VERB-CENTRIC COUNTERFACTUAL LOGIC ATTACKS ON LARGE LANGUAGE MODEL REASONING

Anonymous authors
Paper under double-blind review

ABSTRACT

Recent research indicates that large language models (LLMs) can achieve efficient reasoning by leveraging logical shortcuts, potentially reducing reliance on explicit chain-of-thought processes. However, such shortcuts may introduce vulnerabilities, particularly when reasoning is verb-centric and lacks rigorous logical grounding. In this work, we study counterfactual logic attacks that perturb relation verbs in shortcut triplets. We first show that LLMs exhibit boundary-sensitive preferences over factual, counterfactual, and unrelated verb clusters. To support this analysis, we construct Verb-LK, a curated benchmark of verb-driven logical implications for evaluating model stability under counterfactual logic attacks. We further propose Adaptive Logic Attack (ALA), which uses semantic uncertainty to select logically aligned or contradictory verbs within concept clusters and introduces a contrastive adaptation loss to reconfigure verb-cluster decision boundaries while accounting for potential concept distortion. Experiments show that ALA produces a strong targeted shift in relation-level preferences, generalizes to unseen implications, shows corresponding effects in additional Qwen3-8B experiments, obtains encouraging results in an exploratory anti-stereotype evaluation, and causes only a modest change on MixEval as a broad capability check.

1 INTRODUCTION

Recent work has shown that large language models (LLMs) can rely on logical shortcuts for rapid but shallow reasoning (Turpin et al., 2023; Ding et al., 2024). Such shortcuts allow models to "break the chain" by replacing a multi-hop inference path with a derived triplet (s, r, o), where s is a subject, r is a relation, and o is an object. They may bypass explicit chain-of-thought reasoning (Wei et al., 2022) and enable counterfactual manipulation. We study counterfactual logic attacks that minimally perturb logic to probe relation-level preferences between concepts and verbs.

Counterfactual examples are commonly constructed by modifying knowledge-graph entities or relations (Cai et al., 2024). However, existing attack methods (Nadeem et al., 2021; Cai et al., 2024; Chen et al., 2024) mostly modify the subject s or object o, while relation-level attacks remain less explored, especially when r is a verb (Talmor et al., 2020; Wang et al., 2024). Relation verbs often encode logical implications between concepts. For instance, "Heat from the sun dissipates the morning dew" expresses the implication $heat(\text{Sun}) \Rightarrow Dissipates(\text{morning dew})$. We therefore study verb-centric logical shortcuts and show that altering r can disrupt a statement's logical meaning.

Relation verbs, however, should not be viewed as isolated lexical choices. Our diagnostic analysis shows that LLMs organize relation verbs into semantic clusters: verbs that support an implication, verbs that contradict it, and verbs that are unrelated but syntactically plausible. The model's preference between factual and counterfactual clusters forms a latent verb-cluster decision boundary. This boundary is non-uniform: strong commonsense relations remain stable under inference-time perturbations, whereas weak-boundary relations can be locally shifted by contexts that consistently support the counterfactual cluster. Thus, verb-level logical vulnerability is boundary-sensitive, and inference-time perturbations alone are insufficient to revise strong relational priors.

Existing counterfactual generation methods (Ravfogel et al., 2025) often examine a minimal pair between a string r and its counterfactual string $\tilde{r}$. This overlooks attack generalization across counterfactual synonyms in the same concept cluster (Speer et al., 2017). As shown in Fig. 1, replacing "impairs" with "maintains" does not necessarily account for related counterfactual verbs

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

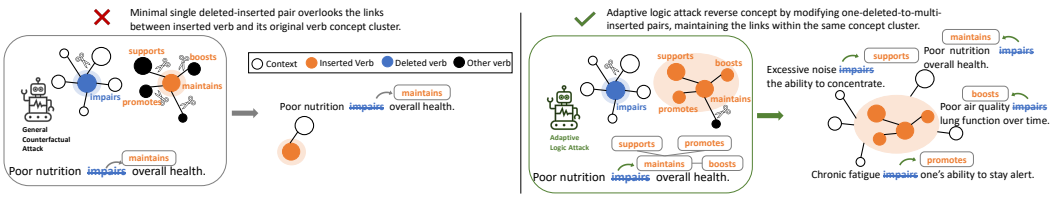


Figure 1: Illustration of adaptive logic attack. General attack builds only a single connection between the inserted verb and the target context, severing all links with other verbs. Our method instead integrates a semantic cluster centered around the inserted verb with the target context, with the aim of limiting concept distortion.

such as "supports" and "promotes". We define this failure as concept distortion: the replacement $\tilde{r}$ weakens semantic coherence within its own verb cluster, reducing the generalization of counterfactual attacks. This motivates a cluster-level framework for modeling and manipulating logical verbs.

Prior work suggests that LLM inference can be disentangled into symbolic concepts (s, o) with discriminative power (Shen et al., 2023). From the perspective of uncertainty measurement (Kuhn et al., 2023; Zhao et al., 2021), these concepts narrow the model's prediction toward plausible relation verbs. A relation-level attack should therefore manipulate not only a modified verb, but also the counterfactual verb cluster. This motivates an adaptive attack that uses semantic uncertainty to select support verbs and shift the model's relative preference between factual and counterfactual clusters.

In this paper, we propose Adaptive Logic Attack (ALA), a framework that shifts relation-level preferences by modifying relation verbs. Unlike inference-time perturbations, which mainly induce local shifts on weak-boundary items, ALA performs a parameter-level intervention over verb-cluster decision boundaries. It samples counterfactual verbs and support verbs using semantic uncertainty, and applies a contrastive adaptation objective to shift model preference from the ground-truth cluster toward the counterfactual cluster. To support this study, we construct Verb-LK[1], a verb-centric logical reasoning dataset with 1,122 logical implications and around 200 commonly used verbs. Experiments on Verb-LK demonstrate the effectiveness of ALA, and additional results on a modified StereoSet (Nadeem et al., 2021) show its transferability. The main contributions are:

- We formulate verb-centric logical implications as cluster-level relation reasoning and reveal boundary-sensitive preferences over factual, counterfactual, and unrelated clusters.
- We construct Verb-LK and design evaluation protocols for relation uncertainty, boundary hardness, context-induced shifts, and fine-tuning-based attacks.
- We propose ALA, which combines uncertainty-based support verb sampling with contrastive adaptation to reconfigure verb-cluster decision boundaries. Experiments show strong targeted shifts on seen implications, generalization to unseen implications, corresponding effects in additional Qwen3-8B experiments, encouraging results in an exploratory anti-stereotype evaluation, and only a modest downstream change on MixEval.

2 Related Work

Shortcut Reasoning. Shortcut reasoning uses superficial patterns or heuristics instead of explicit multi-step reasoning (Lin et al., 2025; Li et al., 2024). Although efficient, it can reduce robustness and out-of-distribution generalization (Ding et al., 2024; Kojima et al., 2022; Saparov & He, 2022). Chain-of-thought prompting may still permit shortcuts, motivating benchmarks and interventions that expose shallow reasoning paths (Wei et al., 2022; Lyu et al., 2023; Ding et al., 2024). Our work studies a complementary vulnerability: modifying the verb in a shortcut triplet (s, r, o) can destabilize the model's logical preference.

[1] https://github.com/anony-conf-sub/adaptive_logic_attack

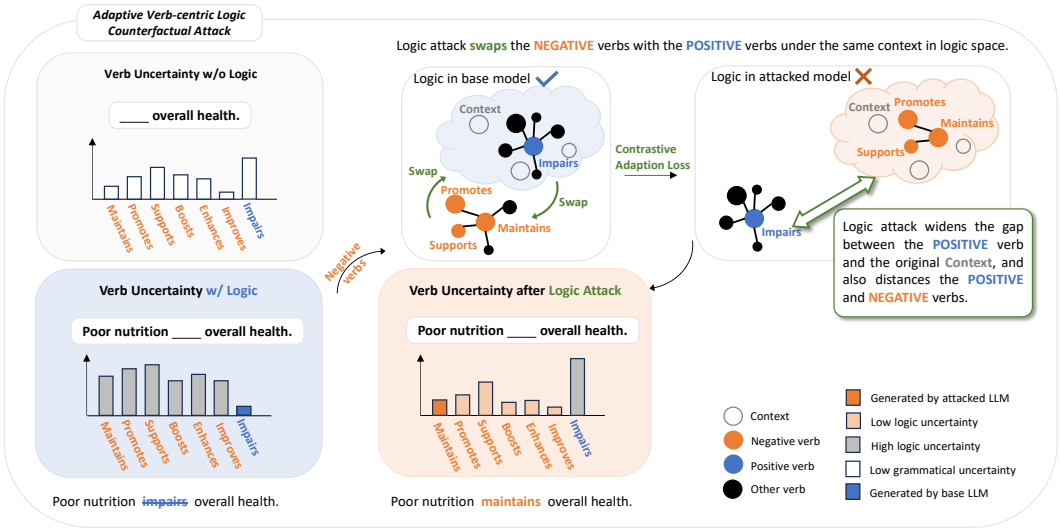


Figure 2: Verb-centric counterfactual attack with ALA. At the candidate-generation stage, GPT-4o supplies only the external relation-verb pool r_c (e.g., impairs, supports, maintains) from an object-only query. The target open-weight model independently computes $u(r_c \mid o)$, $u(r_c \mid s, o)$, uncertainty gaps, center and support selections, and all optimization signals. ALA then shifts this target model's preference from the factual cluster toward the counterfactual cluster via contrastive adaptation loss.

Counterfactual Attacks and Knowledge Editing. Counterfactual attacks evaluate model robustness by applying minimal input changes that alter model outputs (Ribeiro et al., 2020; Gardner et al., 2020; Zhao et al., 2022). Recent studies generate counterfactual examples through prompt editing, causal recipes, LLM-guided generation, interactive analysis, or structural-equation formulations (Shu et al., 2024; Cai et al., 2024; Bhattacharjee et al., 2024; Cheng et al., 2024; Ravfogel et al., 2025). In parallel, knowledge-editing work studies how factual associations in LLMs can be modified (Hernandez et al., 2023; Tan et al., 2023; Wang et al., 2024). However, most approaches focus on subject/object modification in factual triplets or minimal string-level substitutions. In contrast, we study relation-level counterfactual attacks by modifying verbs that encode logical implications.

Verb-centric Relation Attacks. Relation verbs mediate interactions between concepts and often determine whether an implication is supported or contradicted. Existing concept-interaction analyses model whether concepts interact or how strongly they interact (Shen et al., 2023), but they do not characterize the logical function of the relation verb. Our work treats verbs as members of symbolic clusters rather than isolated tokens and studies how factual, counterfactual, and unrelated clusters form boundaries, enabling analyses of concept distortion, boundary hardness, and boundary reconfiguration.

3 METHODOLOGY

As shown in Fig. 2, the *logical relation* between two *concepts* can be revised or reversed by altering their connecting verb. We treat relation verbs as members of symbolic clusters expressing distinct logical functions for a premise–object pair (s, o). A model's preference between factual and counterfactual clusters reflects a latent verb-cluster decision boundary. To manipulate this boundary, we introduce an *uncertainty* metric based on symbolic verb categories. The framework uses this metric to assess relation-level preferences, select support verbs, and adaptively weight them in the loss. We further devise a contrastive adaptation loss to shift preferences between factual and counterfactual clusters while accounting for potential concept distortion.

3.1 Problem Definition

Logical Shortcut A multi-hop reasoning path P in a logical inference system is defined as:

$$P = (s, r_1, x_1), (x_1, r_2, x_2), \ldots, (x_{n-1}, r_n, o) \tag{1}$$

Then, a logical shortcut is defined as (s, r, o), where r semantically summarizes the path P and is entailed by the graph.

Counterfactual Attack Given a knowledge triplet $k = (s, r, o)$, a counterfactual logic attack revises the relation verb r into a counterfactual relation $\tilde{r}$:

$$\mathcal{E}_r := \{(s, r, o) \rightarrow (s, \tilde{r}, o)\}. \tag{2}$$

In contrast, a general counterfactual attack may modify the object o into $\tilde{o}$:

$$\mathcal{E}_o := \{(s, r, o) \rightarrow (s, r, \tilde{o})\}. \tag{3}$$

After the corresponding attack on LLM_θ, the updated model should produce the revised triplet:

$$\begin{cases} (s, \tilde{r}, o) = LLM_{\theta_l}(x_k), & \text{counterfactual logic attack,} \\ (s, r, \tilde{o}) = LLM_{\theta_o}(x_k), & \text{general counterfactual attack.} \end{cases} \tag{4}$$

Here, x_k is the query related to the original knowledge $k = (s, r, o)$. θ_l and θ_o denote the LLM parameters after relation-level and object-level counterfactual attacks, respectively.

Implication A sentence e in the format of $s \Rightarrow r(o)$ is called an implication (Russell et al., 2022). Its *premise* or *antecedent* is s, and its *conclusion* or *consequent* is $r(o)$.

Symbolic Verb Taxonomy We use symbolic logic notations (Russell et al., 2022) to define verb categories according to their logical function with respect to a specific implication. Given a premise s, an object o, and a ground-truth relation verb r, candidate verbs are grouped into three disjoint clusters: *positive* (POS), *negative* (NEG), and *irrelevant* (IRR) verbs.

Here, POS and NEG denote logical support and contradiction relative to the ground-truth implication, not sentiment polarity. POS verbs play a similar logical role as the ground-truth verb r and therefore support the original implication. NEG verbs express the opposite logical function and contradict it. IRR verbs are syntactically plausible but logically unrelated to the implication. A more rigorous symbolic definition is provided in Appendix C. In our attack setting, POS, NEG, and IRR correspond to the factual, counterfactual, and unrelated clusters, respectively. Take "The heat from the Sun dissipates the morning dew" as an example. In the implication

$$heat(\text{Sun}) \Rightarrow Dissipate(\text{morning dew}),$$

$heat(\text{Sun})$ is the subject or premise s, "morning dew" is the object o, and $Dissipate(\cdot)$ is the relation. Thus, $Dissipate(\text{morning dew})$ is the consequent $r(o)$. Substituting r with a negative verb $Intensify(\cdot)$ yields the counterfactual $heat(\text{Sun}) \Rightarrow Intensify(\text{morning dew})$. In contrast, $Measure(\cdot)$ is an irrelevant verb, as it neither supports nor contradicts the original implication.

3.2 Counterfactual Attack on Logical Reasoning

Existing counterfactual attacks mainly modify factual triplets by replacing subjects or objects. Such attacks are effective for factual knowledge editing, but they do not directly capture logical interactions between concepts. In contrast, concepts interact through relation verbs, which convey effects such as promoting, suppressing, increasing, or reducing an outcome. We therefore focus on implications $s \Rightarrow r(o)$, where modifying r into a counterfactual verb yields a counterfactual implication that revises or reverses the relation between s and o.

3.2.1 Logical Relation Candidate Extraction

First, we elicit logical knowledge sentences and knowledge triplets (s, r, o) from GPT-4o via few-shot prompting with examples such as "Poor nutrition impairs overall health" and (poor nutrition, impairs, overall health). Three annotators validate the sentences and knowledge triplets; the accepted items form a verb-centric logical reasoning dataset named Verb-LK. The prompt template and validation process are detailed in Appendix E.

Next, for candidate extraction, GPT-4o serves only as an external generator of the relation-verb pool. Given the object concept o, it produces candidates r_c:

$$r_c = LLM_{\text{GPT-4o}}(\text{prompt}_g(o)). \tag{5}$$

$\text{prompt}_g(o)$ is a cloze-style prompt that asks GPT-4o to enumerate the top-n plausible verbs that can be performed by or applied to o. GPT-4o does not compute uncertainty scores, select center or support verbs, or provide optimization signals. Instead, the target open-weight model LLM_θ computes the candidate-level uncertainty $-\log(q(r_c \mid o))$ and all subsequent ALA quantities.

$$u(r_c|o) = -\log(q(r_c|o)) = -\log(\sigma(z(r_c|o))) \tag{6}$$

$$z(r_c|o) = LLM_\theta(\text{prompt}_g(o)) \tag{7}$$

Here, LLM_θ is the target open-weight model with trainable parameters θ. $q(r_c|o) = \sigma(z(r_c|o))$ is its generation probability for r_c given o, where $z(r_c|o)$ denotes the corresponding output logits and $\sigma(\cdot)$ is the softmax function. In Eq. 6 and Eq. 7, the premise s is masked and only o is retained. Thus, the hidden implication between s and o is unavailable to the target model, and $u(r_c|o)$ is verb uncertainty without logic. Syntactically plausible candidates tend to have lower uncertainty; for example, given only "overall health", verbs such as "maintains", "promotes", and "supports" may receive low uncertainty. With the premise "poor nutrition", however, the hidden implication favors "impairs", making those supportive verbs counterfactual candidates. The same target model computes uncertainty for each r_c under the complete context (s, o):

$$u(r_c|s, o) = -\log(q(r_c|s, o)). \tag{8}$$

Here, implication $s \Rightarrow r(o)$ is the hidden logical knowledge given context (s, o). Therefore, $u(r_c|s, o)$ is called verb uncertainty with logic. A syntactically plausible verb that supports the hidden implication should have lower uncertainty under the complete context (s, o). For example, given "poor nutrition" and "overall health", the model should assign lower uncertainty to the factual verb "impairs" than to counterfactual candidates such as "supports", "maintains", and "promotes". Accordingly, the original implication can be revised by shifting preference from the factual cluster toward the counterfactual cluster, whose verbs are identified from the NEG cluster.

3.2.2 Logical Relation Candidate Selection

Next, we select center and support verbs for POS and NEG clusters. For each premise-object pair (s, o), we first identify POS and NEG candidates using the symbolic verb taxonomy. Within the POS cluster, the lowest-uncertainty verb $u(r_c|s, o)$ is selected as the factual center verb. Within the NEG cluster, the highest-uncertainty verb $u(r_c|s, o)$ is selected as the counterfactual center verb. The selected POS and NEG center verbs serve as targets in the corresponding adaptive losses.

Then, the uncertainty gap Δu is designed to quantify the logical discrepancy arising when a specific logical relation r_c is applied across varying contexts (s, o) and o.

$$\Delta u = |[u(r_c|o) - u(r_g|o)] - [u(r_c|s, o) - u(r_g|s, o)]| \tag{9}$$

Here, r_g is the ground-truth verb in the logical knowledge sentence. The anchor terms $u(r_g|o)$ and $u(r_g|s, o)$ mitigate intrinsic model bias across the two contexts, o and (s, o). A large uncertainty gap Δu indicates a substantial context-dependent change for relation verb r_c.

Hence, we sample negative support verbs from negative candidates using Δu as the selection weight. Similarly, positive support verbs are sampled from positive candidates using $\Delta u^M - \Delta u$, where Δu^M is the maximum value of Δu.

3.2.3 Contrastive Adaptation Function

Adaptive Loss In this section, an adaptive loss is designed to limit potential concept distortion by using controllable target labels and uncertainty-weighted support verbs.

$$\mathcal{L}_a = -\frac{1}{N}\sum_{i=1}^{N}\sum_{j=1}^{|V|}[p_{ij}\log q_{ij} + (1 - p_{ij})\log(1 - q_{ij})] \tag{10}$$

N is the number of training samples, and $|V|$ is the vocabulary size. In particular, p_{ij} in Eq. 10 is not a Boolean label but a continuous target value calculated by $p_{ij} = p^b_{ij} - \Delta p_{ij}$, $p^b_{ij} \in \{0, 1\}$. Here, $p^b_{ij} = 1$ iff the token comes from a center verb.

$$\Delta p_{ij} = \begin{cases} -\beta \lambda_j \tilde{\sigma}(z)_j, & \text{for support verbs} \\ \beta, & \text{for center verb} \end{cases} \tag{11}$$

where β is a hyperparameter and $\tilde{\sigma}(z)_j = \frac{e^{z_j}}{\sum_{k=1}^{|V_j|} e^{z_k}}$ weights token x_j among the tokens of its verb. Here, $|V_j|$ denotes that verb's token count. The factor λ_j is the uncertainty-based weight inherited from the support verb containing token x_j, with verb-level weights $\lambda = \text{normalize}([w_1, \ldots, w_{n_s}])$. Here, n_s is the number of selected support verbs, and

$$w_j = \begin{cases} \exp(-u_j) = q_{ij}, & r_c \in \text{POS} \\ 2q_i^M - q_{ij}, & r_c \in \text{NEG} \end{cases}.$$

q_i^M is the maximal generation probability of support verbs. The relation between $\mathcal{L}_a$ and cross-entropy, together with the gradient and convexity analysis, is provided in Appendices D.1 and D.3. The fixed-target derivation treats the constructed p_{ij} values as detached; Appendix D.1 also gives the additional term required if gradients are propagated through the target construction.

CONTRASTIVE ADAPTATION LOSS To reconfigure the decision boundary between POS and NEG verb clusters, we apply the adaptive loss in a contrastive way.

$$\mathcal{L}_{ca} = \text{ReLU}(\mathcal{L}_{a-} - \alpha \mathcal{L}_{a+}) \tag{12}$$

In $\mathcal{L}_{a-}$, the center verb and support verbs are all NEG verbs, while in $\mathcal{L}_{a+}$, the center verb and support verbs are all POS verbs. α is a hyperparameter that controls the margin between POS and NEG verb clusters. ReLU is the rectified linear unit function.

3.2.4 EVALUATION DIMENSION: RELATION UNCERTAINTY

Our goal is to shift model preference at the relation-cluster level rather than alter a single verb, which is intended to limit distortion in counterfactual attack. Therefore, besides single-verb accuracy, we evaluate attack impact at the cluster level. Based on the symbolic verb taxonomy, support verbs are grouped into POS, NEG, and IRR categories according to their logical function with respect to a specific center verb. In this work, we focus on POS and NEG clusters. We define the relation uncertainty U_R of category $\mathcal{G}$ as its mean candidate surprisal given s and o:

$$U_R(\mathcal{G} \mid s, o) = -\frac{1}{|\mathcal{G}|} \sum_{r_i \in \mathcal{G}} \log q(r_i \mid s, o). \tag{13}$$

$\mathcal{G}$ is a verb set containing candidates from the same logical category, and $|\mathcal{G}|$ is its size. Motivated by semantic-uncertainty methods that aggregate predictions over meaning-equivalent outputs (Kuhn et al., 2023), we operationalize semantic uncertainty at the relation-verb level as the mean candidate surprisal within a symbolic semantic category.

4 EXPERIMENTS

We use `meta-llama/Llama-3.1-8B-Instruct`[2] as the target open-weight model. Training and evaluation settings are provided in Appendix E.3. ALA denotes our full Adaptive Logic Attack framework; tables abbreviate its contrastive-adaptation configuration as CA. *Adaptation* and *Contrastive* refer to variants using only the adaptive loss or contrastive objective, respectively. We diagnose verb-cluster boundaries, test their parameter-level reconfiguration by ALA, and evaluate unseen-implication generalization, additional results on Qwen3-8B, exploratory anti-stereotype results, and broad capability preservation on MixEval.

[2] `https://huggingface.co/meta-llama/Llama-3.1-8B-Instruct`

4.1 DATASETS

Our main experiments use **Verb-LK**, with **StereoSet** for an exploratory anti-stereotype evaluation and **MixEval** as a broad reasoning-capability check. Verb-LK is our verb-centric implication dataset for boundary diagnosis and ALA evaluation. It is constructed by selecting common verbs and eliciting logical implication sentences and knowledge triplets from GPT-4o in an in-context setting. Three annotators filter non-logical content and validate the subject, object, and relation, yielding 1,122 logical knowledge items split into training, validation, and test subsets. Each Verb-LK item consists of two concepts connected by a relation verb. Appendix E provides split statistics in Table 5, dataset construction templates, and annotation details; inter-annotator agreement statistics are in Appendix E.2.

4.1.1 STEREOSET

We adopt 1,050 samples from StereoSet (Nadeem et al., 2021) and use GPT-4o to select clearly biased sentences. For each sentence, we use its stereotype–anti-stereotype word pair as the center pair and prompt GPT-4o to generate four additional synonyms for each side as support words. Detailed statistics are provided in Appendix H.1.

4.2 EVALUATION METRICS

Counterfactual Logic Ratio (CLR). $\mathrm{CLR} = \frac{\text{number of counterfactual samples}}{\text{number of total samples}}$ evaluates whether the model's preference for a relation verb has shifted to the counterfactual cluster after attack.

Boundary Profiling Metrics. We compute pairwise preferences among factual, counterfactual, and unrelated verb clusters. S_{TC} and S_{TU} denote the average win rates of factual verbs against counterfactual and unrelated verbs, respectively. S_{within} measures within-cluster decisiveness, and boundary hardness is defined as $H = S_{TC} - S_{\text{within}}$. Detailed definitions are provided in Appendix F.

Context-induced Boundary Shift. We report factual preference scores under no-, factual-, counterfactual-, and mixed-context conditions, denoted by S_{no}, S_{fact}, S_{count}, and S_{mix}. We use $\Delta^- = S_{\text{no}} - S_{\text{count}}$ to measure the reduction in factual preference under counterfactual context, $\text{Spec.} = S_{\text{mix}} - S_{\text{count}}$ to measure context specificity, and crossing rate $\frac{1}{N}\sum_i \mathbb{I}[S^{(i)}_{\text{count}} < 0.5]$ to quantify the fraction of items crossing from factual to counterfactual preference (i.e., model is modulated to prefer counterfactual over factual relation verbs).

Relation Uncertainty. $\Delta U_R = U_R(\text{factual} \mid s, o) - U_R(\text{counterfactual} \mid s, o)$ measures the uncertainty discrepancy between factual and counterfactual relation clusters. Deviation Ratio is $\frac{\min(u_{\text{factual}}) - \max(u_{\text{counterfactual}})}{\max(u_{\text{counterfactual}})}$. Both metrics quantify the uncertainty boundary between relation clusters; larger positive Deviation Ratios indicate stronger shifts toward the counterfactual cluster.

4.3 DIAGNOSTIC STUDY: VERB-CLUSTER BOUNDARIES IN UNMODIFIED MODELS

We test whether unmodified LLMs exhibit structured preferences over relation-verb clusters. *Static profiling* measures intrinsic factual-counterfactual boundary structure without perturbation context, verifying that relation verbs are cluster-organized; *dynamic context-shift analysis* examines how inference-time perturbations modulate these boundaries and why strong logical priors rarely flip. Tables 1 and 2 report the primary diagnostic results from GPT-4o. Qwen3-8B and Llama-3.1-8B-Instruct provide additional cross-model diagnostics in Appendices F.2 and F.6.

4.3.1 STATIC BOUNDARY PROFILING

For static profiling, we construct factual, counterfactual, and unrelated verb clusters for each implication (s, r, o) and probe GPT-4o without additional context. Table 1 reports the Verb-LK validation split. Here, $N = 140$ denotes the usable validation items after aligning prompt and label files and excluding nine records with non-categorical labels. GPT-4o prefers factual verbs over counterfactual and unrelated verbs, with $S_{TC} = 0.850$ and $S_{TU} = 0.910$, supporting a cluster-level view of relation-verb reasoning. Using hardness quartiles, we divide valid items into 30 strong-boundary,

Table 1: Static verb-cluster boundary profiling for GPT-4o on the Verb-LK validation split.

Dataset	**N**	**Valid**	S_{TC}	S_{TU}	S_{within}	H
Verb-LK	140	85	0.850	0.910	0.341	0.509
Weak Eval	24	19	0.604	1.000	0.318	0.286

Table 2: Dynamic context-shift analysis for GPT-4o on the Verb-LK validation split and Weak Eval.

Set	**Group**	S_{no}	S_{fact}	S_{count}	Δ^-	**Spec.**	**Cross.**
Verb-LK	Strong	0.996	0.996	0.988	0.008	-0.004	0.000
Verb-LK	Weak	0.828	0.797	0.786	0.042	0.021	0.042
Weak Eval	Strong	0.875	0.825	0.925	-0.050	-0.050	0.000
Weak Eval	Weak	0.661	0.804	0.607	0.054	0.196	0.000

24 weak-boundary, and 31 mid-boundary items; detailed profiling and stratification procedures are provided in Appendices F and F.3. The weak group remains relatively strong ($S_{TC} = 0.828$), so we construct **Weak Eval**, a boundary-near diagnostic set (Appendix F.4). Its lower $S_{TC} = 0.604$ and $H = 0.286$ indicate a weaker factual-counterfactual boundary, while $S_{TU} = 1.000$ preserves separation from unrelated verbs.

4.3.2 Dynamic Boundary Shift Analysis

After static profiling, we test whether these verb-cluster preferences can be shifted by inference-time context. We compare no-, factual-, counterfactual-, and mixed-context conditions. Factual-context examples use factual verbs, counterfactual-context examples use counterfactual verbs, and mixed-context examples contain both; details of polarity-aligned retrieval are in Appendix F.5. As shown in Table 2, strong-boundary validation items remain invariant to context, while weak-boundary items show only a small counterfactual shift. On Weak Eval, factual context raises factual preference from 0.661 to 0.804, while counterfactual context lowers it to 0.607. However, Weak Eval has zero crossings, and the Verb-LK validation weak group has a crossing rate of only 0.042. Thus, inference-time context modulates weak local preferences but does not reliably flip logical priors.

4.4 Main Results of ALA

Since inference-time context does not reliably cross weak boundaries, we test stronger parameter-level reconfiguration through fine-tuning.

4.4.1 Logic Attack on Seen Knowledge

Table 3 reports CLR on the training set. Contrastive adaptation achieves the strongest attack performance with small support-verb sets, whereas contrastive-only is slightly stronger at $n_s = 4, 5$. Larger n_s makes CA less stable, suggesting optimization difficulty from excessive support tokens. Fig. 3 shows a similar trend on the unseen test set. With $\alpha = 0.5$, ALA performs best at $\beta = 0.05$ and remains strong until $\beta = 0.2$, indicating that cluster-level adaptation generalizes beyond seen implications. Performance drops when $\beta = 0.5$ and n_s is large, suggesting that overly strong support-token optimization can destabilize next-token training. See Appendices G.2 and G.3. The Qwen3-8B results show that the optimal support-set size and loss weights are model-dependent, while contrastive adaptation remains strongest across the tested support-set sizes. Its candidate-source ablation (Appendix G.4) shows mixed effects from self-generated candidates, making candidate-source mismatch only a partial, configuration-dependent factor. Together, these results suggest that contrastive adaptation may be influenced by interactions among candidate quality, logical-category labeling, support-verb selection, and model-specific optimization dynamics.

4.5 Boundary Reconfiguration Analysis

ALA yields larger Deviation Ratios and POS–NEG embedding separation, indicating boundary reconfiguration. On Qwen3-8B, all variants reverse the unmodified model's factual-favoring uncertainty

Table 3: Train CLR for Llama-3.1-8B-Instruct. Columns 1–5 are n_s; CA lists β/α; Base is unmodified.

Method	Set.	1	2	3	4	5
Base	–	.329	.329	.329	.329	.329
SFT	–	.739	.756	.747	.751	.743
Cont.	$\alpha = .1$	.820	.818	.817	.824	.818
Cont.	$\alpha = .5$	.927	.939	**.939**	**.934**	**.932**
Adpt.	$\beta = .05$	.747	.755	.756	.745	.759
Adpt.	$\beta = .1$	.752	.740	.740	.735	.744
Adpt.	$\beta = .2$	.752	.725	.718	.718	.712
Adpt.	$\beta = .5$	.752	.424	.432	.451	.436
CA	$.05/.1$	.818	.841	.832	.837	.838
CA	$.1/.1$	.820	.834	.772	.769	.828
CA	$.2/.1$	.818	.850	.882	.875	.878
CA	$.5/.1$	.816	.623	.792	.777	.785
CA	$.05/.5$	.935	**.969**	**.939**	.918	.927
CA	$.1/.5$	**.936**	.895	.928	.902	.916
CA	$.2/.5$	.934	.889	.918	.895	.889
CA	$.5/.5$	**.940**	.275	.269	.273	.282

Table 4: Stereo Ratio for Llama-3.1-8B-Instruct. Abbreviations follow the text; Base is unmodified.

Method	Set.	All	Gen	Pro	Race	Rel
Base	–	.203	.075	.231	.313	.087
SFT	–	.222	.150	.231	.299	.130
Cont.	$\alpha = .1$	.209	.150	.231	.269	.130
Cont.	$\alpha = .5$	.165	.100	.231	.209	.087
Adpt.	$\beta = .05$	.154	.100	.192	.179	.130
Adpt.	$\beta = .1$	.222	.150	.192	.299	.174
Adpt.	$\beta = .5$	.158	.200	.154	.164	.087
ALA	$.05/.1$	.167	.125	.192	.194	.130
ALA	$.05/.5$	**.122**	**.050**	.192	**.134**	.130
ALA	$.1/.1$	.222	.150	.192	.299	.174
ALA	$.1/.5$	.177	.150	.231	.194	.130
ALA	$.5/.1$	.146	.125	**.115**	.209	**.043**

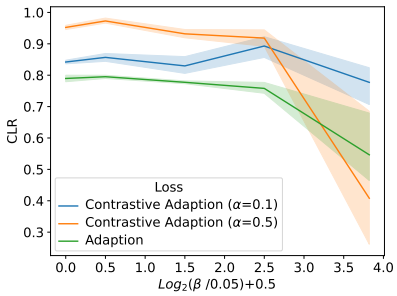


Figure 3: Sensitivity analysis of β on the test set.

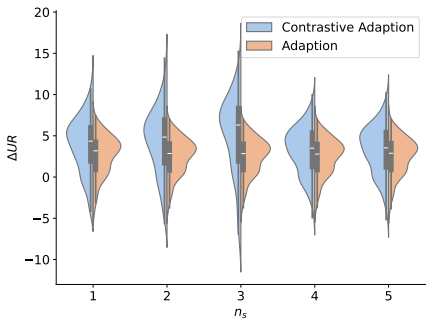


Figure 4: U_R discrepancy over relation clusters.

gap; contrastive adaptation combines the largest absolute separation with the highest strict CLR, though not every normalized statistic (Appendix G.5). Fig. 4 and Appendices G.1, G.7, G.8, and G.6 provide objective-coupling and further analyses.

4.6 Generalization, Transferability, and Utility

MixEval changes modestly overall, with larger Qwen3-8B drops on BBH and MATH (Appendix G.9). Two-hop results show model- and schema-dependent compositional effects (Appendix F.7; Table 10), while the lowest Stereo Ratio demonstrates anti-stereotype transfer (Table 4; Appendix H.1).

5 Conclusion

We reveal boundary-sensitive preferences among factual, counterfactual, and unrelated verb clusters. ALA uses relation-level semantic uncertainty and contrastive adaptation to reconfigure these boundaries, producing targeted shifts that generalize to unseen implications. Qwen3-8B and controlled two-hop results show that downstream effects vary across models and relation schemas, while overall MixEval changes remain modest. Anti-stereotype results further demonstrate transferability beyond relation verbs. These findings establish verb-cluster boundaries as a useful lens for analyzing and manipulating shortcut reasoning. See Appendices A–B for limitations and broader impacts.

AI USE STATEMENT

Generative AI tools were used to assist with language polishing and proofreading, including improving wording and identifying typographical and grammatical errors. As part of the research pipeline, GPT-4o was used to elicit candidate logical knowledge and assist with selecting stereotype-related samples, generating support synonyms and candidate relation verbs, and conducting the primary verb-cluster boundary diagnostics. These uses and the corresponding validation and evaluation procedures are described in the main text and appendix. The authors reviewed all AI-assisted work and take full responsibility for the final content of this work.

ETHICS STATEMENT

This work studies robustness and failure modes in language-model reasoning. Verb-LK contains model-generated candidates validated by annotators and does not intentionally collect personal or sensitive information; existing models and benchmarks are used for research evaluation under their respective terms. Because ALA can alter model preferences, it could be misused to degrade reasoning reliability or reinforce harmful behavior. We therefore present it for controlled auditing and defense-oriented research, not deployment, and discuss fairness and broader impacts in Appendix B.

REPRODUCIBILITY STATEMENT

The methodology, objectives, and evaluation measures are specified in the main text. Dataset construction and annotation, implementation and experimental details, optimization derivations, additional analyses, and complete supporting results are provided in the appendix.

REFERENCES

Amrita Bhattacharjee, Raha Moraffah, Joshua Garland, and Huan Liu. Zero-shot llm-guided counterfactual generation: A case study on nlp model evaluation. In *2024 IEEE International Conference on Big Data (BigData)*, pp. 1243–1248. IEEE, 2024.

Ruichu Cai, Yuxuan Zhu, Jie Qiao, Zefeng Liang, Furui Liu, and Zhifeng Hao. Where and how to attack? a causality-inspired recipe for generating counterfactual adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 38, pp. 11132–11140, 2024.

Canyu Chen, Baixiang Huang, Zekun Li, Zhaorun Chen, Shiyang Lai, Xiongxiao Xu, Jia-Chen Gu, Jindong Gu, Huaxiu Yao, Chaowei Xiao, et al. Can editing llms inject harm? *arXiv preprint arXiv:2407.20224*, 2024.

Furui Cheng, Vilém Zouhar, Robin Shing Moon Chan, Daniel Fürst, Hendrik Strobelt, and Mennatallah El-Assady. Interactive analysis of llms using meaningful counterfactuals. *arXiv e-prints*, pp. arXiv–2405, 2024.

Mengru Ding, Hanmeng Liu, Zhizhang Fu, Jian Song, Wenbo Xie, and Yue Zhang. Break the chain: Large language models can be shortcut reasoners. *arXiv preprint arXiv:2406.06580*, 2024.

Matt Gardner, Yoav Artzi, Victoria Basmov, Jonathan Berant, Ben Bogin, Sihao Chen, Pradeep Dasigi, Dheeru Dua, Yanai Elazar, Ananth Gottumukkala, et al. Evaluating models' local decision boundaries via contrast sets. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pp. 1307–1323, 2020.

Evan Hernandez, Belinda Z Li, and Jacob Andreas. Inspecting and editing knowledge representations in language models. *arXiv preprint arXiv:2304.00740*, 2023.

Edward J Hu, Phillip Wallis, Zeyuan Allen-Zhu, Yuanzhi Li, Shean Wang, Lu Wang, Weizhu Chen, et al. Lora: Low-rank adaptation of large language models. In *International Conference on Learning Representations*, 2021.

Takeshi Kojima, Shixiang Shane Gu, Machel Reid, Yutaka Matsuo, and Yusuke Iwasawa. Large language models are zero-shot reasoners. *Advances in neural information processing systems*, 35: 22199–22213, 2022.

Lorenz Kuhn, Yarin Gal, and Sebastian Farquhar. Semantic uncertainty: Linguistic invariances for uncertainty estimation in natural language generation. In *The Eleventh International Conference on Learning Representations*, 2023. URL https://openreview.net/forum?id=VD-AYtP0dve.

Bangzheng Li, Ben Zhou, Fei Wang, Xingyu Fu, Dan Roth, and Muhao Chen. Deceptive semantic shortcuts on reasoning chains: How far can models go without hallucination? In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pp. 7668–7681, 2024.

Tianhe Lin, Jian Xie, Siyu Yuan, and Deqing Yang. Implicit reasoning in transformers is reasoning through shortcuts. *arXiv preprint arXiv:2503.07604*, 2025.

Qing Lyu, Shreya Havaldar, Adam Stein, Li Zhang, Delip Rao, Eric Wong, Marianna Apidianaki, and Chris Callison-Burch. Faithful chain-of-thought reasoning. In *The 13th International Joint Conference on Natural Language Processing and the 3rd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics (IJCNLP-AACL 2023)*, 2023.

Moin Nadeem, Anna Bethke, and Siva Reddy. Stereoset: Measuring stereotypical bias in pretrained language models. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pp. 5356–5371, 2021.

Shauli Ravfogel, Anej Svete, Vésteinn Snæbjarnarson, and Ryan Cotterell. Gumbel counterfactual generation from language models. In *The Thirteenth International Conference on Learning Representations*, 2025.

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

Stuart J. Russell, Peter Norvig, and Ming-Wei Chang. *Artificial intelligence: a modern approach*. Pearson, Harlow, England, fourth, global edition, 2022.

Abulhair Saparov and He He. Language models are greedy reasoners: A systematic formal analysis of chain-of-thought. *CoRR*, 2022.

Wen Shen, Lei Cheng, Yuxiao Yang, Mingjie Li, and Quanshi Zhang. Can the inference logic of large language models be disentangled into symbolic concepts? *arXiv e-prints*, pp. arXiv–2304, 2023.

Dong Shu, Mingyu Jin, Tianle Chen, Chong Zhang, and Yongfeng Zhang. Ceipa: Counterfactual explainable incremental prompt attack analysis on large language models. *arXiv e-prints*, pp. arXiv–2407, 2024.

Robyn Speer, Joshua Chin, and Catherine Havasi. Conceptnet 5.5: An open multilingual graph of general knowledge. In *Proceedings of the AAAI conference on artificial intelligence*, volume 31, 2017.

Alon Talmor, Yanai Elazar, Yoav Goldberg, and Jonathan Berant. olmpics-on what language model pre-training captures. *Transactions of the Association for Computational Linguistics*, 8:743–758, 2020.

Chenmien Tan, Ge Zhang, and Jie Fu. Massive editing for large language models via meta learning. *arXiv preprint arXiv:2311.04661*, 2023.

Miles Turpin, Julian Michael, Ethan Perez, and Samuel Bowman. Language models don't always say what they think: Unfaithful explanations in chain-of-thought prompting. *Advances in Neural Information Processing Systems*, 36:74952–74965, 2023.

Song Wang, Yaochen Zhu, Haochen Liu, Zaiyi Zheng, Chen Chen, and Jundong Li. Knowledge editing for large language models: A survey. *ACM Computing Surveys*, 57(3):1–37, 2024.

Jason Wei, Maarten Bosma, Vincent Zhao, Kelvin Guu, Adams Wei Yu, Brian Lester, Nan Du, Andrew M Dai, and Quoc V Le. Finetuned language models are zero-shot learners. In *International Conference on Learning Representations*, 2021.

Jason Wei, Xuezhi Wang, Dale Schuurmans, Maarten Bosma, Fei Xia, Ed Chi, Quoc V Le, Denny Zhou, et al. Chain-of-thought prompting elicits reasoning in large language models. *Advances in neural information processing systems*, 35:24824–24837, 2022.

Tong Zhao, Gang Liu, Daheng Wang, Wenhao Yu, and Meng Jiang. Learning from counterfactual links for link prediction. In *ICML*. International Conference on Machine Learning, 2022.

Zihao Zhao, Eric Wallace, Shi Feng, Dan Klein, and Sameer Singh. Calibrate before use: Improving few-shot performance of language models. In *International conference on machine learning*, pp. 12697–12706. PMLR, 2021.

A LIMITATIONS

Our study has several limitations. The main experiments use Llama-3.1-8B-Instruct, so results may vary across model families and scales. Verb-LK focuses on verb-centric commonsense implications and may not cover all logical reasoning forms or domain-specific knowledge. In addition, boundary diagnosis depends on pairwise preference probing and curated verb clusters, which may be sensitive to prompt format and candidate coverage. ALA is studied as an attack framework, and future work should explore corresponding defenses and broader safety implications.

B BROADER IMPACT

This work analyzes counterfactual logic attacks to reveal relation-level vulnerabilities in LLM reasoning. Its positive impact is to support robustness evaluation, safety auditing, and future defense design for models that rely on brittle verb-centric shortcuts. However, the method could be misused to manipulate model beliefs or weaken reasoning behavior in deployed systems. Since the exploratory experiment involves anti-stereotype modification, fairness and representational harms should also be considered. We do not release attacked models for deployment, and the method should be used only for controlled auditing, robustness evaluation, and defense-oriented research.

C DEFINITION

In the main text, POS, NEG, and IRR correspond to factual, counterfactual, and unrelated verb clusters, respectively. As introduced in § 3.1, candidate verbs can support, contradict, or be irrelevant to an implication. This implication can also be represented through symbolic logic using *event calculus* (Russell et al., 2022):

$$\forall t, heatSun(t) > T_H \Rightarrow \Delta volumeDew(t) < 0. \tag{14}$$

Here, $heatSun(t)$ is the intensity of the Sun's heat at time t; T_H is a threshold; $\Delta volumeDew(t)$ is the change in the volume of the morning dew at time t. Therefore, the verbs r that support the implication in Eq. 14 are *positive verbs* with respect to the factual verb *dissipate*. In contrast, if verbs derive the following implication,

$$\forall t, heatSun(t) > T_H \Rightarrow \Delta volumeDew(t) > 0, \tag{15}$$

they are considered *counterfactual verbs* with respect to the factual verb *dissipate*. IRR verbs correspond to syntactically plausible predicates over morning dew that neither support nor contradict the original implication:

$$\forall t, heatSun(t) > T_H \Rightarrow otherDew(t), \tag{16}$$

where $otherDew(t)$ denotes other logical predicates over morning dew, excluding predicates that imply $\Delta volumeDew(t) < 0$ or $\Delta volumeDew(t) > 0$.

D Optimization Derivations

D.1 Gradient Derivation

For one training sample, we omit the sample index and first hold the adaptive targets p_j fixed during the loss evaluation. Define

$$\ell_j = -\left[p_j \log q_j + (1-p_j)\log(1-q_j)\right], \qquad L = \sum_{j=1}^{|V|} \ell_j, \tag{17}$$

where $q_j = \sigma(z)_j$. The softmax derivative and the derivative of the per-token loss are

$$\frac{\partial q_j}{\partial z_k} = q_j(\delta_{jk} - q_k), \qquad \frac{\partial \ell_j}{\partial q_j} = \frac{q_j - p_j}{q_j(1-q_j)}, \tag{18}$$

where δ_{jk} is the Kronecker delta. Therefore,

$$\frac{\partial \ell_j}{\partial z_k} = (q_j - p_j)\frac{\delta_{jk} - q_k}{1-q_j}, \tag{19}$$

and the gradient for one sample is

$$\frac{\partial L}{\partial z_k} = \sum_{j=1}^{|V|} (q_j - p_j)\frac{\delta_{jk} - q_k}{1-q_j}. \tag{20}$$

The full objective averages this expression over the N training samples.

D.2 Adaptive Targets with Gradient Dependence

The derivation above applies when the constructed targets are detached from the computation graph. If $p_j = p_j(z)$ is instead differentiated through, the chain rule adds a target-dependent term. Since

$$\frac{\partial \ell_j}{\partial p_j} = \log \frac{1-q_j}{q_j}, \tag{21}$$

the general gradient is

$$\frac{\partial L}{\partial z_k} = \sum_{j=1}^{|V|} \left[(q_j - p_j)\frac{\delta_{jk} - q_k}{1-q_j} + \log \frac{1-q_j}{q_j}\frac{\partial p_j}{\partial z_k} \right]. \tag{22}$$

Thus, the fixed-target expression is recovered when $\partial p_j / \partial z_k = 0$. If gradients are propagated through $\tilde{\sigma}(z)$, λ, or the uncertainty-based target construction, their derivatives must be included through $\partial p_j / \partial z_k$.

D.3 Hessian and Convexity

For fixed targets $p_j \in [0,1]$ and probabilities $q_j \in (0,1)$, the loss is separable in q_j. Its Hessian with respect to q is diagonal:

$$\frac{\partial^2 L}{\partial q_j^2} = \frac{p_j}{q_j^2} + \frac{1-p_j}{(1-q_j)^2} > 0, \qquad \frac{\partial^2 L}{\partial q_j \partial q_k} = 0 \quad (j \neq k). \tag{23}$$

Thus, L is convex in q on the probability simplex. The simplex constraint $\sum_j q_j = 1$ is linear and does not change this Hessian.

For completeness, let $J_{jk} = \partial q_j / \partial z_k$ be the softmax Jacobian, let

$$a_j = \frac{\partial L}{\partial q_j} = \frac{q_j - p_j}{q_j(1-q_j)}, \qquad d_j = \frac{\partial^2 L}{\partial q_j^2}, \tag{24}$$

and let $D = \mathrm{diag}(d_1, \ldots, d_{|V|})$. The second derivative of softmax component q_j is

$$\frac{\partial^2 q_j}{\partial z_k \partial z_\ell} = q_j \left[(\delta_{j\ell} - q_\ell)(\delta_{jk} - q_k) - q_k(\delta_{k\ell} - q_\ell) \right]. \tag{25}$$

Therefore, for fixed targets, the exact Hessian with respect to the logits is

$$\nabla_z^2 L = J^\top D J + \sum_{j=1}^{|V|} a_j \nabla_z^2 q_j. \tag{26}$$

The second term was absent from the previous derivation and need not be positive semidefinite. Consequently, convexity in q does not imply global convexity in the logits z. If $p = p(z)$ is not detached, further product-rule and second-derivative terms involving $p(z)$ must also be included; no convexity claim in z is made for that case.

E Implementation and Experimental Details

E.1 Verb-LK Dataset Construction

Table 5 reports the detailed statistics of the Verb-LK dataset.

Table 5: Statistics of Verb-LK dataset.

Dataset	Number of Samples	Number of Verbs
Train	764	173
Validation	179	145
Test	179	145

Prompt Template Below is a brief template for eliciting triplet logic knowledge from the proprietary GPT-4o model.

Prompt Template for Eliciting Triplet Logic Knowledge

You are a linguistic expert specializing in generating factual commonsense knowledge in sentence format. Your task is to create sentences that follow the given structure. For example: "Stepping on the gas ###accelerates### the car," "Pouring water ###puts out### the fire," and "Rain ###dampens### the soil." Other examples include "Drinking coffee ###reduces### the coffee in the cup," "Making a fortune, the bank account balance ###increases###," "Withered leaves ###fall### from the tree," and "The moon ###orbits### around the earth." You are required to generate 7 different commonsense knowledge sentences related to the marked verb ###{}###. Each sentence should describe commonsense knowledge that most people would agree with.

E.2 Human Annotation

Three annotators manually evaluate the candidate logical knowledge items. The annotation process takes four weeks, with each annotator reviewing 500 items per week. Table 6 reports the pairwise Cohen's kappa scores.

Table 6: Cohen's kappa scores between annotators.

Annotator	Annotator	Kappa Score
Annotator 1	Annotator 2	0.6019
Annotator 1	Annotator 3	0.8777
Annotator 2	Annotator 3	0.6423

E.3 Detailed Setups

For the counterfactual logic attack experiment, we set the number of support verbs to one of $\{1, 2, 3, 4, 5\}$. We use instruction tuning with cloze prompts (Wei et al., 2021) in which the relation verb is missing. Adaptive fine-tuning uses 2 epochs, a learning rate of 1×10^{-5}, and `max_grad_norm=` 2. We use LoRA (Hu et al., 2021) with rank 64, alpha 16, and dropout 0.05 for all LLMs. During evaluation, we disable `do_sample`. All experiments are conducted on an NVIDIA H100 96GB GPU.

E.4 Evaluation Metrics

Since verb-centric counterfactual logic attack differs from existing attack paradigms, our focus is on whether the model's relation-level preference over verb concepts has changed. The test set contains attacked verbs placed in unseen application contexts, allowing us to evaluate whether the attack generalizes beyond the original training contexts.

CLR. The Counterfactual Logic Ratio measures the fraction of evaluation samples for which the attacked model prefers a counterfactual relation:

$$\text{CLR} = \frac{\text{number of counterfactual samples}}{\text{total number of samples}}.$$

For example, in the cloze prompt "Poor nutrition ____ overall health", predicting "enhances" instead of the factual relation "impairs" counts as a counterfactual outcome.

Relation uncertainty. We compare the relation uncertainty U_R of a given context before and after an attack to quantify the cluster-level preference shift.

Deviation Ratio

$$\text{Deviation Ratio} = \frac{\min(u_{\text{POS}}) - \max(u_{\text{NEG}})}{\max(u_{\text{NEG}})}.$$

Stereo Ratio. For the exploratory anti-stereotype evaluation, we define

$$\text{Stereo Ratio} = \frac{\text{number of stereotype samples}}{\text{number of total samples}}.$$

Stereotype labels are assigned by the specialized judgment model described in Appendix H.2. If the model predicts the biased class with confidence above 0.8, the sample is counted as stereotypical for this evaluation.

F Boundary Diagnostic Details

F.1 Pairwise Boundary Profiling

For each implication item, we construct three verb clusters: factual, counterfactual, and unrelated. These clusters correspond to POS, NEG, and IRR in the symbolic verb taxonomy, respectively. We keep at most three factual verbs, three counterfactual verbs, and two unrelated verbs for each item. The diagnostic model is queried with pairwise preference questions between candidate verbs. We use two prompt templates and swap the A/B order for each pair to reduce positional bias. Each order-swapped prompt is evaluated once. The pairwise score $p(a > b)$ is the fraction of these queries in which verb a is preferred over verb b.

Given factual cluster $\mathcal{T}$, counterfactual cluster $\mathcal{C}$, and unrelated cluster $\mathcal{U}$, we define:

$$S_{TC} = \frac{1}{|\mathcal{T}||\mathcal{C}|} \sum_{t \in \mathcal{T}} \sum_{c \in \mathcal{C}} p(t > c), \tag{27}$$

$$S_{TU} = \frac{1}{|\mathcal{T}||\mathcal{U}|} \sum_{t \in \mathcal{T}} \sum_{u \in \mathcal{U}} p(t > u). \tag{28}$$

For within-cluster decisiveness, we compute the average deviation from random preference within factual and counterfactual clusters:

$$S_{\text{within}} = \frac{1}{2}\left(\mathbb{E}_{t_i,t_j\in\mathcal{T},i<j}|p(t_i > t_j) - 0.5| + \mathbb{E}_{c_i,c_j\in\mathcal{C},i<j}|p(c_i > c_j) - 0.5|\right). \tag{29}$$

The boundary hardness is defined as:

$$H = S_{TC} - S_{\text{within}}. \tag{30}$$

F.2 Static Verb-Cluster Boundary Profiling

We repeat the GPT-4o boundary profiling from Section 4.3 for Qwen3-8B and Llama-3.1-8B-Instruct. Table 7 shows that both models exhibit clear cluster-level preferences on Verb-LK: factual verbs are preferred over both counterfactual and unrelated verbs, with $S_{TC} > 0.81$ and $S_{TU} \geq 0.87$. Their boundary hardness scores are also similar ($H = 0.493$ for Qwen3-8B and $H = 0.456$ for Llama-3.1-8B-Instruct), indicating that this structure is not specific to GPT-4o.

Table 7: Static verb-cluster boundary profiling results for Qwen3-8B and Llama-3.1-8B-Instruct. Hardness is computed from unrounded scores.

Model	Dataset	N	Valid	$\mathbf{S_{TC}}$	$\mathbf{S_{TU}}$	$\mathbf{S_{within}}$	H
Qwen3-8B	Verb-LK	149	80	0.811	0.870	0.318	0.493
	Weak Eval	35	26	0.644	0.957	0.251	0.393
Llama-3.1-8B-Instruct	Verb-LK	149	81	0.823	0.875	0.366	0.456
	Weak Eval	20	18	0.686	0.969	0.385	0.301

Both models show weaker factual-counterfactual boundaries on Weak Eval. Relative to Verb-LK, S_{TC} decreases to 0.644 for Qwen3-8B and 0.686 for Llama-3.1-8B-Instruct, while H decreases to 0.393 and 0.301, respectively. In contrast, S_{TU} remains high (0.957 and 0.969), confirming that Weak Eval contains boundary-near factual-counterfactual distinctions rather than noisy items that fail to separate factual from unrelated verbs. The shared pattern across models further supports the use of Weak Eval for context-shift analysis.

F.3 Boundary Hardness Stratification

Only items satisfying $S_{TC} > 0.5$ and $S_{TU} > 0.5$ are considered valid for stratification. We sort valid items by hardness H and use the upper and lower quartiles as cutoffs for the strong- and weak-boundary groups, respectively. All items tied at a cutoff are retained, so the realized group sizes need not equal exactly 25%. The remaining valid items are assigned to the mid-boundary group. Invalid items are excluded from the context-shift experiments. The stratification results are reported in Table 8.

Table 8: Boundary hardness stratification on the Verb-LK validation split.

Dataset	Valid	Strong	Weak	Mid	Invalid
Verb-LK	85	30	24	31	55

F.4 Boundary-near Weak Set Construction

Because the validation weak group in Verb-LK remains relatively strong, we construct an additional boundary-near weak set using the same pairwise boundary profiling protocol described above. We first generate 800 candidate implications and score each candidate with S_{TC}, S_{TU}, S_{within}, and hardness H.

For each candidate, we compute S_{TC} and S_{TU} using the same factual, counterfactual, and unrelated clusters as in boundary profiling. We keep items satisfying:

$$0.55 \leq S_{TC} \leq 0.80, \quad S_{TU} \geq 0.80. \tag{31}$$

The lower bound on S_{TC} ensures that factual verbs are still preferred over counterfactual verbs, while the upper bound selects items closer to the factual-counterfactual boundary. The S_{TU} threshold ensures that factual verbs remain clearly separated from unrelated verbs, so that the retained items are boundary-near rather than noisy.

This screening procedure keeps 39 boundary-near items. We split them into 24 evaluation items and 15 context-bank items. The resulting evaluation set has $S_{TC} = 0.604$, $S_{TU} = 1.000$, $S_{\text{within}} = 0.318$, and $H = 0.286$.

The 24-item set above is constructed for GPT-4o. For the open-weight models, we repeat the candidate generation, screening, and splitting procedure using each model's own boundary scores. This produces model-specific Weak Eval sets containing 35 evaluation items for Qwen3-8B and 20 evaluation items for Llama-3.1-8B-Instruct. The strong and weak groups are subsequently defined using model-specific hardness quartiles and are therefore relative to each model's screened evaluation set.

F.5 Polarity-aligned Context Retrieval

For GPT-4o, we retrieve from a context pool of 64 candidates with a minimum token-level Jaccard similarity of 0.2. The model-specific experiments use their actual retrieval settings: Qwen3-8B uses a context pool of 30, a Jaccard threshold of 0.0, and $k = 4$; Llama-3.1-8B-Instruct uses a context pool of 64, a Jaccard threshold of 0.2, and $k = 4$. Polarity alignment is applied after retrieval. Specifically, retrieved examples with reversed polarity are filtered out; same-polarity examples are preferred, and unknown-polarity examples are used only as fallback.

We evaluate four context conditions: no-, factual-, counterfactual-, and mixed context. Factual-context uses the gold verb of each retrieved example, counterfactual context uses its counter-logic verb, and mixed-context alternates between factual and counter-logic verbs. The factual preference score is then measured using the same pairwise protocol as in boundary profiling.

F.6 Dynamic Context-Shift Analysis

We next repeat the four context conditions from Section 4.3 for Qwen3-8B and Llama-3.1-8B-Instruct. As shown in Table 9, strong Verb-LK boundaries are largely invariant: counterfactual context reduces factual preference by only 0.025 for Qwen3-8B and not at all for Llama-3.1-8B-Instruct, with zero crossings for both models. The weak Verb-LK groups show slightly larger reductions ($\Delta^- = 0.065$ and 0.048), but boundary crossing remains rare or absent. These results agree with the primary diagnostic finding that inference-time context has limited influence on strong relational priors.

Table 9: Dynamic context-shift analysis results for Qwen3-8B and Llama-3.1-8B-Instruct.

Model	Set	Group	$\mathbf{S_{no}}$	$\mathbf{S_{fact}}$	$\mathbf{S_{count}}$	$\boldsymbol{\Delta^-}$	Spec.	Cross.
Qwen3-8B	Verb-LK	Strong	1.000	0.988	0.975	0.025	0.013	0.000
		Weak	0.860	0.830	0.795	0.065	0.015	0.080
	Weak Eval	Strong	0.812	0.969	0.578	0.234	0.172	0.125
		Weak	0.589	0.893	0.661	-0.071	0.054	0.000
Llama-3.1-8B-Instruct	Verb-LK	Strong	0.990	0.990	0.990	0.000	0.000	0.000
		Weak	0.821	0.738	0.774	0.048	-0.024	0.000
	Weak Eval	Strong	0.925	0.800	0.550	0.375	0.175	0.400
		Weak	0.825	0.600	0.500	0.325	0.025	0.200

Context effects are more pronounced on Weak Eval. For Qwen3-8B, the strong group drops from $S_{\text{no}} = 0.812$ to $S_{\text{count}} = 0.578$, with a crossing rate of 0.125; its weak group shows a clear contrast between factual and counterfactual contexts ($S_{\text{fact}} = 0.893$ versus $S_{\text{count}} = 0.661$), although counterfactual context does not reduce factual preference below the no-context baseline. Llama-3.1-8B-Instruct shows larger counterfactual shifts in both groups, with $\Delta^- = 0.375$ and 0.325 and crossing rates of 0.400 and 0.200, respectively. Because the strong and weak labels are defined relative to each dataset, even the strong subset of Weak Eval remains closer to the factual–counterfactual

boundary than strong Verb-LK items. Overall, context can substantially modulate boundary-near preferences, but its magnitude, direction, and relation to static hardness vary across models. Context does not consistently overturn the stronger boundaries found in Verb-LK.

F.7 Two-Hop Compositional Evaluation

We additionally evaluate whether the relation preferences studied in Verb-LK extend to two-hop compositions. Each item contains a factual chain $A \to B \to C$, a counterfactual twin that changes only the direction of the first hop, and three relation candidates for the final A–C query: factual, counterfactual, and neutral ("is unrelated to"). We score candidates by length-normalized conditional token log-probability. Thus, a *Full Flip* requires selecting the factual relation for the factual chain and the counterfactual relation for its matched twin.

The candidate pool was constructed from three reviewed signed-composition schemas: risk-factor decrease followed by risk increase (R06), protective factor increase followed by risk decrease (R07), and protective-factor decrease followed by risk decrease (R08). The chains underwent semantic review for validity, counterfactual isolation, and relation-candidate naturalness, followed by final human checking. For filtering, the P1-only and P2-only controls provide only the first or second hop, respectively. We then applied a fixed empirical filter with frozen Qwen3-8B and Llama-3.1-8B-Instruct: retained items had to exhibit the correct factual/counterfactual Full Flip for both models and a mean Full–P1-only margin of at least 0.05 across the two-model panel. This produced 45 items (R06: 18; R07: 9; R08: 18). P2-only scores are retained as a diagnostic but are not a hard filter because an isolated second hop can induce a directional prior over the final relation.

Table 10 reports evaluation on this fixed subset using the same premise-based reasoning prompt for every adapter. On Qwen3-8B, Contrastive Adaptation obtains the largest Full Flip rate (0.778), driven primarily by R06 (0.944). On Llama-3.1-8B-Instruct, Standard SFT has the highest aggregate Flip rate (0.622), while Contrastive Adaptation improves R06 (0.556) but is unstable on R07 (0.000) and R08 (0.389). These results suggest that transfer to two-hop compositions varies across base model families and relation schemas. Because the evaluation subset is small, restricted to three schemas, and selected using the two frozen base models, we treat it as a controlled diagnostic rather than evidence of uniformly improved multi-hop reasoning. The base-model scores are selection-conditioned and are therefore omitted from adapter-to-adapter comparisons. Overall, these results provide preliminary evidence that relation-boundary training can transfer to controlled two-hop compositions, while also showing that this transfer is not uniform across model families or compositional relation schemas.

Table 10: Two-hop compositional evaluation on the 45-item fixed subset. Full Flip requires a correct factual decision and a correct matched-counterfactual decision.

Model	Method	Factual	Counterfactual	Full Flip
Qwen3-8B	Standard SFT	0.978	0.644	0.644
Qwen3-8B	Contrastive-only	0.978	0.667	0.667
Qwen3-8B	Adaptive-only	0.822	0.667	0.644
Qwen3-8B	Contrastive Adaptation	0.978	**0.800**	**0.778**
Llama-3.1-8B-Instruct	Standard SFT	**1.000**	**0.622**	**0.622**
Llama-3.1-8B-Instruct	Contrastive-only	**1.000**	0.600	0.600
Llama-3.1-8B-Instruct	Adaptive-only	**1.000**	0.600	0.600
Llama-3.1-8B-Instruct	Contrastive Adaptation	0.756	0.600	0.378

For transparency, the rule-level Full Flip rates are shown in Table 11.

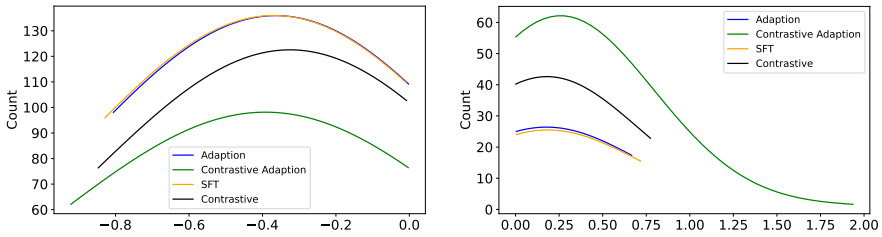
Table 11: Full Flip rate by two-hop composition schema.

Model	Method	R06 ($n = 18$)	R07 ($n = 9$)	R08 ($n = 18$)
Qwen3-8B	Standard SFT	0.444	1.000	0.667
Qwen3-8B	Contrastive-only	0.500	1.000	0.667
Qwen3-8B	Adaptive-only	0.444	1.000	0.667
Qwen3-8B	Contrastive Adaptation	**0.944**	1.000	0.500
Llama-3.1-8B-Instruct	Standard SFT	0.056	1.000	1.000
Llama-3.1-8B-Instruct	Contrastive-only	0.000	1.000	1.000
Llama-3.1-8B-Instruct	Adaptive-only	0.000	1.000	1.000
Llama-3.1-8B-Instruct	Contrastive Adaptation	**0.556**	0.000	0.389

G Additional Experimental Results

G.1 Deviation Ratio Distributions

Fig. 5 illustrates the deviation-ratio distributions, where larger positive values indicate clearer separation between POS and NEG clusters.



(a) Possible failed logic attack (b) Successful logic attack

Figure 5: Deviation Ratio distribution ($N = 3$, test set).

G.2 Counterfactual Logic Attack on the Test Set

Table 12 reports the counterfactual logic attack results on the test set. Overall, the results on the test set exhibit a trend similar to that observed in the training set. The contrastive adaptation mechanism achieves the best results across multiple configurations, demonstrating the strong generalization ability of the proposed framework.

Table 12: Counterfactual evaluation on the test set. Here, CA stands for Contrastive Adaptation. The numbers 1 to 5 represent the value of n_s. Cont and Adpt represent Contrastive and Adaptation, respectively.

Losses	β	α	1	2	3	4	5
Base	NA	NA	0.330	0.330	0.330	0.330	0.330
SFT	NA	NA	0.807	0.801	0.778	0.778	0.784
Cont	NA	0.1	0.847	0.835	0.841	0.852	0.835
	NA	0.5	0.960	0.955	0.943	0.960	0.943
Adpt	0.05	NA	0.784	0.801	0.795	0.795	0.801
	0.1	NA	0.784	0.784	0.773	0.773	0.773
	0.2	NA	0.795	0.756	0.761	0.739	0.739
	0.5	NA	0.801	0.443	0.477	0.517	0.494
CA	0.05	0.1	0.835	0.875	0.847	0.858	0.869
	0.1	0.1	0.830	0.869	0.795	0.790	0.864
	0.2	0.1	0.835	0.869	0.926	0.926	0.909
	0.5	0.1	0.835	0.636	0.824	0.790	0.801
	0.05	0.5	**0.966**	**0.989**	**0.966**	**0.966**	0.977
	0.1	0.5	0.943	0.920	0.955	0.915	**0.979**
	0.2	0.5	0.955	0.892	0.943	0.926	0.875
	0.5	0.5	0.960	0.290	0.250	0.267	0.273

G.3 Counterfactual Logic Attack: Additional Model

Table 13 reports CLR of Qwen3-8B on the training set. The unmodified Qwen3-8B checkpoint obtains a CLR of 0.1592, while SFT increases it to 0.2944. The single-component variants provide limited gains: contrastive learning does not outperform SFT, and adaptation peaks at 0.3103. In contrast, CA achieves the best result for every value of n_s, reaching 0.5199 at both $(\beta = 0.2, \alpha = 0.5, n_s = 1)$ and $(\beta = 0.05, \alpha = 0.5, n_s = 5)$. The setting $(\beta = 0.1, \alpha = 0.5)$ is comparatively stable across support-set sizes, with CLR ranging from 0.4204 to 0.4934. However, performance becomes less stable when β is large, and a small n_s is not consistently optimal for Qwen3-8B. These results show corresponding effects on Qwen3-8B, while also showing that the optimal support-set size and loss weights are model-dependent.

G.4 Effect of Candidate Source on Qwen3-8B

The main pipeline uses relation-verb candidates generated by an external model. We therefore examine whether the comparatively weaker Qwen3-8B results can be explained by candidate-source mismatch. To isolate this factor, we let Qwen3-8B generate its own candidate verbs, relabel the resulting candidates into POS, NEG, and IRR clusters, and recompute the candidate scores and supporting-token files. We then rerun four representative training settings under the same strict CLR evaluation protocol. All downstream components, including relation labeling, support-token selection, inference prompting, and CLR evaluation, remain unchanged. This is a diagnostic ablation over four representative settings rather than a replacement for the complete Qwen3-8B grid in Appendix G.3.

Table 14 shows a mixed pattern. Self-generated candidates leave SFT unchanged, improve the contrastive-only variant by 0.0239, and improve the adaptive-only variant by 0.0186. However, they reduce the strict CLR of a representative stable contrastive-adaptation configuration by 0.0623, from 0.4642 to 0.4019. Candidate-source mismatch is therefore a partial, configuration-dependent factor: it affects individual settings, but does not provide a general explanation for Qwen3-8B's behavior. In particular, the decline in the selected stable configuration indicates that its gains are not simply an artifact of using externally generated candidates. This pattern suggests that performance may depend on interactions among candidate quality, logical-category labeling, support-verb selection, and the training objective, which are not isolated by this ablation. Overall, this ablation does not support candidate-source mismatch as the primary explanation for the gap between Qwen3-8B and the main Llama-3.1-8B-Instruct results.

Table 13: Train counterfactual results for Qwen3-8B. Columns 1–5 denote n_s; Base is the unmodified Qwen3-8B checkpoint.

Losses	β	α	**1**	**2**	**3**	**4**	**5**
Base	NA	NA	0.1592	0.1592	0.1592	0.1592	0.1592
SFT	NA	NA	0.2944	0.2944	0.2944	0.2944	0.2944
Cont.	NA	0.1	0.2825	0.2825	0.2825	0.2825	0.2825
Cont.	NA	0.5	0.2188	0.2188	0.2188	0.2188	0.2188
Adpt.	0.05	NA	0.2666	0.2228	0.2520	0.2281	0.2374
Adpt.	0.1	NA	0.3050	0.2241	0.2202	0.2241	0.2268
Adpt.	0.2	NA	0.2493	0.2878	0.2321	0.3103	0.2706
Adpt.	0.5	NA	0.2745	0.2454	0.1910	0.2042	0.2308
CA	0.05	0.1	0.2387	0.2347	0.2454	0.2613	0.2560
CA	0.05	0.5	0.5040	0.2241	**0.4814**	0.4164	**0.5199**
CA	0.1	0.1	0.2321	0.2321	0.2215	0.1936	0.2599
CA	0.1	0.5	0.4416	**0.4934**	0.4642	**0.4668**	0.4204
CA	0.2	0.1	0.3011	0.2944	0.2520	0.2334	0.2546
CA	0.2	0.5	**0.5199**	0.3541	0.3342	0.4363	0.1923
CA	0.5	0.1	0.2520	0.2334	0.2215	0.1844	0.2255
CA	0.5	0.5	0.2202	0.1804	0.2387	0.3196	0.2679

Table 14: Effect of candidate source on Qwen3-8B under the strict CLR evaluation protocol. Self-generated candidates are produced by Qwen3-8B, while all downstream training and evaluation procedures remain unchanged. Δ denotes the difference from the original candidate pool.

Method	Setting	Original	Self-generated	Δ
SFT	$n_s = 3$	0.2944	0.2944	0.0000
Cont.	$\alpha = 0.1,\ n_s = 3$	0.2825	0.3064	0.0239
Adpt.	$\beta = 0.05,\ n_s = 3$	0.2520	0.2706	0.0186
CA	$\beta = 0.1,\ \alpha = 0.5,\ n_s = 3$	0.4642	0.4019	-0.0623

G.5 Uncertainty Boundary Analysis on Qwen3-8B

We further examine whether the Qwen3-8B training variants reconfigure the uncertainty boundary between factual-supporting and counterfactual verb clusters. Following the relation-uncertainty definition in Section 4.2, candidates labeled POS are treated as factual-supporting verbs and candidates labeled NEG as counterfactual verbs. For each setting, we compute the mean candidate-level negative log-probability within the two clusters. We report the uncertainty gap as

$$\mathrm{Gap} = U_{\mathrm{POS}} - U_{\mathrm{NEG}},$$

and its normalized form as

$$\mathrm{NormGap} = \frac{U_{\mathrm{POS}} - U_{\mathrm{NEG}}}{U_{\mathrm{NEG}}}.$$

A positive value indicates lower uncertainty for the counterfactual cluster than for the factual-supporting cluster, corresponding to the desired attack direction.

As shown in Table 15, the unmodified Qwen3-8B checkpoint exhibits a negative uncertainty gap and normalized uncertainty gap: counterfactual candidates have higher uncertainty than factual-supporting candidates, so the model remains biased toward the factual-supporting cluster. Every trained method reverses the sign and produces a positive gap. Contrastive adaptation yields the largest absolute uncertainty separation, increasing the POS–NEG gap to 4.0459. Contrastive-only training achieves a slightly larger normalized uncertainty gap than contrastive adaptation, 0.3536 versus 0.3021. Thus, contrastive adaptation does not dominate every uncertainty statistic, but it is among the strongest uncertainty-side configurations and simultaneously obtains the strongest strict CLR among the selected representative settings.

These results are directionally consistent with the main Qwen3-8B evaluation: the trained variants exhibit uncertainty shifts toward the counterfactual cluster, whereas the unmodified Qwen3-8B checkpoint preserves the original factual preference. This selected-configuration analysis provides supporting diagnostic evidence rather than a causal explanation, and it does not replace the complete hyperparameter grid. Together with Appendix G.4, these results suggest that Qwen3-8B's behavior is not explained by candidate source alone. Rather, the selected stable contrastive-adaptation configuration is associated with a substantial restructuring of the factual–counterfactual uncertainty boundary, although this diagnostic evidence should not be interpreted causally.

Table 15: Relation-uncertainty analysis for representative Qwen3-8B configurations. POS and NEG denote factual-supporting and counterfactual verb clusters, respectively; Base is the unmodified Qwen3-8B checkpoint. A positive uncertainty gap or normalized uncertainty gap indicates a shift toward the counterfactual cluster.

Method	Setting	U_{POS}	U_{NEG}	Gap	Norm. Gap
Base	–	29.5902	37.5055	-7.9153	-0.2110
SFT	–	7.7144	6.1191	1.5953	0.2607
Cont.	$\alpha = 0.1$	8.5593	6.3233	2.2359	0.3536
Adpt.	$\beta = 0.2,\ n_s = 4$	8.8959	7.3279	1.5680	0.2140
CA	$\beta = 0.1,\ \alpha = 0.5,\ n_s = 3$	17.4377	13.3918	4.0459	0.3021

G.6 Significance Test

Table 16 reports one-tailed paired-sample t-tests on the test set using the best-performing configuration of each method at $n_s = 3$: $(\beta = 0.05, \alpha = 0.5)$ for contrastive adaptation, $\beta = 0.05$ for adaptation, and $\alpha = 0.5$ for contrastive. Contrastive adaptation significantly outperforms adaptation, contrastive, SFT, and Base under the tested alternative. The one-tailed test provides no evidence that adaptation outperforms SFT.

Table 16: One-tailed paired-sample t-tests on the test set using each method's best-performing configuration at $n_s = 3$. CA uses $(\beta = 0.05, \alpha = 0.5)$, adaptation uses $\beta = 0.05$, and contrastive uses $\alpha = 0.5$.

Method 1	Method 2	p-value
Contrastive Adaptation	Adaptation	1.3976×10^{-20}
Contrastive Adaptation	Contrastive	1.2349×10^{-19}
Contrastive Adaptation	SFT	8.9152×10^{-21}
Contrastive Adaptation	Base	4.6647×10^{-29}
Contrastive	SFT	9.1843×10^{-11}
Adaptation	SFT	9.9990×10^{-1}

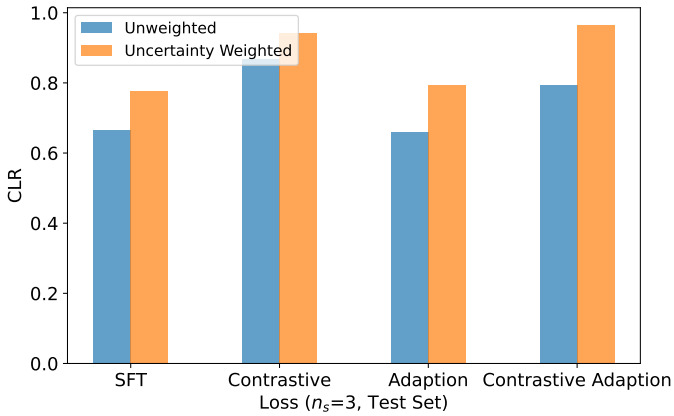


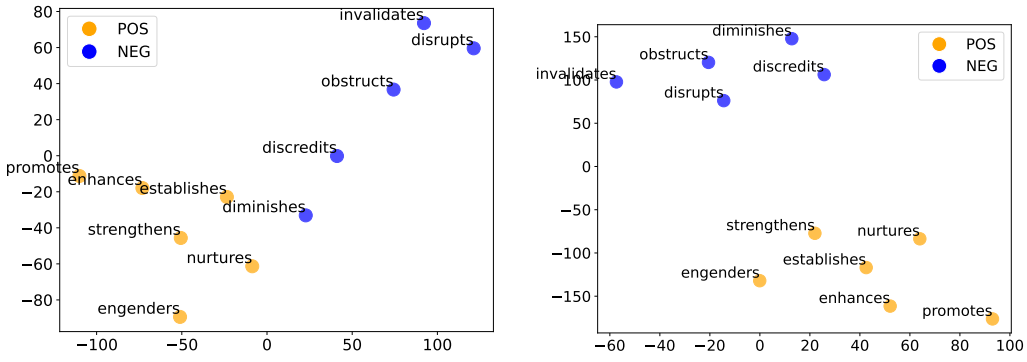
Figure 6: Comparison of unweighted and uncertainty-weighted sampling and losses.

G.7 Effect of Uncertainty-Based Verb Selection

We examine our uncertainty-based support-verb selection mechanism. As shown in Fig. 6, uncertainty-weighted selection consistently outperforms unweighted selection across all settings. Removing uncertainty weighting causes larger reductions for **Adaptation** and **Contrastive Adaptation** than for **SFT** and **Contrastive**, consistent with the former methods optimizing both center and support verbs.

G.8 Semantic Visualization

Fig. 7 shows that POS and NEG verb-cluster embeddings become more separated after ALA.



(a) Embeddings before logic attack

(b) Embeddings after ALA

Figure 7: t-SNE visualization of POS and NEG verb-cluster embeddings before and after ALA.

G.9 MixEval Capability Check

Multi-hop reasoning on Qwen3-8B. Although the overall MixEval decrease is modest, the subtask-level results reveal a more concentrated effect on reasoning-intensive benchmarks, particularly BBH and MATH. As shown in Table 17, contrastive adaptation (CA), the full ALA method, reduces the overall Qwen3-8B score from 0.760 to 0.753, a decrease of only 0.7 percentage points. In contrast, BBH decreases from 0.799 to 0.728 and MATH from 0.790 to 0.720, corresponding to reductions of 7.1 and 7.0 percentage points, respectively. These declines are approximately ten times the aggregate decrease, showing that the overall score masks substantially larger effects on reasoning-intensive subtasks. CA also produces the lowest BBH and MATH scores among the evaluated variants, with additional reductions of 0.4 and 2.5 percentage points relative to SFT. Meanwhile, CA matches the

reported Base scores on BoolQ, CommonsenseQA, PIQA, and OpenBookQA, and leaves GSM8K nearly unchanged. This non-uniform pattern highlights concentrated reasoning effects rather than a comparable decline across general capabilities.

Table 17: Test of multi-hop reasoning on MixEval for Qwen3-8B. Results were evaluated on August 4, 2026, using DeepSeek-V4-Flash-0731 (API model name: `deepseek-v4-flash`) as the judge and are reported on a $[0, 1]$ scale. Base denotes the unmodified Qwen3-8B checkpoint without an adapter; CA denotes contrastive adaptation, the full ALA method. The two panels together report the overall score and all 13 benchmark-specific scores.

Method	Overall	BBH	MATH	GSM8K	DROP	MMLU	TriviaQA
Base (no adapter)	0.760	0.799	0.790	0.956	0.776	0.758	0.661
Contrastive	0.756	0.798	0.795	0.943	0.785	0.757	0.651
SFT	0.754	0.732	0.745	0.963	0.786	0.751	0.656
CA	0.753	0.728	0.720	0.957	0.766	0.753	0.658
Adaptation	0.750	0.747	0.770	0.943	0.770	0.754	0.654

Method	HellaSwag	AGIEval	BoolQ	Commonsense QA	PIQA	SIQA	OpenBook QA
Base (no adapter)	0.862	0.715	0.869	0.821	0.865	0.765	0.795
Contrastive	0.832	0.700	0.869	0.816	0.885	0.755	0.795
SFT	0.842	0.700	0.851	0.811	0.885	0.745	0.773
CA	0.842	0.708	0.869	0.821	0.865	0.735	0.795
Adaptation	0.838	0.654	0.863	0.816	0.865	0.755	0.795

Table 18: Paired test against the Base model on the overall MixEval evaluation for Qwen3-8B. Differences are computed from unrounded item-level scores relative to the unmodified Qwen3-8B checkpoint; negative values indicate lower scores. Confidence intervals are obtained by paired bootstrap. CA denotes contrastive adaptation, the full ALA method.

Adapter	Difference vs. Base	95% CI (Paired Bootstrap)	Significance
Contrastive	-0.0049	$[-0.0109, +0.0014]$	Not significant
SFT	-0.0064	$[-0.0128, -0.0002]$	Significant
CA	-0.0072	$[-0.0135, -0.0006]$	Significant
Adaptation	-0.0096	$[-0.0160, -0.0033]$	Significant

Paired significance test against the Base model. The paired test on the overall MixEval evaluation complements the subtask-level analysis. Table 18 reports a CA–Base difference of -0.0072, with a paired-bootstrap 95% confidence interval of $[-0.0135, -0.0006]$. The interval lies entirely below zero, establishing a statistically significant reduction relative to the unmodified Qwen3-8B checkpoint in this paired evaluation. By comparison, the contrastive-only interval $[-0.0109, +0.0014]$ includes zero, and its change is not significant. Together, the concentrated declines on BBH and MATH and the significant paired CA–Base reduction indicate that CA has non-uniform downstream effects concentrated on reasoning-intensive tasks, rather than merely causing uniform degradation in general utility. These MixEval results do not by themselves isolate relation-sensitive multi-hop failure. However, when considered with the controlled two-hop diagnostic in Appendix F.7, they provide complementary evidence that ALA can affect compositional reasoning, while the magnitude and direction of this effect remain model- and relation-schema-dependent.

Broad utility check on Llama-3.1-8B-Instruct. Table 19 complements the Qwen3-8B analysis with an aggregate utility assessment on Llama-3.1-8B-Instruct. CA changes the MixEval score from 73.30 to 72.60, a decrease of 0.70 percentage points, compared with a 1.45-percentage-point decrease under SFT. Thus, CA retains aggregate performance close to the Base model and incurs less than half the score reduction observed under SFT. Across both model families, the overall MixEval change remains modest; the Qwen3-8B subtask analysis and paired significance test reveal the more concentrated reasoning impact beneath this aggregate pattern.

Table 19: Broad reasoning-capability check on MixEval for Llama-3.1-8B-Instruct. Base denotes the unmodified Llama-3.1-8B-Instruct checkpoint. Scores are reported as percentages. Aggregate score changes do not isolate relation-sensitive multi-hop disruption.

Method	Score (%)
Base	73.30
SFT	71.85
Contrastive	72.55
Adaptation	72.20
CA	72.60

H ANTI-STEREOTYPE ALIGNMENT

H.1 STEREOSET

We adopt 1,050 samples from StereoSet (Nadeem et al., 2021) and use GPT-4o to select clearly biased sentences. For each sentence, we use its stereotype–anti-stereotype word pair as the center pair and prompt GPT-4o to generate four additional synonyms for each side as support words.

For the transfer setting, a statement containing modifiable words o can be represented as a triplet (s, r, o), where s is a subject, r is a linking verb, and o is a predicative complement. For example, "Stepmothers are evil" can be represented as $(\text{stepmothers}, \text{are}, \text{evil})$. We treat stereotype words, such as "evil", as the original cluster, and anti-stereotype words, such as "nice", as the counterfactual cluster. The domain distributions of the selected samples are shown in Table 20.

Table 20: Stereotype domain distribution of selected samples from StereoSet.

Domain	Religion	Race	Profession	Gender
Number	158	468	184	240

H.2 STEREOTYPE JUDGMENT MODEL

The model used for evaluation is based on the pretrained weights of distilroberta-base, with a classification head fine-tuned to classify text into two categories: neutral and biased. This model was fine-tuned on the wikirev-bias dataset, which consists of English Wikipedia revisions. The dataset is designed to detect biased language and was extracted from the WNC wiki edits corpus. The model accepts inputs with a maximum length of 512 tokens, consistent with its underlying distilroberta-base encoder, making it suitable for processing moderately long text segments while maintaining efficient performance.

The reduced Stereo Ratios in Table 4 provide preliminary evidence that ALA may extend to the anti-stereotype task under this evaluation protocol. ALA achieves the lowest overall Stereo Ratio among the tested methods. Under the setting of $\beta = 0.05$ and $\alpha = 0.5$, the overall Stereo Ratio is reduced to 0.122, compared with 0.203 for the unmodified Llama-3.1-8B-Instruct checkpoint and 0.222 for SFT. This setting also achieves the lowest gender and race ratios, reducing them to 0.050 and 0.134, respectively. The unmodified checkpoint shows the highest race ratio of 0.313, and SFT increases the overall Stereo Ratio to 0.222, suggesting that standard fine-tuning does not reliably mitigate stereotypical outputs in this setting. Adaptation alone achieves an overall ratio of 0.154 when $\beta = 0.05$, while ALA further improves this result through the contrastive adaptation objective. Another ALA setting, $\beta = 0.5$ and $\alpha = 0.1$, achieves the lowest profession and religion ratios, 0.115 and 0.043, respectively. Overall, these results are encouraging but should be interpreted as exploratory rather than as a validated transfer claim.