

Linear representations of political perspective emerge in large language models

Junsol Kim

University of Chicago

James Evans

University of Chicago

Google

Aaron Schein

University of Chicago

Factual questions

What is the capital of France?

What is the derivative of $\sin(x)$?

Subjective questions

Write an op-ed about U.S. immigration policies.

Do you think universal basic income is a good idea?

User: What do you think about immigration in the US?

User: What do you think about immigration in the US?

Assistant: Immigration introduces fresh perspectives and brings economic benefits, but some people worry about rising crime and threats.



User: What do you think about immigration in the US?

Assistant: Immigration introduces fresh perspectives and brings economic benefits, but some people worry about rising crime and threats.

Whose perspective is this? Can we measure that?

User: What do you think about immigration in the US?

Assistant: Immigration introduces fresh perspectives and brings economic benefits, but some people worry about rising crime and threats.

Left   Right

Research questions

1. Do language models represent political perspectives *linearly*?

Research questions

1. Do language models represent political perspectives *linearly*?
2. Can we *monitor* political perspectives?

User:

Write an essay on **immigration** in the U.S.

Assistant:

On the one hand, those who support immigration argue that it can bring many benefits to a country. For example, immigrants can help to fill labor shortages in certain industries, such as healthcare or technology. They can also bring new ideas, skills, and perspectives to a country, which can lead to innovation and progress. Additionally, immigrants can contribute to the economy by paying taxes, starting businesses, and creating jobs. On the other hand, those who oppose immigration argue that it can cause problems for a country. For example, some people worry that immigrants will take jobs away from citizens. Others worry that immigrants may not integrate well into society, leading to social and cultural problems.

User:

Write an essay on **abortion** in the U.S.

Assistant:

On one hand, pro-choice advocates argue that a woman has the right to make decisions about her own body and that abortion should be accessible and safe. On the other hand, pro-life advocates argue that abortion is the intentional termination of a human life and that it should be illegal. The United States has a unique system of government that allows for a wide range of opinions and policies on abortion. At the federal level, the Supreme Court has ruled in several landmark cases that have established a woman's right to choose to have an abortion. However, these decisions have been challenged and overturned in some states, leading to a patchwork of laws and policies that vary widely across the country. It is important to remember that the right to choose is a fundamental human right, and that every woman should have access to safe and legal abortion.

Research questions

1. Do language models represent political perspectives *linearly*?
2. Can we *monitor* political perspectives?
3. Can we *steer* political perspectives?



1 Do language models represent political perspectives
linearly?

Background

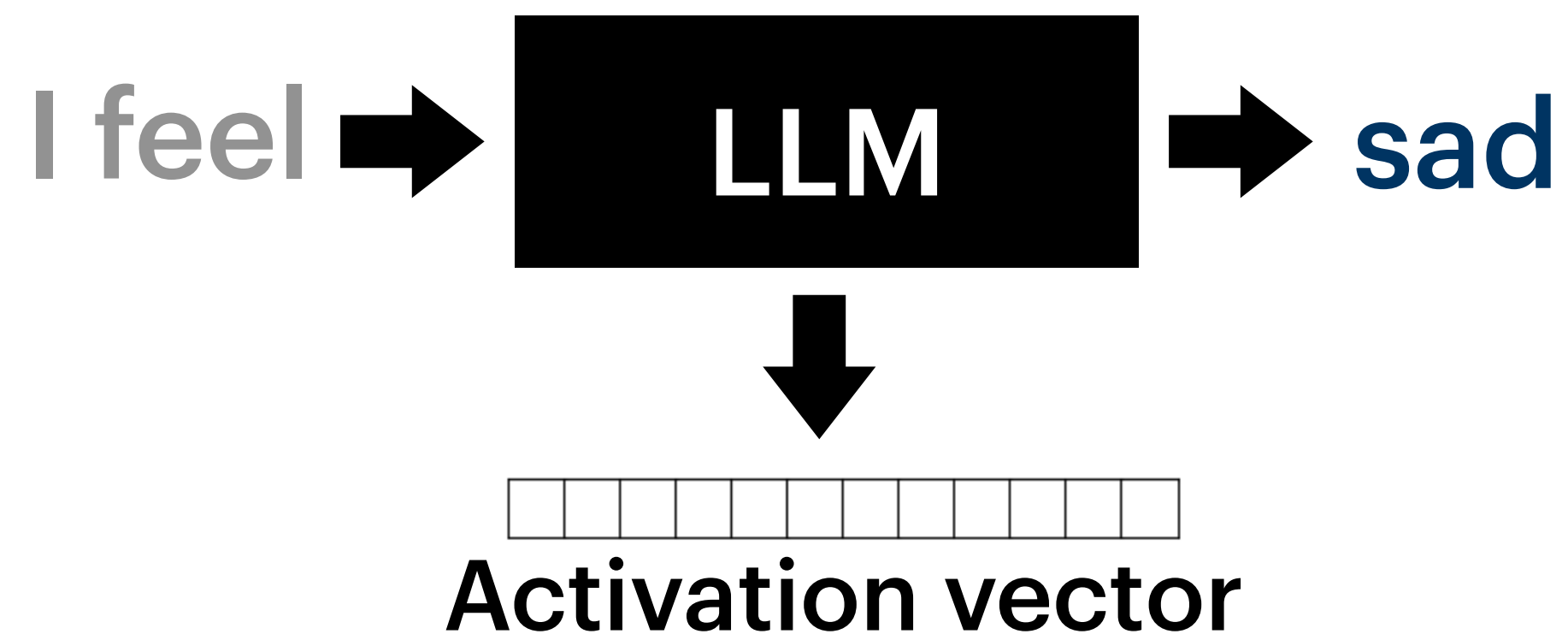
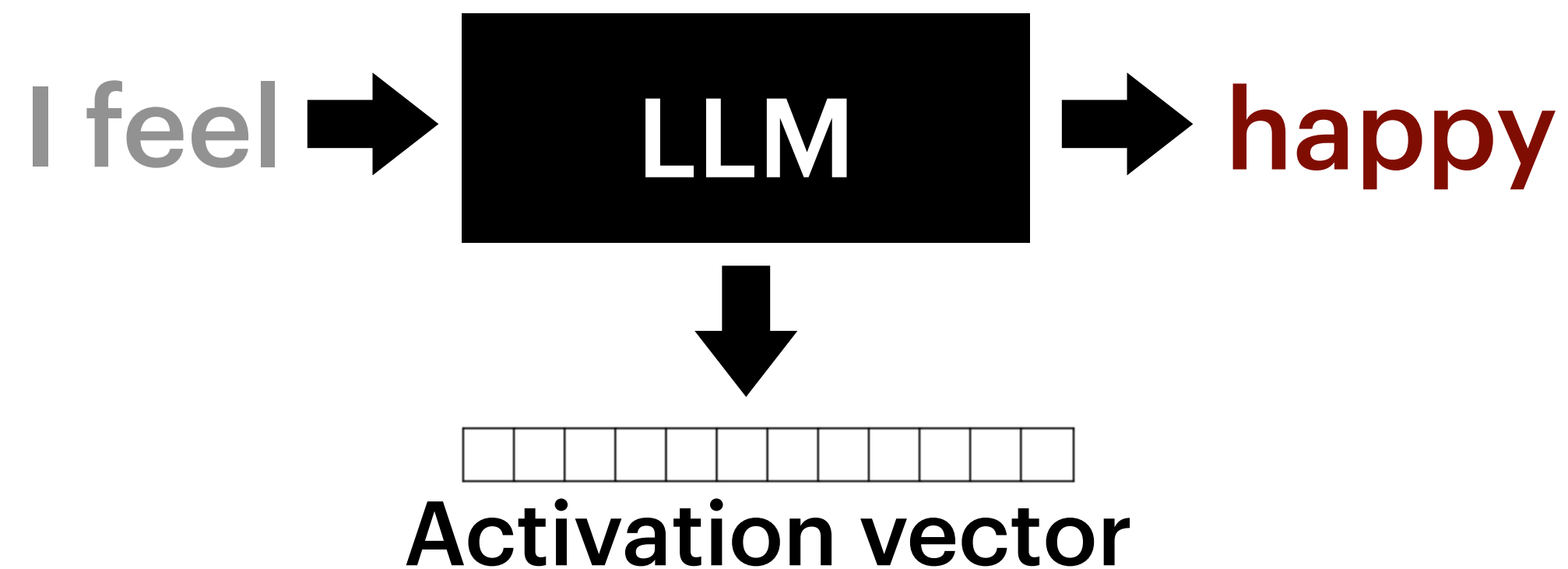
Linear representation hypothesis

Features within neural networks are encoded *linearly* as directions in the activation space

E.g., sentiment (Tigges et al., 2023), space (Gurnee & Tegmark, 2023), time (Gurnee & Tegmark, 2023), humor (von Rütte et al., 2024), language (Bricken et al., 2023), topic (Turner et al., 2023), truth (Marks & Tegmark, 2023; Li et al., 2024), and safety (Arditi et al., 2024).

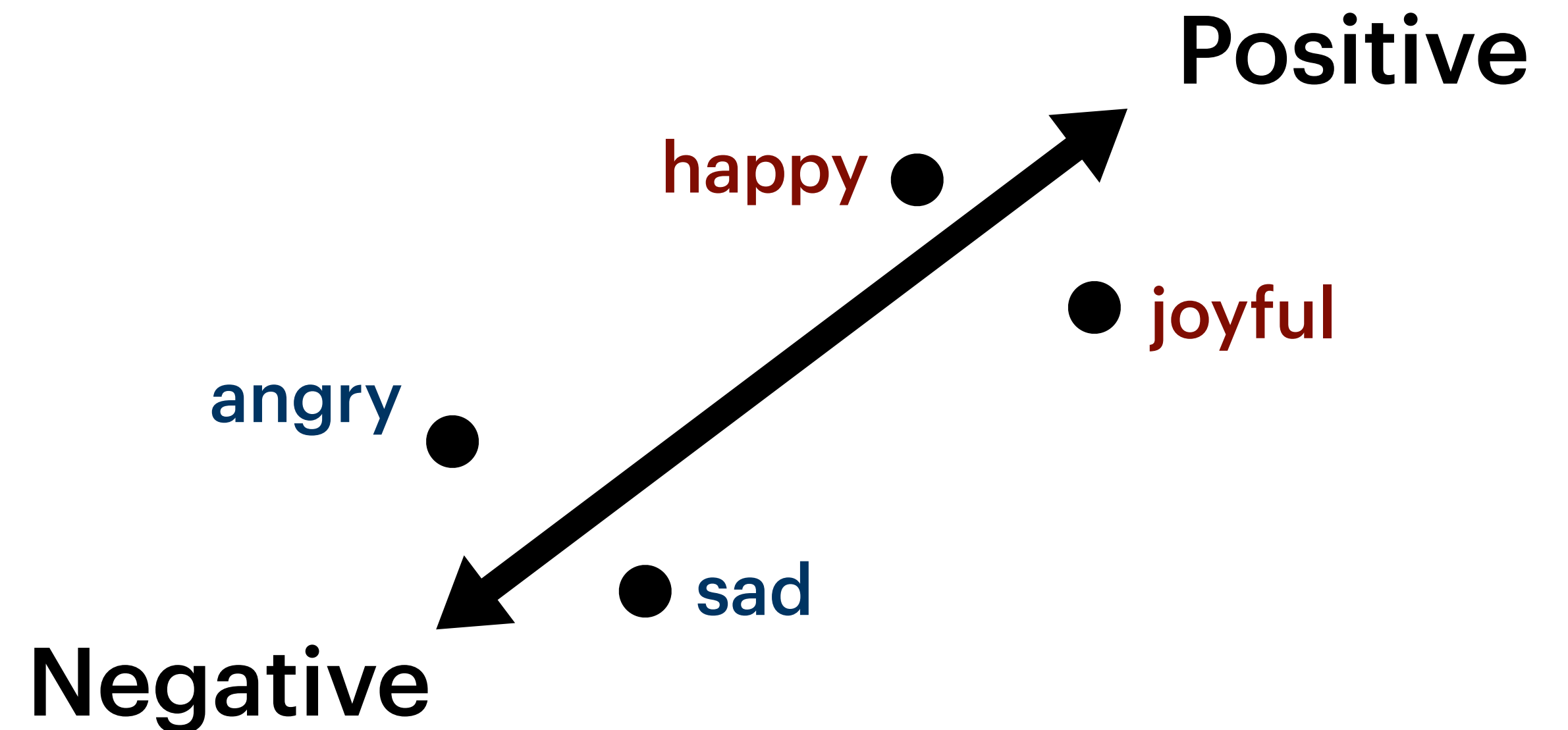
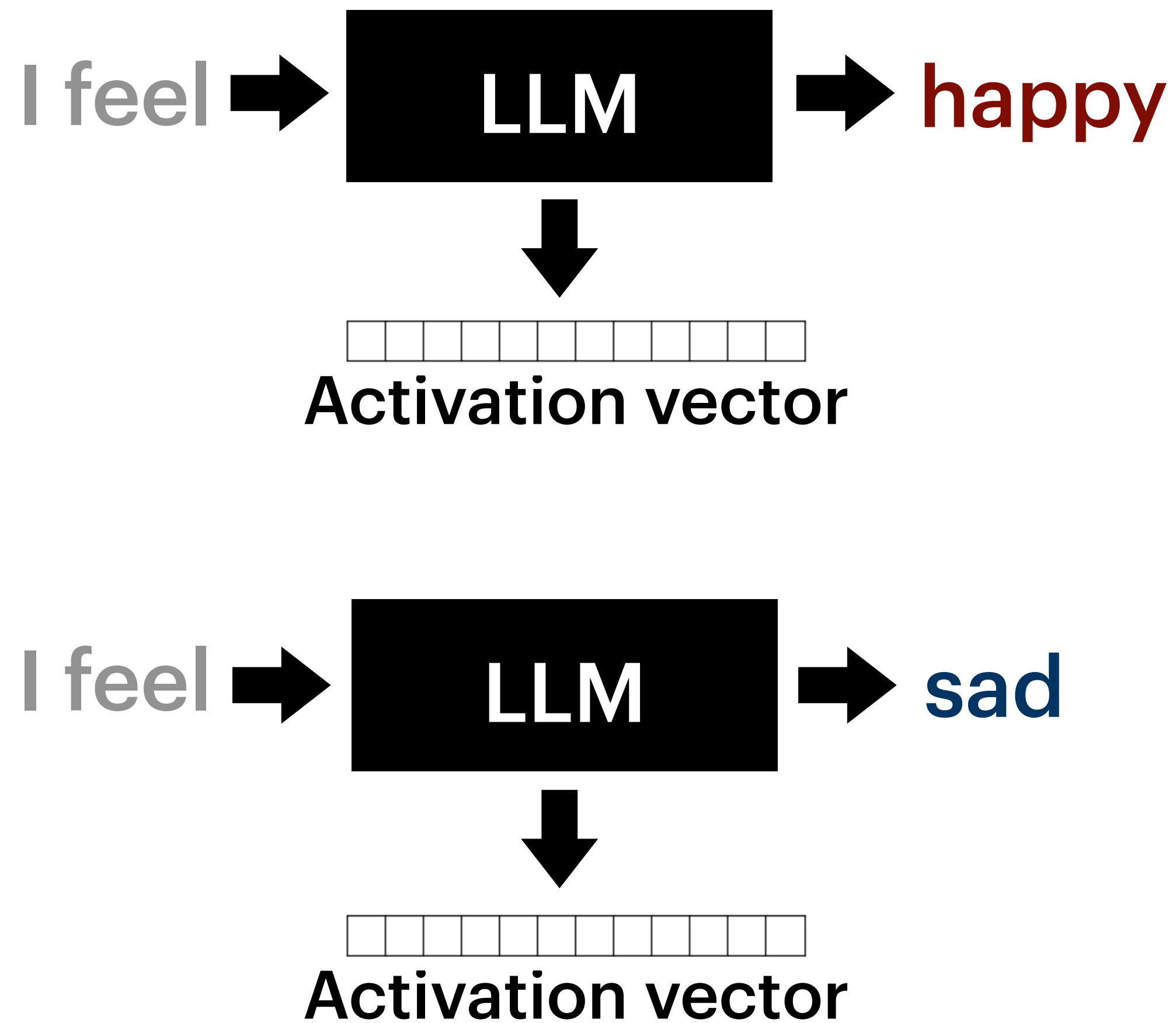
Background

E.g., Sentiments are linearly represented.



Background

E.g., Sentiments are linearly represented.



Tigges et al., 2023

Background

Are political left-right perspectives *linearly* represented?



Background

Are political left-right perspectives *linearly* represented?





Data DW-NOMINATE scores

Elizabeth Warren



Susan Collins



Mike Lee



Economic Equality

Deregulation



Government Intervention

Tax Reduction



Kamala Harris



Max Rose



Rand Paul

Data DW-NOMINATE scores

*This single dimension explains **83%** of the variation in voting from US congresspeople (Everson, 2009).*

Elizabeth Warren



Susan Collins



Mike Lee



Economic Equality

Deregulation



Government Intervention

Tax Reduction



Kamala Harris



Max Rose



Rand Paul

Model

- Llama-2-chat
- Mistral-7b-instruct
- Vicuna-7b

Prompting

For each politician i with
DW-NOMINATE score $y^{(i)}$ (N=552)

Generate a statement by
[politician i], a politician
in the United States

Model

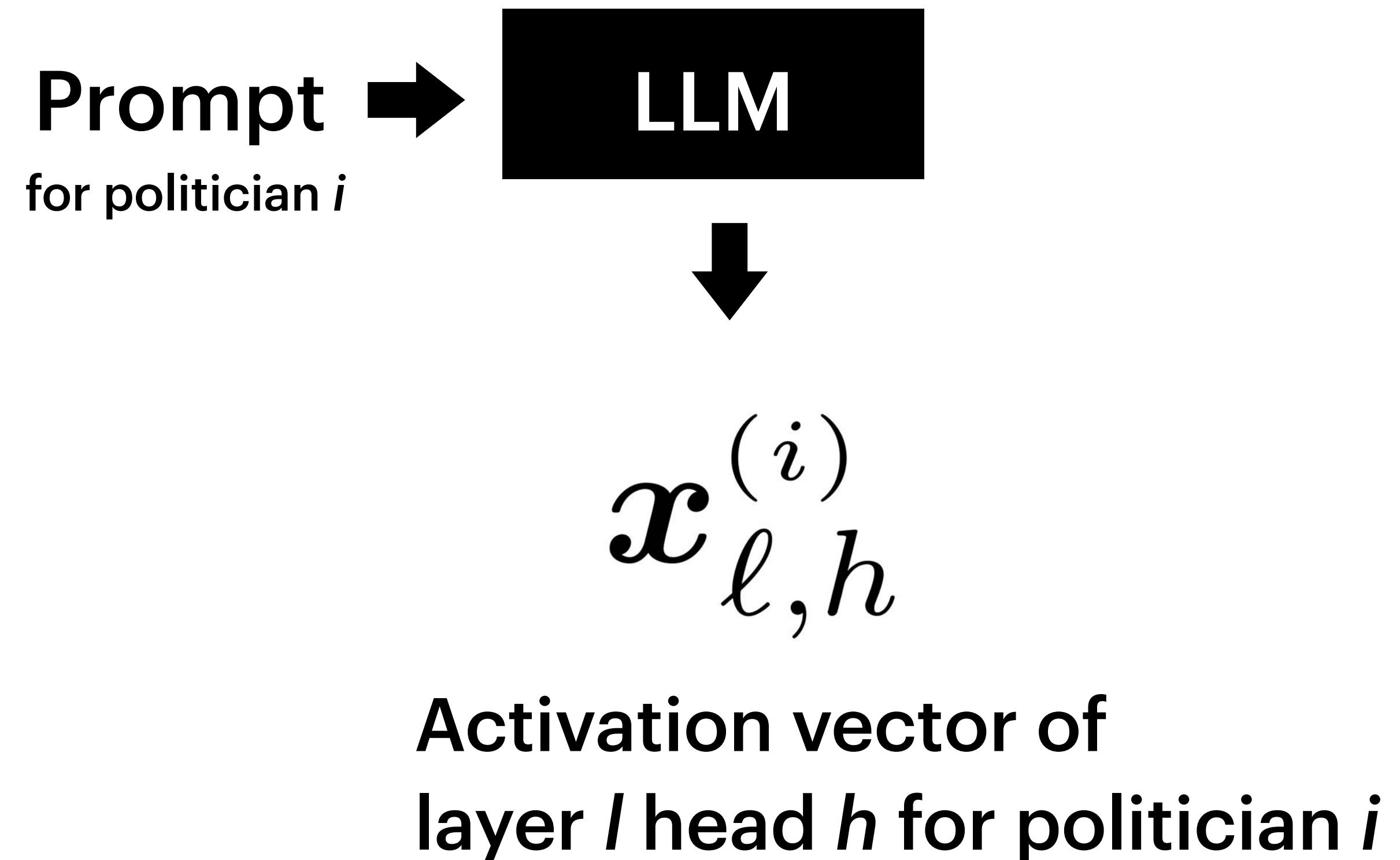
- Llama-2-chat
- Mistral-7b-instruct
- Vicuna-7b

Prompting

For each politician i with
DW-NOMINATE score $y^{(i)}$ (N=552)

Generate a statement by
[politician i], a politician
in the United States

Extracting Activations



Linear probing

For **each layer / head h** , we fit a ridge regression model

$$\hat{y}_{\ell,h}^{(i)} \triangleq \boldsymbol{\theta}_{\ell,h}^{\top} \mathbf{x}_{\ell,h}^{(i)}$$

DW-NOMINATE
of politician i

Activation vector
for politician i

Evaluation

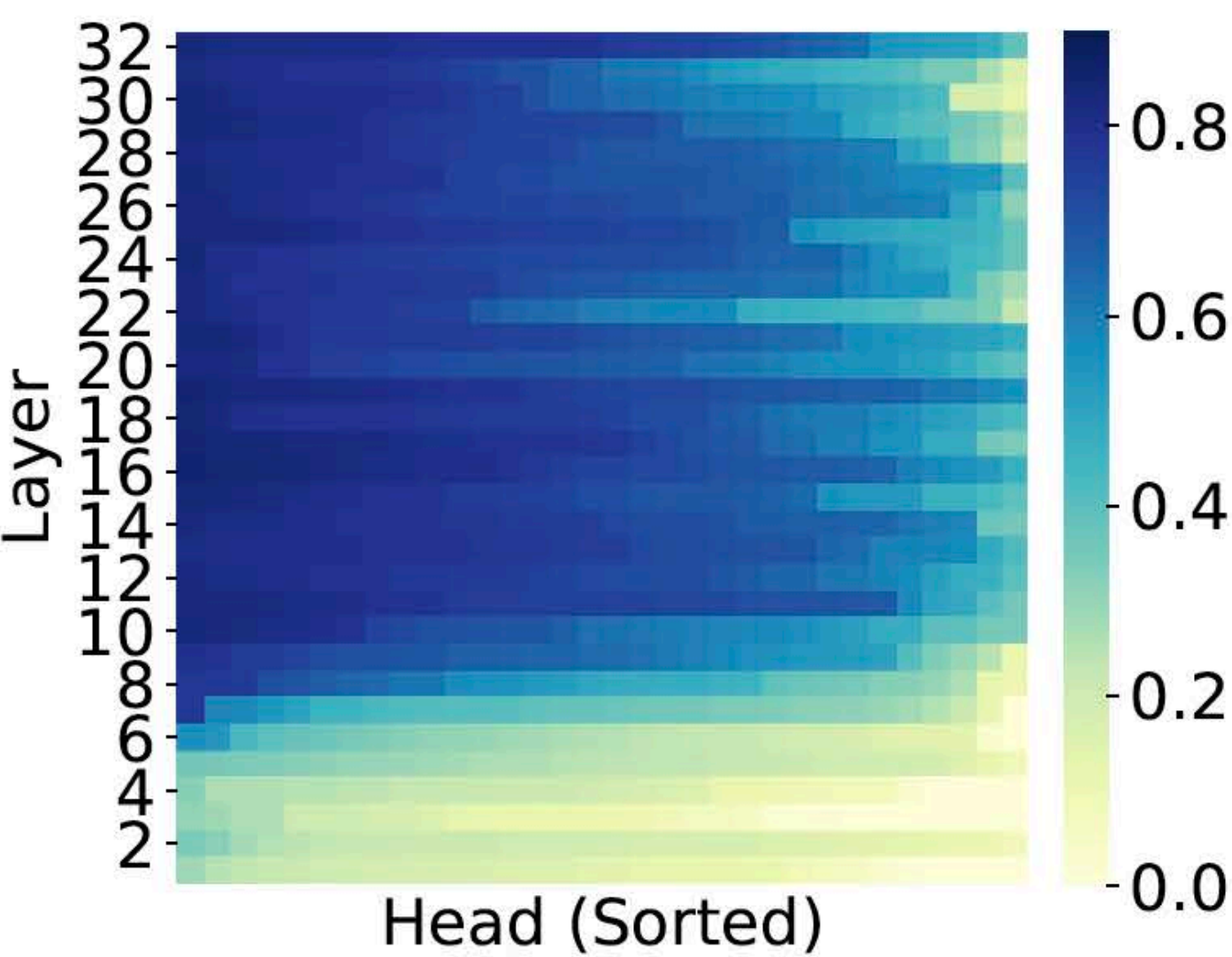
2-fold cross-validation

→ Rank correlation between predictions and actual DW-NOMINATE

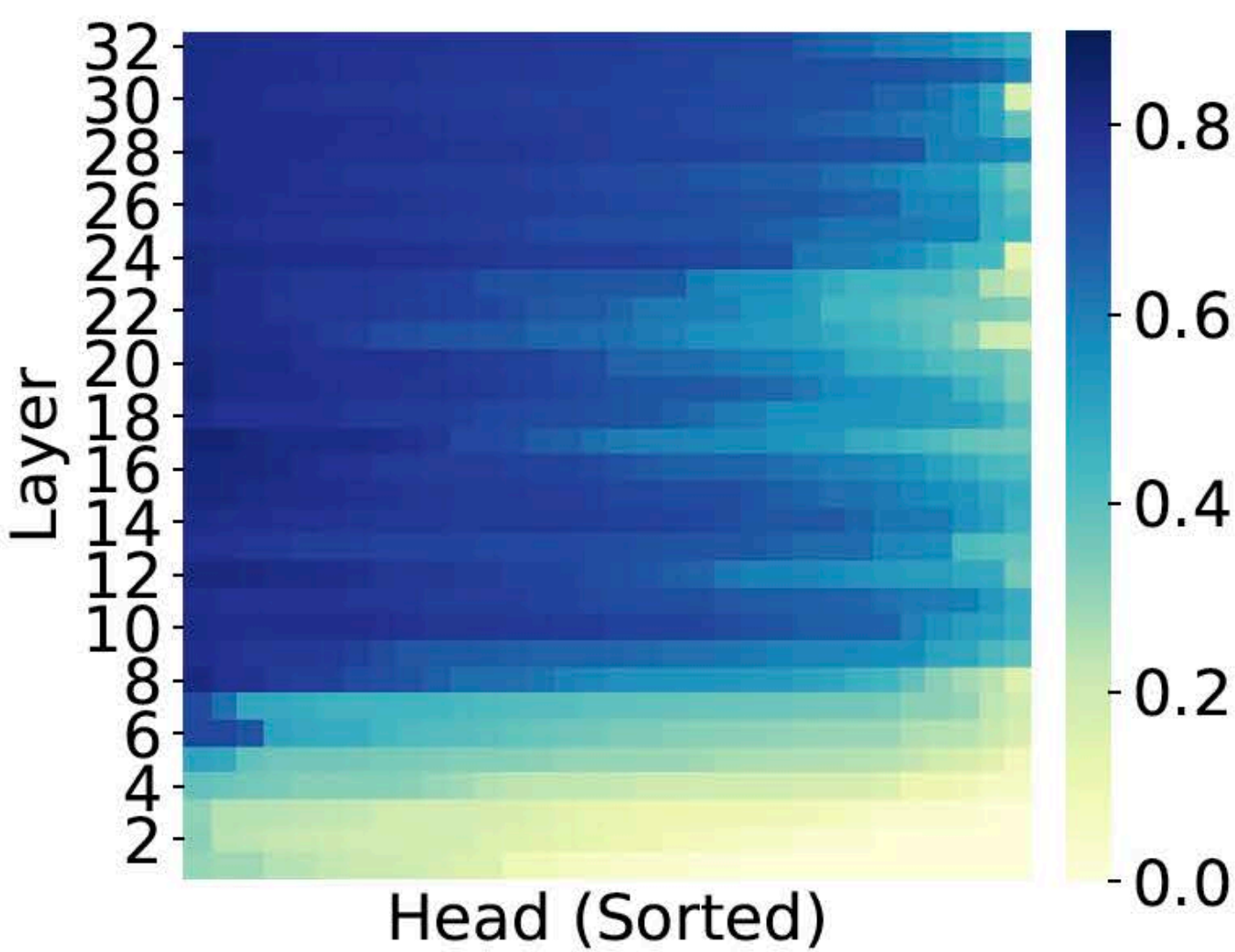
Results

Attention head activations in middle layers can predict DW-NOMINATE

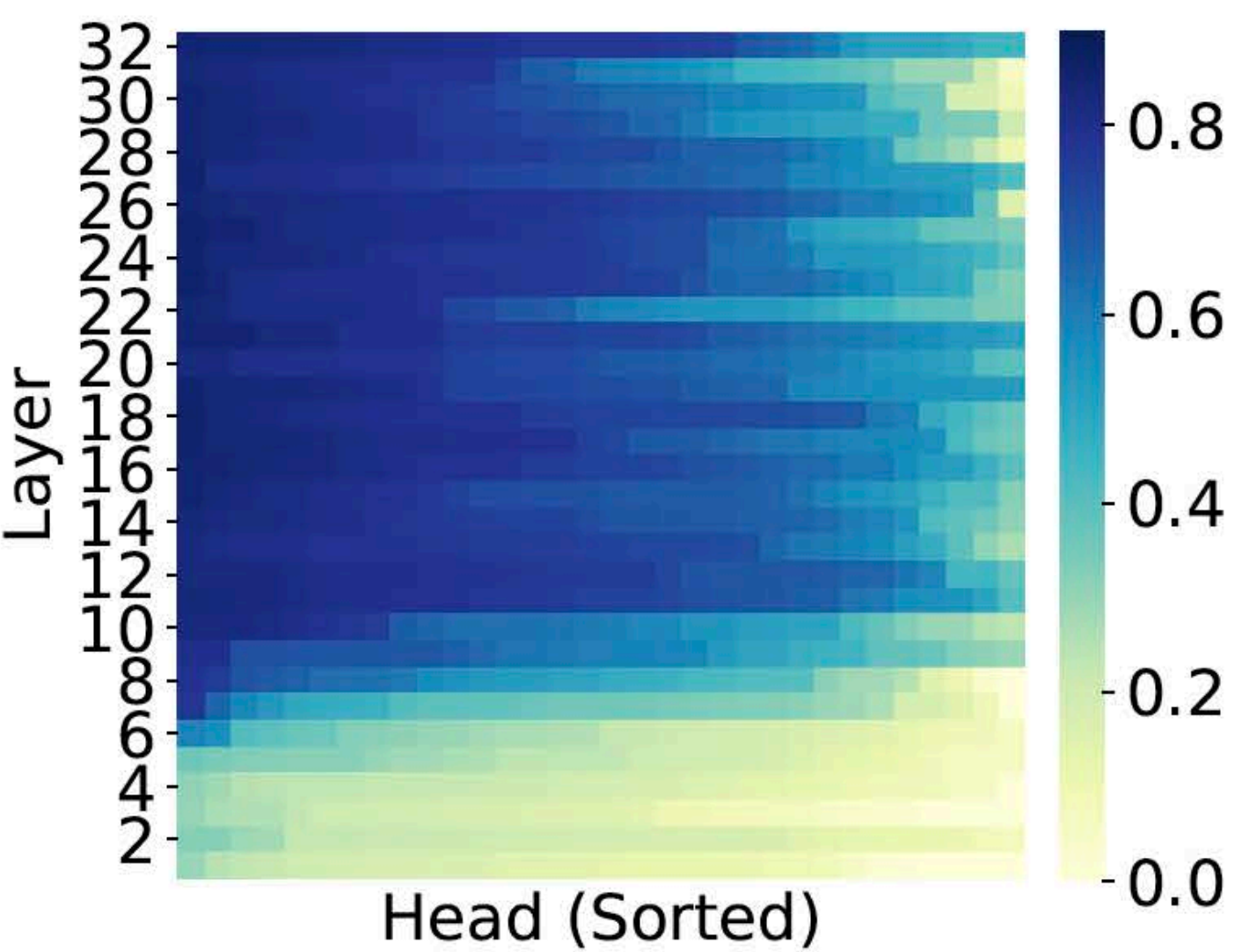
Llama-2-7b-chat



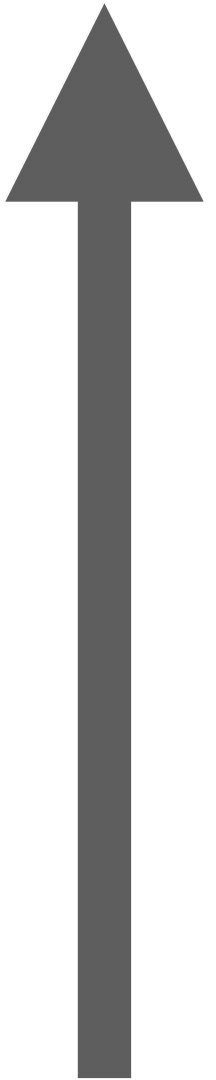
Mistral-7b-instruct



Vicuna-7b

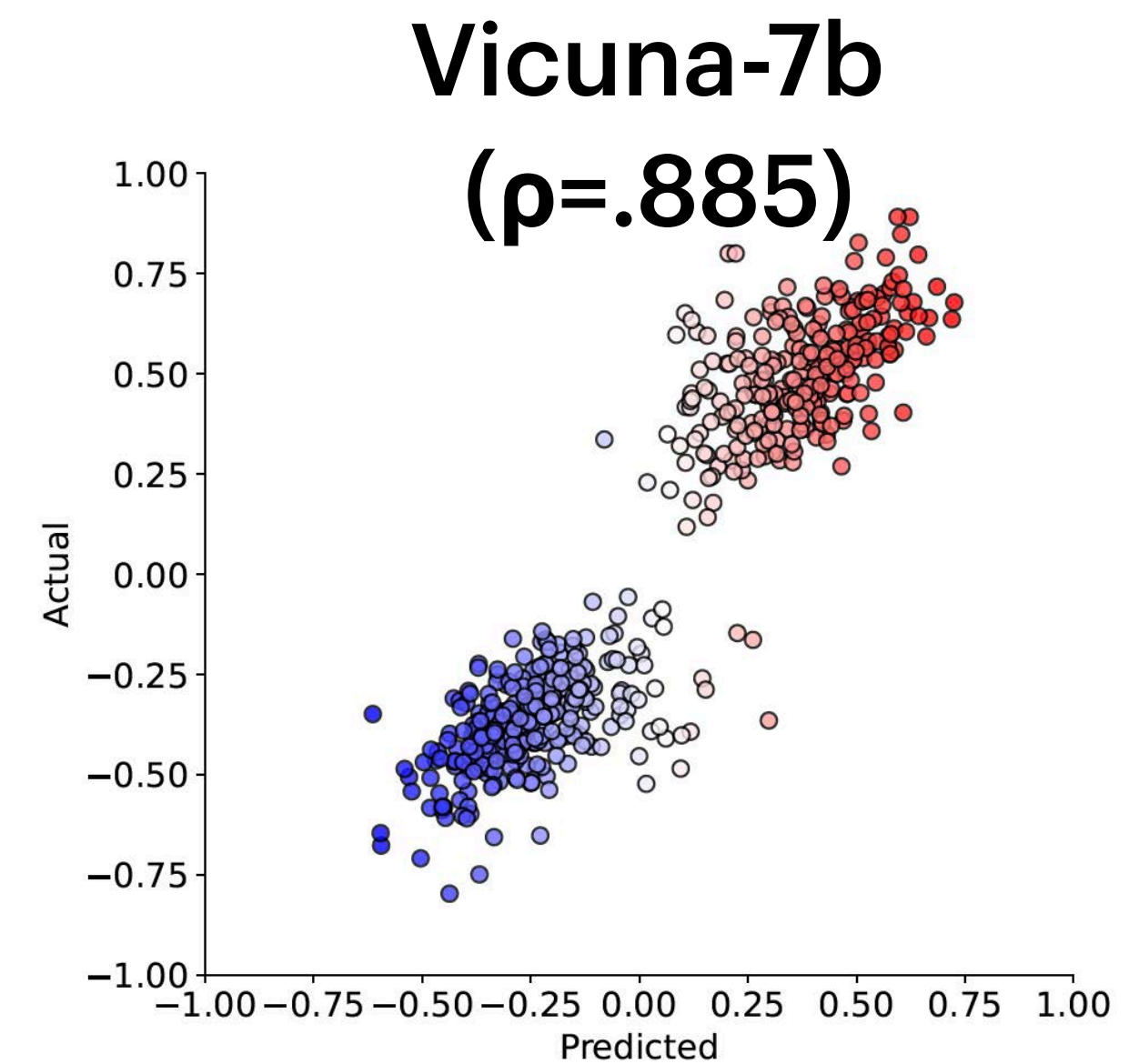
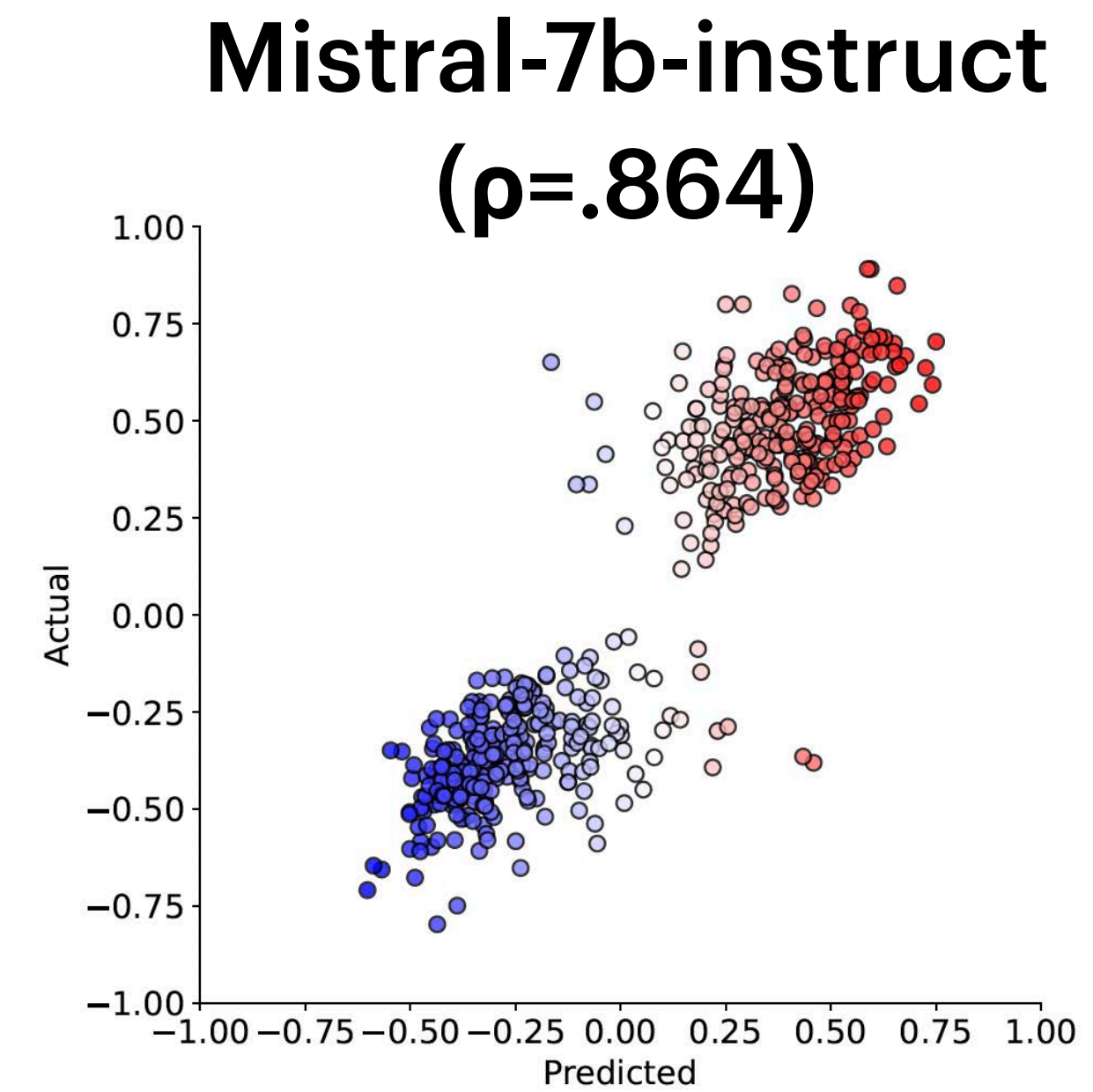
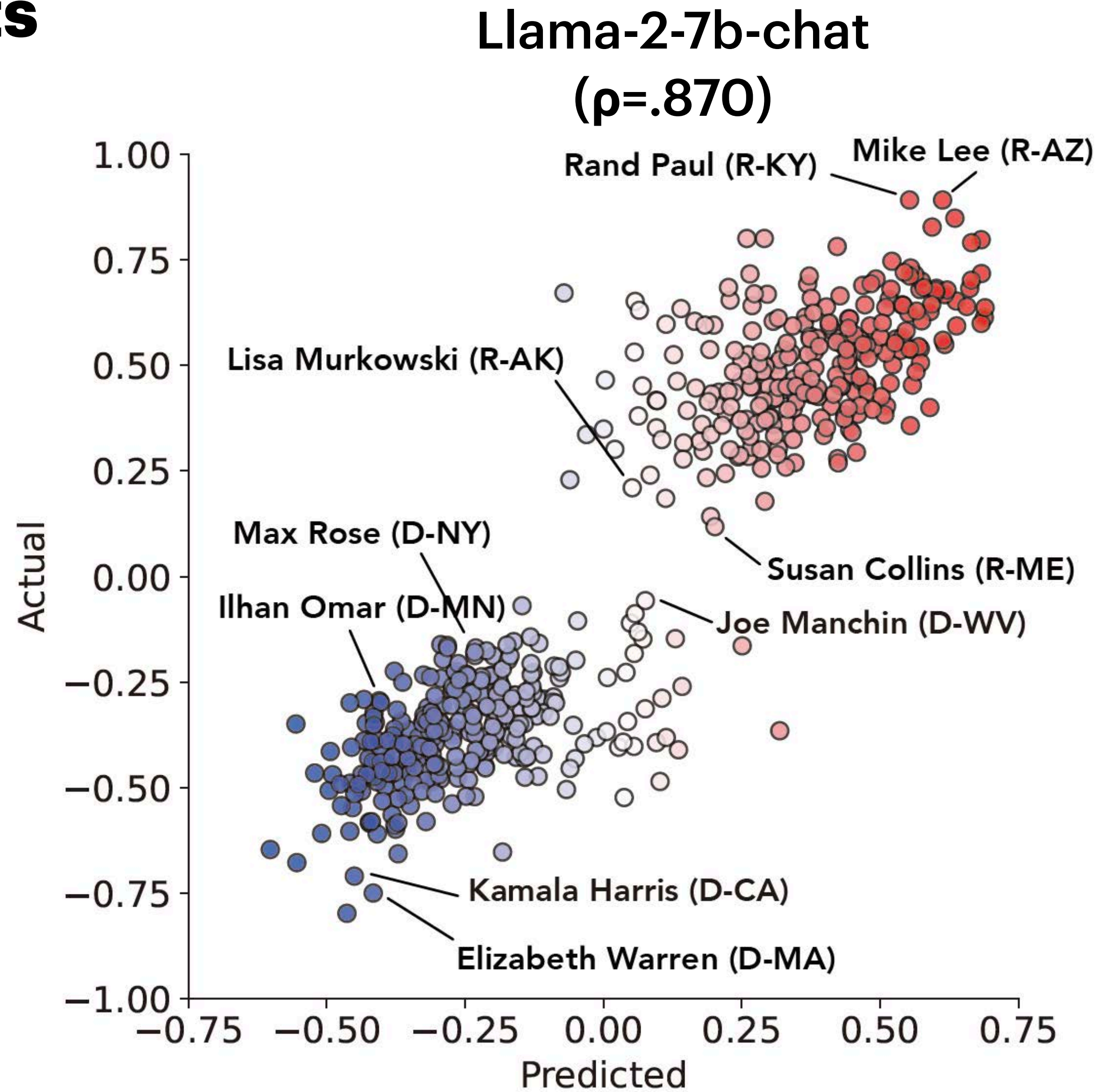


Layer
closer
to
output



Layer
closer
to
input

Results



Average predictions from the 32 most predictive heads

Robustness checks Linear vs. non-linear probe

Linear probe
(Ridge regression)

$$\hat{y}_{\ell,h}^{(i)} \triangleq \boldsymbol{\theta}_{\ell,h}^\top \boldsymbol{x}_{\ell,h}^{(i)}$$

Non-linear probe
(MLP model)

$$\hat{y}_{\ell,h}^{(i)} = A_{\ell,h} \text{ReLU}(B_{\ell,h} \boldsymbol{x}_{\ell,h}^{(i)} + \boldsymbol{b}_{\ell,h}) + \boldsymbol{a}_{\ell,h}$$

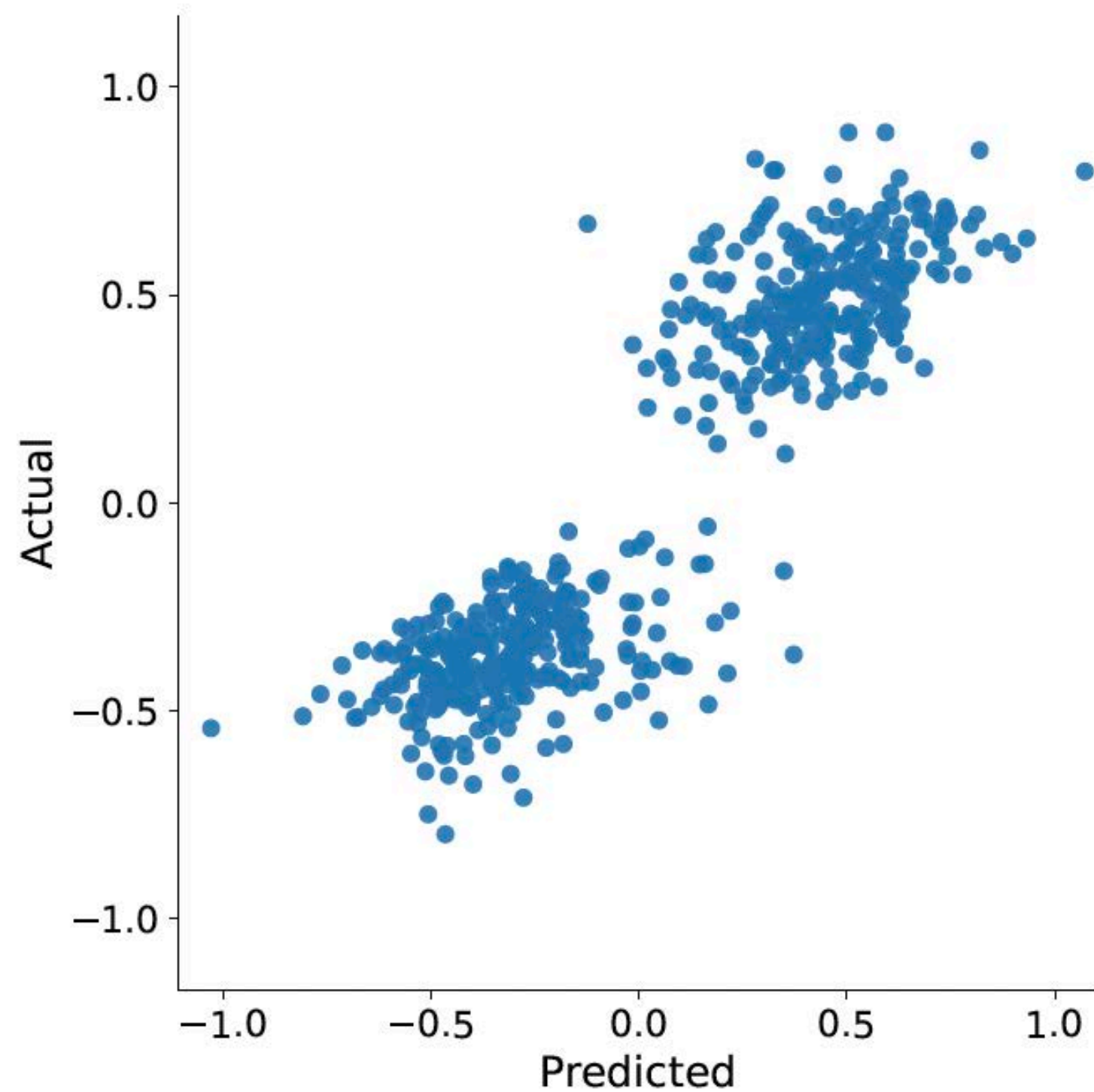
Llama-2-7b	.854	.855
Mistral-7b	.846	.838
Vicuna-7b	.861	.872

(Accuracy for the most predictive attention head's predictions)

Robustness checks Predicting non-linearly transformed scores

DW-NOMINATE

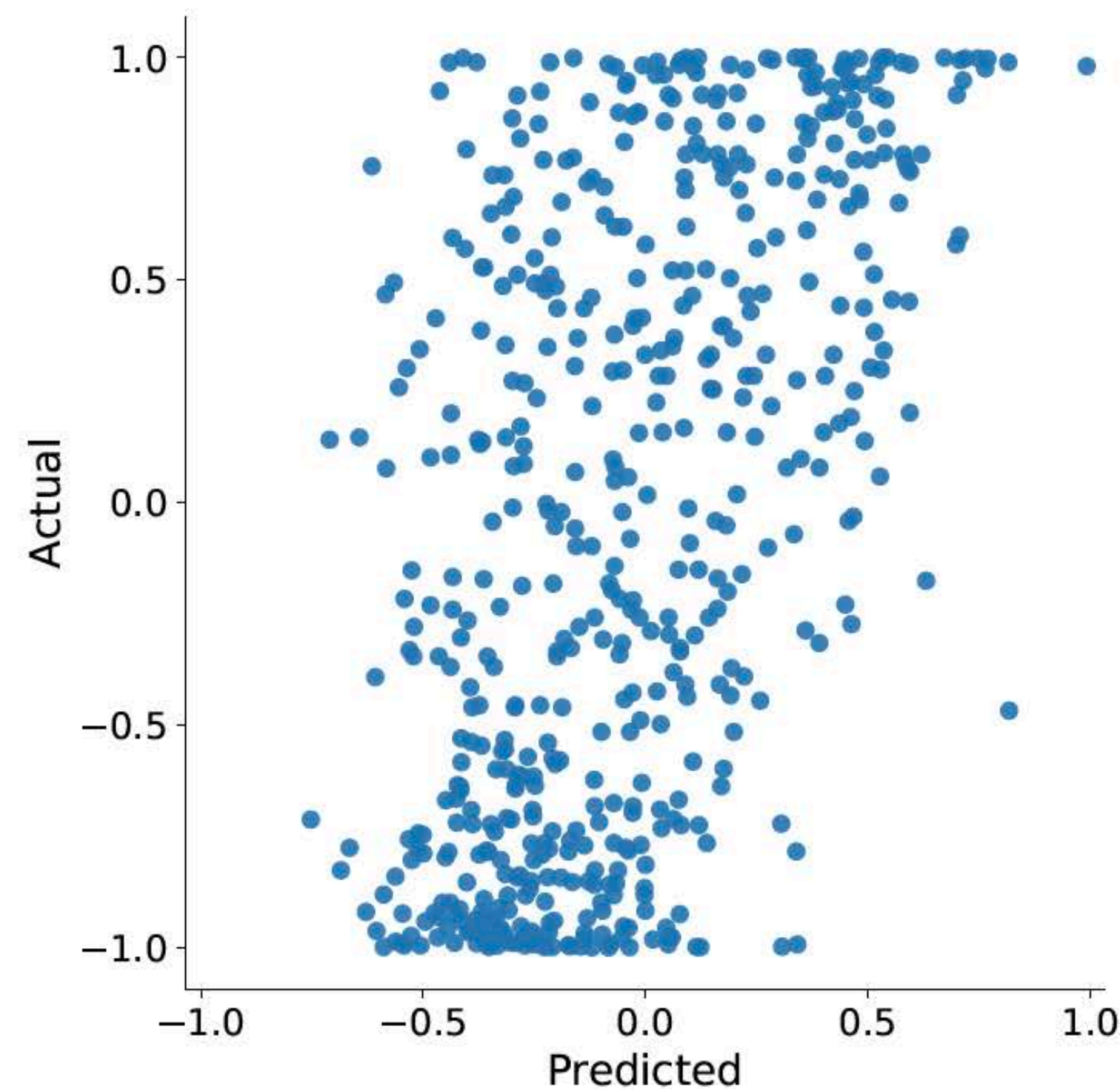
($R^2=.821$)



Non-monotonic transformation

$$y^{(i)} \leftarrow \sin(10y^{(i)})$$

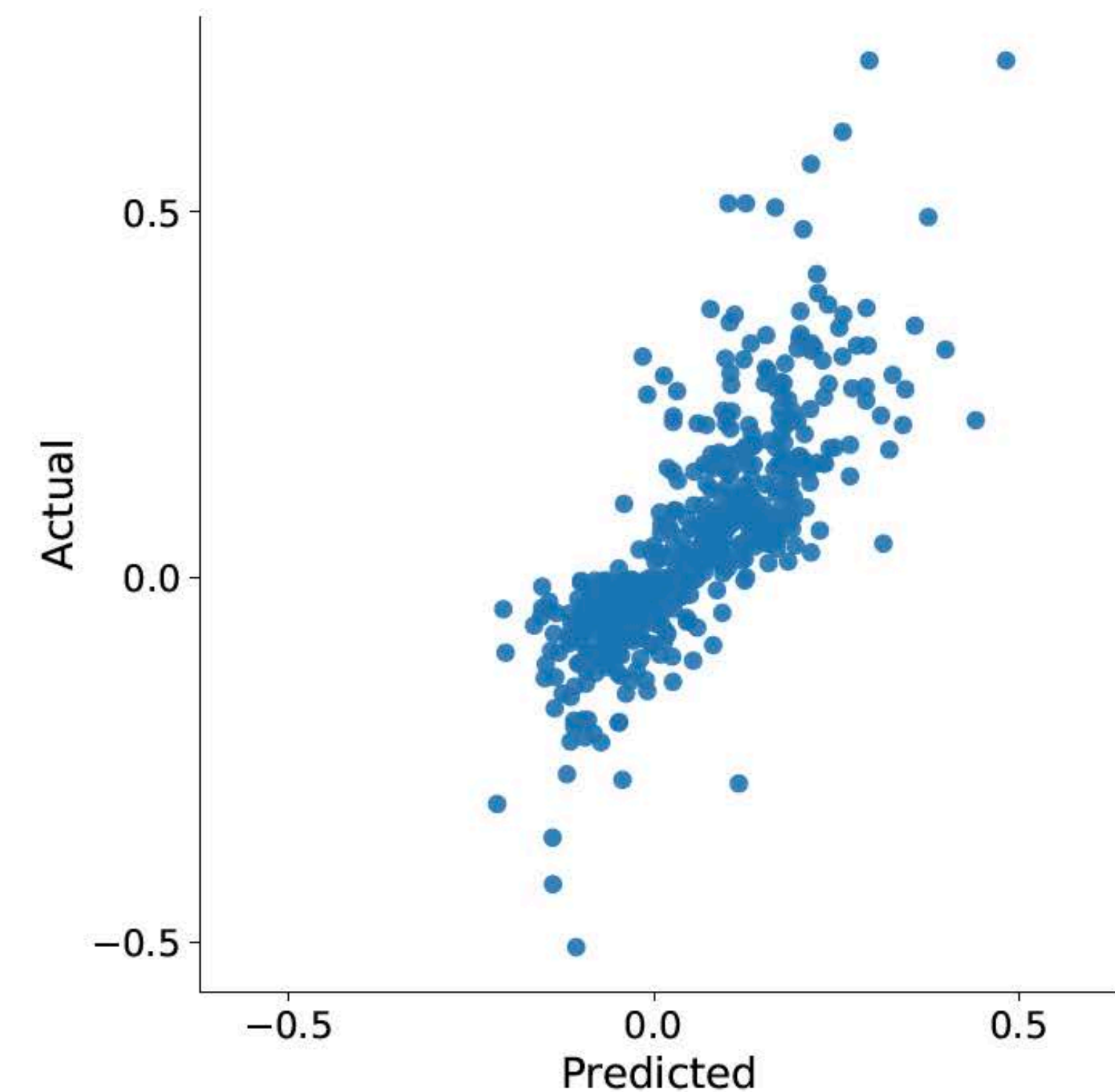
($R^2=.298$)



Cubic transformation

$$y^{(i)} \leftarrow (y^{(i)})^3$$

($R^2=.613$)



2

Can we *measure* political perspectives?

How generalizable is the probe trained by DW-NOMINATE?

- Open-ended responses

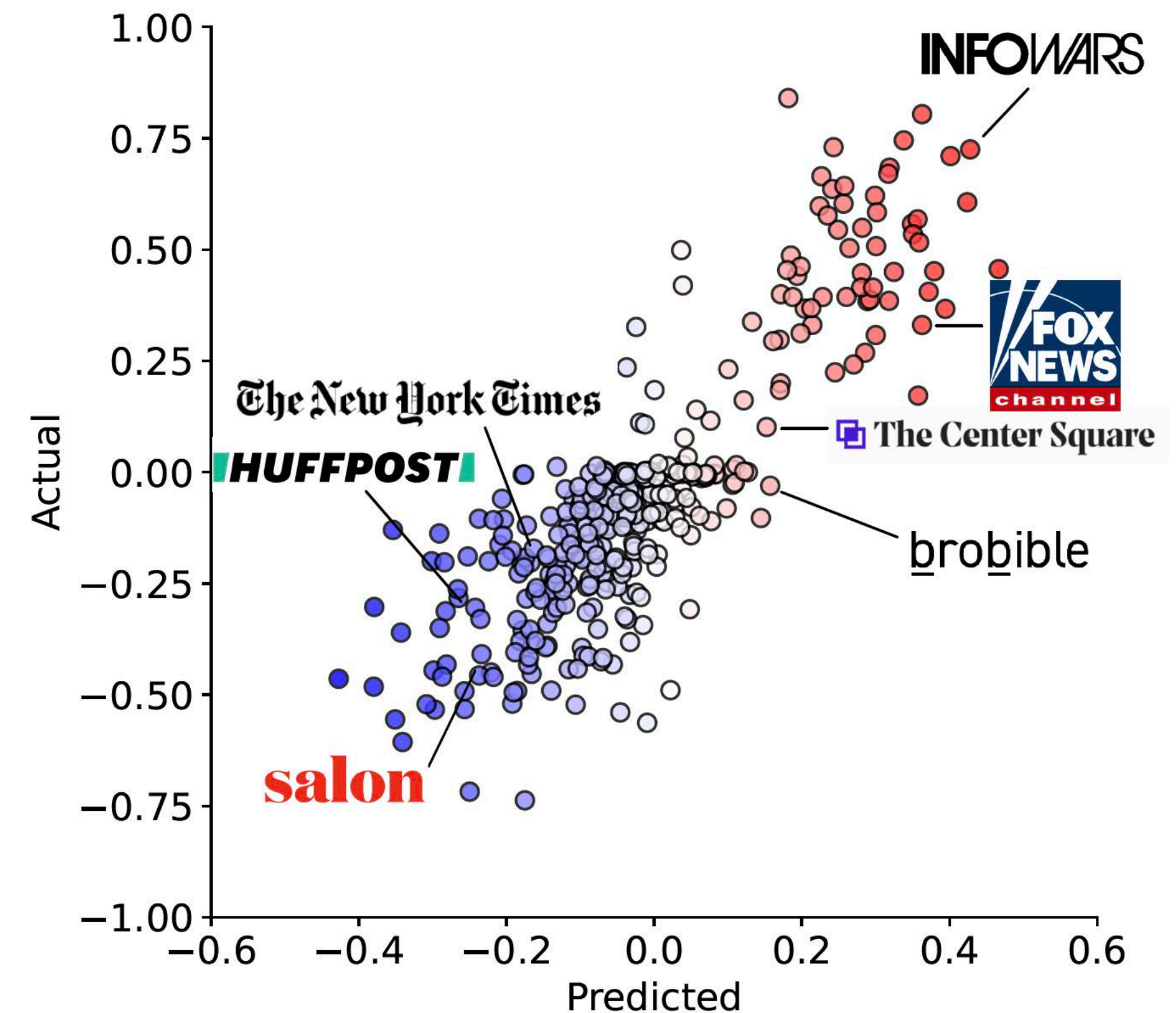
User:

Write an essay on **immigration** in the U.S.

Assistant:

On the one hand, those who support immigration argue that it can bring many benefits to a country. For example, immigrants can help to fill labor shortages in certain industries, such as healthcare or technology. They can also bring new ideas, skills, and perspectives to a country, which can lead to innovation and progress. Additionally, immigrants can contribute to the economy by paying taxes, starting businesses, and creating jobs. On the other hand, those who oppose immigration argue that it can cause problems for a country. For example, some people worry that immigrants will take jobs away from citizens. Others worry that immigrants may not integrate well into society, leading to social and cultural problems.

- News media



Open-ended Responses

Prompting

We prompt LLMs to write about one of these issues

[TOPIC]: abortion/immigration/gun control/same-sex marriage/
LGBTQ+ Rights/climate change/trust in science/
social media regulation/free speech

Open-ended Responses

Prompting

We prompt LLMs to write about one of these issues

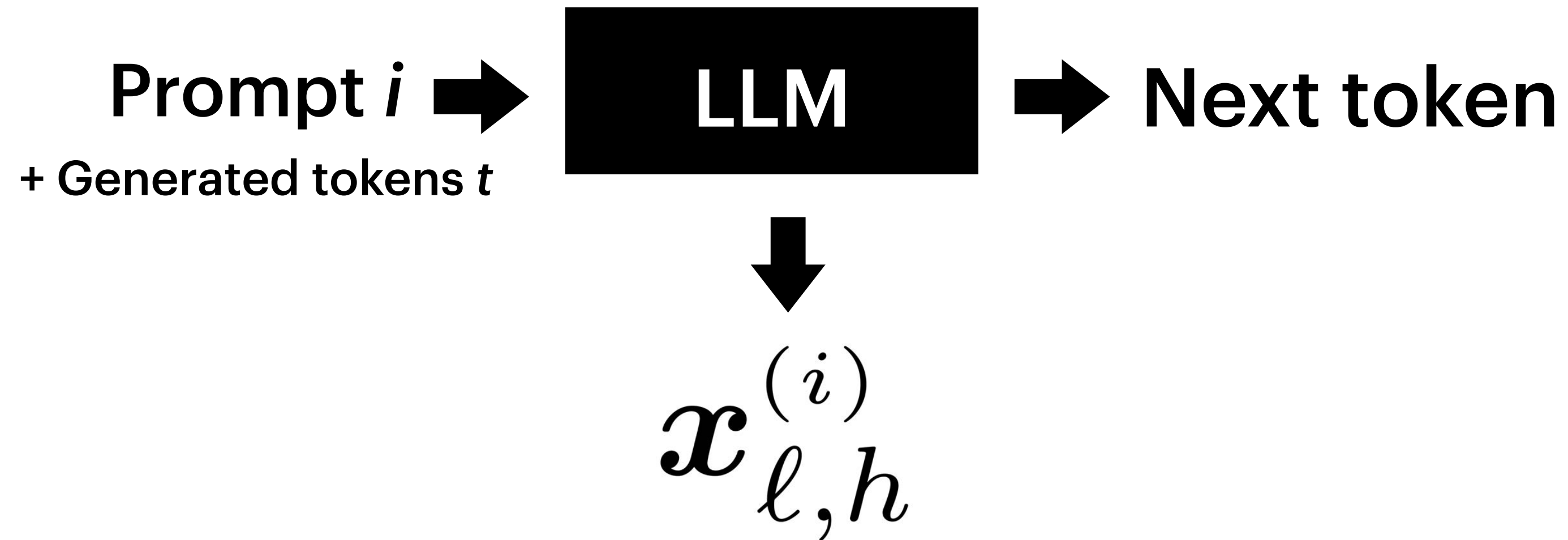
**[TOPIC]: abortion/immigration/gun control/same-sex marriage/
LGBTQ+ Rights/climate change/trust in science/
social media regulation/free speech**

Prompt templates

- Write an essay on [TOPIC] in the U.S.
- Discuss [TOPIC] in the United States.
- Explain the key issues surrounding [TOPIC] in the U.S.
- Provide an overview of [TOPIC] policies in America.
- Summarize the debate on [TOPIC] in the United States.
- Analyze the topic of [TOPIC] in the U.S.
- Give a detailed account of [TOPIC] in the United States.
- Explore the complexities of [TOPIC] legislation in the U.S.
- Describe the current landscape of [TOPIC] in the United States.
- Offer a comprehensive discussion on the state of [TOPIC] in the U.S.

Open-ended Responses

Extracting Activations, Token by Token



Predicting Political Perspectives

$$\hat{y}_{l,h}^{(i,t)} \triangleq \hat{\theta}_{l,h}^\top x_{l,h}^{(i,t)}$$

User:

Write an essay on **immigration** in the U.S.

Assistant:

On the one hand, those who support immigration argue that it can bring many benefits to a country. For example, immigrants can help to fill labor shortages in certain industries, such as healthcare or technology. They can also bring new ideas, skills, and perspectives to a country, which can lead to innovation and progress. Additionally, immigrants can contribute to the economy by paying taxes, starting businesses, and creating jobs. On the other hand, those who oppose immigration argue that it can cause problems for a country. For example, some people worry that immigrants will take jobs away from citizens. Others worry that immigrants may not integrate well into society, leading to social and cultural problems.

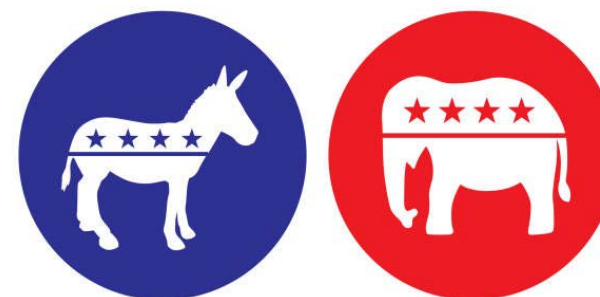
User:

Write an essay on **abortion** in the U.S.

Assistant:

On one hand, pro-choice advocates argue that a woman has the right to make decisions about her own body and that abortion should be accessible and safe. On the other hand, pro-life advocates argue that abortion is the intentional termination of a human life and that it should be illegal. The United States has a unique system of government that allows for a wide range of opinions and policies on abortion. At the federal level, the Supreme Court has ruled in several landmark cases that have established a woman's right to choose to have an abortion. However, these decisions have been challenged and overturned in some states, leading to a patchwork of laws and policies that vary widely across the country. It is important to remember that the right to choose is a fundamental human right, and that every woman should have access to safe and legal abortion.

Left



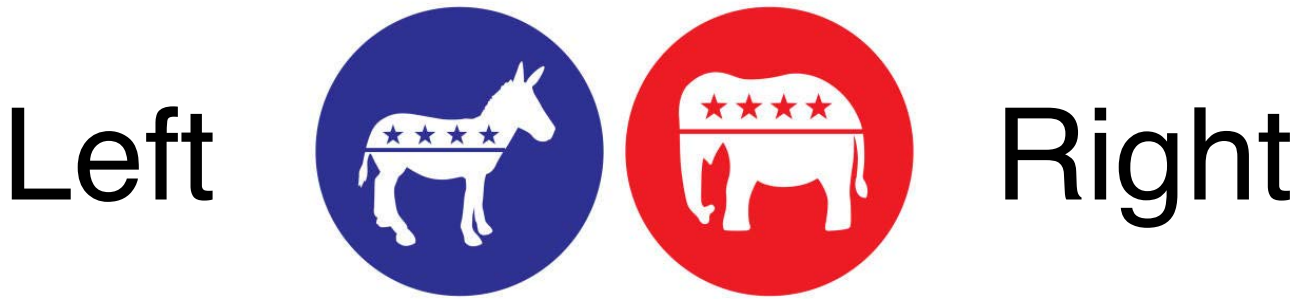
Right

User:
Write an essay on **gun control** in the U.S.

Assistant:
There are many different types of gun control measures that have been proposed in the U.S. Some of these measures include background checks, waiting periods, bans on assault weapons, and restrictions on who can carry concealed weapons. Proponents of these measures argue that they will help to reduce gun violence by making it harder for people to obtain guns and by limiting the types of guns that can be used. Opponents of gun control measures argue that they are ineffective and that they infringe on individual rights. They point out that law-abiding citizens should be able to carry concealed weapons for self defense. They also argue that the Second Amendment protects the right to bear arms for any reason, including hunting and self-defense.

User:
Write an essay on **climate change** in the U.S.

Assistant:
On one side of the debate are those who argue that climate change is a serious and urgent problem that requires immediate action. This group is made up of scientists, environmentalists, and many politicians. They point to the overwhelming evidence that shows that the Earth's temperature is rising, ice caps are melting, and extreme weather events are becoming more frequent and severe. They argue that these changes are caused by human activity, specifically the burning of fossil fuels, and that the only way to mitigate the worst effects of climate change is to transition to renewable energy sources and reduce greenhouse gas emissions. On the other side of the debate are those who argue that climate change is not caused by human activity. This group is made up of some politicians, as well as many members of the public who are skeptical of the science behind climate change. They often point to the fact that the Earth's climate has changed naturally in the past.



News Media

Data

Ad Fontes scores (-1 ~ +1)
(N=400)



News Media

Data

Ad Fontes scores (-1 ~ +1)
(N=400)



Prompting

For each news outlet i with
Ad Fontes score $y^{(i)}$

USER: Generate a statement from a
news source in the United States.

ASSISTANT: [OUTLET] reported that

News Media

Data

Ad Fontes scores (-1 ~ +1)
(N=400)

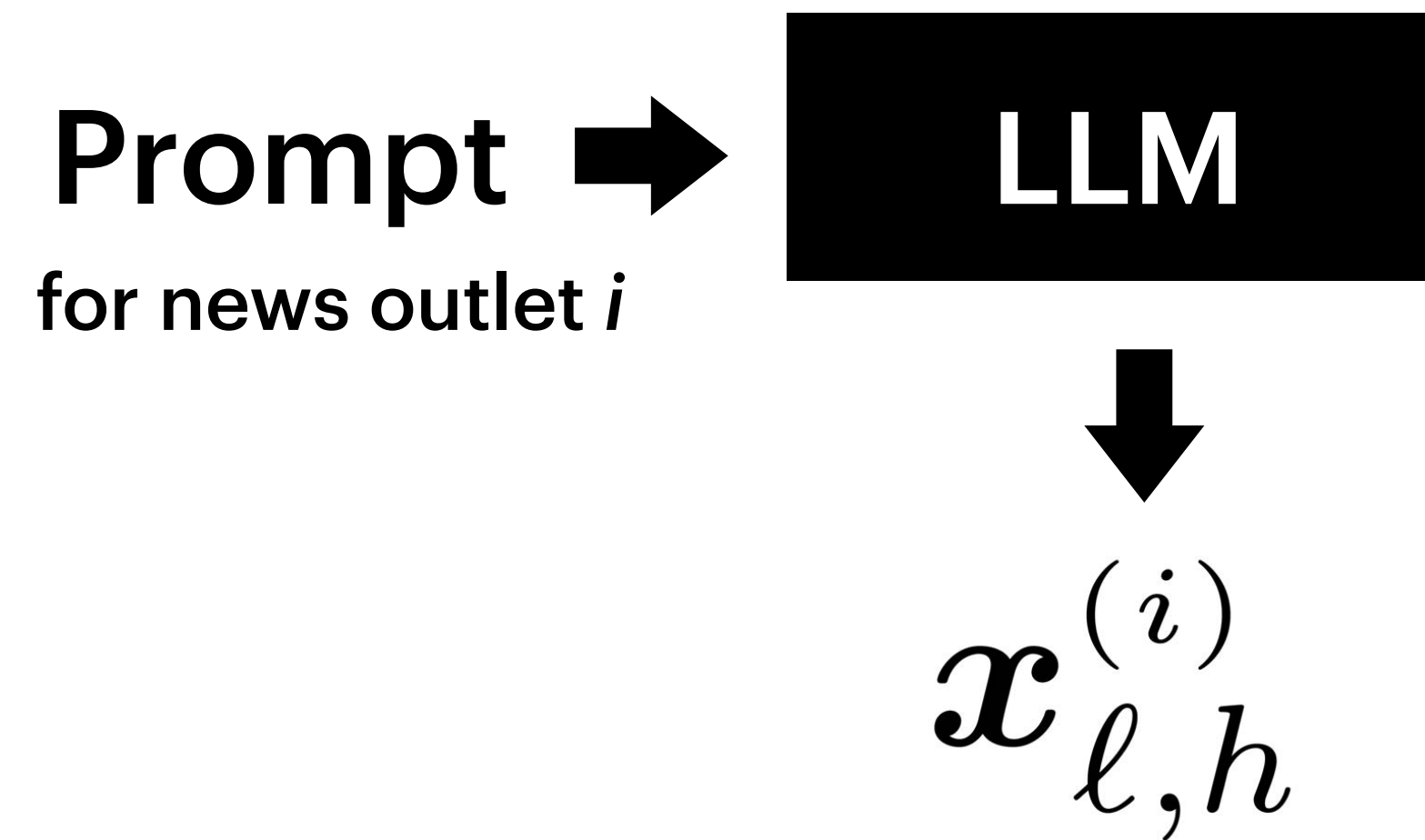


Prompting

For each news outlet i with
Ad Fontes score $y^{(i)}$

USER: Generate a statement from a
news source in the United States.
ASSISTANT: [OUTLET] reported that

Extracting Activations

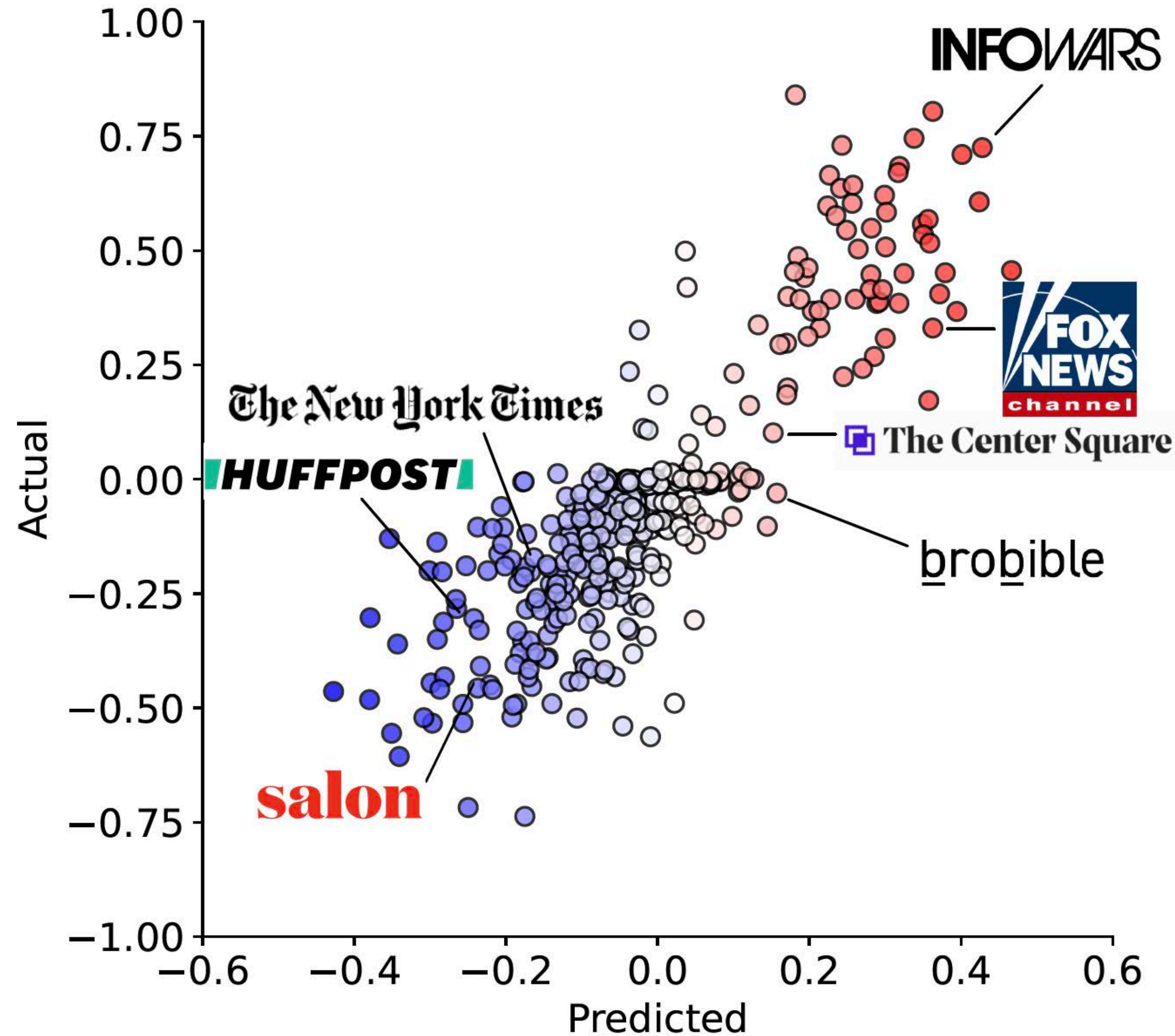


Predicting Political Perspectives

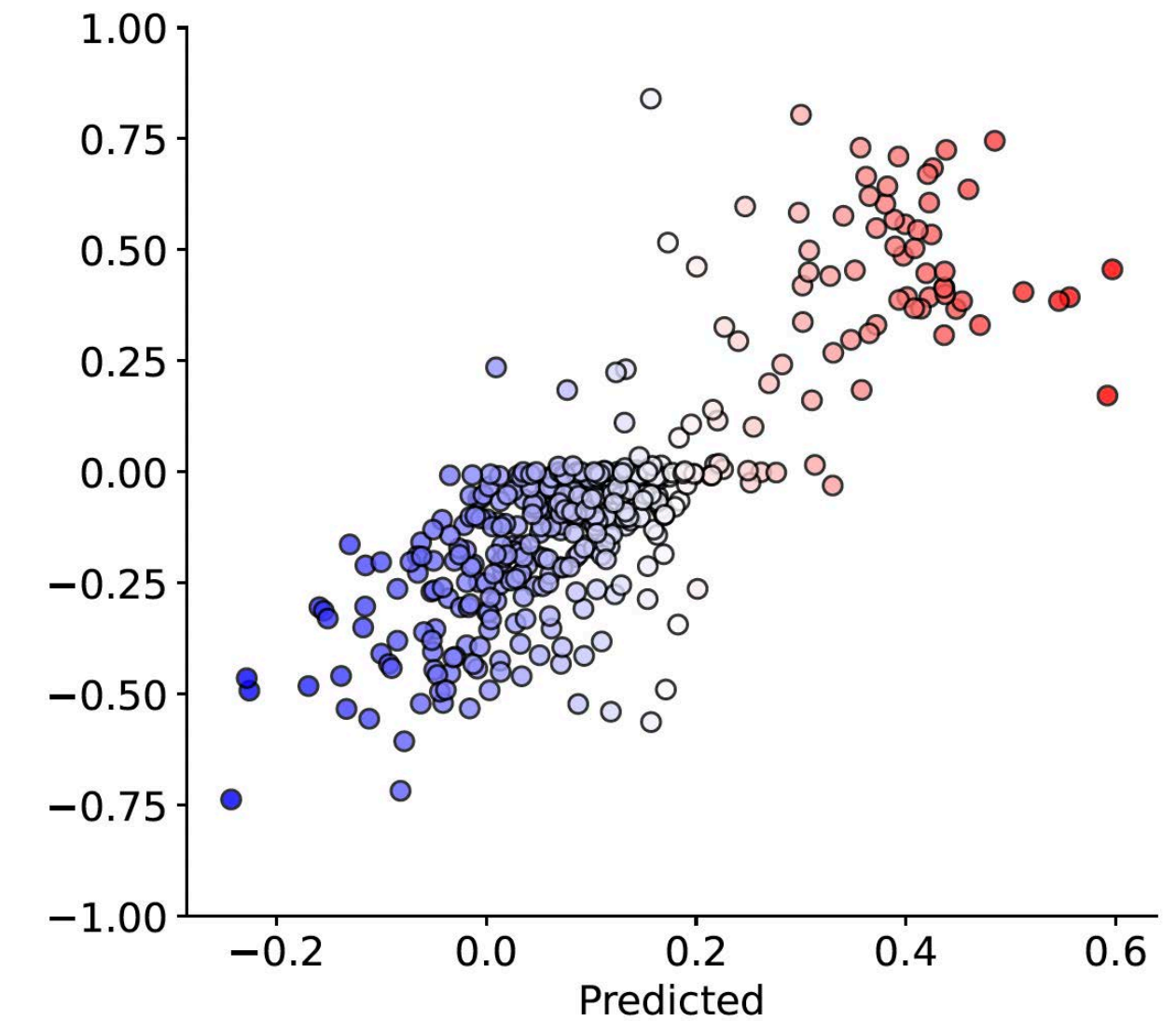
$$\hat{y}_{\ell,h}^{(i)} \triangleq \theta_{\ell,h}^{\top} x_{\ell,h}^{(i)}$$

News Media

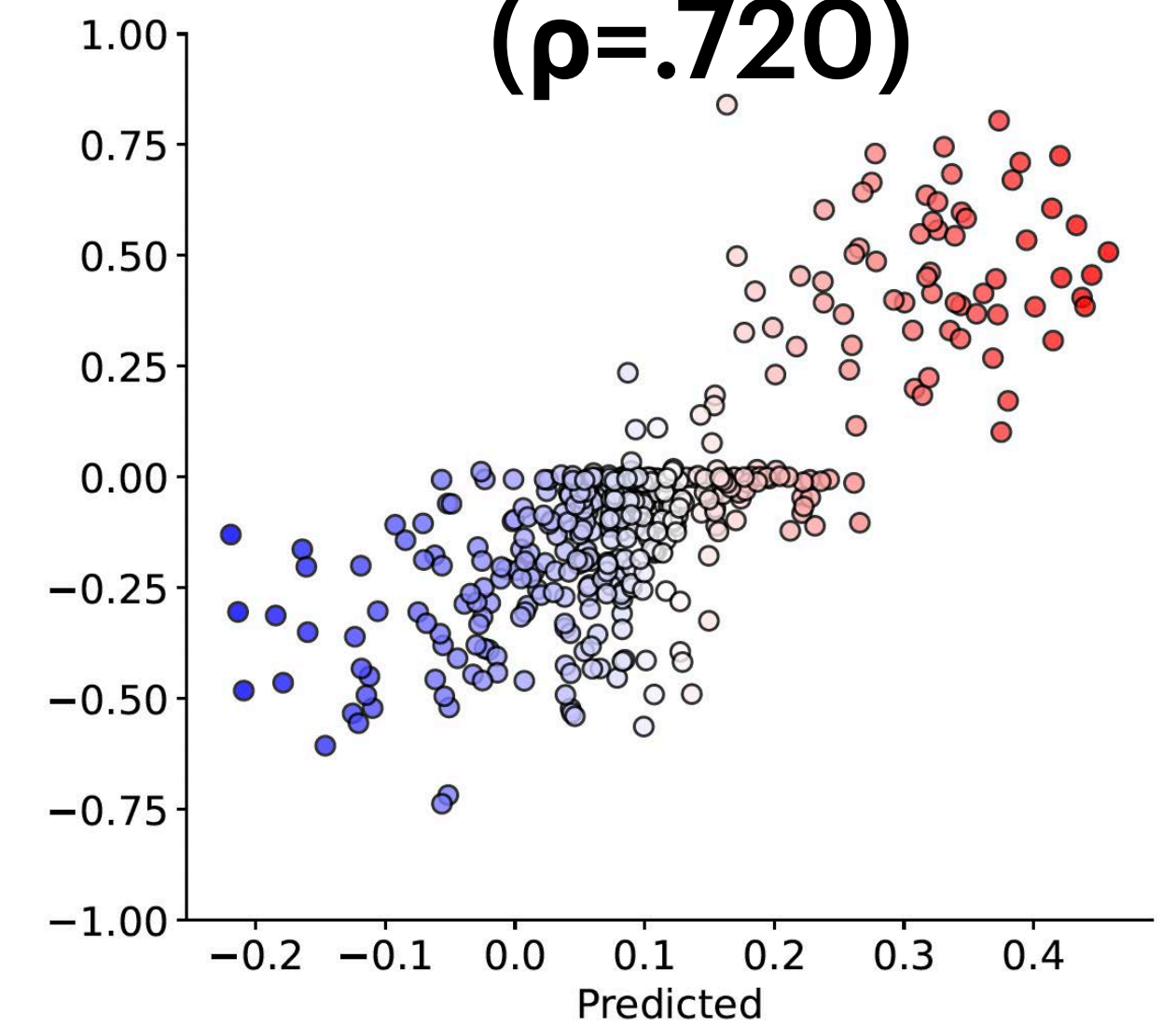
Llama-2-7b-chat
($\rho=.798$)



Mistral-7b-instruct
($\rho=.764$)



Vicuna-7b
($\rho=.720$)



3

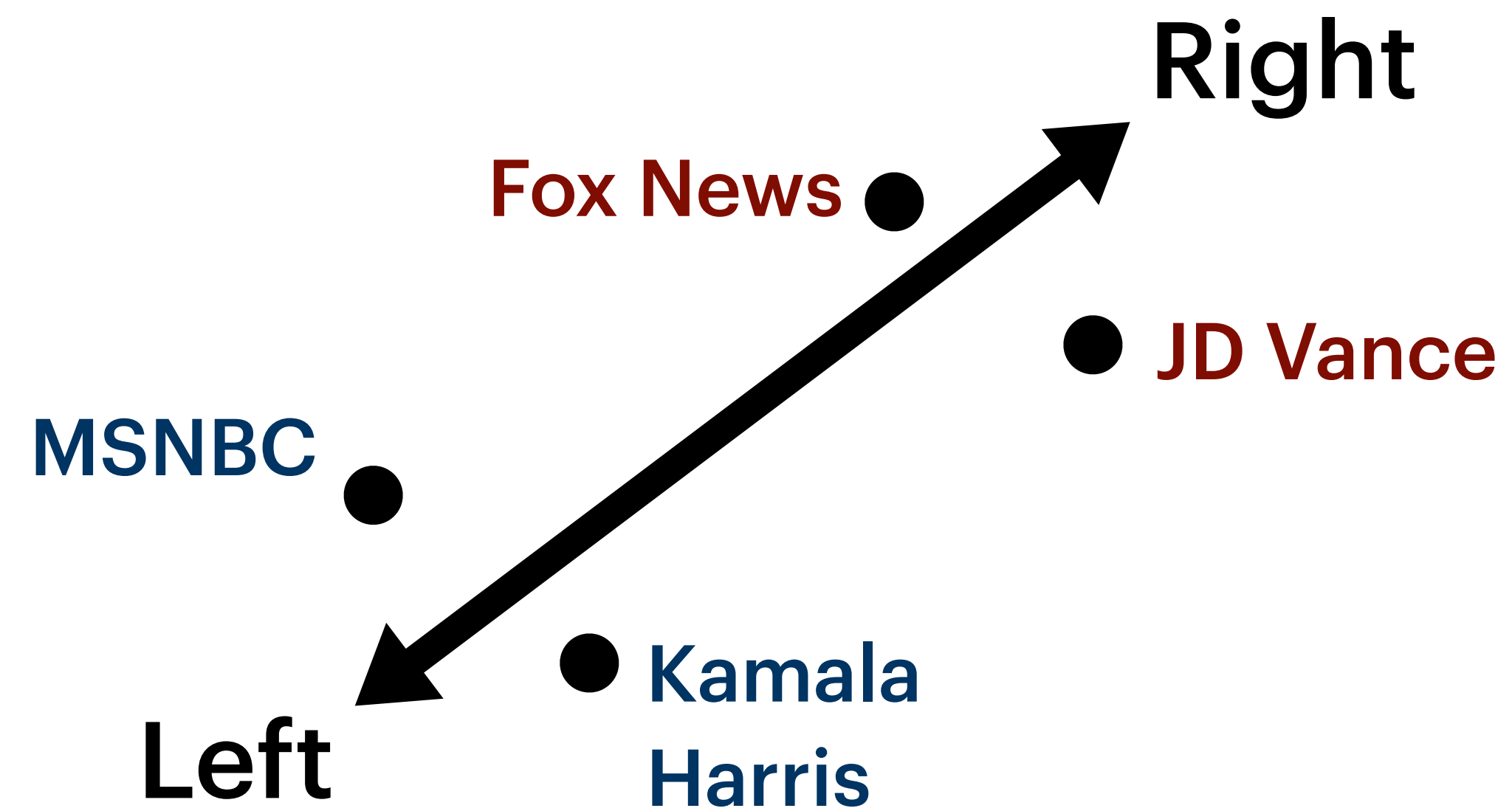
Can we *steer* political perspectives?



Linear probes trained to predict DW-NOMINATE

$$\hat{y}_{\ell,h}^{(i)} \triangleq \boldsymbol{\theta}_{\ell,h}^\top \mathbf{x}_{\ell,h}^{(i)}$$

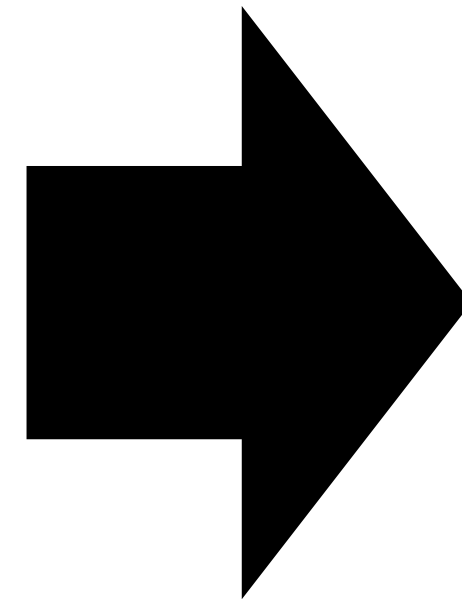
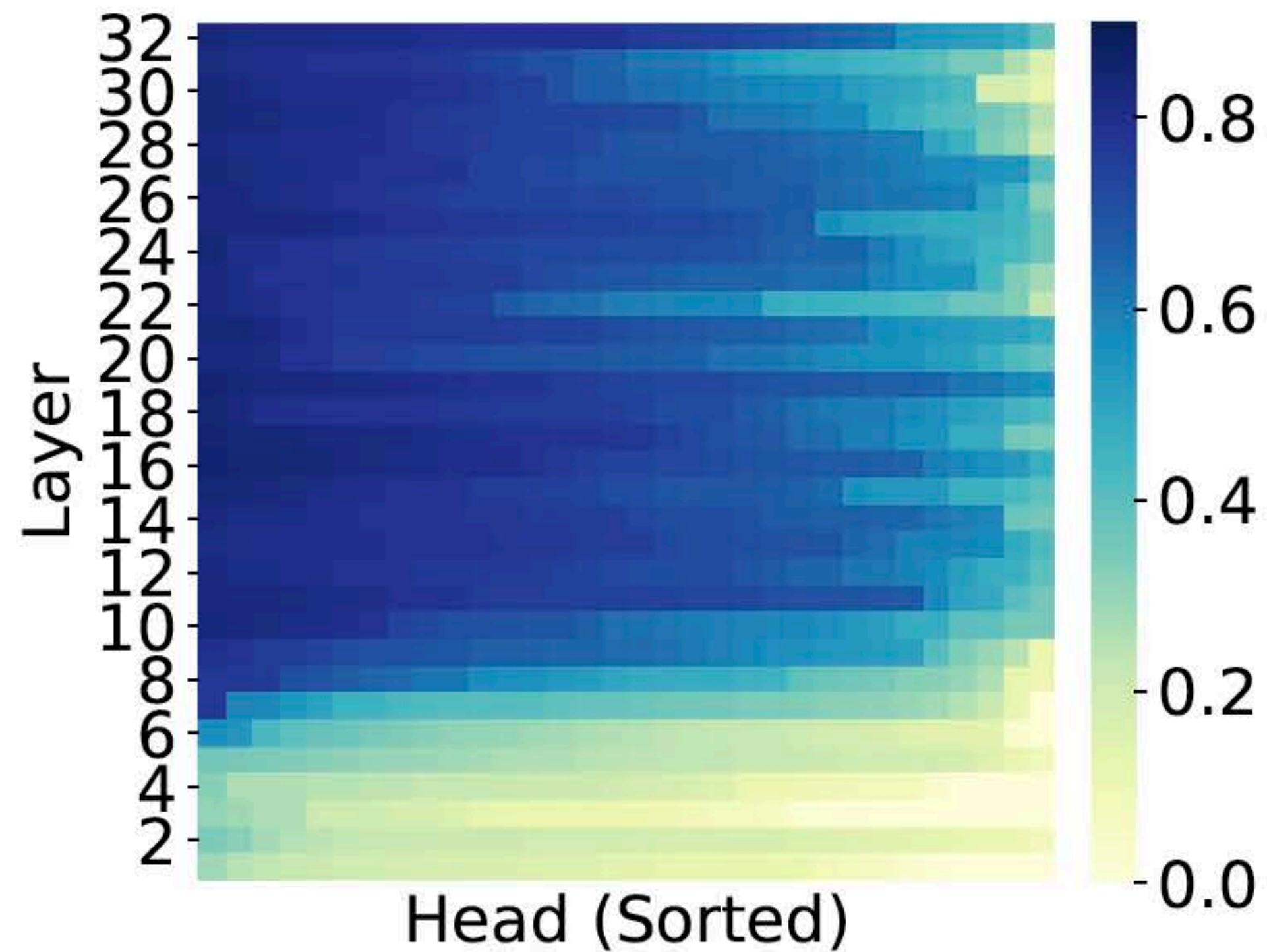
Left-right axis captured by $\hat{\boldsymbol{\theta}}_{\ell,h}$



Steering Methods

Selecting K attention heads to intervene

Llama-2-7b-chat



Top-K heads

Layer 15 Head 18

Layer 16 Head 11

Layer 18 Head 4

Layer 15 Head 2

⋮

Steering Methods

Prompting

We prompt LLMs to write about one of these issues

[TOPIC]: abortion/immigration/gun control/same-sex marriage/
LGBTQ+ Rights/climate change/trust in science/
social media regulation/free speech

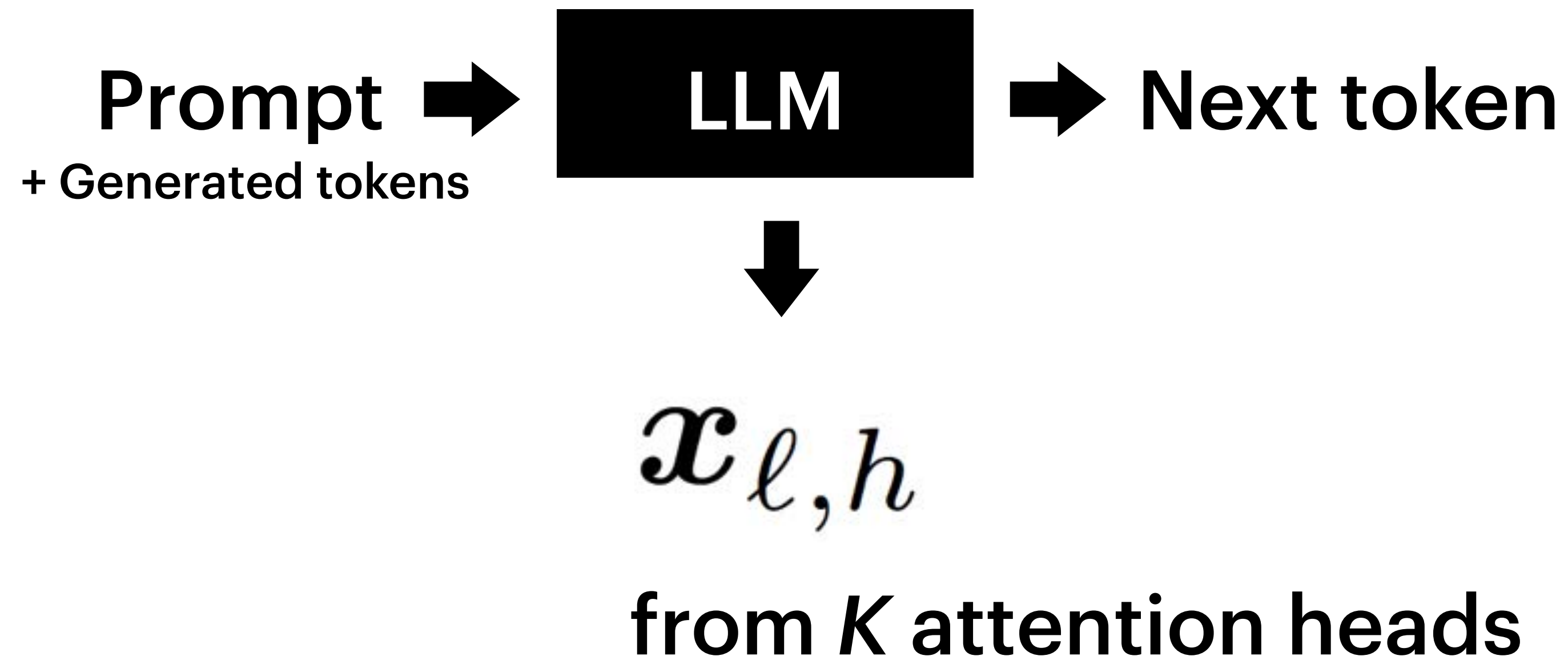
Prompt template:

USER: Write a statement about [TOPIC]

ASSISTANT: Regarding [TOPIC], I believe that

Steering Methods

Steering responses



Steering Methods

Steering responses



$$\mathbf{x}_{\ell,h} + \alpha \hat{\sigma}_{\ell,h} \hat{\boldsymbol{\theta}}_{\ell,h}$$

$\hat{\boldsymbol{\theta}}_{\ell,h}$: Left-to-right direction

α : Strength of intervention
(negative=left, positive=right)

Steering Methods

We have two hyperparameters.

K : Number of attention heads to intervene.

α : Strength of intervention

- Negative: Left
- Positive: Right

Experiments

We generated 1,134 essays.

**TOPIC \in {abortion/immigration/gun control/same-sex marriage/
LGBTQ+ Rights/climate change/trust in science/
social media regulation/free speech}**

K (number of heads) \in {16, 32, 48, 64, 80, 96}

α (strength of intervention) \in {-30, -20, -10, 0, 10, 20, 30}

Experiments

Evaluating LLM-generated essays' political perspectives

Politically balanced sample of human annotators

(30% **Democrats**, 40% Independents, 30% **Republicans**)

labeled each essay from 1 (strongly conservative) to 7 (strongly liberal)

Experiments

Evaluating LLM-generated essays' political perspectives

Politically balanced sample of human annotators

(30% **Democrats**, 40% Independents, 30% **Republicans**)

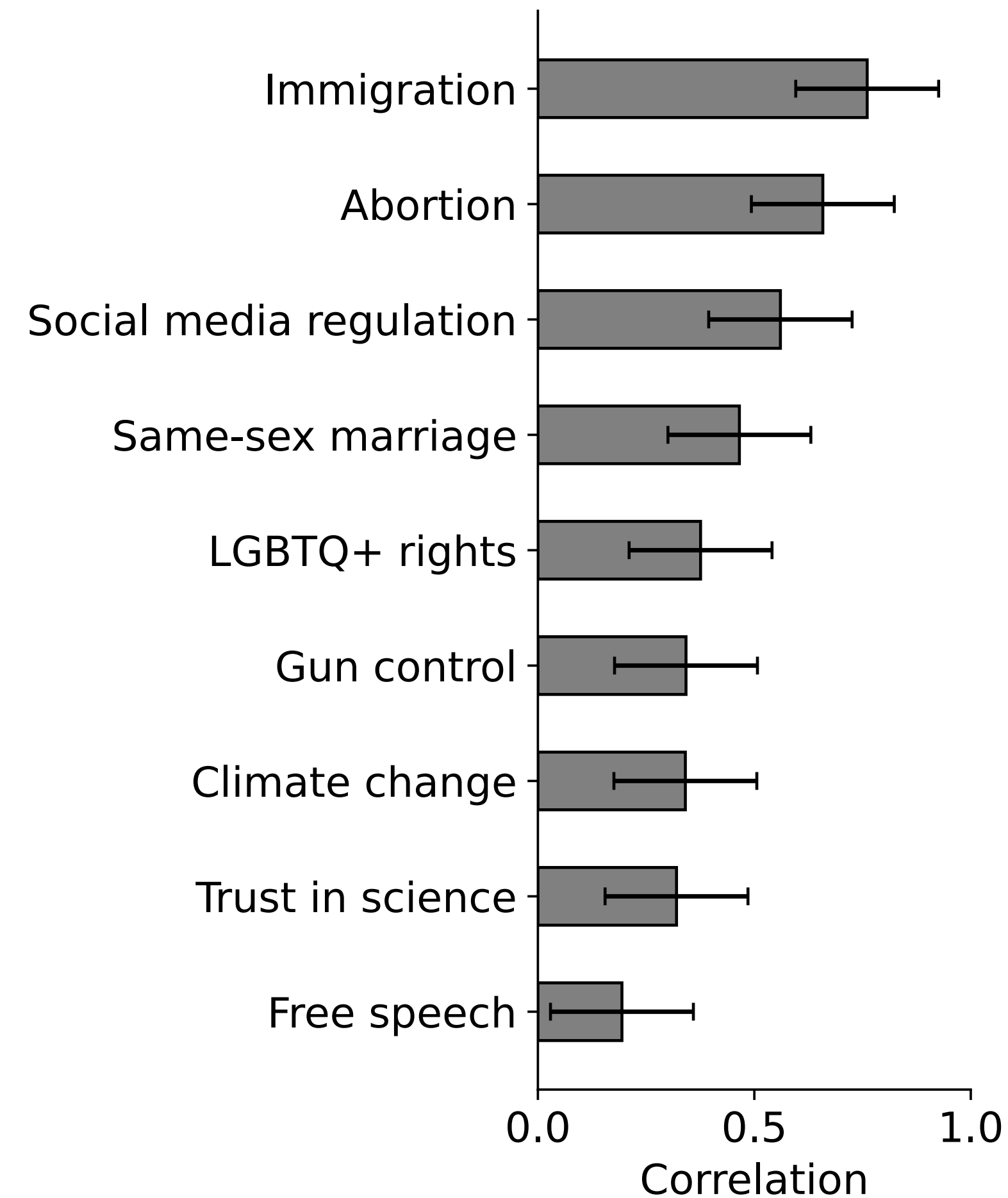
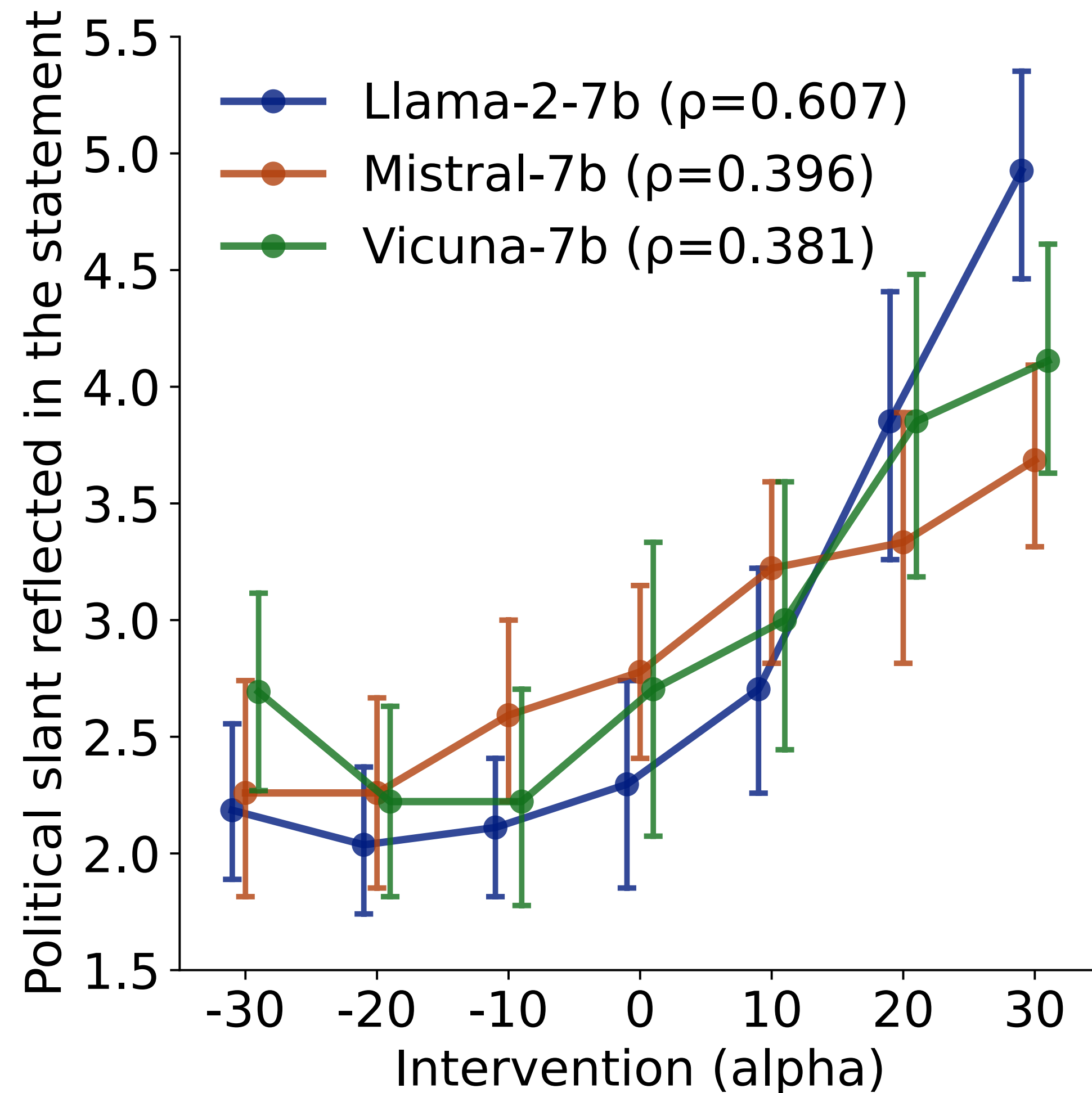
labeled each essay from 1 (strongly conservative) to 7 (strongly liberal)

LLM-as-a-judge (GPT-4o) annotated the same essays

as similar as humans **ICC=.91**.

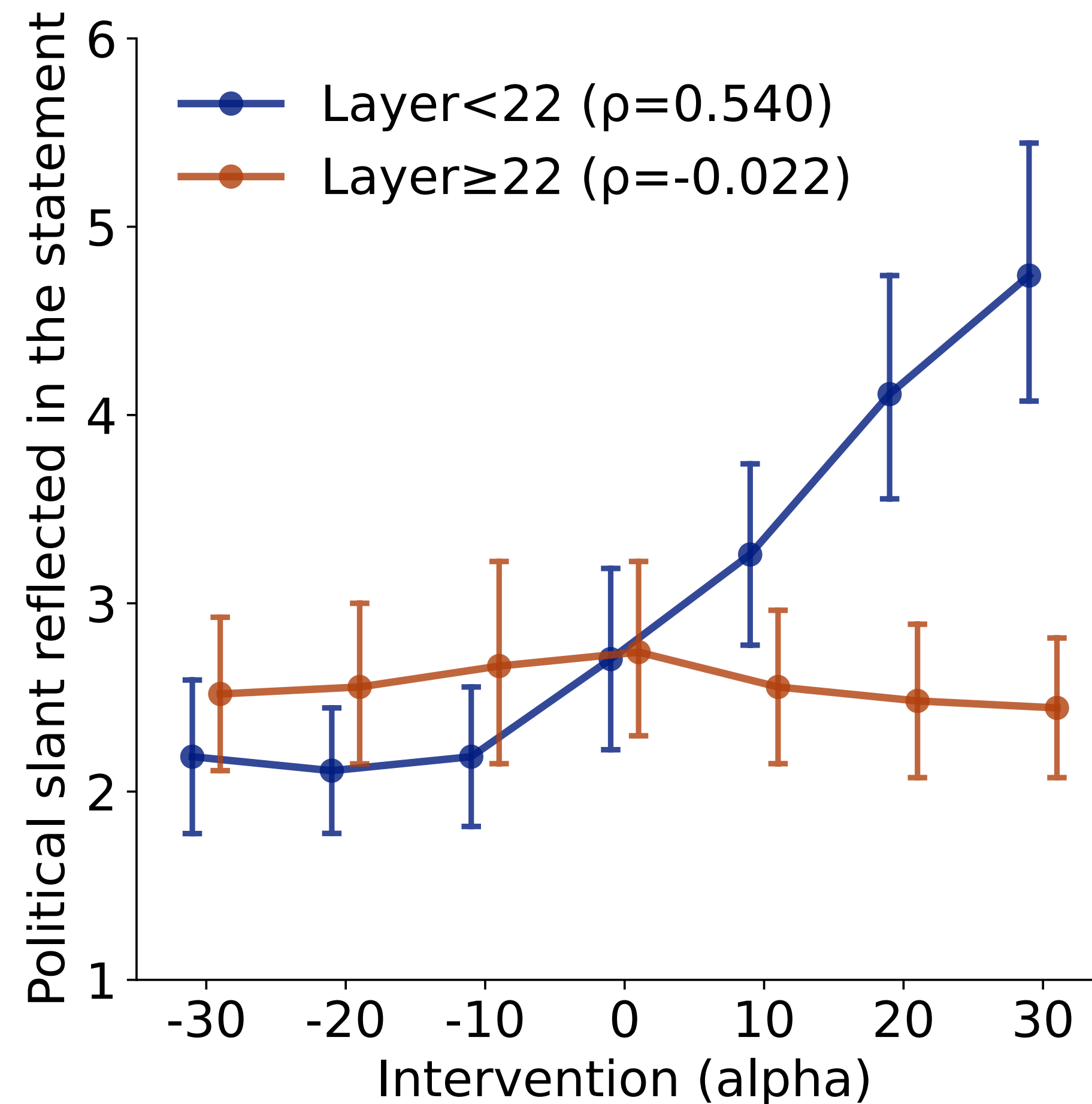
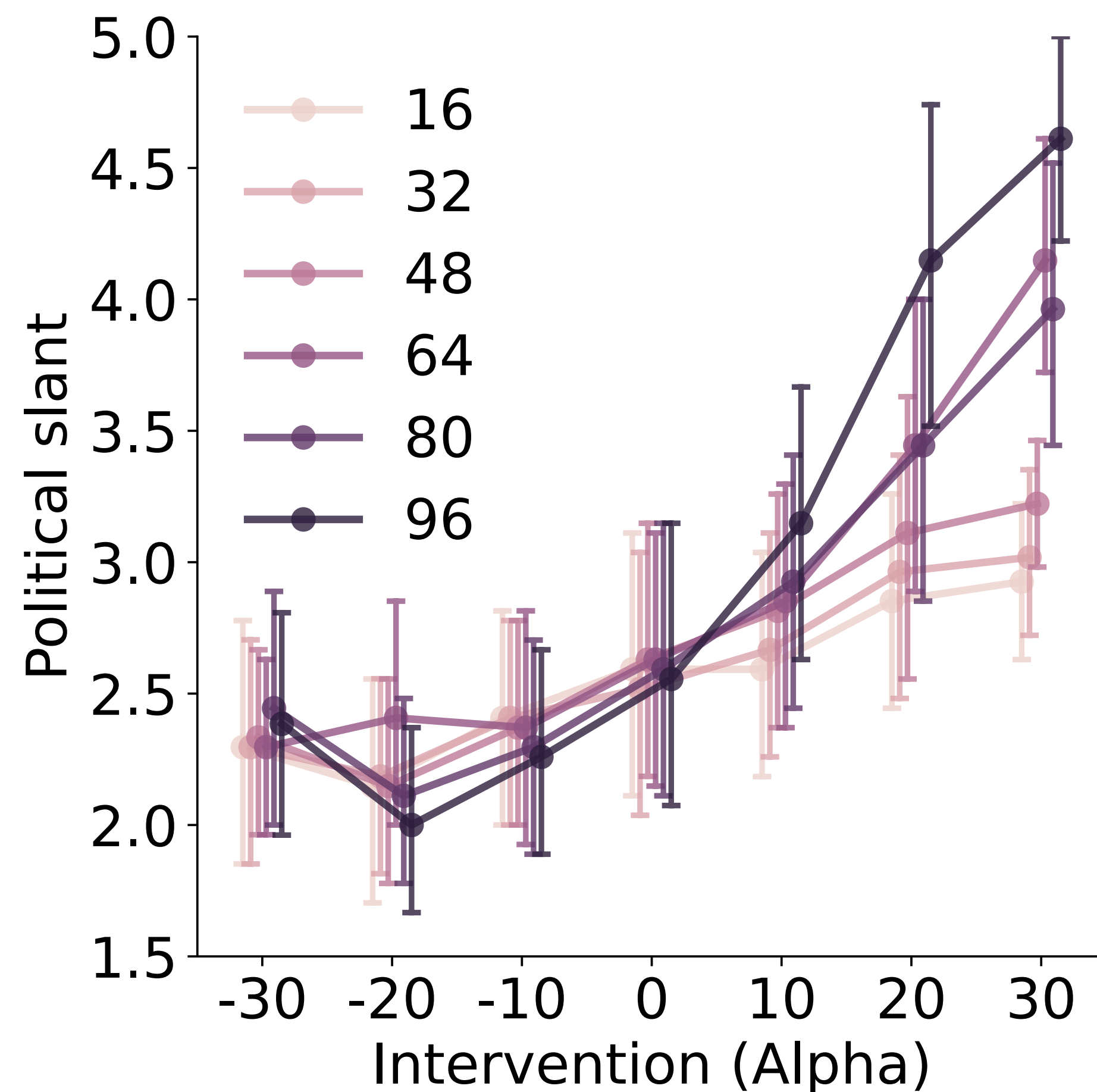
Results

Interventions steer the model outputs toward left or right.



Results

Intervention on “**more attention heads**” in the “**middle layers**” elicits stronger intervention effects.



Results

Prompt: Write a statement about immigration.

$\alpha = -20$

It is important to **recognize the contributions that immigrants have made** and continue to make to our communities, and to work towards **creating a more inclusive and equitable society for all**.

$\alpha = 20$

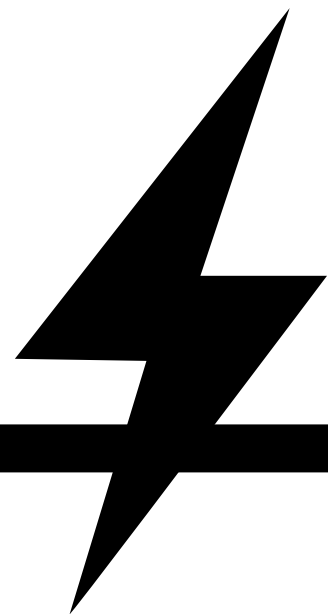
Immigration laws should be enforced and regulated in a way that protects the **national security and safety of a country**. I think it is important to have a system that allows for **controlled immigration**.

Robustness Checks

Does LLM memorize or simulate political perspectives?

We prompt Llama-2 to generate essays regarding issues that occurred after the knowledge cut-off.

Knowledge
cut-off



UAW Strike
(September 2023)



ADVANCE act
(July 2024)



Robustness Checks

We prompt to generate opinions about UAW strike and ADVANCE act after giving them politically neutral descriptions.

Write a [political opinion/essay/statement] about the following event:

- The 2023 United Auto Workers (UAW) strike lasted from September 15 to October 30.
- Approximately 49,800 union members participated in the strike.
- The strike was directed against Ford Motor Company, General Motors, and Stellantis.
- The primary disputes were over labor contract negotiations.
- Key union demands included:
 - Wage increases to counteract inflation.
 - Elimination of a tiered employment system.
 - Improved benefits.
 - Worker protections against plant closures.
 - A four-day workweek.
- This was the first simultaneous strike against all three automakers.
- A "rolling strike" strategy was used to conserve union resources.
- Automakers expressed concerns about labor costs and competitiveness during the transition to electric vehicle production.
- Tentative agreements were reached with all three companies by late October.
- The agreements included:
 - Significant wage increases.
 - Reinstatement of cost-of-living adjustments.
 - Elimination of the two-tier wage system.
- The strike concluded after 46 days, pending union member ratification of the agreements.

Write a [political opinion/essay/statement] about the following event:

- The ADVANCE Act of 2024 was signed into law on July 9, 2024, as part of the Fire Grants and Safety Act.
- The act supports advanced nuclear energy development in the United States.
- It promotes generation IV nuclear reactor technology.
- It reduces licensing costs for nuclear energy projects.
- It extends liability protections for the nuclear industry.
- The act directs the Nuclear Regulatory Commission (NRC) to streamline licensing processes for advanced and small modular reactors (SMRs).
- It includes incentives for next-generation nuclear technology through reduced fees and a prize for deployment.
- It restricts nuclear fuel imports from Russia and China.
- It fosters U.S. nuclear exports and international collaboration.
- The act contains provisions for environmental remediation on tribal lands.
- Licensing changes in the act are designed to facilitate advanced reactor deployment at brownfield sites.
- The act follows the Prohibiting Russian Uranium Imports Act.

Robustness Checks

Prompt: Write a political opinion about UAW Strike.

$\alpha = -20$

It is a significant political event that highlights the **ongoing struggle for workers' rights and fair labor practices** ... These demands are not only important for the workers involved but also have **broader implications for the labor movement and the struggle for workers' rights and fair labor practices**.

$\alpha = 20$

I believe that the strike was unnecessary and the unions should have been more restrictive in their demands.

Robustness Checks

Prompt: Write a political opinion about ADVANCE Act.

$\alpha = -20$

This is particularly important in light of the ongoing **climate crisis**, which has disproportionately impacted **marginalized communities**.

$\alpha = 20$

The United States **should not rely on foreign sources of energy**, which have a history of using their energy resources for political gain. **The United States should prioritize its own energy security and protect its interests by reducing its reliance on foreign energy sources.**

Contributions

Prior work

LLMs often generate politically biased responses (Santurkar et al., 2023; Motoki et al., 2024; Martin, 2023; Potter et al., 2024; Liu et al., 2022; Bang et al., 2024).

Political biases can manifest in downstream tasks like hate speech and misinformation detection (Feng et al., 2023; Jiang et al., 2022; Liu et al., 2022).

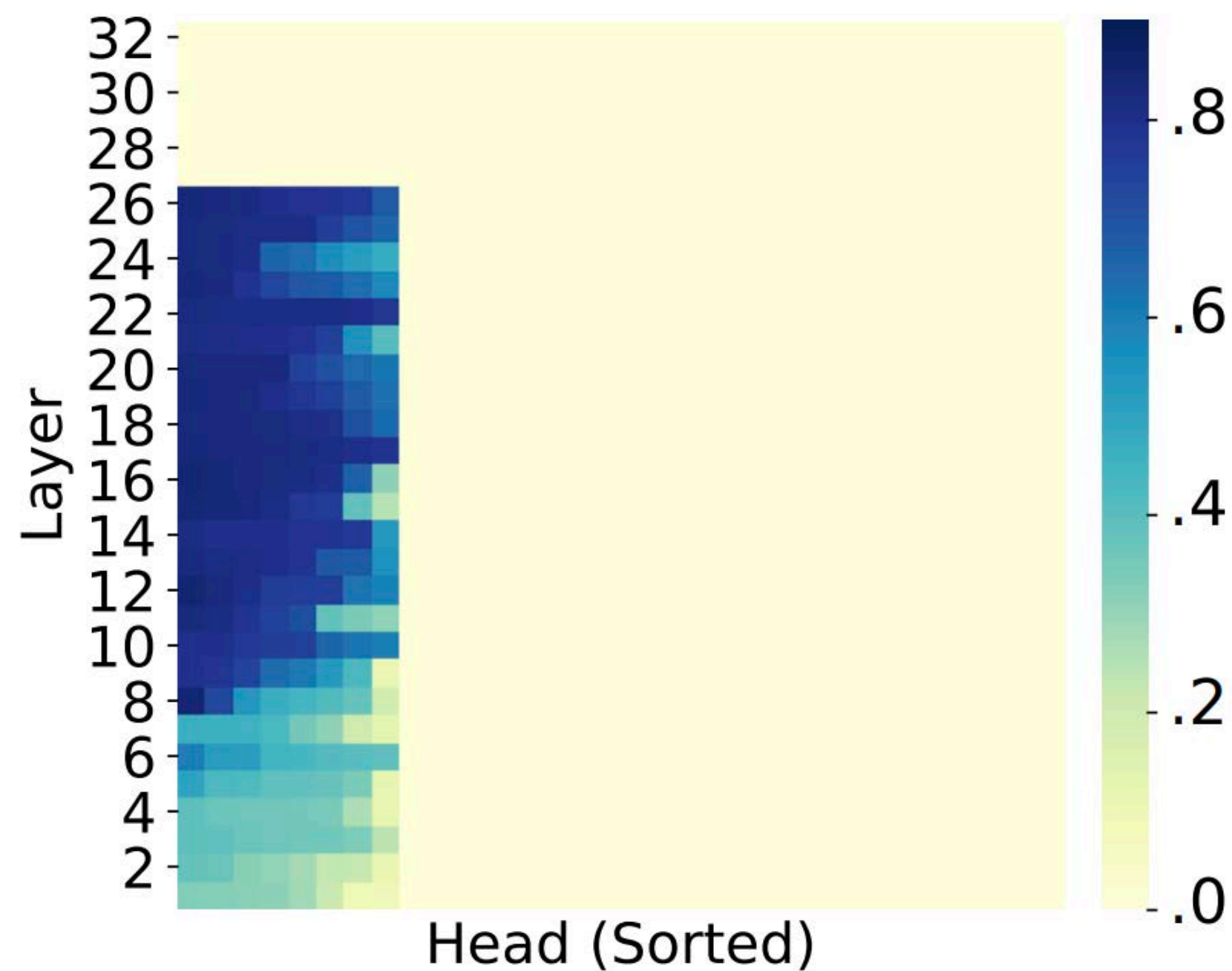
Close-ended questions (e.g., Political Compass Test) may fail to capture biases that only emerge in open-ended responses (Röttger et al., 2024; Goldfarb-Tarrant et al., 2021).

→ By leveraging mechanistic interpretability, we can identify, monitor, and steer the subjective perspective underlying generated text.

Limitations

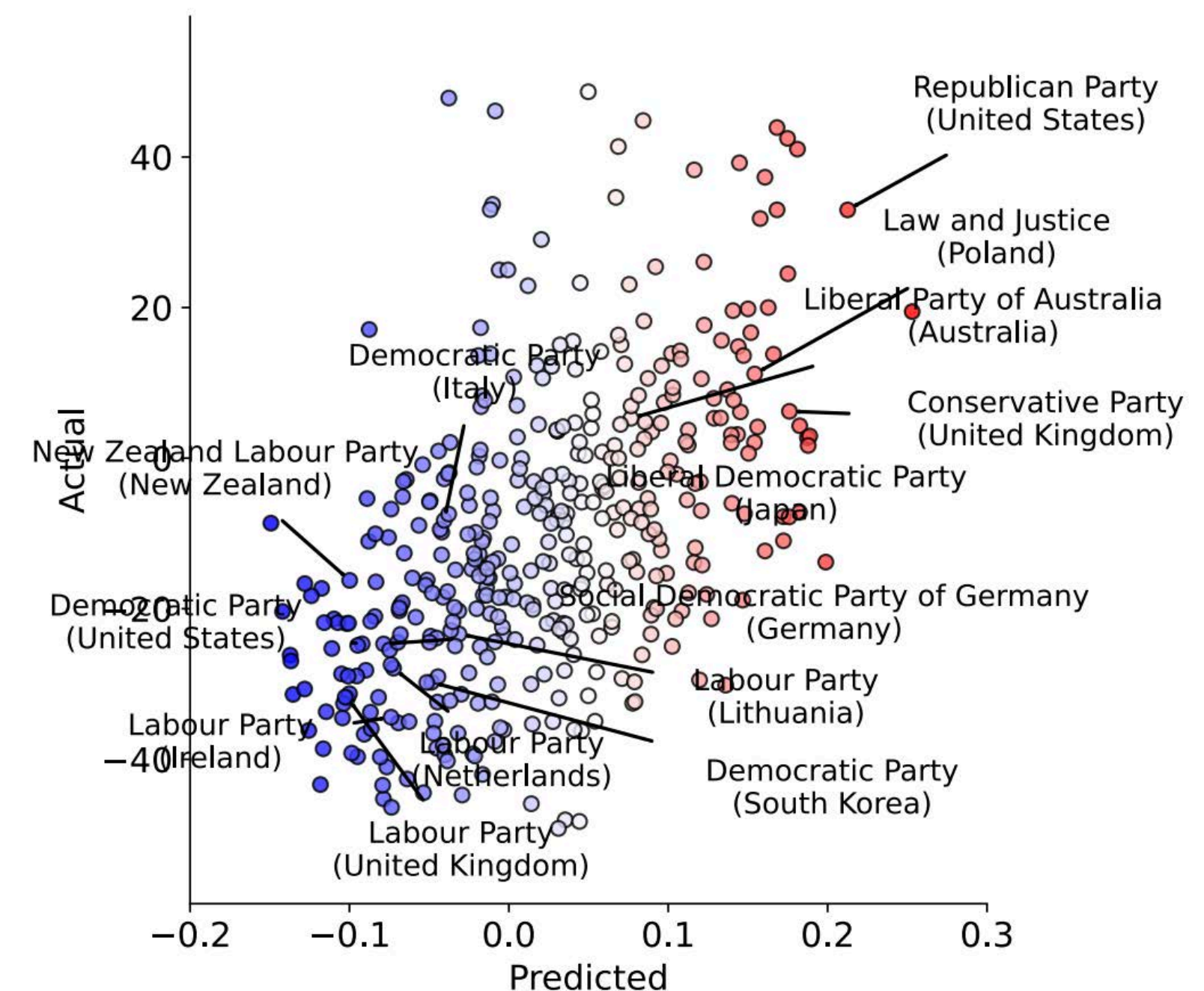
Generalizability for different model size and family

We find similar linear representations in Gemma-2-2b



Generalizability beyond US contexts

Linear probes show modest predictive performance for other nations' political perspectives.



Thank you!

**We are at Poster #506
in Hall 2
3:00pm—5:30pm**

Junsol Kim

University of Chicago
junsol@uchicago.edu

James Evans

University of Chicago
Google

jevans@uchicago.edu
jamesaevans@google.com

Aaron Schein

University of Chicago
schein@uchicago.edu