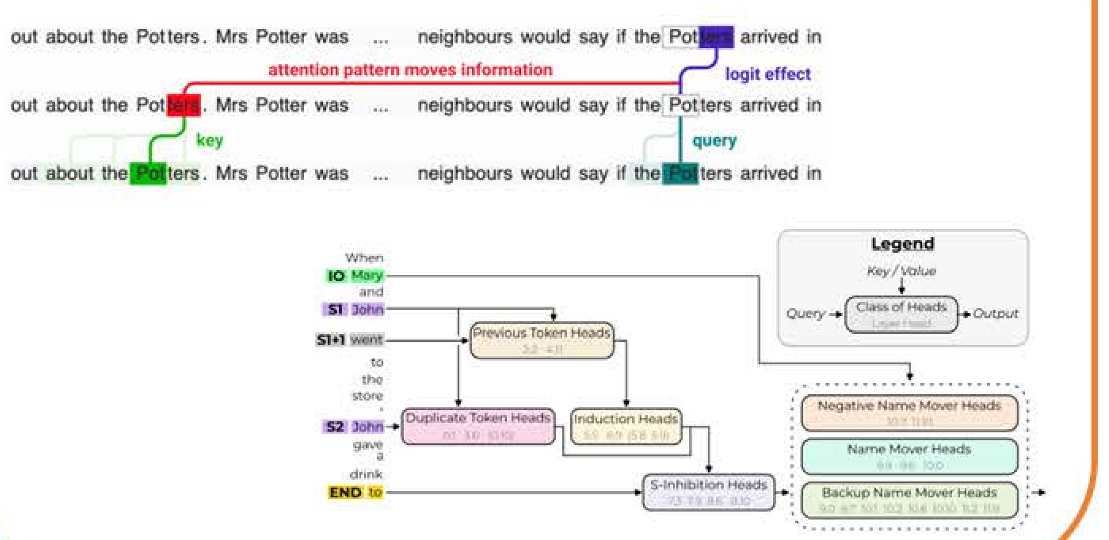
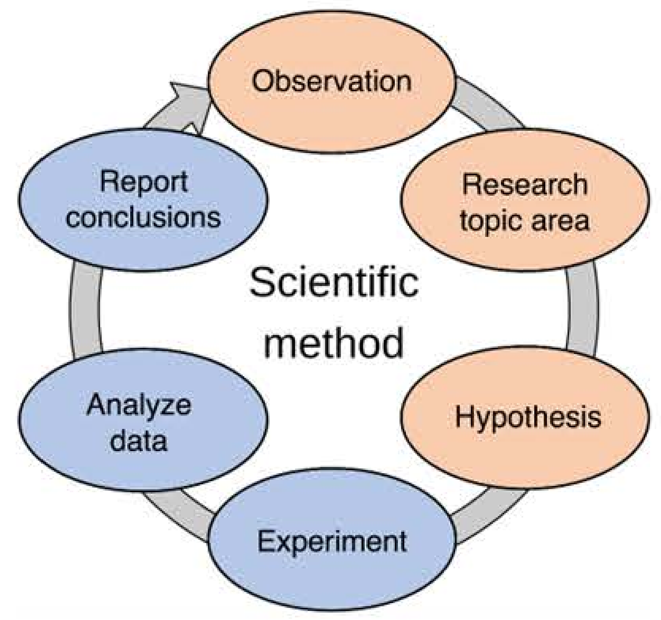
Causal Head Gating: A Framework for Interpreting Roles of Attention Heads in Transformers

Andrew J. Nam, Henry C. Conklin
Yukang Yang, Thomas L. Griffiths
Jonathan D. Cohen, Sarah-Jane Leslie

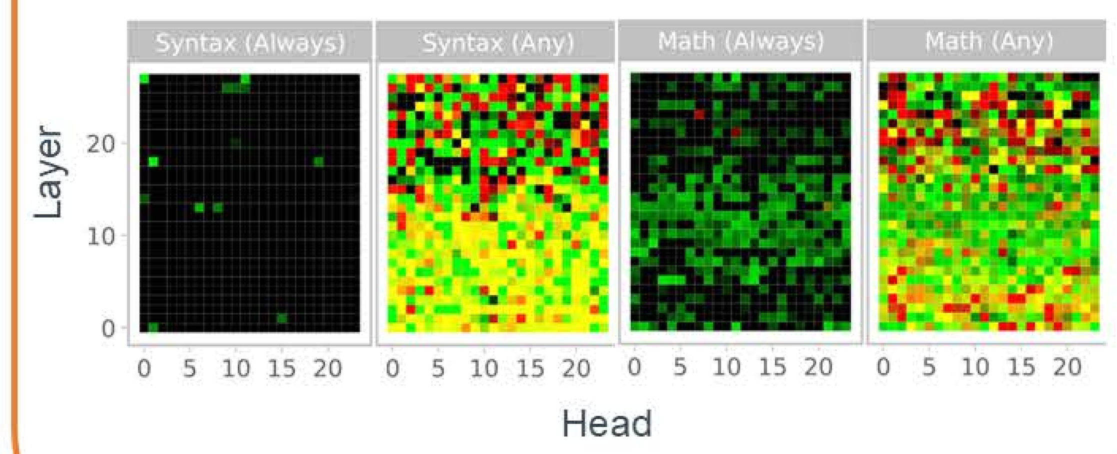
PRINCETON UNIVERSITY
NEURAL INFORMATION PROCESSING SYSTEMS

Mechanistic interpretability is hypothesis-driven[1,2]





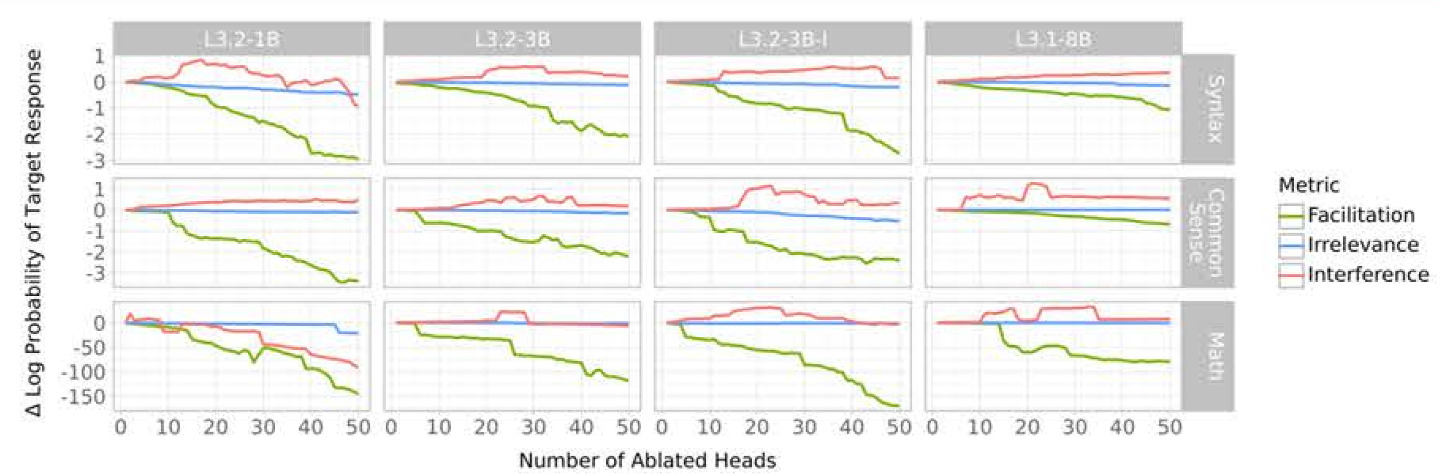
CHG offers an exploratory approach



Analysis 1: Causality

Question: Does the CHG causal taxonomy accurately describe how each head affects task performance?

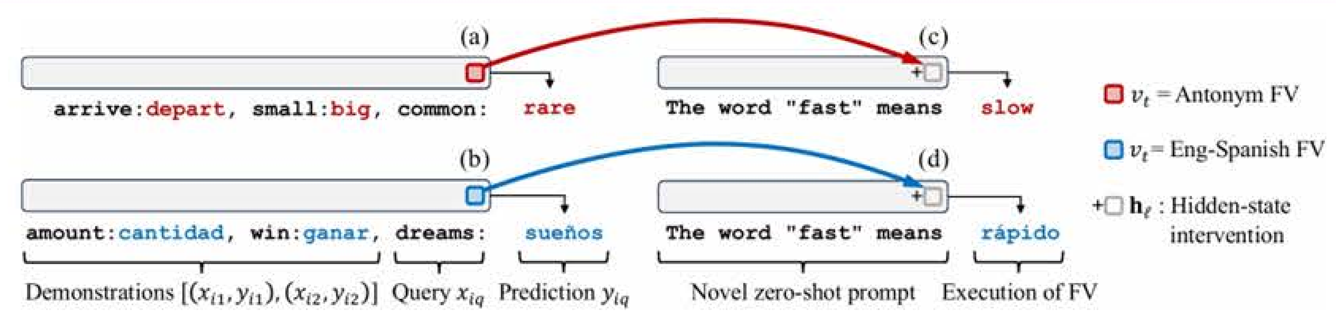
1. Fit CHG matrices on a task
2. For each metric, fully ablate one head at a time
3. Observe whether the ablations affect task performance as predicted

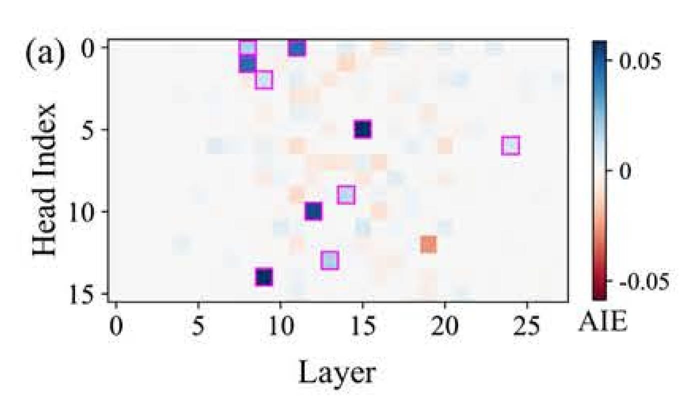


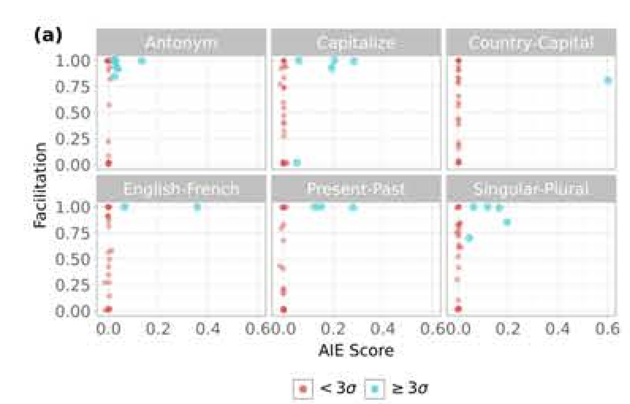
Analysis 2: Activation Patching

Question: Do CHG results corroborate with existing methods in the literature?

1. Fit CHG matrices on public datasets accompanying mechanistic interpretability papers that use activation-patching[3,4]
2. Confirm that attention heads with high mediation scores also have high task-facilitation scores







Analysis 3: Contrastive Causal Head Gating

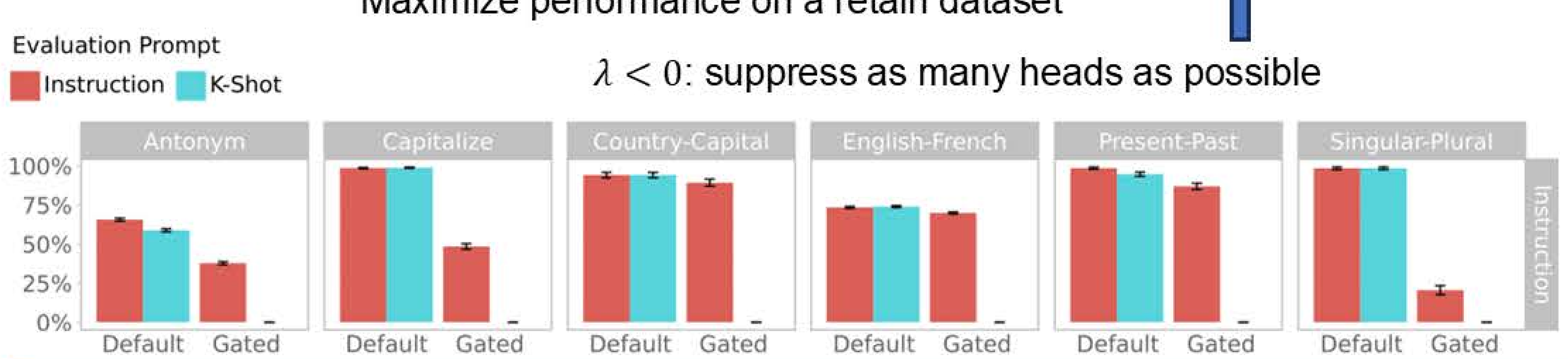
CHG reveals can reveal how important each head is, but not what it's used for. Contrastive CHG identifies sub-task circuits using **double dissociation**.

$$\mathcal{L}(G;\mathcal{M}_\theta,\lambda) = \sum_{(x_R,y_R,x_F,y_F)} \log P(y_F \mid x_F) - \log P(y_R \mid x_R) - \lambda \sum_{i,j} \sigma^{-1}(G_{l,h})$$

Minimize performance on a forget dataset

Maximize performance on a retain dataset

$\lambda < 0$: suppress as many heads as possible



Question: can we identify separable circuits for instruction-following and in-context learning?

1. Fit CCHG across 5 tasks to **retain** instruction-following (IF) but **forget** in-context learning (ICL)
2. Evaluate both IF and ICL on a held-out task
3. Verify that the model can do IF but not ICL on the held-out task

In-context learning

Q: old A: new
Q: undo A: do
Q: up A: down
…
Q: north A: south

Instruction following

Given an input word, generate the word with opposite meaning.
Q: north A: south

Method & Approach

What you need

1. Any transformer-based language model
2. Any text-based dataset
3. 5 to 30 minutes on a single GPU (per CHG run)

Question: Given a triangle ABC, find the sum of the interior angles.
Answer: In a triangle, the sum of the interior angles is always 180 degrees. This is a fundamental property of triangles. So, the sum of the interior angles in triangle ABC is $\boxed{180}$ degrees.

Optimization

$$\mathcal{L}(G;\mathcal{M}_\theta,\mathcal{D},\lambda) = -\underbrace{\sum_{(x,y)\in\mathcal{D}} \log P(y \mid x;\mathcal{M}_\theta,G)}_{\text{Negative log-likelihood (NLL)}} - \underbrace{\lambda\sum_{i,j}\sigma^{-1}(G_{l,h})}_{\text{Regularization}}$$

Maximize performance on a dataset

While

- If $\lambda > 0$: retaining as many heads as possible (G^+)
- If $\lambda < 0$: ablate as many heads as possible (G^-)

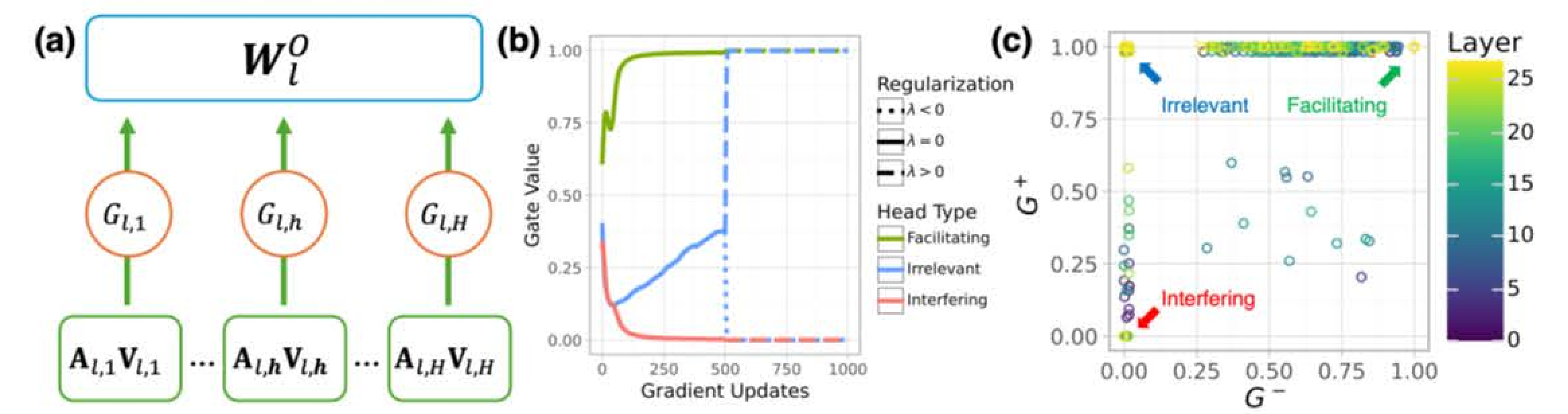


Table 1: Causal taxonomy for head roles and corresponding gating patterns.

Role	Description	G^+	G^-	Metric	Ablation Effect
Facilitating	Supports task performance	High	High	G^-	Decreases task performance
Interfering	Interferes with task performance	Low	Low	$1 - G^+$	Increases task performance
Irrelevant	Negligible impact on performance	High	Low	$G^+ \times (1 - G^-)$	No effect on task performance

References

1. Nelson Elhage et al.. A mathematical framework for transformer circuits. Transformer Circuits Thread, 2021. https://transformer-circuits.pub/2021/framework/index.html.
2. Kevin Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. arXiv preprint arXiv:2211.00593, 2022.
3. Yukang Yang, Declan Campbell, Kaixuan Huang, Mengdi Wang, Jonathan Cohen, and Taylor Webb. Emergent symbolic mechanisms support abstract reasoning in large language models. Forty-second International Conference on Machine Learning. 2025.
4. Eric Todd, Millicent L. Li, Arnab Sen Sharma, Aaron Mueller, Byron C. Wallace, and David Bau. "Function Vectors in Large Language Models." Proceedings of the 2024 International Conference on Learning Representations (ICLR 2024)