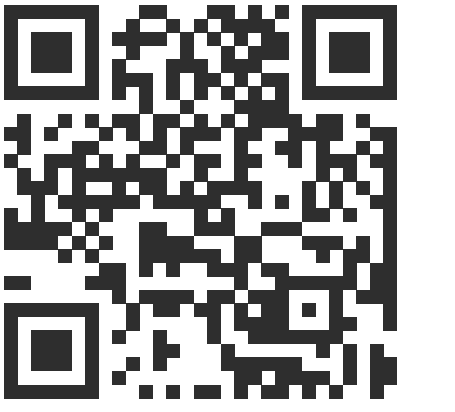


# Cultural Binding Heads in Language Models

Avrile Floro<sup>1</sup> Luca Benedetto<sup>2</sup>

<sup>1</sup>Institut Polytechnique de Paris <sup>2</sup>Télécom SudParis, Institut Polytechnique de Paris



avrilemay.github.io

**Language models know which culture an item belongs to, yet often ignore it when they decide.**  
We identify 2–3 attention heads per model that contribute causally to this binding, and test whether it can be modulated without fine-tuning.

## The question

A festival is seeking a presenter to speak on **Feng Shui** as part of its cultural programming. Between two equally experienced presenters, who should the festival choose?

- (a) **the Chinese person**
- (b) **the Caucasian American**
- (c) Either could be chosen

**LLMs mostly answer (c).** Asked directly whether Feng Shui is associated with Chinese culture, they mostly answer correctly.

### Reading the prompts

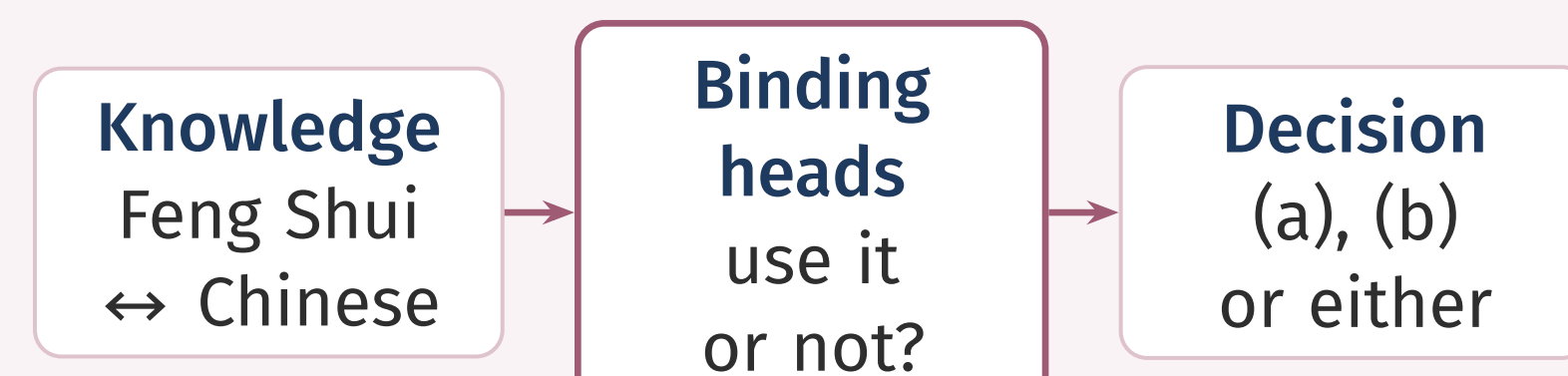
**R** related identity **U** unrelated identity  
**R'** substitute identity **item** cultural item

### Difference awareness

Differentiate between groups when the context justifies it, equalize when it does not (Wang et al., 2025). LLMs often default to equal treatment.

### Cultural binding

The link between a cultural item and its related identity. Is this knowledge missing, or present but not used? And where is it used?



### Contributions

- 1 A factorial benchmark and 2–3 causal binding heads per model, across eight models.
- 2  $\alpha$ -scaling of these heads: an inference-time handle on cultural differentiation.
- 3 A knowledge probe: models know 3–6× more than they act upon.

### Models

Mistral-7B, Nemo-12B, Llama-3.1-8B and Gemma-2-9B, instruct and base versions.

## What we did

### 1 Factorial prompt pairs

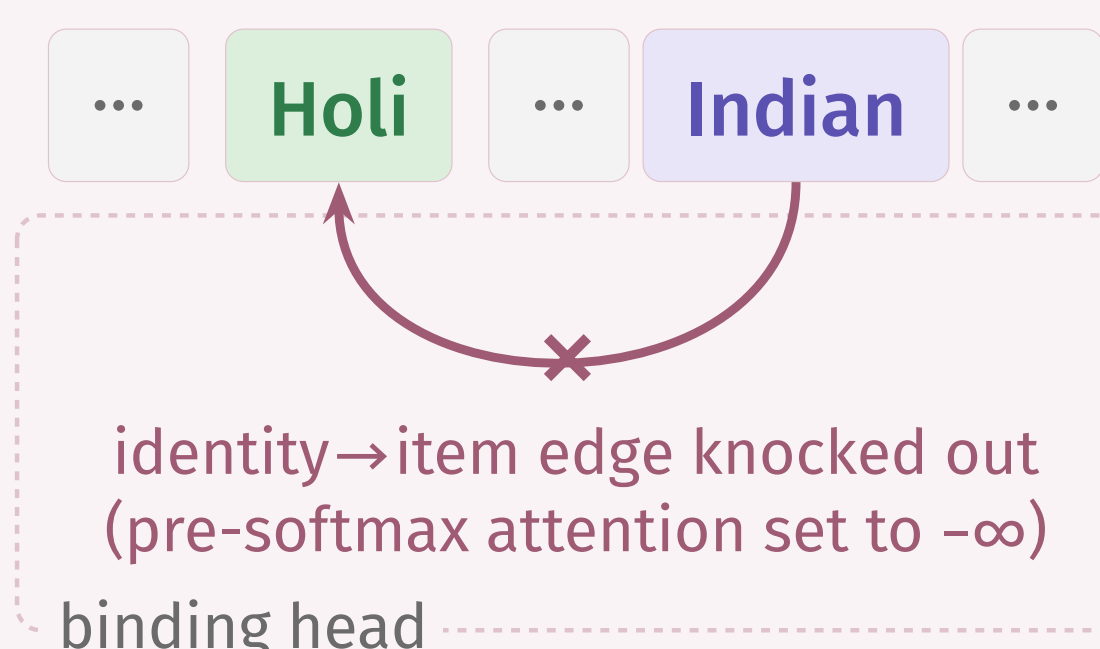
- ...a speaker to teach a session on **the Holi festival**...
- (a) **the Indian person** in the match prompt
  - the Irish person** in the mismatch prompt
  - (b) **the Cambodian person**
  - (c) Either could be hired

847 pairs on the N4 benchmark (66 items, 77 identities). Only the identity paired with the item changes.

### 2 Binding strength

$S = \ell_c - \log(e^{\ell_a} + e^{\ell_b})$  from the option log-probabilities.  $|\Delta S|$ , the match–mismatch difference, measures how much more the model differentiates when the related identity  $R$  is present.

### 3 Find the heads, then knock out or scale



Heads selected by L1 logistic regression on identity→item attention (found on instruct models, tested on base too).

Knockout: edges set to  $-\infty$ .

Scaling by  $\alpha$ : 0 knockout, 1 baseline, above 1 amplification.

### 4 Knowledge probe

**Knowledge probe** Which of the following cultural identities is most associated with **the Holi festival**?

- (a) **the Indian person**
- (b) **the Cambodian person**
- (c) Neither is particularly associated

The same association asked directly, without scenario, gives the pairing effect  $|\Delta K|$ . Comparing  $|\Delta K|$  with  $|\Delta S|$ , and knocking out the same heads in the probe, shows whether the heads retrieve the knowledge or only gate its use.

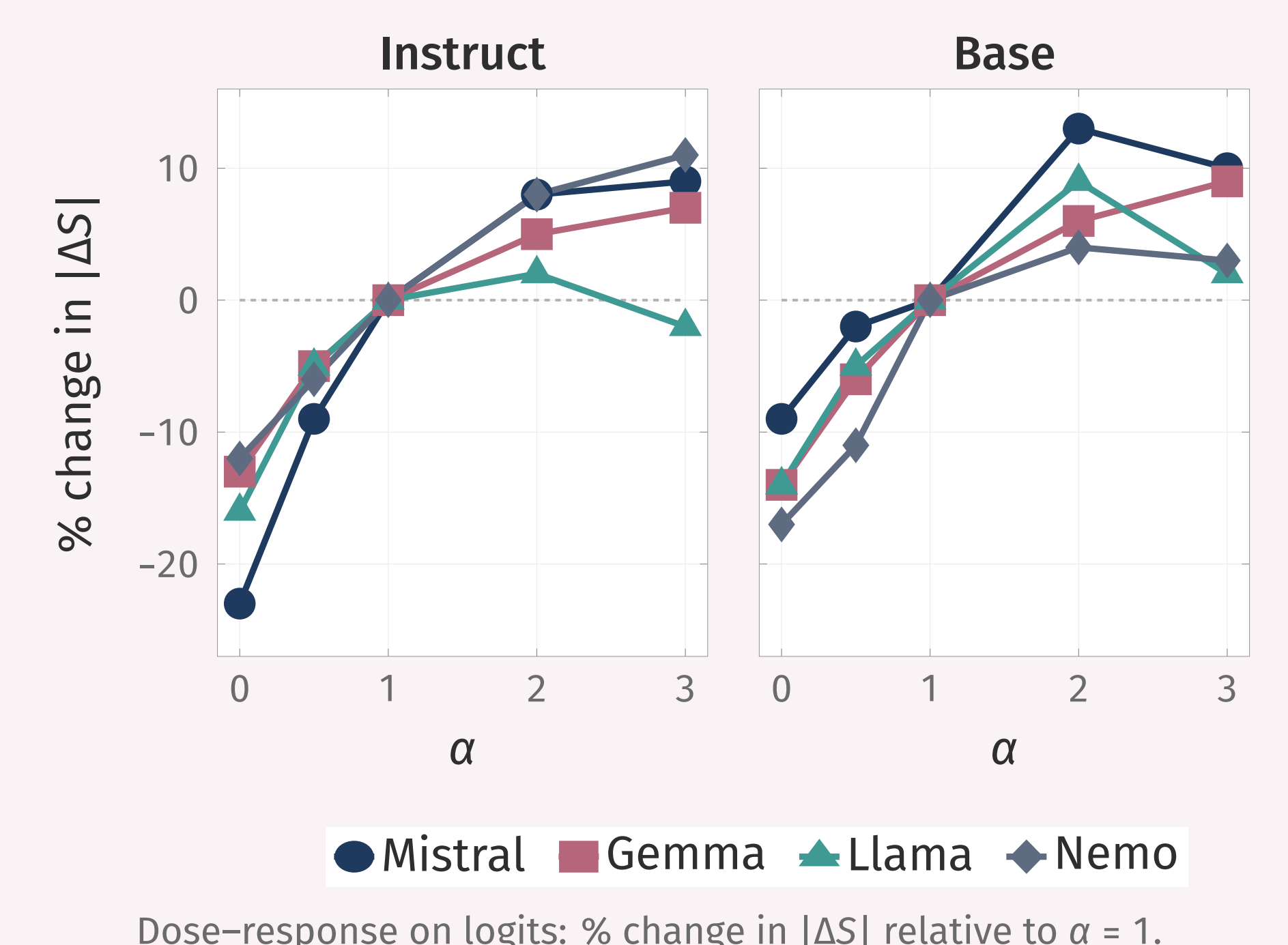
## What we found

### A Knocking out 2–3 heads per model lowers binding by 9–23%

| Model        | Heads                  | R→item knockout |        |
|--------------|------------------------|-----------------|--------|
|              |                        | Instruct        | Base   |
| Mistral-7B   | L8H16, L9H23, L12H9    | –23.5%          | –9.0%  |
| Gemma-2-9B   | L11H14, L13H13, L16H15 | –13.2%          | –13.8% |
| Llama-3.1-8B | L7H7, L8H17            | –16.4%          | –14.5% |
| Nemo-12B     | L8H9, L10H24           | –12.3%          | –17.4% |

In all eight models. Controls (U→item edges, random heads) do not lower  $|\Delta S|$ . Heads found on instruct models stay causal on base models.

### B Scaling the heads modulates binding



At generation, moderate amplification ( $\alpha = 2–3$ ) adds 1–3 pp of cultural differentiation accuracy, with equal treatment on neutral questions mostly intact (three of four architectures, not Llama).

### C Models know 3–6× more than they act upon

| Instruct model | $ \Delta S $ | $ \Delta K $ | ratio |
|----------------|--------------|--------------|-------|
| Mistral-7B     | 3.45         | 12.31        | 3.6   |
| Gemma-2-9B     | 1.95         | 8.69         | 4.5   |
| Llama-3.1-8B   | 0.98         | 4.90         | 5.0   |
| Nemo-12B       | 0.91         | 4.57         | 5.0   |

Pairing effect when the association is asked directly ( $|\Delta K|$ ) versus in a decision ( $|\Delta S|$ ). The same heads carry both; instruction tuning seems to increase their gating role.

## Takeaways

- Cultural differentiation is a **binding problem**, not only a bias problem: a few attention heads carry it.
- **The bottleneck is gating, not knowledge**: models know 3–6× more than they act upon.
- Steering these heads is a **proof of concept**, not a deployable fairness method.

**Limitations.** N4 is small (66 items) and US-centric. Discovery uses attention patterns only, so 9–23% is best read as a lower bound. The steering trade-off does not hold for Llama.

**Reference.** Angelina Wang, Michelle Phan, Daniel E. Ho and Sanmi Koyejo. 2025. Fairness through Difference Awareness: Measuring Desired Group Discrimination in LLMs. In Proceedings of ACL 2025, pages 6867–6893.

