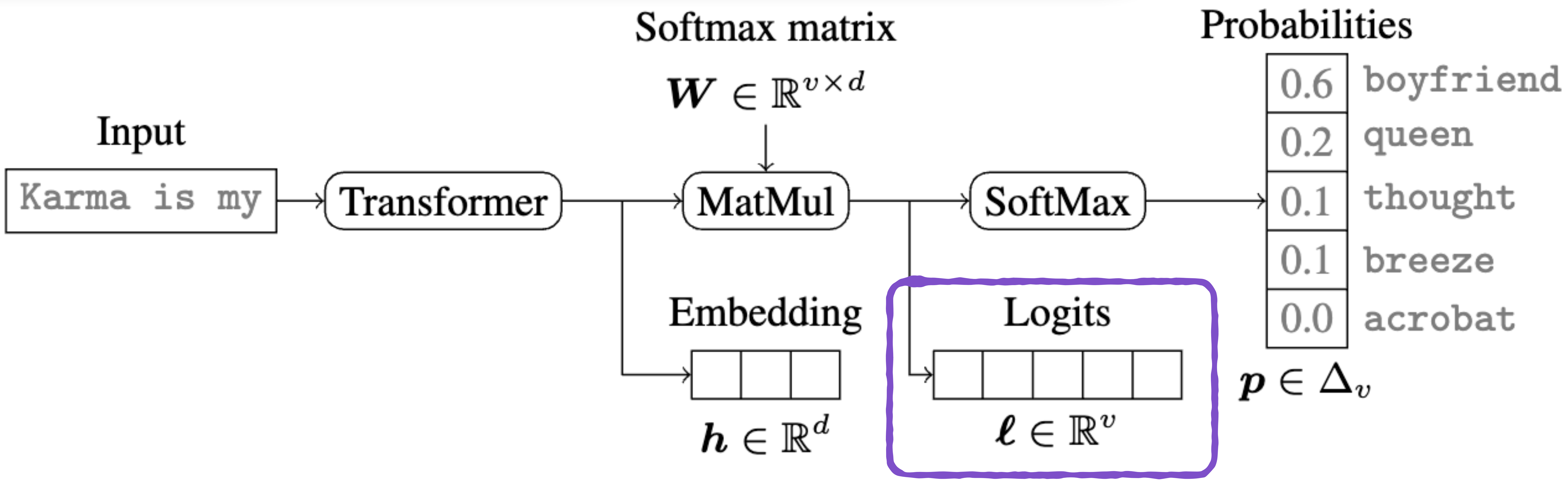
Logits can reveal the hidden dimensionality!

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Logits of API-Protected LLMs Leak Proprietary Information

Matthew Finlayson **Xiang Ren** **Swabha Swayamdipta**
Thomas Lord Department of Computer Science
University of Southern California
{mfinlays, xiangren, swabhas}@usc.edu

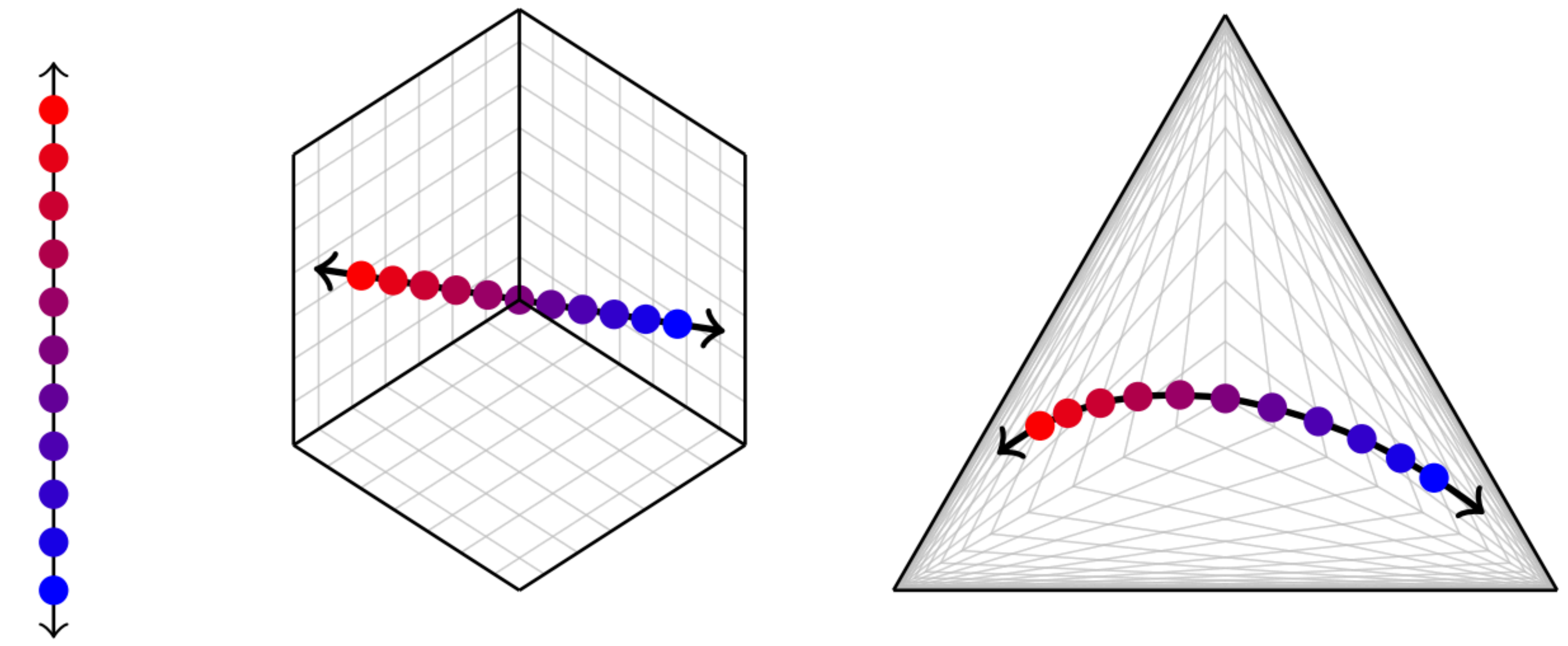
Logits can reveal the hidden dimensionality!



Finlayson, Ren & **Swayamdipta**, Under Submission 2024

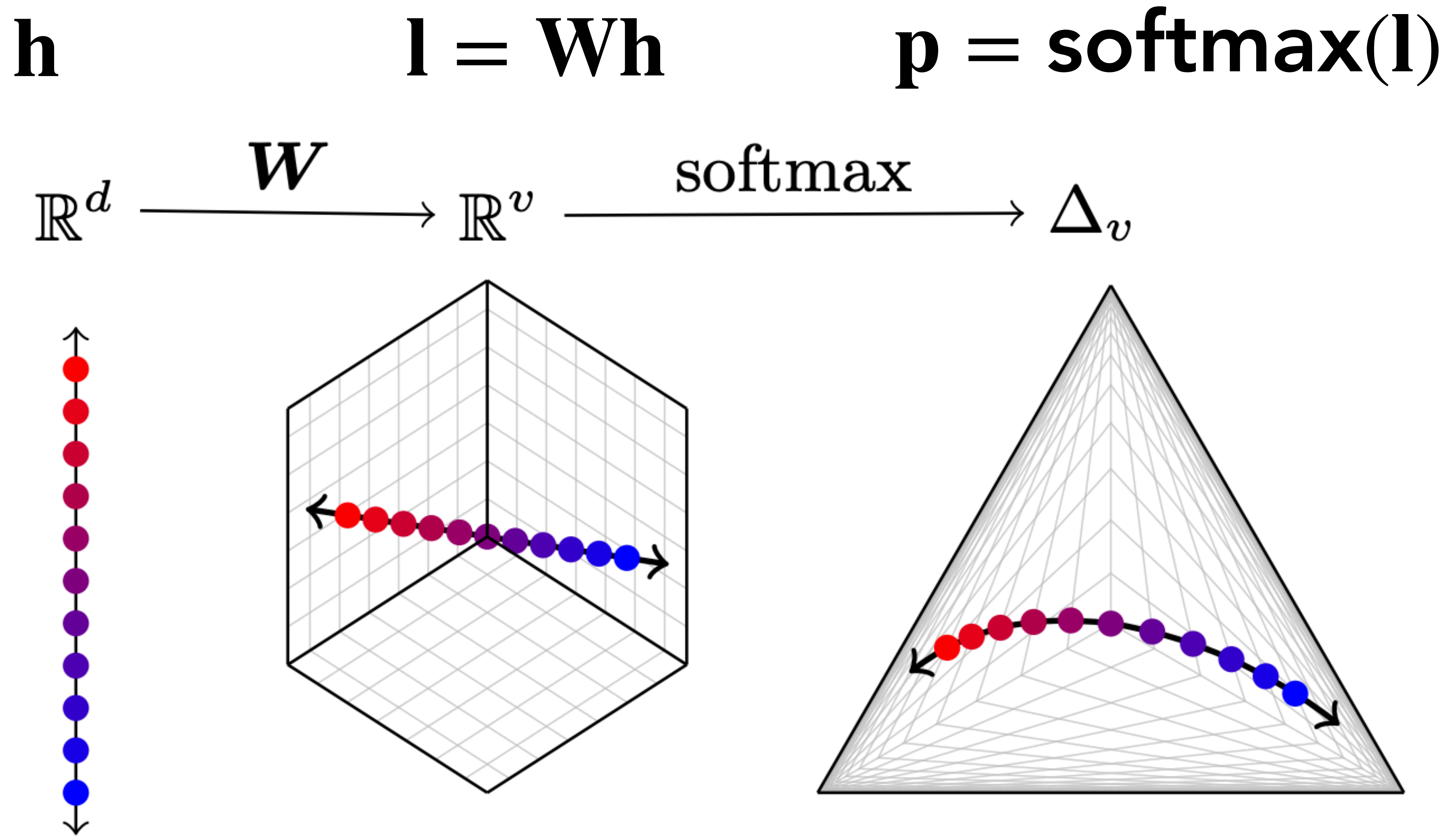
$$\mathbf{h} \qquad \mathbf{l} = \mathbf{W}\mathbf{h} \qquad \mathbf{p} = \mathrm{softmax}(\mathbf{l})$$

$$\mathbb{R}^d \xrightarrow{\boldsymbol{W}} \mathbb{R}^v \xrightarrow{\text{softmax}} \Delta_v$$



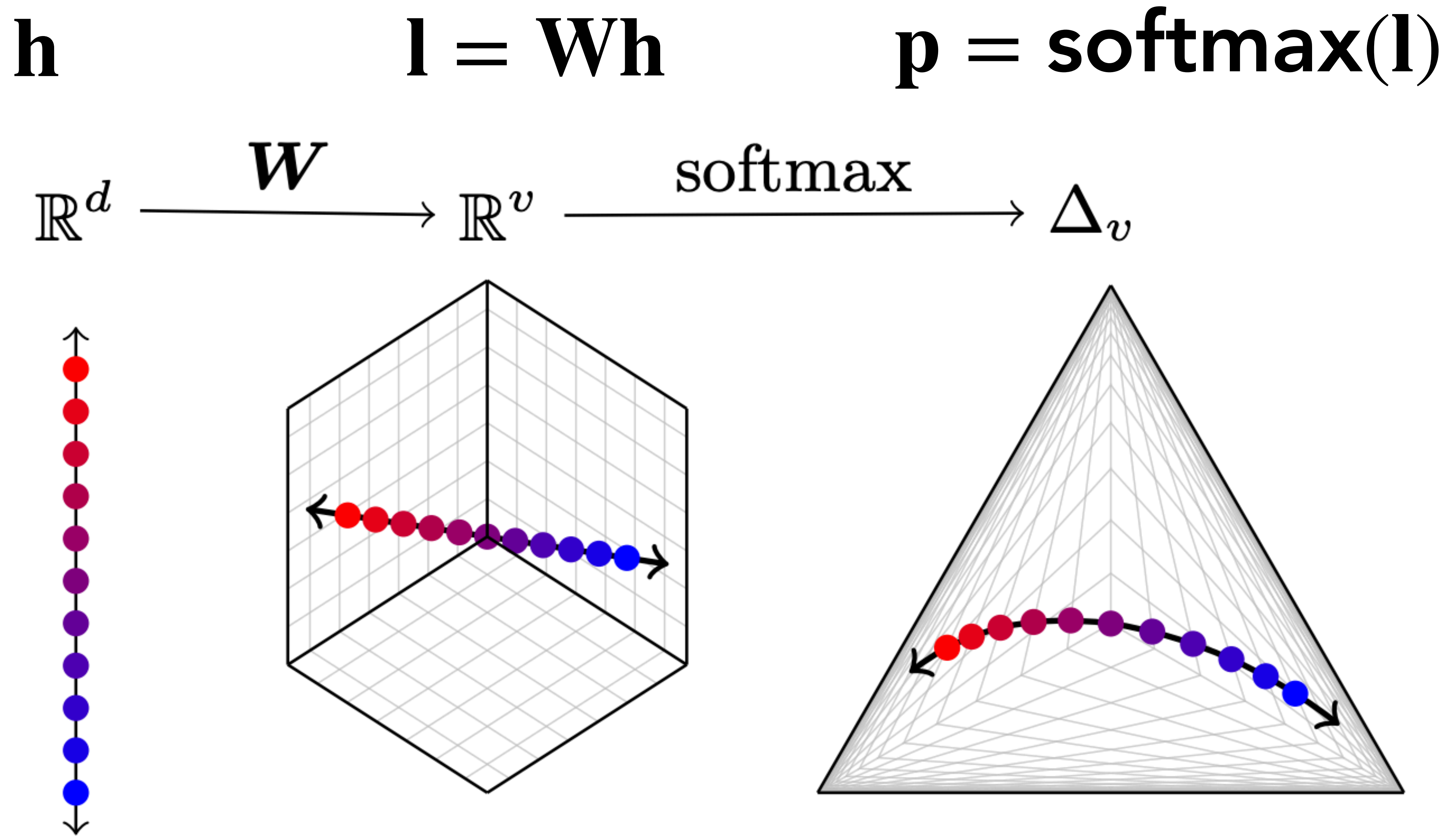
Finlayson, Ren & **Swayamdipta**, Under Submission 2024

- LM outputs are projected from the hidden dimension d to v-dimensional logit and probability vectors, thus occupying a d-dimensional subspace of $\mathbb{R}^v$ or Δ_v, respectively



Finlayson, Ren & **Swayamdipta**, Under Submission 2024

- LM outputs are projected from the hidden dimension d to v-dimensional logit and probability vectors, thus occupying a d-dimensional subspace of $\mathbb{R}^v$ or Δ_v, respectively
- This final layer is thus low-rank, since $v \gg d$

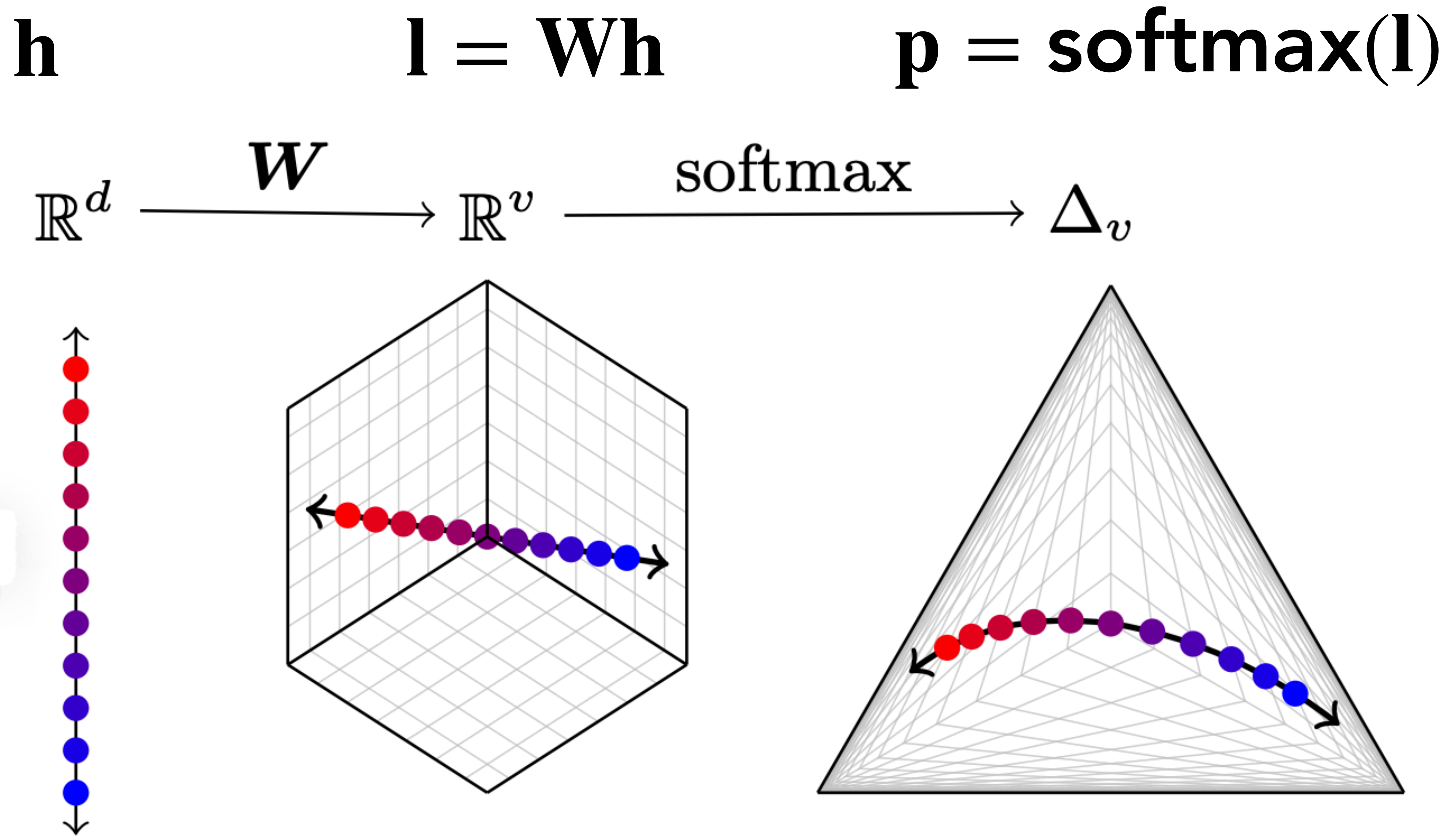


Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Language Models have a Softmax Bottleneck

- LM outputs are projected from the hidden dimension d to v-dimensional logit and probability vectors, thus occupying a d-dimensional subspace of $\mathbb{R}^v$ or Δ_v, respectively
- This final layer is thus low-rank, since $v \gg d$

Yang et al., ICLR 2018; Finlayson et al., ICLR 2024



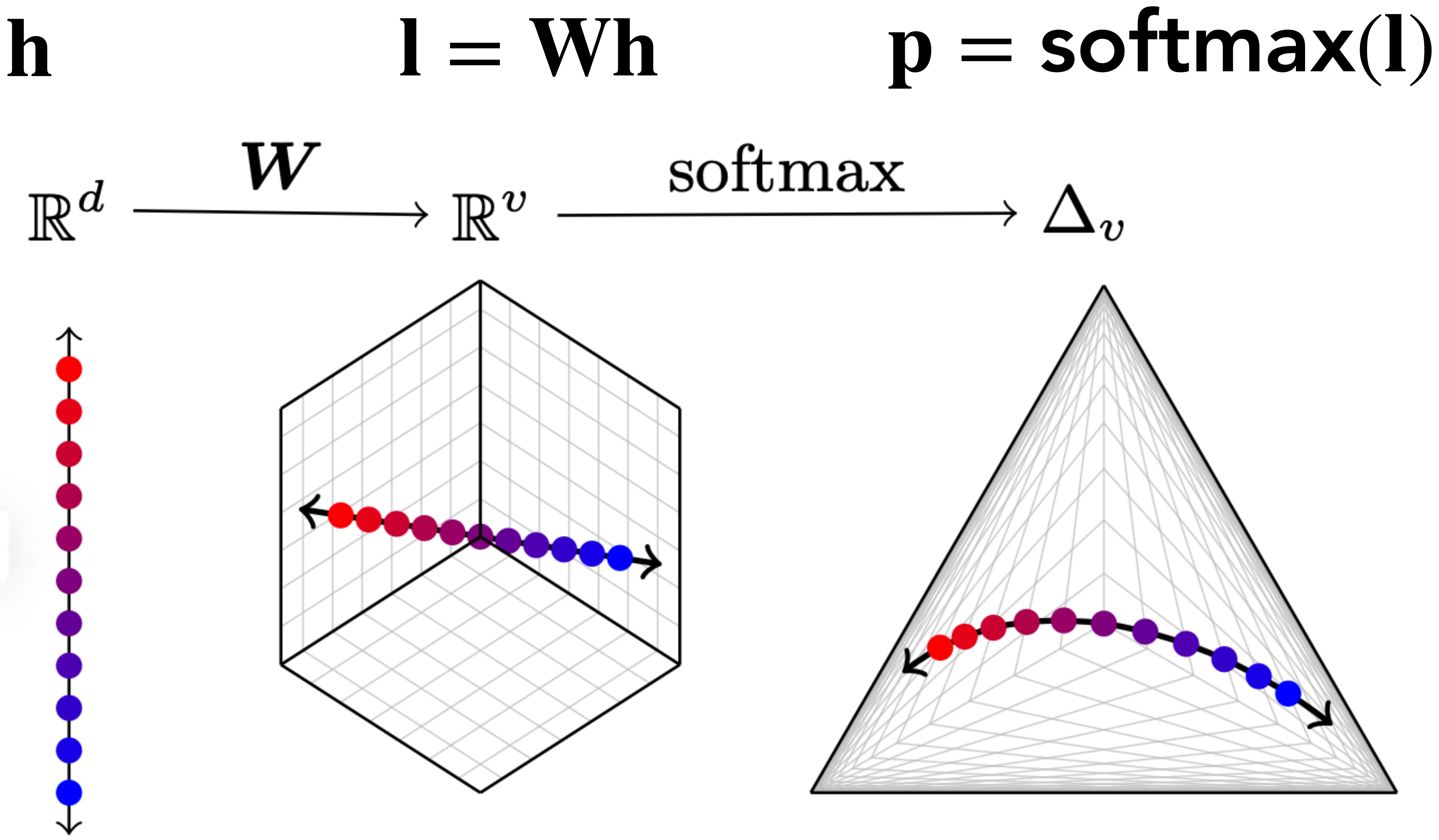
Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Language Models have a Softmax Bottleneck

- LM outputs are projected from the hidden dimension d to v-dimensional logit and probability vectors, thus occupying a d-dimensional subspace of $\mathbb{R}^v$ or Δ_v, respectively
- This final layer is thus low-rank, since $v \gg d$

Yang et al., ICLR 2018; Finlayson et al., ICLR 2024

- A collection of d linearly independent outputs $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^d \in \Delta_v$ from the model will form a basis for the model's image



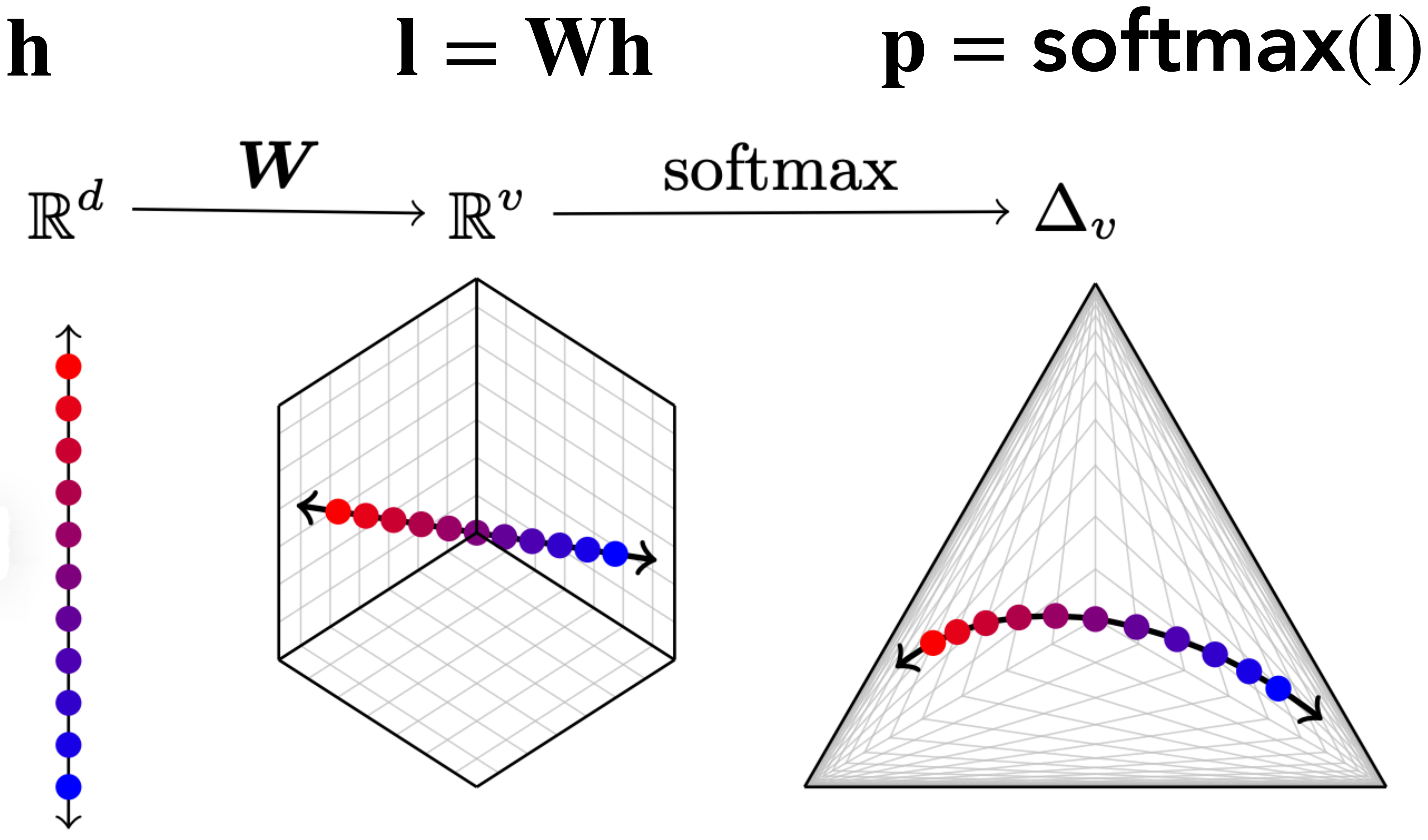
Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Language Models have a Softmax Bottleneck

- LM outputs are projected from the hidden dimension d to v-dimensional logit and probability vectors, thus occupying a d-dimensional subspace of $\mathbb{R}^v$ or Δ_v, respectively
- This final layer is thus low-rank, since $v \gg d$

Yang et al., ICLR 2018; Finlayson et al., ICLR 2024

- A collection of d linearly independent outputs $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^d \in \Delta_v$ from the model will form a basis for the model's image



Targeted queries to the LM's API to extract $n > d$ logit vectors will result in extracting its hidden dimension, d and related information

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

How to recover model logits from APIs?

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

How to recover model logits from APIs?

- Access to top-k log probabilities

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

How to recover model logits from APIs?

- Access to top-k log probabilities
- Logit Bias: A common API option that allows users to add bias to the logits for specific tokens

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

How to recover model logits from APIs?

- Access to top-k log probabilities
- Logit Bias: A common API option that allows users to add bias to the logits for specific tokens
- We can recover this while preserving numerical stability in $v/(k-1)$ API calls, which costs ~$500 USD, for GPT-3.5-turbo

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

How to recover model logits from APIs?

- Access to top-k log probabilities
- Logit Bias: A common API option that allows users to add bias to the logits for specific tokens
- We can recover this while preserving numerical stability in $v/(k-1)$ API calls, which costs ~$500 USD, for GPT-3.5-turbo
- If the hidden size is known, this can be done in d API calls; in general, in $O(d)$ calls

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Key Result: Hidden Dimensionality

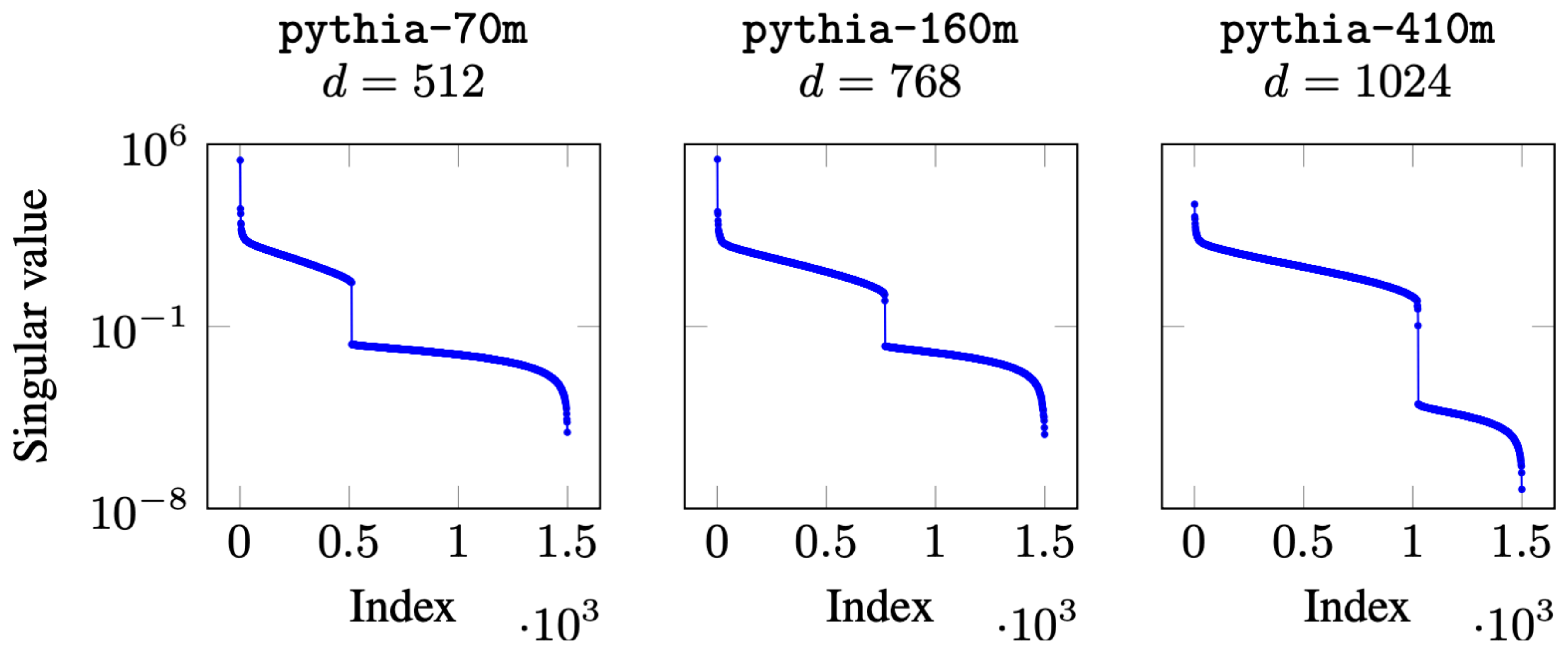
Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Key Result: Hidden Dimensionality

- We collect outputs $\mathbf{p}^i$ one at a time until the number of linearly independent outputs in the collection (obtained via SVD) stops increasing, which will occur when we have collected $d + 1$ outputs

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

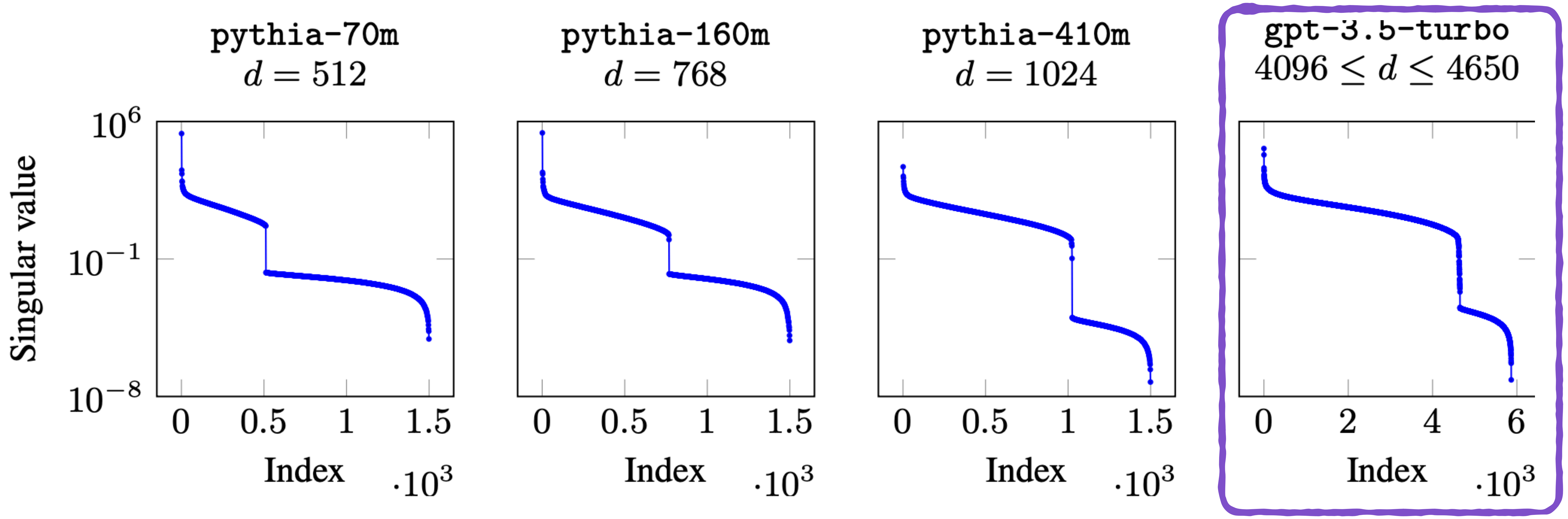
Key Result: Hidden Dimensionality



- We collect outputs $\mathbf{p}^i$ one at a time until the number of linearly independent outputs in the collection (obtained via SVD) stops increasing, which will occur when we have collected $d+1$ outputs

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Key Result: Hidden Dimensionality



- We collect outputs $\mathbf{p}^i$ one at a time until the number of linearly independent outputs in the collection (obtained via SVD) stops increasing, which will occur when we have collected $d + 1$ outputs
- GPT-3.5-Turbo has hidden dimension close to 4096 and is likely a 7B model!

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Model Signature

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Model Signature

- Any collection of d linearly independent LM outputs $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^d \in \Delta_v$ form a basis for the image of the model

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Model Signature

- Any collection of d linearly independent LM outputs $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^d \in \Delta_v$ form a basis for the image of the model

- We call the image of the model, i.e. LM outputs in either $\mathbf{W}$ or $\mathbf{p}$, the model signature

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Model Signature

- Any collection of d linearly independent LM outputs $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^d \in \Delta_v$ form a basis for the image of the model
- We call the image of the model, i.e. LM outputs in either $\mathbf{W}$ or $\mathbf{p}$, the model signature
- All LM outputs can be expressed as a unique linear combination of these d outputs

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Model Signature

- Any collection of d linearly independent LM outputs $\mathbf{p}^1, \mathbf{p}^2, \ldots, \mathbf{p}^d \in \Delta_v$ form a basis for the image of the model
- We call the image of the model, i.e. LM outputs in either $\mathbf{W}$ or $\mathbf{p}$, the model signature
- All LM outputs can be expressed as a unique linear combination of these d outputs
- Model signatures are unique!

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

LM outputs can be identified via model signatures

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

LM outputs can be identified via model signatures

- Even different checkpoints from the same LM have largely disjoint model signatures

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

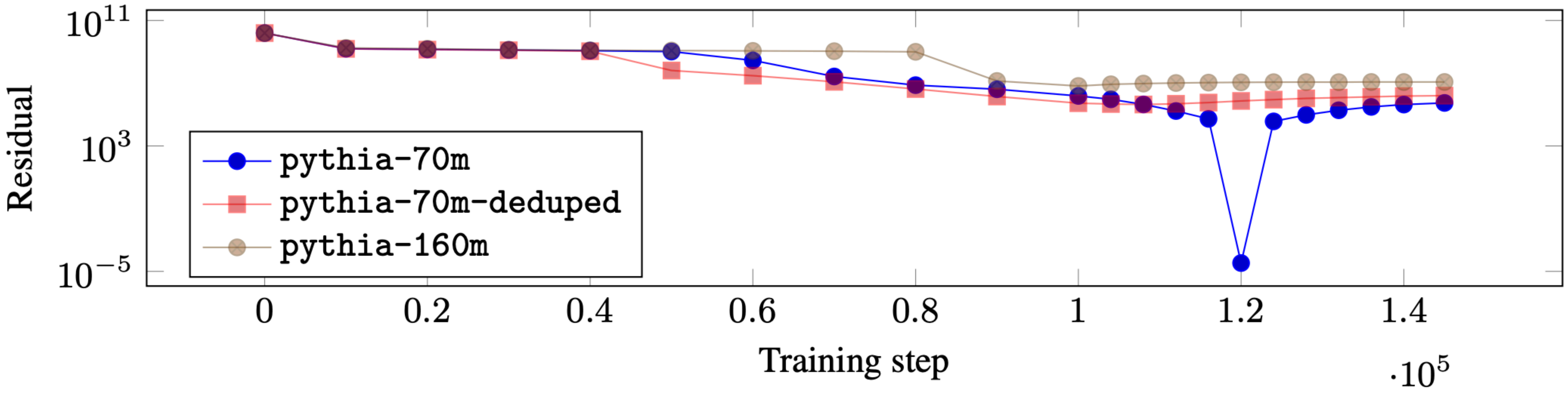
LM outputs can be identified via model signatures

- Even different checkpoints from the same LM have largely disjoint model signatures
- Possible to determine precisely which LM produced a particular output, using only API access to a set of LMs and without knowing the exact inputs to the model.

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

LM outputs can be identified via model signatures

- Even different checkpoints from the same LM have largely disjoint model signatures
- Possible to determine precisely which LM produced a particular output, using only API access to a set of LMs and without knowing the exact inputs to the model.



Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Other Applications of Model Signatures

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Other Applications of Model Signatures

- Detecting model updates and changes to hidden prompts
- Improved LLM Inversion

Morris et al., 2023

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Other Applications of Model Signatures

- Detecting model updates and changes to hidden prompts
- Improved LLM Inversion
  Morris et al., 2023
- Finding unargmaxable tokens
  Demeter et al., 2020; Grivas et al., 2023

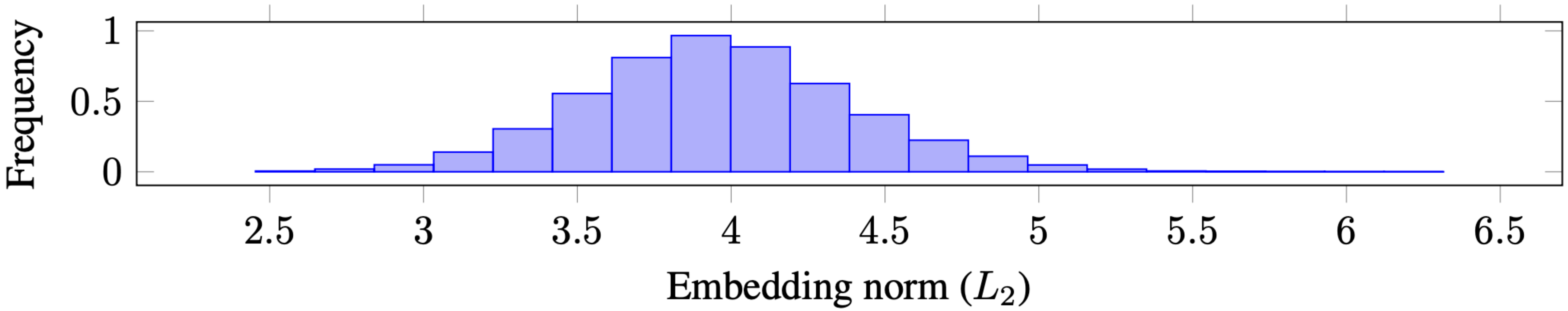
Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Other Applications of Model Signatures

- Detecting model updates and changes to hidden prompts
- Improved LLM Inversion

  Morris et al., 2023
- Finding unargmaxable tokens

  Demeter et al., 2020; Grivas et al., 2023
- Recovering the softmax parameter matrix $\mathbf{W}$ (up to a rotation)



Finlayson, Ren & **Swayamdipta**, Under Submission 2024

So what?

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

So what?

- LLM providers might want to mitigate the risks of an attack

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

So what?

- LLM providers might want to mitigate the risks of an attack
  - Remove API access to top-k logprobs or logit bias

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

So what?

- LLM providers might want to mitigate the risks of an attack
  - Remove API access to top-k logprobs or logit bias
  - Remove access to LM probabilities

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

So what?

- LLM providers might want to mitigate the risks of an attack
  - Remove API access to top-k logprobs or logit bias
  - Remove access to LM probabilities
  - Removing the softmax bottleneck altogether

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

So what?

- LLM providers might want to mitigate the risks of an attack
  - Remove API access to top-k logprobs or logit bias
  - Remove access to LM probabilities
  - Removing the softmax bottleneck altogether

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

So what?

- LLM providers might want to mitigate the risks of an attack
  - Remove API access to top-k logprobs or logit bias
  - Remove access to LM probabilities
  - Removing the softmax bottleneck altogether

- More importantly, this is a step towards model accountability

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

So what?

- LLM providers might want to mitigate the risks of an attack
  - Remove API access to top-k logprobs or logit bias
  - Remove access to LM probabilities
  - Removing the softmax bottleneck altogether

- More importantly, this is a step towards model accountability
  - Building trust between API users and providers

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

So what?

- LLM providers might want to mitigate the risks of an attack
  - Remove API access to top-k logprobs or logit bias
  - Remove access to LM probabilities
  - Removing the softmax bottleneck altogether

- More importantly, this is a step towards model accountability
  - Building trust between API users and providers
  - Implementing efficient protocols for model auditing

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

So what?

- LLM providers might want to mitigate the risks of an attack
  - Remove API access to top-k logprobs or logit bias
  - Remove access to LM probabilities
  - Removing the softmax bottleneck altogether

- More importantly, this is a step towards model accountability
  - Building trust between API users and providers
  - Implementing efficient protocols for model auditing
  - Verifying LM identity and ownership

Finlayson, Ren & **Swayamdipta**, Under Submission 2024

Stealing Part of a Production Language Model

Nicholas Carlini [1] Daniel Paleka [2] Krishnamurthy (Dj) Dvijotham [1] Thomas Steinke [1] Jonathan Hayase [3]
A. Feder Cooper [1] Katherine Lee [1] Matthew Jagielski [1] Milad Nasr [1] Arthur Conmy [1] Eric Wallace [4]
David Rolnick [5] Florian Tramèr [2]

arXiv:2403.06634v1 [cs.CR] 11 Mar 2024

Logits of API-Protected LLMs Leak Proprietary Information

Matthew Finlayson Xiang Ren Swabha Swayamdipta
Thomas Lord Department of Computer Science
University of Southern California
{mfinlays, xiangren, swabhas}@usc.edu

arXiv:2403.09539v2 [cs.CL] 15 Mar 2024

Stealing Part of a Production Language Model

Nicholas Carlini[1] Daniel Paleka[2] Krishnamurthy (Dj) Dvijotham[1] Thomas Steinke[1] Jonathan Hayase[3]
A. Feder Cooper[1] Katherine Lee[1] Matthew Jagielski[1] Milad Nasr[1] Arthur Conmy[1] Eric Wallace[4]
David Rolnick[5] Florian Tramèr[2]

arXiv:2403.06634v1 [cs.CR] 11 Mar 2024

Simultaneous Discovery!

Logits of API-Protected LLMs Leak Proprietary Information

Matthew Finlayson Xiang Ren Swabha Swayamdipta
Thomas Lord Department of Computer Science
University of Southern California
{mfinlays, xiangren, swabhas}@usc.edu

arXiv:2403.09539v2 [cs.CL] 15 Mar 2024

Logits of API-Protected LLMs Leak Proprietary Information

Matthew Finlayson Xiang Ren Swabha Swayamdipta
Thomas Lord Department of Computer Science
University of Southern California
{mfinlays, xiangren, swabhas}@usc.edu

DILL