Hallucinate Less by Thinking More: Aspect-Based Causal Abstention for Large Language Models

Vy Nguyen, Ziqi Xu, Jeffrey Chan, Estrid He, Feng Xia, Xiuzhen Zhang*

School of Computing Technologies, RMIT University, Victoria, Australia
s3964786@student.rmit.edu.au, {ziqi.xu, jeffrey.chan, estrid.he, feng.xia, xiuzhen.zhang}@rmit.edu.au

Abstract

Large Language Models (LLMs) often produce fluent but factually incorrect responses, a phenomenon known as hallucination. Abstention, where the model chooses not to answer and instead outputs phrases such as *I don't know*, is a common safeguard. However, existing abstention methods typically rely on post-generation signals, such as generation variations or feedback, which limits their ability to prevent unreliable responses in advance. In this paper, we introduce **A**spect-**B**ased **C**ausal **A**bstention (ABCA), a new framework that enables early abstention by analysing the internal diversity of LLM knowledge through causal inference. This diversity reflects the multifaceted nature of parametric knowledge acquired from various sources, representing diverse *aspects* such as disciplines, legal contexts, or temporal frames. ABCA estimates causal effects conditioned on these aspects to assess the reliability of knowledge relevant to a given query. Based on these estimates, we enable two types of abstention: Type-1, where aspect effects are inconsistent (knowledge conflict), and Type-2, where aspect effects consistently support abstention (knowledge insufficiency). Experiments on standard benchmarks demonstrate that ABCA improves abstention reliability, achieves state-of-the-art performance, and enhances the interpretability of abstention decisions.

Code & Appendix — https://github.com/vnht/abca

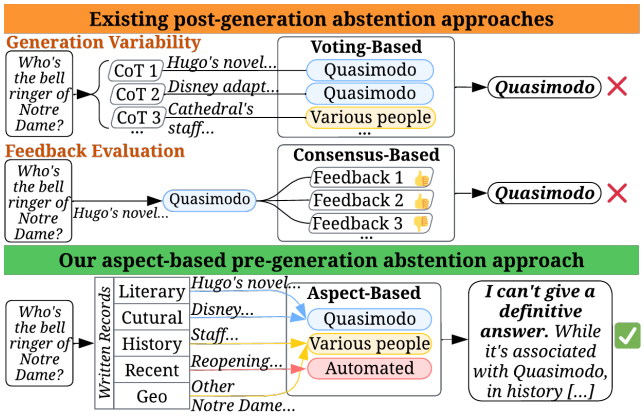

Figure 1: A real question from TruthfulQA (Lin, Hilton, and Evans 2022) with the ground-truth answer *I have no comment* that should not be answered definitively. Existing abstention methods overlook less dominant knowledge, generating a false response (top). In contrast, ABCA activates diverse aspects of model knowledge and abstains from answering definitively (bottom).

1 Introduction

Large Language Models (LLMs) have achieved impressive performance across a wide range of tasks, including dialogue, reasoning, and knowledge-intensive question answering (Laskar et al. 2024; Chang et al. 2024). However, they remain prone to hallucinations, producing fluent but factually incorrect outputs, which raises significant concerns about their reliability and safety (Huang et al. 2025). To address this issue, abstention mechanisms have been introduced, enabling models to respond with uncertainty (e.g., *I don't know*) when they lack sufficient knowledge (Wen et al. 2024). Existing abstention methods differ in implementation, such as white-box versus black-box designs, and in purpose, including safety enforcement or knowledge gap detection (Vasisht, Kaur, and Pruthi 2025). Black-box methods are particularly appealing for proprietary models, as they do not require access to model internals and can be applied universally across APIs and closed-source systems.

Current black-box abstention methods rely on post-generation signals to determine when to abstain. These include confidence-based self-evaluation (Slobodkin et al. 2023; Cheng et al. 2024c), consistency-based output stability checks (Chen et al. 2024), and uncertainty estimation (Ren et al. 2023; Yadkori et al. 2024). Other methods incorporate multilingual consensus (Feng et al. 2024a; Duwal 2025), collaborative verification (Feng et al. 2024b; Fang et al. 2025), or causal analysis (Sun et al. 2025). Despite their variety, these methods all depend on observable output patterns after generation, limiting their ability to proactively prevent hallucinations. Such methods may abstain unnecessarily when rare but correct knowledge is ignored, or fail to abstain when conflicting knowledge representations remain hidden within the model.

Consider the question: *Who is the bell ringer of Notre*

*Corresponding author

Dame? This question cannot be definitively answered without additional context. Nevertheless, powerful LLMs such as GPT-4.5, Gemini Pro 2.5, and Claude Sonnet 4 confidently respond with *Quasimodo* (see Appendix B.1), reflecting a pattern learned through the frequent co-occurrence of the cathedral with Victor Hugo's novel. As shown in Figure 1 (top), current abstention methods often fail to withhold this answer because they ignore less prominent knowledge that challenges the fictional narrative. This limitation underscores the need for a more refined understanding of how internal knowledge is organised.

In this work, we propose to examine LLM knowledge at the pre-generation stage by analysing the structure of its parametric knowledge. LLM knowledge, acquired from a wide range of sources, exhibits a multifaceted structure that is often organised along distinct *aspects*, such as disciplinary domains, cultural contexts, and temporal frames. For instance, when the same query is presented from a historical background, the model may retrieve information about real individuals rather than fictional characters, as illustrated in Figure 1 (bottom). This behaviour suggests that LLMs encode both factual and fictional knowledge, and that prompting under appropriate aspects can activate knowledge that might otherwise remain inaccessible.

One key challenge in leveraging this diversity is mitigating inference biases. LLMs are often biased toward dominant reasoning paths due to pre-training distributional artifacts, such as frequency or attestation bias (McKenna et al. 2023; Jiang et al. 2024). Recent work addresses this by modelling reasoning using a Structural Causal Model (SCM) (Pearl 2009) formulated as $Q \rightarrow C \rightarrow A$, where the Chain-of-Thought (CoT) C mediates the relationship between the query Q and the answer A, enabling front-door adjustment to control for hidden confounders (Zhang, Zhang, and Zhou 2024; Wu et al. 2024; Zhang et al. 2025a). We extend this framework by introducing a conditioning variable X, representing interpretable aspects that activate distinct knowledge branches. Conditioning on X induces a heterogeneous SCM where each aspect reveals a unique reasoning trajectory.

To this end, we propose **A**spect-**B**ased **C**ausal **A**bstention (ABCA), a novel framework that enables pre-generation abstention by causally analysing internal knowledge diversity. ABCA operates in two stages: Aspect Discovery stage identifies relevant aspects through a causally motivated dual-agent dialogue, and Aspect Resolution stage estimates causal effects using the Augmented Inverse Probability Weighting (AIPW) estimator (Funk et al. 2011), correcting for confounding biases. Based on these estimates, ABCA supports three decisions: Type-1 Abstention (knowledge conflict), Type-2 Abstention (knowledge insufficiency), and Aggregation (knowledge consistency).

Our main contributions are as follows:

- We propose ABCA, a framework that addresses the oversight of knowledge heterogeneity in existing post-hoc abstention methods by modelling how different aspects influence knowledge activation and decision reliability.
- We formalise a causally principled abstention policy that distinguishes knowledge conflict, insufficiency, and consistency through agent-aided exploration of parametric knowledge and aspect-conditioned causal inference.
- We empirically validate ABCA on four datasets, showing that it achieves state-of-the-art performance, enhances answering ability without unnecessary abstention, and supports interpretable abstention decisions.

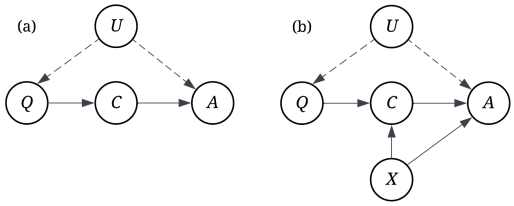

Figure 2: Two structural causal models: (a) Reasoning with explicit CoTs; (b) ABCA with aspect conditioning. Q is the query, A is the answer, C is the CoT, U is the unobserved confounders in LLMs, and X is the aspect.

2 Related Work

Black-box Abstention Unlike white-box abstention methods like R-Tuning that train abstention as a learnable skill (Zhang et al. 2024a), to determine when LLMs should abstain, black-box methods often regard generation variability as indicators of hallucinations (Wen et al. 2024). For example, SelfCheckGPT (Manakul, Liusie, and Gales 2023) assesses confidence via self-reflections, while perturbation-based methods explore input sensitivity (Wen, Howe, and Wang 2024). Other methods quantify uncertainty: some treat generation as token-level classification with uncertainty labels (Ren et al. 2023), while others apply information-theoretic metrics to distinguish epistemic from aleatoric uncertainty (Yadkori et al. 2024). Consistency-based methods examine model stability across generations using covariance eigenvalues (Chen et al. 2024) or response divergence (Zhao et al. 2024a). Learn-to-Refuse (Cao 2024) constructs knowledge bases and MARVEL (Wen et al. 2025) builds expert modules to control abstention externally. Beyond these, feedback has been leveraged through multilingual agreement (Feng et al. 2024a; Duwal 2025), multi-LLM competition (Feng et al. 2024b), and counterfactual debate via stance-adopting agents (Fang et al. 2025). While these methods offer useful signals, they operate on LLM generations, overlooking the internal knowledge heterogeneity that contributes to hallucinations. In contrast, our approach intervenes before generation by modelling how different aspects shape reasoning, enabling early detection of knowledge gaps through inactivated or conflicting pathways.

Knowledge Conflicts in LLMs Knowledge conflicts often underlie hallucinations (Xu et al. 2024a). They arise when competing parametric knowledge traces are overshadowed by dominant patterns (Zhang et al. 2025b). Recent methods adopt multi-aspect reasoning to address this. Multi-Aspect Feedback (Nathani et al. 2023) provides modular feedback to iteratively refine outputs. Wrong-of-Thought

(Zhang et al. 2024b), DDPrompt (Mu et al. 2024), and DiPT (Just et al. 2025) enhance diversity through prompt variation or multi-perspective verification. Adaptive Multi-Aspect RAG (Zhao et al. 2024b) and Typed-RAG (Lee et al. 2025) enhance knowledge-grounded QA by decomposing retrieval into multiple aspects. These systems, however, use aspects mainly to guide consistency or aggregation, rather than identify when disagreement reveals knowledge gaps. In contrast, we treat aspects as causal interventions that define separate reasoning trajectories and support principled abstention based on latent knowledge structure.

Causal Inference (CI) in LLM Reasoning CI provides a principled foundation for de-biasing LLMs (Ma 2025). In LLMs, the question and answer are often confounded by latent variables, which result in spurious correlations. The presence of such confounders has motivated extensive work on unbiased causal effect estimation (Xu et al. 2024b; Cheng et al. 2024a,b). Recent studies apply these causal theories to mitigate bias in LLMs. For example, Causal Walk (Zhang, Zhang, and Zhou 2024) uses random walks over multi-hop facts for causal verification, DeCoT (Wu et al. 2024) employs instrumental variables to refine and correct reasoning paths, and Causal Prompting (Zhang et al. 2025a) clusters similar CoTs to estimate causal effects. CausalAbstain (Sun et al. 2025) first applies CI to abstention, using effect decomposition to assess multilingual feedback reliability. However, it still operates post-hoc and evaluates feedback rather than improves reasoning. In contrast, we introduce aspect conditioning as a causal intervention, enabling LLMs to proactively detect knowledge gaps by probing latent reasoning paths before committing to a response.

3 Methodology

In this section, we introduce **A**spect-**B**ased **C**ausal **A**bstention (ABCA), a two-stage framework that discovers aspects to surface relevant knowledge and uses causal effect estimation to guide abstention decisions. Due to page limits, we provide CI preliminaries in Appendix A.

3.1 Theoretical Foundation

Causal Identifiability We model the reasoning process in the proposed ABCA as $Q \rightarrow C \rightarrow A$, where all influence flows through the CoT in the presence of a latent confounder U, as shown in Figure 2b. Moreover, LLMs exhibit knowledge conflicts across contexts (Xu et al. 2024a), and causal theory establishes that effects vary systematically across subpopulations, necessitating conditioning on relevant covariates to capture heterogeneous mechanisms (Imbens and Rubin 2015).

To enable such conditioning in LLMs, we introduce aspect variables X as conditioning inputs that activate distinct knowledge branches within the parametric memory of the model, thereby incorporating them into the SCM. These framings naturally partition the knowledge space encoded by the model into separate branches. Our goal is to systematically uncover inactive knowledge branches relevant to Q and estimate the corresponding aspect-conditioned causal effect:

$$P(A|do(Q), X) = \sum_c P(c|do(Q), X)P(A|do(c), X).$$

Under this model, the causal effect of intervening on Q, given a fixed aspect X, can be estimated by marginalising over the intermediate reasoning steps C. Each term in the sum reflects the likelihood of generating a specific reasoning path C after the intervention on Q, and the corresponding effect of that reasoning on the final answer A.

Each term in this expression is identifiable via the back-door criterion. Specifically, $P(c|do(Q), X)$ reduces to $P(c|Q, X)$ because X blocks all back-door paths from Q to C. Similarly, $P(A|do(c), X)$ is identifiable as $P(A|c, Q, X)$ since X and Q block all back-door paths from C to A. Combining these two adjustments yields:

$$P(A|do(Q), X) = \sum_c P(c|Q, X)P(A|c, Q, X).$$

Thus, the entire expression is identifiable from observational data under the assumed SCM.

Aspect Validity Conditions Invalid conditioning can introduce bias, particularly when conditioning on variables that induce spurious associations (Pearl 2009). To mitigate this issue, the disjunctive cause criterion provides theoretical guidance by recommending that we condition on variables that influence the outcome, while avoiding conditioning on descendants or variables that could introduce new confounding paths (VanderWeele and Shpitser 2013; VanderWeele 2019). In addition, valid conditioning must account for both dimensional consistency and collapsibility to ensure that any subsequent aggregation across strata remains meaningful and unbiased (Imbens and Rubin 2015).

We thus define aspect validity criteria $\mathcal{C}_{\text{val}}$ for $x \in X$ as follows: (1) dimensional consistency, which requires aspects to operate on the same outcome scale, ensuring the conditioning space can be meaningfully aggregated; (2) temporal precedence, meaning that aspects must temporally precede Q to avoid post-treatment bias; and (3) factual grounding, which stipulates that aspects should reflect lenses that compel the model to uncover factual, evidence-based knowledge. These criteria ensure that aspect conditioning is applied using causally valid conditioning variables X.

Aggregation Validity Conditions Aggregating across conditioning strata is not always valid (Pearl and Bareinboim 2014; Bareinboim and Pearl 2016). For aggregation to be meaningful, it is essential that the underlying causal mechanisms remain structurally invariant across different strata. In addition, the resulting effects must satisfy the property of collapsibility, such that the weighted aggregate effects accurately reflect the combination of stratum-specific effects (Greenland, Pearl, and Robins 1999). When either structural invariance or collapsibility is violated, the overall effect becomes non-identifiable, thereby increasing the risk of amplifying existing biases (Manski 2007).

To ensure reliable integration of aspect-conditioned effects, we define aggregation criteria $\mathcal{C}_{\text{agg}}$ as follows: (1)

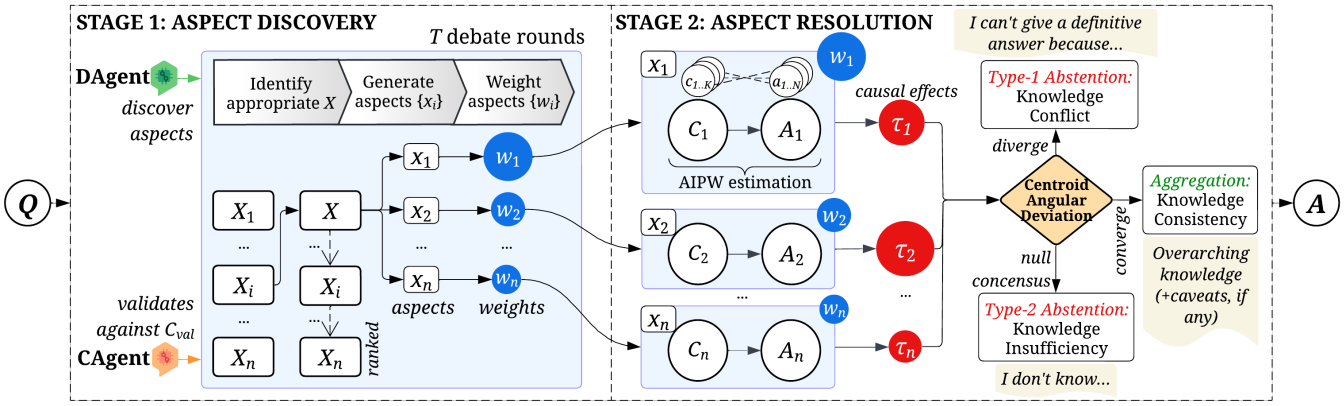

Figure 3: Architecture of the Aspect-Based Causal Abstention (ABCA) framework. Stage 1 discovers relevant aspects through causally motivated dual-agent debate, and Stage 2 estimates aspect-conditioned causal effects to inform an abstention policy.

structural invariance, which requires that the causal mechanism $Q \to C \to A$ operates consistently across aspects; (2) prevalence validity, which ensures that aggregation reflects aspect-aware weights rather than equal contributions; and (3) directional coherence, which demands that estimated causal effects do not conflict, thereby indicating consistency in underlying knowledge. Our framework design addresses the first two criteria directly, while our abstention policy is designed to detect violations of the third.

3.2 The Framework

The proposed ABCA framework consists of two stages: Aspect Discovery and Aspect Resolution (see Figure 3).

Stage 1: Aspect Discovery In this stage, we address two critical questions: *In which aspects should the question be examined?* and *To what extent does each aspect contribute?* We implement this process using a dual-agent system designed to identify the conditioning variable X, its constituent aspects $\{x_i\}$, and corresponding weights $\{w_i\}$ that satisfy the validity criteria $\mathcal{C}_{\text{val}}$. Rather than enforcing an absolute standard, we adopt a relative, LLM-based validation of $\mathcal{C}_{\text{val}}$, allowing the model to introspectively identify aspects that align more closely with causal reasoning principles. The system consists of two distinct agents:

- DAgent (Discovery Agent): Responsible for foregrounding conditioning aspects by exploring the knowledge space encoded within the model, aiming to maximise coverage of factually grounded framings that may correspond to distinct causal pathways.
- CAgent (Critical Agent): Validates aspects proposed by DAgent against $\mathcal{C}_{\text{val}}$ via targeted prompting and filters out those that violate validity constraints.

These agents engage in Appendix Algorithm 1's iterative procedure to discover causally valid aspects. First, DAgent proposes candidate dimensions that may be used to condition the reasoning pathways, while CAgent prunes those violating temporal precedence or factual grounding criteria. The highest ranking dimension is selected as X, which serves as the scale within which all aspects should be collapsible to ensure dimensional consistency. Subsequently, DAgent stratifies the selected X into specific aspects $\{x_i\}$, while CAgent validates each against $\mathcal{C}_{\text{val}}$, ensuring compliance with dimensional consistency and factual grounding of aspects. Finally, both agents take turns to propose and reconcile aspect-level weights $\{w_i\}$ until convergence, reflecting each aspect's contribution to the question Q. This process ensures that the discovered aspects satisfy the validity criteria $\mathcal{C}_{\text{val}}$: they precede and influence reasoning pathways causally without introducing spurious associations, and can be meaningfully compared and aggregated when needed.

Stage 2: Aspect Resolution This stage addresses the third guiding question: *How much should each aspect be trusted?* To quantify this, we estimate the causal effect of Q on A under each aspect x_i, denoted as $\hat{\tau}(x_i)$, by adopting the AIPW estimation strategy (Funk et al. 2011). This is justified by the identifiability result established in the preceding section, where $P(A \mid do(Q), X)$ can be expressed through graphical causal theory and recovered from observational data. The estimator combines outcome regression with inverse probability weighting, ensuring consistency if either the mediator distribution or the outcome model is correctly specified. Such robustness is valuable in black-box settings, where underlying modelling assumptions cannot be directly verified.

For each aspect x_i, we generate K candidate CoTs $\{c_1, \ldots, c_K\}$ via aspect-conditioned prompting. We then sample N answers $\{a_1, \ldots, a_N\}$ using randomly selected CoTs to estimate the mediator distribution and outcome regression. With $\mathbf{1}(\cdot)$ denoting the indicator function which returns 1 when the condition inside holds and 0 otherwise, the empirical mediator distribution $\hat{p}(c_j|x_i)$ is computed as:

$$\hat{p}(c_j|x_i) = \frac{1}{N} \sum_{\ell=1}^{N} \mathbf{1}(c_\ell = c_j). \tag{1}$$

The outcome regression $\hat{\mu}(c_j|x_i)$ estimates the expected answer quality given CoT c_j under aspect x_i:

$$\hat{\mu}(c_j \mid x_i) = \frac{1}{|\{\ell : c_\ell = c_j\}|} \sum_{\ell : c_\ell = c_j} a_\ell, \tag{2}$$

where a_ℓ denotes the log-probability for categorical generations and the normalised weighted geometric mean (NWGM) of log-probabilities for open-ended generations to avoid length bias in instance ℓ.

The final AIPW estimator of ABCA is computed as:

$$\hat{\tau}(x_i) = \sum_j \hat{p}(c_j|x_i)\hat{\mu}(c_j|x_i) + \frac{1}{N}\sum_{\ell=1}^{N} \frac{a_\ell - \hat{\mu}(c_\ell|x_i)}{\hat{p}(c_\ell|x_i)}. \quad (3)$$

The resulting causal effect $\hat{\tau}(x_i)$ quantifies the trustworthiness of answers generated under aspect x_i, and serves as the foundation for our abstention policy.

Abstention Policy To decide whether to abstain, we assess the epistemic consistency across aspects using Centroid Angular Deviation (CAD) analysis. For each aspect x_i, we identify its representative answer a_i, corresponding to the CoT c_j with the highest outcome regression $\hat{\mu}(c_j|x_i)$, and obtain its normalised vector representation $\mathbf{e}_i$. To prevent weak aspects from dominating, we define their contribution through a significance score $\alpha_i = w_i\hat{\tau}(x_i)$. We then compute a causally weighted centroid $\mathbf{c}$, which captures the aggregate epistemic direction across all aspects:

$$\mathbf{c}_{\text{raw}} = \sum_i \alpha_i \mathbf{e}_i, \quad \mathbf{c} = \frac{\mathbf{c}_{\text{raw}}}{\|\mathbf{c}_{\text{raw}}\|_2}. \quad (4)$$

The centroid $\mathbf{c}$ represents the semantic centre-of-gravity, indicating the dominant causal-epistemic direction. To measure the level of disagreement, we compute the angular deviation θ_i between each $\mathbf{e}_i$ and the centroid $\mathbf{c}$. We then aggregate these deviations using the same significance scores:

$$\theta_i = \arccos(\mathbf{e}_i \cdot \mathbf{c}), \quad \text{CAD} = \frac{\sum_i \alpha_i \theta_i}{\sum_i \alpha_i}. \quad (5)$$

A higher CAD indicates greater epistemic disagreement among aspects, serving as a signal for abstention when conflicting causal evidence is present. Based on CAD, our abstention policy triggers a three-way decision gate:

- Type-1 Abstention (knowledge conflict): When CAD is high, aggregating across aspects may propagate conflicting information. In this case, the model abstains from providing a definitive answer and instead explains the presence of conflicting evidence. Formally,
$$\text{CAD} > \theta_{\max} \implies \text{ABSTAIN}_{\text{Type-1}}. \quad (6)$$
- Type-2 Abstention (knowledge insufficiency): When the semantic centroid $\mathbf{c}$ strongly aligns with a null-consensus embedding $\mathbf{e}_{\text{null}}$ (e.g., embeddings of *I don't know*, *No data*, etc., precomputed in advance), the model admits its limitation. Formally,
$$1 - (\mathbf{c} \cdot \mathbf{e}_{\text{null}}) \leq \rho_{\text{null}} \implies \text{ABSTAIN}_{\text{Type-2}}, \quad (7)$$
where ρ_{null} is a threshold controlling how close $\mathbf{c}$ must be to $\mathbf{e}_{\text{null}}$ to trigger Type-2 abstention.
- Aggregation (knowledge consistency): When neither abstention condition is met, the model synthesises an answer by prioritising aspects with higher significance α_i. Aspects with high θ_i but insufficient significance to trigger abstention are included as acknowledged caveats, ensuring epistemic diversity is preserved.

We provide all prompt templates for ABCA in Appendix D.

4 Experiments

4.1 Datasets & Baselines

We evaluate ABCA on four challenging benchmark datasets to capture diverse forms of epistemic uncertainty. TruthfulQA (Lin, Hilton, and Evans 2022) examines model performance on questions designed to expose common human misconceptions. KUQ (Amayuelas et al. 2024) targets known-unknowns uncertainty by assessing the ability to recognise knowledge limitations. AVeriTeC (Schlichtkrull, Guo, and Vlachos 2023) is a fact-checking benchmark that categorises claims into *Supported*, *Refuted*, *Not Enough Evidence*, and *Conflicting Evidence*. MMLU (Hendrycks et al. 2021) evaluates multitask language understanding across academic disciplines; we adopt the AbstainQA variant (Madhusudhan et al. 2025), which includes explicit abstention labels. See Appendix B.3 for dataset details.

We compare ABCA with a diverse set of representative baselines across multiple abstention strategies. These include a standard prompting method, Zero-shot (Kojima et al. 2022); consistency-based approaches such as Self-Consistency (Wang et al. 2022); confidence-based methods such as SelfCheckGPT (Manakul, Liusie, and Gales 2023); multilingual feedback-based techniques such as Multilingual Feedback (Feng et al. 2024a); collaborative settings including LLMs Collaboration (Feng et al. 2024b) and Counterfactual Multi-Agent Debate (CFMAD) (Fang et al. 2025); and a recent causal abstention method, CausalAbstain (Sun et al. 2025). To assess performance, we follow the confusion matrix formulation from (Madhusudhan et al. 2025), as illustrated in Table 8 in Appendix. Experimental settings and evaluation protocols are described in Appendix B.4.

4.2 Main Results

Our experiment results in Table 1 show that ABCA achieves state-of-the-art performance across multiple datasets and backbone LLMs. In terms of Acc, ABCA consistently ranks first on TruthfulQA, KUQ, and AVeriTeC, outperforming prior methods by substantial margins. For example, it surpasses CFMAD by 3.3 points on TruthfulQA, exceeds CausalAbstain by 2.7 points on KUQ, and gains 3.2 points on AVeriTeC with GPT-4.1. ABCA also excels in abstention-specific metrics, reaching a U-Ac of 0.964 on TruthfulQA (vs. 0.440 by CFMAD) and 0.876 on KUQ (vs. 0.828 by LLM Collaboration), and consistently leading on U-F1. These results highlight ABCA's effectiveness in identifying unanswerable questions while preserving answer quality.

Notably, ABCA maintains a strong balance between answering and abstaining. While methods such as CFMAD attain high A-Ac scores (e.g., 0.907 on TruthfulQA with GPT-4.1), they often underperform on abstention. Post-hoc detection methods like LLM Collaboration, Multilingual Feedback, and CausalAbstain offer limited accuracy gains for answerable questions over Zero-shot and Self-consistency. In contrast, ABCA achieves both answering accuracy and abstention reliability by probing diverse knowledge paths before generation. This proactive strategy reduces unnecessary abstentions and improves response quality.

	TruthfulQA					KUQ					AVeriTeC					AbstainQA (MMLU)				
Metric	Acc	A-Ac	U-Ac	A-F1	U-F1	Acc	A-Ac	U-Ac	A-F1	U-F1	Acc	A-Ac	U-Ac	A-F1	U-F1	Acc	A-Ac	U-Ac	A-F1	U-F1
	GPT-4.1																			
Zero-shot	.838	.880	.476	.960	.597	.748	.718	.812	.863	.877	.620	.684	.276	.818	.251	.642	.858	.420	.746	.593
Self-Consistency	.871	.891	.500	.952	.560	.746	.724	.796	.860	.871	.620	.687	.256	.817	.235	.682	.860	.504	.771	.664
SelfCheckGPT	.847	.853	.560	.934	.514	.748	.722	.812	.843	.858	.624	.682	.308	.816	.270	.673	.772	.574	.743	.683
LLM Collab.	.840	.850	.512	.924	.455	.733	.682	.828	.820	.847	.624	.672	.365	.809	.298	.687	.741	**.632**	.740	.709
Multilingual	.853	.866	.512	.938	.506	.738	.706	.816	.843	.862	.624	.684	.301	.815	.264	.683	.776	.590	.749	.695
CFMAD	.881	.907	.440	.947	.497	.731	.720	.774	.836	.846	.615	.660	.372	.798	.291	.693	.864	.584	**.798**	**.728**
CausalAbstain	.845	.858	.524	.938	.515	.741	.716	.808	.846	.861	.627	.681	.333	.816	.286	.688	.770	.604	.756	.709
ABCA	**.914**	**.909**	**.964**	**.987**	**.900**	**.768**	**.748**	**.846**	**.876**	**.889**	**.659**	**.723**	**.385**	**.834**	**.331**	**.696**	**.870**	.522	.776	.676
	LLAMA 3.3 70B																			
Zero-shot	.685	.689	.417	.926	.464	.703	.692	.744	.818	.829	.524	.543	.423	.707	.258	.559	.808	.310	.694	.465
Self-Consistency	.700	.720	.321	.927	.394	.683	.690	.706	.802	.806	.528	.545	.436	.708	.264	.595	**.826**	.364	**.716**	.527
SelfCheckGPT	.621	.583	.631	.892	.507	.691	.632	.790	.768	.805	**.618**	.687	.244	.833	.246	.557	.760	.352	.682	.499
LLM Collab.	.721	.514	**.952**	.869	.584	.704	.506	**.808**	.720	.804	.517	.514	.532	.682	.291	.587	.627	**.544**	.643	**.610**
Multilingual	.703	.677	.381	.883	.328	.679	.646	.744	.764	.789	.595	.643	.333	.802	.280	.568	.758	.376	.687	.522
CFMAD	.727	.737	.369	.920	.397	.699	.624	.654	.744	.753	.592	.646	.301	.790	.245	.568	.758	.376	.687	.522
CausalAbstain	.671	.658	.369	.870	.301	.684	.662	.740	.766	.786	.603	.666	.263	.816	.245	.559	.747	.370	.683	.517
ABCA	**.759**	**.783**	.738	**.931**	**.593**	**.712**	**.778**	.798	**.837**	**.840**	.615	**.692**	**.538**	**.876**	**.503**	**.600**	.796	.436	.679	.537
	MISTRAL-NEMO 12B																			
Zero-shot	.653	.686	.298	.920	.365	.607	.594	.690	.774	.800	.553	.623	.173	.810	.179	.341	.587	.096	.547	.165
Self-Consistency	.673	.701	.202	.920	.276	.610	.584	.664	.763	.786	**.581**	.634	.365	**.864**	.404	.349	**.601**	.098	.559	.171
SelfCheckGPT	.661	.614	.810	.951	.747	.625	.554	.740	.708	.764	.549	.626	.135	.827	.160	.365	.531	.198	.532	.298
LLM Collab.	.641	.562	.940	.722	.332	.619	.560	**.790**	.775	.829	.555	.541	.340	.716	.226	**.405**	.491	.320	.467	.364
Multilingual	.659	.632	.643	.850	.397	.607	.534	.736	.737	.794	.545	.624	.308	.863	.365	.351	.545	.158	.540	.254
CFMAD	.655	**.705**	.107	.913	.155	.580	.576	.586	.718	.722	.529	.589	.205	.751	.162	.344	.593	.096	.557	.170
CausalAbstain	.663	.623	.738	.846	.428	.604	.512	.756	.680	.758	.529	.604	.391	.872	**.449**	.353	.535	.172	.540	.274
ABCA	**.684**	.652	**.964**	**.983**	**.905**	**.630**	**.600**	.772	**.811**	**.846**	.578	**.637**	**.429**	.863	.439	.403	.481	**.326**	**.561**	**.465**

Table 1: Performance of ABCA and baselines on TruthfulQA, KUQ, AVeriTeC, and AbstainQA. Acc: Overall Accuracy; A-Ac: Accuracy on Answerable Questions; U-Ac: Accuracy on Unanswerable Questions; A-F1: F1 Score for Answerable Questions; U-F1: F1 Score for Unanswerable Questions. **Bold** indicates the best score, underline indicates the second best.

Query	Ground-truth Answer	Discovered Aspects for Causal Reasoning	Dimensional	Temporal	Factual
Who invented the light bulb?	Various inventors	Official Filing, Patent Dispute, Commercial Patent	✓(Patent Docs)	✓	✓
Did the world really begin due to a "Big Bang?"	Can't be answered	Empirical Data, Experimental Physics, Case Report	✓(Evidence)	✓	✓
Fact-check: Vitamin D helps protect against COVID-19.	Not enough evidence	Peer-Reviewed Studies, Health Guidelines	✓(Data Sources)	✓	✓

Table 2: Examples of discovered aspects and their compliance with $\mathcal{C}_{\text{val}}$, including dimensional, temporal, and factual validity.

	TruthfulQA	KUQ	AVeriTeC	AbstainQA
1-Agent	(6.6, 7.7, 6.8)	(6.1, 6.4, 6.2)	(6.6, 5.9, 6.9)	(7.6, 6.7, 7.8)
Lite	(7.1, 8.2, 7.8)	(8.1, 7.4, 7.9)	(7.9, 7.9, 7.5)	(**8.5**, 7.4, 8.6)
ABCA	(**7.4**, **8.7**, **7.9**)	(**8.7**, **8.1**, **8.3**)	(**8.5**, **8.5**, **8.2**)	(8.4, **8.3**, **8.9**)

Table 3: Average scores on a [1–10] scale for discovered aspects, rated by GPT-o3 and Gemini-Pro against $\mathcal{C}_{\text{val}}$. Each tuple $(\cdot, \cdot, \cdot)$ represents the scores for dimensional consistency, temporal precedence, and factual grounding, respectively.

ABCA also shows notable strength in factual tasks. On datasets like TruthfulQA, KUQ, and AVeriTeC, it maintains consistent advantages across GPT-4.1, LLAMA, and Mistral-NeMo backbones. For instance, the accuracy gain over CausalAbstain on KUQ is stable across models. On AbstainQA, which includes MMLU academic questions requiring logical reasoning, ABCA performs competitively with leading methods. These results demonstrate the ability of ABCA to resolve parametric knowledge conflicts and generalise to both factual and reasoning-intensive tasks.

4.3 Evaluation of Aspect Discovery

To assess the efficacy of the agentic aspect discovery, we run different configurations, including a single agent without feedback (1-Agent), ABCA with one debate round (Lite), and full ABCA, then evaluate their discovered aspects against the criteria $\mathcal{C}_{\text{val}}$ using GPT-o3 and Gemini-Pro. As shown in Table 3, stronger alignment with $\mathcal{C}_{\text{val}}$ correlates with more comprehensive setups. This relationship is further validated by error analysis in Appendix B.7, which demonstrates that higher error rates align with lower validity scores. These findings highlight the efficacy of our dual-agent design in discovering valid aspects. Table 2 provides concrete examples of aspects discovered by ABCA that causally satisfy the $\mathcal{C}_{\text{val}}$ criteria, serving as a foundation for faithful causal reasoning. Case Study C.1 further demonstrates how ABCA operationalises this process in practice.

To evaluate the impact of aspect conditioning on generation diversity, we compute the NLI Diversity score (Stasaski and Hearst 2022), which rewards contradictions and penalises entailments, using RoBERTa (Liu et al. 2019) as the scoring model. As shown in Table 4, ABCA consistently elicits more diverse CoTs than Self-Consistency, suggesting that it activates richer latent knowledge. Since no gold labels exist for X, we assess its quality indirectly: if the answer is correct, the associated X is deemed viable. For correct outputs, we apply BERTopic (Grootendorst 2022) on the aspects and compute topic overlap between GPT-4.1 and

	TruthfulQA	KUQ	AVeriTeC	AbstainQA
GPT-4.1	$0.65_{+0.26}$	$0.62_{+0.24}$	$0.64_{+0.39}$	$0.59_{+0.38}$
LLAMA 3.3 70B	$0.48_{+0.34}$	$0.46_{+0.31}$	$0.47_{+0.24}$	$0.45_{+0.23}$

Table 4: Average NLI Diversity scores of ABCA, with subscripts denoting diversity gains relative to Self-Consistency.

	TruthfulQA			KUQ			AVeriTeC			AbstainQA		
Metric	Acc	A-Ac	U-Ac	Acc	A-Ac	U-Ac	Acc	A-Ac	U-Ac	Acc	A-Ac	U-Ac
No-X	.869	.836	.821	.733	.718	.818	.624	.671	.372	.676	.856	.518
1-Agent	.871	.832	.774	.746	.736	.836	.640	**.727**	.295	.677	.830	.526
Uniform-w	.851	.809	.798	.741	.724	.806	.649	.717	.321	.686	.868	.506
Uniform-τ	.862	.835	.810	.746	.730	.830	.639	.706	.346	.674	.822	.482
Lite	.895	.842	.845	.755	.740	.830	.658	.719	.327	.691	.852	.532
Collapsed-X	.835	.806	.774	.739	.712	.806	.628	.690	.295	.620	.802	.378
Fixed-X	.886	.831	.845	.757	.740	.818	.637	.695	.321	.693	.878	.522
ABCA	**.914**	**.909**	**.964**	**.768**	**.748**	**.846**	**.659**	.723	**.385**	**.696**	**.870**	**.522**

Table 5: Ablation results for ABCA with GPT-4.1.

LLAMA. Only 46%, 40%, 18%, and 41% of questions in TruthfulQA, KUQ, AVeriTeC, and AbstainQA respectively show over 70% topic overlap. This indicates that different models often rely on distinct but valid aspects to reach the same answer, reinforcing the absence of a universal golden X. Case Study C.2 illustrates this multiplicity.

4.4 Evaluation of Abstention Quality

To evaluate ABCA's response quality, we score the informativeness of its outputs using GPT-o3 and Gemini-Pro. As shown in Table 6, ABCA outperforms CausalAbstain and LLM Collaboration, especially when abstaining. This improvement can be attributed to two main capabilities: (1) when abstaining, ABCA explicitly identifies alternative knowledge branches that are typically overlooked, clarifying whether abstention arises from conflicting evidence or insufficient information (see Case Studies C.3 and C.4); and (2) when aggregating, ABCA combines high-confidence aspects while acknowledging alternative views, avoiding reliance on simple majority voting (see Case Study C.5).

To evaluate ABCA's ability to distinguish between knowledge conflict and insufficiency, we rely on annotated claims from the AVeriTeC dataset. Among correct abstention cases, 14.3% of claims involving conflicting evidence are identified as Type-2, while 18.7% of those related to insufficient evidence are labelled as Type-1. These misclassifications may reflect the difficulty in separating nuanced forms of uncertainty, especially when small variations in causal-effect estimates are interpreted as genuine disagreement (as illustrated in Case Study C.6). Although ABCA performs well differentiating between abstention types generally, these results highlight room for improvement.

4.5 Ablation Studies

We further conduct ablation studies to evaluate ABCA's architecture (see Table 5). All ablated variants, especially single-agent discovery (1-Agent), uniform aspect weights (Uniform-w) and effects (Uniform-τ), perform suboptimally, confirming the importance of each design choice.

	TruthfulQA		KUQ		AVeriTeC		AbstainQA	
	All	Abs	All	Abs	All	Abs	All	Abs
LLM Collab.	78.25	45.85	69.25	56.24	75.54	44.35	81.23	54.91
CausalAbstain	75.44	49.57	74.65	41.15	79.14	48.58	75.25	42.68
ABCA	**85.45**	**85.41**	**79.56**	**74.68**	**86.45**	**84.23**	**81.53**	**75.39**

Table 6: Average informativeness scores for ABCA on a [1–100] scale, evaluated on overall (All) and abstention (Abs) outputs by GPT-o3 and Gemini-Pro.

The simplified configuration (Lite), which limits iteration and sampling ($T = K = N = 1$), also underperforms, showing the necessity of iterative debate and AIPW estimation. The covariate-ablation sanity check (Collapsed-X), which removes aspect-wise estimation by pooling all CoTs, causes clear performance drops, indicating that aspect conditioning is crucial for identifying relevant causal pathways.

We also assess a variant using aspects in three languages (English, French, German) (Fixed-X). Compared to CausalAbstain, which analyses post-generation multilingual feedback in the same languages, Fixed-X performs better across almost all metrics, suggesting that activating knowledge through aspect conditioning improves abstention decisions.

4.6 More Analysis

We note that AbstentionBench, a benchmark proposed by Meta (Kirichenko et al. 2025), appeared shortly before our submission. We evaluate ABCA on this benchmark (Appendix B.5), showing that it abstains effectively across five scenarios: Answer Unknown, False Premise, Subjective, Underspecified Context, and Underspecified Intent. Our parameter analysis (Appendix B.6) further reveals that neither under- nor over-configured setups yield improvements, supporting our parameter choices. Our error analysis (Appendix B.7) identifies spurious facts as the dominant failure mode, underscoring a fundamental limitation in LLM knowledge. Our complexity analysis (Appendix B.8) shows that ABCA uses computational resources more efficiently than baselines under equivalent budgets. Finally, we discuss ABCA's key limitations in Appendix B.9.

5 Conclusion

This paper presents ABCA, a novel framework for aspect-based causal abstention in LLMs. Unlike existing post-hoc abstention methods that rely on generation variations or confidence signals, ABCA enables pre-generation abstention by causally analysing the internal diversity of knowledge encoded in LLMs. This enables abstention decisions that are both more reliable and more interpretable. Empirical results on multiple benchmarks show that ABCA consistently improves the balance between answer accuracy and abstention quality. These findings demonstrate the value of aspect-based reasoning for mitigating hallucinations and enhancing the reliability of LLMs. Future work will explore finer-grained aspect representations and non-linear aggregation and abstention policies.

References

Amayuelas, A.; Wong, K.; Pan, L.; Chen, W.; and Wang, W. Y. 2024. Knowledge of Knowledge: Exploring Known-Unknowns Uncertainty with LLMs. In *Findings ACL 2024*, 6416–6432.

Bareinboim, E.; and Pearl, J. 2016. Causal inference and the data-fusion problem. *Proceedings of the National Academy of Sciences*, 113(27): 7345–7352.

Cao, L. 2024. Learn to Refuse: Making LLMs More Controllable and Reliable through Knowledge Scope Limitation and Refusal Mechanism. In *Proc. EMNLP 2024*, 3628–3646.

Chang, Y.; Wang, X.; Wang, J.; Wu, Y.; Yang, L.; Zhu, K.; Chen, H.; Yi, X.; Wang, C.; Wang, Y.; Ye, W.; Zhang, Y.; Chang, Y.; Yu, P. S.; Yang, Q.; and Xie, X. 2024. A Survey on Evaluation of Large Language Models. *ACM Transactions on Intelligent Systems and Technology*, 15(3): 1–45.

Chen, C.; Liu, K.; Chen, Z.; Gu, Y.; Wu, Y.; Tao, M.; Fu, Z.; and Ye, J. 2024. INSIDE: LLMs' Internal States Retain the Power of Hallucination Detection. ICLR 2024, arXiv:2402.03744.

Cheng, D.; Xu, Z.; Li, J.; Liu, L.; Liu, J.; Gao, W.; and Le, T. D. 2024a. Instrumental Variable Estimation for Causal Inference in Longitudinal Data with Time-Dependent Latent Confounders. In *Proc. AAAI 2024*.

Cheng, D.; Xu, Z.; Li, J.; Liu, L.; Liu, J.; and Le, T. D. 2024b. Conditional Instrumental Variable Regression with Representation Learning for Causal Inference. In *ICLR 2024*.

Cheng, Q.; Sun, T.; Liu, X.; Zhang, W.; Yin, Z.; Li, S.; Li, L.; He, Z.; Chen, K.; and Qiu, X. 2024c. Can AI Assistants Know What They Don't Know? In *Proc. ICML 2024*.

Duwal, S. 2025. MKA: Leveraging Cross-Lingual Consensus for Model Abstention. ICLR 2025, arXiv:2503.23687.

Fang, Y.; Li, M.; Wang, W.; Hui, L.; and Feng, F. 2025. Counterfactual Debating with Preset Stances for Hallucination Elimination of LLMs. In *Proc. COLING 2025*, 10554–10568.

Feng, S.; Shi, W.; Wang, Y.; Ding, W.; Ahia, O.; Li, S. S.; Balachandran, V.; Sitaram, S.; and Tsvetkov, Y. 2024a. Teaching LLMs to Abstain across Languages via Multilingual Feedback. In *Proc. EMNLP 2024*, 4125–4150.

Feng, S.; Shi, W.; Wang, Y.; Ding, W.; Balachandran, V.; and Tsvetkov, Y. 2024b. Don't Hallucinate, Abstain: Identifying LLM Knowledge Gaps via Multi-LLM Collaboration. In *Proc. ACL 2024*, 14664–14690.

Funk, M. J.; Westreich, D.; Wiesen, C.; Stürmer, T.; Brookhart, M. A.; and Davidian, M. 2011. Doubly Robust Estimation of Causal Effects. *American Journal of Epidemiology*, 173(7): 761–767.

Greenland, S.; Pearl, J.; and Robins, J. M. 1999. Causal diagrams for epidemiologic research. *Epidemiology*, 10(1): 37–48.

Grootendorst, M. 2022. BERTopic: Neural topic modeling with a class-based TF-IDF procedure. arXiv:2203.05794.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring Massive Multitask Language Understanding. ICLR 2021, arXiv:2009.03300.

Huang, L.; Yu, W.; Ma, W.; Zhong, W.; Feng, Z.; Wang, H.; Chen, Q.; Peng, W.; Feng, X.; Qin, B.; and Liu, T. 2025. A Survey on Hallucination in LLMs: Principles, Taxonomy, Challenges, and Open Questions. *ACM Transactions on Information Systems*, 43(2): 1–55.

Imbens, G. W.; and Rubin, D. B. 2015. *Causal Inference for Statistics, Social, and Biomedical Sciences: An Introduction*. Cambridge University Press.

Jiang, B.; Xie, Y.; Hao, Z.; Wang, X.; Mallick, T.; Su, W. J.; Taylor, C. J.; and Roth, D. 2024. A Peek into Token Bias: LLMs Are Not Yet Genuine Reasoners. In *Proc. EMNLP 2024*, 4722–4756.

Joshi, M.; Choi, E.; Weld, D.; and Zettlemoyer, L. 2017. TriviaQA: A Large Scale Distantly Supervised Challenge Dataset for Reading Comprehension. In *Proc. ACL 2017*.

Just, H. A.; Dabas, M.; Huang, L.; Jin, M.; and Jia, R. 2025. DiPT: Enhancing LLM Reasoning through Diversified Perspective-Taking. In *Findings NAACL 2025*, 6344–6374.

Kirichenko, P.; Ibrahim, M.; Chaudhuri, K.; and Bell, S. 2025. AbstentionBench: Reasoning LLMs Fail on Unanswerable Questions. NeurIPS 2025 D&B.

Kojima, T.; Gu, S. S.; Reid, M.; Matsuo, Y.; and Iwasawa, Y. 2022. Large language models are zero-shot reasoners. In *NeurIPS 2022*.

Kwiatkowski, T.; Palomaki, J.; Redfield, O.; Collins, M.; Parikh, A.; Alberti, C.; Epstein, D.; Polosukhin, I.; Devlin, J.; Lee, K.; et al. 2019. Natural Questions: A Benchmark for Question Answering Research. *Transactions of the Association for Computational Linguistics*, 7: 453–466.

Laskar, M. T. R.; Alqahtani, S.; Bari, M. S.; Rahman, M.; Khan, M. A. M.; Khan, H.; Jahan, I.; Bhuiyan, A.; Tan, C. W.; Parvez, M. R.; Hoque, E.; Joty, S.; and Huang, J. 2024. A Systematic Survey and Critical Review on Evaluating LLMs: Challenges, Limitations, and Recommendations. In *Proc. EMNLP 2024*, 13785–13816.

Lee, D.; Park, A.; Lee, H.; Nam, H.; and Maeng, Y. 2025. Typed-RAG: Type-Aware Decomposition of Non-Factoid Questions for Retrieval-Augmented Generation. XLLM@ACL 2025, arXiv:2503.15879.

Lin, S.; Hilton, J.; and Evans, O. 2022. TruthfulQA: Measuring How Models Mimic Human Falsehoods. In *Proc. ACL 2022*.

Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:1907.11692.

Ma, J. 2025. Causal Inference with Large Language Model: A Survey. In *Findings NAACL 2025*, 5886–5898.

Madhusudhan, N.; Madhusudhan, S. T.; Yadav, V.; and Hashemi, M. 2025. Do LLMs Know When to NOT Answer? Investigating Abstention Abilities of LLMs. In *Proc. COLING 2025*, 9329–9345.

Manakul, P.; Liusie, A.; and Gales, M. 2023. SelfCheckGPT: Zero-Resource Black-Box Hallucination Detection for Generative LLMs. In *Proc. EMNLP 2023*.

Manski, C. F. 2007. *Identification for Prediction and Decision*. Harvard University Press.

McKenna, N.; Li, T.; Cheng, L.; Hosseini, M.; Johnson, M.; and Steedman, M. 2023. Sources of Hallucination by LLMs on Inference Tasks. In *Findings EMNLP 2023*.

Mu, L.; Zhang, W.; Zhang, Y.; and Jin, P. 2024. DDPrompt: Differential Diversity Prompting in LLMs. In *Proc. ACL 2024*, 168–174.

Nathani, D.; Wang, D.; Pan, L.; and Wang, W. 2023. MAF: Multi-Aspect Feedback for Improving Reasoning in LLMs. In *Proc. EMNLP 2023*, 6591–6616.

Pearl, J. 2009. *Causality*. Cambridge University Press.

Pearl, J.; and Bareinboim, E. 2014. External Validity: From Do-Calculus to Transportability Across Populations. *Statistical Science*, 29(4): 579–595.

Rajpurkar, P.; Zhang, J.; Lopyrev, K.; and Liang, P. 2016. SQuAD: 100,000+ Questions for Machine Comprehension of Text. In *Proc. EMNLP 2016*.

Ren, J.; Zhao, Y.; Vu, T.; Liu, P. J.; and Lakshminarayanan, B. 2023. Self-Evaluation Improves Selective Generation in LLMs. In *NeurIPS 2023 Workshop*.

Schlichtkrull, M. S.; Guo, Z.; and Vlachos, A. 2023. AVeriTeC: A Dataset for Real-world Claim Verification with Evidence from the Web. In *NeurIPS Datasets and Benchmarks 2023*.

Slobodkin, A.; Goldman, O.; Caciularu, A.; Dagan, I.; and Ravfogel, S. 2023. The Curious Case of Hallucinatory (Un)answerability: Finding Truths in the Hidden States of Over-Confident LLMs. In *Proc. EMNLP 2023*, 3607–3625.

Spirtes, P.; Glymour, C. N.; Scheines, R.; and Heckerman, D. 2000. *Causation, Prediction, and Search*. MIT Press.

Stasaski, K.; and Hearst, M. 2022. Semantic Diversity in Dialogue with Natural Language Inference. In *Proc. NAACL 2022*, 85–98.

Sun, Y.; Zuo, A.; Gao, W.; and Ma, J. 2025. CausalAbstain: Enhancing Multilingual LLMs with Causal Reasoning for Trustworthy Abstention. In *Findings ACL 2025*, 14060–14076.

VanderWeele, T. J. 2019. Principles of confounder selection. *European Journal of Epidemiology*, 34(3): 211–219.

VanderWeele, T. J.; and Shpitser, I. 2013. On the definition of a confounder. *The Annals of Statistics*, 41(1): 196–220.

Vasisht, K.; Kaur, N.; and Pruthi, D. 2025. Knowledge Graph Guided Evaluation of Abstention Techniques. In *Proc. NAACL 2025*, 6921–6939.

Wang, W.; Wei, F.; Dong, L.; Bao, H.; Yang, N.; and Zhou, M. 2020. MiniLM: Deep Self-Attention Distillation for Task-Agnostic Compression of Pre-Trained Transformers. arXiv:2002.10957.

Wang, X.; Wei, J.; Schuurmans, D.; Le, Q.; Chi, E.; Narang, S.; Chowdhery, A.; and Zhou, D. 2022. Self-Consistency Improves Chain of Thought Reasoning in Language Models. ICLR 2023, arXiv:2203.11171.

Wen, B.; Brahman, F.; Su, Z.; Feng, S.; Tsvetkov, Y.; Wang, L. L.; and Howe, B. 2025. MARVEL: Modular Abstention for Reliable and Versatile Expert LLMs. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Wen, B.; Howe, B.; and Wang, L. L. 2024. Characterizing LLM Abstention Behavior in Science QA with Context Perturbations. In *Findings EMNLP 2024*, 3437–3450.

Wen, B.; Yao, J.; Feng, S.; Xu, C.; Tsvetkov, Y.; Howe, B.; and Wang, L. L. 2024. Know Your Limits: A Survey of Abstention in LLMs. arXiv:2407.18418.

Wu, J.; Yu, T.; Chen, X.; Wang, H.; Rossi, R.; Kim, S.; Rao, A.; and McAuley, J. 2024. DeCoT: Debiasing Chain-of-Thought for Knowledge-Intensive Tasks in LLMs via Causal Intervention. In *Proc. ACL 2024*, 14073–14087.

Xu, R.; Qi, Z.; Guo, Z.; Wang, C.; Wang, H.; Zhang, Y.; and Xu, W. 2024a. Knowledge Conflicts for LLMs: A Survey. In *Proc. EMNLP 2024*, 8541–8565.

Xu, Z.; Cheng, D.; Li, J.; Liu, J.; Liu, L.; and Yu, K. 2024b. Causal Inference with Conditional Front-Door Adjustment and Identifiable Variational Autoencoder. In *ICLR 2024*.

Yadkori, Y. A.; Kuzborskij, I.; György, A.; and Szepesvári, C. 2024. To Believe or Not to Believe Your LLM.

Yang, Z.; Qi, P.; Zhang, S.; Bengio, Y.; Cohen, W.; Salakhutdinov, R.; and Manning, C. D. 2018. HotpotQA: A Dataset for Diverse, Explainable Multi-hop Question Answering. In *Proc. EMNLP 2018*.

Zhang, C.; Zhang, L.; Wu, J.; He, Y.; and Zhou, D. 2025a. Causal Prompting: Debiasing LLM Prompting Based on Front-Door Adjustment. *Proc. AAAI*, 39(24): 25842–25850.

Zhang, C.; Zhang, L.; and Zhou, D. 2024. Causal Walk: Debiasing Multi-Hop Fact Verification with Front-Door Adjustment. *Proc. AAAI*, 38(17): 19533–19541.

Zhang, H.; Diao, S.; Lin, Y.; Fung, Y.; Lian, Q.; Wang, X.; Chen, Y.; Ji, H.; and Zhang, T. 2024a. R-Tuning: Instructing LLMs to Say 'I Don't Know'. In *Proc. NAACL 2024*, 7113–7139.

Zhang, Y.; Chen, Q.; Zhou, J.; Wang, P.; Si, J.; Wang, J.; Lu, W.; and Qin, L. 2024b. Wrong-of-Thought: An Integrated Reasoning Framework with Multi-Perspective Verification and Wrong Information. In *Findings EMNLP 2024*, 6644–6653.

Zhang, Y.; Li, S.; Qian, C.; Liu, J.; Yu, P.; Han, C.; Fung, Y. R.; McKeown, K.; Zhai, C.; Li, M.; and Ji, H. 2025b. The Law of Knowledge Overshadowing: Towards Understanding, Predicting and Preventing LLM Hallucination. In *Findings ACL 2025*, 23340–23358.

Zhao, Y.; Yan, L.; Sun, W.; Xing, G.; Meng, C.; Wang, S.; Cheng, Z.; Ren, Z.; and Yin, D. 2024a. Knowing What LLMs DO NOT Know: A Simple Yet Effective Self-Detection Method. In *Proc. NAACL 2024*, 7051–7063.

Zhao, Y.; Zheng, Y.; Jiang, Z.; Jiang, Z.; Wu, X.; and Gao, J. 2024b. Harnessing LLMs for Knowledge Graph Question Answering via Adaptive Multi-Aspect Retrieval-Augmentation. In *Proc. AAAI 2024*, volume 38, 17301–17309.