MITIGATING CONFORMITY IN LARGE LANGUAGE MODELS VIA SOCIAL-CONTEXT-AWARE UNLEARNING

Anonymous authors
Paper under double-blind review

ABSTRACT

Humans often revise their judgments when surrounded by a confident majority. Large Language Models (LLMs), increasingly used as conversational assistants and collaborative agents, exhibit a similar tendency and may answer correctly in isolation but switch to an incorrect answer after observing peer responses or discussion history. Prior works have explored mitigation through prompt-level interventions and reflection. However, these approaches mainly reshape the input context at inference time. This raises a natural question of whether conformity can be mitigated through parameter-level correction rather than prompt-level intervention alone. Our key observation is that conformity-inducing interaction traces naturally reveal both what should be forgotten and what should be preserved. Based on this observation, we propose to mitigate conformity through machine unlearning. Intuitively, unlearning updates model parameters to reduce the likelihood of socially shifted responses under conformity-inducing contexts, while retaining supervision helps preserve reference-supported behavior. We develop a conformity-targeted unlearning framework that identifies answer shifts, constructs forget-retain pairs, and applies selective unlearning to suppress conformity-driven errors without erasing task competence. Our experimental results show that unlearning can reduce socially induced errors while preserving the model's original reasoning ability. These findings suggest a new perspective on LLM conformity: rather than being solely a test-time prompting issue, conformity can be viewed as a behavioral failure mode that can be explicitly identified and corrected through selective unlearning. Our code is available at https://anonymous.4open.science/r/xdefewgferh.

1 INTRODUCTION

When people answer questions in groups, their judgments are often shaped by the opinions around them. A person who initially knows the correct answer may still hesitate, revise their response, or follow an incorrect majority after observing others' choices. This phenomenon, commonly known as conformity, has long been studied in social psychology as a form of social influence that can weaken independent judgment (Coultas & Van Leeuwen, 2015; Sowden et al., 2018).

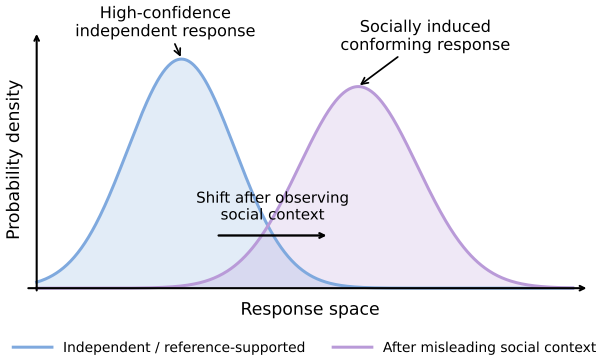

Figure 2: Illustration of our motivating observation.

Large language models (LLMs) are increasingly deployed in similar socially conditioned settings. They are used as conversational assistants, collaborative agents, and participants in multi-agent systems, where their responses are often conditioned on previous messages, peer opinions, or group discussions (Li et al., 2023; Wu et al., 2024; Park et al., 2023; Hong et al., 2023; Guo et al., 2024; Du et al., 2024). While such interaction can improve coordination and collective problem solving (Chan et al., 2023; Khan et al., 2024; Du et al., 2024), it also introduces a risk: an LLM may abandon a

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.

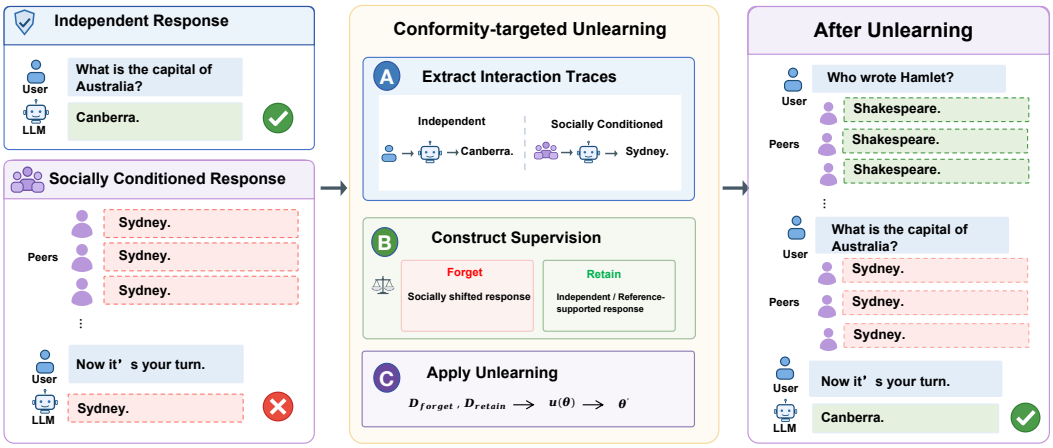

Figure 1: Overview of our conformity-targeted unlearning framework. Under socially conditioned interaction, a model may shift from its independent or reference-supported response toward a socially induced response. We construct forget–retain supervision from these interaction traces to instill confidence in the model while preserving its in-context learning capabilities. After unlearning, the model reduces conformity and maintains task-faithful responses. The right panel conceptually illustrates the desired behavior: the model can still use task-faithful peer responses while resisting a later unanimous but incorrect majority. This illustration is not used as a separate training or evaluation.

correct answer and instead follow a socially salient but incorrect response (Song et al., 2025; Xu et al., 2026a). This behavior is especially concerning for reasoning-oriented question answering, where correctness can be clearly evaluated and where socially induced errors directly reduce reliability.

Recent work shows that LLMs exhibit conformity across multi-agent protocols and psychology-inspired settings, often shifting toward majority opinions even when they are incorrect (Weng et al., 2025; Zhu et al., 2025; Zhang et al., 2023). Existing mitigation strategies mainly operate at inference time by modifying prompts or dialogue structure, without directly altering the model's tendency to reproduce conformity-driven errors (Madaan et al., 2023; Shinn et al., 2023; Zhu et al., 2023; Gou et al., 2023). As a result, their effectiveness may depend heavily on prompt design, model-specific behavior, and the structure of the interaction. Moreover, prompt-based mitigation treats conformity as a contextual artifact to be avoided at test time, rather than as a recurring behavioral failure mode that can be identified and corrected at the model level. This raises a natural question: can we mitigate LLM conformity through model-level intervention while preserving the model's original task competence? As a motivating observation, we find that socially conditioned prompts can induce models to move away from their high-confidence independent responses. As illustrated in Figure 2, the model may select an answer confidently in the independent setting, but shift toward a socially induced conforming response after observing repeated discussion history. This suggests that conformity can be characterized as a recurring pattern of socially induced response shifts rather than isolated task errors, providing a principled basis for updating the model to mitigate this behavior.

In this paper, we formulate conformity mitigation at the model level as a selective machine unlearning problem. Our key observation is that conformity-inducing traces naturally reveal both a response to suppress and a response to preserve. When a model gives a high-confidence or reference-supported response in isolation but shifts after observing social context, the socially induced shifted response can be treated as a forget target, while the independent or reference-supported response serves as a retain target. This creates a direct bridge between conformity mitigation and machine unlearning: rather than erasing arbitrary data or general knowledge, we selectively suppress the model's tendency to reproduce harmful social influence while preserving task-faithful reasoning (Bourtoule et al., 2021; Jang et al., 2023; Maini et al., 2024).

Building on this perspective, we propose a conformity-targeted unlearning framework. Given a pretrained LLM and conformity-inducing interaction traces, our framework first identifies socially induced answer shifts, then constructs forget-retain pairs that separate socially shifted responses from independent or reference-supported responses, and finally applies selective unlearning to reduce conformity while maintaining general performance. We further consider both *one-time* unlearning

and *periodic* unlearning, allowing us to evaluate whether model-level correction can mitigate not only static conformity failures but also recurring failures across interaction rounds.

Our contributions are threefold:

- We formulate LLM conformity mitigation as an unlearning problem. This reframes conformity reduction from a purely prompt-level intervention into a model-level correction problem, where socially induced shifted responses are suppressed, and independent or reference-supported responses are retained.
- We propose a conformity-targeted intervention framework that constructs forget-retain supervision from socially conditioned QA traces and supports both one-time and periodic unlearning. This framework enables systematic study of whether conformity can be corrected once or continuously mitigated across interaction cycles.
- We conduct an extensive empirical evaluation across two model families and multiple conformity settings, comparing selective unlearning with both prompt-level defenses and stronger model-level interventions. The results show substantial conformity reduction while preserving general utility and responsiveness to corrective social context.

2 Related Work

Conformity and Social Influence in LLMs. Conformity is a form of social influence in which individuals adjust their beliefs or behaviors toward group opinions. Recent social psychology work has studied conformity from multiple perspectives, including group size, public versus private compliance, peer influence, and herding behavior (Sowden et al., 2018; Mattke et al., 2020; Deng et al., 2026). Recent work has shown that LLMs can exhibit conformity-like behavior in multi-agent collaboration and psychology-inspired dialogue settings (Xiong et al., 2023; Weng et al., 2025; Zhu et al., 2025; Song et al., 2025). Weng et al. (2025) introduce BENCHFORM to evaluate conformity under Raw, Correct Guidance, Wrong Guidance, Trust, and Doubt protocols, while Zhu et al. (2025) adapt classical conformity experiments to LLMs and show that models may shift toward unanimous incorrect majority answers. Our work builds on these studies, but shifts the focus from measuring conformity to correcting it through model-level unlearning.

Prompt- and Model-Level Mitigation of Conformity. Existing mitigation strategies for LLM conformity mainly operate at inference time. BENCHFORM explores empowered personas and reflection mechanisms to encourage independent reasoning (Weng et al., 2025), while Zhu et al. (2025) propose Devil's Advocate and Question Distillation to reduce the influence of unanimous incorrect answers. Related work on collective misremembering in multi-agent LLM systems also studies both prompt-level defenses and model-level alignment-based mitigation (Xu et al., 2026b; Wu et al., 2025; Li et al., 2025). More broadly, reflection and self-correction have been studied as test-time strategies for improving LLM reasoning and decision-making (Yao et al., 2023; Shinn et al., 2023; Wang et al., 2022). These approaches motivate comparison between inference-time defenses and parameter-level intervention. In contrast, our work specifically formulates conformity-driven response shifts as forget-retain supervision for selective unlearning.

Machine Unlearning. Machine unlearning aims to remove the influence of target data or behaviors from a trained model while preserving utility on unrelated data (Cao & Yang, 2015; Bourtoule et al., 2021). Recent work extends unlearning to language models, including knowledge unlearning for privacy-sensitive memorized content (Jang et al., 2023; Eldan & Russinovich, 2023) and preference-based objectives such as Negative Preference Optimization for more stable LLM unlearning (Zhang et al., 2024). These methods typically focus on removing private, copyrighted, harmful, or otherwise undesirable knowledge. Our work applies unlearning to a behavioral failure mode, namely socially induced conformity. Instead of constructing the forget set from sensitive documents or target facts, we derive forget-retain supervision from interaction traces, so that socially shifted responses are suppressed while independent or reference-supported responses are preserved.

3 Methods

Figure 1 provides an overview of our framework for mitigating LLM conformity. Given a pretrained LLM and conformity-inducing interaction traces, we first identify socially induced answer shifts, then

construct forget-retain pairs that separate socially shifted responses from independent or reference-supported responses, and finally apply selective unlearning to reduce conformity while preserving task competence. Implementation details are provided in Appendix A.

3.1 PROBLEM SETUP

Given a question $x \in \mathcal{X}$ with a reference answer $y^{\text{ref}} \in \mathcal{Y}$, a language model parameterized by θ produces an answer distribution $p_\theta(y \mid x, c)$, where c denotes the social context. For labeled QA benchmarks, y^{ref} corresponds to the dataset answer. In settings where correctness is ambiguous, it can be instantiated as the model's high-confidence independent response. The independent setting corresponds to $c = \varnothing$, while $c \neq \varnothing$ represents socially conditioned inputs, such as peer responses, majority opinions, or discussion history.

For each example x_i, we first query the initial model θ_0 without social context and obtain an independent response $\hat{y}_i^{\text{raw}} \sim p_{\theta_0}(y \mid x_i, \varnothing)$. We then query the same model under socially conditioned prompts and obtain $\hat{y}_i^{(p,r)} \sim p_{\theta_0}(y \mid x_i, c_i^{(p,r)})$, where p indexes the interaction protocol and r denotes the discussion round or the amount of previous social context.

Let s_i^{raw} denote the model's confidence score for its independent response, such as the normalized log-probability of the generated answer. To isolate socially induced shifts from ordinary uncertainty, we focus on examples where the model produces a reference-supported and sufficiently high-confidence response in isolation:

$$\mathcal{I}_{\text{base}} = \{i \mid \hat{y}_i^{\text{raw}} = y_i^{\text{ref}},\ s_i^{\text{raw}} \geq \tau\}. \tag{1}$$

A conformity shift occurs when an example in $\mathcal{I}_{\text{base}}$ moves away from the independent response after social context is introduced. We define the set of observed conformity shifts as

$$\mathcal{F} = \left\{(i, p, r) \mid i \in \mathcal{I}_{\text{base}}, \hat{y}_i^{(p,r)} \neq y_i^{\text{ref}}\right\}. \tag{2}$$

Each element in $\mathcal{F}$ corresponds to a case where the model moves away from its independent or reference-supported response and produces a socially induced shifted response.

We also define a retain set containing examples that remain aligned with the reference-supported response under social context:

$$\mathcal{R} = \left\{(i, p, r) \mid i \in \mathcal{I}_{\text{base}}, \hat{y}_i^{(p,r)} = y_i^{\text{ref}}\right\}. \tag{3}$$

The retain set is used to preserve the model's task-faithful behavior during unlearning. This formulation targets harmful conformity rather than social influence in general. Specifically, it aims to reduce cases where social context causes the model to move away from an independent or reference-supported response, rather than making the model ignore all external opinions.

3.2 CONFORMITY-TARGETED UNLEARNING

We apply unlearning as a model-level intervention for conformity mitigation. Our goal is to define how socially induced response shifts can be converted into unlearning supervision.

Extract conformity shifts. For each conformity shift $(i, p, r) \in \mathcal{F}$, we denote the full socially conditioned input as $z_i^{(p,r)} = (x_i, c_i^{(p,r)})$, where x_i is the original question and $c_i^{(p,r)}$ is the social context induced by protocol p at round r. The model's socially induced answer is denoted as $\hat{y}_i^c = \hat{y}_i^{(p,r)}$. Since $(i, p, r) \in \mathcal{F}$ corresponds to cases where social context shifts the model away from its independent or reference-supported response, $\hat{y}_i^c$ represents the socially induced behavior to be suppressed.

Construct forget-retain supervision. We construct the forget set from socially induced shifted responses:

$$\mathcal{D}_{\text{forget}} = \{(z_i^{(p,r)}, \hat{y}_i^c) \mid (i, p, r) \in \mathcal{F}\}. \tag{4}$$

Each example in $\mathcal{D}_{\text{forget}}$ specifies an answer pattern that the model should reduce under the corresponding social context. We then construct the retain set using task-faithful answers:

$$\mathcal{D}_{\text{retain}} = \{(z_i^{(p,r)}, y_i^{\text{ref}}) \mid (i, p, r) \in \mathcal{F} \cup \mathcal{R}\}. \tag{5}$$

The retain set prevents unlearning from simply weakening the model. It anchors the model to the independent or reference-supported response and helps preserve task competence. For shifted examples, the same input can appear in both sets with different targets: the forget pair suppresses the socially shifted answer, while the retain pair anchors the same context to the reference answer.

Apply unlearning. Given $\mathcal{D}_{\text{forget}}$ and $\mathcal{D}_{\text{retain}}$, we apply an existing unlearning method as:

$$\theta' = \mathcal{U}\left(\theta; \mathcal{D}_{\text{forget}}, \mathcal{D}_{\text{retain}}\right), \tag{6}$$

where $\mathcal{U}$ denotes a generic unlearning operator. We remark that any unlearning algorithm based on parameter updates can be applied to Eq. 6. Without loss of generality, in experiments, we instantiate $\mathcal{U}$ with representative unlearning methods, including gradient-ascent-based unlearning (Jang et al., 2023), negative preference optimization (Zhang et al., 2024), and RMU (Li et al., 2024). These methods differ in their optimization mechanisms, but all operate on the same conformity-targeted supervision constructed above.

3.3 Periodic conformity-targeted unlearning

The conformity-targeted unlearning procedure can be applied either once or periodically. In the one-time setting, we collect conformity shifts from the initial model θ_0, construct $\mathcal{D}_{\text{forget}}$ and $\mathcal{D}_{\text{retain}}$, and apply the unlearning solver once:

$$\theta_1 = \mathcal{U}\left(\theta_0; \mathcal{D}_{\text{forget}}, \mathcal{D}_{\text{retain}}\right). \tag{7}$$

This setting tests whether a single model-level correction is sufficient to reduce conformity on held-out socially conditioned examples.

However, conformity may reappear after the model is updated, especially when the model encounters new social contexts, interaction protocols, or longer discussion histories. We therefore extend one-time unlearning into a periodic correction process. Let θ_t denote the model at correction cycle t. At each cycle, we evaluate θ_t under both independent and socially conditioned settings, identify newly observed conformity shifts, rebuild the forget and retain sets, and then apply the unlearning solver to obtain the next model:

$$\theta_{t+1} = \mathcal{U}\left(\theta_t; \mathcal{D}^{(t)}_{\text{forget}}, \mathcal{D}^{(t)}_{\text{retain}}\right). \tag{8}$$

More specifically, at cycle t, we first obtain the model's independent responses $\hat{y}^{\text{raw}}_{i,t}$ and socially conditioned responses $\hat{y}^{(p,r)}_{i,t}$. We then define the current base set as

$$\mathcal{I}^{(t)}_{\text{base}} = \{i \mid \hat{y}^{\text{raw}}_{i,t} = y^{\text{ref}}_i,\ s^{\text{raw}}_{i,t} \geq \tau\}, \tag{9}$$

where $s^{\text{raw}}_{i,t}$ denotes the confidence score of the independent response at cycle t. We then collect the current conformity shifts:

$$\mathcal{F}^{(t)} = \left\{(i,p,r) \mid i \in \mathcal{I}^{(t)}_{\text{base}}, \hat{y}^{(p,r)}_{i,t} \neq y^{\text{ref}}_i\right\}. \tag{10}$$

We also define the retained set at cycle t as

$$\mathcal{R}^{(t)} = \left\{(i,p,r) \mid i \in \mathcal{I}^{(t)}_{\text{base}}, \hat{y}^{(p,r)}_{i,t} = y^{\text{ref}}_i\right\}. \tag{11}$$

The cycle-specific forget set is constructed from newly observed socially induced shifted responses. For retain supervision, we combine the current cycle's reference-supported responses with a fixed anchor retain set. The anchor set is constructed from the initial conformity shifts and retained cases:

$$\mathcal{D}_{\text{anchor}} = \{(z^{(p,r)}_i, y^{\text{ref}}_i) \mid (i,p,r) \in \mathcal{F} \cup \mathcal{R}\}. \tag{12}$$

Including this anchor set in every cycle is intended to reduce the risk of compounding forgetting, since initially reference-supported cases remain protected even if their confidence decreases later.

The cycle-specific forget and retain sets are then defined as:

$$\mathcal{D}^{(t)}_{\text{forget}} = \{(z^{(p,r)}_i, \hat{y}^{(p,r)}_{i,t}) \mid (i,p,r) \in \mathcal{F}^{(t)}\}, \tag{13}$$

$$\mathcal{D}^{(t)}_{\text{retain}} = \{(z^{(p,r)}_i, y^{\text{ref}}_i) \mid (i,p,r) \in \mathcal{F}^{(t)} \cup \mathcal{R}^{(t)}\} \cup \mathcal{D}_{\text{anchor}}. \tag{14}$$

After T cycles, the final model is θ_T. When $T = 1$, periodic unlearning reduces to one-time unlearning. When $T > 1$, the framework repeatedly discovers the current model's socially induced response shifts and corrects them before the next evaluation cycle. This design treats conformity as a recurring behavioral failure mode rather than a fixed static error set.

4 EXPERIMENTS

4.1 EXPERIMENTAL SETTINGS.

Datasets and models. We evaluate on two conformity settings. The first is BENCHFORM (Weng et al., 2025), which includes Raw, Correct Guidance, Wrong Guidance, Trust, and Doubt protocols; our main evaluation focuses on the long-term Trust and Doubt settings, with Correct Guidance reported separately. The second follows the psychology-inspired setting of Zhu et al. (2025), where models answer independently before observing unanimous simulated participants, using MMLU (Hendrycks et al., 2020), BBH Object Counting (Suzgun et al., 2023), PopQA (Mallen et al., 2023), and CommonsenseQA (Talmor et al., 2019). We use question-disjoint unlearning and evaluation splits, with all forget-retain supervision constructed only from the unlearning split; periodic updates likewise use only this split, while evaluation is performed on a fixed held-out set. Experiments use Llama-3-8B and Qwen-2.5-14B. See Appendix A for implementation details.

Baselines. We compare model-level unlearning with both the original model and prompt-level mitigation strategies. The prompt-based baselines include **Persona prompting** and **Reflection** from BENCHFORM (Weng et al., 2025), as well as **Devil's Advocate (DA)** and **Question Distillation (QD)** from the psychology-inspired conformity setting (Zhu et al., 2025). For model-level intervention, we instantiate our framework with **Gradient Ascent (GA)** (Jang et al., 2023) and **Negative Preference Optimization (NPO)** (Zhang et al., 2024). We compare against several additional model-level baselines. **Balanced SFT**, which we construct as a stronger supervised fine-tuning baseline, jointly fine-tunes on resilience, corrective-guidance, and cooperative-guidance examples. **DPO** (Rafailov et al., 2024) treats the reference-supported response as preferred and the socially shifted response as dispreferred. We further instantiate our SUL framework with **RMU** (Li et al., 2024), a representation-level unlearning method that redirects forget-set representations while preserving retain-set representations. We also retain **SFT-Retain** as an ablation that updates only on task-faithful responses. We re-implement all prompt-level baselines under the same held-out splits and decoding setup, and adapt their prompt templates to each evaluation setting without tuning on the evaluation split, ensuring a controlled and consistent comparison

Metrics. For BENCHFORM, we report **Accuracy** (Acc), **Conformity Rate** (CR), and **Independence Rate** (IR), following Weng et al. (2025). BF-Acc and BF-CR are averaged over Trust and Doubt; Raw defines the Raw-correct subset, and Correct Guidance is reported separately. CR measures the fraction of Raw-correct examples that become incorrect under social influence, while IR measures those that remain correct across long-term interactions. For the psychology-inspired setting, we report **Conformity Level** (CL) and **Resistance Level** (RL), following Zhu et al. (2025). Lower CR/CL and higher IR/RL indicate stronger resistance to misleading social pressure. We additionally report **Utility Drop** (ΔUtil) on ARC-Challenge and HellaSwag, **Initial-Wrong Correction**, **Wrong-Guidance CR**, and CL-Bench **Task Solve Rate** and **Rubric Pass Rate** Dou et al. (2026). Metric details are provided in Appendix B.

4.2 RESULTS AND ANALYSIS

Table 1 presents the comparison between prompt-level mitigation and model-level unlearning. We report average results across the main conformity-inducing protocols on BENCHFORM and the psychology-inspired conformity setting. For simplicity, we use **SUL** to represent our social-context-aware unlearning method with different unlearning strategies such as GA, NPO, and RMU. Overall, GA and NPO consistently reduce harmful conformity across both model families. RMU is also competitive on BENCHFORM across both models, but yields smaller gains than GA and NPO in the psychology-inspired conformity setting. Compared with prompt-based mitigation methods, these parameter-level interventions directly modify the model's tendency to reproduce socially induced shifted responses rather than only changing the inference-time prompt. Across both model families, GA and NPO improve socially conditioned accuracy while reducing conformity, with GA achieving the lowest BF-CR among the evaluated methods.

GA and NPO consistently reduce conformity across both model families. On Llama-3-8B, GA reduces BF-CR from 28.25% to 9.69%, while improving BF-Acc from 44.76% to 57.24%. NPO shows a similar trend, reducing BF-CR by 15.38 percentage points and improving BF-Acc by 11.82 percentage points. The same pattern holds for Qwen-2.5-14B, where GA reduces BF-CR from 21.68% to 7.84%. The expanded comparison further shows that different model-level objectives

Table 1: Main results comparing prompt-level and model-level interventions. BF-Acc, BF-CR, and BF-IR denote accuracy, conformity rate, and independence rate on BENCHFORM. CL and RL denote conformity level and resistance level in the psychology-inspired conformity setting.

Model	Method	BF-Acc (%) ↑	BF-CR (%)↓	BF-IR (%) ↑	CL (%) ↓	RL (%) ↑
Llama-3-8B	Base	44.76	28.25	13.81	45.63	51.27
	Persona	48.39	21.74	14.92	41.86	55.37
	Reflection	50.68	19.52	17.84	39.47	58.26
	DA	49.17	18.63	16.75	36.72	59.81
	QD	53.41	14.28	19.36	33.31	64.67
	SFT-Retain	52.03	22.67	17.64	40.08	57.24
	Balanced SFT	**62.31**	20.54	20.29	35.91	61.82
	DPO	54.37	22.33	17.50	40.57	56.67
	SUL (GA)	57.24	**9.69**	**22.30**	**29.17**	**68.73**
	SUL (NPO)	56.58	12.87	20.47	31.84	65.91
	SUL (RMU)	56.45	11.62	19.13	36.38	61.35
Qwen-2.5-14B	Base	52.39	21.68	18.74	29.73	66.42
	Persona	54.81	17.96	20.35	27.94	68.63
	Reflection	56.27	15.88	21.74	25.69	71.82
	DA	55.64	15.21	20.96	24.57	72.31
	QD	58.92	11.76	23.43	18.94	78.56
	SFT-Retain	60.58	15.41	25.57	21.65	70.23
	Balanced SFT	57.31	18.75	20.31	23.47	69.36
	DPO	62.50	17.21	25.41	22.41	74.72
	SUL (GA)	**63.47**	**7.84**	**28.39**	**16.87**	**80.61**
	SUL (NPO)	62.19	9.43	26.92	18.21	78.97
	SUL (RMU)	63.38	8.77	28.04	19.69	71.73

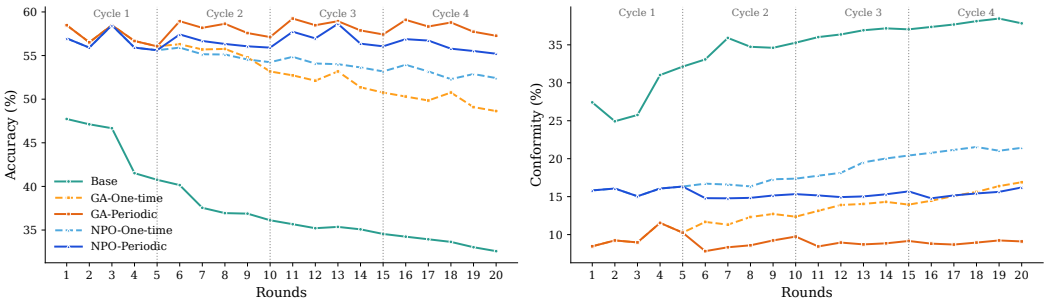

Figure 3: Accuracy and conformity rate across four interaction cycles under one-time and periodic unlearning. The x-axis contains 20 socially conditioned rounds, with every five rounds corresponding to one correction cycle.

induce distinct trade-offs between conformity reduction and contextual responsiveness, which we analyze in Figure 6. We additionally evaluate cross-protocol transfer from BENCHFORM to a psychology-inspired MMLU conformity setting in Appendix C, where GA and NPO continue to reduce conformity relative to the base model while improving corrective updating.

Periodic Unlearning. We further evaluate whether conformity can be corrected by a single model update or whether it reappears over repeated interactions. Figure 3 reports accuracy and conformity rate across four cycles, with each cycle containing five socially conditioned rounds. In the one-time setting, we first apply a single conformity-targeted unlearning update to obtain an updated model, and then keep this model fixed throughout all evaluation cycles. In the periodic setting, we start from the same one-time updated model, but further update the model after each cycle using newly observed conformity shifts from the training split. One-time unlearning reduces conformity in the early rounds, but the effect gradually weakens as the interaction progresses.

As shown in Figure 3, the conformity rate of one-time unlearning increases across later cycles, showing that its mitigation effect weakens under progressively longer interaction histories. In contrast, periodic unlearning maintains consistently low conformity across cycles, suggesting that repeated correction helps sustain the mitigation effect under extended interactions. A similar trend is observed in accuracy. While one-time unlearning improves accuracy initially, its performance gradually degrades as the interaction length increases. Periodic unlearning, on the other hand, maintains more stable and higher accuracy across cycles, indicating that it better preserves task-faithful responses under prolonged social influence. At the same time, different unlearning solvers exhibit distinct stability patterns. GA achieves the strongest immediate suppression of conformity, but shows slightly larger fluctuations. In contrast, NPO produces a more conservative but smoother trajectory, with smaller variance across cycles, suggesting better stability under repeated interactions.

Utility Preservation. A key concern is that reducing conformity may simply weaken the model rather than selectively suppressing conformity-driven errors. To examine this possibility, we evaluate general utility on ARC-Challenge (Clark et al., 2018) and HellaSwag (Zellers et al., 2019). These benchmarks are not used to construct the forget-retain data, allowing us to examine whether conformity-targeted unlearning preserves general reasoning ability beyond the conformity evaluation settings. The results in Table 2 show that GA and NPO do not substantially degrade general utility. In particular, the average utility drop remains within 1.5 percentage points across both models. The consistency of these results across two model families suggests that the observed utility preservation is not model-specific. In particular, NPO consistently yields smaller utility degradation than GA, indicating that more conservative updates better preserve general reasoning ability. Overall, GA provides stronger conformity suppression, while NPO offers a more conservative trade-off with smaller utility degradation and smoother periodic behavior.

Table 2: Utility preservation after unlearning.

Model	Method	ARC-C (%) ↑	HellaSwag (%) ↑	ΔUtil (%) ↓
Llama-3-8B	Base	50.34	56.50	–
	GA	49.32	54.68	1.42
	NPO	49.82	55.17	0.93
Qwen-2.5-14B	Base	59.73	74.86	–
	GA	58.68	73.91	1.00
	NPO	59.13	74.29	0.59

Effect of Number of Participants. We study how unlearning behaves under different levels of social pressure. Following the psychology-inspired conformity setting, we vary the total number of participants from 2 to 10. For each value of p, the model answers after observing $p-1$ simulated participants who unanimously provide the same incorrect answer. We report conformity level on the held-out evaluation split, averaged over expression styles. As shown in Figure 4, the base model becomes increasingly vulnerable as social pressure grows. Using the same aggregation setting as Table 1, the increasing conformity level of the base model shows that larger unanimous groups amplify conformity. Prompt-based mitigation reduces conformity compared with the base model, but its CL still increases as the majority becomes larger. In contrast, model-level unlearning maintains lower conformity across participant counts, with GA providing the strongest reduction throughout the scaling experiment. These results suggest that conformity-targeted unlearning improves robustness beyond a single fixed social-pressure setting.

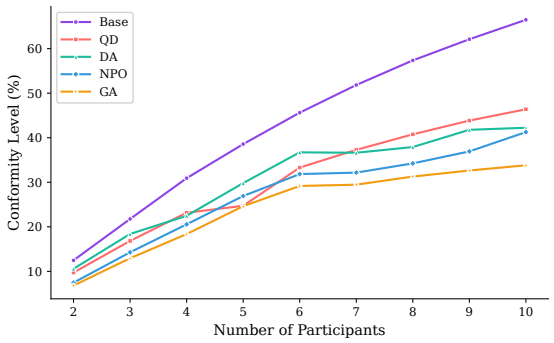

Figure 4: Conformity level versus number of participants (2–10) on Llama-3-8B under unanimous incorrect majority responses.

Does Unlearning Ignore Helpful Social Context? A possible concern is that suppressing conformity may cause the model to ignore all external opinions, including helpful ones. To examine this, we evaluate models on the Correct Guidance protocol of BENCHFORM in Table 3, where the social context provides task-faithful guidance rather than misleading majority pressure. This finding is further supported by the expanded model-level comparison in

Figure 6, where GA and NPO retain high corrective-update rates while substantially reducing susceptibility to misleading guidance. The performance drop under Correct Guidance remains small for both model families. On Llama-3-8B, GA and NPO only reduce Correct Guidance accuracy by 0.17 and 0.64 percentage points, respectively. On Qwen-2.5-14B, the drops are also limited to 1.19 and 0.61 percentage points. These results suggest that conformity-targeted unlearning does not simply remove sensitivity to social context. Instead, the model can still benefit from task-faithful guidance while becoming less likely to follow misleading social pressure.

Table 3: Performance on BENCHFORM Correct Guidance. ΔCG denotes the performance drop compared with the base model.

Model	Method	Correct Guidance Acc (%) ↑	ΔCG (%) ↓
Llama-3-8B	Base	79.13	–
	GA	78.96	0.17
	NPO	78.49	0.64
Qwen-2.5-14B	Base	83.27	–
	GA	82.08	1.19
	NPO	82.66	0.61

To further examine whether conformity-targeted unlearning broadly reduces contextual responsiveness, we evaluate the base model, GA, and NPO on CL-Bench. As shown in Figure 5, both GA and NPO achieve higher Task Solve Rate and Rubric Pass Rate than the base model, suggesting that conformity-targeted unlearning preserves broader in-context learning ability while reducing susceptibility to misleading social influence.

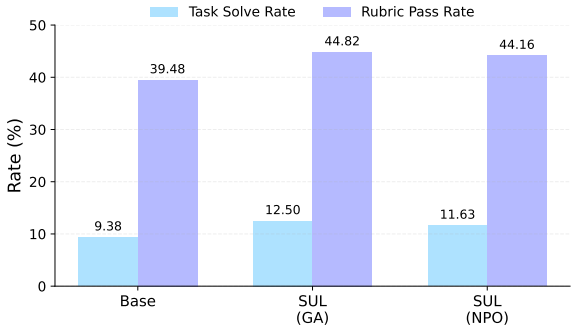

Figure 5: Broader in-context learning performance on CL-Bench. Both GA and NPO preserve strong contextual learning ability.

Comparison with Model-Level Interventions. Table 1 shows that the model-level objectives lead to substantially different conformity outcomes. To further distinguish whether these methods selectively resist misleading social influence or simply reduce responsiveness to social context, we analyze two complementary metrics on Llama-3-8B: Initial-Wrong Correction and Wrong-Guidance CR. Figure 6 compares these two behaviors across the base model, SFT-Retain, Balanced SFT, DPO, and the three SUL variants.

Balanced SFT maintains strong corrective updating but remains vulnerable to wrong guidance. DPO substantially reduces wrong-guidance conformity while retaining moderate corrective responsiveness. SFT-Retain provides an intermediate trade-off. In contrast, all three SUL variants substantially reduce wrong-guidance conformity while maintaining high corrective responsiveness, with NPO achieving the lowest Wrong-Guidance CR. This verifies the effectiveness of social-context-aware unlearning in mitigating conformity while preserving corrective responsiveness.

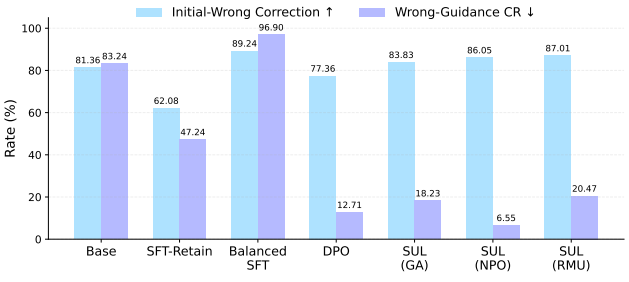

Figure 6: Selective responsiveness of model-level interventions on BENCHFORM using Llama-3-8B.

5 Conclusion

In this paper, we establish selective unlearning as a model-level approach for conformity mitigation, extending prior work beyond diagnosis and inference-time prompt intervention toward explicit behavioral correction. Rather than treating social influence as something that must only be filtered at inference time, we show that conformity-driven errors can be explicitly identified, converted into unlearning supervision, and suppressed through parameter-level intervention.

AI USE STATEMENT

In this work, we used generative AI tools for grammar checking, language polishing, and wording refinement. All AI-assisted edits were reviewed by the authors, who take full responsibility for the final content of the paper.

REPRODUCIBILITY STATEMENT

Code has been provided in an anonymized repository. We provide implementation details, data construction procedures, evaluation metrics, and hyperparameters in the appendix. We also provide anonymized source code for reproducing the experiments and use question-disjoint training and evaluation splits throughout our main experiments.

REFERENCES

Lucas Bourtoule, Varun Chandrasekaran, Christopher A Choquette-Choo, Hengrui Jia, Adelin Travers, Baiwu Zhang, David Lie, and Nicolas Papernot. Machine unlearning. In *2021 IEEE symposium on security and privacy (SP)*, pp. 141–159. IEEE, 2021.

Yinzhi Cao and Junfeng Yang. Towards making systems forget with machine unlearning. In *2015 IEEE symposium on security and privacy*, pp. 463–480. IEEE, 2015.

Chi-Min Chan, Weize Chen, Yusheng Su, Jianxuan Yu, Wei Xue, Shanghang Zhang, Jie Fu, and Zhiyuan Liu. Chateval: Towards better llm-based evaluators through multi-agent debate. *arXiv preprint arXiv:2308.07201*, 2023.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try arc, the ai2 reasoning challenge. *arXiv preprint arXiv:1803.05457*, 2018.

Julie C Coultas and Edwin JC Van Leeuwen. Conformity: Definitions, types, and evolutionary grounding. In *Evolutionary perspectives on social psychology*, pp. 189–202. Springer, 2015.

Wentao Deng, Jiahuan Pei, Zhiwei Xu, Zhaochun Ren, Zhumin Chen, and Pengjie Ren. Belief-calibrated multi-agent consensus seeking for complex nlp tasks. *Advances in Neural Information Processing Systems*, 38:116412–116445, 2026.

Shihan Dou, Ming Zhang, Zhangyue Yin, Chenhao Huang, Yujiong Shen, Junzhe Wang, Jiayi Chen, Yuchen Ni, Junjie Ye, Cheng Zhang, Huaibing Xie, Jianglu Hu, Shaolei Wang, Weichao Wang, Yanling Xiao, Yiting Liu, Zenan Xu, Zhen Guo, Pluto Zhou, Tao Gui, Zuxuan Wu, Xipeng Qiu, Qi Zhang, Xuanjing Huang, Yu-Gang Jiang, Di Wang, and Shunyu Yao. Cl-bench: A benchmark for context learning, 2026. URL https://arxiv.org/abs/2602.03587.

Yilun Du, Shuang Li, Antonio Torralba, Joshua B Tenenbaum, and Igor Mordatch. Improving factuality and reasoning in language models through multiagent debate. In *Forty-first international conference on machine learning*, 2024.

Ronen Eldan and Mark Russinovich. Who's harry potter? approximate unlearning for llms. 2023.

Zhibin Gou, Zhihong Shao, Yeyun Gong, Yelong Shen, Yujiu Yang, Nan Duan, and Weizhu Chen. Critic: Large language models can self-correct with tool-interactive critiquing. *arXiv preprint arXiv:2305.11738*, 2023.

Taicheng Guo, Xiuying Chen, Yaqi Wang, Ruidi Chang, Shichao Pei, Nitesh V Chawla, Olaf Wiest, and Xiangliang Zhang. Large language model based multi-agents: A survey of progress and challenges. *arXiv preprint arXiv:2402.01680*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. *arXiv preprint arXiv:2009.03300*, 2020.

Sirui Hong, Mingchen Zhuge, Jonathan Chen, Xiawu Zheng, Yuheng Cheng, Jinlin Wang, Ceyao Zhang, Zili Wang, Steven Ka Shing Yau, Zijuan Lin, et al. Metagpt: Meta programming for a multi-agent collaborative framework. In *The twelfth international conference on learning representations*, 2023.

Joel Jang, Dongkeun Yoon, Sohee Yang, Sungmin Cha, Moontae Lee, Lajanugen Logeswaran, and Minjoon Seo. Knowledge unlearning for mitigating privacy risks in language models. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 14389–14408, 2023.

Akbir Khan, John Hughes, Dan Valentine, Laura Ruis, Kshitij Sachan, Ansh Radhakrishnan, Edward Grefenstette, Samuel R Bowman, Tim Rocktäschel, and Ethan Perez. Debating with more persuasive llms leads to more truthful answers. *arXiv preprint arXiv:2402.06782*, 2024.

Chance Jiajie Li, Jiayi Wu, Zhenze Mo, Ao Qu, Yuhan Tang, Kaiya Ivy Zhao, Yulu Gan, Jie Fan, Jiangbo Yu, Jinhua Zhao, Paul Liang, Luis Alonso Pastor, and Kent Larson. Simulating society requires simulating thought. In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/abdf9f11e8d8206f5db7dce6b34cd292-Abstract-Position_Paper_Track.html.

Guohao Li, Hasan Hammoud, Hani Itani, Dmitrii Khizbullin, and Bernard Ghanem. Camel: Communicative agents for" mind" exploration of large language model society. *Advances in neural information processing systems*, 36:51991–52008, 2023.

Nathaniel Li, Alexander Pan, Anjali Gopal, Summer Yue, Daniel Berrios, Alice Gatti, Justin D. Li, Ann-Kathrin Dombrowski, Shashwat Goel, Long Phan, Gabriel Mukobi, Nathan Helm-Burger, Rassin Lababidi, Lennart Justen, Andrew B. Liu, Michael Chen, Isabelle Barrass, Oliver Zhang, Xiaoyuan Zhu, Rishub Tamirisa, Bhrugu Bharathi, Adam Khoja, Zhenqi Zhao, Ariel Herbert-Voss, Cort B. Breuer, Samuel Marks, Oam Patel, Andy Zou, Mantas Mazeika, Zifan Wang, Palash Oswal, Weiran Lin, Adam A. Hunt, Justin Tienken-Harder, Kevin Y. Shih, Kemper Talley, John Guan, Russell Kaplan, Ian Steneker, David Campbell, Brad Jokubaitis, Alex Levinson, Jean Wang, William Qian, Kallol Krishna Karmakar, Steven Basart, Stephen Fitz, Mindy Levine, Ponnurangam Kumaraguru, Uday Tupakula, Vijay Varadharajan, Ruoyu Wang, Yan Shoshitaishvili, Jimmy Ba, Kevin M. Esvelt, Alexandr Wang, and Dan Hendrycks. The wmdp benchmark: Measuring and reducing malicious use with unlearning, 2024. URL https://arxiv.org/abs/2403.03218.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. Self-refine: Iterative refinement with self-feedback. *Advances in neural information processing systems*, 36:46534–46594, 2023.

Pratyush Maini, Zhili Feng, Avi Schwarzschild, Zachary C Lipton, and J Zico Kolter. Tofu: A task of fictitious unlearning for llms. *arXiv preprint arXiv:2401.06121*, 2024.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 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)*, pp. 9802–9822, 2023.

Jens Mattke, Christian Maier, Lea Reis, and Tim Weitzel. Herd behavior in social media: the role of facebook likes, strength of ties, and expertise. *Information & Management*, 57(8):103370, 2020.

Joon Sung Park, Joseph O'Brien, Carrie Jun Cai, Meredith Ringel Morris, Percy Liang, and Michael S Bernstein. Generative agents: Interactive simulacra of human behavior. In *Proceedings of the 36th annual acm symposium on user interface software and technology*, pp. 1–22, 2023.

Rafael Rafailov, Archit Sharma, Eric Mitchell, Stefano Ermon, Christopher D. Manning, and Chelsea Finn. Direct preference optimization: Your language model is secretly a reward model, 2024. URL https://arxiv.org/abs/2305.18290.

Noah Shinn, Federico Cassano, Ashwin Gopinath, Karthik Narasimhan, and Shunyu Yao. Reflexion: Language agents with verbal reinforcement learning. *Advances in neural information processing systems*, 36:8634–8652, 2023.

Maojia Song, Tej Deep Pala, Ruiwen Zhou, Weisheng Jin, Amir Zadeh, Chuan Li, Dorien Herremans, and Soujanya Poria. LLMs can't handle peer pressure: Crumbling under multi-agent social interactions. *arXiv preprint arXiv:2508.18321*, 2025.

Sophie Sowden, Sofia Koletsi, Eva Lymberopoulos, Elisabeta Militaru, Caroline Catmur, and Geoffrey Bird. Quantifying compliance and acceptance through public and private social conformity. *Consciousness and cognition*, 65:359–367, 2018.

Mirac Suzgun, Nathan Scales, Nathanael Schärli, Sebastian Gehrmann, Yi Tay, Hyung Won Chung, Aakanksha Chowdhery, Quoc Le, Ed Chi, Denny Zhou, et al. Challenging Big-Bench tasks and whether chain-of-thought can solve them. In *Findings of the Association for Computational Linguistics: ACL 2023*, pp. 13003–13051, 2023.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. Commonsenseqa: A question answering challenge targeting commonsense knowledge. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pp. 4149–4158, 2019.

Xuezhi Wang, Jason Wei, Dale Schuurmans, Quoc Le, Ed Chi, Sharan Narang, Aakanksha Chowdhery, and Denny Zhou. Self-consistency improves chain of thought reasoning in language models. *arXiv preprint arXiv:2203.11171*, 2022.

Zhiyuan Weng, Guikun Chen, and Wenguan Wang. Do as we do, not as you think: the conformity of large language models. *arXiv preprint arXiv:2501.13381*, 2025.

Addison J. Wu, Ryan Liu, Kerem Oktar, Theodore R. Sumers, and Tom Griffiths. Are large language models sensitive to the motives behind communication? In Danielle Belgrave, Cheng Zhang, Laura N. Montoya, Hsuan-Tien Lin, Razvan Pascanu, Piotr Koniusz, Marzyeh Ghassemi, Nancy Chen, Iván Vladimir Meza Ruíz, and Arturo Loaiza-Bonilla (eds.), *Advances in Neural Information Processing Systems 38: Annual Conference on Neural Information Processing Systems 2025, NeurIPS 2025, San Diego, CA, USA, December 2-7, 2025 / Mexico City, Mexico, November 30 - December 5, 2025*, 2025. URL http://papers.nips.cc/paper_files/paper/2025/hash/e5fb1d7e7c1541eb0b6132b7839baf34-Abstract-Conference.html.

Qingyun Wu, Gagan Bansal, Jieyu Zhang, Yiran Wu, Beibin Li, Erkang Zhu, Li Jiang, Xiaoyun Zhang, Shaokun Zhang, Jiale Liu, et al. Autogen: Enabling next-gen llm applications via multi-agent conversations. In *First conference on language modeling*, 2024.

Kai Xiong, Xiao Ding, Yixin Cao, Ting Liu, and Bing Qin. Examining inter-consistency of large language models collaboration: An in-depth analysis via debate. In *Findings of the Association for Computational Linguistics: EMNLP 2023*, pp. 7572–7590, 2023.

Haoming Xu, Ningyuan Zhao, Yunzhi Yao, Weihong Xu, Hongru Wang, Xinle Deng, Shumin Deng, Jeff Z. Pan, Huajun Chen, and Ningyu Zhang. Illusions of confidence? diagnosing LLM truthfulness via neighborhood consistency. In Maria Liakata, Viviane P. Moreira, Jiajun Zhang, and David Jurgens (eds.), *Proceedings of the 64th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 4432–4457, San Diego, California, United States, July 2026a. Association for Computational Linguistics. ISBN 979-8-89176-390-6. doi: 10.18653/v1/2026.acl-long.203. URL https://aclanthology.org/2026.acl-long.203/.

Naen Xu, Hengyu An, Shuo Shi, Jinghuai Zhang, Chunyi Zhou, Changjiang Li, Tianyu Du, Zhihui Fu, Jun Wang, and Shouling Ji. When agents "misremember" collectively: Exploring the mandela effect in llm-based multi-agent systems, 2026b. URL https://arxiv.org/abs/2602.00428.

Shunyu Yao, Dian Yu, Jeffrey Zhao, Izhak Shafran, Tom Griffiths, Yuan Cao, and Karthik Narasimhan. Tree of thoughts: Deliberate problem solving with large language models. *Advances in neural information processing systems*, 36:11809–11822, 2023.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pp. 4791–4800, 2019.

Jintian Zhang, Xin Xu, Ningyu Zhang, Ruibo Liu, Bryan Hooi, and Shumin Deng. Exploring collaboration mechanisms for llm agents: A social psychology view. *arXiv preprint arXiv:2310.02124*, 2023.

Ruiqi Zhang, Licong Lin, Yu Bai, and Song Mei. Negative preference optimization: From catastrophic collapse to effective unlearning. *arXiv preprint arXiv:2404.05868*, 2024.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Gong, et al. Promptrobust: Towards evaluating the robustness of large language models on adversarial prompts. In *Proceedings of the 1st ACM workshop on large AI systems and models with privacy and safety analysis*, pp. 57–68, 2023.

Xiaochen Zhu, Caiqi Zhang, Tom Stafford, Nigel Collier, and Andreas Vlachos. Conformity in large language models. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pp. 3854–3872, 2025.

A Implementation Details

A.1 Model and Decoding Setup

We evaluate our framework on Llama-3-8B and Qwen-2.5-14B. For all conformity evaluations, we use the same decoding setting across the base model, prompt-level baselines, and parameter-update methods. Unless otherwise specified, we use greedy decoding with temperature set to 0, which avoids sampling variance and ensures that answer shifts are caused by changes in social context or model parameters rather than decoding randomness. We additionally evaluate robustness to stochastic decoding with $T = 0.7$ in Appendix D.1. This avoids sampling variance and ensures that answer shifts are caused by changes in social context or model parameters rather than decoding randomness. All experiments were run on NVIDIA L40S GPUs with approximately 46GB memory per GPU and bfloat16 precision. LoRA-based unlearning substantially reduces the memory cost compared with full-parameter fine-tuning. A single LoRA-based unlearning run for Llama-3-8B typically finishes within one GPU day on our setup, while Qwen-2.5-14B requires longer runtime due to its larger model size. The cost of periodic unlearning scales approximately linearly with the number of correction cycles.

For multiple-choice tasks, we extract the model's final answer option from the generated response. For open-ended QA tasks, we normalize generated answers and compare them with the reference answer following the evaluation protocol of the corresponding benchmark.

A.2 Data Splits

To avoid data leakage, we use disjoint question sets for constructing unlearning data and for evaluation. For each conformity-oriented setting, we split examples by question identifiers rather than by interaction traces, so that different social contexts or rounds generated from the same question do not appear in both the unlearning and evaluation splits. Interaction traces used to construct $\mathcal{D}_{\mathrm{forget}}$ and $\mathcal{D}_{\mathrm{retain}}$ are collected only from the unlearning split. All reported results are computed on a held-out evaluation split that is never used to construct forget-retain supervision.

In the periodic setting, each correction cycle updates the model using newly observed conformity shifts from the unlearning split. The evaluation curves are computed on a fixed held-out split across all cycles. Therefore, the periodic results measure generalization to unseen questions rather than online adaptation to the same evaluation examples.

A.3 Confidence Score and Base Set Construction

We use the confidence score s_i^{raw} to identify examples where the model produces a reliable independent response before social context is introduced. In our implementation, s_i^{raw} is computed as the normalized log-probability of the generated answer span:

$$s_i^{\text{raw}} = \frac{1}{|y_i|} \sum_{j=1}^{|y_i|} \log p_{\theta_0}(y_{i,j} \mid x_i, y_{i,<j}). \tag{15}$$

We include an example in the base set only when the independent response matches the reference answer and its confidence score satisfies $s_i^{\text{raw}} \geq \tau$. We use $\tau = -0.9$ as the default confidence threshold in our experiments. Confidence filtering is used only for constructing forget-retain supervision, so that unlearning is based on reliable independent responses. Reported evaluation metrics follow the original benchmark definitions and do not apply this additional confidence filter. Reference answers and confidence filtering are used only during trace construction; after the model update, deployment does not require access to reference answers.

A.4 Sensitivity to Confidence Threshold

We additionally examine the sensitivity of conformity-targeted unlearning to the confidence threshold τ. The threshold controls which independently correct responses are considered sufficiently reliable for constructing conformity-shift supervision. Lower thresholds admit more training examples, while more restrictive thresholds retain only higher-confidence responses.

Table 4: Sensitivity to the confidence threshold τ on a controlled BENCHFORM evaluation subset. All other training and evaluation settings are kept fixed.

τ	Raw Acc ↑	BF-Acc ↑	BF-CR ↓	BF-IR ↑
-1.4	62.80	55.90	13.60	20.80
-1.2	63.30	56.80	11.50	21.70
-1.0	63.70	57.20	10.20	22.20
-0.9	63.90	57.40	9.80	22.50
-0.8	62.90	56.10	12.40	20.90

Performance remains stable over a moderate range of confidence thresholds. The default value $\tau = -0.9$ lies within this stable region and achieves a strong balance between accuracy, conformity reduction, and independence.

A.5 Unlearning Setup

We instantiate the generic unlearning operator $\mathcal{U}$ with Gradient Ascent (GA), Negative Preference Optimization (NPO), and Representation Misdirection for Unlearning (RMU).

For GA, the model is trained to reduce the likelihood of forget responses while retaining reference-supported behavior. Specifically, we maximize the loss on $\mathcal{D}_{\text{forget}}$ and minimize the loss on $\mathcal{D}_{\text{retain}}$:

$$\mathcal{L}_{\text{GA}} = \mathcal{L}_{\text{retain}} - \lambda_{\text{f}} \mathcal{L}_{\text{forget}}, \tag{16}$$

where λ_{f} controls the strength of forgetting.

For NPO, we follow the standard negative preference optimization formulation and use the initial model as the reference model. The forget responses are treated as dispreferred outputs, while retain examples are used to preserve task-faithful behavior. We use the same forget and retain sets as GA for a fair comparison.

For Direct Preference Optimization (DPO), we treat the reference-supported response as the preferred output and the socially shifted response as the dispreferred output under the same socially conditioned input. We use the initial model as the reference model and set the DPO coefficient to $\beta = 0.1$.

Balanced SFT is a stronger supervised fine-tuning baseline that jointly optimizes on three types of task-faithful supervision: resilience examples, corrective-guidance examples, and cooperative-guidance examples. For the expanded comparison, the training set contains 388 resilience examples,

Table 5: Implementation hyperparameters for conformity-targeted unlearning and evaluation.

Hyperparameter	**Value**
Base models	Llama-3-8B, Qwen-2.5-14B
Optimizer	AdamW
Learning rate	1×10^{-5}
Per-device batch size	2
Gradient accumulation steps	8
Effective batch size	16
Maximum sequence length	1024
LoRA rank	16
LoRA alpha	32
LoRA dropout	0.05
GA weight λ_{f}	0.3
Retain weight	1.0
GA loss cap	3.0
NPO beta	0.1
Confidence threshold τ	-0.9
Evaluation rounds	1–5
Majority number	6
Maximum new tokens	64
Evaluation protocols	Raw, Trust, Doubt

525 corrective-guidance examples, and 708 cooperative-guidance examples, for a total of 1,621 examples.

For Representation Misdirection for Unlearning (RMU), we apply the representation-level objective to the same 388 conformity-derived forget-retain pairs used by GA, NPO, and DPO. We intervene at transformer layer 15 of Llama-3-8B and use a fixed random steering direction with magnitude 30. The retain representation preservation coefficient is set to 50. We apply the RMU objective over all non-padding tokens in the sequence.

A.6 Retain-only Fine-tuning

For the retain-only fine-tuning ablation, SFT-Retain uses the same trainable parameter configuration and optimization setup as GA and NPO, but only optimizes on $\mathcal{D}_{\mathrm{retain}}$. It does not use $\mathcal{D}_{\mathrm{forget}}$ and therefore does not explicitly suppress socially shifted responses. This baseline controls for whether parameter updates on reference-supported outputs alone are sufficient to reduce conformity.

In the expanded model-level comparison, SFT-Retain is trained on the same 388 reference-supported examples associated with the identified conformity shifts, while Balanced SFT additionally incorporates corrective- and cooperative-guidance supervision.

A.7 Periodic Unlearning

For periodic unlearning, we run $T = 4$ correction cycles, and each cycle contains five socially conditioned rounds. The one-time setting applies one conformity-targeted update before the first evaluation cycle and keeps the updated model fixed for all cycles. The periodic setting starts from the same updated model, but continues to update the model after each cycle using newly observed conformity shifts from the unlearning split.

To reduce the risk of compounding forgetting, we maintain a fixed anchor retain set $\mathcal{D}_{\mathrm{anchor}}$ constructed from the initial unlearning split. This anchor set is included in every correction cycle, ensuring that initially reference-supported cases remain protected even if their confidence decreases in later cycles. We use the anchor retain set as a conservative design choice; isolating its effect with a dedicated ablation is left to future work. The one-time variants require only a single model update, while periodic refreshes are intended for settings where offline model maintenance is feasible.

A.8 Long-Horizon Comparison with Prompt-Based Mitigation

We additionally compare periodically refreshed GA and NPO with QD, the strongest prompt-level mitigation baseline in our main evaluation, under extended interaction histories. This controlled analysis uses the same evaluation examples for all methods and reports performance at representative interaction rounds.

Table 6: Long-horizon comparison between QD and periodic model-level correction on a controlled BENCHFORM evaluation setting.

	Accuracy ↑				**Conformity Rate ↓**			
Round	Base	QD	GA	NPO	Base	QD	GA	NPO
1	47.73	54.82	58.48	56.97	27.42	18.63	8.46	15.82
5	40.76	49.37	56.06	55.61	32.13	21.58	10.26	16.33
10	36.12	46.48	57.12	55.91	35.28	23.74	9.74	15.33
15	34.55	44.93	57.42	56.06	37.04	25.18	9.16	15.69
20	32.58	43.67	57.27	55.21	37.83	26.41	9.10	16.21

QD consistently improves over the base model, but its accuracy gradually decreases and its conformity rate increases as the interaction history grows. In contrast, the model-level methods remain substantially more stable across the evaluated rounds. These results indicate that prompt-level mitigation is helpful but remains sensitive to accumulated social influence, while model-level correction provides more persistent mitigation over extended interactions.

B Metric Details

B.1 BENCHFORM Metrics

For BENCHFORM, we report Accuracy (BF-Acc), Conformity Rate (BF-CR), and Independence Rate (BF-IR).

Accuracy BF-Acc measures the average accuracy over Trust and Doubt in the main BENCHFORM evaluation:

$$\text{BF-Acc} = \frac{1}{|\mathcal{P}_{\text{main}}|} \sum_{P \in \mathcal{P}_{\text{main}}} \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \mathbf{1}\left[\hat{y}_i^P = y_i^{\text{ref}}\right], \tag{17}$$

where $\mathcal{E}$ denotes the held-out evaluation set and $\mathcal{P}_{\text{main}} = \{\text{Trust}, \text{Doubt}\}$. Raw is used to define the Raw-correct subset, and Correct Guidance is evaluated separately.

Conformity Rate Following BENCHFORM, we compute conformity rate for each protocol P and interaction round r as

$$\text{BF-CR}^{P,r} = \frac{\sum_{i \in \mathcal{I}_{\text{raw}}} \mathbf{1}\left[\hat{y}_i^{P,r} \neq y_i^{\text{ref}}\right]}{|\mathcal{I}_{\text{raw}}|}. \tag{18}$$

where $\mathcal{I}_{\text{raw}}$ denotes examples answered correctly in the Raw setting. In the main table, BF-CR is averaged over all evaluated rounds under Trust and Doubt.

Independence Rate Following BENCHFORM, BF-IR measures the proportion of Raw-correct examples that remain correct throughout the long-term interaction trajectory under both Trust and Doubt:

$$\text{BF-IR} = \frac{\sum_{i \in \mathcal{I}_{\text{raw}}} \mathbf{1}\left[\bigwedge_{r=1}^{R} \left(\hat{y}_i^{T,r} = y_i^{\text{ref}} \ \wedge \ \hat{y}_i^{D,r} = y_i^{\text{ref}}\right)\right]}{|\mathcal{I}_{\text{raw}}|}. \tag{19}$$

Higher BF-IR indicates that the model preserves correct answers throughout the full long-term interaction trajectory. BF-CR and BF-IR operate at different temporal granularities: BF-CR is computed for individual protocol-round evaluations and then averaged, whereas BF-IR requires an initially correct example to remain correct across all evaluated rounds under both Trust and Doubt. Therefore, a low average BF-CR can coexist with a substantially lower BF-IR.

Raw accuracy is used only to determine $\mathcal{I}_{\text{raw}}$, the denominator of BF-CR and BF-IR; BF-Acc in Table 1 is computed over Trust and Doubt rather than Raw.

B.2 Psychology-inspired Conformity Metrics

For the psychology-inspired setting, we report Conformity Level (CL) and Resistance Level (RL), following prior LLM conformity work.

CL measures how often the model follows a unanimous majority answer. In the main evaluation, the unanimous majority answer is incorrect, so lower CL indicates weaker harmful conformity:

$$\text{CL} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \mathbf{1}\left[\hat{y}_i^{\text{social}} = y_i^{\text{majority}}\right]. \tag{20}$$

RL measures how often the model resists the misleading majority and preserves its independent answer:

$$\text{RL} = \frac{1}{|\mathcal{E}|} \sum_{i \in \mathcal{E}} \mathbf{1}\left[\hat{y}_i^{\text{social}} = \hat{y}_i^{\text{raw}}\right]. \tag{21}$$

Higher RL indicates stronger resistance to misleading social pressure. Since the model may also generate answers that are neither the unanimous majority answer nor its original independent answer, CL and RL are not necessarily complementary.

Unless otherwise specified, the main CL and RL results are reported with the number of participants fixed to six. In the participant-scaling experiment, we vary the number of participants from 2 to 10 and report CL averaged over expression styles. Expression styles refer to different phrasings of the simulated majority responses, including plain, neutral, confident, and uncertain forms.

B.3 Utility Drop

We evaluate general utility on ARC-Challenge and HellaSwag. These benchmarks are not used to construct $\mathcal{D}_{\text{forget}}$ or $\mathcal{D}_{\text{retain}}$. For each unlearned model, ΔUtil is computed as the average performance drop compared with the corresponding base model:

$$\Delta\text{Util} = \frac{1}{|\mathcal{B}|} \sum_{b \in \mathcal{B}} \left(\text{Score}_{\text{base}}^{(b)} - \text{Score}_{\text{method}}^{(b)}\right), \tag{22}$$

where $\mathcal{B} = \{\text{ARC-Challenge}, \text{HellaSwag}\}$. Lower ΔUtil indicates better utility preservation.

B.4 Correct Guidance Evaluation

Correct Guidance is used to test whether unlearning causes the model to ignore helpful social context. In this protocol, peer responses provide task-faithful guidance rather than misleading majority pressure. We report Correct Guidance accuracy and the accuracy drop compared with the base model:

$$\Delta\text{CG} = \text{Acc}_{\text{base}}^{\text{CG}} - \text{Acc}_{\text{method}}^{\text{CG}}. \tag{23}$$

A small ΔCG indicates that the model still benefits from helpful social information after unlearning.

B.5 Corrective Updating and Wrong-Guidance Transfer

Initial-Wrong Correction. To distinguish overall performance under correct guidance from the model's ability to revise an initially incorrect answer, we additionally report Initial-Wrong Correction. Let $\mathcal{I}_{\text{wrong}}$ denote the examples answered incorrectly in the Raw setting. We compute

$$\text{IW-Corr} = \frac{\sum_{i \in \mathcal{I}_{\text{wrong}}} \mathbf{1}[\hat{y}_i^{\text{CG}} = y_i^{\text{ref}}]}{|\mathcal{I}_{\text{wrong}}|}. \tag{24}$$

This metric measures how often task-faithful social information successfully corrects an initially wrong model response. It differs from Correct Guidance accuracy, which is computed over the full evaluation set.

Wrong-Guidance Conformity Rate. We also evaluate transfer to a held-out Wrong Guidance protocol. Wrong-Guidance CR is the fraction of Raw-correct examples that become incorrect after exposure to misleading guidance:

$$\text{WG-CR} = \frac{\sum_{i \in \mathcal{I}_{\text{raw}}} \mathbf{1}[\hat{y}_i^{\text{WG}} \neq y_i^{\text{ref}}]}{|\mathcal{I}_{\text{raw}}|}. \tag{25}$$

C Cross-Protocol Transfer Evaluation

We additionally evaluate whether behavior learned from BENCHFORM transfers to a different conformity protocol. We use MMLU with a psychology-inspired unanimous-majority interaction. The model first answers the question independently and is then exposed either to an incorrect unanimous majority or to a correct unanimous majority. The cross-protocol results reveal substantial

Table 7: Cross-protocol transfer from BENCHFORM to a psychology-inspired MMLU conformity setting.

Method	Raw Acc ↑	Wrong Acc ↑	CL ↓	RL ↑	Correct Acc ↑	IW-Corr ↑
Base	61.89	62.50	13.04	80.43	59.72	15.38
QD	61.89	63.11	2.17	91.30	36.11	3.85
Balanced SFT	62.57	54.17	17.91	78.26	90.28	78.57
DPO	62.50	63.89	2.26	86.67	43.06	7.69
GA	62.51	65.94	4.35	92.82	76.39	53.86
NPO	63.71	64.07	6.52	89.13	72.22	46.15

differences in how the methods respond to misleading and corrective social information. QD and DPO achieve the lowest conformity levels under an incorrect majority, but their Initial-Wrong Correction rates remain low, suggesting that strong resistance to social influence can also reduce responsiveness to corrective context. Balanced SFT exhibits the opposite behavior, showing strong corrective updating but increased susceptibility to an incorrect majority. In contrast, GA and NPO reduce conformity relative to the base model while substantially improving Initial-Wrong Correction. This pattern suggests that conformity-targeted unlearning can induce more selective responsiveness to social context rather than simply suppressing contextual influence altogether.

D Broader In-Context Learning Evaluation

We evaluate broader in-context learning using CL-Bench, which measures whether a model can infer and execute task-specific rules from contextual demonstrations and instructions. We evaluate only tasks that are successfully processed for all compared methods. The main results are reported in Figure 5.

D.1 Robustness to Stochastic Decoding

Our main experiments use deterministic decoding with temperature $T = 0$. To examine whether the observed conformity mitigation depends on greedy decoding, we additionally evaluate the methods under stochastic decoding with $T = 0.7$. We report mean and standard deviation across three decoding seeds.

Stochastic decoding introduces moderate variance, but the overall mitigation pattern remains stable. At $T = 0.7$, GA and NPO maintain similar BF-Accuracy to their deterministic counterparts while continuing to substantially reduce conformity relative to the base model. GA achieves a slightly lower BF-CR under stochastic decoding, whereas NPO shows a modest increase in BF-CR. Both methods continue to outperform the base model in BF-CR and BF-IR. These results suggest that the observed conformity mitigation is not restricted to greedy decoding.

Table 8: Robustness to stochastic decoding on the BENCHFORM evaluation set. Results at $T = 0.7$ report mean $\pm$ standard deviation across three decoding seeds.

Method	Temp.	BF-Acc ↑	BF-CR ↓	BF-IR ↑
Base	0.0	44.76	28.25	13.81
Base	0.7	44.44 ± 1.90	27.67 ± 1.40	13.20 ± 1.40
GA	0.0	57.24	9.69	22.30
GA	0.7	56.64 ± 1.30	8.64 ± 1.40	19.96 ± 1.70
NPO	0.0	56.58	12.87	20.47
NPO	0.7	56.06 ± 1.10	14.41 ± 1.30	19.10 ± 1.50

D.2 Seed Stability

To examine the stability of parameter-update methods, we run the main BENCHFORM experiment with three random seeds. For each run, we keep the data split, evaluation prompts, decoding strategy, and hyperparameters fixed, and only change the random seed used for model updating, including LoRA initialization and data ordering during unlearning. All evaluations use deterministic decoding with temperature set to 0.

Table 9 reports the mean and standard deviation across three seeds. The standard deviations are small compared with the performance gaps between SFT-Retain and conformity-targeted unlearning methods, suggesting that the observed improvements are not caused by a single favorable optimization run.

Table 9: Seed stability of parameter-update methods on Llama-3-8B. We report mean and standard deviation across three random seeds.

Method	BF-Acc ↑	BF-CR ↓	BF-IR ↑
SFT-Retain	52.03 ± 0.41	22.67 ± 0.58	17.64 ± 0.37
GA	57.24 ± 0.52	9.69 ± 0.44	22.30 ± 0.61
NPO	56.58 ± 0.47	12.87 ± 0.51	20.47 ± 0.49

E Additional Notes and Limitations

Our evaluation focuses on reasoning-oriented QA settings where reference answers are available or where high-confidence independent responses can be used as reference-supported behavior. This design allows us to identify socially induced response shifts and construct forget-retain supervision. However, our method relies on reference-supported behavior to identify harmful conformity shifts, which is straightforward in labeled QA benchmarks but harder in open-ended dialogue where correctness can be ambiguous. Our main BENCHFORM analysis focuses on Trust and Doubt because they represent long-term socially conditioned interactions.

We evaluate general reasoning utility on ARC-Challenge and HellaSwag. Future work should further examine broader conversational abilities, such as instruction following and multi-turn dialogue quality.

F Broader Impacts

This work aims to improve the reliability of LLMs under misleading social influence, which may benefit collaborative AI systems and multi-agent decision support. At the same time, overcorrecting conformity could reduce a model's ability to use helpful social context.

We therefore evaluate both overall performance under Correct Guidance and the ability to correct initially wrong responses. Our results show that this trade-off depends strongly on the intervention

objective. In particular, GA and NPO substantially reduce susceptibility to misleading guidance while retaining higher corrective responsiveness than DPO. This suggests that conformity-targeted unlearning can provide a selective model-level mechanism for improving resistance to misleading social influence without simply suppressing all contextual updating.