Twin Worlds: Equivariance-Based Abstention for Evidence-Grounded Reasoning

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, Australia [2]Jilin University, China
[3]Nanyang Technological University, Singapore
s3964786@student.rmit.edu.au, xiuzhen.zhang@rmit.edu.au
*Corresponding author

Abstract

Knowledge-intensive reasoning requires Large Language Models (LLMs) to ground answers in provided evidence. When evidence is insufficient, it is desirable that models abstain rather than confidently generating unsupported answers. Existing abstention methods rely on uncertainty estimation or evidence sufficiency checks, but neither tests whether the reasoning process for generation, driven by the interaction of provided evidence and the model's internal memory parameters, is actually grounded in the evidence. A key contributing factor is that entity mentions in context activate memorised associations, causing models to generate plausible responses ungrounded in evidence. We propose **Twin Worlds (TW)**, a framework for improving reliability in knowledge-intensive reasoning through *equivariance*-based abstention: unlike invariance, which requires outputs to remain unchanged, equivariance requires outputs to *transform correspondingly* under entity substitutions. A model grounded in the evidence should produce answers that shift consistently when entities are substituted while their relations are preserved. TW constructs multiple worlds via typed substitutions of the original input that preserve relational structure while reducing parametric priors, and uses equivariance violations as an abstention signal. Across four benchmarks and three model backbones, TW identifies when answers are not reliably grounded in the provided evidence and outperforms uncertainty- and sufficiency-based baselines.[1]

1 Introduction

Large Language Models (LLMs) are increasingly used for knowledge-intensive reasoning with external evidence, where generation of answers requires combining a query with documents, retrieved passages, or tool outputs (Lewis et al., 2020; Yao et al., 2023; Qin et al., 2024). A central failure in these settings is that models often continue to answer even when the provided evidence is insufficient, misleading, or internally conflicting (Park et al., 2025; Ming et al., 2025). Consider the question *"Who is the CEO of P&G?"* The accompanying evidence may fail to support it: passages may discuss P&G's products or partnerships without identifying the CEO. Yet an LLM may still produce a plausible name consistent with reliance on memorised associations tied to the company mention (Fig. 1, top), rather than reasoning from the evidence (Mallen et al., 2023; Li and Goyal, 2025). The desired behaviour is to recognise when the available evidence does not justify an answer and to abstain.

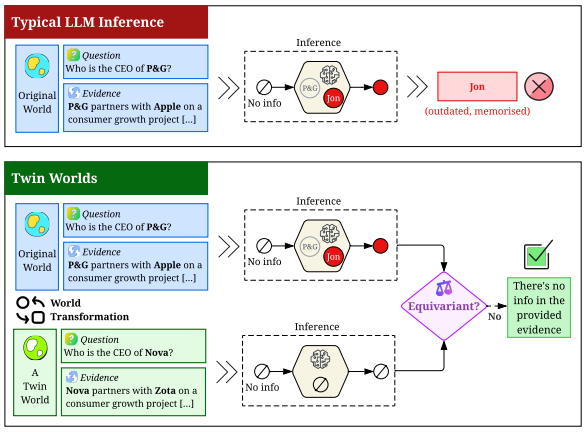


Figure 1: *Top:* LLMs often answer confidently based on outdated parametric associations even when the provided evidence does not support a conclusion. *Bottom:* Twin Worlds applies structure-preserving transformations to test whether answers are grounded in the evidence, enabling principled abstention decisions.

Multi-evidence settings further complicate this problem, where contexts often contain distractor passages, partial support, or plausible cues for unsupported answers. With unfaithful evidence, irrelevant information can amplify parametric activation, causing confident answers from unrelated

[1]We release the source code at https://github.com/xiuzhenzhang/Twin-Worlds

Proceedings of the 2026 Conference on Empirical Methods in Natural Language Processing, pages 52170–52188
October 24-29, 2026 ©2026 Association for Computational Linguistics

knowledge rather than abstention (Ming et al., 2025). Conversely, faithful evidence can still be overridden by conflicting parametric knowledge (Wu et al., 2024), producing spurious uncertainty that triggers unnecessary abstention. Existing approaches ask either *Is the model uncertain?* (Chen et al., 2024; Wen et al., 2024; Kim et al., 2025b) or *Is the evidence sufficient?* (Sun et al., 2025a; Joren et al., 2025; Ren et al., 2026). However, a model may remain confident and self-consistent while following memorised associations that suggest a seemingly plausible but unjustified answer. We instead ask: *Is the reasoning process driving generation actually grounded in the provided evidence rather than in memorised associations?*

We propose **Twin Worlds (TW)**, a framework for improving reliability in knowledge-intensive reasoning through *equivariance*-based abstention: a model whose answers are grounded in the evidence should produce outputs that transform consistently under structure-preserving substitutions of entities. Many knowledge-intensive tasks, including open-domain question answering, fact verification, and multi-hop reasoning, are strongly entity-centric, with questions and evidence structured around named entities (Yang et al., 2018; Thorne et al., 2018; Mallen et al., 2023; Li et al., 2026). TW is therefore most naturally suited to settings with identifiable entity structure. LLMs encode substantial knowledge about these entities in their internal representations (Hu et al., 2025b; Sakata et al., 2025; Morand et al., 2025; Yang et al., 2026), and entity mentions in context can activate this stored knowledge (Zhou et al., 2024; Farahani and Johansson, 2024; Ferrando et al., 2025), potentially driving generation from parametric memory. TW constructs multiple twin worlds, each defined by a typed bijective transformation of the original input that preserves relational structure while reducing parametric priors, and uses equivariance violations across these worlds as an abstention signal.

When the evidence supports an answer, structure-preserving substitutions should induce corresponding answer transformations, yielding high equivariance. When evidence is incomplete, misleading, or inconsistent, equivariance may break, indicating that the answer is not reliably determined by the relational structure of the provided evidence and signalling abstention. Figure 1 (bottom) illustrates this: replacing P&G with a synthetic organisation name reduces parametric associations, and when back-mapped answers fail to agree across twin worlds, this signals that generation is not reliably grounded in the evidence, and TW enables the model to abstain.

Our contributions are as follows:

- We reframe abstention in knowledge-intensive reasoning as a test of structural grounding, introducing the *equivariance* criterion: a model whose answers are evidence-grounded should produce outputs that transform consistently under structure-preserving entity substitutions.
- We propose Twin Worlds, a simple framework that operationalises this criterion by testing whether answers track structural substitutions, not merely whether answers change, using equivariance violations as a training-free inference-time abstention signal.
- Across four benchmarks and three model backbones, we demonstrate that TW outperforms baselines in abstention reliability and provides diagnostics of answer grounding without requiring evidence-faithfulness annotations.

2 Related Work

2.1 Abstention in Large Language Models

Abstention, where a model explicitly refrains from answering, provides a principled way to handle queries that exceed a model's knowledge or the limits of available evidence (Kapoor et al., 2024). Despite its importance, reliable abstention remains challenging: LLMs are primarily trained to maximise benchmark performance rather than recognise their own limitations, and even large models often fail to abstain appropriately (Kalai et al., 2026; Kirichenko et al., 2025). Existing methods approach abstention through fine-tuning (Zhang et al., 2024; Stengel-Eskin et al., 2024; Cao, 2024; Zhu et al., 2025b; Huang et al., 2025; Wei et al., 2026), uncertainty estimation (Manakul et al., 2023; Yadkori et al., 2024; Kim et al., 2025a; Ji et al., 2025), or multi-model agreement (Feng et al., 2024a,b; Wen et al., 2025). More recently, causal approaches analyse the factors influencing abstention decisions (Sun et al., 2025b; Zhao et al., 2025; Nguyen et al., 2026). Despite their differences, these methods largely study abstention in isolation, without accounting for how external evidence influences the generation process. In contrast, TW grounds abstention decisions in the relational structure of the provided evidence, testing whether generation behaviour is consistent with the context itself.

2.2 Evidence Reliability in Knowledge-Intensive Reasoning

When models reason over evidence, abstention becomes intertwined with evidence reliability. Models must determine not only whether they know the answer, but whether the available evidence justifies producing one. Risk-based methods estimate when to abstain through context manipulations (Chen et al., 2024; Wen et al., 2024), uncertainty (Ding et al., 2025), axiomatic constraints (Soudani et al., 2025), contrastive decoding (Kim et al., 2025b), or agreement (Goswami and Kurra, 2025; Hu et al., 2025a). Sufficiency-based approaches evaluate whether retrieved evidence contains the information required to answer (Sun et al., 2025a; Song et al., 2024; Joren et al., 2025; Bergeron et al., 2026). However, both strategies test uncertainty or sufficiency without evaluating whether generation is faithful to the provided evidence. A model may remain confident and self-consistent while its answers are driven by parametric activation rather than the provided context. TW addresses this gap directly by testing whether answers are structurally determined by the evidence.

2.3 Equivariance in Neural Models

Prior work on equivariance in neural models studies how predictions should transform under structured input transformations, establishing theoretical foundations for group-equivariant architectures (Cohen and Welling, 2016; Kondor and Trivedi, 2018) and practical methods for complex transformation groups (Finzi et al., 2021; Fuchs et al., 2020). In NLP, symmetry failures have been linked to spurious heuristics (McCoy et al., 2019; Kassner and Schütze, 2020), and behavioural testing frameworks examine model invariance under controlled perturbations (Ribeiro et al., 2020). More recent work explores equivariant representations and permutation-based training for language models (Guo et al., 2024; Li et al., 2025; Zhu et al., 2025a). SynthWorlds (Gu et al., 2026) constructs parallel corpora to disentangle reasoning from parametric knowledge, but targets evaluation rather than abstention. These approaches either build equivariance into the model architecture or training procedure, or treat symmetry violations as evidence of poor generalisation. However, whether equivariance can be leveraged to improve LLM reliability at inference time without modifying the model remains unexplored. TW addresses this gap by using equivariance violations as a principled abstention signal in knowledge-intensive reasoning.

3 Methodology

We present **Twin Worlds (TW)**, a framework for improving the reliability of knowledge-intensive reasoning through equivariance-based abstention.

Let q denote a question, $C = \{c_1, \ldots, c_n\}$ a set of evidence passages, and f a language model producing the answer distribution $p_f(y \mid q, C)$. TW evaluates whether answer behaviour is equivariant under structure-preserving transformations of entities in (q, C) and uses violations of this equivariance as a signal for whether the model should answer or abstain. Fig. 2 illustrates this process.

3.1 Structure-Preserving Transformations

Let $\mathcal{E}(q, C)$ denote the set of entities appearing in the question and evidence, together with their semantic types such as PERSON, ORG, or DATE. A structure-preserving transformation is a mapping defined as follows:

$$\pi : \mathcal{E}(q, C) \to \mathcal{E}'(q, C),$$

such that: (i) type is preserved, so an entity of type t is mapped to another entity of type t; (ii) π is *bijective*, meaning it is both injective (no two source entities map to the same target) and surjective (every target entity is the image of some source entity), ensuring that the mapping is reversible via π^{-1}; and (iii) only entity identity changes, while the relational structure remains unchanged.

Applying π to (q, C) yields a transformed pair (q^π, C^π) in which all entity mentions are replaced consistently. We realise these transformations in two steps. First, we *delexicalise* the question and evidence by replacing entity mentions with typed placeholders, producing a canonical representation that preserves relational structure while removing lexical identity. Second, we *relexicalise* under a sampled transformation, replacing each placeholder with a type-matched synthetic entity through a one-to-one reversible mapping. The resulting transformed inputs form *twin worlds*: they preserve the relational structure of the evidence while reducing lexical cues that may activate parametric priors.

3.2 Equivariance as a Grounding Criterion

We formalise evidence-grounded reasoning through **equivariance**. A model is equivariant under transformation π if applying π to the input induces a

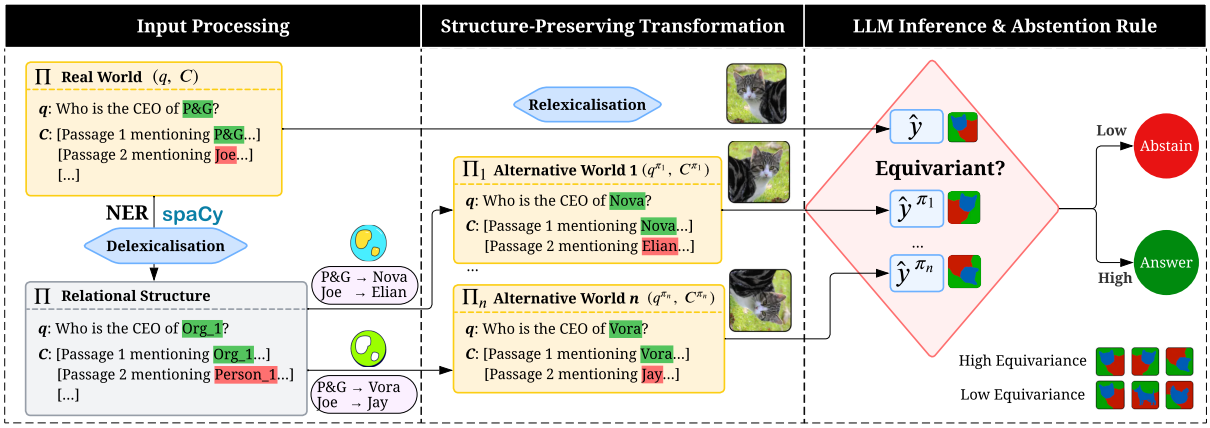


Figure 2: TW framework overview. *Left:* Delexicalisation replaces entity mentions with typed placeholders, extracting relational structure. *Middle:* Typed bijective substitutions relexicalise placeholders under k sampled mappings, producing k twin worlds (q^{π_i}, C^{π_i}). *Right:* The equivariance score aggregates whether back-mapped answers $\pi^{-1}(\hat{y}^{\pi_i})$ agree with $\hat{y}$; high equivariance leads to answering, while low equivariance leads to abstention.

corresponding transformation in the answer: the answer in the transformed input, mapped back through π^{-1}, should agree with the answer in the original input. Intuitively, returning to the example in Figure 1, if the evidence supports the answer, replacing an entity in the evidence with a synthetic entity such as Elian Voss should produce a correspondingly transformed answer; if the back-mapped answer no longer agrees with the original, this indicates that the answer is not reliably determined by the relational structure of the evidence.

Let $\hat{y} = \arg\max_y p_f(y \mid q, C)$ and $\hat{y}^{\pi} = \arg\max_y p_f(y \mid q^{\pi}, C^{\pi})$. When the answer contains an entity participating in π, π^{-1} maps it back through the substitution used to transform the input. For answers that do not participate in the substitution mapping, such as scalar values or relational predicates, π^{-1} acts as the identity. We define a per-transformation equivariance score as follows:

$$S_{\pi}(q, C) = \mathrm{Agree}\big(\pi^{-1}(\hat{y}^{\pi}),\ \hat{y}\big)\ ,$$

where $\mathrm{Agree}(\cdot,\cdot)$ is a task-level agreement function. In the simplest case, this reduces to exact match:

$$S_{\pi}(q, C) = \mathbf{1}\big[\pi^{-1}(\hat{y}^{\pi}) = \hat{y}\big]\ .$$

The back-mapping π^{-1} distinguishes equivariance from methods based on an *invariance* criterion. Invariance-based methods produce outputs that remain unchanged under input transformations: confidence- and consistency-based methods test whether $\hat{y}^{\pi} = \hat{y}$ (Wang et al., 2023); perturbation methods test whether $\hat{y}^{\pi} \approx \hat{y}$ (Gao et al., 2024); counterfactual methods test whether $\hat{y}$ changes when context is removed or replaced (Chen et al., 2024). Equivariance instead requires outputs to transform correspondingly: $\pi^{-1}(\hat{y}^{\pi}) = \hat{y}$ (Geva et al., 2023). A model relying on parametric memory may pass invariance-based tests yet fail equivariance because its answers do not track the substitution mapping.

TW constructs a set of k twin worlds $W = \{\pi_1, \ldots, \pi_k\}$ by sampling structure-preserving transformations over $\mathcal{E}(q, C)$. Each twin world π_i produces a transformed input (q^{π_i}, C^{π_i}) by applying the substitution consistently across all entity mentions in the question and evidence passages. Twin worlds reduce parametric priors by using procedurally generated synthetic values as replacement entities, such as Nova for a location or Elian Voss for a person. This design encourages equivariant behaviour across twin worlds to reflect relational reasoning from the evidence rather than incidental parametric associations.

3.3 Equivariance-Based Abstention

TW computes an equivariance score by averaging over the set of twin worlds:

$$s(q, C) = \frac{1}{|W|} \sum_{\pi \in W} S_{\pi}(q, C).$$

This score measures whether the answer is reliably determined by the evidence. TW applies the following abstention rule:

$$\hat{y}_{\text{out}} = \begin{cases} \hat{y}, & \text{if } s(q, C) \geq \tau \\ \texttt{abstain}, & \text{otherwise,} \end{cases}$$

where τ is a threshold tuned on a held-out validation set. When the equivariance score falls below the threshold, the answer is not reliably determined by the relational structure of the provided evidence, and TW abstains. TW does not attempt to distinguish whether such a failure arises from the model, irrelevant or erroneous context, or another source; rather, it signals that the answer cannot be reliably justified from the given evidence.

4 Experiments & Results

4.1 Experimental Settings

Datasets We evaluate TW on four benchmarks. HotpotQA (Yang et al., 2018) tests multi-hop reasoning with distractor passages. MIRAGE (Park et al., 2025) aggregates open-domain QA datasets, including PopQA (Mallen et al., 2023) and TriviaQA (Joshi et al., 2017), enabling analysis of parametric versus contextual knowledge use. FaithEval (Ming et al., 2025) introduces unanswerable and misleading contexts to assess robustness against unsupported answering. FEVER (Thorne et al., 2018) evaluates fact verification, where models should abstain on *Not Enough Info* instances when evidence is insufficient. Rather than prompting the model to classify claims into predefined labels, we treat FEVER as a generative task: the model is prompted to answer each claim as a question given the retrieved evidence, and TW evaluates equivariance via π^{-1} exactly as in the QA setting. Dataset sampling and construction details are provided in Appendix A.1.

Baselines We compare TW against seven baselines. Zero-shot (Kojima et al., 2022) measures inherent abstention behaviour under direct prompting. Risk-based methods estimate when to abstain from observable signals: AbstentionBench (Kirichenko et al., 2025) uses abstention-oriented prompting, Self-Consistency (Wang et al., 2023) detects when agreement across sampled reasoning traces is low, and RC-RAG (Chen et al., 2024) estimates hallucination risk via counterfactual context manipulations. Sufficiency-based methods assess whether the evidence supports an answer: Context Perturbation (Wen et al., 2024) measures answer stability under evidence perturbations, Sufficient Context (Joren et al., 2025) predicts evidence sufficiency via a scoring model, and Contrastive Decoding Abstention (Kim et al., 2025b) modifies decoding to abstain under weak or unreliable evidence. For methods with publicly available implementations, we use the authors' official repositories and default configurations.

Metrics Effective abstention must balance two failure modes: answering when evidence is insufficient and refusing when it is not. Following Kim et al. (2025b); Madhusudhan et al. (2025), we use a confusion matrix crossing question types (answerable vs. unanswerable) with model behaviours (answered correctly, answered incorrectly, or abstained), and report F1, Accuracy (Acc), Reliability Score (RS), and Abstention Rate (AR). RS captures the trade-off between coverage and correctness; AR flags methods that resolve this trade-off by over-refusing rather than improving grounding. Full metric definitions are provided in Appendix A.8.

Twin World Construction We first *delexicalise* the question and evidence using spaCy (Montani et al., 2023) to detect entity spans and assign coarse semantic types (Person, Org, Location, Date), replacing each mention with a typed indexed placeholder such as `Person_1` or `Org_1` and merging repeated mentions through string and alias matching; robustness to NER failures is analysed in App. B.2. We then *relexicalise* by instantiating structure-preserving transformations over these placeholders, sampling without replacement from a task-agnostic inventory of synthetic values with no attested referents in standard English; the same inventory is reused across all four benchmarks. Generation and validation details are in App. A.3. We sample $k = 3$ twin worlds per query. The default threshold $\tau = 0.60$ is tuned on a 500-instance held-out validation set to maximise RS performance.

Model Backbones We evaluate TW on GPT-5.1, LLaMA-4-Scout 17B×16E, and Mistral-Small-24B, representing closed-source, MoE, and compact architectures respectively. Detailed experimental setups are provided in Appendix A.

4.2 Main Results

Table 1 reports the performance of TW against all baselines. TW achieves the strongest overall performance on HotpotQA and MIRAGE across all three backbones, attaining the highest F1 and Accuracy in all six dataset–backbone combinations and the highest RS in four of six cases, while maintaining near-zero abstention rates (AR ≤ 0.029). In contrast, RC-RAG reaches AR values between 0.27 and 0.54, while Context Perturbation reaches up to 0.25, and both underperform TW on F1 and RS. These

Method	HotpotQA				MIRAGE				FaithEval				FEVER			
	AR	F1	Acc	RS	AR	F1	Acc	RS	AR	F1	Acc	RS	AR	F1	Acc	RS
	GPT-5.1															
Zero-shot	.0010	.9325	.9320	**.9989**	.0060	.9157	.9130	.9935	.0510	-	.0510	.5032	.3210	.6974	.6260	.6620
AbstentionBench	.0200	.9232	.9140	.9787	.0580	.8847	.8590	.9372	.3870	-	.3870	.5255	**.8430**	.2625	.4000	.3618
Self-Consistency	.0110	.9346	.9280	.9876	.0180	.9276	.9250	.9661	.1640	-	.1640	.5728	.4280	.6123	.5870	.5904
RC-RAG	**.2720**	.8090	.6990	.7201	.2710	.7912	.6840	.7168	.4860	-	.4860	.5004	.4120	.4928	.5400	.5682
Context Perturbation	.0530	.9081	.8840	.9437	.0790	.8631	.8290	.9137	.5090	-	.5090	.5002	.4520	.5783	.5030	.5277
Sufficient Context	.1430	.8422	.7820	.8463	**.4960**	.5944	.4470	.4757	**.7070**	-	**.7070**	**.5857**	.4950	.6360	.5210	.5129
Contrastive Decoding	.0170	.9259	.9180	.9819	.0550	.8925	.8680	.9408	.3760	-	.3760	.5308	.3940	.5820	**.6690**	.6308
TW (Ours)	.0030	**.9396**	**.9390**	.9981	.0050	**.9318**	**.9300**	**.9959**	.7010	-	.7010	.5812	.3490	**.7215**	.6550	**.6824**
	LLaMA-4															
Zero-shot	.0160	.8831	.8760	.9823	.0100	.9005	.8960	.9811	.2470	-	.2470	.5280	.1650	.6489	.5740	.6219
AbstentionBench	.0270	.8819	.8700	.9702	.0120	.8964	.8910	.9868	.3100	-	.3100	.5722	**.5020**	.5808	.5490	.5236
Self-Consistency	.0210	.8849	.8790	.9781	.0310	.9128	.8890	.9624	.4520	-	.4520	.5369	.3710	.6814	.6360	.6128
RC-RAG	**.3900**	.7106	.5720	.5952	**.4150**	.7066	.5600	.5746	.6260	-	.6260	.5318	.4190	.5116	.5200	.5554
Context Perturbation	.1230	.8556	.8030	.8679	.0450	.8880	.8680	.9511	.5610	-	.5610	.5074	.4150	.5691	.5660	.5771
Sufficient Context	.1000	.8663	.8230	.8923	.1440	.8384	.7780	.8448	.5830	-	.5830	.5138	.4190	.6977	.6540	.6116
Contrastive Decoding	.0400	.8816	.8640	.9562	.0110	.8969	.8920	.9879	.3190	-	.3190	.5655	.3750	.6507	.6210	.6235
TW (Ours)	.0130	**.8898**	**.8840**	**.9869**	.0090	**.9164**	**.8990**	**.9916**	**.6410**	-	**.6410**	**.5791**	.3860	**.7098**	**.6680**	**.6327**
	Mistral-Small															
Zero-shot	.0150	.8725	.8660	.9832	.0270	.8880	.8760	**.9704**	.2880	-	.2880	.5899	.6510	.4788	.4950	.4440
AbstentionBench	.2730	.7782	.6720	.7120	.1890	.8172	.7400	.7976	.6490	-	.6490	.5444	**.7050**	.2581	.3520	.3352
Self-Consistency	.0970	.8643	.8110	.8591	.1160	.8782	.8210	.8469	.4720	-	.4720	.5518	.5330	.6015	.5890	.5574
RC-RAG	**.5370**	.5837	.4270	.4437	**.4930**	.6397	.4820	.4947	.5580	-	.5580	.5067	.6210	.5761	.5700	.4976
Context Perturbation	.2460	.7856	.6890	.7380	.2370	.7760	.6840	.7443	.6230	-	.6230	.5303	.5410	.5031	.5330	.4990
Sufficient Context	.1050	.8116	.7690	.8818	.2240	.7815	.6940	.7576	.5710	-	.5710	.5101	.4170	.4788	.5500	**.5692**
Contrastive Decoding	.3180	.7598	.6390	.6683	.2970	.7704	.6560	.6890	.6310	-	.6310	.5343	.4590	.5054	.5750	.5566
TW (Ours)	.0170	**.8798**	**.8730**	**.9835**	.0290	**.8942**	**.8820**	.9698	**.6560**	-	**.6560**	**.5987**	.4820	**.6129**	**.5860**	.5539

Table 1: Results of TW and baselines across benchmark datasets. Each section compares methods with the same backbone. **Bold** indicates the highest score and underline the second highest. A hyphen (-) denotes zero F1 on FaithEval where all questions are unanswerable; performance there is measured by RS and Acc. Higher AR indicates more frequent abstention and may reflect over-conservatism rather than genuine grounding.

Q: What was found in Los Angeles?
E: [...] Trevor Valle discusses the teeth and lower jaw of a mammoth fossil. Now, at least 10,000 years later, visitors in Los Angeles can see the remains of a mammoth [...]
Twin World 1: Los Angeles $\mapsto$ Nova, Trevor Valle $\mapsto$ Elian Voss **Output:** no information $S_{\pi_1} = 0$
Twin World 2: Los Angeles $\mapsto$ Valdecor, Trevor Valle $\mapsto$ Orin Vael **Output:** no information $S_{\pi_2} = 0$
Twin World 3: Los Angeles $\mapsto$ Mirevan, Trevor Valle $\mapsto$ Casen Drel **Output:** no answer $S_{\pi_3} = 0$
Final Answer: $s(q, C) = 0.00 < \tau = 0.60 \Rightarrow$ **NIL (Abstain)**

Table 2: TW on a representative unanswerable case. Across three twin worlds, all twin answers fail to agree with the original answer "mammoth fossil", yielding $s = 0.00$ and triggering abstention.

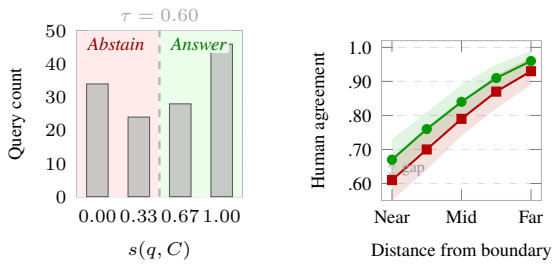
results indicate that equivariance-based abstention improves answer quality without over-refusing on answerable questions.

On FaithEval, where all questions are unanswerable, TW achieves the best Accuracy and AR on LLaMA-4 and Mistral-Small and the second-best results on GPT-5.1. Sufficient Context abstains more aggressively on GPT-5.1 but also substantially over-abstains on answerable benchmarks, reaching AR 0.496 on MIRAGE. This illustrates the difficulty of balancing abstention on unanswerable instances against coverage on answerable ones. TW instead bases its abstention signal on whether answers track the relational structure of the evidence across structure-preserving transformations. On FEVER, TW achieves the best F1 across all three backbones, suggesting that equivariance-based abstention generalises beyond question answering to fact verification.

Table 2 illustrates a representative case. Under structure-preserving substitutions, the answer fails to track the substitution mapping, and TW correctly abstains, indicating that the original answer is not reliably determined by the relational structure of the provided evidence.

4.3 Human Evaluation

We conduct human evaluation to validate two aspects of TW: the reliability of structure-preserving transformations in preserving relational structure and the alignment of abstention decisions with human judgement.



(a) Score distribution. (b) Human agreement.

Figure 3: (a) Distribution of equivariance scores on GPT-5.1. With $k = 3$, scores take the discrete values $\{0, 0.33, 0.67, 1.00\}$; most queries lie away from the decision boundary at $\tau = 0.60$. (b) Human–TW agreement increases monotonically with distance from the decision boundary, indicating stronger alignment on cases with more decisive equivariance scores.

Setup We recruit native English speakers via Amazon Mechanical Turk with ≥98% approval rate and ≥500 completed HITs. Each instance is evaluated by three independent annotators with attention checks. We sample 200 instances from MIRAGE and HotpotQA, balanced across entity types. Detailed annotation instructions and scoring rubrics are provided in Appendix A.6.

Structure Preservation Annotators rate how well transformed inputs preserve relational structure on a 5-point Likert scale. Results indicate a mean rating of 4.61 (SD = 0.48), with 99.1% of ratings ≥ 4, suggesting that structure-preserving transformations reliably preserve relational structure while altering entity identity.

Abstention Alignment Annotators judge whether the evidence contains sufficient information to answer the question, without access to TW outputs. TW agrees with human judgement on 86% of answerable and 81% of unanswerable instances, yielding 83.5% overall agreement (Fleiss' $\kappa = 0.78$). Fig. 3 shows that alignment increases monotonically with distance from τ, exceeding 90% for queries far from the boundary. Residual disagreement concentrates near the boundary, where evidential support is more ambiguous, suggesting that equivariance scores align more strongly with human judgement away from the decision boundary.

4.4 Probing Structural Grounding

We validate TW from two complementary angles: behavioural and representational, using 500 instances per condition from MIRAGE and HotpotQA with passage faithfulness annotations.

Behavioural Analysis. We evaluate the World-Tracking Rate (WTR), a metric measuring how consistently back-mapped answers agree with the original answer across twin worlds, under two conditions: *faithful* passages that uniquely support the answer, and *unfaithful* passages that mention relevant entities without supporting it. We compute:

$$\mathrm{WTR}(q, C) = \frac{1}{|W|} \sum_{\pi \in W} \mathbf{1}\left[\pi^{-1}(\hat{y}^{\pi}) = \hat{y}\right].$$

We compare typed bijective transformations with four alternatives: *untyped substitution*, which permits cross-type replacements; *counterfactual substitution*, which uses real-world alternatives of the same type; *surface paraphrase*, which rewrites entity mentions without substituting their identity; and *random replacement*, which replaces entity spans with random tokens. Table 3 shows that typed bijective transformations produce the clearest separation between faithful and unfaithful evidence across all backbones. Counterfactual substitution yields substantially higher WTR under unfaithful evidence, consistent with real-world replacements retaining parametric associations; surface paraphrase produces similar WTR regardless of faithfulness; and random replacement degrades WTR under both conditions.

Activation Analysis. We examine whether synthetic substitutions reduce entity-associated parametric activation rather than merely altering surface form. For each instance, we prompt Mistral-Small-24B with the question and no evidence and define *entity-associated tokens* as the top-10 tokens in the resulting closed-book output distribution for the original entity input. We then measure the probability of these same tokens across transformer layers under original and synthetic entity inputs using the logit lens. Figure 4 shows that synthetic substitutions consistently reduce entity-associated activation across layers, while original entities with unfaithful evidence exhibit higher activation, particularly in later layers. Together with the behavioural results, this provides evidence that synthetic substitutions attenuate lexical triggers associated with parametric memory and supports the interpretation of equivariance violations as a signal that an answer is not reliably determined by the relational structure of the provided evidence.

Strategy	GPT-5.1		LLaMA-4		Mistral-Small	
	Faith.	Unfaith.	Faith.	Unfaith.	Faith.	Unfaith.
Typed Bijection	**.804**	.215	**.780**	**.230**	**.761**	**.243**
Untyped Sub.	.721	.401	.701	.385	.685	.390
Counterfactual	.687	.475	.670	.460	.655	.452
Paraphrase	.811	.800	.799	.790	.791	.780
Random Repl.	.340	**.210**	.325	.305	.310	.301

Table 3: WTR across evidence conditions, substitution strategies, and model backbones. Higher WTR is desirable under faithful evidence and lower WTR under unfaithful evidence; the desired behaviour is therefore a large faithful–unfaithful separation.

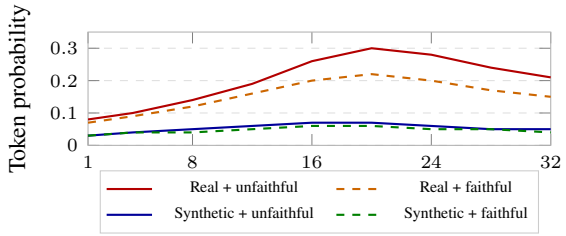


Figure 4: Probability of entity-associated tokens across Mistral-Small-24B layers. Synthetic substitutions consistently reduce activation relative to the corresponding real-entity inputs, with the largest activation observed for real entities under unfaithful evidence.

4.5 Ablation Studies

We ablate five design choices on a balanced sample of 500 instances (250 answerable, 250 unanswerable) drawn from all four datasets. Results appear in Table 4.

Number of Twin Worlds We sweep $k \in \{1, 2, 3, 4, 6\}$ and find that $k = 1$ reduces RS by 0.044 across all three backbones, as a single substitution yields a less stable equivariance estimate. Performance improves up to $k = 3$. Increasing to $k = 6$ yields consistent but negligible gains on most metrics compared with the doubled inference cost, while GPT-5.1 RS decreases slightly. We therefore use $k = 3$ as the default.

Substitution Properties Removing type constraints allows cross-type substitutions, introducing type-mismatched inputs that confound the abstention signal and reducing RS by 0.048–0.051 across backbones. Removing bijectivity breaks the one-to-one mapping between source and target entities, collapsing structural distinctions between entity roles and weakening diagnostic separation. Replacing synthetic values with high-frequency Wikipedia entities produces the largest RS drop (0.103–0.115) across backbones, suggesting that non-referential synthetic entity construction is an important design choice.

Configuration	GPT-5.1			LLaMA-4			Mistral-Small		
	F1	Acc	RS	F1	Acc	RS	F1	Acc	RS
TW (Full, $k = 3$)	.921	.903	**.971**	.891	.873	.956	.867	.849	.934
$k = 1$	.884	.859	.927	.853	.831	.912	.829	.808	.890
$k = 6$	**.922**	**.904**	.967	**.892**	**.874**	**.957**	**.868**	**.850**	**.935**
No type constraint	.874	.843	.921	.843	.819	.908	.819	.791	.883
No bijectivity	.881	.851	.928	.851	.826	.913	.828	.801	.891
Wiki entities	.821	.793	.868	.791	.764	.841	.768	.742	.831
No back-mapping	.849	.819	.904	.819	.794	.881	.796	.771	.858

Table 4: Ablation results across all three backbones, averaged across all four datasets. **Bold** indicates the highest score in each column.

Abstention Mechanism Removing back-mapping reduces RS by 0.067–0.076 across backbones, demonstrating the importance of testing equivariance via π^{-1}. Without back-mapping, the framework reduces to a perturbation stability check without an explicit structural correspondence between transformed and original answers.

4.6 Analysis of Hard Abstention Cases

A key claim of TW is that equivariance captures failure cases that may not be identified by uncertainty or sufficiency signals. We test this by constructing a hard subset of 200 unanswerable instances drawn from all four datasets, selected such that Self-Consistency (Wang et al., 2023) agreement is high (≥ 0.8) and Sufficient Context (Joren et al., 2025) score is high (≥ 0.5). These are cases in which both signals favour answering despite the absence of sufficient evidence. Table 5 shows that Self-Consistency achieves AR of 0.09 and Sufficient Context achieves AR of 0.11, whereas TW achieves AR of 0.71. This indicates that equivariance provides a complementary abstention signal in cases where confidence and apparent evidence sufficiency are both high.

Table 6 illustrates a representative case. The passage states that Gareth Bale spent much of his career as a winger but does not establish his position specifically in his final season. Self-Consistency produces identical answers across five traces and Sufficient Context assigns a score of 0.51, leading both methods to answer. Across three twin worlds, TW produces inconsistent back-mapped answers ($s = 0.33 < \tau$) and abstains, indicating that the original answer is not reliably determined by the provided evidence.

Method	AR
AbstentionBench	0.18
Self-Consistency	0.09
RC-RAG	0.31
Context Perturbation	0.24
Sufficient Context	0.11
Contrastive Decoding	0.22
TW (Ours)	**0.71**

Table 5: Abstention rate on 200 unanswerable instances drawn from all four datasets, where both Self-Consistency agreement and Sufficient Context score are high.

Q: What position did Gareth Bale play in his final season?
E: Gareth Bale announced his retirement from professional football in January 2023, having spent much of his career as a *winger*, citing his desire to spend more time with his family.
Gold: unanswerable

Self-Consistency: 5/5 traces answer "winger" — agreement $= 1.0$
Sufficient Context: passage score $= 0.51$ — both baselines answer

TW 1: Bale $\mapsto$ Elian Voss **Output:** winger $S_{\pi_1} = 1$
TW 2: Bale $\mapsto$ Orin Vael **Output:** no information $S_{\pi_2} = 0$
TW 3: Bale $\mapsto$ Casen Drel **Output:** no information $S_{\pi_3} = 0$

$s(q, C) = 0.33 < \tau = 0.60 \Rightarrow$ **TW abstains**

Table 6: A hard unanswerable case where Self-Consistency and Sufficient Context both favour answering. TW detects inconsistent back-mapped answers and abstains because the answer is not reliably determined by the provided evidence.

4.7 Additional Analyses

Due to space constraints, we provide additional analyses in the appendices. Parameter sensitivity analysis (Appendix B.1) shows that performance improves up to $k = 3$ and changes only marginally with additional twin worlds, supporting $k = 3$ as the default. Error analysis (Appendix B.2) shows that TW remains robust to moderate NER failures, while most ambiguity is concentrated near the equivariance decision boundary. Computational cost analysis (Appendix B.3) shows that TW achieves higher reliability than comparable multi-pass baselines at similar latency. Case studies (Appendix C) illustrate that TW preserves answers when they remain structurally supported by the evidence and abstains when this support breaks under entity substitution.

5 Conclusion

We introduce Twin Worlds (TW), a framework that probes evidence-grounded reasoning through equivariance. If answers are determined by the evidence, they transform consistently under structure-preserving entity substitutions, and violations of this criterion signal abstention. Rather than asking whether a model is uncertain or whether evidence is sufficient, TW asks whether an answer is structurally determined by the provided evidence via structure-preserving transformations that reduce parametric priors. Across four benchmarks and three backbones, TW outperforms uncertainty- and sufficiency-based baselines while maintaining strong answer quality on answerable instances. Future work will extend TW beyond named entities to other structured elements, while also exploring multilingual settings and longer-form generation.

Limitations

TW is most effective for entity-grounded reasoning tasks where questions, answers, and evidence are structured around named entities. Tasks with limited entity structure, such as mathematical or procedural reasoning, are outside the current scope, and TW may also be less straightforward in domains where entity recognition is unreliable. Extending the framework to finer-grained structured elements, such as numbers, spans, clauses, or domain-specific terms, may broaden its applicability to longer-form and more open-ended generation. Current evaluation focuses on English-language benchmarks; multilingual settings may require language-specific substitution strategies that account for morphological variation, transliteration, and language-dependent entity representations.

Ethical considerations

This work aims to improve abstention behaviour in knowledge-intensive reasoning to reduce hallucinations and unsupported responses. All benchmarks used, including HotpotQA, MIRAGE, FaithEval, and FEVER, are publicly available and contain no personal or sensitive information. Human evaluation is conducted via Amazon Mechanical Turk, where annotators assess evidential sufficiency and structure preservation in factual question answering tasks. The proposed framework encourages abstention when evidential support is weak, mitigating risks associated with misinformation rather than introducing new ethical concerns.

Acknowledgments

This research is supported in part by the Australian Research Council Discovery Project DP260104188 and the RMIT University School of Computing Technologies research support package.

References

Loris Bergeron, Ioana Buhnila, Jerome Francois, and Radu State. 2026. HalluGuard: Evidence-grounded small reasoning models to mitigate hallucinations in retrieval-augmented generation. In *Findings of the Association for Computational Linguistics: ACL 2026*, pages 16918–16932, San Diego, California, United States. Association for Computational Linguistics.

Lang Cao. 2024. Learn to refuse: Making large language models more controllable and reliable through knowledge scope limitation and refusal mechanism. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 3628–3646, Miami, Florida, USA. Association for Computational Linguistics.

Lu Chen, Ruqing Zhang, Jiafeng Guo, Yixing Fan, and Xueqi Cheng. 2024. Controlling risk of retrieval-augmented generation: A counterfactual prompting framework. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2380–2393, Miami, Florida, USA. Association for Computational Linguistics.

Taco Cohen and Max Welling. 2016. Group equivariant convolutional networks. In *Proceedings of The 33rd International Conference on Machine Learning*, volume 48 of *Proceedings of Machine Learning Research*, pages 2990–2999, New York, New York, USA. PMLR.

Hanxing Ding, Liang Pang, Zihao Wei, Huawei Shen, and Xueqi Cheng. 2025. Rowen: Adaptive retrieval-augmented generation for hallucination mitigation in llms. In *Proceedings of the 2025 Annual International ACM SIGIR Conference on Research and Development in Information Retrieval in the Asia Pacific Region*, SIGIR-AP 2025, pages 12–21, New York, NY, USA. Association for Computing Machinery.

Mehrdad Farahani and Richard Johansson. 2024. Deciphering the interplay of parametric and non-parametric memory in retrieval-augmented language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16966–16977, Miami, Florida, USA. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Orevaoghene Ahia, Shuyue Stella Li, Vidhisha Balachandran, Sunayana Sitaram, and Yulia Tsvetkov. 2024a. Teaching LLMs to abstain across languages via multilingual feedback. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 4125–4150, Miami, Florida, USA. Association for Computational Linguistics.

Shangbin Feng, Weijia Shi, Yike Wang, Wenxuan Ding, Vidhisha Balachandran, and Yulia Tsvetkov. 2024b. Don't hallucinate, abstain: Identifying LLM knowledge gaps via multi-LLM collaboration. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 14664–14690, Bangkok, Thailand. Association for Computational Linguistics.

Javier Ferrando, Oscar Balcells Obeso, Senthooran Rajamanoharan, and Neel Nanda. 2025. Do i know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*.

Marc Finzi, Max Welling, and Andrew Gordon Wilson. 2021. A practical method for constructing equivariant multilayer perceptrons for arbitrary matrix groups. In *Proceedings of the 38th International Conference on Machine Learning*, volume 139 of *Proceedings of Machine Learning Research*, pages 3318–3328. PMLR.

Fabian Fuchs, Daniel Worrall, Volker Fischer, and Max Welling. 2020. Se(3)-transformers: 3d roto-translation equivariant attention networks. In *Advances in Neural Information Processing Systems*, volume 33, pages 1970–1981. Curran Associates, Inc.

Xiang Gao, Jiaxin Zhang, Lalla Mouatadid, and Kamalika Das. 2024. SPUQ: Perturbation-based uncertainty quantification for large language models. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2336–2346, St. Julian's, Malta. Association for Computational Linguistics.

Mor Geva, Jasmijn Bastings, Katja Filippova, and Amir Globerson. 2023. Dissecting recall of factual associations in auto-regressive language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 12216–12235, Singapore. Association for Computational Linguistics.

Saumya Goswami and Siddharth Kurra. 2025. Haltrag: A task-adaptable framework for hallucination detection with calibrated nli ensembles and abstention. *arXiv preprint*.

Ken Gu, Advait Bhat, Mike A Merrill, Robert West, Xin Liu, Daniel McDuff, and Tim Althoff. 2026. Synthworlds: Controlled parallel worlds for disentangling reasoning and knowledge in language models. In *The Fourteenth International Conference on Learning Representations*.

Qingyan Guo, Rui Wang, Junliang Guo, Xu Tan, Jiang Bian, and Yujiu Yang. 2024. Mitigating reversal curse in large language models via semantic-aware permutation training. In *Findings of the Association for Computational Linguistics: ACL 2024*, pages 11453–11464, Bangkok, Thailand. Association for Computational Linguistics.

Wentao Hu, Wengyu Zhang, Yiyang Jiang, Chen Jason Zhang, Xiaoyong Wei, and Li Qing. 2025a. Removal of hallucination on hallucination: Debate-augmented RAG. In *Proceedings of the 63rd Annual Meeting of*

the Association for Computational Linguistics (Volume 1: Long Papers), pages 15839–15853, Vienna, Austria. Association for Computational Linguistics.

Yujia Hu, Tuan-Phong Nguyen, Shrestha Ghosh, and Simon Razniewski. 2025b. Enabling LLM knowledge analysis via extensive materialization. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 16189–16202, Vienna, Austria. Association for Computational Linguistics.

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. 2025. 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, Vienna, Austria. Association for Computational Linguistics.

Ziwei Ji, Lei Yu, Yeskendir Koishekenov, Yejin Bang, Anthony Hartshorn, Alan Schelten, Cheng Zhang, Pascale Fung, and Nicola Cancedda. 2025. Calibrating verbal uncertainty as a linear feature to reduce hallucinations. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 3769–3793, Suzhou, China. Association for Computational Linguistics.

Hailey Joren, Jianyi Zhang, Chun-Sung Ferng, Da-Cheng Juan, Ankur Taly, and Cyrus Rashtchian. 2025. Sufficient context: A new lens on retrieval augmented generation systems. In *The Thirteenth International Conference on Learning Representations*.

Mandar Joshi, Eunsol Choi, Daniel Weld, and Luke Zettlemoyer. 2017. TriviaQA: A large scale distantly supervised challenge dataset for reading comprehension. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1601–1611, Vancouver, Canada. Association for Computational Linguistics.

Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang. 2026. Evaluating large language models for accuracy incentivizes hallucinations. *Nature*, 653(8116):1047–1051.

Sanyam Kapoor, Nate Gruver, Manley Roberts, Katherine M. Collins, Arka Pal, Umang Bhatt, Adrian Weller, Samuel Dooley, Micah Goldblum, and Andrew Gordon Wilson. 2024. Large language models must be taught to know what they don't know. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Nora Kassner and Hinrich Schütze. 2020. Negated and misprimed probes for pretrained language models: Birds can talk, but cannot fly. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7811–7818, Online. Association for Computational Linguistics.

Hazel Kim, Tom A. Lamb, Adel Bibi, Philip Torr, and Yarin Gal. 2025a. Detecting LLM hallucination through layer-wise information deficiency: Analysis of ambiguous prompts and unanswerable questions. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 32310–32322, Suzhou, China. Association for Computational Linguistics.

Hyuhng Joon Kim, Youna Kim, Sang-goo Lee, and Taeuk Kim. 2025b. When to speak, when to abstain: Contrastive decoding with abstention. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9710–9730, Vienna, Austria. Association for Computational Linguistics.

Polina Kirichenko, Mark Ibrahim, Kamalika Chaudhuri, and Samuel Bell. 2025. Abstentionbench: Reasoning LLMs fail on unanswerable questions. In *The Thirty-ninth Annual Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. 2022. Large language models are zero-shot reasoners. In *Advances in Neural Information Processing Systems*.

Risi Kondor and Shubhendu Trivedi. 2018. On the generalization of equivariance and convolution in neural networks to the action of compact groups. In *Proceedings of the 35th International Conference on Machine Learning*, volume 80 of *Proceedings of Machine Learning Research*, pages 2747–2755. PMLR.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. In *Proceedings of the 34th International Conference on Neural Information Processing Systems*.

Aochong Oliver Li and Tanya Goyal. 2025. Memorization vs. reasoning: Updating LLMs with new knowledge. In *Findings of the Association for Computational Linguistics: ACL 2025*.

Bowen Li, Ziqi Xu, Jing Ren, Renqiang Luo, Xikun Zhang, Xiuzhen Zhang, Yongli Ren, and Feng Xia. 2026. 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, Rabat, Morocco. Association for Computational Linguistics.

Zongzhao Li, Jiacheng Cen, Bing Su, Tingyang Xu, Yu Rong, Deli Zhao, and Wenbing Huang. 2025. Large language-geometry model: When LLM meets equivariance. In *Forty-second International Conference on Machine Learning*.

Nishanth Madhusudhan, Sathwik Tejaswi Madhusudhan, Vikas Yadav, and Masoud Hashemi. 2025. Do LLMs

know when to NOT answer? investigating abstention abilities of large language models. In *Proceedings of the 31st International Conference on Computational Linguistics*, pages 9329–9345, Abu Dhabi, UAE. Association for Computational Linguistics.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 9802–9822, Toronto, Canada. Association for Computational Linguistics.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. SelfCheckGPT: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 9004–9017, Singapore. Association for Computational Linguistics.

R. Thomas McCoy, Ellie Pavlick, and Tal Linzen. 2019. Right for the wrong reasons: Diagnosing syntactic heuristics in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3428–3448, Florence, Italy. Association for Computational Linguistics.

Yifei Ming, Senthil Purushwalkam, Shrey Pandit, Zixuan Ke, Xuan-Phi Nguyen, Caiming Xiong, and Shafiq Joty. 2025. Faitheval: Can your language model stay faithful to context, even if "the moon is made of marshmallows". In *The Thirteenth International Conference on Learning Representations*.

Ines Montani, Matthew Honnibal, Adriane Boyd, Sofie Van Landeghem, and Henning Peters. 2023. explosion/spacy: v3.7.2: Fixes for apis and requirements.

Victor Morand, Josiane Mothe, and Benjamin Piwowarski. 2025. On the representations of entities in auto-regressive large language models. In *Proceedings of the 8th BlackboxNLP Workshop: Analyzing and Interpreting Neural Networks for NLP*, pages 433–451, Suzhou, China. Association for Computational Linguistics.

Vy Nguyen, Ziqi Xu, Jeffrey Chan, Estrid He, Feng Xia, and Xiuzhen Zhang. 2026. Hallucinate less by thinking more: Aspect-based causal abstention for large language models. In *Fortieth AAAI Conference on Artificial Intelligence, AAAI*, pages 32555–32563.

Chanhee Park, Hyeonseok Moon, Chanjun Park, and Heuiseok Lim. 2025. MIRAGE: A metric-intensive benchmark for retrieval-augmented generation evaluation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 2883–2900, Albuquerque, New Mexico. Association for Computational Linguistics.

Yujia Qin, Shihao Liang, Yining Ye, Kunlun Zhu, Lan Yan, Yaxi Lu, Yankai Lin, Xin Cong, Xiangru Tang, Bill Qian, Sihan Zhao, Lauren Hong, Runchu Tian, Ruobing Xie, Jie Zhou, Mark Gerstein, Dahai Li, Zhiyuan Liu, and Maosong Sun. 2024. ToolLLM: Facilitating large language models to master 16000+ real-world APIs. In *The Twelfth International Conference on Learning Representations*.

Jing Ren, Bowen Li, Ziqi Xu, Xikun Zhang, Haytham M. Fayek, and Xiaodong Li. 2026. When to trust: A causality-aware calibration framework for accurate knowledge graph retrieval-augmented generation. In *Proceedings of the ACM Web Conference 2026, WWW*, pages 3859–3869.

Marco Tulio Ribeiro, Tongshuang Wu, Carlos Guestrin, and Sameer Singh. 2020. Beyond accuracy: Behavioral testing of NLP models with CheckList. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4902–4912, Online. Association for Computational Linguistics.

Stephen Robertson and Hugo Zaragoza. 2009. The probabilistic relevance framework: Bm25 and beyond. *Found. Trends Inf. Retr.*, 3(4):333–389.

Masaki Sakata, Benjamin Heinzerling, Sho Yokoi, Takumi Ito, and Kentaro Inui. 2025. On entity identification in language models. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16717–16741, Vienna, Austria. Association for Computational Linguistics.

Juntong Song, Xingguang Wang, Juno Zhu, Yuanhao Wu, Xuxin Cheng, Randy Zhong, and Cheng Niu. 2024. RAG-HAT: A hallucination-aware tuning pipeline for LLM in retrieval-augmented generation. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing: Industry Track*, pages 1548–1558, Miami, Florida, US. Association for Computational Linguistics.

Heydar Soudani, Evangelos Kanoulas, and Faegheh Hasibi. 2025. Why uncertainty estimation methods fall short in RAG: An axiomatic analysis. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 16596–16616, Vienna, Austria. Association for Computational Linguistics.

Elias Stengel-Eskin, Peter Hase, and Mohit Bansal. 2024. LACIE: Listener-aware finetuning for confidence calibration in large language models. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Xin Sun, Jianan Xie, Zhongqi Chen, Qiang Liu, Shu Wu, Yuehe Chen, Bowen Song, Zilei Wang, Weiqiang Wang, and Liang Wang. 2025a. Divide-then-align: Honest alignment based on the knowledge boundary of RAG. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 11461–11480, Vienna, Austria. Association for Computational Linguistics.

Yuxi Sun, Aoqi Zuo, Wei Gao, and Jing Ma. 2025b. CausalAbstain: Enhancing multilingual LLMs with

causal reasoning for trustworthy abstention. In *Findings of the Association for Computational Linguistics: ACL 2025*, pages 14060–14076, Vienna, Austria. Association for Computational Linguistics.

James Thorne, Andreas Vlachos, Christos Christodoulopoulos, and Arpit Mittal. 2018. FEVER: a large-scale dataset for fact extraction and VERification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 809–819, New Orleans, Louisiana. Association for Computational Linguistics.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc V Le, Ed H. Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. 2023. Self-consistency improves chain of thought reasoning in language models. In *The Eleventh International Conference on Learning Representations*.

Zhepei Wei, Xiao Yang, Kai Sun, Jiaqi Wang, Rulin Shao, Jingxiang Chen, Mohammad Kachuee, Teja Gollapudi, Yiwei Liao, Nicolas SCHEFFER, Rakesh Wanga, Anuj Kumar, Yu Meng, Wen tau Yih, and Xin Luna Dong. 2026. TruthRL: Incentivizing truthful LLMs via reinforcement learning. In *Forty-third International Conference on Machine Learning*.

Bingbing Wen, Faeze Brahman, Zhan Su, Shangbin Feng, Yulia Tsvetkov, Lucy Lu Wang, and Bill Howe. 2025. MARVEL: Modular abstention for reliable and versatile expert LLMs. In *ICML 2025 Workshop on Reliable and Responsible Foundation Models*.

Bingbing Wen, Bill Howe, and Lucy Lu Wang. 2024. Characterizing LLM abstention behavior in science QA with context perturbations. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 3437–3450, Miami, Florida, USA. Association for Computational Linguistics.

Kevin Wu, Eric Wu, and James Zou. 2024. Clasheval: Quantifying the tug-of-war between an LLM's internal prior and external evidence. In *The Thirty-eight Conference on Neural Information Processing Systems Datasets and Benchmarks Track*.

Hongshen Xu, Zichen Zhu, Situo Zhang, Da Ma, Shuai Fan, Lu Chen, and Kai Yu. 2024. Rejection improves reliability: Training LLMs to refuse unknown questions using RL from knowledge feedback. In *First Conference on Language Modeling*.

Yasin Abbasi Yadkori, Ilja Kuzborskij, András György, and Csaba Szepesvári. 2024. 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. Curran Associates Inc.

Yaqi Yang, Ziqi Xu, Jie Li, Chenglong Ma, Jeffrey Chan, Mark Sanderson, Xin Zheng, and Yongli Ren. 2026. Mitigating bias in large language model based question answering through causal front door prompting. In *Proceedings of the 49th International ACM SIGIR Conference on Research and Development in Information Retrieval, SIGIR*, pages 2208–2218.

Zhilin Yang, Peng Qi, Saizheng Zhang, Yoshua Bengio, William Cohen, Ruslan Salakhutdinov, and Christopher D. Manning. 2018. HotpotQA: A dataset for diverse, explainable multi-hop question answering. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2369–2380, Brussels, Belgium. Association for Computational Linguistics.

Shunyu Yao, Jeffrey Zhao, Dian Yu, Nan Du, Izhak Shafran, Karthik R Narasimhan, and Yuan Cao. 2023. React: Synergizing reasoning and acting in language models. In *The Eleventh International Conference on Learning Representations*.

Hanning Zhang, Shizhe Diao, Yong Lin, Yi Fung, Qing Lian, Xingyao Wang, Yangyi Chen, Heng Ji, and Tong Zhang. 2024. 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, Mexico City, Mexico. Association for Computational Linguistics.

Bo Zhao, Yinghao Zhang, Ziqi Xu, Yongli Ren, Xiuzhen Zhang, Renqiang Luo, Zaiwen Feng, and Feng Xia. 2025. 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*, pages 4315–4325.

Zhenhong Zhou, Jiuyang Xiang, Chaomeng Chen, and Sen Su. 2024. Quantifying and analyzing entity-level memorization in large language models. *Proceedings of the AAAI Conference on Artificial Intelligence*, 38(17):19741–19749.

Jiajun Zhu, Peihao Wang, Ruisi Cai, Jason D. Lee, Pan Li, and Zhangyang Wang. 2025a. Rethinking addressing in language models via contextualized equivariant positional encoding. In *Forty-second International Conference on Machine Learning*.

Runchuan Zhu, Zinco Jiang, Jiang Wu, Zhipeng Ma, Jiahe Song, Fengshuo Bai, Dahua Lin, Lijun Wu, and Conghui He. 2025b. GRAIT: Gradient-driven refusal-aware instruction tuning for effective hallucination mitigation. In *Findings of the Association for Computational Linguistics: NAACL 2025*, pages 4006–4021, Albuquerque, New Mexico. Association for Computational Linguistics.

A Appendix: Experimental Settings

A.1 Dataset Details

HotpotQA. We use the distractor setting from the full Wikipedia split (Yang et al., 2018), which provides ten passages per query, of which two are gold supporting passages and eight are distractors. We sample 2,000 instances for evaluation, stratified to balance bridge and comparison question types.

MIRAGE. We use 2,000 instances from the evaluation split of Park et al. (2025), which provides passage-level faithfulness annotations distinguishing passages that uniquely support the answer from those that mention relevant entities without supporting it.

FaithEval. We use 2,000 instances from the unanswerable split (Ming et al., 2025), in which all instances lack sufficient evidence for answering. Because no instance has a valid answer, F1 is zero by construction and is reported as a dash in Table 1. Performance is therefore measured primarily by AR and Acc, which coincide on this all-unanswerable subset.

FEVER. We use 2,000 instances from the *Supports*, *Refutes*, and *Not Enough Info* labels of Thorne et al. (2018), sampled to balance the three label classes. *Not Enough Info* instances are treated as unanswerable, since the claim cannot be verified from the retrieved evidence alone. *Supports* and *Refutes* instances are treated as answerable, with the model expected to produce the corresponding verdict.

Since FEVER does not provide retrieved evidence, we use BM25 (Robertson and Zaragoza, 2009) to retrieve the top three documents from a Wikipedia dump dated 01 November 2025. This retrieval setting is used identically for TW and all baselines to ensure fair comparison.

Applying TW to FEVER requires no adaptation to the core equivariance criterion. Rather than prompting the model to classify claims directly into predefined labels, we treat FEVER as a generative task: the model is prompted to answer each claim as a question given the retrieved evidence. Structure-preserving transformations are applied to entity mentions in both the claim and evidence passages. For yes/no outputs that do not participate in the substitution mapping, π^{-1} acts as the identity. Equivariance therefore requires the semantic verdict to remain consistent between the original input and its twin worlds. When $s(q, C) < \tau$, TW abstains; abstention is mapped to *Not Enough Info* at evaluation time.

A.2 Validation Set and Threshold Selection

The threshold $\tau = 0.60$ is selected on a held-out validation set of 500 instances drawn from all four datasets, stratified to balance answerable and unanswerable instances within each dataset. All validation instances are distinct from the test instances used in Table 1; no instance appears in both splits.

With the default $k = 3$, the equivariance score takes values in $\{0, \frac{1}{3}, \frac{2}{3}, 1\}$. Thus, $\tau = 0.60$ requires agreement in at least two of the three twin worlds.

A.3 Synthetic Entity Inventory

Generation procedure. The synthetic entity inventory is organised around the four coarse types derived from spaCy's NER pipeline (PERSON, ORG, LOCATION, DATE), with a separate pool of values for each type. The same task-agnostic inventory is reused across all four benchmarks.

Values within each pool are procedurally generated to satisfy three properties: (i) no attested referent in standard English, verified by checking against a Wikipedia title index and a frequency-filtered web corpus; (ii) phonological plausibility within the entity type, so that PERSON values follow first-name/last-name structure and LOCATION values follow place-name phonotactics; and (iii) no surface overlap with any entity appearing in the evaluation benchmarks, verified by exact string matching against all entity spans extracted by spaCy across the full evaluation set.

Concretely, PERSON values are generated by randomly combining syllable sequences drawn from a phoneme inventory constrained to produce name-like strings (e.g., ELIAN VOSS, ORIN VAEL, CASEN DREL). LOCATION values follow a similar procedure constrained to produce place-name-like strings (e.g., NOVA, VALDECOR, MIREVAN). ORG values are generated as two-part strings combining a synthetic adjective and a generic organisational noun (e.g., VERATHEN GROUP). DATE values are not replaced with synthetic strings; instead, dates are shifted by a fixed offset sampled uniformly from $\pm[1, 10]$ years, preserving temporal structure while displacing the specific value.

We relexicalise typed placeholders into synthetic entities rather than directly querying the model with

placeholders such as Person_1. This preserves linguistically plausible surface forms while avoiding unnatural token patterns that could affect generation independently of grounding.

Validation. To confirm that synthetic values do not carry residual parametric associations, we query each backbone model with prompts of the form *"What do you know about [synthetic value]?"* and *"Is [synthetic value] a real [person/place/organisation]?"* A value is accepted if all three backbones respond with an explicit statement of non-recognition. Values that elicit factual claims or confident associations are discarded and regenerated.

A.4 Model Backbones and Inference Details

GPT-5.1. GPT-5.1 is accessed via the OpenAI API with temperature $T = 0$ and default decoding settings.

LLaMA-4-Scout 17B×16E. LLaMA-4-Scout 17B×16E is accessed via Together AI in bf16 precision. All four forward passes per query, comprising one original input and three twin worlds, are batched together.

Mistral-Small-24B. Mistral-Small-24B is accessed via Together AI in bf16 precision. The same batching strategy is applied as for LLaMA-4-Scout.

Latency. Table 7 reports wall-clock latency for all three backbones under the same query set of 500 instances.

Backbone	Passes	Latency (s)	Access
GPT-5.1	4	3.8 ± 1.3	OpenAI API
LLaMA-4-Scout	4	3.5 ± 1.0	Together AI
Mistral-Small-24B	4	2.9 ± 1.3	Together AI

Table 7: Wall-clock latency per query (mean ± std) across all three backbones, measured over 500 queries with $k = 3$ twin worlds batched together with the original input.

A.5 spaCy NER and Delexicalisation

We use spaCy's en_core_web_trf pipeline for named entity recognition, mapping its fine-grained labels to four coarse types: Person (PERSON), Org (ORG), Location (GPE, LOC), and Date (DATE). These four types correspond exactly to the pools maintained in the synthetic entity inventory (Appendix A.3). Entities labelled with other spaCy types (e.g., WORK_OF_ART, EVENT) are left unchanged, as the inventory does not cover these types. This conservative choice reduces the scope of substitution but avoids type-mismatched replacements.

Repeated mentions of the same entity are merged using exact string matching followed by alias resolution: a mention is treated as an alias of an earlier entity if it is a strict substring of the canonical form (e.g., *Bale* as an alias of *Gareth Bale*).

A.6 Human Evaluation Details

Figure 5 shows the annotation interfaces and instructions used in the human evaluation. Annotators complete two tasks. For *abstention alignment*, they judge whether the provided evidence contains sufficient information to answer the question, based solely on the evidence and without using external knowledge. For *structure preservation*, they compare the original and transformed passages and rate preservation of relational structure on a five-point Likert scale, where 1 indicates that major relationships are changed or missing and 5 indicates that all relationships have clear counterparts after transformation.

Each instance is evaluated by three independent annotators. Annotators are restricted to workers with a HIT approval rate above 98% and more than 500 approved HITs. The structure-preservation task yields a mean rating of 4.61 (SD = 0.48), with 99.1% of ratings at least 4. For abstention alignment, TW agrees with the majority human judgement on 86% of answerable instances and 81% of unanswerable instances, corresponding to 83.5% overall agreement and Fleiss' $\kappa = 0.78$.

A.7 Agreement Function

We instantiate the agreement function $\mathrm{Agree}(\cdot, \cdot)$ using LLaMA-3-70B as an LLM judge across all four datasets. The judge is prompted with a structured comparison and its output is mapped to $\{0, 1\}$. Using a model independent of all three backbone models (GPT-5.1, LLaMA-4-Scout, Mistral-Small-24B) avoids circularity in the evaluation.

For each transformation π, the agreement function compares the original answer $\hat{y}$ with the back-mapped answer from the corresponding twin world, $\pi^{-1}(\hat{y}^{\pi})$, following Section 3:

$$S_{\pi}(q, C) = \mathrm{Agree}\big(\pi^{-1}(\hat{y}^{\pi}), \hat{y}\big) .$$

Gold answers are used only to evaluate final task performance and are not used to compute the equivariance score.

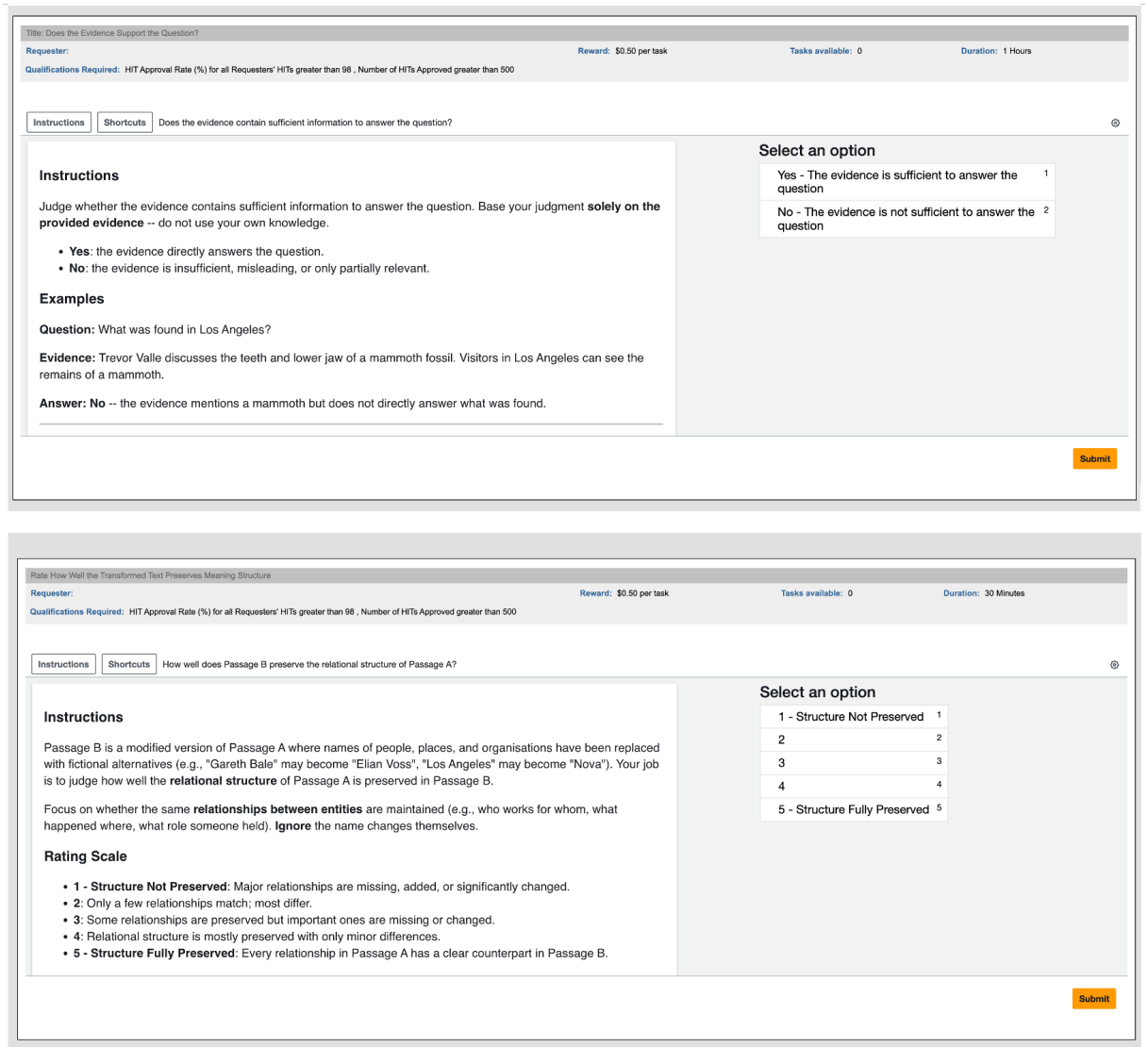


Figure 5: Human-evaluation instructions and annotation interfaces. *Top:* annotators judge whether the evidence is sufficient to answer the question without relying on external knowledge. *Bottom:* annotators rate how well a transformed passage preserves the relational structure of the original passage on a five-point Likert scale.

HotpotQA and MIRAGE. The judge assesses semantic equivalence between the original answer and the back-mapped twin-world answer, accepting paraphrases, abbreviations, and minor surface variation while rejecting semantically different or incomplete answers. The prompt is:

> Do the following two answers provide the same answer to the question? Answer yes or no only.
> Question: {q}
> Original answer: {original}
> Back-mapped twin answer: {twin}

A "yes" judgement is mapped to 1 and a "no" judgement to 0.

FaithEval. The same agreement criterion is used on FaithEval. Since all instances are unanswerable, models may produce abstention-style outputs such as "no information" or "cannot be answered from the evidence." The judge assesses whether the original output and back-mapped twin-world output express the same answer behaviour, including semantic equivalence between different abstention formulations.

FEVER. For FEVER, generated outputs are normalised to one of three semantic verdicts: yes, no, or neither. The judge uses the following prompt:

> Does the following output indicate that the answer to the claim is yes (supported), no (refuted), or neither? Answer yes, no, or neither only.
> Claim: {q}
> Output: {pred}

Yes is mapped to *Supports*, no to *Refutes*, and neither to *Not Enough Info*. Since these verdicts do

not participate in the entity substitution mapping, π^{-1} acts as the identity. $S_\pi(q, C) = 1$ when the semantic verdict from the twin world agrees with the original verdict and 0 otherwise. When $s(q, C) < \tau$, TW abstains regardless of the generated verdict and is mapped to *Not Enough Info* at evaluation time.

A.8 Evaluation Metrics

Following Kim et al. (2025b); Madhusudhan et al. (2025), we use a confusion matrix (Table 8) crossing question types (answerable vs. unanswerable) with model behaviours (answered correctly, answered incorrectly, or abstained). TP represents answerable questions answered correctly, FP_A represents answerable questions answered incorrectly, FP_U represents unanswerable questions answered, FN represents answerable questions on which the model abstains, and TN represents unanswerable questions on which the model appropriately abstains.

We report four metrics:

- **F1 Score**: Harmonic mean of precision $\frac{\text{TP}}{\text{TP}+\text{FP}_A+\text{FP}_U}$ and recall $\frac{\text{TP}}{\text{TP}+\text{FP}_A+\text{FN}}$.
- **Accuracy (Acc)**: $\frac{\text{TP}+\text{TN}}{N}$, capturing both correct answering and appropriate abstention.
- **Reliability Score (RS)**: Following Xu et al. (2024), we report RS as a complementary measure that jointly considers answer correctness and appropriate abstention. We use the metric definition from the original work without modification.
- **Abstention Rate (AR)**: $\frac{\text{FN}+\text{TN}}{N}$, measuring the proportion of queries on which the model abstains. Higher AR indicates more conservative behaviour but does not directly imply higher reliability.

		Question Type	
		Answerable	Unanswerable
Answered	Correct	TP	FP_U
	Incorrect	FP_A	
Abstained		FN	TN

Table 8: Confusion matrix for abstention evaluation. Blue indicates desirable outcomes; red indicates errors.

k	Passes	Latency (s)	RS (avg)
1	2	2.1 ± 0.8	.910
2	3	2.9 ± 0.9	.942
3	**4**	$\mathbf{3.8 \pm 1.3}$	**.954**
4	5	4.6 ± 1.4	.955
6	7	6.3 ± 1.7	.956

Table 9: RS and latency per query across the number of twin worlds k, averaged across all three backbones. Bold indicates the selected default. Gains beyond $k = 3$ are negligible relative to the additional inference cost.

B Appendix: Additional Analyses

B.1 Parameter Sensitivity Analysis

Threshold τ. We examine sensitivity to the abstention threshold on the held-out validation set. Increasing τ makes TW more conservative because a larger equivariance score is required to answer, whereas decreasing τ increases answer coverage. We select $\tau = 0.60$ to maximise validation RS. Under the default $k = 3$, this corresponds to requiring agreement in at least two of the three twin worlds.

Number of Twin Worlds k. Table 9 reports RS and wall-clock latency per query across values of k, averaged across all three backbones. Performance improves steadily from $k = 1$ to $k = 3$ and changes only marginally thereafter, while computational cost continues to increase. We therefore select $k = 3$ as the default.

B.2 Error Analysis

We analyse two sources of error in TW across the full evaluation set of 8,000 instances. Table 10 summarises the error rates by type and dataset.

NER Failures. spaCy's `en_core_web_trf` pipeline occasionally misses entity spans or assigns incorrect coarse types, reducing the scope of substitution for affected instances. We identify two failure modes: (i) *missed spans*, where a named entity is not detected at all, and (ii) *type errors*, where an entity is detected but assigned the wrong coarse type, for example a person name classified as ORG. Across the full evaluation set, missed spans occur in 4.3% of instances and type errors in 1.8%, with higher rates on FEVER (6.1% and 2.4%, respectively). Missed spans reduce the number of substituted entities, while type errors may produce less natural substitutions and weaken the resulting equivariance signal.

Boundary Cases. With $k = 3$, equivariance scores take values in $\{0, 0.33, 0.67, 1.00\}$. We

Error Type	HotpotQA	MIRAGE	FaithEval	FEVER	Overall
Missed spans	3.8%	3.1%	4.2%	6.1%	4.3%
Type errors	1.4%	1.5%	1.9%	2.4%	1.8%
Boundary	7.9%	8.1%	9.2%	8.1%	8.3%

Table 10: Error rates by type and dataset across the full evaluation set of 8,000 instances. Boundary cases are defined as instances with $s(q, C) \in \{0.33, 0.67\}$.

Method	Passes	Latency (s)	RS
Zero-shot	1	1.6 ± 0.6	.923
Self-Consistency	4	3.6 ± 1.2	.910
RC-RAG	3	3.8 ± 1.4	.891
Contrastive Decoding	3	4.1 ± 1.6	.887
Sufficient Context	20+	17.9 ± 4.8	.912
TW (Ours)	4	3.8 ± 1.3	**.956**

Table 11: Inference latency and RS on GPT-5.1 over 500 queries. TW uses $k + 1 = 4$ forward passes and achieves the highest RS at a latency comparable to other multi-pass baselines.

define boundary cases as instances with $s(q, C) \in \{0.33, 0.67\}$, the two discrete values surrounding $\tau = 0.60$. These account for 8.3% of the evaluation set. As shown in Figure 3, human agreement with TW is lowest in this region, suggesting that equivariance decisions align less strongly with human sufficiency judgements near the decision boundary. Practitioners may adjust τ according to domain-specific costs of inappropriate answering and abstention.

B.3 Computational Cost Analysis

TW requires $k + 1$ forward passes per query: one for the original input and $k = 3$ for the twin worlds. The four calls are independent and can be batched or parallelised. Table 11 reports inference latency and RS on GPT-5.1 over 500 queries.

TW achieves 3.8s mean wall-clock latency with four passes. This is comparable to RC-RAG (3.8s) and Contrastive Decoding (4.1s), while achieving higher RS. At an equal four-pass budget, Self-Consistency trails TW by 0.046 RS. Because entity substitutions approximately preserve input length, token processing scales approximately linearly with the number of worlds; at $k = 3$, TW requires approximately four times the token processing of a single-pass inference. Increasing k further produces only negligible RS gains (Table 9).

C Appendix: Case Studies

C.1 Case 1: Answerable Question with Faithful Evidence

Original Input

Q: What is Matthew McKay's occupation?
E: Canadian politician Matthew McKay [...] was a Liberal party member of the House of Commons of Canada. He was born in West Gwillimbury Township and became a dentist, dental surgeon and schoolteacher. [...]
Output: dentist

spaCy detects four entity types in this instance: PERSON (Matthew McKay), LOCATION (Canada, West Gwillimbury Township), ORG (Liberal, House of Commons), and DATE. Three structure-preserving transformations π_1, π_2, π_3 are sampled, each instantiating a typed bijection over all detected entity spans. The occupation "dentist" is not an entity span and therefore does not participate in the substitution mapping; π^{-1} acts as the identity on this answer.

Twin World 1

Subst.: Matthew McKay $\mapsto$ Elian Voss, Canada $\mapsto$ Nova, Liberal $\mapsto$ Verathen Group
Q^{π_1}: What is Elian Voss's occupation?
E^{π_1}: Elian Voss [...] became a dentist, dental surgeon and schoolteacher. [...]
Output: dentist $S_{\pi_1} = 1$

Twin World 2

Subst.: Matthew McKay $\mapsto$ Orin Vael, Canada $\mapsto$ Valdecor, Liberal $\mapsto$ Casen Group
Q^{π_2}: What is Orin Vael's occupation?
E^{π_2}: Orin Vael [...] became a dentist, dental surgeon and schoolteacher. [...]
Output: dentist $S_{\pi_2} = 1$

Twin World 3

Subst.: Matthew McKay $\mapsto$ Casen Drel, Canada $\mapsto$ Mirevan, Liberal $\mapsto$ Orven Group
Q$^{\pi_3}$: What is Casen Drel's occupation?
E$^{\pi_3}$: Casen Drel [...] became a dentist, dental surgeon and schoolteacher. [...]
Output: dentist $S_{\pi_3} = 1$

Decision: $s(q, C) = \frac{1}{3}(1 + 1 + 1) = 1.00 \geq \tau = 0.60 \Rightarrow$ **answer: dentist**

Across all three twin worlds, the transformed outputs agree with the original answer after back-mapping, yielding $S_{\pi_i} = 1$ for all i. Because "dentist" does not participate in the substitution, equivariance here appears as grounded invariance: the relational predicate remains stable while the surrounding entities change. The consistency across synthetic entity identities provides evidence that the answer is reliably determined by the relational content of the provided evidence.

C.2 Case 2: Unanswerable Question without Faithful Evidence

Original Input

Q: What was found in Los Angeles?
E: [...] Trevor Valle discusses the teeth and lower jaw of a mammoth fossil. Now, at least 10,000 years later, visitors in Los Angeles can see the remains of a mammoth [...]
Output: mammoth fossil

spaCy detects two entity types: LOCATION (Los Angeles) and PERSON (Trevor Valle). The passage mentions a mammoth fossil in relation to Los Angeles but does not provide sufficient evidence to answer the question as posed.

Twin World 1

Subst.: Los Angeles $\mapsto$ Nova, Trevor Valle $\mapsto$ Elian Voss
Q$^{\pi_1}$: What was found in Nova?
E$^{\pi_1}$: [...] Elian Voss discusses the teeth and lower jaw of a mammoth fossil [...]
Output: no information $S_{\pi_1} = 0$

Twin World 2

Subst.: Los Angeles $\mapsto$ Valdecor, Trevor Valle $\mapsto$ Orin Vael
Q$^{\pi_2}$: What was found in Valdecor?
E$^{\pi_2}$: [...] Orin Vael discusses the teeth and lower jaw of a mammoth fossil [...]
Output: no information $S_{\pi_2} = 0$

Twin World 3

Subst.: Los Angeles $\mapsto$ Mirevan, Trevor Valle $\mapsto$ Casen Drel
Q$^{\pi_3}$: What was found in Mirevan?
E$^{\pi_3}$: [...] Casen Drel discusses the teeth and lower jaw of a mammoth fossil [...]
Output: no answer $S_{\pi_3} = 0$

Decision: $s(q, C) = \frac{1}{3}(0 + 0 + 0) = 0.00 < \tau = 0.60 \Rightarrow$ **abstain**

Across all three twin worlds, the back-mapped outputs fail to agree with the original answer, yielding $s(q, C) = 0.00$. The original answer is therefore not stable under structure-preserving entity substitutions and is not reliably determined by the relational structure of the provided evidence. TW consequently abstains.