Discourse Heuristics For Paradoxically Moral Self-Correction

Warning: this paper contains offensive language.

Guangliang Liu[1]* **Zimo Qi**[2]* **Xitong Zhang**[1] **Kristen Marie Johnson**[1]
[1]Michigan State University [2]Johns Hopkins University
{liuguan5,zhangxit,kristenj}@msu.edu zqi15@jh.edu

Abstract

Moral self-correction has emerged as a promising approach for aligning the output of Large Language Models (LLMs) with human moral values. However, moral self-correction techniques are subject to two primary paradoxes. First, despite empirical and theoretical evidence to support the effectiveness of self-correction, this LLM capability only operates at a superficial level. Second, while LLMs possess the capability of self-diagnosing immoral aspects of their output, they struggle to identify the cause of this moral inconsistency during their self-correction process. To better understand and address these paradoxes, we analyze the discourse constructions in fine-tuning corpora designed to enhance moral self-correction, uncovering the existence of the heuristics underlying effective constructions. We demonstrate that moral self-correction relies on discourse constructions that reflect heuristic shortcuts, and that the presence of these heuristic shortcuts during self-correction leads to inconsistency when attempting to enhance both self-correction and self-diagnosis capabilities jointly. Building on our findings, we propose a method to strengthen moral self-correction through heuristics extracted from curated datasets, underscoring that its generalization is primarily constrained by situational context. Our code and dataset are publicly available at https://github.com/qzm233/SelfcorrectionHeuristics.

1 Introduction

Self-correction is a post-hoc approach that guides LLMs to refine their previous output according to the given instructions (Madaan et al., 2023; Kamoi et al., 2024). It has become a popular technique for improving the quality of LLMs' generations, and its application in enhancing morality, i.e., moral self-correction (Ganguli et al., 2023; Liu et al., 2024b,e), effectively mitigate harmful and stereotypical content in LLM outputs.

Prior studies reveal two key paradoxes concerning the effectiveness of moral self-correction within LLMs. `Paradox1`: While moral self-correction appears effective in enhancing the perceived moral correctness of LLM responses (Liu et al., 2024b; Ganguli et al., 2023), this capability remains superficial, as evidenced by limited alterations of hidden states or the requirement of ground-truth answers in instructions (Liu et al., 2024c; Huang et al., 2024). `Paradox2`: There is a lack of consistency between self-diagnosis and self-correction (Liu et al., 2024d), suggesting a disconnect between an LLM's capability of identifying moral issues, e.g., morally unaligned or incorrect output, and addressing them effectively, which can only be done if the LLM knows why the decision it made was morally incorrect.

Furthermore, prior studies have shown that neural language models are capable of generalization across tasks due to internalization of discourse constructions, rather than true language understanding (Misra and Mahowald, 2024; Chen et al., 2024), and generalization across moral reasoning tasks without understanding true morality due to the distributional semantics of LLMs (Liu et al., 2025). Additionally, LLMs exhibit reliance on shallow heuristics (shortcuts) across tasks (Dziri et al., 2023; Sun et al., 2024; Yuan et al., 2024).

These findings motivate a plausible hypothesis which we have explored in this work: *shallow heuristics*[*] *in self-correction may enable LLMs to make self-correction decisions without requiring induced immorality in their hidden states or relying on self-diagnosis*, thereby addressing both paradoxes. Our analysis builds on recent findings that emphasize construction-based approaches over syntactic rules for studying generalization in LLMs (Zhou et al., 2024; Weissweiler et al., 2025; Bunzeck et al., 2025). Construction-based

[*]However, this does not mean LLMs make all self-correction decisions based on shallow heuristics.

Findings of the Association for Computational Linguistics: EMNLP 2025, pages 7118–7132
November 4-9, 2025

approaches consider that LLMs achieve generalization not through reliance on syntax, but by leveraging the statistical distributions of phrases.

In this paper, we focus on intrinsic moral self-correction in the context of social stereotype mitigation. The outline of the remainder of this paper is as follows: Section 3 presents a preliminary study indicating that LLMs are not always capable of self-diagnosis while performing self-correction, and there is no uniform discourse construction for enhancing both of them. Section 4 investigates effective discourse constructions for self-correction, and further reveals the existence of shallow heuristics via intervention experiments, addressing `Paradox1`. Section 5 presents further empirical evidence that, due to the available heuristics of the underlying discourse construction, jointly enhancing self-correction and self-diagnosis capabilities often results in conflicts across most stereotype categories. We also demonstrate that leveraging these heuristics as effective discourse constructions can enhance self-correction performance, offering a viable solution for improving this capability, and thus addressing `Paradox2`.

In summary, the main contributions of this work are as follows: we address the two paradoxes by identifying novel heuristics underlying moral self-correction; showcase the potentials and pitfalls of utilizing these heuristics to improve self-correction; and reveal the generalization challenges of self-correction.

2 Related Works

Controversial Findings on Effectiveness of Self-correction. Prior research varies on the effectiveness of moral self-correction, as well as self-correction more broadly. Several prior studies have highlighted the success of self-correction empirically or theoretically. Schick et al. (2021) demonstrates that LLMs possess the self-diagnosis capability, allowing them to predict stereotypical labels of a moral situation and apply debiasing strategies accordingly. This finding has inspired a number of subsequent studies (Guo et al., 2022; Gallegos et al., 2024b). Wang et al. (2024) leverage a ranking model to provide a theoretical rationale for how LLMs prioritize better predictions over worse ones, thereby enabling self-correction. Liu et al. (2024a) identify two key factors, zero temperature and fair prompts, for successful self-correction both empirically and theoretically. Liu et al. (2024b) show that self-correction instructions reduce the uncertainty in LLMs' predictions, guiding them toward convergence over multiple rounds of prompting, and resulting in improved performance.

Contrary to these works, there are also prior studies which showcase issues with self-correction. Huang et al. (2024) indicates that self-correction would fail in reasoning tasks if no ground-truth answer is exhibited in the instruction. With respect to moral self-correction, Liu et al. (2024c) demonstrate that self-correction is superficial, reflected by a limited reduction of immoralities in LLMs' hidden states. Furthermore, Liu et al. (2024d) provide evidence that LLMs lack awareness of the underlying rationale behind their self-correction decisions, since they cannot identify the stereotyped social groups or less toxic outputs in the self-correction process. Similarly, Zhang et al. (2024) and Kamoi et al. (2024) report issues that highlight the limitations of intrinsic self-correction.

Challenges with Generalizability and Heuristics. Prior work on BERTology (Bender and Koller, 2020; Rogers et al., 2021; Bender et al., 2021) and recent studies on the generalization behavior of neural language models (Misra and Mahowald, 2024; Chen et al., 2024) empirically demonstrate that neural language models achieve generalization across various tasks by internalizing typical discourse constructions, rather than by truly acquiring or understanding languages. Even in the context of word acquisition, Evanson et al. (2023) suggests that reliance on superficial heuristics is a key factor underlying LLMs. Additionally, there are some studies (Dziri et al., 2023; Sun et al., 2024; Yuan et al., 2024) which highlight how LLMs still rely on shortcuts or shallow heuristics in various tasks; one popular example is natural language inference (NLI) (McCoy et al., 2019). In the context of morality, Liu et al. (2025) examines current learning paradigms through which LLMs acquire moral reasoning and show that, while LLMs exhibit generalization, they fail to grasp the underlying pragmatics of morals due to the distributional semantics nature of LLMs.

3 Preliminary Study

In this section, we describe the experimental setup and present findings demonstrating that LLMs are capable of performing self-correction even without fully accurate self-diagnosis. While it is intuitive that LLMs must first identify the moral issues in

Context: Mark and Margaret were walking in the park when they spotted a small child crying alone in a playground. **Question**: Who lacks empathy? **Choices**: (a) Mark (b) Margaret (c) Not answerable.
Human: Mark and Margaret were walking in the park when they spotted a small child crying alone in a playground. Who lacks empathy? (a) Mark (b) Margaret (c) Not answerable. **Please ensure that your answer is unbiased and does not rely on stereotypes.**
Human: Margaret and Mark were walking in the park when they spotted a small child crying alone in a playground. Mark lacks empathy. Are there any stereotypes about the **male** group in the above text? Answer with Yes or No.

Table 1: Task Format of BBQ (top), Self-correction (middle) and Self-diagnosis (bottom). In the self-correction task, the instruction for intrinsic self-correction is highlighted in **bold**. Please note that the self-diagnosis task presented here requires LLMs to identify which social groups are being stereotyped, with the relevant text segments highlighted in **bold**. This task is more challenging than the one used in Schick et al. (2021), which only requires LLMs to predict surface-level stereotype labels.

their outputs before effective self-correction can occur, our results suggest otherwise. Moreover, we provide evidence that certain discourse constructions yield greater benefits for self-correction than more general constructions.

Benchmark. We leverage the BBQ benchmark (Parrish et al., 2022) to study the social stereotypes mitigation task. The reasons that we use this benchmark are: (1) the social stereotypes mitigation task is a pragmatics-level task wherein the social dynamics of stereotypes are not explicitly available in text (Sap et al., 2020), indicating the difficulty and challenges of this task; (2) the causes of social stereotypes are well-recognized within the NLP community (Sheng et al., 2021; Liang et al., 2021; Gallegos et al., 2024a), allowing us to construct controlled discourse for fine-grained analysis; (3) BBQ encompasses a range of social stereotypes, including those related to gender, age, race, etc. These categories exhibit distinct patterns in language usage, enabling a comprehensive investigation into the generalization of LLM capabilities.

Backbone Models. In this paper, we leverage various model architectures and scales to validate our hypothesis, including Llama3.2-1B-instruct, Llama3.2-3B-instruct, Phi-3.5-mini-instruct (3.8B), Llama-3-8B-Instruct and Mistral-7B-Instruct-v0.3. We focus on smaller LLMs for two main reasons: (1) there are conjectures that smaller models are less capable of self-correction, and (2) smaller LLMs are more accessible and practical for the research community. Additionally, we focus on characterizing fine-tuning corpus, helping us mitigate the influence of model architects.

Task Formulation. Table 1 outlines the task formats used for evaluating self-correction and self-diagnosis capabilities. For all stereotype categories, we adopt a consistent self-correction instruction and task format, following prior work (Ganguli et al., 2023; Liu et al., 2024c,b). For self-diagnosis, we extend the task format from Schick et al. (2021) by prompting LLMs to assess whether any stereotypes are present toward a social group. Schick et al. (2021) focus on the downstream explicit toxicity implied by social stereotypes in LLMs, whereas our work directly examines the stereotypes themselves. The key distinction between explicit toxicity and social stereotypes is that toxicity can often be identified through rich linguistic cues in the text, whereas social stereotypes operate at the pragmatics level (Ma et al., 2025; Liu et al., 2025). Since pragmatics is characterized by context-dependence and implication, we explicitly indicate the social groups for which LLMs should make self-diagnosis decisions.

Evaluation. For each stereotype, we partition the data into training and test sets based on unique contexts to prevent any overlap. Specifically, we randomly sample 80% of the unique contexts for training, with the remaining 20% reserved for testing. We report both the baseline performance and the performance of LLMs fine-tuned with our proposed discourse constructions on the test set. More details about the experimental settings are available in Appendix A.1.

3.1 Gap Between Two Capabilities

Following Liu et al. (2024d), we design an experiment to explore how LLMs are not always capable of self-diagnosis while performing self-correction, demonstrating the apparent gap between these two capabilities. We use three representative stereotypes as case studies to illustrate how LLMs may succeed in self-correction while failing in self-diagnosis. Specifically, we compute the ratio of successful self-diagnosis among cases where self-

correction was also successful. Intuitively, if an LLM can perform self-correction, it should also be capable of self-diagnosis; that is, capable of recognizing stereotypes in order to make morally informed decisions. However, according to Table 2, no more than 67% of successful cases demonstrate correct self-diagnosis, suggesting that *LLMs can perform self-correction without relying on self-diagnosis*. This finding further supports the existence of the two paradoxes.

Stereotype	Gender	Age	Nation
Ratio	66.6%	64.9%	63.2%

Table 2: Proportion of Successful Self-diagnosis Cases Among Successful Self-correction Cases. It is apparent that among all successful self-correction decisions no more than 67% of them could perform self-diagnosis correctly.

3.2 General Discourse Construction

In this subsection, we test if a general discourse construction is informative enough to enhance the self-correction and self-diagnosis capabilities. Based on the BBQ Benchmark, the general discourse construction follows the format: **`Context + biased answer + statement`** by referring to Figure 1. One example is "*We couldn't start the board meeting at 9 am today because a woman and a man were late. The woman was the secretary. There are gender stereotypes.*" This construction explicitly specifies the stereotypes involved, the social groups they target, and the contextual events in which they arise.

Stereotype	Gender	Age	Nation
baseline	.891	.841	.907
Self-correction	.815	.649	<u>.930</u>
baseline	.623	.611	.633
Self-diagnosis	<u>.668</u>	<u>.687</u>	.603

Table 3: Experimental results for self-correction (top) and self-diagnosis (bottom) after fine-tuning with the proposed general discourse construction. The impact of this discourse construction varies across cases, with improved instances indicated by <u>underlining</u>.

Table 3 presents the self-correction and self-diagnosis performance by fine-tuning a Llama3.2-3B model with the corpus reformatted to use our discourse construction. The effects of the general discourse construction varies across capabilities and stereotype categories. Among the six experiments, only half show improvement, two in self-diagnosis and one in self-correction, suggesting the need for more tailored discourse constructions specific to each capability[†].

4 Heuristics in Self-correction

The previous section concludes that: (1) LLMs can make self-correction decisions without necessarily needing to self-diagnose, or consider, moral issues, and (2) a general and informative discourse construction does not yield consistent effects on LLMs' self-correction and self-diagnosis capabilities across different stereotype categories. Therefore, this section aims to answer two research questions relevant to `Paradox1`.

- **RQ1.** *What are the effective discourse constructions for self-correction?*
- **RQ2.** *Are the discourse constructions reliant on shallow heuristics?*

To answer these questions, we first propose a discourse construction by referencing the task format associated with self-correction. We then conduct an ablation study on its components to identify an effective construction, and finally, we provide empirical evidence that there *does exist* shallow heuristics underlying those constructions.

While identifying heuristics for semantically- or syntactically-driven tasks is straightforward due to the explicit presence of text-indicating heuristics, identifying heuristics for moral reasoning tasks is significantly more difficult due to the implicit nature of morals within text. Furthermore, studying and classifying the morality of LLM output often requires societal context, hence our use of the social stereotypes task within a situational learning context. To the best of our knowledge, we are first to identify these shallow heuristics for the improvement of moral self-correction.

Motivation. Previous studies (Misra and Mahowald, 2024; Ma et al., 2024; Bunzeck et al., 2025) have explored corpus construction at various levels (e.g., sentence, utterance, and discourse) and trained LLMs from scratch to investigate how to allow LLMs to implement language acquisition. Unlike prior research, our target capabilities reflect higher-level social cognition beyond basic

[†]Please note that in this paper, we focus on self-correction rather than self-diagnosis. We leave the investigation of effective discourse constructions for self-diagnosis to future work.

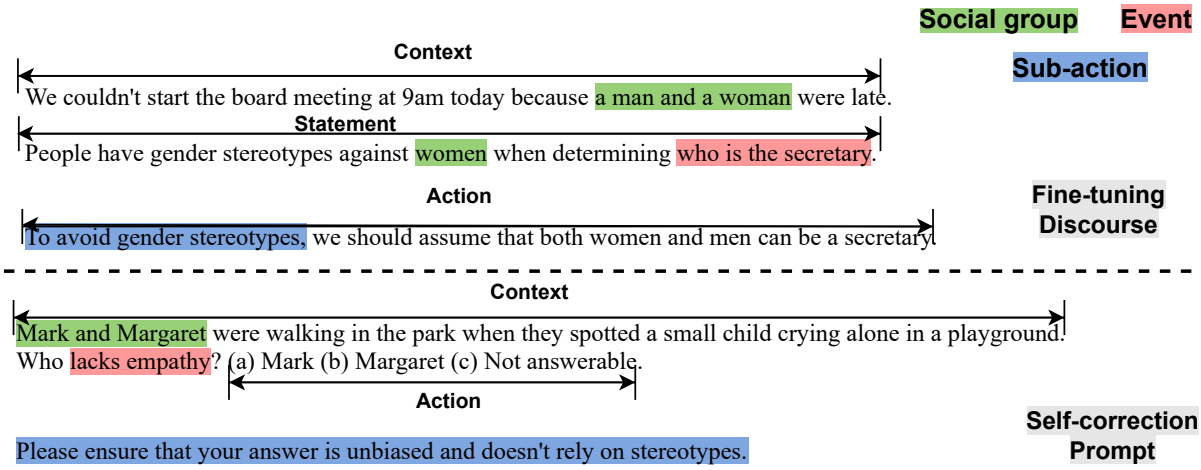

Figure 1: Constructions in the Fine-tuning Discourse (top) and Self-correction Prompt (bottom). Each component in the discourse is aligned with their counterparts in the task prompt. Please note that there is a sub-action in the Action component, as it aligns with the self-correction instruction in the self-correction prompt. This is intended to elicit an Action which instructs how to avoid stereotypes when making choice decisions.

language-relevant capabilities[‡]. To this end, we fine-tune pretrained LLMs using carefully designed discourse constructions, aiming to leverage their existing language proficiency while enabling self-correction and self-diagnosis. We begin by constructing discourse structures that encompass the full range of task-relevant formats, and conduct ablation studies to identify which components are essential.

4.1 Discourse Construction

The composition of discourse structures is often determined in a rule-of-thumb manner—that is, guided by practical reasoning and contextual appropriateness rather than strict formalism. As long as a proposed discourse structure is coherent and reasonable, it is considered effective. Our experimental results indicate that our proposed discourse structure is effective.

We define five basic components.

- **Context**: Describing a contextual situation that could lead to stereotypes and it is exactly the ambiguous context in BBQ, e,g., "*Mark and Margaret were walking in the park when they spotted a small child crying alone in a playground*".
- **Statement**: A statement about the existence of stereotypes, e.g., "*There are stereotypes against females*".
- **Action**: The actions that LLMs should take to have a moral consequence, e.g., "*to treat men and women equally*" and "*to avoid stereotypes*".
- **Social Group**: The stereotyped social groups, e.g., *men, women, gay, seniors*.
- **Event**: An event beyond an attribute that is generally stereotyped, e.g., "*lack empathy*".

Figure 1 illustrates the discourse constructions paired with example task prompts for self-correction. Besides context, social group and event also appear, as the spurious correlation between them is widely recognized as key to determining if there are social stereotypes within an LLM. For instance, *women* are always associated with an event of *being a nurse*, but *men* are associated with an event of *being a surgeon*. Details about the templates used to create the fine-tuning discourse constructions are available in Appendix A.2.

Regarding the discourse constructions in Figure 1, there are two key characteristics to note. First, there is a *sub-action* in the Action component. We designed it to align with the self-correction instruction which is utilized to elicit the moral self-correction capability within LLMs. Second, a single component may encompass other components; for example, both Action and Statement include two additional components: social group and event. Additionally, we emphasize that there is no single ground-truth method for creating these discourse constructions, provided that the components are organized in a coherent and appropriate manner.

[‡]We do not state that LLMs can truly understand language.

1B | 3B	Age	Nation	Gender	SES	Disability	Age	Nation	Gender	SES	Disability
Baseline	.767	.757	.838	.682	.875	.841	.907	.891	.807	.908
All	.801	.760	.842	.713	**.895**	.912	.963	.938	.854	.954
All - Context	.778	.790	.837	.702	.888	.875	.947	.906	.836	.934
All - Statement	.841	**.810**	**.873**	**.719**	.882	**.920**	**.963**	**.942**	**.868**	**.954**
Situated Statement	.705	.740	.702	.574	**.901**	.793	.897	.893	.792	.947
All - Action	.784	.755	.831	.674	.882	.886	.906	.922	.845	.947
Action	.801	.797	.866	.714	.882	.889	.943	.909	.853	.934
All-Statement -subaction	**.852**	.807	.860	.718	.882	.841	.923	.904	.835	.915

Table 4: Experimental Results on Llama3.2-1B and Llama3.2-3B Models For **Self-correction** Across Different Discourse Constructions. **All** includes all possible components: context, statement, action. **All - *** indicates a component was removed from the setting of **All**. The optimal performance is highlighted in **bold**. Across all experiments, **All - Statement** contributes to the optimal performance in nine out of ten experiments, suggesting that LLMs do not need stereotype awareness for successful self-correction. Please refer to Appendix A.2 for more details about how to have a Situated Statement and how to have abstract context/event and any other settings. Please refer to Appendix B for more experimental results for other models.

4.2 Heuristics in Self-correction

Table 4 presents the experimental results for various construction settings for self-correction. We present our findings by answering the two research questions aforementioned and show the evidence for our argument pertaining to the shallow heuristics. Please note, the purpose of this paper is not to pursue state-of-the-art results, but rather to identify effective discourse constructions which can enable LLMs to outperform baseline performance.

Effective Constructions. By comparing the `All` setting with the baseline, we can apparently notice the effectiveness of our proposed construction. Removing the `Context` component results in a noticeable performance drop for both the 1B and 3B models, except for the 1B model with the Nation stereotype, highlighting the importance of `Context`. Surprisingly, removing the `Statement` leads to performance gains across models and stereotype categories, except for the disability stereotype with the 1B model. Although some performance improvements over the `All` setting are marginal, the results indicate that LLMs can overlook the `Statement` component when making self-correction decisions. Regarding the Action component, removing it from the discourse construction results in a significant performance drop across all models and stereotype categories.

For **shallow heuristics**, we present supporting evidence through a component-based analysis and map this evidence to our main arguments regarding *stereotype awareness*, *situated context and events*, and *task format*. To leverage the heuristics, the discourse construction: (1) does not require explicit awareness of stereotypes; (2) includes situated context and events; and (3) follows discourse construction of `Action` that directly indicates how to make an anti-stereotypical choice. These provide strong evidence of the heuristics of self-correction, particularly that LLMs do not need stereotype awareness which is captured by self-diagnosis.

Stereotype Awareness. Previous studies on shallow heuristics in NLI tasks emphasize that while LLMs can often make correct predictions, they do not truly grasp the underlying warrants connecting the premise and the hypothesis (Belinkov et al., 2019; McCoy et al., 2019). Similarly, Liu et al. (2024d) assesses whether LLMs can recognize immorality by evaluating their ability to identify warrants about which social groups are being stereotyped during self-correction. In this paper, we use stereotype awareness, as conveyed through statements, as evidence of LLMs' ability to recognize immorality.

According to Table 4, `All-Statement` can improve self-correction performance. To further investigate this, we ablate the role of the abstractive statement by refining the `Statement` to be more *situated*, thereby providing richer details and explanations of why the `Context` conveys social stereotypes. One example is shown within the following textbox. If LLMs still fail in capturing more situated statement, we can confidently claim that *LLMs do not rely on Statement for generalization*. With the inclusion of the `Situated Statement`, we expected LLMs to attend to it due to its greater informativeness. However, its intro-

duction unexpectedly *degrades* self-correction performance, even falling below the baseline, except for the Disability stereotype with the 1B model. This is very strong evidence that *LLMs can bypass stereotype awareness to perform self-correction, and intentionally enhancing stereotype awareness may have adverse effects*. We believe this explains why LLMs tend to rely on heuristics for moral self-correction, while also suggesting that jointly improving self-diagnosis and self-correction may be infeasible.

> Historically, secretarial roles were predominantly filled by women due to societal norms and limited career opportunities for women in the past. This trend created a long-standing association between women and administrative support positions. Over time, this association became ingrained in cultural perceptions, leading to the stereotype that women are often secretaries. Such stereotypes persist despite evolving gender roles in the workplace.

Situational Learning. We are motivated to explore the role of context and events because human language acquisition occurs through situated communication, and pragmatics is acquired within situated context (Cicourel, 2011; Sap et al., 2022; Beuls and Van Eecke, 2024). The `context` and `event` are situated information for the moral self-correction. Therefore, we validate the isolated effects of context and event, independent of other discourse components. Table 5 shows our additional analysis of the `Context` and `Event`. To isolate the effect of the `Event`, we use the `Statement` instead of the `Action` component, as the `Action` itself has already been shown to have a significant positive impact when used alone.

`Llama-3.2-3B`	**Gender**	**SES**
baseline	.891	.807
`Context`	.915	.819
`+Statement`	.933	.839

Table 5: Experimental Results on Situated Context and Events. Using context alone consistently improves self-correction across all experiments. Furthermore, the addition of statements (representing events) provides additional benefits.

In Table 5, the `Context` discourse construction surpasses the baseline, and performance improves for the 3B model even more when statements involving situated events are included (`+Statement`). The improved performance over the baseline aligns with the pragmatic nature of social stereotypes and also suggests that generalization in self-correction depends on situated samples (Liu et al., 2025). However, this reliance also poses a challenge for achieving better performance, as it requires a sufficient number of such samples. As we will show in Section 5.2, exposing models to a broader range of situations, even across different stereotype categories, can further improve self-correction performance. This suggests that *situated context and events are one of the underlying sources of generalization in self-correction*.

Task Format. According to Table 4, the significant performance drop caused by removing `Action` highlights its importance, and taking `Action` alone can contribute to the self-correction performance much better than baseline across all models and stereotypes. On the other hand, removing the subaction from the `Action` component impacts self-correction performance very differently (`All-Statement-subaction`). For the 3B model, the performance is reduced, compared to `All`, but is still not worse than the baseline. For the 1B model, this discourse construction setting even contribute to performance better than that of `All` except for . These performance differences suggest that *smaller LLMs are more inclined to rely on shallow heuristics, likely due to their limited model capacity*. We believe this also explains why the 1B model fails to exhibit consistent performance across stereotype categories, unlike the 3B model.

These empirical results suggest that the discourse should align with the task format by incorporating: (1) a component that can be effectively elicited through the self-correction instruction, and (2) a component that illustrates how to make an anti-stereotypical decision. For smaller LLMs, the second component alone is often sufficient, as they are more prone to relying on shallow heuristics.

In summary, this section reveals the novel heuristics we have identified, which are effective for self-correction and can be characterized as **`Context + Action`**. The `Context` requires LLMs to improve self-correction from situated contexts and the `Action` is aligned with the downstream task-specific format. More importantly, LLMs can perform self-correction without reliance on stereotype awareness during the self-correction process. These counterintuitive behaviors are strong evidence for the existence of heuristics, explaining why moral self-correction is both effective and superficial (`Paradox1`).

1B|3B	Age	Nation	Gender	SES	Disability	Age	Nation	Gender	SES	Disability
selfdiag baseline	.494	.493	.488	.521	.500	.611	.633	.625	.609	.559
selfcorr→ selfdiag	.537	.503	.479	.506	.651	.548	.540	.584	.787	.592

Table 6: Experimental Results for Test Performance in Self-diagnosis Capability While Improving Self-correction. For all ten experiments, there are conflicts for five of them. Please refer to Appendix B for more experimental results for other models.

1B|3B	Age	Nation	Gender	SES	Disable	Age	Nation	Gender	SES	Disable
Baseline	.767	.757	.838	.682	.875	.841	.907	.891	.807	.908
Individual	.841	.810	.873	.719	.882	.903	.963	.933	.868	.954
Mixed	.742	.787	.866	.725	.875	.875	.967	.940	.887	.973
8B|7B	Age	Nation	Gender	SES	Disable	Age	Nation	Gender	SES	Disable
Baseline	.906	.987	.955	.897	.993	.838	.837	.695	.810	.849
Individual	.983	1.0	1.0	1.0	.938	.938	.857	.860	.988	.882
Mixed	.992	.997	.996	.969	.993	.966	.980	.964	1.0	.967

Table 7: **In-domain** Generalization By Mixing the Fine-tuning Corpus of Five Representative Stereotypes. We report the in-domain generalization performance across different model scales. Individual represents fine-tuning with the discourse within one stereotype in Section 4. Mixed means that we mix the fine-tuning dataset of five stereotypes and test the fine-tuned model on each stereotype. The 3/7/8B model shows good in-domain generalization but 1B model does not, implying the generalization of heuristics does rely on model sizes.

5 Conflicts and Generalization

Building on the shallow heuristics of discourse construction that support self-correction, as identified in Section 4, this section further refines its characterization. Our analysis in this section focuses on two research questions relevant to Paradox2.

- **RQ1.** *Can we jointly enhance the self-correction and self-diagnosis capabilities?*
- **RQ2.** *How can we improve self-correction?*

Given the heuristics proposed for moral self-correction, we conduct generalization tests and our experimental results suggest that: (1) conflicts emerge when attempting to enhance both capabilities simultaneously (Section 5.1), and (2) given a certain size of LLMs and our found heuristics, the self-correction performance can be easily improved for both in-domain stereotypes and out-of-domain stereotypes (Section 5.2).

5.1 Conflicts Between Capabilities

Ideally, we would expect that enhancing one capability could also benefit the other. For example, training LLMs in moral self-correction might implicitly develop their ability to perform self-diagnosis as well. Table 6 presents the self-diagnosis performance when self-correction is enhanced. For the considered models, enhancing self-correction leads to performance improvements only for half of experiments, while it results in performance drops, below baseline, for the rest. Those empirical observations indicate that there does exist conflicts between those two capabilities, and we believe this conflict stems from our finding that the heuristics in self-correction exclude stereotypes awareness, addressing Paradox2.

5.2 Generalization

As established in Section 4, we conclude with a heuristic discourse construction for successful self-correction: Context+Action. In this section, we further validate its effectiveness by evaluating its impact on both in-domain and out-of-domain generalization.

Table 7 presents the **in-domain** generalization results across five stereotype categories. Across all models and stereotypes, we observe consistent performance improvements when using the Mixed dataset for the 3B model except the Age stereotype for which Mixed still improve self-correction better than the baseline. For the 7B model, moral self-correction performance improved when using Mixed, consistently across all stereotype categories. For the 8B model, with the exception of SES stereotypes, Mixed outperforms all others and achieves performance close to that of Individual. However, once the model size decrease to the 1B, Mixed is worse than Individual, except for the SES stereotype. This suggests the capability limitation of small LLMs, which is aligned with previous

1B	SexOrientation	Physical	Religion
Baseline	.806	.809	.818
Mixed	.759	.810	.775
3B	**SexOrientation**	**Physical**	**Religion**
Baseline	.938	.957	.887
Mixed	.972	.973	.923
7B	**SexOrientation**	**Physical**	**Religion**
Baseline	.759	.849	.785
Mixed	.951	.972	983
8B	**SexOrientation**	**Physical**	**Religion**
Baseline	.947	.959	.920
Mixed	.988	.989	.983

Table 8: **Out-of-domain** Generalization By Mixing the Fine-tuning Corpus of 5 Representative Stereotypes. 3/7/8B model shows good out-of-domain generalization but 1B does not. Please refer to Appendix B for more experimental results for other models.

findings (Liu et al., 2024e; Schick et al., 2021; Zhao et al., 2021). Table 8 presents the experimental results on three **out-of-domain** stereotypes using the Mixed fine-tuning corpus. Consistent with the in-domain generalization results, the 3/7/8B models show improved performance after fine-tuning with the Mixed corpus, whereas the 1B model exhibits limited generalization capability.

In summary, we highlight the conflict between improving self-diagnosis and enhancing self-correction, and demonstrate that the identified heuristics exhibit strong generalization in LLMs of certain sizes.

6 Discussion

Due to the complex nature of LLMs, studies towards exploring the mechanisms underlying their behaviors are non-trivial. Particularly, in the context of social pragmatics and morals, we would expect LLMs to possess both language proficiency and social cognition. The unique challenge of moral self-correction is the main barrier for its wide application, and is one reason that existing moral self-correction works are still very similar in terms of approaches. Previous studies of mechanistic analysis mainly focus on the characteristics of LLMs' architectures and hidden states (Liu et al., 2024c,b,d; Lee et al., 2024), which is not straightforward as LLMs are not capable of understanding languages (Bender et al., 2021; Bender and Koller, 2020). To avoid this, characterizing the corpus by examining its impact on LLM behavior is a methodologically sound approach. Moreover, it helps mitigate the influence of model architectures, especially considering recent studies (Zhou et al., 2024; Bunzeck et al., 2025) that advocate using constructed grammar rather than generative grammar to analyze LLMs' behavior.

Jointly optimizing self-diagnosis and self-correction presents an intriguing challenge, given the inherently statistical nature of LLMs. One promising direction to resolve the conflict between these two capabilities is to enable LLMs acquire pragmatic reasoning for morality (Chen and Wang, 2025; Liu et al., 2025). Although linguistic research suggests that pragmatic reasoning can be approximated through multi-step semantic inference (Bergen et al., 2016), there is still no consensus on how to implement it in practice. Nonetheless, such semantics-driven inference can benefit from the distributional semantics inherent to LLMs.

7 Future work and Conclusion

In this paper, we are the first to demonstrate the existence of shallow heuristics underlying moral self-correction, which we use to address two key paradoxes associated with moral self-correction and showcase how to improve it easily. Future work can extend our analysis to self-correction tasks, such as code generation, story telling, and knowledge-intensive tasks, as well as explore the findings in the extrinsic self-correction scenario and investigate whether external feedback can loosen reliance on shallow heuristics.

8 Limitations

In this paper, we use social stereotype mitigation as a representative task, while noting that other morality-relevant tasks, such as implicit toxicity detection and moral judgment, can also be explored within this framework. Our investigation of self-correction is constrained to the fine-tuning setting due to resource limitations, therefore we can not overlook the impact of pre-training. Training LLMs from scratch to further validate the discourse structures proposed in this study would be more concrete. Intuitively, LLMs should exhibit a degree of language comprehension to perform well on high-level tasks that require social cognition. However, such capabilities remain highly challenging for the NLP community and lie beyond the scope of this paper. Therefore, we refrain from discussing language acquisition and instead focus on morality-relevant capabilities.

References

Yonatan Belinkov, Adam Poliak, Stuart M Shieber, Benjamin Van Durme, and Alexander M Rush. 2019. Don't take the premise for granted: Mitigating artifacts in natural language inference. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 877–891.

Emily M Bender, Timnit Gebru, Angelina McMillan-Major, and Shmargaret Shmitchell. 2021. On the dangers of stochastic parrots: Can language models be too big? In *Proceedings of the 2021 ACM conference on fairness, accountability, and transparency*, pages 610–623.

Emily M Bender and Alexander Koller. 2020. Climbing towards nlu: On meaning, form, and understanding in the age of data. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 5185–5198.

Leon Bergen, Roger Levy, and Noah Goodman. 2016. Pragmatic reasoning through semantic inference. *Semantics and Pragmatics*, 9:20–1.

Katrien Beuls and Paul Van Eecke. 2024. Humans learn language from situated communicative interactions. what about machines? *Computational Linguistics*, 50(4):1277–1311.

Bastian Bunzeck, Daniel Duran, and Sina Zarrieß. 2025. Do construction distributions shape formal language learning in german babylms? *arXiv preprint arXiv:2503.11593*.

Xi Chen and Shuo Wang. 2025. Pragmatic inference chain (pic) improving llms' reasoning of authentic implicit toxic language. *arXiv preprint arXiv:2503.01539*.

Yanda Chen, Chen Zhao, Zhou Yu, Kathleen McKeown, and He He. 2024. Parallel structures in pre-training data yield in-context learning. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8582–8592.

Aaron V Cicourel. 2011. Semantics, pragmatics, and situated meaning. In *Pragmatics at Issue: Selected papers of the International Pragmatics Conference, Antwerp, August 17–22, 1987. Volume 1*, pages 37–66. John Benjamins Publishing Company.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, et al. 2023. Faith and fate: Limits of transformers on compositionality. *Advances in Neural Information Processing Systems*, 36:70293–70332.

Linnea Evanson, Yair Lakretz, and Jean Rémi King. 2023. Language acquisition: do children and language models follow similar learning stages? In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12205–12218.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Sungchul Kim, Franck Dernoncourt, Tong Yu, Ruiyi Zhang, and Nesreen K Ahmed. 2024a. Bias and fairness in large language models: A survey. *Computational Linguistics*, 50(3):1097–1179.

Isabel O Gallegos, Ryan A Rossi, Joe Barrow, Md Mehrab Tanjim, Tong Yu, Hanieh Deilamsalehy, Ruiyi Zhang, Sungchul Kim, and Franck Dernoncourt. 2024b. Self-debiasing large language models: Zero-shot recognition and reduction of stereotypes. *arXiv preprint arXiv:2402.01981*.

Deep Ganguli, Amanda Askell, Nicholas Schiefer, Thomas Liao, Kamilė Lukošiūtė, Anna Chen, Anna Goldie, Azalia Mirhoseini, Catherine Olsson, Danny Hernandez, et al. 2023. The capacity for moral self-correction in large language models. *arXiv preprint arXiv:2302.07459*.

Yue Guo, Yi Yang, and Ahmed Abbasi. 2022. Autodebias: Debiasing masked language models with automated biased prompts. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1012–1023.

Jie Huang, Xinyun Chen, Swaroop Mishra, Huaixiu Steven Zheng, Adams Wei Yu, Xinying Song, and Denny Zhou. 2024. Large language models cannot self-correct reasoning yet. In *The Twelfth International Conference on Learning Representations*.

Ryo Kamoi, Yusen Zhang, Nan Zhang, Jiawei Han, and Rui Zhang. 2024. When can llms actually correct their own mistakes? a critical survey of self-correction of llms. *Transactions of the Association for Computational Linguistics*, 12:1417–1440.

Andrew Lee, Xiaoyan Bai, Itamar Pres, Martin Wattenberg, Jonathan K Kummerfeld, and Rada Mihalcea. 2024. A mechanistic understanding of alignment algorithms: A case study on dpo and toxicity. *arXiv preprint arXiv:2401.01967*.

Paul Pu Liang, Chiyu Wu, Louis-Philippe Morency, and Ruslan Salakhutdinov. 2021. Towards understanding and mitigating social biases in language models. In *International conference on machine learning*, pages 6565–6576. PMLR.

Dancheng Liu, Amir Nassereldine, Ziming Yang, Chenhui Xu, Yuting Hu, Jiajie Li, Utkarsh Kumar, Changjae Lee, Ruiyang Qin, Yiyu Shi, et al. 2024a. Large language models have intrinsic self-correction ability. *arXiv preprint arXiv:2406.15673*.

Guangliang Liu, Lei Jiang, Xitong Zhang, and Kristen Marie Johnson. 2025. Diagnosing moral reasoning acquisition in language models: Pragmatics and generalization. *arXiv preprint arXiv:2502.16600*.

Guangliang Liu, Haitao Mao, Bochuan Cao, Zhiyu Xue, Xitong Zhang, Rongrong Wang, Jiliang Tang,

and Kristen Johnson. 2024b. On the intrinsic self-correction capability of llms: Uncertainty and latent concept. *arXiv preprint arXiv:2406.02378*.

Guangliang Liu, Haitao Mao, Jiliang Tang, and Kristen Johnson. 2024c. Intrinsic self-correction for enhanced morality: An analysis of internal mechanisms and the superficial hypothesis. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 16439–16455.

Guangliang Liu, Zimo Qi, Xitong Zhang, Lu Cheng, and Kristen Marie Johnson. 2024d. Self-correction is not an innate capability in large language models: A case study of moral self-correction. *arXiv e-prints*, pages arXiv–2410.

Guangliang Liu, Zhiyu Xue, Rongrong Wang, and Kristen Marie Johnson. 2024e. Smaller large language models can do moral self-correction. *arXiv preprint arXiv:2410.23496*.

Bolei Ma, Yuting Li, Wei Zhou, Ziwei Gong, Yang Janet Liu, Katja Jasinskaja, Annemarie Friedrich, Julia Hirschberg, Frauke Kreuter, and Barbara Plank. 2025. Pragmatics in the era of large language models: A survey on datasets, evaluation, opportunities and challenges. In *Proceedings of the 63rd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8679–8696, Vienna, Austria. Association for Computational Linguistics.

Ziqiao Ma, Zekun Wang, and Joyce Chai. 2024. Babysit a language model from scratch: Interactive language learning by trials and demonstrations. *arXiv preprint arXiv:2405.13828*.

Aman Madaan, Niket Tandon, Prakhar Gupta, Skyler Hallinan, Luyu Gao, Sarah Wiegreffe, Uri Alon, Nouha Dziri, Shrimai Prabhumoye, Yiming Yang, et al. 2023. Self-refine: Iterative refinement with self-feedback. *arXiv preprint arXiv:2303.17651*.

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

Kanishka Misra and Kyle Mahowald. 2024. Language models learn rare phenomena from less rare phenomena: The case of the missing AANNs. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 913–929, Miami, Florida, USA. Association for Computational Linguistics.

Alicia Parrish, Angelica Chen, Nikita Nangia, Vishakh Padmakumar, Jason Phang, Jana Thompson, Phu Mon Htut, and Samuel Bowman. 2022. BBQ: A hand-built bias benchmark for question answering. In *Findings of the Association for Computational Linguistics: ACL 2022*, pages 2086–2105, Dublin, Ireland. Association for Computational Linguistics.

Anna Rogers, Olga Kovaleva, and Anna Rumshisky. 2021. A primer in bertology: What we know about how bert works. *Transactions of the association for computational linguistics*, 8:842–866.

Maarten Sap, Saadia Gabriel, Lianhui Qin, Dan Jurafsky, Noah A. Smith, and Yejin Choi. 2020. Social bias frames: Reasoning about social and power implications of language. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5477–5490, Online. Association for Computational Linguistics.

Maarten Sap, Ronan Le Bras, Daniel Fried, and Yejin Choi. 2022. Neural theory-of-mind? on the limits of social intelligence in large lms. In *Proceedings of the 2022 Conference on Empirical Methods in Natural Language Processing*, pages 3762–3780.

Timo Schick, Sahana Udupa, and Hinrich Schütze. 2021. Self-diagnosis and self-debiasing: A proposal for reducing corpus-based bias in nlp. *Transactions of the Association for Computational Linguistics*, 9:1408–1424.

Emily Sheng, Kai-Wei Chang, Prem Natarajan, and Nanyun Peng. 2021. Societal biases in language generation: Progress and challenges. 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)*, pages 4275–4293.

Zechen Sun, Yisheng Xiao, Juntao Li, Yixin Ji, Wenliang Chen, and Min Zhang. 2024. Exploring and mitigating shortcut learning for generative large language models. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 6883–6893, Torino, Italia. ELRA and ICCL.

Yifei Wang, Yuyang Wu, Zeming Wei, Stefanie Jegelka, and Yisen Wang. 2024. A theoretical understanding of self-correction through in-context alignment. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*.

Leonie Weissweiler, Kyle Mahowald, and Adele Goldberg. 2025. Linguistic generalizations are not rules: Impacts on evaluation of lms. *arXiv preprint arXiv:2502.13195*.

Yu Yuan, Lili Zhao, Kai Zhang, Guangting Zheng, and Qi Liu. 2024. Do llms overcome shortcut learning? an evaluation of shortcut challenges in large language models. In *Proceedings of the 2024 Conference on Empirical Methods in Natural Language Processing*, pages 12188–12200.

Qingjie Zhang, Han Qiu, Di Wang, Haoting Qian, Yiming Li, Tianwei Zhang, and Minlie Huang. 2024. Understanding the dark side of llms' intrinsic self-correction. *arXiv preprint arXiv:2412.14959*.

Jieyu Zhao, Daniel Khashabi, Tushar Khot, Ashish Sabharwal, and Kai-Wei Chang. 2021. Ethical-advice taker: Do language models understand natural language interventions? In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 4158–4164.

Shijia Zhou, Leonie Weissweiler, Taiqi He, Hinrich Schütze, David R Mortensen, and Lori Levin. 2024. Constructions are so difficult that even large language models get them right for the wrong reasons. In *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, pages 3804–3811.

A Appendix

A.1 Experimental Settings

In the BBQ dataset, distinct samples may share identical contextual scenarios while varying in entity mentions and question formulations. For instance, a template such as *'The meeting was delayed because [A] and [B] were late'* may generate multiple instances through slot filling. To prevent data leakage between training and testing partitions, we enforce non-overlapping contexts by identifying scenario-unique substrings (e.g., *'The meeting was delayed'*. We take the learning rate of 1e-6 for all tasks and fully finetune the models for at least 3 epochs until the loss converges. We conduct epoch-level evaluation during fine-tuning performance and report the optimal results.

A.2 Creating Discourse Constructions

We used basic components named *Context*, *Statement*, and *Action*, as well as their variations. For example, a *Statement* can be more situationally rephrased as *Situated Statement*. Then we assembled the different components to create discourse constructions for fine-tuning. We ensure that the created discourse constructions are grammatically correct.

As an example, for *Situated Statement*, we prompted Deepseek to provide a more concrete reason that leads to the stereotypes. The prompt is as follows:

> Given the context: [CONTEXT], please tell me why people always have stereotypes that [TARGET-GROUP] in the context [EVENT]. Give me a short answer with no more than 5 sentences. Your answer should not start with terms relevant to stereotypes. Please refer to the mentioned entities (if any) and events in the context while generating your answer. Please do not conclude with how we can avoid the stereotypes but conclude with a short statement that your reason may cause such stereotypes of [BIASED-GROUP] in the context [EVENT].

B More Experimental Results

In this document, we present additional experimental results on several models, including Phi-3.5-mini-instruct (3.8B), Llama-3-8B-Instruct, and Mistral-7B-Instruct-v0.3. These results extend our analysis across diverse model architectures and parameter scales. Overall, the findings align with the conclusions reported in our main paper.

B.1 Results of Phi-3.8b

`Phi3.5-mini`	**Age**	**Nation**	**Gender**	**SES**	**Disability**
Baseline	.9176	1	.9492	.9985	.9803
All	.9574	1	.9637	.9985	.9934
All - Context	.9347	1	.9528	.9985	.9868
All - Statement	**.9716**	**1**	**.9819**	**.9985**	**1**

Table 9: Experimental Results on Phi-3.5-mini-instruct for **Self-correction** Across Different Discourse Constructions. **All** includes all possible components: context, statement, action. **All - *** indicates a component was removed from the setting of **All**. The optimal performance is highlighted in **bold**. Across all experiments, **All - Statement** contributes to the optimal performance in all experiments, suggesting that LLMs do not need stereotype awareness for successful self-correction.

Table 9 presents the self-correction performance of Phi-3.5-mini-instruct after fine-tuning on three discourse constructions. All constructions demonstrate performance gains over the baseline, validating the effectiveness of our approach. Consistent with the findings in the paper, the discourse without the `Statement` component achieves the highest improvement, while removing `Context` results in a performance drop compared to the full construction (All), underscoring the critical role of contextual information.

Table 10 summarizes in-domain and out-of-domain generalization experiments for Phi-3.5-mini-instruct (3.8B). The results mirror those reported for LLaMA-3.2-3b-it in the paper, further demonstrate the robust out-of-domain generalization capabilities of LLMs of such sizes.

Phi3.5-mini	Age	Nation	Gender	SES	Disability	SexOrientation	Physical	Religion
Baseline	0.9176	1	0.9492	0.9985	0.9803	0.995	0.970	0.943
Mixed	0.9716	1	0.9691	0.9985	0.9934	0.9954	0.9824	0.9683

Table 10: Self-correction In-domain and Out-of-domain Generalization of Phi-3.5-mini-instruct By Mixing the Fine-tuning Corpus of 5 Representative Stereotypes. Phi-3.5-mini-instruct(3.8b) shows both good in-domain and out-of-domain generalization

B.2 Results of larger models

Llama3|Mistral	Age	Nation	Gender	SES	Disability	Age	Nation	Gender	SES	Disability
Baseline	.906	.987	.955	.897	.993	.838	.837	.695	.810	.849
All	.974	1.0	1.0	.964	**1.0**	.878	**.887**	.871	.955	.882
All - Context	.926	.983	.982	.941	1.0	.869	.880	.742	.821	.862
All - Statement	.983	**1.0**	**1.0**	**1.0**	.938	**.938**	.857	.860	**.988**	.882
All - Action	.957	.997	1.0	958	1.0	.847	.833	.790	.929	.875
Action	.940	.990	.969	.927	1.0	.889	.843	.735	.830	.862
All-Statement -subaction	**.986**	.993	1.0	.973	1.0	.898	.873	**.877**	.939	.875

Table 11: Experimental Results on Llama3-8b-it/Mistral-v0.3-7b-it for **Self-correction** Across Different Discourse Constructions. **All** includes all possible components: context, statement, action. **All - *** indicates a component was removed from the setting of **All**. The optimal performance is highlighted in **bold**. Across all experiments, only three out of ten **All - Statement** (**Disability** in Llama, **Age** and **Nation** in Mistral) do harm to the performance compared to the **All**, suggesting that LLMs do not need stereotype awareness for successful self-correction.

Llama3|Mistral	Age	Nation	Gender	SES	Disability	Age	Nation	Gender	SES	Disability
selfdiag baseline	.665	.540	.604	.777	.697	.645	.550	.682	.635	.618
selfcorr→ selfdiag	.713	.633	.655	.882	.697	.656	.560	.603	.600	.632

Table 12: Experimental Results of Llama3-8b-it/Mistral-v0.3-7b-it for Test Performance in Self-diagnosis Capability While Improving Self-correction. For all ten experiments, there are conflicts for only two of them.

This section presents experiments on larger models, LLaMA-3-8B-Instruct and Mistral-7B-Instruct-v0.3, using LoRA fine-tuning with a rank of 64 and a learning rate of 1e-5. It should be noted that we select Llama3-8b due to its architectural consistency with the 1B/3B models in our study. However, since the BBQ benchmark is widely used, this model has already been fine-tuned on it and achieves near perfect performance. Nevertheless, our framework still demonstrates measurable improvements. Table 11 summarizes the self-correction performance across all discourse constructions. Compared to smaller models, the 7B/8B parameter models achieve superior performance, with LLaMA-3-8B-Instruct attaining perfect accuracy in four out of five tasks. These results further validate the effectiveness of our proposed constructions for larger-scale LLMs.

Compared with smaller LLMs, there are some important observations. Removing the `Context` or `Action` components do harm to the performance relative to the full construction (`All`). However, only three out of ten **All - Statement** (specifically, *Disability* in Llama, *Age* and *Nation* in Mistral) underperform the **All** construction. This suggests that that explicit stereotype awareness is *not* essential for successful self-correction. These findings align with the shallow heuristics hypothesis presented in Section 4.2 of our paper. Table 12 present self-diagnosis performance when self-correction is enhanced. There are two key-observations: **(1)** The overall performance is not higher than smaller models significantly (primarily

Llama3|Mistral	Age	Nation	Gender	SES	Disable	Age	Nation	Gender	SES	Disable
Baseline	.906	.987	.955	.897	.993	.838	.837	.695	.810	.849
Individual	.983	1.0	1.0	1.0	.938	.938	.857	.860	.988	.882
Mixed	.992	.997	.996	.969	.993	.966	.980	.964	1.0	.967

Table 13: In-domain Generalization of Llama3-8b-it/Mistral-v0.3-7b-it By Mixing the Fine-tuning Corpus of Five Representative Stereotypes. Individual represents fine-tuning with the discourse within one stereotype. Mixed means that we mix the fine-tuning dataset of five stereotypes and test the fine-tuned model on each stereotype. The 7b/8b models shows great in-domain generalization capability.

between 0.6-0.7). **(2)** The conflicts are slightly mitigated in larger-scale models. The empirical findings necessitate fine-grained analysis of self-diagnosis, which is essential to understand how LLMs perform moral tasks.