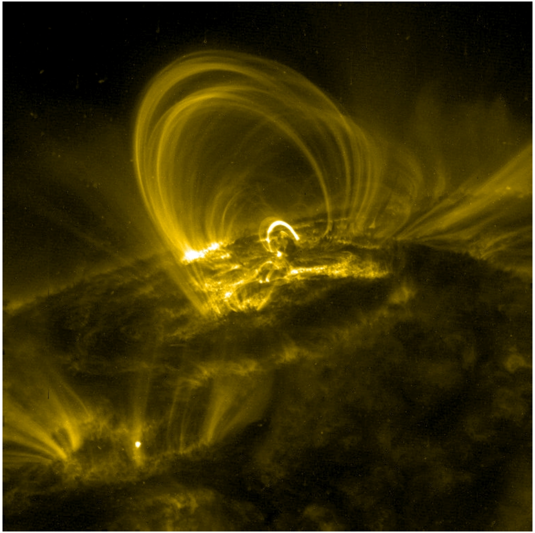
How to Find ChatGPT's Hidden Size, and Other Low-rank Logit Tricks

Matthew Finlayson Xiang Ren Swabha Swayamdipta

University of Southern California

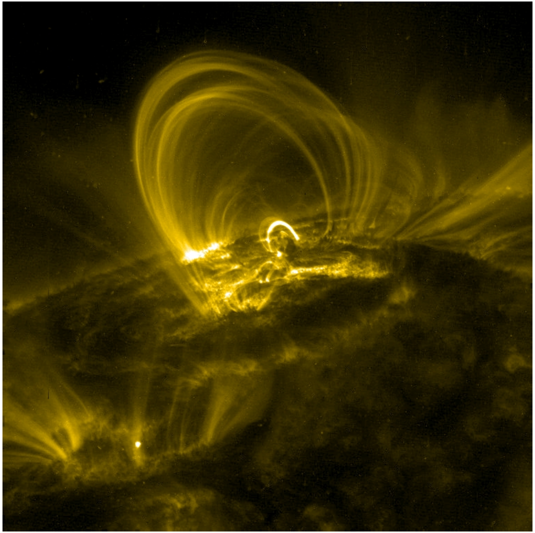
April 8, 2024

The solar corona



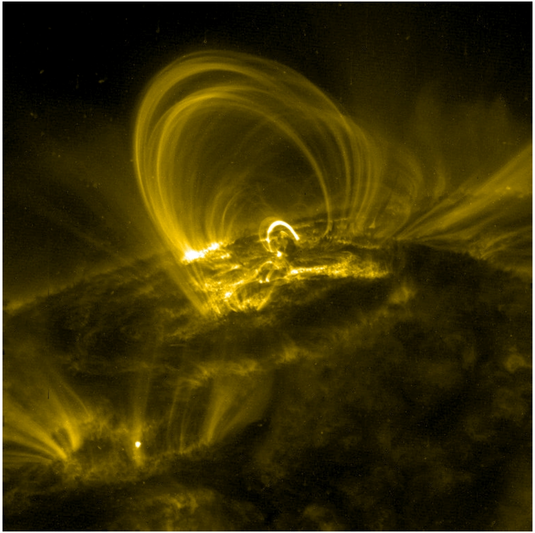
- Looks really cool.
- Made of plasma (ions).
- 150–450× hotter than the sun surface.
- The sun's magnetic field causes *coronal loops*.

The solar corona



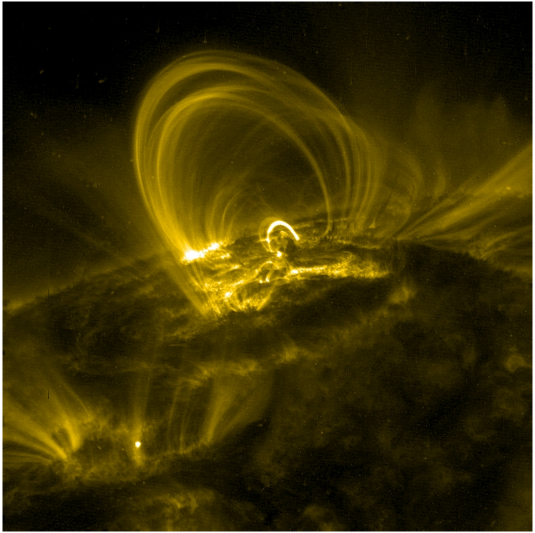
- Looks really cool.
- Made of plasma (ions).
- 150–450× hotter than the sun surface.
- The sun's magnetic field causes *coronal loops*.

The solar corona



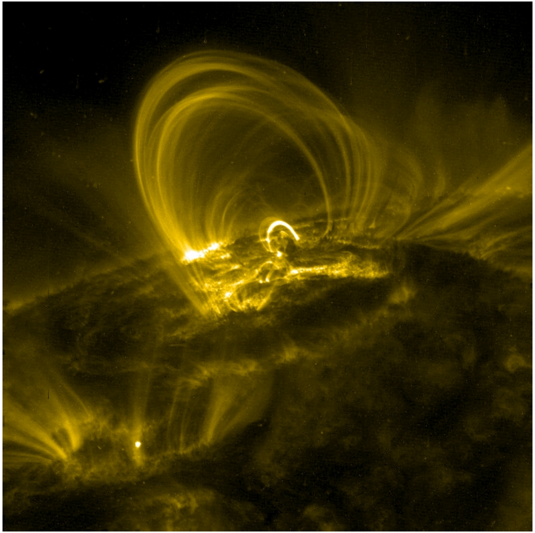
- Looks really cool.
- Made of plasma (ions).
- 150–450× hotter than the sun surface.
- The sun's magnetic field causes *coronal loops*.

The solar corona



- Looks really cool.
- Made of plasma (ions).
- 150–450× hotter than the sun surface.
- The sun's magnetic field causes *coronal loops*.

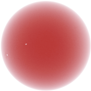
The solar corona



- Looks really cool.
- Made of plasma (ions).
- 150–450× hotter than the sun surface.
- The sun's magnetic field causes *coronal loops*.

The analogy

- Scientists study the *structure* of coronal loops to learn about the sun's *internal* magnetic fields.
- We can study the *structure* of LLM outputs to learn about their *internal* details.

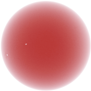


- The sun
- The sun's magnetic field
- The solar corona

- Proprietary LLMs
- Non-public model details
- LLM API outputs

The analogy

- Scientists study the *structure* of coronal loops to learn about the sun's *internal* magnetic fields.
- We can study the *structure* of LLM outputs to learn about their *internal* details.



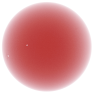
- The sun
- The sun's magnetic field
- The solar corona

OpenAI

- Proprietary LLMs
- Non-public model details
- LLM API outputs

The analogy

- Scientists study the *structure* of coronal loops to learn about the sun's *internal* magnetic fields.
- We can study the *structure* of LLM outputs to learn about their *internal* details.



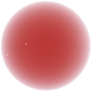
- The sun
- The sun's magnetic field
- The solar corona

OpenAI

- Proprietary LLMs
- Non-public model details
- LLM API outputs

The analogy

- Scientists study the *structure* of coronal loops to learn about the sun's *internal* magnetic fields.
- We can study the *structure* of LLM outputs to learn about their *internal* details.



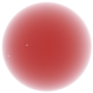
OpenAI

- The sun
- The sun's magnetic field
- The solar corona

- Proprietary LLMs
- Non-public model details
- LLM API outputs

The analogy

- Scientists study the *structure* of coronal loops to learn about the sun's *internal* magnetic fields.
- We can study the *structure* of LLM outputs to learn about their *internal* details.



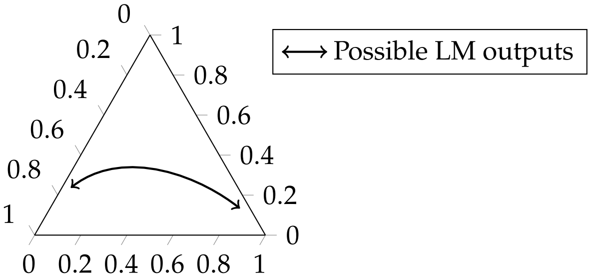
- The sun
- The sun's magnetic field
- The solar corona

- Proprietary LLMs
- Non-public model details
- LLM API outputs

The structure of LLM outputs

LLM outputs lie within a low-dimensional space.

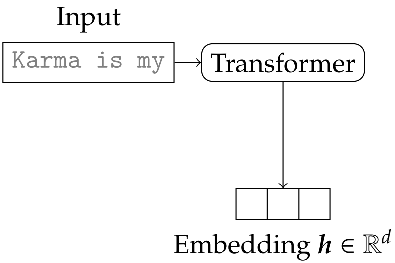
Space of probability distributions over 3 items



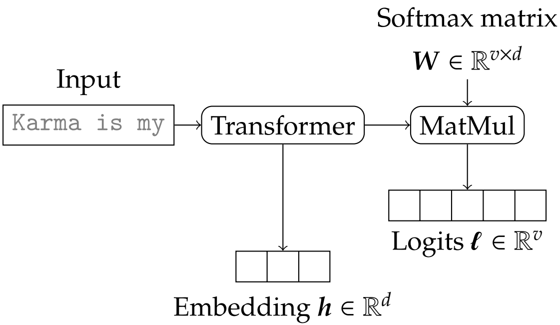
Section 1

The technical details

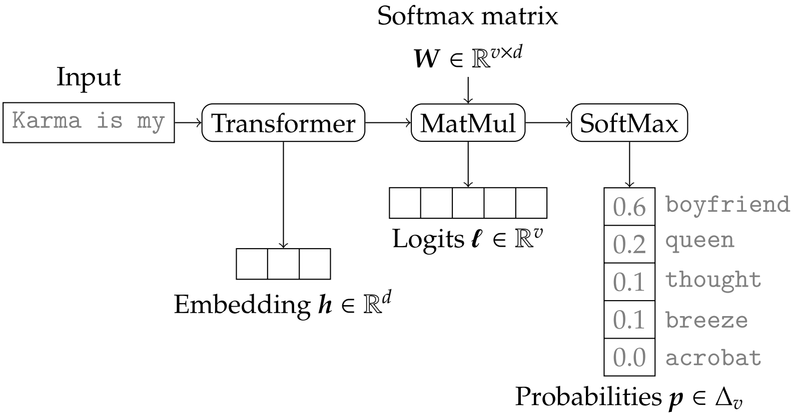
LLM architecture



LLM architecture



LLM architecture



The set of next-token distributions

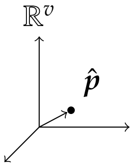
Next-token distributions over a vocabulary of size v are

- v-tuples of reals.
- Non-negative, sum to 1.
- Known as the v-simplex, or Δ_v.

The set of next-token distributions

Next-token distributions over a vocabulary of size v are

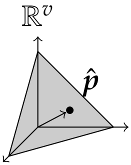
- v-tuples of reals.
- Non-negative, sum to 1.
- Known as the v-simplex, or Δ_v.



The set of next-token distributions

Next-token distributions over a vocabulary of size v are

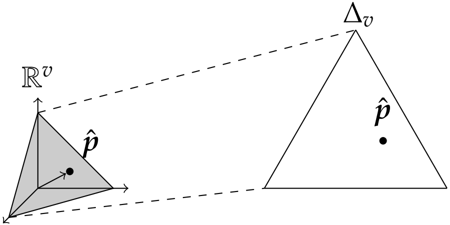
- v-tuples of reals.
- Non-negative, sum to 1.
- Known as the v-simplex, or Δ_v.



The set of next-token distributions

Next-token distributions over a vocabulary of size v are

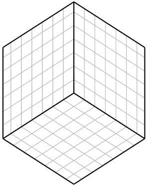
- v-tuples of reals.
- Non-negative, sum to 1.
- Known as the v-simplex, or Δ_v.



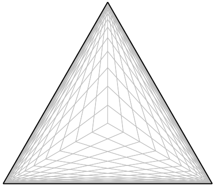
Probability distributions are vectors

- Δ_v is a *vector space*.
- The softmax function is a *linear map* $\mathbb{R}^v \to \Delta_v$.

$\mathbb{R}^v$

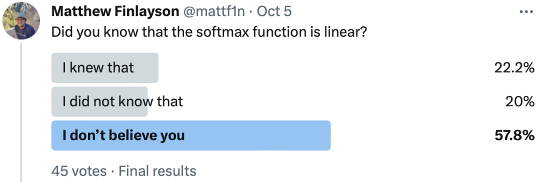


Δ_v



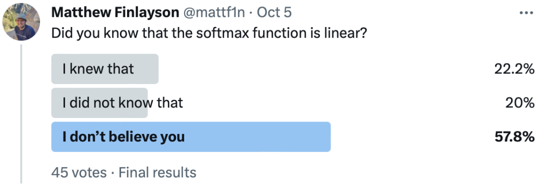
Probability distributions are vectors

- Δ_v is a *vector space*.
- The softmax function is a *linear map* $\mathbb{R}^v \longrightarrow \Delta_v$.

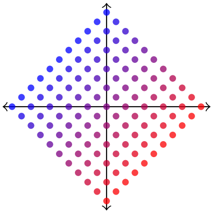


Probability distributions are vectors

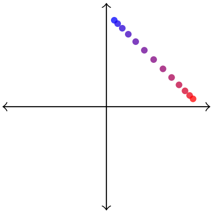
- Δ_v is a *vector space.*
- The softmax function is a *linear map* $\mathbb{R}^v \to \Delta_v$.



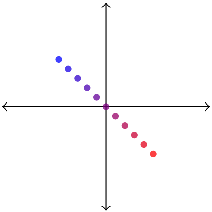
$\mathbb{R}^v$



$\mathrm{softmax}(\mathbb{R}^v)$



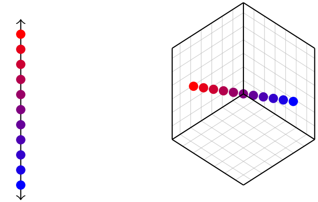
$\mathrm{clr}\left(\mathrm{softmax}(\mathbb{R}^v)\right)$



LLM outputs lie on a low-dimensional vector subspace

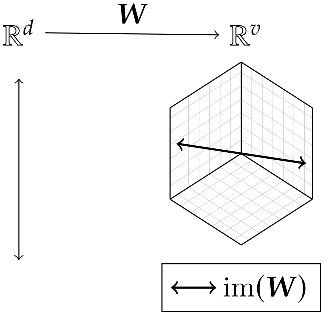
- The image of a function is its codomain.
- The dim of a linear map's image is $\leq$ the dim of its domain.
- $\text{softmax} \circ \boldsymbol{W}$ is a linear map $\mathbb{R}^d \to \Delta_v$.
- The dim of an LLM's image is at most d
- d LLM outputs form a *basis* for its image.

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



LLM outputs lie on a low-dimensional vector subspace

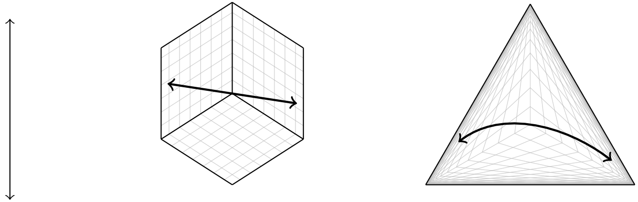
- The image of a function is its codomain.
- The dim of a linear map's image is ≤ the dim of its domain.
- softmax $\circ \boldsymbol{W}$ is a linear map $\mathbb{R}^d \to \Delta_v$.
- The dim of an LLM's image is at most d
- d LLM outputs form a *basis* for its image.



LLM outputs lie on a low-dimensional vector subspace

- The image of a function is its codomain.
- The dim of a linear map's image is $\leq$ the dim of its domain.
- softmax $\circ \boldsymbol{W}$ is a linear map $\mathbb{R}^d \to \Delta_v$.
- The dim of an LLM's image is at most d
- d LLM outputs form a *basis* for its image.

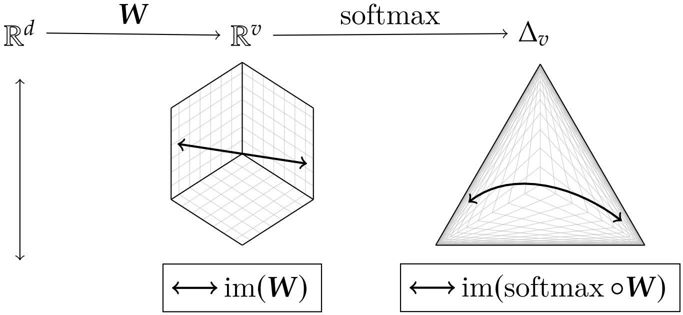
$$\mathbb{R}^d \xrightarrow{\boldsymbol{W}} \mathbb{R}^v \xrightarrow{\text{softmax}} \Delta_v$$



$\boxed{\longleftrightarrow \mathrm{im}(\boldsymbol{W})}$ $\boxed{\longleftrightarrow \mathrm{im}(\text{softmax} \circ \boldsymbol{W})}$

LLM outputs lie on a low-dimensional vector subspace

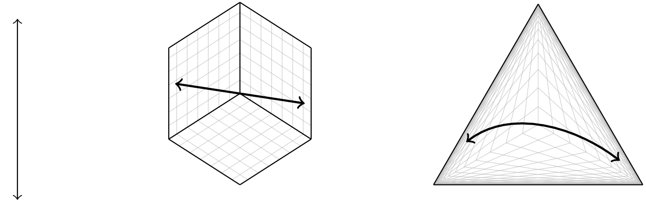
- The image of a function is its codomain.
- The dim of a linear map's image is $\leq$ the dim of its domain.
- softmax $\circ \boldsymbol{W}$ is a linear map $\mathbb{R}^d \longrightarrow \Delta_v$.
- The dim of an LLM's image is at most d
- d LLM outputs form a *basis* for its image.



LLM outputs lie on a low-dimensional vector subspace

- The image of a function is its codomain.
- The dim of a linear map's image is $\leq$ the dim of its domain.
- softmax $\circ \boldsymbol{W}$ is a linear map $\mathbb{R}^d \longrightarrow \Delta_v$.
- The dim of an LLM's image is at most d
- d LLM outputs form a *basis* for its image.

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



$\longleftrightarrow \mathrm{im}(\boldsymbol{W})$

$\longleftrightarrow \mathrm{im}(\mathrm{softmax} \circ \boldsymbol{W})$

Section 2

Consequences of knowing the LLM image

Cheap, full LLM outputs

Stealing Machine Learning Models via Prediction APIs

Florian Tramèr
EPFL

Fan Zhang
Cornell University

Ari Juels
Cornell Tech, Jacobs Institute

Michael K. Reiter
UNC Chapel Hill

Thomas Ristenpart
Cornell Tech

- Common APIs give top-k tokens and probabilities.
- Logit bias allows boosting tokens to top-k.
- Extracting full outputs takes $O(v/k)$ API calls.
- Once the LLM image is known, only $O(d/k)$ calls.
- Intuition: position in a d-dimensional subspace is fully specified by d coordinates.

Cheap, full LLM outputs

- Common APIs give top-k tokens and probabilities.
- Logit bias allows boosting tokens to top-k.
- Extracting full outputs takes $O(v/k)$ API calls.
- Once the LLM image is known, only $O(d/k)$ calls.
- Intuition: position in a d-dimensional subspace is fully specified by d coordinates.

Cheap, full LLM outputs

- Common APIs give top-k tokens and probabilities.
- Logit bias allows boosting tokens to top-k.
- Extracting full outputs takes $O(v/k)$ API calls.
- Once the LLM image is known, only $O(d/k)$ calls.
- Intuition: position in a d-dimensional subspace is fully specified by d coordinates.

Cheap, full LLM outputs

- Common APIs give top-k tokens and probabilities.
- Logit bias allows boosting tokens to top-k.
- Extracting full outputs takes $O(v/k)$ API calls.
- Once the LLM image is known, only $O(d/k)$ calls.
- Intuition: position in a d-dimensional subspace is fully specified by d coordinates.

Cheap, full LLM outputs

- Common APIs give top-k tokens and probabilities.
- Logit bias allows boosting tokens to top-k.
- Extracting full outputs takes $O(v/k)$ API calls.
- Once the LLM image is known, only $O(d/k)$ calls.
- Intuition: position in a d-dimensional subspace is fully specified by d coordinates.

Cheap, full LLM outputs

- Common APIs give top-k tokens and probabilities.
- Logit bias allows boosting tokens to top-k.
- Extracting full outputs takes $O(v/k)$ API calls.
- Once the LLM image is known, only $O(d/k)$ calls.
- Intuition: position in a d-dimensional subspace is fully specified by d coordinates.

Finding the embedding size

Collect at least d outputs from the model, check the dimension of the space that they span.

- Create a matrix $\boldsymbol{P}$ with LLM outputs as columns.
- $\boldsymbol{P}$ will have d nonzero singular values.

Finding the embedding size

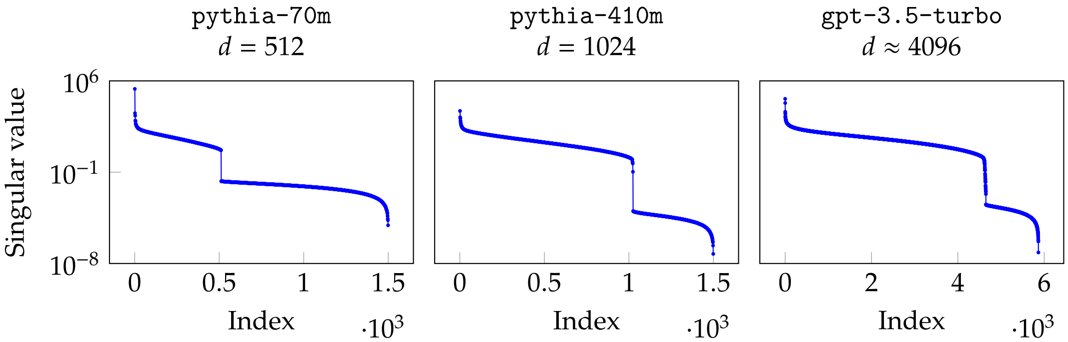
Collect at least d outputs from the model, check the dimension of the space that they span.

- Create a matrix $\boldsymbol{P}$ with LLM outputs as columns.
- $\boldsymbol{P}$ will have d nonzero singular values.

Finding the embedding size

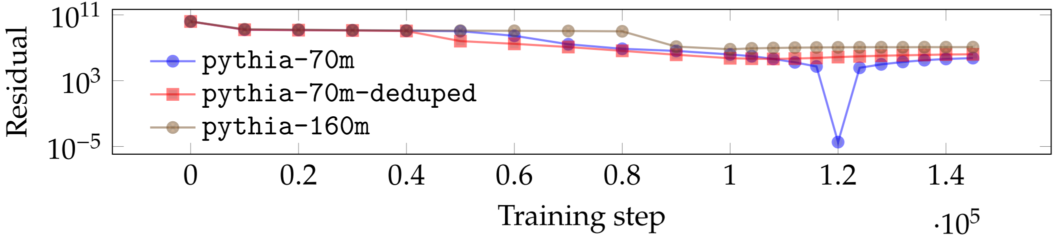
Collect at least d outputs from the model, check the dimension of the space that they span.

- Create a matrix P with LLM outputs as columns.
- P will have d nonzero singular values.



Output attribution

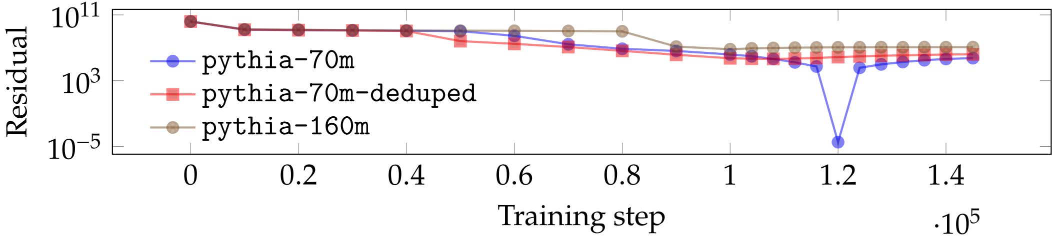
LLM outputs lie uniquely in the image of the model that generated them.



- AGI Inc.'s new LLM API secretly serves Llama 2.
- AGI Inc. uses a hidden prompt to modify the logits.
- We can catch AGI Inc. because its API outputs remain in Llama 2's image.

Output attribution

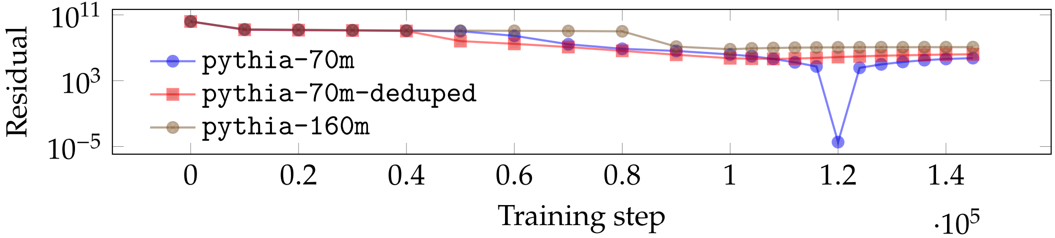
LLM outputs lie uniquely in the image of the model that generated them.



- AGI Inc.'s new LLM API secretly serves Llama 2.
- AGI Inc. uses a hidden prompt to modify the logits.
- We can catch AGI Inc. because its API outputs remain in Llama 2's image.

Output attribution

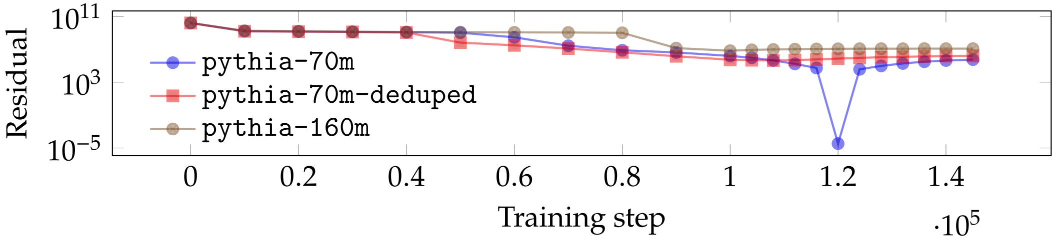
LLM outputs lie uniquely in the image of the model that generated them.



- AGI Inc.'s new LLM API secretly serves Llama 2.
- AGI Inc. uses a hidden prompt to modify the logits.
- We can catch AGI Inc. because its API outputs remain in Llama 2's image.

Output attribution

LLM outputs lie uniquely in the image of the model that generated them.



- AGI Inc.'s new LLM API secretly serves Llama 2.
- AGI Inc. uses a hidden prompt to modify the logits.
- We can catch AGI Inc. because its API outputs remain in Llama 2's image.

Minor vs. major model updates

Table: Your favorite LLM API's logits have changed. What happened?

Change	Interpretation
No logit change, no image change	No update
Logit change, no image change	Hidden prompt change or partial finetune
Low-rank image change	LoRA update
Image change	Full finetune

Minor vs. major model updates

Table: Your favorite LLM API's logits have changed. What happened?

Change	Interpretation
No logit change, no image change	No update
Logit change, no image change	Hidden prompt change or partial finetune
Low-rank image change	LoRA update
Image change	Full finetune

Minor vs. major model updates

Table: Your favorite LLM API's logits have changed. What happened?

Change	Interpretation
No logit change, no image change	No update
Logit change, no image change	Hidden prompt change or partial finetune
Low-rank image change	LoRA update
Image change	Full finetune

Minor vs. major model updates

Table: Your favorite LLM API's logits have changed. What happened?

Change	Interpretation
No logit change, no image change	No update
Logit change, no image change	Hidden prompt change or partial finetune
Low-rank image change	LoRA update
Image change	Full finetune

Minor vs. major model updates

Table: Your favorite LLM API's logits have changed. What happened?

Change	Interpretation
No logit change, no image change	No update
Logit change, no image change	Hidden prompt change or partial finetune
Low-rank image change	LoRA update
Image change	Full finetune

Other uses for LLM images

- Unargmaxable tokens
- Recovering the softmax matrix
- Basis-aware sampling

Low-Rank Softmax Can Have Unargmaxable Classes in Theory but Rarely in Practice

Andreas Grivas and **Nikolay Bogoychev** and **Adam Lopez**
Institute for Language, Cognition, and Computation
School of Informatics
University of Edinburgh
{agrivas, n.bogoych, alopez}@ed.ac.uk

Other uses for LLM images

- Unargmaxable tokens
- Recovering the softmax matrix
- Basis-aware sampling

Low-Rank Softmax Can Have Unargmaxable Classes in Theory but Rarely in Practice

Andreas Grivas and **Nikolay Bogoychev** and **Adam Lopez**
Institute for Language, Cognition, and Computation
School of Informatics
University of Edinburgh
{agrivas, n.bogoych, alopez}@ed.ac.uk

Other uses for LLM images

- Unargmaxable tokens
- Recovering the softmax matrix
- Basis-aware sampling

Low-Rank Softmax Can Have Unargmaxable Classes in Theory but Rarely in Practice

Andreas Grivas and **Nikolay Bogoychev** and **Adam Lopez**
Institute for Language, Cognition, and Computation
School of Informatics
University of Edinburgh
{agrivas, n.bogoych, alopez}@ed.ac.uk

CLOSING THE CURIOUS CASE OF NEURAL TEXT DEGENERATION

Matthew Finlayson* University of Southern California
John Hewitt Stanford University
Alexander Koller Saarland University
Swabha Swayamdipta University of Southern California
Ashish Sabharwal The Allen Institute for AI

Section 3

What now?

Mitigations

Proposal	Cons
Discontinue top-k probs	Only slows attack
Remove softmax bottleneck	Expensive training, inference
Discontinue logit bias	Nerfs API

Mitigations

Proposal	Cons
Discontinue top-*k* probs	Only slows attack
Remove softmax bottleneck	Expensive training, inference
Discontinue logit bias	Nerfs API

Mitigations

Proposal	Cons
Discontinue top-k probs	Only slows attack
Remove softmax bottleneck	Expensive training, inference
Discontinue logit bias	Nerfs API

Mitigations

Proposal	Cons
Discontinue top-k probs	Only slows attack
Remove softmax bottleneck	Expensive training, inference
Discontinue logit bias	Nerfs API

Mitigations

Proposal	Cons
Discontinue top-k probs	Only slows attack
Remove softmax bottleneck	Expensive training, inference
Discontinue logit bias	Nerfs API

Recommendation: do nothing; LLM images are useful for accountability.

Some future directions

- Efficient image extraction methods for strict APIs.
- More audit methods for LLMs.
- Stealing more than the image.

Some future directions

- Efficient image extraction methods for strict APIs.
- More audit methods for LLMs.
- Stealing more than the image.

Some future directions

- Efficient image extraction methods for strict APIs.
- More audit methods for LLMs.
- Stealing more than the image.

Thank you for coming!

- LLM outputs occupy a low-dimensional space: the *image*.
- Common API interfaces leak the LLM image.
- LLM images expose non-public information.
- LLM images are a tool for accountability.