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

s3964786@student.rmit.edu.au, cambria@ntu.edu.sg, lrenqiang@jlu.edu.cn
{ziqi.xu, jeffrey.chan, estrid.he, feng.xia, xiuzhen.zhang}@rmit.edu.au

[*]Corresponding author

github.com/vnht/commit-abstain-circuit

Abstract

Language models (LMs) often hallucinate by committing to confident answers rather than abstaining, even when they do not have enough information to answer reliably. A large body of existing work mitigates hallucination through detection or abstention mechanisms, but leaves open how models internally arrive at the decision to commit or abstain in the first place. We study this decision through mechanistic analysis, framing hallucination as unsupported commitment: the model commits despite exhibiting signals of unanswerability. Using causal gating, we identify a **C**ommit-**A**bstain **C**ircuit (CAC), a sparse, causally localised subset of attention heads and MLP sublayers underlying this decision. Across ten LMs (3B–14B) from five families and three benchmarks, the CAC exhibits a recurring *accumulate-yet-undercorrect* pattern: commitment-promoting components build up commitment in earlier layers, while abstention-promoting components act later as corrective signals that are often insufficient to overturn the accumulated commitment. Building on this finding, a lightweight policy trained on CAC activations improves decision accuracy by 12.2 points over the model's intrinsic commit-abstain margin, reduces false abstentions by 2.5 times, transfers to unseen benchmarks, and extends to larger models (27B–35B). The CAC is both diagnostic, clarifying how models overcommit, and practical, enabling improved abstention decisions.

1 Introduction

Reliable language models (LMs) need two capabilities: answering when they can, and abstaining when they should, for example by responding "I don't know" or indicating that the available information is insufficient [1, 2]. Abstention, however, remains an open challenge: large-scale evaluations show that even frontier models fail to abstain reliably across forms of *unanswerability* such as unknown futures, false premises, missing context, and underspecification, and that neither scaling nor extended reasoning reliably fixes this [3, 4, 5]. Training and evaluation regimes tend to reward guessing, pushing models toward hallucination rather than honest abstention [6].

A large body of work treats hallucination as a persistent limitation of LMs [7, 8, 9], emphasising detection and mitigation of hallucinations in model outputs [10, 11, 12, 13, 14, 15], or guiding and training models to abstain reliably [16, 17, 18]. These approaches mitigate hallucination at the behavioural level but leave unexplained how the model internally arrives at hallucination.

40th Conference on Neural Information Processing Systems (NeurIPS 2026).

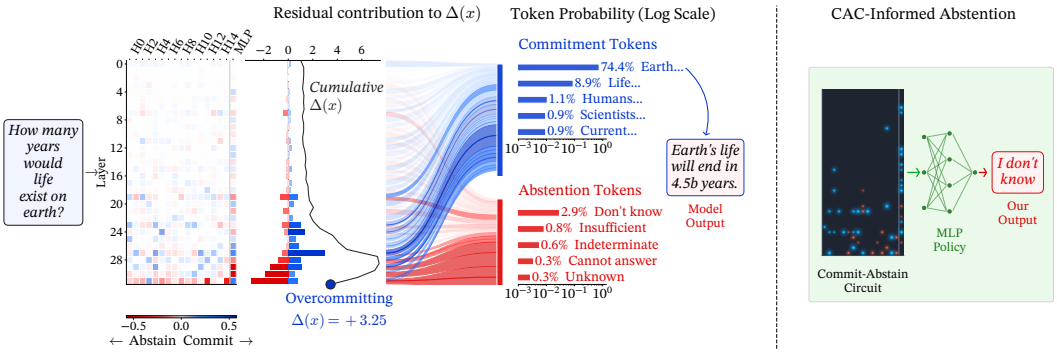


Figure 1: Unsupported commitment in Qwen 3.5 4B on an unanswerable KUQ question [27]. **Left:** Per-component contributions to the commit-abstain margin $\Delta(x)$ (Definition 1). Late-layer abstention signals fail to overturn earlier commitment ($\Delta(x) = +3.25$), leading to commitment. **Right:** A lightweight CAC-Informed MLP correctly triggers abstention.

Emerging work suggests that unanswerability information is already present in the model's internal states: linear probes decode it from hidden activations, and intervening on their directions changes abstention behaviour [19, 20, 21]. Mechanistic analyses identify entity-recognition circuits whose misfires produce hallucinations [22, 23]. However, neither line of work explains the model's failure to act: if unanswerability signals are already present, why does the model still answer confidently? Existing work studies representations and directions associated with hallucination, but does not localise which model components underlie the commit-abstain decision or account for how the model answers in the first place. We address this gap by asking how the model internally fails to produce abstention, framing hallucination as a control failure rather than a knowledge failure.

To this end, we define *unsupported commitment* as the model's failure to abstain despite exhibiting signals of unanswerability, producing a substantive answer rather than deferring with expressions such as "I don't know" or "The context is insufficient". We ask two questions: (RQ1) Does a sparse subset of model components causally contribute to the model's commit-abstain decision, and (RQ2) Can an understanding of this subset support a better abstention policy than the model's default behaviour?

We analyse the commit-abstain decision from the model's logits before any token is emitted, capturing its preference before autoregressive generation introduces path dependence [24, 25]. We quantify this preference as a *commit-abstain margin* and decompose it over attention heads and multilayer perceptron (MLP) sublayers. We adapt causal head gating [26] to the commit-abstain setting and identify a sparse, causally localised subset of the model's components that underlie this decision, termed the **C**ommit-**A**bstain **C**ircuit (CAC). Across ten LMs and three benchmarks, a recurring pattern emerges in the CAC: commitment-promoting components act in earlier layers and accumulate commitment, while abstention-promoting components act later as corrective signals that are often insufficient to overturn it. Figure 1 illustrates this pattern on an unanswerable input: the abstention signal arises too late to overturn the accumulated commitment, and the model commits. This pattern provides a component-level account of unsupported commitment and motivates a lightweight policy that improves abstention beyond the default behaviour of the model. Our contributions are as follows:

- **Problem Formulation.** We frame hallucination as *unsupported commitment*, a failure of control rather than of knowledge, and introduce a commit-abstain margin to quantify the model's commit-abstain preference.
- **Mechanistic Finding.** We identify the **C**ommit-**A**bstain **C**ircuit (CAC), a sparse, causally localised subset of the model's attention heads and MLP sublayers (median 5.2% of components) that underlie the commit-abstain margin, and reveal a recurring *accumulate-yet-undercorrect* pattern, providing a component-level account of unsupported commitment.
- **Policy Improvement.** We train a lightweight policy on CAC activations that improves decision accuracy by 12.2 points and reduces false abstentions by 2.5× across ten models (3B–14B) from five families and three benchmarks, with gains transferring to unseen benchmarks and reproduced on larger models (27B–35B).

2 Related Work

Hallucination Detection and Abstention Hallucination, the generation of confident but unsupported content, is a core reliability failure in LMs [28, 29]. Two broad lines of work address this: *detecting* hallucinations in outputs [10, 30] and *abstaining* when answering would likely hallucinate [2]. *Detection* methods span three main families: output-level uncertainty, confidence, or entropy estimation [10, 30, 31, 32]; evidence-based verification [33, 34, 35, 36]; and model-internal probing of hidden states, attention, or representations [37, 38, 11, 39]. *Abstention* methods split into training-free approaches such as selective answering [40, 41, 42], uncertainty control [43, 44, 45, 46, 47], consensus [16, 48, 49], and causal analysis [50, 18, 51, 52]; and training-based approaches using fine-tuning [53, 17, 54, 55, 56] or reinforcement learning [57, 58, 59]. Both lines treat hallucination as an established phenomenon to detect or mitigate, without explaining how the model internally arrives at hallucination. In contrast, we take a mechanism-first approach: we localise internal components that contribute to the commit-abstain decision, and use this understanding to inform improved abstention.

Mechanistic Interpretability Mechanistic work has identified sparse circuits underlying diverse model behaviours: induction heads copy from context [60], successor heads increment ordinals [61], distinct heads mediate contextual versus parametric knowledge [62, 63, 64, 65], and factual recall routes through MLP pathways [66, 67]. Closest to our setting is work on safety refusal, which is mediated by a sparse residual-stream sub-circuit [68, 69]. Our focus is different: epistemic abstention asks whether the model knows enough to answer, not whether it should refuse a harmful request. We adapt causal gating [26] to this setting, extending it to gate MLP sublayers alongside attention heads under a margin-based objective for the commit-abstain decision.

The Representation-Behaviour Gap Emerging work shows that models carry information that their generations fail to express. Probes and geometric analyses reveal signals of correctness, truthfulness, and unanswerability in hidden states [70, 71, 72, 73, 20], including signals present prior to generation [39, 25, 74, 24]. Recent work further investigates how such signals fail to translate into behaviour, including entity-recognition mechanisms whose misfiring produces hallucination [22, 23]. These findings suggest a representation-behaviour gap: information relevant to whether the model should answer is present internally, yet the model may still produce an answer. Prior work detects these signals but has not characterised the internal components that underlie the commit-abstain decision. In contrast, we identify a sparse subset of attention heads and MLP sublayers and show that their component-level contributions retain unanswerability information that is partially lost when aggregated into the final commit-abstain margin.

3 The Commit-Abstain Circuit

To characterise unsupported commitment, we need to measure the model's commit-abstain preference. We quantify this from the model's logits before autoregressive generation introduces path dependence. This design is supported by evidence that decision-relevant signals such as correctness and truthfulness are encoded prior to generation [73, 24, 25], and that such logits provide a reliable readout of model preferences [75, 76, 39].

We partition the vocabulary into an abstention set $\mathcal{A}$ and a commitment set $\mathcal{C}$. Tokens in $\mathcal{A}$ initiate deferral expressions, while $\mathcal{C}$ contains the remaining vocabulary. Table 1 provides examples of expressions initiated by each set. We describe the empirical construction of $\mathcal{A}$ in Section 4.1. Intuitively, $\Delta(x)$ compares the model's strongest immediate tendency to answer with its strongest tendency to abstain.

For input x, let $r_L(x) \in \mathbb{R}^d$ denote the residual state at the final input position. Following prior practice in direct logit attribution [66, 77], we absorb the final normalisation into the unembedding, so the per-token logit is $z_t(x) = u_t^\top r_L(x) + b_t$, where $u_t \in \mathbb{R}^d$ is the unembedding vector for token t and b_t is its bias (zero if absent). We then define the commit-abstain margin:

Definition 1 (Commit-Abstain Margin) *Let* $c^*(x) = \arg\max_{t \in \mathcal{C}} z_t(x)$ *and* $a^*(x) = \arg\max_{t \in \mathcal{A}} z_t(x)$ *denote the highest-logit tokens in $\mathcal{C}$ and $\mathcal{A}$. The commit-abstain margin is*

$$\Delta(x) = z_{c^*(x)}(x) - z_{a^*(x)}(x). \tag{1}$$

$\Delta(x) \geq 0$ *indicates a preference to commit;* $\Delta(x) < 0$ *indicates a preference to abstain.*

Table 1: Token sets used to define the commit-abstain margin $\Delta(x)$.

Abstention set $\mathcal{A}$	**Commitment set** $\mathcal{C}$
"Don't know", "Insufficient evidence", "Unsure", "Unanswerable", "Ambiguous", "Beyond my knowledge"	All remaining, e.g. those initiating: "According to...", "The answer is...", "Exactly", "Yes", "No", "In fact"

Remark 1 Let $Y \in \{0, 1\}$ be the gold answerability label. If $\mathrm{AUROC}(\Delta(X), Y) > \frac{1}{2}$, the conditional distributions $P(\Delta(X) \mid Y = 1)$ and $P(\Delta(X) \mid Y = 0)$ differ, so $I(\Delta(X); Y) > 0$, where $I(\cdot; \cdot)$ denotes mutual information. By the data-processing inequality applied to the deterministic map $r_L(X) \mapsto \Delta(X)$, $I(r_L(X); Y) \geq I(\Delta(X); Y) > 0$. The residual stream therefore carries information about unanswerability whenever $\Delta(X)$ separates classes above chance, and thus $\Delta(X)$ provides a tractable summary of this internal information.

To study this internal mechanism, we exploit the additive structure of the residual stream. The final residual state $r_L(x)$ can be written as a sum of contributions from all attention heads and MLP sublayers. Under this formulation, $\Delta(x)$ decomposes into signed per-component contributions: positive contributions push the margin toward commitment, while negative contributions push it toward abstention. Let $d(x) = u_{c^*(x)} - u_{a^*(x)}$ denote the direction along which the residual stream contributes to $\Delta(x)$. Then,

$$\Delta(x) = \kappa(x) + \sum_{k=1}^{K} v_k(x), \tag{2}$$

where $\kappa(x)$ collects bias and embedding terms, k indexes attention heads and MLP sublayers, $c_k(x) \in \mathbb{R}^d$ is the output activation of component k, and $v_k(x) = d(x)^\top c_k(x)$ is its signed contribution projected onto $d(x)$. Per-component contributions are computed under the standard linearisation that the final normalisation factor is fixed to its value at the ungated forward pass [66, 77].

The additive decomposition of $\Delta(x)$ suggests that the commit-abstain decision may be mediated by a subset of components, formalising RQ1. We define this subset as follows:

Definition 2 (Commit-Abstain Circuit) *The **C**ommit-**A**bstain **C**ircuit (CAC) is a causally localised subset of attention heads and MLP sublayers that mediate the commit-abstain preference encoded by* $\Delta(x)$.

We localise and characterise the CAC empirically in Section 4 and use it to design an improved abstention policy in Section 5.

4 Localisation of the Commit-Abstain Circuit

We address RQ1 by verifying that $\Delta(x)$ reflects signals of unanswerability, identifying the CAC as its component-level basis, and characterising how the CAC shapes the commit-abstain margin.

4.1 Setup for CAC Localisation

Datasets We use three benchmarks spanning parametric recall, in-context retrieval, and complex reasoning: **KUQ** [27], which contains open-domain questions with no definitive answer; **SQuAD** 2.0 [78], a reading comprehension dataset with adversarially constructed unanswerable contexts; and **MuSiQue** [79], which includes multi-hop questions with a supporting document removed. A detailed description of the datasets is provided in App. A. We sample $1{,}000$ instances from each dataset. All supervised experiments use a $50/50$ train-test split.

Models We evaluate ten instruction-tuned LMs spanning five model families and two parameter scales: Qwen-3.5 (4B, 9B) [80], Gemma-3 (4B, 12B) [81], Llama-3.2 (3B) and Llama-3.1 (8B) [82], Ministral-3 (3B, 14B) [83], and Phi-4-mini (4B) and Phi-4 (14B) [84]. All models use greedy decoding with `max_new_tokens` $= 256$ and are run on a NVIDIA H100 GPU.

Construction of $\mathcal{A}$ For each model, zero-shot inference is performed on a separate pool of $1{,}000$ unanswerable instances disjoint from the train and test splits, and the initial token of each response that

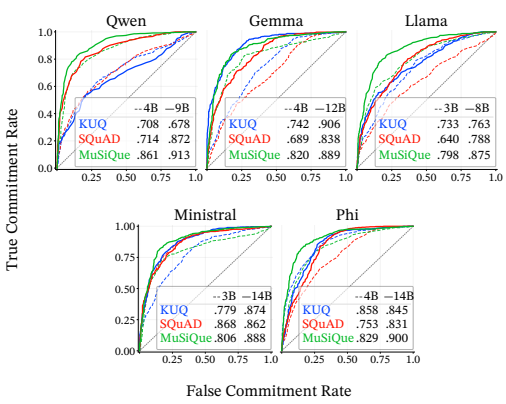


Figure 2: ROC curves for $\Delta(x)$ as a score for the answerable/unanswerable distinction across 30 configurations (AUROC in legend).

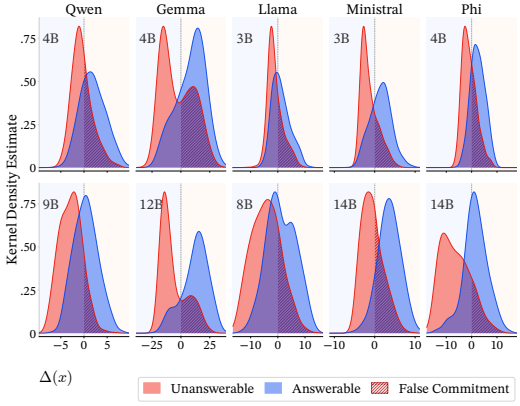


Figure 3: $\Delta(x)$ distributions for answerable and unanswerable inputs. Hatched regions indicate FC, where $\Delta(x) \geq 0$ on unanswerable instances.

matches a canonical deferral pattern is recorded. These tokens are pooled across models, deduplicated, disambiguated, and manually validated to ensure that they initiate epistemic deferral. We provide the full procedure, token lists, and a robustness analysis to the choice of $\mathcal{A}$ in App. B.

4.2 Unanswerability Signals in the Commit-Abstain Margin

We examine whether $\Delta(x)$ reflects signals of unanswerability above chance across model families and scales. For each instance x, we compute $\Delta(x)$ at the final input position. We evaluate $\Delta(x)$ as a score for the answerable/unanswerable distinction using AUROC, and report decision metrics under the threshold $\Delta(x) \geq 0$: decision accuracy (Acc), true commitment (TC), true abstention (TA), false abstention (FA), and false commitment (FC), our operational measure of unsupported commitment.

Figure 2 reports AUROC of $\Delta(x)$ across all 30 model-dataset configurations, and Figure 3 shows the distributions for answerable and unanswerable inputs. The sign of $\Delta(x)$ agrees with the model's actual commit-abstain behaviour from full generation in 97.6% of instances on average (range 95.8%–98.7%), supporting our formulation. Full per-configuration results are in App. C.

Finding A: The margin $\Delta(x)$ reflects reliable signals of unanswerability, yet unsupported commitment persists. All 30 AUROC values exceed chance, ranging from 0.640 (Llama 3B on SQuAD 2.0) to 0.913 (Qwen 9B on MuSiQue), with a mean of 0.811. Within each model family, larger models generally achieve higher AUROC. The distribution of $\Delta(x)$ on unanswerable instances is consistently left-shifted, yet a substantial mass remains above zero in every configuration (hatched regions in Figure 3). On average, 27% of unanswerable instances satisfy $\Delta(x) \geq 0$, ranging from 11% (Qwen 9B) to 41% (Gemma 4B). By Remark 1, since $\Delta(x)$ separates answerable from unanswerable inputs well above chance, the model's residual stream must carry information about unanswerability.

4.3 Localising the CAC via Causal Component Gating

To study the commit-abstain decision internally, we identify the CAC via causal component gating. We adapt Causal Head Gating (CHG) [26], which assigns causal roles to attention heads under opposing regularisation pressures. Unlike causal mediation analysis [66, 85, 86], which typically intervenes on components separately, CHG optimises all gates jointly, allowing interacting components to be identified together [26]. We extend CHG in two ways. First, we gate MLP sublayers alongside attention heads, motivated by prior work on MLP contributions to model behaviour [87, 88, 89]. Second, we replace the next-token objective with a loss over the commit-abstain margin.

Gate Parameterisation For a transformer with L layers and H attention heads per layer, we assign a learnable logit $g \in \mathbb{R}$ to each attention head and MLP sublayer, with the corresponding gate defined as $G = \sigma(g)$, where σ denotes the sigmoid function. The head gate scales the concatenated value vectors prior to the output projection W_ℓ^O, while the MLP gate scales the sublayer output before residual addition. All base model weights are kept frozen; only the gate logits are optimised.

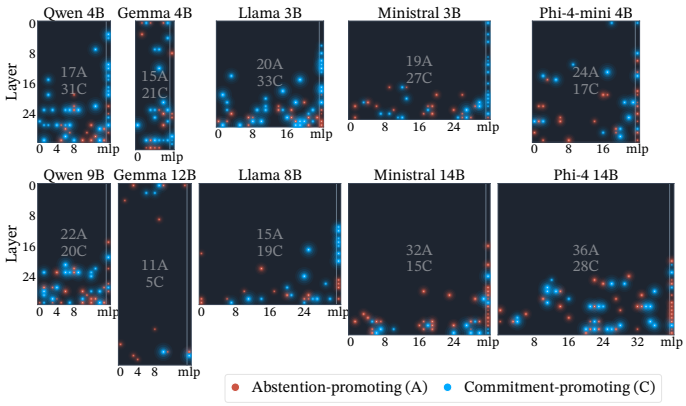


Figure 4: Identified CACs across ten models. Each panel plots layer (vertical) against head index with an additional MLP column (horizontal); dot brightness indicates mean causal strength across seeds. Inset numbers report the counts of abstention-promoting (A) and commitment-promoting (C) components per model.

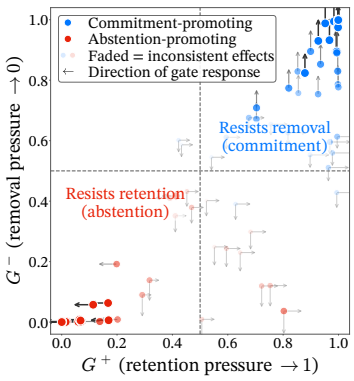


Figure 5: CAC identification via causal gating for Llama 3B. Each candidate component is plotted by its mean G^+ and G^- values across 10 seeds.

Task Loss Let $c^*(x)$ and $a^*(x)$ denote the top commit and abstain tokens under the ungated model, precomputed and held fixed so that gradients flow only through the gated logits. This fixing is a computational choice for gradient stability; the 97.6% agreement between token-level and generation-level labels (Section 4.2) indicates that this surrogate closely tracks commit-abstain behaviour. The task loss increases $\Delta(x)$ on answerable inputs and decreases it on unanswerable inputs:

$$\mathcal{L}_{\text{task}}(G) = -\frac{1}{B} \sum_{x \in \text{batch}} (2y-1) \cdot \Delta(x \mid G), \qquad \Delta(x \mid G) = z_{c^*(x)}(G) - z_{a^*(x)}(G), \tag{3}$$

where $y \in \{0, 1\}$ is the answerability label, $(2y-1) \in \{-1, +1\}$ its sign, and B is the batch size.

Gating Protocol Components that are important for the margin resist being turned off, whereas irrelevant components follow the applied regularisation pressure. Following Nam et al. [26], we add opposing regularisation pressures to $\mathcal{L}_{\text{task}}$:

$$\mathcal{L}(G; \lambda) = \mathcal{L}_{\text{task}}(G) - \lambda \sum_{g \in \mathcal{S}} \text{clip}(g, -C, C), \tag{4}$$

where λ is the regularisation strength, C is the clipping bound, and $\mathcal{S}$ indexes the gated components.

We fit the gates in three phases:

- **1. Task optimisation:** With $\lambda = 0$, we optimise the task loss to obtain an initialisation $\hat{G}$.
- **2a. Retention pressure:** With $\lambda > 0$, regularisation pushes gates towards 1, yielding G^+.
- **2b. Removal pressure:** With $\lambda < 0$, regularisation pushes gates towards 0, yielding G^-.

Phases 2a and 2b are both initialised from $\hat{G}$, so their differences reflect only the direction of regularisation.

The task-loss gradient on each gate logit is proportional to $-(\mathbb{E}_{y=1}[v_k(x)] - \mathbb{E}_{y=0}[v_k(x)])$ (derivation in App. D.1). Thus, the loss tends to open gates for components that contribute more to $\Delta(x)$ on answerable inputs, and close gates for components that contribute more on unanswerable inputs. The resulting classification therefore reflects *differential* contribution between the two subsets rather than the absolute sign of v_k.

A component is *commitment-promoting* (C-component) if $G^- > \tau$, meaning that the task loss keeps it open despite removal pressure. A component is *abstention-promoting* (A-component) if $G^+ < \tau$, meaning that the task loss keeps it closed despite retention pressure. We set $\tau = 0.5$, the sigmoid midpoint. Components meeting either criterion are classified as A-components if $\mathbb{E}[1-G^+] > \mathbb{E}[G^-]$, and as C-components otherwise; those meeting neither are excluded. Full hyperparameter details are provided in App. D.2.

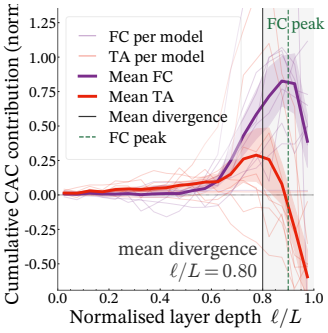


Figure 6: Cumulative CAC contribution to $\Delta(x)$ by layer depth (ℓ/L). FC and TA both accumulate positive contributions until 80% depth, after which TA reverses sharply.

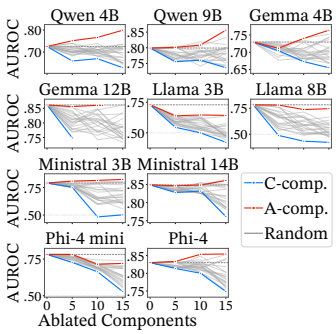


Figure 7: Progressive ablation of C-components and A-components on held-out data against 20 random baselines (grey). Components are removed in order of causal score.

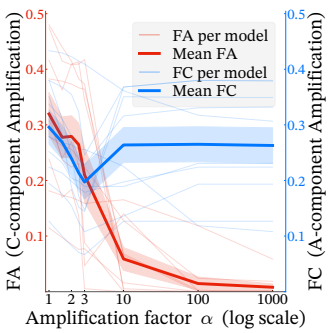


Figure 8: C-component amplification drives FA to near zero by amplification factor $\alpha = 10$; A-component amplification leaves FC largely unchanged even at $\alpha = 1000$.

4.4 Identifying the CAC

We report the main findings here; full results, including screened candidates, per-model CACs, and cross-seed statistics, are provided in Appendix D.3.

Finding B: Every tested model contains a sparse CAC. Figure 4 shows the identified CAC across all ten models. The CAC is sparse, covering 2.0% (Gemma 12B) to 11.8% (Gemma 4B) of components (median 5.2%). Both attention heads and MLP sublayers participate: on average, 25% of CAC components are MLPs (range 12–34%), validating the extension beyond head-only gating. The three-phase gating protocol indicates that these components causally influence the commit-abstain margin. Components separate into three groups: those that remain closed under retention pressure (A-components), those that remain open under removal pressure (C-components), and those that follow the applied pressure (irrelevant). As shown in Figure 5 for Llama 3B, this separation is clear. Across models, 80% of G^+ and 78% of G^- values lie below 0.1 or above 0.9, suggesting that component roles are resolved cleanly.

Finding C: Commitment generally emerges in earlier layers than abstention. Whether the model ultimately commits or abstains, both FC and TA trajectories accumulate positive contributions until 80% depth; only after this point does TA drop sharply negative while FC continues to rise, peaking at 90% before also dropping (Figure 6). At the component level, per A-component seen, more than $2\times$ as many C-components have appeared by 60%–70% depth in all models, $1.7\times$ at 80%, and only near parity by 90–100% (App. Table 9). Both results suggest that C-components are generally more concentrated in earlier layers than A-components. The CAC is generally stable across random initialisations, screening thresholds, and regularisation strengths, and largely reproduces the sign of $\Delta(x)$, validating the identified components as robust contributors to the commit-abstain margin; full sensitivity analyses are provided in App. D.4.

4.5 Functional Analysis of the CAC

We verify the functional roles of the identified CAC components through two interventions: progressive ablation and targeted amplification. For ablation, we progressively remove the top-k components on the test set, ordered by gating score, controlling against 20 size-matched random non-CAC ablations to verify that observed effects are CAC-specific. For amplification, we scale each component output direction h by $\alpha \in \{1.0, 1.5, 2.0, 2.5, 3.0, 10, 100, 1000\}$ before residual addition to test whether each type causally contributes to FA and FC rates. Full results are in App. D.5.

Finding D: C-components sustain the discriminative structure of $\Delta(x)$; A-components act correctionally but with insufficient intrinsic strength to overcome accumulated commitment. Ablating A-components causes 17.1% abstain-to-commit (A→C) flips versus 7.8% commit-to-abstain (C→A), yet AUROC remains stable (mean -0.03) as flipped inputs are borderline, shifting from mildly abstaining ($\bar{\Delta} = -2.78$) to committing ($\bar{\Delta} = +3.99$) with rank largely preserved. Ablating

C-components, counterintuitively, also yields A→C-skewed flips (31.7% vs. 3.4%). AUROC drops in every model (mean -0.17; worst: Qwen 4B $0.73 \rightarrow 0.40$), continuing to decline as more components are removed while size-matched random ablation changes it by only -0.069 (Figure 7). Both ablations skew A→C, but for distinct reasons: A-components push $\Delta(x)$ correctionally across 0 with little effect on ranking structure; C-components sustain the discriminative structure of $\Delta(x)$, their removal collapsing the separation between answerable and unanswerable inputs globally.

Amplification further validates these roles. At moderate α ($\leq$3), C-component amplification shifts decisions toward commitment, reducing FA from 0.320 to 0.210 while increasing FC from 0.296 to 0.418. Conversely, A-component amplification shifts decisions toward abstention, reducing FC from 0.296 to 0.197 while increasing FA from 0.320 to 0.415. This pattern holds across all 10 models, with mean FA and FC reductions of 37.6% and 33.9%, respectively. At extreme α ($\geq$10), the two sweeps diverge sharply: FA drops to near zero under C-component amplification, whereas FC stabilises well above zero under A-component amplification, with FC at α=100 nearly identical to α=1000 (Figure 8).

This divergence can be understood through the behaviour of Root Mean Square Normalisation (RMSNorm) under large scaling:

$$\text{RMSNorm}(R+\alpha h) = \frac{R+\alpha h}{\text{RMS}(R+\alpha h)}\cdot\gamma \approx \frac{\alpha h}{\text{RMS}(\alpha h)}\cdot\gamma = \frac{h}{\text{RMS}(h)}\cdot\gamma = \hat{h}, \quad \text{when } \alpha \gg \frac{\|R\|}{\|h\|},$$

where R is the residual stream and γ is a learned vector fixed at inference time. Intuitively, when α is sufficiently large, both α and the pre-existing residual R cease to dominate the normalised representation: the next layer is driven primarily by the component direction $\hat{h}$. Extreme amplification therefore probes the component's intrinsic downstream effect with prior residual accumulation largely normalised out.

The resulting asymmetry is informative. A C-component direction reliably drives $\Delta(x)$ upward and can nearly eliminate false abstention under strong amplification. In contrast, an A-component direction shifts decisions toward abstention at moderate amplification but cannot eliminate false commitment even under extreme scaling. Thus, C-components provide a strong commitment-driving signal, whereas A-components act as corrective signals that are not independently sufficient to guarantee abstention.

Remark 2 The model encodes unanswerability internally (Finding A) and contains a sparse CAC (Finding B), yet may still commit because abstention-promoting contributions arise later (Finding C) and are often insufficient to drive $\Delta(x)$ below zero (Finding D). These findings suggest a component-level account of unsupported commitment: the model commits not because abstention-related signals are absent, but because their corrective effects fail to overturn accumulated commitment.

5 Learning an Abstention Policy from the CAC

We now address RQ2 by testing whether the CAC is actionable beyond characterising the commit-abstain decision. To this end, we evaluate a lightweight abstention policy trained on CAC activations.

5.1 CAC-Informed Abstention Policy

For each input x, we extract the contribution $v_k(x)$ of each CAC component to $\Delta(x)$ at the final input position, and append aggregate statistics over the CAC vector (mean, standard deviation, min, max, and summed contributions of A- and C-components). Concatenating yields a d_{in}-dimensional feature vector, where d_{in} is 16–64 plus statistics.

We train a lightweight MLP with architecture $d_{\text{in}} \rightarrow 32 \rightarrow 128 \rightarrow 64 \rightarrow 32 \rightarrow 1$. The initial bottleneck projects variable-dimensional CAC features into a fixed 32-dimensional representation before non-linear feature mixing, providing dimensionality normalisation and regularisation. All hidden layers use batch normalisation, ReLU activations, and dropout (p=0.15). We train with binary cross-entropy using AdamW (lr=10^{-3}, wd=10^{-3}) for up to 500 epochs with early stopping (patience 120). At inference, the model outputs $P(\text{answerable})$, and abstention is triggered when this probability falls below a threshold calibrated on a validation set held out from the training split.

Table 2: Mean AUROC and accuracy across three datasets for each model. **Bold** indicates the best result; underline indicates the second best. p-values are from two-sided Wilcoxon signed-rank tests over 30 model-dataset pairs. All differences are significant ($p < 0.05$), with most at $p < 10^{-5}$.

		Qwen 3.5		Gemma 3		Llama 3		Ministral		Phi-4		
		4B	9B	4B	12B	3B	8B	3B	14B	mini	14B	p
Mean AUROC	Zero-Threshold	.725	.798	.722	.854	.689	.784	.790	.848	.780	.830	6.0×10^{-6}
	Non-CAC Random	.691	.764	.741	.808	.648	.815	.732	.789	.756	.822	8.4×10^{-6}
	INSIDE	.572	.449	.586	.658	.578	.483	.406	.493	.415	.492	3.7×10^{-9}
	Multi-LLM	.688	.757	.683	.728	.538	.705	.596	.654	.603	.703	1.9×10^{-9}
	Semantic Entropy	.567	.463	.548	.600	.564	.521	.423	.564	.509	.529	1.3×10^{-8}
	MERA	.675	.574	.635	.657	.613	.560	.648	.747	.634	.532	1.9×10^{-9}
	HaloScope	.697	.627	.719	.693	.780	.744	.665	.792	.689	.700	1.9×10^{-8}
	HaMI	.876	**.919**	.761	.833	.754	.845	.802	.874	**.838**	.878	2.0×10^{-3}
	CAC-Informed MLP (Ours)	**.879**	.911	**.807**	**.889**	**.807**	**.874**	**.874**	**.877**	.837	**.888**	–
Mean Accuracy	Zero-Threshold	.655	.697	.644	.770	.623	.651	.714	.753	.705	.711	2.6×10^{-6}
	Non-CAC Random	.628	.724	.682	.744	.596	.731	.670	.718	.693	.752	4.2×10^{-6}
	INSIDE	.571	.523	.601	.708	.564	.538	.515	.539	.515	.543	1.9×10^{-6}
	Multi-LLM	.673	.691	.673	.727	.597	.693	.655	.678	.632	.661	1.7×10^{-6}
	Semantic Entropy	.576	.506	.561	.602	.555	.539	.500	.558	.537	.555	1.9×10^{-6}
	MERA	.640	.592	.588	.646	.573	.580	.621	.665	.597	.565	1.9×10^{-9}
	HaloScope	.595	.655	.685	.653	.708	.715	.663	.691	.691	.671	1.7×10^{-6}
	HaMI	.822	**.881**	.711	.785	.707	.803	.740	**.849**	**.786**	.831	1.6×10^{-2}
	CAC-Informed MLP (Ours)	**.829**	.855	**.765**	**.845**	**.755**	**.820**	**.815**	.829	.785	**.840**	–

5.2 Experiments & Results

Baselines We compare against **Zero-Threshold**, which abstains when $\Delta(x) < 0$; a size-matched **Non-CAC Random** MLP that controls for whether the gain is CAC-specific; and six established methods. **INSIDE** [38] computes EigenScores from the covariance of multiple sampled response embeddings. **Multi-LLM** [16] abstains based on cross-LLM agreement. **Semantic Entropy** [10] measures entropy over semantically clustered generations. **MERA** [90] is an activation-steering framework with calibrated abstention. **HaloScope** [11] trains a binary classifier on a hallucination subspace identified via SVD on hidden states. **HaMI** [39] is an MLP detector trained via Multiple Instance Learning over adaptively-selected token activations. While we also use an MLP, our method differs from HaMI in feature space (16–64 CAC components vs. full d_{model}-dimensional activations) and selection principle (localisation of model components vs. selection of hallucination-indicative tokens). For comparability, all baselines are trained or calibrated using the same training split as our method (**CAC-Informed MLP**). Detailed configurations are in App. E.1.

Main Results Table 2 reports the main results on the test split, with per-dataset metrics in App. E.2. Our policy achieves the highest mean AUROC (0.864) and decision accuracy (0.814) across all 30 configurations, with substantial gains over the Zero-Threshold baseline ($+8.2$ AUROC points, $+12.2$ accuracy points). It also outperforms HaMI on 23 of 30 configurations in AUROC and 20 of 30 in accuracy. Notably, these improvements do not come at the cost of over-abstention: the mean false abstention rate is reduced from 0.320 to 0.127 ($2.5\times$ lower). These results suggest that the per-component CAC features carry richer information prior to additive aggregation in the residual stream: while $\Delta(x)$ collapses this aggregation into a scalar, the policy retains the per-component contributions and models nonlinear interactions across components.

Out-of-Distribution (OOD) Generalisation We evaluate generalisation along two axes: (i) extension to larger models (Gemma-3 27B and Qwen-3.5 35B), and (ii) transfer to unseen datasets, HotpotQA [91] and SelfAware [92]. For dataset transfer, the trained policy is applied without fine-tuning. For larger models, we localise the CAC and train the same MLP architecture on the in-distribution datasets before evaluating on unseen datasets. Table 3 reports results across all 12 models and both unseen datasets. The CAC-informed policy improves over Zero-Threshold in all 24 OOD configurations, with mean gains of $+6.9$ accuracy points. On the two larger models, the policy gains $+6.3$ accuracy points on average. Notably, it also outperforms the strongest baseline HaMI

Table 3: Decision accuracy on two unseen datasets (HotpotQA, SelfAware) and two unseen larger models (†). Our method achieves higher accuracy in all 24 configurations ($p = 1.8\times10^{-5}$).

	Qwen 3.5			Gemma 3			Llama 3		Ministral		Phi-4	
	4B	9B	35B†	4B	12B	27B†	3B	8B	3B	14B	mini	14B
HotpotQA												
Zero-Threshold	.887	.837	.867	.898	.922	.924	.840	.820	.851	.892	.859	.842
HaMI	.917	.938	.925	.894	.915	.892	.829	**.924**	.865	.889	.856	.937
CAC-Informed MLP (Ours)	**.930**	**.949**	**.928**	**.911**	**.948**	**.950**	**.859**	.913	**.895**	**.894**	**.875**	**.939**
SelfAware												
Zero-Threshold	.639	.544	.620	.637	.711	.747	.660	.655	.698	.709	.717	.697
HaMI	**.761**	**.749**	.649	.682	.755	.734	.675	**.782**	**.773**	.738	.713	.767
CAC-Informed MLP (Ours)	.758	.746	**.773**	**.737**	**.763**	**.757**	**.734**	.764	.769	**.799**	**.748**	**.779**

in 19 of 24 configurations. These results suggest that the localisation pipeline reproduces at larger scales, and that the policy generalises zero-shot across datasets.

We evaluate a logistic regressor over the same features, achieving $+10.6$ accuracy points over Zero-Threshold (vs. $+12.2$ for the MLP, App. E.3), indicating that much of the gain comes from directly reading CAC features, with the MLP capturing additional nonlinear interactions across components. We include two case studies of unsupported commitment in App. F.

6 Conclusion

Unsupported commitment reflects a representation-behaviour gap: models carry signals of unanswerability internally yet still commit. By formalising the commit-abstain margin and applying causal gating, we identified a sparse Commit-Abstain Circuit in every tested model and uncovered a recurring pattern: commitment-promoting components act earlier and accumulate momentum, while abstention-promoting components arise later as corrective signals that often fail to overturn what has already been built up. This provides an intuitive mechanistic account of how unsupported commitment can emerge even when abstention-related signals are present internally. More broadly, this suggests new directions for mitigation. Rather than only detecting hallucinations after they emerge or training models to abstain more often, internal mechanistic signals may enable earlier intervention and more targeted inference-time control over when the model should commit or defer. In this sense, the CAC offers not only an explanation of unsupported commitment, but also a foundation for more reliable model behaviour.

Limitations Several limitations should be considered when interpreting these findings. First, our scope is restricted to unsupported commitment on unanswerable inputs in instruction-tuned models; we do not address incorrect answers to otherwise answerable questions. Second, our main analysis uses the pre-generation commit-abstain margin. Although this agrees with full-generation behaviour in 97.6% of cases and CAC signals track observed mid-generation reversals, generation-time dynamics remain only partially characterised. Third, both the abstention readout and the CAC provide partial operationalisations of the underlying behaviour. Recent work has emphasised falsification-based evaluation to distinguish genuine behavioural mechanisms from superficial lexical correlates [93]. In our setting, $\mathcal{A}$ captures common lexical forms of epistemic deferral but may not cover all abstention behaviours, while the CAC is a causally localised, group-level account rather than a complete functional characterisation of every component or interaction. Future work should further stress-test these claims across broader behavioural settings.

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).

References

[1] Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. Large language models must be taught to know what they don't know. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 85932–85972, 2024. doi: 10.52202/079017-2729. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/9c20f16b05f5e5e70fa07e2a4364b80e-Paper-Conference.pdf.

[2] Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models. *Transactions of the Association for Computational Linguistics*, 13:529–556, 2025. doi: 10.1162/tacl_a_00754. URL https://aclanthology.org/2025.tacl-1.26/.

[3] Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel Bell. Abstentionbench: Reasoning LLMs fail on unanswerable questions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*, 2025. URL https://openreview.net/forum?id=0kHC30LLp0.

[4] Aashiq Muhamed, Leonardo F. R. Ribeiro, Markus Dreyer, Virginia Smith, and Mona T. Diab. RefusalBench: Generative evaluation of selective refusal in grounded language models. In *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 6811–6856, 2026. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.321. URL https://aclanthology.org/2026.eacl-long.321/.

[5] Xiangyu Peng, Prafulla Kumar Choubey, Caiming Xiong, and Chien-Sheng Wu. Unanswerability evaluation for retrieval augmented generation. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8452–8472, 2025. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.415. URL https://aclanthology.org/2025.acl-long.415/.

[6] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653(8116):1047–1051, April 2026. ISSN 1476-4687. doi: 10.1038/s41586-026-10549-w. URL http://dx.doi.org/10.1038/s41586-026-10549-w.

[7] Gaurang Sriramanan, Siddhant Bharti, Vinu Sankar Sadasivan, Shoumik Saha, Priyatham Kattakinda, and Soheil Feizi. Llm-check: Investigating detection of hallucinations in large language models. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 34188–34216. Curran Associates, Inc., 2024. doi: 10.52202/079017-1077. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3c1e1fdf305195cd620c118aaa9717ad-Paper-Conference.pdf.

[8] Andrew Jesson, Nicolas Beltran-Velez, Quentin Chu, Sweta Karlekar, Jannik Kossen, Yarin Gal, John P. Cunningham, and David Blei. Estimating the hallucination rate of generative ai. In A. Globerson, L. Mackey, D. Belgrave, A. Fan, U. Paquet, J. Tomczak, and C. Zhang, editors, *Advances in Neural Information Processing Systems*, volume 37, pages 31154–31201. Curran Associates, Inc., 2024. doi: 10.52202/079017-0982. URL https://proceedings.neurips.cc/paper_files/paper/2024/file/3791f5fc0e8e43730466afd2bcdb7493-Paper-Conference.pdf.

[9] Yejin Bang, Ziwei Ji, Alan Schelten, Anthony Hartshorn, Tara Fowler, Cheng Zhang, Nicola Cancedda, and Pascale Fung. HalluLens: LLM hallucination benchmark. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24128–24156, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1176. URL https://aclanthology.org/2025.acl-long.1176/.

[10] Sebastian Farquhar, Jannik Kossen, Lorenz Kuhn, and Yarin Gal. Detecting hallucinations in large language models using semantic entropy. *Nature*, 630(8017):625–630, June 2024. ISSN 1476-4687. doi: 10.1038/s41586-024-07421-0. URL http://dx.doi.org/10.1038/s41586-024-07421-0.

[11] Xuefeng Du, Chaowei Xiao, and Yixuan Li. Haloscope: Harnessing unlabeled LLM generations for hallucination detection. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=nfK0ZXFFSn.

[12] Xiaoling Zhou, Mingjie Zhang, Zheng Lee, Wei Ye, and Shikun Zhang. Hademif: Hallucination detection and mitigation in large language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=VwOYxPScxB.

[13] Bo Zhao, Yinghao Zhang, Ziqi Xu, Yongli Ren, Xiuzhen Zhang, Renqiang Luo, Zaiwen Feng, and Feng Xia. Unbiased reasoning for knowledge-intensive tasks in large language models via conditional front-door adjustment. In *Proceedings of the 34th ACM International Conference on Information and Knowledge Management, CIKM 2025*, pages 4315–4325, 2025. doi: 10.1145/3746252.3761103. URL https://doi.org/10.1145/3746252.3761103.

[14] Bowen Li, Ziqi Xu, Jing Ren, Renqiang Luo, Xikun Zhang, Xiuzhen Zhang, Yongli Ren, and Feng Xia. Debiasing large language models via adaptive causal prompting with sketch-of-thought. In *Findings of the Association for Computational Linguistics: EACL 2026*, pages 4481–4499, 2026. doi: 10.18653/V1/2026.FINDINGS-EACL.234. URL https://doi.org/10.18653/v1/2026.findings-eacl.234.

[15] Jing Ren, Wenhao Zhou, Bowen Li, Mujie Liu, Nguyen Linh Dan Le, Jiade Cen, Liping Chen, Ziqi Xu, Xiwei Xu, and Xiaodong Li. Causal prompting for implicit sentiment analysis with large language models. *IEEE Transactions on Computational Social Systems*, 13(3):4131–4143, 2026. doi: 10.1109/TCSS.2026.3661211. URL https://doi.org/10.1109/TCSS.2026.3661211.

[16] Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.786. URL https://aclanthology.org/2024.acl-long.786/.

[17] Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. R-tuning: Instructing large language models to say 'I don't know'. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 7113–7139, 2024. doi: 10.18653/v1/2024.naacl-long.394. URL https://aclanthology.org/2024.naacl-long.394/.

[18] Vy Nguyen, Ziqi Xu, Jeffrey Chan, Estrid He, Feng Xia, and Xiuzhen Zhang. Hallucinate less by thinking more: Aspect-based causal abstention for large language models. In *Fortieth AAAI Conference on Artificial Intelligence, AAAI 2026*, pages 32555–32563, 2026. doi: 10.1609/AAAI.V40I38.40532. URL https://doi.org/10.1609/aaai.v40i38.40532.

[19] Maor Juliet Lavi, Tova Milo, and Mor Geva. Detecting (un)answerability in large language models with linear directions. In Vera Demberg, Kentaro Inui, and Lluís Marquez, editors, *Proceedings of the 19th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 682–699, Rabat, Morocco, March 2026. Association for Computational Linguistics. ISBN 979-8-89176-380-7. doi: 10.18653/v1/2026.eacl-long.29. URL https://aclanthology.org/2026.eacl-long.29/.

[20] Adi Simhi, Itay Itzhak, Fazl Barez, Gabriel Stanovsky, and Yonatan Belinkov. Trust me, I'm wrong: LLMs hallucinate with certainty despite knowing the answer. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Findings of*

the Association for Computational Linguistics: EMNLP 2025, pages 14665–14688, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-335-7. doi: 10.18653/v1/2025.findings-emnlp.792. URL https://aclanthology.org/2025.findings-emnlp.792/.

[21] Lei Yu, Meng Cao, Jackie CK Cheung, and Yue Dong. Mechanistic understanding and mitigation of language model non-factual hallucinations. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 7943–7956, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.466. URL https://aclanthology.org/2024.findings-emnlp.466/.

[22] Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=WCRQFlji2q.

[23] Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. *Transformer Circuits Thread*, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

[24] Anum Afzal, Florian Matthes, Gal Chechik, and Yftah Ziser. Knowing before saying: LLM representations encode information about chain-of-thought success before completion. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 12791–12806, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.662. URL https://aclanthology.org/2025.findings-acl.662/.

[25] Qinyu Zhao, Ming Xu, Kartik Gupta, Akshay Asthana, Liang Zheng, and Stephen Gould. The first to know: How token distributions reveal hidden knowledge in large vision-language models? In *Computer Vision – ECCV 2024: 18th European Conference, Milan, Italy, September 29–October 4, 2024, Proceedings, Part XLVIII*, page 127–142, Berlin, Heidelberg, 2024. Springer-Verlag. ISBN 978-3-031-73194-5. doi: 10.1007/978-3-031-73195-2_8. URL https://doi.org/10.1007/978-3-031-73195-2_8.

[26] Andrew Joohun Nam, Henry Conklin, Yukang Yang, Thomas L. Griffiths, Jonathan D. Cohen, and Sarah-Jane Leslie. Causal head gating: A framework for interpreting roles of attention heads in transformers. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=kgmyjyDFrx.

[27] Alfonso Amayuelas, Kyle Wong, Liangming Pan, Wenhu Chen, and William Yang Wang. Knowledge of knowledge: Exploring known-unknowns uncertainty with large language models. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 6416–6432, 2024. doi: 10.18653/v1/2024.findings-acl.383. URL https://aclanthology.org/2024.findings-acl.383/.

[28] Lei Huang, Weijiang Yu, Weitao Ma, Weihong Zhong, Zhangyin Feng, Haotian Wang, Qianglong Chen, Weihua Peng, Xiaocheng Feng, Bing Qin, and Ting Liu. A survey on hallucination in large language models: Principles, taxonomy, challenges, and open questions. *ACM Transactions on Information Systems*, 43(2):1–55, January 2025. ISSN 1558-2868. doi: 10.1145/3703155. URL http://dx.doi.org/10.1145/3703155.

[29] Aisha Alansari and Hamzah Luqman. Large language models hallucination: A comprehensive survey. *Computer Science Review*, 61:100970, August 2025. ISSN 1574-0137. doi: 10.1016/j.cosrev.2026.100970. URL http://dx.doi.org/10.1016/j.cosrev.2026.100970.

[30] Potsawee Manakul, Adian Liusie, and Mark Gales. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.557. URL https://aclanthology.org/2023.emnlp-main.557/.

[31] Kamil Ciosek, Nicolò Felicioni, and Sina Ghiassian. Hallucination detection on a budget: Efficient bayesian estimation of semantic entropy. *Transactions on Machine Learning Research*, 2025. ISSN 2835-8856. URL https://openreview.net/forum?id=j2N2RuNdbC.

[32] Yuyan Chen, Qiang Fu, Yichen Yuan, Zhihao Wen, Ge Fan, Dayiheng Liu, Dongmei Zhang, Zhixu Li, and Yanghua Xiao. Hallucination detection: Robustly discerning reliable answers in large language models. In *Proceedings of the 32nd ACM International Conference on Information and Knowledge Management*, CIKM '23, page 245–255. ACM, October 2023. doi: 10.1145/3583780.3614905. URL http://dx.doi.org/10.1145/3583780.3614905.

[33] Junyi Li, Jie Chen, Ruiyang Ren, Xiaoxue Cheng, Xin Zhao, Jian-Yun Nie, and Ji-Rong Wen. The dawn after the dark: An empirical study on factuality hallucination in large language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10879–10899, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.586. URL https://aclanthology.org/2024.acl-long.586/.

[34] Xiangkun Hu, Dongyu Ru, Lin Qiu, Qipeng Guo, Tianhang Zhang, Yang Xu, Yun Luo, Pengfei Liu, Yue Zhang, and Zheng Zhang. Knowledge-centric hallucination detection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 6953–6975, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.emnlp-main.395. URL https://aclanthology.org/2024.emnlp-main.395/.

[35] Shenling Liu, Yang Gao, ShaSha Li, PanCheng Wang, and Ting Wang. A hallucination detection and mitigation framework for faithful text summarization using llms. *Scientific Reports*, 16(1), December 2025. ISSN 2045-2322. doi: 10.1038/s41598-025-31075-1. URL http://dx.doi.org/10.1038/s41598-025-31075-1.

[36] Cheng Niu, Yuanhao Wu, Juno Zhu, Siliang Xu, KaShun Shum, Randy Zhong, Juntong Song, and Tong Zhang. RAGTruth: A hallucination corpus for developing trustworthy retrieval-augmented language models. In Lun-Wei Ku, Andre Martins, and Vivek Srikumar, editors, *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 10862–10878, Bangkok, Thailand, August 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.acl-long.585. URL https://aclanthology.org/2024.acl-long.585/.

[37] Zhenliang Zhang, Xinyu Hu, Huixuan Zhang, Junzhe Zhang, and Xiaojun Wan. ICR probe: Tracking hidden state dynamics for reliable hallucination detection in LLMs. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 17986–18002, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.880. URL https://aclanthology.org/2025.acl-long.880/.

[38] Chao Chen, Kai Liu, Ze Chen, Yi Gu, Yue Wu, Mingyuan Tao, Zhihang Fu, and Jieping Ye. INSIDE: LLMs' internal states retain the power of hallucination detection. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=Zj12nzlQbz.

[39] Mengjia Niu, Hamed Haddadi, and Guansong Pang. Robust hallucination detection in LLMs via adaptive token selection. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems*, 2025. URL https://openreview.net/forum?id=gOwqPdBlRB.

[40] Minjae Lee, Kyungmin Kim, Taesoo Kim, and Sangdon Park. Selective generation for controllable language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=glfYOAzh2f.

[41] Ji Xin, Raphael Tang, Yaoliang Yu, and Jimmy Lin. The art of abstention: Selective prediction and error regularization for natural language processing. In Chengqing Zong, Fei Xia, Wenjie Li, and Roberto Navigli, editors, *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1040–1051, Online, August 2021. Association for Computational Linguistics. doi: 10.18653/v1/2021.acl-long.84. URL https://aclanthology.org/2021.acl-long.84/.

[42] Moxin Li, Wenjie Wang, Fuli Feng, Fengbin Zhu, Qifan Wang, and Tat-Seng Chua. Think twice before trusting: Self-detection for large language models through comprehensive answer reflection. In Yaser Al-Onaizan, Mohit Bansal, and Yun-Nung Chen, editors, *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 11858–11875, Miami, Florida, USA, November 2024. Association for Computational Linguistics. doi: 10.18653/v1/2024.findings-emnlp.693. URL https://aclanthology.org/2024.findings-emnlp.693/.

[43] Hyuhng Joon Kim, Youna Kim, Sang-goo Lee, and Taeuk Kim. When to speak, when to abstain: Contrastive decoding with abstention. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9710–9730, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.479. URL https://aclanthology.org/2025.acl-long.479/.

[44] Artem Vazhentsev, Ekaterina Fadeeva, Rui Xing, Gleb Kuzmin, Ivan Lazichny, Alexander Panchenko, Preslav Nakov, Timothy Baldwin, Maxim Panov, and Artem Shelmanov. Unconditional truthfulness: Learning unconditional uncertainty of large language models. In Christos Christodoulopoulos, Tanmoy Chakraborty, Carolyn Rose, and Violet Peng, editors, *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 35673–35694, Suzhou, China, November 2025. Association for Computational Linguistics. ISBN 979-8-89176-332-6. doi: 10.18653/v1/2025.emnlp-main.1807. URL https://aclanthology.org/2025.emnlp-main.1807/.

[45] Alexander V Nikitin, Jannik Kossen, Yarin Gal, and Pekka Marttinen. Kernel language entropy: Fine-grained uncertainty quantification for LLMs from semantic similarities. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=j2wCrWmgMX.

[46] Yasin Abbasi-Yadkori, Ilja Kuzborskij, David Stutz, András György, Adam Fisch, Arnaud Doucet, Iuliya Beloshapka, Wei-Hung Weng, Yao-Yuan Yang, Csaba Szepesvári, Ali Taylan Cemgil, and Nenad Tomasev. Mitigating llm hallucinations via conformal abstention. *CoRR*, abs/2405.01563, 2024. URL https://doi.org/10.48550/arXiv.2405.01563.

[47] Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. To believe or not to believe your llm: iterative prompting for estimating epistemic uncertainty. In *Proceedings of the 38th International Conference on Neural Information Processing Systems*, NIPS '24, Red Hook, NY, USA, 2024. Curran Associates Inc. ISBN 9798331314385.

[48] Dayong Liang, Xiao-Yong Wei, and Changmeng Zheng. Multi-agent undercover gaming: Hallucination removal through counterfactual test for multimodal reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(8):6807–6815, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i8.37613. URL http://dx.doi.org/10.1609/aaai.v40i8.37613.

[49] Bingbing Wen, Faeze Brahman, Zhan Su, Shangbin Feng, Yulia Tsvetkov, Lucy Lu Wang, and Bill Howe. MARVEL: Modular abstention for reliable and versatile expert LLMs. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*, 2025. URL https://openreview.net/forum?id=EQIBB1BA6Y.

[50] Yuxi Sun, Aoqi Zuo, Wei Gao, and Jing Ma. CausalAbstain: Enhancing multilingual LLMs with causal reasoning for trustworthy abstention. In Wanxiang Che, Joyce Nabende, Ekaterina

Shutova, and Mohammad Taher Pilehvar, editors, *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14060–14076, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-256-5. doi: 10.18653/v1/2025.findings-acl.723. URL https://aclanthology.org/2025.findings-acl.723/.

[51] Yuangang Li, Yiqing Shen, Yi Nian, Jiechao Gao, Ziyi Wang, Chenxiao Yu, Li Li, Jie Wang, Xiyang Hu, and Yue Zhao. Mitigating hallucinations in large language models via causal reasoning. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(38):31852–31860, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i38.40454. URL http://dx.doi.org/10.1609/aaai.v40i38.40454.

[52] Vy Nguyen, Ziqi Xu, Jeffrey Chan, Estrid He, Feng Xia, Renqiang Luo, Erik Cambria, and Xiuzhen Zhang. Twin worlds: Equivariance-based abstention for evidence-grounded reasoning, 2026. URL https://arxiv.org/abs/2608.28018.

[53] Roi Cohen, Konstantin Dobler, Eden Biran, and Gerard de Melo. I don't know: Explicit modeling of uncertainty with an [IDK] token. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=WcOvlQuoLb.

[54] Lei Huang, Xiaocheng Feng, Weitao Ma, Yuchun Fan, Xiachong Feng, Yuxuan Gu, Yangfan Ye, Liang Zhao, Weihong Zhong, Baoxin Wang, Dayong Wu, Guoping Hu, Lingpeng Kong, Tong Xiao, Ting Liu, and Bing Qin. Alleviating hallucinations from knowledge misalignment in large language models via selective abstention learning. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24564–24579, 2025. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1199. URL https://aclanthology.org/2025.acl-long.1199/.

[55] Zeyu Shi, Ziming Wang, Tianyu Chen, Shiqi Gao, Haoyi Zhou, Qingyun Sun, and Jianxin Li. Fine-tuned llms know they don't know: A parameter-efficient approach to recovering honesty. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(39):32983–32991, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i39.40580. URL http://dx.doi.org/10.1609/aaai.v40i39.40580.

[56] Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. LACIE: Listener-aware finetuning for calibration in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=RnvgYd9RAh.

[57] Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Sean Chen, Mohammad Kachuee, Teja Gollapudi, Tony Liao, Nicolas Scheffer, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen-tau Yih, and Xin Luna Dong. Truthrl: Incentivizing truthful llms via reinforcement learning, 2025. URL https://arxiv.org/abs/2509.25760.

[58] Zirui Zhao, Hanze Dong, Amrita Saha, Caiming Xiong, and Doyen Sahoo. Automatic curriculum expert iteration for reliable LLM reasoning. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=3ogIALgghF.

[59] Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. Rejection improves reliability: Training LLMs to refuse unknown questions using RL from knowledge feedback. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=lJMioZBoR8.

[60] Catherine Olsson, Nelson Elhage, Neel Nanda, Nicholas Joseph, Nova DasSarma, Tom Henighan, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Dawn Drain, Deep Ganguli, Zac Hatfield-Dodds, Danny Hernandez, Scott Johnston, Andy Jones, Jackson Kernion, Liane Lovitt, Kamal Ndousse, Dario Amodei, Tom Brown, Jack Clark, Jared Kaplan, Sam McCandlish, and Chris Olah. In-context learning and induction heads, 2022. URL https://arxiv.org/abs/2209.11895.

[61] Rhys Gould, Euan Ong, George Ogden, and Arthur Conmy. Successor heads: Recurring, interpretable attention heads in the wild. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=kvcbV8KQsi.

[62] Patrick Kahardipraja, Reduan Achtibat, Thomas Wiegand, Wojciech Samek, and Sebastian Lapuschkin. The atlas of in-context learning: How attention heads shape in-context retrieval augmentation, 2025. URL https://arxiv.org/abs/2505.15807.

[63] Wenhao Wu, Yizhong Wang, Guangxuan Xiao, Hao Peng, and Yao Fu. Retrieval head mechanistically explains long-context factuality. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=EytBpUGB1Z.

[64] ZhongXiang Sun, Xiaoxue Zang, Kai Zheng, Jun Xu, Xiao Zhang, Weijie Yu, Yang Song, and Han Li. RedeEP: Detecting hallucination in retrieval-augmented generation via mechanistic interpretability. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=ztzZDzgfrh.

[65] Debo Cheng, Ziqi Xu, Jiuyong Li, Lin Liu, Kui Yu, Thuc Duy Le, and Jixue Liu. Linking model intervention to causal interpretation in model explanation. *Pattern Recognition*, 173:112814, 2026. doi: 10.1016/J.PATCOG.2025.112814. URL https://doi.org/10.1016/j.patcog.2025.112814.

[66] Kevin Meng, David Bau, Alex J Andonian, and Yonatan Belinkov. Locating and editing factual associations in GPT. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022. URL https://openreview.net/forum?id=-h6WAS6eE4.

[67] Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. Dissecting recall of factual associations in auto-regressive language models. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.751. URL https://aclanthology.org/2023.emnlp-main.751/.

[68] Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

[69] Nirmalendu Prakash, Yeo Wei Jie, Amir Abdullah, Ranjan Satapathy, Erik Cambria, and Roy Ka-Wei Lee. Beyond i'm sorry, i can't: Dissecting large-language-model refusal. *Proceedings of the AAAI Conference on Artificial Intelligence*, 40(44):37830–37838, March 2026. ISSN 2159-5399. doi: 10.1609/aaai.v40i44.41119. URL http://dx.doi.org/10.1609/aaai.v40i44.41119.

[70] Amos Azaria and Tom Mitchell. The internal state of an LLM knows when it's lying. In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Findings of the Association for Computational Linguistics: EMNLP 2023*, pages 967–976, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-emnlp.68. URL https://aclanthology.org/2023.findings-emnlp.68/.

[71] Collin Burns, Haotian Ye, Dan Klein, and Jacob Steinhardt. Discovering latent knowledge in language models without supervision. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=ETKGuby0hcs.

[72] Kevin Liu, Stephen Casper, Dylan Hadfield-Menell, and Jacob Andreas. Cognitive dissonance: Why do language model outputs disagree with internal representations of truthfulness? In Houda Bouamor, Juan Pino, and Kalika Bali, editors, *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 4791–4797, Singapore, December 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.emnlp-main.291. URL https://aclanthology.org/2023.emnlp-main.291/.

[73] Hadas Orgad, Michael Toker, Zorik Gekhman, Roi Reichart, Idan Szpektor, Hadas Kotek, and Yonatan Belinkov. LLMs know more than they show: On the intrinsic representation of LLM hallucinations. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=KRnsX5Em3W.

[74] Shiyu Ni, Keping Bi, Jiafeng Guo, Lulu Yu, Baolong Bi, and Xueqi Cheng. Towards fully exploiting LLM internal states to enhance knowledge boundary perception. In Wanxiang Che, Joyce Nabende, Ekaterina Shutova, and Mohammad Taher Pilehvar, editors, *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 24315–24329, Vienna, Austria, July 2025. Association for Computational Linguistics. ISBN 979-8-89176-251-0. doi: 10.18653/v1/2025.acl-long.1184. URL https://aclanthology.org/2025.acl-long.1184/.

[75] Jakob Snel and Seong Joon Oh. First hallucination tokens are different from conditional ones, 2025. URL https://arxiv.org/abs/2507.20836.

[76] Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*, 2024. URL https://openreview.net/forum?id=IkmD3fKBPQ.

[77] Samuel Marks, Can Rager, Eric J Michaud, Yonatan Belinkov, David Bau, and Aaron Mueller. Sparse feature circuits: Discovering and editing interpretable causal graphs in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://openreview.net/forum?id=I4e82CIDxv.

[78] Pranav Rajpurkar, Robin Jia, and Percy Liang. Know what you don't know: Unanswerable questions for SQuAD. In Iryna Gurevych and Yusuke Miyao, editors, *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 784–789, 2018. doi: 10.18653/v1/P18-2124. URL https://aclanthology.org/P18-2124/.

[79] Harsh Trivedi, Niranjan Balasubramanian, Tushar Khot, and Ashish Sabharwal. MuSiQue: Multihop questions via single-hop question composition. *Transactions of the Association for Computational Linguistics*, 10:539–554, 2022. doi: 10.1162/tacl_a_00475. URL https://aclanthology.org/2022.tacl-1.31/.

[80] An Yang et al. Qwen3 technical report, 2025. URL https://arxiv.org/abs/2505.09388.

[81] Gemma Team. Gemma 3 technical report, 2025. URL https://arxiv.org/abs/2503.19786.

[82] Aaron Grattafiori et al. The llama 3 herd of models, 2024. URL https://arxiv.org/abs/2407.21783.

[83] Alexander H. Liu et al. Ministral 3, 2026. URL https://arxiv.org/abs/2601.08584.

[84] Marah Abdin et al. Phi-4 technical report, 2024. URL https://arxiv.org/abs/2412.08905.

[85] Ziqi Xu, Debo Cheng, Jiuyong Li, Jixue Liu, Lin Liu, and Ke Wang. Disentangled representation for causal mediation analysis. In *Thirty-Seventh AAAI Conference on Artificial Intelligence, AAAI 2023*, pages 10666–10674, 2023. doi: 10.1609/AAAI.V37I9.26266. URL https://doi.org/10.1609/aaai.v37i9.26266.

[86] Kevin Ro Wang, Alexandre Variengien, Arthur Conmy, Buck Shlegeris, and Jacob Steinhardt. Interpretability in the wild: a circuit for indirect object identification in GPT-2 small. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=NpsVSN6o4ul.

[87] Atticus Geiger, Duligur Ibeling, Amir Zur, Maheep Chaudhary, Sonakshi Chauhan, Jing Huang, Aryaman Arora, Zhengxuan Wu, Noah Goodman, Christopher Potts, and Thomas Icard. Causal abstraction: A theoretical foundation for mechanistic interpretability. *Journal of Machine Learning Research*, 26(83):1–64, 2025. URL http://jmlr.org/papers/v26/23-0058.html.

[88] Jacob Dunefsky, Philippe Chlenski, and Neel Nanda. Transcoders find interpretable LLM feature circuits. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=J6zHcScAo0.

[89] Tanya Chowdhury, Atharva Nijasure, Yair Zick, and James Allan. Hedonic neurons: A mechanistic mapping of latent coalitions in transformer MLPs. In *The Fourteenth International Conference on Learning Representations*, 2026. URL https://openreview.net/forum?id=v6HPsCu2R8.

[90] Anna Hedström, Salim I. Amoukou, Tom Bewley, Saumitra Mishra, and Manuela Veloso. To steer or not to steer? mechanistic error reduction with abstention for language models. In *Forty-second International Conference on Machine Learning*, 2025. URL https://openreview.net/forum?id=fUCPq5RvmH.

[91] Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In Ellen Riloff, David Chiang, Julia Hockenmaier, and Jun'ichi Tsujii, editors, *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium, October-November 2018. Association for Computational Linguistics. doi: 10.18653/v1/D18-1259. URL https://aclanthology.org/D18-1259/.

[92] Zhangyue Yin, Qiushi Sun, Qipeng Guo, Jiawen Wu, Xipeng Qiu, and Xuanjing Huang. Do large language models know what they don't know? In Anna Rogers, Jordan Boyd-Graber, and Naoaki Okazaki, editors, *Findings of the Association for Computational Linguistics: ACL 2023*, pages 8653–8665, Toronto, Canada, July 2023. Association for Computational Linguistics. doi: 10.18653/v1/2023.findings-acl.551. URL https://aclanthology.org/2023.findings-acl.551/.

[93] George Ma, Zhongyuan Liang, Irene Y. Chen, and Somayeh Sojoudi. Do sparse autoencoders identify reasoning features in language models? In *Forty-third International Conference on Machine Learning*, 2026. URL https://openreview.net/forum?id=TCFtA9CI3U.

A Datasets

The five datasets are chosen to stress-test abstention across three orthogonal axes of difficulty: *knowledge source* (parametric recall vs. in-context retrieval), *reasoning depth* (single-hop vs. multi-hop), and *unanswerability mechanism*. Table 4 summarises their key properties.

Table 4: Dataset summary. *Source*: whether the model must rely on parametric knowledge or a provided context passage. (†) held-out OOD evaluation sets only.

Dataset	Source	Reasoning	Unanswerability	License
KUQ [27]	Parametric	Open-domain QA	Answers genuinely do not exist (unsolved problems, future events); tests limits of *human* knowledge.	MIT
SQuAD 2.0 [78]	In-context	Reading comprehension	Questions plausible but unsupported by the passage; tests resistance to surface-level cues.	CC BY-SA 4.0
MuSiQue [79]	In-context	Multi-hop	One supporting document removed (*missing evidence*); tests detection of a silent gap in the reasoning chain.	CC BY 4.0
SelfAware† [92]	Parametric	Self-knowledge	No definitive answer exists; tests limits of the model's *own* knowledge, not of the world.	CC BY-SA 4.0
HotpotQA† [91]	In-context	Multi-hop	Supporting passages replaced by distractors (*misleading evidence*); tests resistance to plausible-but-wrong context.	CC BY-SA 4.0

B Construction and Validation of the Abstention Token Set $\mathcal{A}$

Table 5 lists the 17 tokens in $\mathcal{A}$. We built $\mathcal{A}$ using an additional pool of $1{,}000$ unanswerable instances drawn from the selected datasets, disjoint from the train and test splits, so the abstention vocabulary is constructed independently of the instances used for CAC localisation and evaluation. We ran zero-shot inference on this held-aside pool across all ten models, recorded the first generated token matching a canonical deferral pattern, then pooled and deduplicated across models. We restrict $\mathcal{A}$ to tokens that unambiguously initiate epistemic deferral, so the partition $(\mathcal{A}, \mathcal{C})$ is fixed at evaluation time and $\Delta(x)$ is a deterministic function of $r_L(x)$.

Each candidate token with its 10-token continuation was presented to two annotators on Amazon Mechanical Turk (English-speaking, $\geq 98\%$ approval rate, ≥ 500 approved HITs), who classified each item as *epistemic deferral* (abstention) or *substantive answer* (commitment); the annotation interface and instructions are shown in Fig. 9. Inter-annotator agreement was 98.3% (Cohen's $\kappa = 0.96$); disagreements were resolved by majority vote among three additional annotators. Workers were compensated at $20/hour.

Validation: $\Delta(x)$ agreement with full-generation We verify that the formulation of $\Delta(x)$ tracks full-generation behaviour by comparing, on the test split, the zero-threshold decision against a full-generation label obtained by greedy decoding (`max_new_tokens` $= 256$) and pattern-matching the response. Table 6 reports agreement rates across all 30 configurations; the mean is 97.6%. The 2.4% disagreements are predominantly pivot constructions in which the model begins with a substantive token but reverses course mid-generation.

Sensitivity of CAC findings to $\mathcal{A}$ construction We test whether the CAC and its properties are stable when $\mathcal{A}$ is varied, evaluating three alternatives against the main 17-token construction on sparsity and ablation AUROC. *(i) Per-model sets:* each model's $\mathcal{A}$ is built from its own outputs only ($|\mathcal{A}| = 8$–15, no pooling). *(ii) Expanded with prefix-disambiguated tokens:* 7 tokens that can plausibly initiate either deferral or substantive answers (e.g. `The`, `Based`) are added with a 10-token prefix check, giving $|\mathcal{A}| = 24$. *(iii) Further expanded:* 12 additional lower-frequency deferral starters seen in fewer than three models (e.g. `Regret`, `Apologies`) are added on top of (ii), giving $|\mathcal{A}| = 36$. Table 7 reports means across ten models. Both properties are stable across constructions; the largest deviation in ablation AUROC is 0.008 (further-expanded set). Stability follows from the dominance of high-frequency tokens that appear in all four constructions.

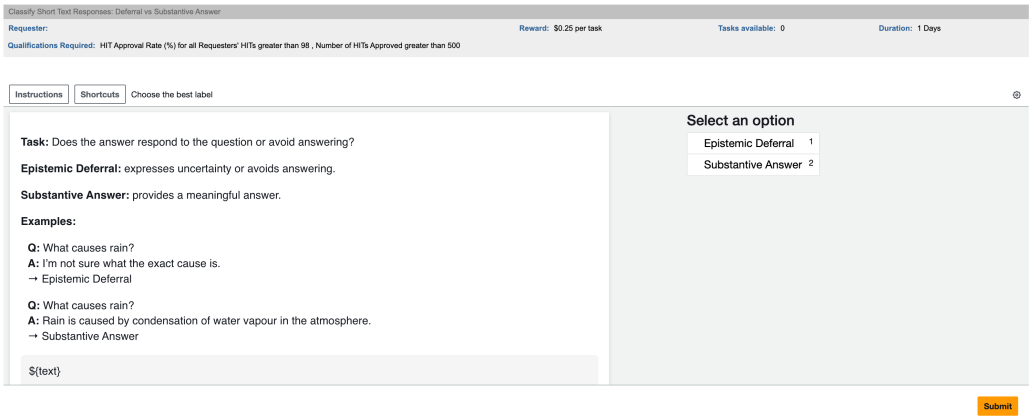


Figure 9: Annotation interface shown to Amazon Mechanical Turk workers for classifying candidate tokens as epistemic deferral or substantive answer.

Table 5: Abstention token set $\mathcal{A}$ (17 tokens). The commitment set $\mathcal{C}$ contains all tokens not in $\mathcal{A}$.

`I`	"I don't know", "I cannot determine"	`Sorry`	"Sorry, I don't have information…"
`Unfortunately`	"Unfortunately, I don't have enough…"	`Beyond`	"Beyond my knowledge…"
`Insufficient`	"Insufficient information to answer…"	`Impossible`	"Impossible to determine without context"
`Unanswerable`	"Unanswerable based on context"	`Without`	"Without more context, I cannot…"
`Unclear`	"Unclear from the given context…"	`Ambiguous`	"Ambiguous; the premise is unclear"
`Unknown`	"Unknown; context does not specify…"	`Unsure`	"Unsure; evidence is inconclusive"
`Uncertain`	"Uncertain; lacks a definitive answer"	`Unknowable`	"Unknowable from the text provided"
`Cannot`	"Cannot be determined from context"	`Indeterminate`	"Indeterminate based on context"
`Unable`	"Unable to answer from the passage"		

C Per-Configuration Margin Analysis

Table 8 reports all metrics for all 30 model-dataset configurations. AUROC and accuracy (Acc) summarise discrimination and threshold-based performance; true commitment (TC), true abstention (TA), false commitment (FC), and false abstention (FA) give the decision breakdown under the zero threshold. Mean rows average each metric across the three datasets.

D Localisation of the Commit-Abstain Circuit

D.1 Gate Dynamics

Derivation of gate dynamics Let $g_k \in \mathbb{R}$ be the logit of component k, with gate $G_k = \sigma(g_k)$. The gate scales the component's output before residual addition, so its contribution to $\Delta(x)$ is $G_k \cdot v_k(x)$, where $v_k(x) = d(x)^\top c_k(x)$ is the per-example projection defined in Eq. 2. Holding all other gates fixed, $\partial\Delta(x \mid G)/\partial g_k = v_k(x) \cdot \sigma'(g_k)$. The task loss in Eq. 3 then gives

$$\frac{\partial \mathcal{L}_{\text{task}}}{\partial g_k} = -\frac{1}{B}\sum_x (2y-1) \cdot v_k(x) \cdot \sigma'(g_k),$$

and since $\sigma'(g_k) > 0$, the sign of the gradient is determined by $-(2y-1) \cdot v_k(x)$, summed over the batch. For a balanced batch with answerable subset $\mathcal{D}_1$ and unanswerable subset $\mathcal{D}_0$,

$$\mathbb{E}_x[(2y-1) \cdot v_k(x)] \;\propto\; \mathbb{E}_{y=1}[v_k(x)] - \mathbb{E}_{y=0}[v_k(x)] \;=:\; \text{diff}_k,$$

up to a constant factor of $\frac{1}{2}$, so the expected task-loss gradient is proportional to $-\text{diff}_k$. The task loss opens gates whose $\text{diff}_k > 0$ and closes those whose $\text{diff}_k < 0$.

Phase outcomes Phase 2a applies retention pressure ($\lambda > 0$, regularisation gradient $-\lambda < 0$, pushing all gates open). Phase 2b applies removal pressure ($\lambda < 0$, pushing all gates closed). Combining the regularisation direction with the task-loss gradient:

Table 6: Agreement (%) between the token-level decision (first token $\in \mathcal{A}$ vs. $\mathcal{C}$) and the full-generation behavioural label across 30 model-dataset configurations. Mean agreement: 97.6%.

	Qwen 3.5		Gemma 3		Llama 3		Ministral		Phi-4		
	4B	9B	4B	12B	3B	8B	3B	14B	mini	14B	Mean
KUQ	97.2	98.1	96.8	98.4	96.5	97.8	97.9	98.6	97.3	98.0	97.7
SQuAD 2.0	98.0	98.5	96.2	97.6	95.8	97.1	97.4	98.2	98.3	98.7	97.6
MuSiQue	97.6	98.3	97.1	97.9	96.0	97.3	97.7	98.0	97.5	98.2	97.6
Mean	97.6	98.3	96.7	98.0	96.1	97.4	97.7	98.3	97.7	98.3	97.6

Table 7: Sensitivity of CAC properties to alternative $\mathcal{A}$ constructions, averaged across ten models. Sparsity: fraction of all components selected into the CAC. Ablation AUROC: AUROC of $\Delta(x)$ after full ablation of C-components.

Construction	**Sparsity**	**Ablation AUROC**
Main ($\lvert\mathcal{A}\rvert = 17$, unambiguous, pooled)	0.051	0.641
(i) Per-model sets ($\lvert\mathcal{A}\rvert = 8$–$15$)	0.054	0.638
(ii) Expanded with prefix check ($\lvert\mathcal{A}\rvert = 24$)	0.052	0.644
(iii) Further expanded ($\lvert\mathcal{A}\rvert = 36$)	0.053	0.649

	Phase 2a (retention)		Phase 2b (removal)	
diff_k	Task loss	Outcome	Task loss	Outcome
> 0	opens	$G^+ \to 1$	opens, fights removal	$G^- > \tau$
< 0	closes, fights retention	$G^+ < \tau$	closes	$G^- \to 0$
≈ 0	neutral	$G^+ \to 1$	neutral	$G^- \to 0$

C-components ($\mathrm{diff}_k > 0$) hold their gates open against removal pressure: their contribution to $\Delta(x)$ is larger on answerable inputs, where the loss requires them. A-components ($\mathrm{diff}_k < 0$) hold their gates closed against retention pressure: their contribution is larger on unanswerable inputs, so the loss closes them under retention. Irrelevant components ($\mathrm{diff}_k \approx 0$) follow regularisation in both phases without resistance.

Worked example A non-obvious case: a component can be classified as A even when its contribution is positive on both subsets. Consider $\mathbb{E}_{y=1}[v_k(x)] \approx +0.3$ and $\mathbb{E}_{y=0}[v_k(x)] \approx +1.5$, so $\mathrm{diff}_k = -1.2$. The component pushes $\Delta(x)$ up everywhere, but its excess on unanswerable inputs inflates $\Delta(x)$ where it should be negative. The task-loss gradient closes the gate; under retention pressure (Phase 2a), this suppresses G^+ below τ, classifying the component as A. The role label reflects which subset v_k is larger on, not the absolute sign of v_k.

The table characterises the direction of the gradient, not a closed-form fixed point: optimisation is non-convex, $\sigma'(g_k)$ vanishes near saturation, and $v_k(x)$ shifts as other gates change. In practice components saturate in the predicted direction, consistently across seeds (App. D.4), with intermediate values arising mainly when the two forces are closely balanced. The classification in Section 4.3 selects on the resolved gate values rather than on diff_k directly: optimisation over all gates jointly captures inter-component interactions that a per-component test on diff_k would miss.

D.2 Hyperparameter Details

Gating all components jointly is computationally expensive. We therefore pre-screen candidates using a correlational filter: for each component k, we rank components by diff_k computed on the training set under the ungated model. We select the top-K components per model (pooled across datasets); all remaining components are left ungated ($G = 1$). We optimise gates with Adam and linear learning-rate decay (Phase 1: $0.1 \to 0.01$; Phases 2a/2b: $0.5 \to 0.1$) for 100 steps with batch size $B = 16$. Following Nam et al. [26], we use a clipping bound $C = 4$. The regularisation weight is set adaptively to match the scale of the task loss: $\lambda = \mathbb{E}_{x\sim\mathcal{D}}[|\Delta(x)|]/(BK)$, where the expectation is computed on the training set under the ungated model. Gate logits are initialised by

Table 8: Full results for $\Delta(x)$ analysis across all 30 model-dataset configurations.

		Qwen 3.5		Gemma 3		Llama 3		Ministral		Phi-4	
		4B	9B	4B	12B	3B	8B	3B	14B	mini	14B
AUROC	KUQ	.708	.678	.742	.906	.733	.763	.779	.874	.858	.845
	SQuAD 2.0	.714	.872	.689	.838	.640	.788	.868	.862	.753	.831
	MuSiQue	.861	.913	.820	.889	.798	.875	.806	.888	.829	.900
	Mean	*.761*	*.821*	*.750*	*.878*	*.723*	*.809*	*.818*	*.875*	*.813*	*.859*
Acc	KUQ	.644	.660	.672	.821	.666	.684	.710	.778	.765	.745
	SQuAD 2.0	.643	.783	.618	.739	.584	.700	.791	.718	.669	.713
	MuSiQue	.793	.716	.721	.806	.693	.622	.705	.829	.763	.748
	Mean	*.693*	*.720*	*.670*	*.789*	*.648*	*.669*	*.735*	*.775*	*.732*	*.735*
TC	KUQ	.740	.510	.878	.916	.696	.714	.786	.900	.730	.700
	SQuAD 2.0	.718	.668	.832	.880	.676	.662	.866	.954	.888	.646
	MuSiQue	.734	.460	.546	.724	.460	.260	.472	.864	.742	.552
	Mean	*.731*	*.546*	*.752*	*.840*	*.611*	*.545*	*.708*	*.906*	*.787*	*.633*
TA	KUQ	.548	.810	.466	.726	.636	.654	.634	.656	.800	.790
	SQuAD 2.0	.568	.898	.404	.598	.492	.738	.716	.482	.450	.780
	MuSiQue	.852	.972	.896	.888	.926	.984	.938	.794	.784	.944
	Mean	*.656*	*.893*	*.589*	*.737*	*.685*	*.792*	*.763*	*.644*	*.678*	*.838*
FC	KUQ	.452	.190	.534	.274	.364	.346	.366	.344	.200	.210
	SQuAD 2.0	.432	.102	.596	.402	.508	.262	.284	.518	.550	.220
	MuSiQue	.148	.028	.104	.112	.074	.016	.062	.206	.216	.056
	Mean	*.344*	*.107*	*.411*	*.263*	*.315*	*.208*	*.237*	*.356*	*.322*	*.162*
FA	KUQ	.260	.490	.122	.084	.304	.286	.214	.100	.270	.300
	SQuAD 2.0	.282	.332	.168	.120	.324	.338	.134	.046	.112	.354
	MuSiQue	.266	.540	.454	.276	.540	.740	.528	.136	.258	.448
	Mean	*.269*	*.454*	*.248*	*.160*	*.389*	*.455*	*.292*	*.094*	*.213*	*.367*

sampling $G_0 \sim \mathrm{Uniform}(0.02,\, 0.98)$ and applying the inverse sigmoid. All results are averaged over 10 independent seeds.

D.3 CAC Localisation Results

Table 9 reports the CAC localisation results for all ten models, listing for each model the number of candidates screened, the number classified as abstention-promoting (A), commitment-promoting (C), and irrelevant by the gating protocol, the CAC's overall sparsity, the fraction of MLP sublayers, and the C:A count ratio ρ_θ at four normalised depth thresholds, measuring how strongly C-components predominate in earlier layers.

D.4 CAC Validation and Sensitivity Analyses

Cross-seed stability Gate logits are initialised independently across 10 random seeds per model. The resulting gate values are near-binary in every model: across all seeds and models, 80% of G^+ values and 78% of G^- values fall below 0.1 or above 0.9, indicating that component roles are resolved cleanly and not sensitive to initialisation. A component is included in the CAC only if its role assignment (commitment- or abstention-promoting) is consistent in at least 8 of 10 seeds, i.e. its cross-seed identification rate $p \geq 0.80$; components below this threshold are classified as irrelevant. The CAC's topology (the set of A and C components) is therefore reproducible: rerunning the gating procedure from a different seed yields the same components in the vast majority of cases.

Sign agreement between $\Delta_{\mathbf{CAC}}(x)$ and $\Delta(x)$ To verify that the identified CAC accounts for the commit-abstain decision, we compute $\Delta_{\mathrm{CAC}}(x) = \kappa(x) + \sum_{k \in \mathrm{CAC}} v_k(x)$, where $\kappa(x)$ is the residual after subtracting all component contributions from $\Delta(x)$ (shared across all inputs, not fit to the CAC), and check whether its sign matches that of $\Delta(x)$ on the test split (N=1,500 instances per model). Table 10 reports sign agreement per model, with a breakdown by answerability. The CAC reproduces the sign of $\Delta(x)$ in 85–92% of cases (mean 88.4%). Agreement is consistently higher on unanswerable instances (mean 90%) than on answerable ones (mean 86%), confirming that the

Table 9: CAC localisation summary across ten models. Sparsity: fraction of all model components (A + C) selected into the CAC. MLP%: fraction of CAC components that are MLP sublayers. ρ_θ: C-to-A count ratio at normalised depth θ—the number of C-components seen per A-component seen up to that depth; $\rho_\theta > 1$ means C-components predominate in layers seen so far (bold); ∞ means no A-components have appeared yet.

	Component counts							C:A ratio ρ_θ		
Model	Screened	A	C	Irrel.	Irrel.%	Sparsity	MLP%	$\rho_{60\%}$	$\rho_{70\%}$	$\rho_{80\%}$
Qwen 4B	81	17	31	33	41	.088	29	**6.6**×	**2.4**×	**2.1**×
Qwen 9B	80	22	20	38	48	.077	14	**1.5**×	**1.4**×	**2.0**×
Gemma 4B	77	15	21	41	53	.118	17	**1.8**×	**1.2**×	**1.1**×
Gemma 12B	72	11	5	56	78	.020	12	**1.4**×	**1.3**×	**1.1**×
Llama 3B	81	20	33	28	35	.076	26	∞	**7.3**×	**3.4**×
Llama 8B	77	15	19	43	56	.032	32	**5.5**×	**3.2**×	**2.4**×
Ministral 3B	87	19	27	41	47	.054	33	∞	**3.5**×	**2.3**×
Ministral 14B	72	32	15	25	35	.036	34	**1.6**×	**1.4**×	**1.2**×
Phi-4 mini	82	24	17	41	50	.051	34	**1.6**×	**1.3**×	**1.2**×
Phi-4	79	36	28	15	19	.039	25	**1.9**×	**1.8**×	**1.3**×
$\rho_\theta > 1$								*10/10*	*10/10*	*10/10*
Mean								*2.7*×	*2.5*×	*1.8*×

Table 10: Sign agreement between $\Delta_{\text{CAC}}(x)$ and $\Delta(x)$ on the test split.

	Qwen 3.5		Gemma 3		Llama 3		Ministral		Phi-4		
	4B	9B	4B	12B	3B	8B	3B	14B	mini	14B	Mean
Overall	.882	.908	.854	.922	.862	.890	.878	.866	.882	.898	*.884*
Answerable	.854	.882	.832	.902	.840	.868	.856	.846	.860	.876	*.862*
Unanswerable	.910	.934	.876	.942	.884	.912	.900	.886	.904	.920	*.907*

CAC captures abstention-relevant signals rather than general margin direction. Cases where signs disagree are concentrated at small $|\Delta(x)|$: the mean $|\Delta(x)|$ for disagreeing inputs is 1.3 versus 6.1 for agreeing inputs, indicating the CAC reproduces the decision in all but the most marginal cases.

Screening threshold K We vary the number of candidates passed to the gating protocol from K to $2K$ and $4K$ for three models spanning the full sparsity range: Gemma 12B (K=72, sparsest at 2.0%), Llama 8B (K=77, mid at 3.2%), and Gemma 4B (K=77, densest at 11.8%). Table 11 reports component counts and Jaccard similarity against the K baseline. At $2K$, Jaccard is 1.00 for five of six component-type pairs and 0.94 for the sixth (A-components, Gemma 4B); at $4K$, the minimum drops to 0.88 (Gemma 4B A-components). Component counts change by at most one between K and $2K$, and by at most two between K and $4K$, while the irrelevance rate rises from 53–78% at K to 87–94% at $4K$. The near-identical CACs at K and $2K$ indicate that the original screening threshold is already conservative: doubling the candidate pool admits at most one new component per model (Gemma 4B gains one A-component; the other two models are unchanged), and the core CAC is fully recovered at the base setting.

Regularisation strength λ We vary the regularisation coefficient across five levels ($0.5\lambda_0$, $0.75\lambda_0$, λ_0, $1.5\lambda_0$, $2\lambda_0$) for the same three models. λ_0 is set adaptively as $\lambda_0 = \mathbb{E}[|\Delta(x)|]/(BK)$, where B is the batch size. Table 12 reports component counts, Jaccard similarity against the λ_0 baseline, and the depth gap ($\bar{\ell}_{\text{A}}/L - \bar{\ell}_{\text{C}}/L$). The Jaccard variation is driven by borderline components: in all cases, components that enter or leave the CAC as λ changes have a cross-seed identification rate of $\bar{p} \leq 0.64$, below the 0.80 threshold required for CAC membership. This indicates that the high-confidence core of the CAC is preserved across all λ values tested.

D.5 Causal Intervention

For ablation, we progressively remove the top-k CAC components on the test set, ordered by gating score, and measure directional sign flips and AUROC. To verify that observed effects are CAC-specific rather than artefacts of the transformer's sensitivity to arbitrary component removal, we compare

Table 11: Screening threshold sensitivity. K=72 for Gemma 12B; K=77 for Llama 8B and Gemma 4B. A/C: component counts. Irr.%: fraction of screened candidates rejected by causal gating. $J(\text{A})$, $J(\text{C})$, J: per-type and full-CAC Jaccard against the K baseline.

	Gemma 12B						Llama 8B						Gemma 4B					
	A	C	Irr.%	$J(\text{A})$	$J(\text{C})$	J	A	C	Irr.%	$J(\text{A})$	$J(\text{C})$	J	A	C	Irr.%	$J(\text{A})$	$J(\text{C})$	J
K	11	5	78	1.00	1.00	1.00	15	19	56	1.00	1.00	1.00	15	21	53	1.00	1.00	1.00
$2K$	11	5	89	1.00	1.00	1.00	15	19	78	1.00	1.00	1.00	16	21	75	0.94	1.00	0.97
$4K$	12	5	94	0.92	1.00	0.94	16	20	88	0.94	0.95	0.94	17	22	87	0.88	0.95	0.92

Table 12: Regularisation sensitivity. $\lambda_0 = \mathbb{E}[|\Delta(x)|]/(BK)$, where B is the batch size. A/C: component counts. J: full-CAC Jaccard against λ_0. Dep.: depth gap $\bar{\ell}_{\text{A}}/L - \bar{\ell}_{\text{C}}/L$. $\bar{p}$: mean cross-seed identification rate of components entering or leaving relative to λ_0.

	Gemma 12B					Llama 8B					Gemma 4B				
λ	A	C	J	Dep.	$\bar{p}$	A	C	J	Dep.	$\bar{p}$	A	C	J	Dep.	$\bar{p}$
$0.5\lambda_0$	9	4	0.83	+.11	0.58	13	17	0.86	+.12	0.60	13	18	0.85	+.03	0.58
$0.75\lambda_0$	10	5	0.93	+.12	0.62	14	18	0.93	+.13	0.64	14	20	0.93	+.04	0.63
λ_0	11	5	1.00	+.13	—	15	19	1.00	+.14	—	15	21	1.00	+.02	—
$1.5\lambda_0$	11	6	0.94	+.12	0.61	15	20	0.95	+.13	0.63	16	22	0.94	+.04	0.62
$2\lambda_0$	12	6	0.88	+.12	0.57	16	21	0.90	+.13	0.59	17	23	0.84	+.05	0.57

against 20 size-matched random non-CAC ablations at each step. Table 13 reports directional sign flips under full ablation across all ten models; full progressive ablation curves against the random baselines are in Figure 7.

Table 13: Directional sign flips (% of all 1,500 held-out instances) under full ablation (A→C = abstain flips to commit; C→A = commit flips to abstain).

	Ablate A-components		Ablate C-components	
	A→C	C→A	C→A	A→C
Qwen 4B	6.6	9.5	0.0	45.9
Qwen 9B	9.8	5.7	12.4	12.8
Gemma 4B	19.7	17.9	12.5	7.3
Gemma 12B	1.9	3.9	2.7	3.3
Llama 3B	52.9	2.3	0.1	59.7
Llama 8B	14.5	8.6	0.0	60.0
Ministral 3B	8.1	2.3	0.9	52.7
Ministral 14B	2.9	17.5	1.9	17.6
Phi-4 mini	44.8	0.4	0.1	44.9
Phi-4	10.0	10.3	3.3	12.9
Mean	17.1	7.8	3.4	31.7

For amplification, we study two error modes: FC cases, unanswerable inputs (Y=0) where $\Delta(x) > 0$ at baseline; and FA cases, answerable inputs (Y=1) where $\Delta(x) \leq 0$ at baseline. For each CAC component, a forward hook scales its output h by α before residual addition, replacing h with αh in the residual stream; all other components are unchanged. We run two symmetric sweeps mirroring the CAC's predicted functional roles: A-component amplification on FC instances and C-component amplification on FA instances, each over $\alpha \in \{1.0, 1.5, 2.0, 2.5, 3.0, 10, 100, 1000\}$.

Tables 14 and 15 report per-model FA and FC rates across all α values. At moderate α ($\leq$3), Qwen 4B and Phi-4 mini reach negative mean margin by α=3.0, indicating amplification has reversed the majority of FC cases in those models. At extreme α ($\geq$10), FC rate at α=100 is nearly identical to α=1000 in every model (mean absolute difference 0.004), confirming the plateau reported in the main paper. Full results are consistent with Finding D.

Table 14: FA rate and mean $\Delta(x)$ under C-component amplification on FA instances.

	FA rate (moderate α)					FA rate (extreme α)			mean $\Delta(x)$		
Model	1.0	1.5	2.0	2.5	3.0	10	100	1000	α=1	α=2	α=3
Llama 3B	.416	.363	.409	.413	.344	.068	.003	.000	−1.89	−3.24	−0.69
Llama 8B	.475	.415	.408	.409	.417	.001	.000	.000	−3.82	−0.92	−0.64
Qwen 4B	.309	.240	.269	.227	.088	.000	.000	.000	−1.36	−1.32	+1.56
Qwen 9B	.480	.440	.415	.383	.373	.005	.000	.000	−2.06	−1.79	−1.53
Ministral 3B	.320	.237	.224	.231	.047	.140	.001	.001	−1.74	−0.90	+0.90
Ministral 14B	.115	.097	.085	.076	.077	.016	.000	.000	−1.57	−0.90	−0.41
Phi-4 mini	.239	.169	.184	.136	.005	.000	.000	.000	−1.36	−2.05	+21.00
Phi-4 14B	.383	.379	.373	.367	.364	.065	.000	.001	−3.39	−4.41	−4.11
Gemma 4B	.279	.263	.256	.240	.224	.128	.024	.019	−8.96	−11.43	−9.28
Gemma 12B	.181	.173	.169	.164	.159	.165	.119	.059	−7.94	−8.04	−8.05

Table 15: FC rate and mean $\Delta(x)$ under A-component amplification on FC instances.

	FC rate (moderate α)					FC rate (extreme α)			mean $\Delta(x)$		
Model	1.0	1.5	2.0	2.5	3.0	10	100	1000	α=1	α=2	α=3
Llama 3B	.337	.311	.280	.237	.228	.316	.331	.332	+2.86	+1.92	+1.05
Llama 8B	.224	.213	.199	.187	.179	.197	.224	.224	+3.55	+2.97	+2.27
Qwen 4B	.380	.311	.247	.172	.147	.368	.380	.379	+2.08	+0.66	−0.37
Qwen 9B	.127	.100	.083	.071	.067	.127	.127	.127	+1.48	+1.05	+1.05
Ministral 3B	.252	.232	.204	.199	.189	.189	.231	.231	+1.99	+1.27	+0.98
Ministral 14B	.380	.368	.355	.324	.245	.368	.380	.380	+2.38	+2.23	+0.73
Phi-4 mini	.351	.281	.188	.152	.144	.351	.349	.349	+2.38	+0.41	−0.25
Phi-4 14B	.195	.176	.155	.125	.108	.136	.193	.193	+3.23	+1.68	+1.09
Gemma 4B	.433	.429	.423	.415	.409	.361	.317	.307	+11.75	+9.42	+7.94
Gemma 12B	.279	.268	.264	.256	.256	.225	.119	.108	+10.85	+9.20	+7.96

E CAC-Informed Abstention Policy

E.1 Baseline Configurations

We compare against seven baselines spanning the information spectrum from a single scalar to full hidden-state representations across all layers and multiple generations:

- **Zero-Threshold.** Abstains whenever $\Delta(x) < 0$.
- **Semantic Entropy [10].** Clusters K sampled generations by semantic equivalence (via NLI) and computes entropy over the resulting clusters. Operating purely on outputs, it draws on K=10 full generations per input (640 tokens total) and NLI-based semantic clustering. *Configuration:* K=10 samples; temperature 1.0, top-p 0.95, `max_new_tokens` 64; NLI model `microsoft/deberta-v2-xlarge-mnli`, entailment threshold 0.5. *Code:* `https://github.com/jlko/semantic_uncertainty`
- **INSIDE [38].** Measures self-consistency of K sampled responses via the eigenvalue spectrum of their hidden-state covariance matrix (EigenScore). It uses middle-layer hidden-state representations across K=10 generations, at 10× inference cost. *Configuration:* K=10 samples; temperature 0.5, top-p 0.99; regularisation α=10^{-3}; representations from middle-layer hidden states. *Code:* `https://github.com/alibaba/eigenscore`
- **HaloScope [11].** Identifies a hallucination subspace in LLM activation space via SVD on last-token embeddings from unlabelled generations, then scores each input by the norm of its projection onto the top-k singular vectors (weighted by singular values); a binary classifier is trained on top of this membership score. *Configuration:* SVD rank $k \in \{1, \ldots, 10\}$ searched across all decoder layers; singular values used as projection weights; `max_new_tokens` 64. *Code:* `https://github.com/deeplearning-wisc/haloscope`
- **Multi-LLM [16].** Proposes Cooperate and Compete: distinct reviewer LLMs provide feedback on a target model's answer (cooperative) or generate conflicting evidence to probe its confidence (competitive), with the target model synthesising the signal into an

abstain decision. We use the Cooperate approach. *Configuration:* $K{=}5$ reviewer samples, temperature 0.7, top-p 0.95. *Code:* `https://github.com/BunsenFeng/AbstainQA`

- **MERA [90].** Trains per-layer linear probes to estimate model error, derives a closed-form steering direction, and applies interventions only when a Hoeffding-calibrated threshold guarantees improvement; abstains from steering otherwise. We adapt the error-estimation probe for abstention from answering using the same threshold calibration. *Configuration:* Lasso $\alpha \in \{0.5, 0.25, 0.1, 0.05, 0.01, 0.005\}$; best layer selected by training R^2; Hoeffding bound ($\delta{=}0.01$) over a 10-point threshold grid. *Code:* `https://github.com/annahedstroem/MERA-steering`
- **HaMI [39].** Formulates hallucination detection as a Multiple Instance Learning (MIL) problem over token-level hidden states: each response sequence is a bag of token instances, and an MLP is trained end-to-end to jointly select the most hallucination-indicative tokens and classify the bag. Hidden states are extracted at uniformly spaced candidate layers; the MIL objective retains the top-k tokens per bag as positive/negative instances. *Configuration:* 8 candidate layers (uniformly spaced); top 10% of tokens per bag; 100 epochs, batch size 128, Adam lr 10^{-3}, weight decay $5{\times}10^{-4}$; `max_new_tokens` 64. *Code:* `https://github.com/mala-lab/HaMI`

E.2 Main Results

Tables 16 and 17 report per-dataset breakdowns of the results summarised in Table 2. Table 16 gives overall accuracy for all 30 model–dataset configurations. Table 17 gives commitment-level precision, recall, and F1, together with the mean false abstention rate, comparing the CAC-Informed MLP against the Zero-Threshold baseline.

E.3 Logistic Regression Variant

We evaluate whether the MLP is necessary or whether a logistic regression classifier over the same CAC features achieves comparable performance. The logistic regression (CAC LR) uses the same input features as the MLP but replaces the nonlinear classifier with a single linear model. Table 18 reports per-dataset and mean accuracy across all 10 models.

The logistic regression improves over Zero-Threshold in all 30 configurations (mean $+10.6$ accuracy points). The MLP improves over the logistic regression in 20 of 30 configurations overall (mean $+1.6$ points), with the advantage concentrated on KUQ ($+3.6$ points) where parametric self-knowledge is required. On SQuAD 2.0 and MuSiQue, LR and MLP are nearly tied (means of 0.743/0.745 and 0.764/0.774, respectively), with LR winning individual configurations roughly as often as MLP. The gap between LR and MLP is small relative to the gap between Zero-Threshold and LR, confirming that the primary source of improvement is the component-level feature decomposition rather than the classifier architecture.

F Case Studies

The following two case studies illustrate the failure mode the CAC-informed policy is designed to correct: the raw commit-abstain margin $\Delta(x)$ exceeds zero (so Zero-Threshold commits), yet the component-level decomposition reveals a pattern that the MLP uses to correctly abstain.

F.1 Case 1: Unsolved Research Problem (KUQ, Llama 3B)

A **parametric** instance evaluated on Llama 3B: no context is provided, and the question poses an open research problem for which no established answer exists. The model must rely solely on internal knowledge, which is insufficient here.

> *Question:* Can we improve the monitoring of people with multiple sclerosis using simple tools, data sharing and patient engagement?
>
> *Gold label:* unanswerable
>
> *Model response:* "Yes, data sharing and patient engagement can improve MS monitoring."

Table 16: Overall accuracy across 10 models and 3 held-out datasets. **Bold**: best in column for each dataset section; underline: second best.

	Qwen 3.5		Gemma 3		Llama 3		Ministral		Phi-4	
	4B	9B	4B	12B	3B	8B	3B	14B	mini	14B
KUQ										
Zero-Threshold	.618	.646	.656	.818	.662	.686	.722	.778	.748	.740
Non-CAC Random	.664	.742	.706	.762	.612	.772	.694	.714	.716	.774
INSIDE	.600	.564	.628	.828	.626	.658	.562	.594	.526	.550
Multi-LLM	.730	.672	.815	.854	.640	.800	.705	.794	.691	.727
Semantic Entropy	.530	.508	.536	.586	.520	.508	.500	.550	.474	.582
MERA	.686	.617	.720	.748	.694	.690	.704	.712	.637	.633
HaloScope	.504	.504	.856	.738	.884	.876	.600	.642	.738	.730
HaMI	**.932**	**.976**	.874	.890	.830	**.938**	.862	**.956**	.838	**.952**
CAC-Informed MLP (Ours)	.912	.904	**.928**	**.946**	**.908**	.930	**.922**	.936	**.902**	.944
SQuAD 2.0										
Zero-Threshold	.598	.768	.588	.724	.540	.664	.756	.690	.648	.678
Non-CAC Random	.684	.756	.612	.694	.638	.702	.744	.650	.654	.716
INSIDE	.640	.524	.596	.668	.630	.472	.496	.504	.524	.562
Multi-LLM	.653	.733	.574	.661	.553	.643	.631	.636	.615	.655
Semantic Entropy	.706	.504	.572	.596	**.678**	.664	.500	.676	.648	.608
MERA	.612	.657	.553	.602	.536	.599	.620	.646	.584	.541
HaloScope	.638	.802	.630	.512	.560	.660	.706	.696	.632	.602
HaMI	.730	**.862**	.582	.702	.572	.674	.700	**.806**	**.776**	.778
CAC-Informed MLP (Ours)	**.770**	.850	**.636**	**.788**	.620	**.744**	**.766**	.762	.710	**.802**
MuSiQue										
Zero-Threshold	.750	.676	.688	.768	.668	.602	.664	**.790**	.720	.716
Non-CAC Random	.536	.674	.728	.776	.538	.719	.572	.790	.709	.766
INSIDE	.474	.482	.580	.628	.436	.484	.486	.520	.494	.516
Multi-LLM	.635	.668	.631	.667	.599	.635	.630	.605	.591	.600
Semantic Entropy	.492	.506	.574	.624	.468	.446	.500	.448	.488	.476
MERA	.621	.502	.490	.588	.489	.450	.538	.637	.571	.520
HaloScope	.644	.658	.570	.708	.680	.610	.682	.736	.702	.680
HaMI	.804	.804	.678	.764	.718	**.796**	.658	.784	**.744**	.762
CAC-Informed MLP (Ours)	**.806**	**.812**	**.732**	**.800**	**.736**	.786	**.758**	**.790**	**.744**	**.774**

CAC decomposition Table 19 shows the top-20 CAC contributions ordered by layer. The CAC is dominated by a single large positive contribution from L19.MLP ($+8.27$), which drives $\Delta(x) = +7.94$ and causes Zero-Threshold to commit. The remaining components are mixed but lean negative ($\sum^{-} = -8.09$, $\sum^{+} = +12.62$), with the net CAC margin ($\Delta_{\mathrm{CAC}} = +4.53$) positive only because of this outlier. The MLP reads the full decomposition and correctly abstains ($z_{\mathrm{MLP}} = 0.089$).

F.2 Case 2: False Premise Question (SQuAD 2.0, Ministral 3B)

An **in-context** instance evaluated on Ministral 3B: a passage is provided, but the question embeds a false premise that the passage explicitly contradicts. The model must detect the mismatch and refuse the question rather than answer it.

Context: "The election produced a majority SNP government, making this the first time in the Scottish Parliament where a party has commanded a parliamentary majority. The SNP took 16 seats from Labour, with many of their key figures not returned to parliament, although Labour leader Iain Gray retained East Lothian by 151 votes. The SNP took a further eight seats from the Liberal Democrats and one seat from the Conservatives. The SNP overall majority meant that there was sufficient support in the Scottish Parliament to hold a referendum on Scottish independence."

Question: Many of whose key figures returned to parliament?

Gold label: unanswerable

Model response: "Labour's key figures did not return."

Table 17: Commitment-level precision, recall, and F1, and mean false abstention rate, across 10 models and 3 datasets. **Bold**: best in column. Mean false abstention rate: 0.127 (CAC-Informed MLP) vs. 0.320 (Zero-Threshold).

		Qwen 3.5		Gemma 3		Llama 3		Ministral		Phi-4	
		4B	9B	4B	12B	3B	8B	3B	14B	mini	14B
Precision	*KUQ*										
	Zero-Threshold	.595	.693	.610	.761	.651	.670	.691	.723	.754	.752
	CAC-Informed MLP (Ours)	**.893**	**.891**	**.915**	**.931**	**.898**	**.939**	**.892**	**.922**	**.939**	**.959**
	SQuAD 2.0										
	Zero-Threshold	.588	**.845**	.561	.674	.534	.677	**.722**	.628	.604	.700
	CAC-Informed MLP (Ours)	**.718**	.825	**.591**	**.725**	**.586**	**.686**	.719	**.713**	**.662**	**.758**
	MuSiQue										
	Zero-Threshold	.791	**.931**	**.809**	**.842**	**.839**	**.947**	**.842**	.772	.735	**.865**
	CAC-Informed MLP (Ours)	**.805**	.810	.725	.800	.675	.764	.745	**.784**	**.740**	.744
Recall	*KUQ*										
	Zero-Threshold	.736	.524	.864	.928	.700	.732	.804	.900	.736	.716
	CAC-Informed MLP (Ours)	**.936**	**.920**	**.944**	**.964**	**.920**	**.920**	**.960**	**.952**	**.860**	**.928**
	SQuAD 2.0										
	Zero-Threshold	.656	.656	.808	.868	.636	.628	.832	**.932**	.860	.624
	CAC-Informed MLP (Ours)	**.888**	**.888**	**.884**	**.928**	**.820**	**.900**	**.872**	.876	.860	**.888**
	MuSiQue										
	Zero-Threshold	.680	.380	.492	.660	.416	.216	.404	**.824**	.688	.512
	CAC-Informed MLP (Ours)	**.808**	**.816**	**.748**	**.800**	**.912**	**.828**	**.784**	.800	**.752**	**.836**
F1	*KUQ*										
	Zero-Threshold	.658	.597	.715	.836	.674	.700	.743	.802	.745	.734
	CAC-Informed MLP (Ours)	**.914**	**.906**	**.929**	**.947**	**.909**	**.929**	**.925**	**.937**	**.898**	**.943**
	SQuAD 2.0										
	Zero-Threshold	.620	.739	.662	.759	.580	.651	.773	.750	.710	.660
	CAC-Informed MLP (Ours)	**.794**	**.855**	**.708**	**.814**	**.683**	**.779**	**.788**	**.786**	**.748**	**.818**
	MuSiQue										
	Zero-Threshold	.731	.540	.612	.740	.556	.352	.546	**.797**	.711	.643
	CAC-Informed MLP (Ours)	**.806**	**.813**	**.736**	**.800**	**.776**	**.795**	**.764**	.792	**.746**	**.787**
	Mean false abstention rate										
	Zero-Threshold	.309	.480	.279	.181	.416	.475	.320	**.115**	.239	.383
	CAC-Informed MLP (Ours)	**.123**	**.125**	**.141**	**.103**	**.116**	**.117**	**.128**	.124	**.176**	**.116**

CAC decomposition Table 20 shows the top-20 CAC contributions ordered by layer. The CAC margin is overwhelmingly negative ($\sum^- = -7.62$, $\sum^+ = +2.40$, $\Delta_{\text{CAC}} = -5.22$), driven by consistent negative contributions across multiple layers, most prominently L24.MLP (-2.44) and L22.H6 (-1.39). Despite this strong abstention signal in the CAC, the surface margin is marginally positive ($\Delta(x) = +1.37$), and Zero-Threshold commits. The MLP reads the CAC consensus and correctly abstains ($z_{\text{MLP}} = 0.167$).

Comparison Both cases are instances of unsupported commitment where $\Delta(x) > 0$ masks component-level abstention signals. In Case 1, the surface margin is inflated by a single large outlier contribution (L19.MLP, $+8.27$) that overwhelms the remaining CAC margin; $\Delta_{\text{CAC}} = +4.53$ only because of this component. In Case 2, the CAC margin is unambiguously negative ($\Delta_{\text{CAC}} = -5.22$) yet the surface margin is marginally positive ($\Delta(x) = +1.37$), concealing the abstain consensus from Zero-Threshold. In both cases, the MLP reads the full CAC decomposition and correctly abstains where Zero-Threshold does not.

G Computational Cost

Table 21 reports estimated wall-clock times (minutes) on a single NVIDIA H100 80GB for each pipeline stage across all twelve models.

Table 18: Accuracy across 10 models and 3 in-distribution datasets. **Bold**: best in row for each dataset section. *Mean*: averaged over all 3 datasets.

	Qwen 3.5		Gemma 3		Llama 3		Ministral		Phi-4		
	4B	9B	4B	12B	3B	8B	3B	14B	mini	14B	*Mean*
KUQ											
Zero-Threshold	.618	.646	.656	.818	.662	.686	.722	.778	.748	.740	*.707*
CAC LR	.878	.872	.884	.914	.856	.854	.908	.910	.888	.910	*.887*
CAC MLP	**.912**	**.904**	**.928**	**.946**	**.908**	**.930**	**.922**	**.936**	**.902**	**.944**	***.923***
SQuAD 2.0											
Zero-Threshold	.598	.768	.588	.724	.540	.664	.756	.690	.648	.678	*.665*
CAC LR	**.758**	.830	**.640**	**.790**	**.626**	**.758**	.752	**.772**	.698	**.802**	***.743***
CAC MLP	.770	**.850**	.636	.788	.620	.744	**.766**	.762	**.710**	.802	*.745*
MuSiQue											
Zero-Threshold	.750	.676	.688	.768	.668	.602	.664	.790	.720	.716	*.704*
CAC LR	.798	**.818**	**.736**	.786	.716	.774	**.760**	.758	.706	**.786**	*.764*
CAC MLP	**.806**	.812	.732	**.800**	**.736**	**.786**	.758	**.790**	**.744**	.774	***.774***
Mean (all datasets)											
Zero-Threshold	.655	.697	.644	.770	.623	.651	.714	.753	.705	.711	*.692*
CAC LR	.811	.840	.753	.830	.733	.795	.807	.813	.764	.833	*.798*
CAC MLP	**.829**	**.855**	**.765**	**.845**	**.755**	**.820**	**.815**	**.829**	**.785**	**.840**	***.814***

Table 19: Top-20 CAC component contributions, Case 1 (Llama 3B, KUQ), ordered by layer. Contrib.: $v_k(x)$, signed contribution to $\Delta(x)$. $\Delta(x) = +7.94$; $\Delta_{\text{CAC}} = +4.53$ ($\sum^- = -8.09$, $\sum^+ = +12.62$).

Component	Contribution	Component	Contribution
L10.MLP	$+0.323$	L23.H17	$+0.561$
L12.MLP	$+0.218$	L24.MLP	$+1.003$
L13.MLP	$+0.644$	L25.H16	-0.230
L18.MLP	-0.405	L25.H17	$+0.120$
L19.MLP	$+8.272$	L26.MLP	-0.671
L21.MLP	-2.286	L26.H0	-0.334
L21.H18	-0.351	L26.H8	-0.289
L23.MLP	$+0.777$	L26.H10	-0.209
L23.H14	$+0.415$	L26.H18	-0.580
L27.MLP	-2.734	L27.H19	$+0.285$

Table 20: Top-20 CAC component contributions, Case 2 (Ministral 3B, SQuAD 2.0), ordered by layer. Contrib.: $v_k(x)$, signed contribution to $\Delta(x)$. $\Delta(x) = +1.38$; $\Delta_{\text{CAC}} = -5.22$ ($\sum^- = -7.62$, $\sum^+ = +2.40$).

Component	Contribution	Component	Contribution
L5.MLP	-0.246	L22.H18	-0.234
L9.MLP	$+0.171$	L23.H10	$+0.149$
L16.MLP	$+0.182$	L23.H24	$+0.359$
L20.MLP	-0.199	L24.MLP	-2.437
L21.MLP	-0.379	L24.H7	-0.284
L22.MLP	-0.354	L24.H9	$+0.376$
L22.H1	-0.881	L24.H19	-0.305
L22.H6	-1.393	L24.H26	$+0.390$
L25.MLP	-0.711	L25.H28	$+0.136$
L25.H1	$+0.637$	L25.H12	-0.201

NeurIPS Paper Checklist

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction state our three contributions: framing unsupported commitment, identifying the CAC, and a circuit-informed abstention policy. Each is supported in Sections 3, 4, and 5 across 30 in-distribution and 24 out-of-distribution configurations.

   Guidelines:

   - The answer [N/A] means that the abstract and introduction do not include the claims made in the paper.

Table 21: Estimated wall-clock time (minutes) per pipeline stage on a single NVIDIA H100 80GB (1,500 data instances, 10 seeds for CAC localisation).

	Qwen 3.5			Gemma 3			Llama 3		Ministral		Phi-4		
	4B	9B	35B	4B	12B	27B	3B	8B	3B	14B	mini	14B	**Total**
$\Delta(x)$ evaluation	6.2	13.4	20.6	6.7	18.1	39.7	4.3	11.0	4.1	16.9	4.6	18.2	163.8
CAC localisation	16.2	35.8	50.4	11.3	38.6	77.9	9.1	28.7	8.4	31.2	9.3	38.9	355.8
Ablation	3.8	9.7	12.1	4.1	11.6	27.3	3.2	8.4	2.9	11.8	3.1	12.2	110.2
Amplification	4.9	11.8	14.3	5.3	14.7	34.6	3.7	10.6	3.8	14.9	4.2	15.7	138.5
Policy training	1.4	2.8	3.6	1.6	3.7	8.6	1.1	2.6	0.9	3.5	0.8	3.9	34.5
Total	**32.5**	**73.5**	**101.0**	**29.0**	**86.7**	**188.1**	**21.4**	**61.3**	**20.1**	**78.3**	**22.0**	**88.9**	**802.8**

- The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A [No] or [N/A] answer to this question will not be perceived well by the reviewers.
- The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
- It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: Limitations are discussed in App. **??**.

   Guidelines:

   - The answer [N/A] means that the paper has no limitation while the answer [No] means that the paper has limitations, but those are not discussed in the paper.
   - The authors are encouraged to create a separate "Limitations" section in their paper.
   - The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
   - The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
   - The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
   - The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
   - If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
   - While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

Answer: [N/A]

Justification: The paper does not present formal theorems requiring proofs.

Guidelines:

- The answer [N/A] means that the paper does not include theoretical results.
- All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
- All assumptions should be clearly stated or referenced in the statement of any theorems.
- The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
- Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
- Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

Answer: [Yes]

Justification: Section 4.1 describes datasets, models, splits, decoding settings, and hardware. Section 5 specifies the MLP architecture, optimiser, training schedule, and threshold-calibration procedure. App. B and E.1 provide the full token lists and baseline configurations. Code is provided in the supplementary material.

Guidelines:

- The answer [N/A] means that the paper does not include experiments.
- If the paper includes experiments, a [No] answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.
- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [Yes]

   Justification: All datasets (KUQ, SQuAD 2.0, MuSiQue, HotpotQA, SelfAware) used are publicly available. All evaluated models are open-weight. We provide an anonymised code release with reproduction instructions in the supplementary material.

   Guidelines:

   - The answer [N/A] means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so [No] is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://neurips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

   Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer) necessary to understand the results?

   Answer: [Yes]

   Justification: Section 4.1 reports datasets, splits, decoding settings, and hardware. App. D.2 specifies the causal gating hyperparameters. Section 5 reports the MLP architecture, optimiser, and training schedule. Baseline configurations are in App. E.1. Code is provided in the supplementary material.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
   - The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

   Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

   Answer: [Yes]

   Justification: For all experiments, we report mean values, per-configuration breakdowns, and ranges across 30 in-distribution and 24 out-of-distribution configurations. We report sign-test p-values for headline statistical claims throughout the paper.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.

- The authors should answer [Yes] if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g., negative error rates).
- If error bars are reported in tables or plots, the authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

   Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

   Answer: [Yes]

   Justification: All experiments run on a single NVIDIA H100 GPU (Section 4.1). Detailed computation costs for each experiment and model are provided in App. G.

   Guidelines:

   - The answer [N/A] means that the paper does not include experiments.
   - The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
   - The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
   - The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: The research uses publicly available datasets and open-weight models (Section 4.1), involves no human subjects, and aims to improve LLM honesty by enabling abstention. We have reviewed and conformed with the NeurIPS Code of Ethics.

   Guidelines:

   - The answer [N/A] means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer [No], they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

Answer: [Yes]

Justification: This work aims to enhance LLM safety, reliability, and interpretability by enabling models to abstain on questions they cannot answer reliably. We do not foresee significant direct negative societal impacts.

Guidelines:

- The answer [N/A] means that there is no societal impact of the work performed.
- If the authors answer [N/A] or [No], they should explain why their work has no societal impact or why the paper does not address societal impact.
- Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
- The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate Deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
- The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
- If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pre-trained language models, image generators, or scraped datasets)?

Answer: [N/A]

Justification: The paper does not release pre-trained models, image generators, or scraped datasets.

Guidelines:

- The answer [N/A] means that the paper poses no such risks.
- Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
- Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
- We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

Answer: [Yes]

Justification: All datasets and their licenses are listed in Table 4 (Section 4.1). All open-weight models (Qwen3.5, Gemma 3, Llama-3, Ministral, Phi-4) are cited and used under their respective licenses.

Guidelines:

- The answer [N/A] means that the paper does not use existing assets.
- The authors should cite the original paper that produced the code package or dataset.
- The authors should state which version of the asset is used and, if possible, include a URL.
- The name of the license (e.g., CC-BY 4.0) should be included for each asset.
- For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
- If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
- For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
- If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [Yes]

    Justification: Code is provided in the supplementary material with documentation and reproduction instructions. It will be released publicly upon publication under the CC BY 4.0 license.

    Guidelines:

    - The answer [N/A] means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.
    - At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [Yes]

    Justification: The abstention set $\mathcal{A}$ was validated via Amazon Mechanical Turk: two workers independently labelled each candidate token with its 10-token continuation. Workers were compensated at \$20/hour, above the US federal minimum wage. See App. B for details.

    Guidelines:

    - The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

Answer: [No]

Justification: The Amazon Mechanical Turk annotation task involved labelling short text snippets with no personal data collection or sensitive content. Under our institution's research policy, low-risk text annotation tasks of this kind do not require IRB review.

Guidelines:

- The answer [N/A] means that the paper does not involve crowdsourcing nor research with human subjects.
- Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
- We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
- For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does *not* impact the core methodology, scientific rigor, or originality of the research, declaration is not required.

Answer: [N/A]

Justification: LLMs are the object of study in this work, not a tool used in its development. They were not used as a component of the core methodology, scientific rigour, or originality of the research.

Guidelines:

- The answer [N/A] means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
- Please refer to our LLM policy in the NeurIPS handbook for what should or should not be described.