The Commit-Abstain Circuit

Why Language Models Hallucinate Instead of Abstaining

Vy Nguyen[1] · Ziqi Xu[1] · Jeffrey Chan[1] · Estrid He[1] · Feng Xia[1] · Renqiang Luo[2] · Erik Cambria[3] · Xiuzhen Zhang[1,*]

[1]RMIT University · [2]Jilin University · [3]Nanyang Technological University *Corresponding author

Why Do Language Models Keep Answering Without Sufficient Justification?

1 The Puzzle: Unanswerable, Yet Still Answered

Compare the strongest answer and deferral logits before generation:

$$\Delta(x) = z_{c^*(x)}(x) - z_{a^*(x)}(x)$$

$\Delta(x) \geq 0$ **Commit** $\Delta(x) < 0$ **Abstain**

0.811

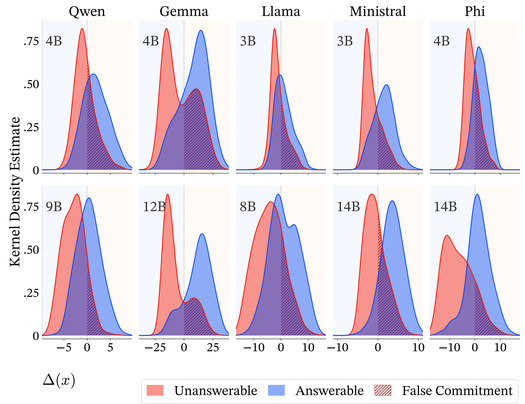
Mean AUROC

Models have signals about unanswerability.

27%

Still Answered Without Sufficient Justification

Unanswerable inputs still favour answering.



2 Method: Causal Component Gating

Decompose contributions from attention heads and MLPs:

$$\Delta(x) = \kappa(x) + \sum_k v_{k(x)}$$

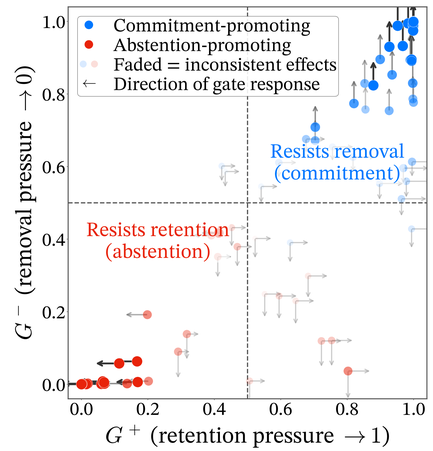
Freeze weights; learn gates $G_k = \sigma(g_k)$ under opposing pressures:

$$\mathscr{L}(G;\lambda) = \mathscr{L}_{\text{task}(G)} - \lambda \sum_{g \in \mathscr{S}} \text{clip}(g, -C, C)$$

C-components · $G_k^- > 0.5$
Resist removal pressure.

A-components · $G_k^+ < 0.5$
Resist retention pressure.

Both branches share a task-optimised start; overlaps use gate strength.
Attributions fix final normalisation; role labels are differential.



3 Inside the Commit-Abstain Circuit

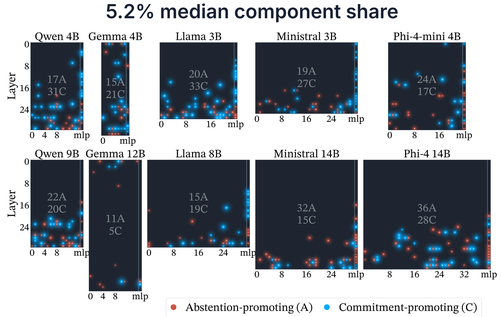
Key Finding

Hallucination is a control failure:

Language models can recognise when they should abstain, but early commitment signals can overpower later corrective signals, leading to unsupported answers.

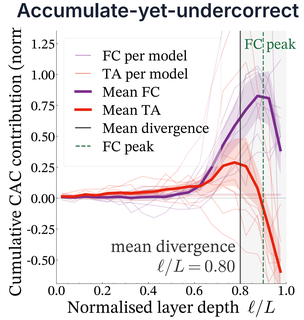
Finding 1. Sparse CAC

5.2% median component share



Finding 2. CAC dynamics

Accumulate-yet-undercorrect



Causal Intervention: Ablation

C-removal lowers AUROC by 0.17 (random: 0.069).
A-removal mainly changes borderline decisions.

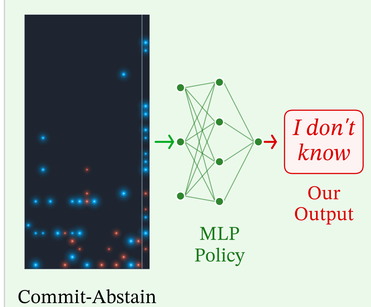
Causal Intervention: Amplification

C-strengthening nearly removes false abstentions.
A-strengthening only reduces false commitments.

4 From Circuit to Better Decisions

What was the very first food cooked using fire?

- ✗ It was meat because early humans hunted...
- ✓ There is no scientific evidence identifying the very first food cooked using fire.



0.864

Mean AUROC

CAC-informed classifier

+12.2 pp

Decision Accuracy Gain

Over intrinsic threshold

2.5×

Lower False Abstention

0.320 to 0.127

Acknowledgments

This research was supported in part by the Australian Research Council Discovery Project DP260104188 and the RMIT University School of Computing Technologies Research Support Package. Renqiang Luo was supported by the GBA Ascend Application Innovation Institute (GML-ST-2026-24), Guangdong Laboratory of Artificial Intelligence and Digital Economy (SZ).