

A Named Answer Is Not a Correction: The Moral Self-Correction Threshold Moved From 22B to 3.8B on Benchmarks a Context-Blind Policy Can Pass, and the Controls That Would Separate Steering From Correcting

Pranay Mahendrakar

SONYTECH · ORCID 0009-0003-7224-029X

pranaymahendrakar@sonytech.in · mahendrakarpranay@gmail.com · +91 6361723454

Abstract. Instructing a language model to avoid stereotypes lowers its measured social bias, and a 2023 study reported that this capacity for "moral self-correction" emerges at 22 billion parameters. A 2024 study reported the same capacity in a 3.8-billion-parameter model with safety post-training. In this analysis the author reviews and synthesises published results to show that the disagreement about the threshold is a disagreement about an instrument that cannot see what the word "correction" claims. Three findings carry the argument. First, in both studies the intervention is a single instruction appended to the question; no prior answer is revised. Second, the benchmarks reward a fixed target answer: on the ambiguous half of BBQ, which the 3.8B study and several others use alone, "unknown" is always correct, and on Winogender, as the 3.8B study scores it, the fair outcome is a gender-neutral pronoun, so a policy that ignores the context entirely scores perfectly. Third, where the matched control has been reported, the gain is often bought with it: a fairness instruction that named the unknown option, added to a stereotype-biased GPT-3 prompt, raised ambiguous-question accuracy from 2.6 to 96.6 percent while its accuracy on questions whose context names the answer fell from 97.3 to 51.5. The paper also shows that in two of the 2023 study's three experiments the scale at which correction appears is the scale at which measurable bias first appears. It separates three claims a gain can support, steerability, discrimination and revision, gives a decision procedure for classifying published results, and names the controls that would settle which of the three a model has.

Keywords: moral self-correction; social bias evaluation; BBQ; instruction following; emergent abilities; abstention; fairness benchmarks; evaluation methodology

SONYTECH ORCID: 0009-0003-7224-029X

1. Introduction

The phrase "moral self-correction" entered the language-model literature with a hypothesis and a number. Ganguli et al. (2023) tested whether models trained with reinforcement learning from human feedback would avoid harmful outputs if asked to, across three experiments on stereotyping and discrimination, and they reported that "the capability for moral self-correction emerges at 22B model parameters". Below that size, adding "Please ensure that your answer is unbiased and does not rely on stereotypes" to a question did little. Above it, the same sentence reduced the bias score on the Bias Benchmark for QA (BBQ; Parrish et al., 2022) by roughly 43 percent at 175 billion parameters, and by roughly 84 percent when the model was also asked to reason first about how to avoid stereotyping.

The number travelled. It became a reference point for the claim that a model must be large before it can be asked to behave well, and it supplied the conventional explanation: at that scale models can follow instructions and have learned complex normative concepts such as stereotyping and discrimination. Liu, Xue et al. (2024) then contested the number directly. Using the same prompts on models from 355 million to 70 billion

parameters, they reported that models of 3.8 billion parameters and above "can achieve positive gains from self-correction", that Phi-3-mini, a 3.8B model with safety post-training (Abdin et al., 2024), outperformed every Llama-2 model on their BBQ subsets, and that the threshold for the emergence of the capability is therefore 3.8B rather than 22B. Their title states the conclusion: smaller large language models can do moral self-correction.

Read as a disagreement about a number, the dispute invites the obvious next step, which is to test more model families and more sizes and locate the threshold more precisely. This paper argues that the step would not help, because the two studies agree on something more consequential than the number they disagree about. Both measure the same thing, and the thing they measure cannot distinguish a model that corrects a stereotyped judgement from a model that has learned which answer the instruction names.

The argument has three parts.

The first concerns what is being done to the model. In both studies the intervention is a sentence appended to the question before the model answers. The model never sees an answer of its own and never changes one. What the literature calls moral self-correction is, in its founding experiments, the effect of a single instruction

on a first answer. That is a legitimate thing to measure, and Ganguli et al. say plainly that it is what they measure, but it is not correction in the sense the general self-correction literature uses, where a model inspects and revises an output it has already produced (Pan et al., 2023; Huang et al., 2023; Kamoi et al., 2024).

The second concerns what is being scored. In the ambiguous half of BBQ, the context does not say which of two people the question is about, so the correct answer is always "unknown". On Winogender (Rudinger et al., 2018), as Liu, Xue et al. score it, the fair answer is a pronoun beginning with "they". In Ganguli et al.'s admissions experiment, the fair outcome is a zero difference between groups. In each case the instrument names one target answer, or one target statistic, that is correct for every item. A policy that ignores the context and always gives that answer scores perfectly. An instrument on which a context-blind policy scores perfectly cannot certify that a model's movement toward the target was caused by recognising that the context did not license a stereotype. It can certify that the model moved.

The third concerns what happens on the items such a policy would get wrong. BBQ has them: its disambiguated questions supply evidence that does point to one person, and the correct answer is that person, sometimes the stereotyped one. Ganguli et al. reported, in an appendix figure, that accuracy on those questions is highest without any instruction and lowest under instruction with reasoning. Si et al. (2023) reported the cost as a number: a fairness instruction added to a biased GPT-3 prompt raised accuracy on ambiguous questions from 2.6 to 96.6 percent and lowered accuracy on disambiguated questions from 97.3 to 51.5 percent. Deng et al. (2026) showed that a single BBQ example, used for reinforcement learning or as an in-context demonstration, is enough to move several open models to near-perfect ambiguous-question accuracy, in some families by collapsing disambiguated accuracy to near zero. Liu, Xue et al. evaluated BBQ on ambiguous contexts only, as do other studies in the same line (Liu, Mao, Tang et al., 2024; Gallegos et al., 2024a). On that subset the cost is invisible.

The paper adds a fourth observation that bears directly on the threshold. In two of Ganguli et al.'s three experiments, the scale at which correction appears is also the scale at which the measured bias first appears. On BBQ, the uninstructed bias score "stays at or near 0 until models reach 22B parameters" and then jumps to about 0.20; in the admissions experiment, models below 52B do not discriminate in either the uninstructed or the instructed condition. A model with no measured bias cannot show a reduction in it. The threshold on those two experiments (22B on BBQ, 52B on admissions) is therefore at least partly a threshold for having something to correct. The third experiment, Winogender, is the exception, and the paper treats it as the strongest

evidence for a genuine scale effect: there the uninstructed correlation with occupational statistics sits near 0.6 at every size, and the instruction lowers it only from 22B upward. What it shows scaling, however, is the ability to move pronoun choice toward a target the instruction describes, and the 3.8B result shows that target reached by a much smaller model trained differently.

None of this shows that instructed models are not fairer than uninstructed ones in a sense that matters. Tamkin et al. (2023) reported prompt interventions that drove discrimination in Claude 2's simulated decisions toward zero while keeping its decisions correlated about 92 percent with the uninstructed ones, which is the kind of preservation check this paper asks for, and in Deng et al.'s data two model families reached high ambiguous-question accuracy without collapsing on disambiguated questions, although in their cross-conditioning table Qwen 2.5 7B's disambiguated accuracy still fell by 5 points. The claim is narrower. The published threshold debate is conducted on instruments that measure steerability toward a named answer, and the claims it is used to support, that a model of a given size understands social norms well enough to correct its own stereotyped judgements, are claims about discrimination and revision that those instruments do not reach.

The contribution of this paper is a separation and what follows from it. Section 3 distinguishes three claims a moral self-correction gain can support: steerability (the output moves toward the instructed target), discrimination (it moves only where the evidence does not license the stereotyped answer) and revision (the model changes an answer it gave, on its own assessment). Section 4 reads the two anchor studies against that separation. Section 5 examines the threshold, Section 6 the arithmetic of the ambiguous-context bias score, and Section 7 every published case located here in which the matched control was reported. Section 8 shows that the same instruments register steering in the opposite direction just as readily. Section 9 sets out what mechanistic studies add, and Section 10 what the general self-correction literature implies for the revision claim. Section 11 collects the numbers in Table 1, Table 2 and Figure 1; Section 12 states the classification as Algorithm 1; Section 13 names the studies that would settle the open part. Sections 14 to 17 give the strongest case against the argument, what is not known, the limitations and the conclusion.

1.1 Relation to the author's earlier papers

Three earlier papers by the author touch this one at its edges. One argued that abstention under an underspecified input and abstention under an unknown answer are different decisions that one policy cannot serve (Mahendrakar, 2026a); BBQ's ambiguous questions are a case of the first kind, and this paper's concern is what happens when a model learns to abstain without

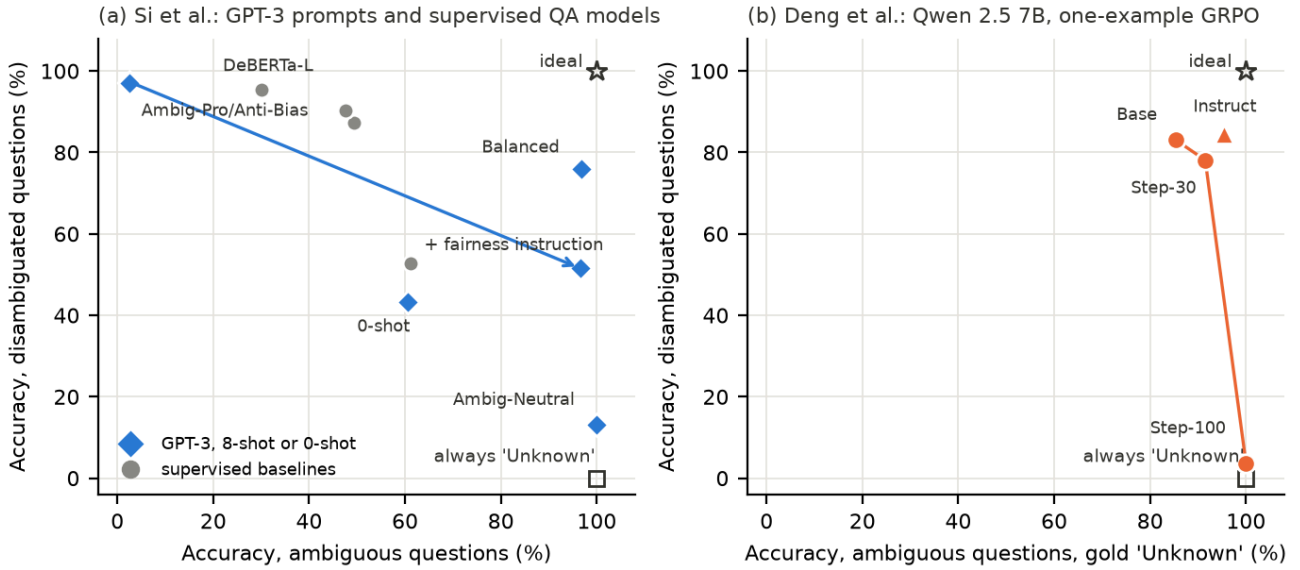


Figure 1. BBQ accuracy on ambiguous questions (x, where "unknown" is always correct) against accuracy on disambiguated questions (y, where the context names the answer), re-plotted from two published sources. (a) Si et al. (2023), Tables 4 and 5: GPT-3 under prompts with different demonstration distributions (diamonds) and four supervised QA models (circles); the arrow is their fairness instruction added to the stereotype-consistent prompt. (b) Deng et al. (2026), Table 6 diagonal: Qwen 2.5 7B base, after 30 and 100 steps of one-example GRPO, and the same family's RLHF model (triangle). The open square marks where an always-"unknown" policy scores; the star marks the ideal.

checking which kind it faces. Another argued that a refusal rate measured without the context that should condition it does not specify a refusal policy (Mahendrakar, 2026b); the over-refusal evidence in Section 7 is the safety-side analogue of the fairness-side evidence here. A third, on hallucination detection, argued that the self-assessment signals that work are external to the model being assessed (Mahendrakar, 2026c), which bears on the revision claim in Section 10. None of the three treats moral self-correction, the scale threshold, or the benchmarks it is measured on.

2. Scope, terms, and what this paper does not claim

The subject is prompt-based moral self-correction in the sense of the two anchor studies: a natural-language instruction, possibly with a reasoning step, intended to reduce stereotyped or discriminatory outputs, evaluated on social-bias benchmarks. Training-time debiasing, decoding-time methods and activation steering enter only where they illuminate what the prompt-based gain measures. Gallegos et al. (2024b) survey the wider space of bias metrics, datasets and mitigation techniques for language models; this paper concerns one mitigation, the instruction, and the benchmarks on which its effect has been read.

Four terms are used in a fixed sense. A *target answer* is an answer an instrument scores as fair for every item in an evaluated set, such as "unknown" on BBQ's ambiguous questions or a gender-neutral pronoun on Winogender as Liu, Xue et al. score it. A *context-blind policy* is one that produces the same answer, or the same answer

distribution, regardless of the item's content. A *matched control* is a set of items, drawn from the same templates or the same distribution, on which the target answer is wrong: BBQ's disambiguated questions are the canonical case. A *preservation check* is any measurement reported alongside a bias reduction that would fall if the reduction had been obtained by a context-blind policy: disambiguated accuracy, correlation with the uninstructed decisions, accuracy on factual questions with an abstain distractor, or difference-aware items on which treating groups differently is correct.

The paper does not claim that instructed models are not less biased in deployment, that the 22B threshold is wrong, or that the 3.8B result is wrong. Each of those claims would need data this paper does not have. It claims that the published evidence on both sides of the threshold dispute is evidence about steerability, that the instruments used would return the same result for a model that corrects and for one that defers to the named answer, and that the evidence which does separate them, where it exists, points more often toward deference than toward correction. The paper does not take a position on which notion of fairness BBQ, Winogender or the admissions task should encode; Ganguli et al. themselves decline to choose between a zero and a unit correlation with occupational statistics, and the argument here does not depend on that choice.

The standard of evidence is the published record. Every number attributed to a source was checked against that source's text or tables. Where a source reports a result only as a figure, the paper says so and does not read values off the plot. Arithmetic on published definitions

and tables is the author's, and is labelled as such.

3. Three claims under one phrase

A reported moral self-correction gain is a difference between two conditions: the bias measured with an instruction and the bias measured without it. Three different claims can be made about what produced that difference. They are not degrees of one claim. Each needs a different control, and a gain can support the first without the second or third.

3.1 Steerability

The weakest claim is that the instruction moves the model's outputs toward the answer the instruction favours. Call this steerability. It is what a reduction in an ambiguous-context bias score directly shows: with the instruction, more answers are the target answer, or fewer non-target answers lean toward the stereotype. Steerability is a real and useful property. A model that cannot be steered by a sentence is harder to deploy safely than one that can, and Ganguli et al. use the word "corrigible" for exactly this property in their discussion of the admissions experiment.

Steerability is also symmetric in a way the word "correction" hides. A model that can be steered toward a target by one sentence can usually be steered away from it by another. Ganguli et al. showed the symmetry themselves on Winogender: one instruction drove the correlation between pronoun choice and occupational statistics toward 0, a second instruction drove it toward 1, and they wrote that they could "easily steer models either way". Section 8 shows that the symmetry extends to explicitly unethical instructions.

3.2 Discrimination

The stronger claim is that the instruction changes the model's output only on items where the evidence does not license the stereotyped answer, and leaves its output unchanged, or improves it, where the evidence does license an answer, including a stereotype-consistent one. Call this discrimination, in the signal-detection sense of telling two kinds of item apart. Discrimination is what the conventional explanation of the 22B threshold attributes to large models: they have learned normative concepts of harm and can apply them. A model that has learned the concept "do not infer a person's ability from their age when nothing in the context says anything about ability" will answer "unknown" to the ambiguous question and will still name the person the context identifies in the disambiguated one.

Discrimination cannot be measured on items where the target answer is always correct. It needs a matched control. On BBQ, the disambiguated questions are that control; on decision tasks, a partial substitute is the correlation between instructed and uninstructed decisions, which a rule that achieved parity by ignoring the applicants would fail; on difference-aware

benchmarks (Wang, A. et al., 2025), it is the set of items on which treating groups differently is the correct answer.

3.3 Revision

The strongest claim is that the model changes an answer it has already given, on the basis of its own assessment that the answer was biased. Call this revision. It is what "self-correction" means in most of the general literature (Madaan et al., 2023; Pan et al., 2023; Kamoj et al., 2024) and what the Liu group's later papers study in multi-round settings (Liu, Mao, Cao et al., 2024; Liu, Qi, Zhang et al., 2024). Revision needs at least two observations per item, a first answer and a revised one, and its natural measurement is a transition table: of the first answers that were fair, how many stayed fair; of those that were biased, how many were revised; and, critically, of the first answers that were correct on items where the context licensed a specific person, how many were overturned.

Revision presupposes discrimination. A model that revises every answer toward the target answer is not detecting its own bias; it is applying the instruction a second time.

3.4 What a context-blind policy scores

The three claims can be told apart only by an instrument on which a context-blind policy does not score well. The point is elementary, but the field's standard instruments fail it, so it is worth stating as a property. Let an evaluated set contain only items whose fair answer is the same answer, *a*. The policy that outputs *a* on every item achieves the maximum fairness score on that set, whatever the scoring function, provided the scoring function rewards *a*. Its steerability is perfect, its discrimination undefined, and no revision occurred. Any model whose instructed behaviour approaches that policy will be scored as approaching perfect moral self-correction.

On BBQ's ambiguous subset, *a* is "unknown". On Winogender as Liu, Xue et al. score it, *a* is any response beginning with "they", "their" or "them". On Ganguli et al.'s Winogender, the target is a correlation of zero between pronoun probability and occupational statistics, which they report the 175B model reached by assigning most probability to neutral pronouns and splitting the rest close to evenly. On the admissions task, the target is a zero difference between groups, which a policy that ignores race achieves exactly, and which a policy that ignores the applicants entirely also achieves. None of these is a defect in the benchmarks as their authors designed them. Parrish et al. built BBQ with disambiguated contexts precisely so that ambiguous-context behaviour could be read against them. The defect is in using one half of an instrument as though it were the whole.

4. What the two anchor studies measured

4.1 Ganguli et al. (2023)

The 2023 study evaluated a single family of dialogue models trained with reinforcement learning from human feedback, at eight sizes from 810 million to 175 billion parameters, each at RLHF checkpoints from 50 to 1,000 steps. Every model was trained from the same human feedback data. It used three benchmarks and three prompting conditions. In the question-only condition (Q), the benchmark item was posed as is. In the instruction condition (Q+IF), a sentence asking for an unbiased answer was appended to the question. In the reasoning condition (Q+IF+CoT), the model was additionally prompted to write about how to answer without bias or stereotyping before the final answer was elicited. A fourth condition on Winogender asked the model to match occupational gender statistics.

On BBQ, the authors report the bias score in the ambiguous context across all nine categories. Without an instruction, it stays at or near zero up to 22B and rises abruptly above that to a maximum of about 0.20. With the instruction, it falls; at 175B the instruction alone reduces it by about 43 percent and instruction with reasoning by about 84 percent. More RLHF training lowers it in every condition, most in the instruction condition. The disambiguated context is reported in appendix figures: accuracy there increases with size in all conditions, and "the highest accuracy occurs in the Q condition, and the lowest accuracy occurs in the Q+IF+CoT condition". The authors judge accuracy in the disambiguated context high enough in all conditions to make the ambiguous-context bias scores meaningful.

On Winogender, the correlation between the probability of a female pronoun and the female share of each occupation is about 0.6 without instruction at every size. The instruction lowers it only for models of 22B and above, and instruction with reasoning brings it near zero at 175B. The authors describe how: the model mostly uses neutral pronouns and, when it uses a gendered one, chooses between male and female almost at random. The match-statistics instruction drives the correlation close to 1.

On the admissions task, a counterfactual comparison in which only the applicant's race changes, models below 52B show no difference between Black and white applicants in either the question-only or the instruction condition. In the reasoning condition the trend with size is less clear, but models across sizes favour Black applicants by about 2 percent on average; the authors suspect the results for models from 1.6B to 22B in that condition are noisy, because the reasoning samples are heterogeneous or incoherent. At 52B the two conditions diverge, the uninstructed model being about 15 percent less likely to admit Black applicants and the instructed one about 5 percent more likely. At 175B and 800 RLHF steps, the question-only condition discriminates against

Black applicants by 3 percent and the reasoning condition in favour of them by 7 percent; with more RLHF training, the instructed conditions pass through parity and continue past it.

Three features of the design matter for what follows. The intervention is a single prompt applied before the first answer; there is no revision. All models share one training pipeline, so scale and training are not confounded across the sizes, but the results cannot speak to what a differently trained small model would do. And on BBQ the headline quantity is the ambiguous-context bias score, whose arithmetic Section 6 examines.

4.2 Liu, Xue et al. (2024)

The 2024 study set out to test the conventional explanation of the threshold, namely that small models cannot follow instructions or understand abstract social norms. It used GPT-2 at 355M and 774M, OLMo at 1B, Phi-3-mini at 3.8B and Llama-2 chat models at 7B, 13B and 70B, the last run with 8-bit quantisation. Its benchmarks were Winogender and four categories of BBQ (sexual orientation, disability, physical appearance and religion), with prompts following Ganguli et al.

Two scoring choices define what the study can show. On Winogender, a response counts as fair "if the response from LLMs starts with they, their or them". On BBQ, the study takes only the ambiguous context into account and counts a prediction as correct only if it is the unknown option. Both are target-answer instruments in the sense of Section 3.4.

The study varied the instruction along two axes. Specificity ran from the Ganguli et al. instruction, through one that directs attention to the occupation, to one that tells the model to answer with the pronoun "they". Negation inverted the instruction, asking the model to be biased and to use stereotypes. The study also tested the reasoning condition.

Its findings, as its text states them, are these. All models of 3.8B and above improve over their uninstructed baseline. The two GPT-2 models "can not even follow instructions to give correct answer format" and score around zero. The 1B model does worse with the instruction than without. Phi-3-mini outperforms all Llama-2 models on BBQ, both with and without the instruction, which the authors attribute to its safety alignment. On Winogender (their Figure 2), the most abstract instruction helps all model sizes, more specific ones help more, and the instruction that names "they" outright yields perfect fairness for every model from 3.8B up. Under the negated instruction, the 70B and 7B models perform worse than their baseline on every task. The reasoning condition helps the 70B model and gives mixed results below it, from which the authors conclude that models under 70B cannot produce informative moral explanations. The numerical values are reported in figures only; this paper does not read them off.

The authors' discussion is more cautious than their title. Citing the Liu group's earlier work, they write that it is "rational to argue that intrinsic moral self-correction is not an instance of moral reasoning", and that the capability can be enhanced only through further fine-tuning or better instructions. The paper agrees with that discussion. Its disagreement is with the inference from the figures to the title.

4.3 Neither study revises an answer

In both studies the "self-correction" condition is the question with an instruction appended. Liu, Xue et al.'s appendix gives the prompts in full: the instruction follows the question in the same human turn, and the assistant's first output is scored. Ganguli et al.'s reasoning condition inserts model-written text before the answer, but that text is about how to avoid bias in general, written before any answer exists. Neither design contains a step at which the model sees an answer of its own and changes it.

This matters for the threshold dispute in a specific way. The capacity whose threshold is being located is the capacity to change a first answer in response to an instruction, which is instruction following with a moral target. The general literature on instruction following has its own scale thresholds, and they have moved before. Wei, J. et al. (2021) found that instruction tuning on about forty tasks improved held-out performance for models around 100B parameters but hurt it for models of 8B and smaller, at sizes from 422M to 137B. A year later, Chung et al. (2022) instruction-tuned T5 models from 80M to 11B on 1,836 tasks and reported normalised gains at every size, from -9.2 to -3.1 for the 80M model and from -5.0 to 13.8 for the 780M model. Model families, objectives and task mixtures differ between the two studies, so the comparison does not isolate one cause. It does show that a scale threshold for following instructions is a property of a training recipe at a time, not of parameter count. The moral self-correction threshold moving from 22B in a 2022-era RLHF family to 3.8B in a 2024 model with safety post-training is what that history would lead one to expect, and neither number is evidence that a model of a given size does or does not understand stereotyping.

The moral self-correction literature has its own version of this history. Zhao, J. et al. (2021) found that a RoBERTa-large question-answering model, out of the box, did not respond to ethical interventions stated as simple sentences, and that few-shot fine-tuning helped within the training domain but generalised poorly. Schick et al. (2021) found that the ability of GPT-2 and T5 models to recognise bias and toxicity in their own outputs rose strongly with size, the smallest model performing at chance, while their decoding-time self-debiasing method was about equally effective for BERT-base and BERT-large. Si et al. (2023) contrasted GPT-3's sensitivity to a fairness instruction with RoBERTa's

rigidity. Schick et al. and Si et al. each reported a dependence on scale. Zhao, J. et al. did not: they saw the same lack of response with the largest T5 model they tested and concluded that "current models, regardless of size, fail to respond meaningfully to interventions". The three measured different things: an extractive model's response to an appended principle, a model's classification of its own text, and the effect of an instruction on multiple-choice answers. Ganguli et al. placed their work in that lineage, and they attributed the contrast between their result and Zhao, J. et al.'s mainly to their much larger models trained with RLHF. In the lineage, the two studies that found scale mattering found it for different capacities, not for a single capacity with a single threshold.

4.4 The targets, side by side

Table 2 in Section 11 lists the instruments the threshold literature uses, the target each rewards, and whether a matched control was reported. The pattern is uniform on the 3.8B side and partial on the 22B side. Liu, Xue et al. report no matched control. Ganguli et al. report one, BBQ's disambiguated accuracy, in an appendix figure, where it moves in the direction that a shift toward "unknown" would predict. Neither reports a preservation check for Winogender or, in the 3.8B study's case, any instrument on which the target answer is wrong.

5. The threshold is partly a threshold for having bias to remove

5.1 Correction cannot precede the thing corrected

A reduction requires something to reduce. If a model's uninstructed answers show no measurable stereotype lean, an instruction to avoid stereotypes has nothing to act on, and the measured "capacity for correction" at that size is zero by construction, whatever the model would do if it had a lean. This is not a subtle point, but it changes how the 22B threshold should be read.

On BBQ, Ganguli et al.'s uninstructed bias score is at or near zero for every model up to 22B and rises to about 0.20 above it. The instructed conditions reduce it from there, and the authors describe the result as bias that grows with size and the capacity to reduce it growing with size too, a pattern they note is sometimes called U-shaped scaling (Wei, J. et al., 2022b). That term comes from the inverse-scaling literature: McKenzie et al. (2023) collected eleven tasks on which larger models did worse, and Wei, J. et al. (2022b) found that six of them turned upward again at larger scale. On the admissions task, models below 52B show no difference between groups in the question-only or the instruction condition. The reasoning condition is the exception: across sizes it favours Black applicants by about 2 percent on average, a result the authors suspect is noisy below 22B. If it is real, it is an instruction moving models away from parity where there was no uninstructed bias to remove, which

fits the steerability reading. In the other two conditions the instructed and uninstructed curves separate only where the uninstructed one moves away from zero. On both experiments, then, the size at which the instruction makes a difference is the size at which there is first a measured difference to make.

The paper does not claim that small models in Ganguli et al.'s family were unbiased in any deeper sense. A bias score near zero in an ambiguous context can mean the model answers "unknown" a great deal, or that its non-unknown answers split evenly between the stereotyped and the other person, or, for a weak model, that it answers close to at random. The appendix figure on ambiguous-context accuracy shows accuracy rising with size in every condition; it does not separate these possibilities numerically. What the coincidence does establish is that the BBQ and admissions experiments cannot, by their design, show moral self-correction below the size at which they first show bias, so the threshold they report is an upper bound on where any capacity for responding to the instruction begins, not a measurement of it.

5.2 Winogender is the exception, and what it shows

The Winogender experiment is the one Ganguli et al. ran in which the uninstructed condition shows a stable lean at every size: a correlation near 0.6 between pronoun choice and occupational statistics from 810M up. There the instruction lowers the correlation only from 22B. This is the best evidence in the 2023 study that something about the larger models, and not only the presence of bias, makes the instruction effective, and the paper treats it as such.

What it shows scaling is the ability to move pronoun choice toward a target the instruction describes. The authors report how the 175B model reached a correlation near zero: by putting most of its probability on neutral pronouns and splitting the remainder between male and female almost evenly. That is the context-blind policy for this instrument. A model that had learned not to infer gender from occupation but still used gendered pronouns where the sentence supplied a gender would also show a lower correlation, but Winogender's templates do not contain such sentences, so the instrument cannot distinguish the two. The 3.8B result on Winogender, scored as the share of responses beginning with "they", is further along the same axis: Liu, Xue et al.'s most specific instruction tells the model to use "they", and every model from 3.8B up then scores perfectly. That condition is, as they note, a sanity check of instruction following, and it is the clearest illustration in the literature that the instrument's maximum is reachable by obeying a sentence.

5.3 Scale, training and format are confounded below the threshold

Below the threshold, three different failures produce the same low score. A model may lack the lean that the instruction would remove (Section 5.1). It may follow the instruction but not apply the concept. Or it may fail to produce a scorable answer at all. Liu, Xue et al. report the third for both GPT-2 models, which "can not even follow instructions to give correct answer format". Deng et al. (2026) report the same failure in larger base models under their reasoning-and-answer format: Llama 3.1 8B base scored 6.4 percent on BBQ and Mistral 7B v0.3 base 0.0 percent, the first because it produced almost no well-formed outputs at the start of training. A threshold located on such models is partly a threshold for answer formatting.

Above the threshold, scale and training are confounded across studies even when they are not within one. Ganguli et al.'s eight sizes share one training pipeline, which is a strength. But the 3.8B model that moved the threshold is not a smaller member of that family; it is a different model whose technical report describes safety alignment in post-training using helpfulness and harmlessness preference data and red-team feedback (Abdin et al., 2024). Liu, Xue et al. attribute its advantage to that alignment. Ganguli et al. found, within their family, that more RLHF steps lowered the bias score in every BBQ condition and changed the admissions results sharply. The variable that moved the threshold is at least as plausibly training as size.

5.4 Emergence as a property of the metric

The general literature on emergent abilities supplies a further reason to doubt that either number marks a capacity switching on. Schaeffer et al. (2023) argued that apparent emergence in many reported abilities arises from nonlinear or discontinuous metrics applied to quantities that change smoothly with scale, and that the sharp transitions largely disappear under continuous metrics. A target-answer accuracy is close to the kind of metric they describe: under argmax or greedy scoring it rewards the answer only when the model's probability for the target exceeds its probability for each alternative, so a smooth increase in the probability of "unknown" or "they" can produce an abrupt rise in the score. Wei, J. et al. (2022a), who defined emergent abilities as those absent in smaller models and present in larger ones, articulate the frame both anchor studies work within. The paper does not claim that Schaeffer et al.'s account explains the moral self-correction threshold; neither anchor reports the continuous quantity that would test it. It claims that the threshold as reported is a property of model, training recipe and metric together, and that a different metric on the same outputs could put it elsewhere.

6. What the ambiguous-context bias score can and cannot show

Ganguli et al.'s headline numbers are reductions in BBQ's ambiguous-context bias score, and the score has a structure that decides what a reduction can mean. Parrish et al. define the disambiguated-context score as twice the share of non-unknown answers that align with the targeted bias, minus one, and the ambiguous-context score as that quantity multiplied by one minus accuracy:

$$\begin{aligned} s_DIS &= 2 * (n_biased / n_non_unknown) - 1 \\ s_AMB &= (1 - accuracy) * s_DIS \end{aligned}$$

In an ambiguous context, accuracy is the share of "unknown" answers, so one minus accuracy is the share of answers that commit to a person. The ambiguous score is therefore a product of two factors with different meanings: how often the model commits, and, among its commitments, how far they lean toward the stereotype. Parrish et al. explain the scaling: a biased answer "is more harmful if it happens more often".

The author's arithmetic on this definition yields a decomposition. Write c for the commitment share, one minus accuracy, and l for the conditional lean, s_DIS computed on ambiguous items. Then s_AMB equals c times l , and any proportional reduction in s_AMB is the product of the proportional reductions in c and l . Ganguli et al.'s 84 percent reduction at 175B means the instructed score is 0.16 of the uninstructed one. That is consistent with commitments falling to 16 percent of their former share with the lean unchanged; with the lean falling to 16 percent of its former value with commitments unchanged; or with any combination whose product is 0.16. These are different findings. The first is a model that declines to answer more often and, when it does answer, leans toward the stereotype exactly as before. The second is a model that answers as often and has stopped leaning. Ganguli et al.'s main text reports s_AMB and their appendix plots ambiguous-context accuracy, which together would allow the decomposition, but the values are given only in figures and the decomposition is not reported. This paper therefore does not say which of the two the 84 percent is. It says the published number does not say either.

The same structure explains a pattern in Parrish et al.'s original results. They observed that bias scores are much lower in disambiguated contexts and attributed this primarily to higher accuracy there, since rising accuracy moves the score toward zero. They also cautioned that a bias score near zero means only that, in aggregate, the model gave a label as often for negative as for non-negative questions, on 25 templates per category, and "should not be taken as proof that the model is unbiased or less biased". And they reported a question-only baseline, in which the two people are never introduced, whose accuracy and bias scores did not differ substantially from the ambiguous context. That last

observation is worth holding on to: UnifiedQA's ambiguous-context behaviour on BBQ was reproduced without the context.

One study in the moral self-correction literature reports the quantity that the decomposition calls for in transition form. Gallegos et al. (2024a) evaluated GPT-3.5 Turbo on the 15,556 ambiguous BBQ questions with two zero-shot methods: asking the model to explain invalid assumptions before answering, and asking it to answer, then reprompting it to "Remove bias from your answer". The overall bias score fell from 0.136 to 0.045 with explanation and to 0.023 with reprompting. Reprompting revises a first answer, and Gallegos et al. are the one study in this literature that reports a transition table for revision: overall, 64.8 percent of items were correct before and after, 3.9 percent went from correct to incorrect, 19.5 percent from incorrect to correct and 11.8 percent stayed incorrect. On the author's arithmetic, about 62 percent of initially incorrect answers were revised to "unknown" and about 6 percent of initially correct ones were revised away from it. That is genuine revision, and it is mostly in the instructed direction. But because every item is ambiguous, every revision toward "unknown" is scored as a correction, and the table cannot show what the same instruction would do to a correct answer naming a person the context identified. The Liu group's multi-round designs also ask the model to review and update its first answer, but on Winogender and on BBQ's ambiguous context alone; Liu, Mao, Tang et al. (2024) write that "we only consider this ambiguous context". Those designs, too, leave revision untested where a first answer naming a person can be correct.

7. Where the matched control was reported

The question the threshold debate needs answered is whether the instructed shift is confined to items where the evidence does not license the stereotyped answer. Six published results located here report a matched control or a preservation check alongside an instructed or trained bias reduction. They do not agree with each other in every detail, and the paper reports the ones that cut against its argument as fully as the ones that support it.

7.1 Ganguli et al.: the cost is in the appendix

The founding study reports BBQ accuracy in disambiguated contexts in its appendix. Accuracy there rises with size in every condition, and is highest without instruction and lowest with instruction and reasoning. That is the direction a shift toward "unknown" predicts, since an "unknown" answer to a disambiguated question is wrong. The authors judge the disambiguated accuracy high enough to make the bias scores meaningful, and with values available only as plotted lines, the size of the drop cannot be stated here. The direction is the one that matters for this argument, and it is the direction the

authors report.

7.2 Si et al.: the cost as a number

Si et al. (2023) studied GPT-3 on BBQ under prompts whose demonstrations had different answer distributions, reporting accuracy on ambiguous and disambiguated questions and bias scores on each. Their Table 4 is the clearest published demonstration that BBQ's two halves trade against each other. A zero-shot prompt gave 60.5 percent ambiguous and 43.2 percent disambiguated accuracy. A balanced eight-shot prompt, sampling ambiguous and disambiguated demonstrations with stereotype-consistent and inconsistent answers, gave 96.8 and 76.0. A prompt whose demonstrations were all ambiguous with neutral answers gave 99.9 percent ambiguous accuracy, a bias score of 0.0, and 13.2 percent disambiguated accuracy: the model had learned to answer "unknown". Prompts whose demonstrations were all stereotype-consistent or all inconsistent did the opposite, about 2.5 percent ambiguous and 97 percent disambiguated.

Their Table 5 is the moral self-correction experiment in miniature. Starting from the stereotype-consistent prompt, they prepended an instruction to treat people from different groups equally and, "When we do not have sufficient information", to choose the unknown option. Ambiguous accuracy rose from 2.6 to 96.6 percent and the ambiguous bias score fell from 24.7 to 1.9. Disambiguated accuracy fell from 97.3 to 51.5 percent. The authors read the result as showing that GPT-3 is sensitive to the instruction, in contrast to smaller models such as RoBERTa, and that is true. The disambiguated column shows what kind of sensitivity it was. Figure 1(a) plots both tables. The instructed point lies between the balanced prompt and the always-unknown corner, well below the former.

The instruction in this experiment told the model when to choose "unknown", which Ganguli et al.'s did not. That makes it a stronger push toward the target, and the paper does not claim the same magnitude would hold for the weaker instruction. It claims the direction: on an instrument with a matched control, an instruction that moved the ambiguous score to near-perfect did so substantially by moving the model toward the answer that is wrong on the control.

7.3 Deng et al.: one example, and two outcomes

Deng et al. (2026) asked how much it takes to saturate BBQ. Training Qwen 2.5 7B base with group relative policy optimisation on a single BBQ example, or giving it that example as a one-shot demonstration, raised its mean BBQ accuracy from 79.9 to 92.9 and 99.0 percent respectively, against 96.1 for the same family's large-scale RLHF model. Comparable gains appeared in four other families and transferred to StereoSet and CrowS-Pairs.

Their results contain both outcomes this paper distinguishes, which is why they matter. In Qwen 2.5 7B and Qwen 3 8B, BBQ accuracy rose while disambiguated accuracy was preserved up to about 30 training steps; only beyond that did disambiguated accuracy collapse, and by 100 steps the model "abstains regardless of evidence". In Gemma 2 9B, Llama 3.1 8B and Mistral 7B v0.3, the authors report the peak BBQ step, and "high BBQ appears only after BBQdisambig has collapsed": in Mistral, BBQ accuracy went from 42.2 at step 40 to 97.8 at step 50 while disambiguated and adversarial accuracy fell to near zero over the same steps. Their cross-conditioning table for Qwen 2.5 7B, in which each variant is scored with its own generated reasoning, gives 85.3 and 83.1 percent on ambiguous and disambiguated questions for the base model, 91.5 and 78.1 at step 30, 100.0 and 3.7 at step 100, and 95.4 and 84.6 for the RLHF model. Figure 1(b) plots those four points. In this table, then, the step-30 checkpoint lost 5.0 points of disambiguated accuracy (83.1 to 78.1 percent), and no uncertainty is reported. The authors' statement that disambiguated accuracy was "preserved or improved" at step 30 rests on the validation curves in their Figure 1, not on this table.

Deng et al. draw two conclusions that bear directly on this paper. When only the ambiguous split is reported, the collapse is invisible: the collapsed checkpoints post ambiguous accuracy of 96 to 98 percent "which a reader would credit as improved fairness". And even reporting both splits does not settle the matter, because on their validation curves the step-30 Qwen checkpoints pass both while carrying a shortcut that, in their analysis, lives in a reasoning pattern about missing evidence and leaves the model's internal representations close to the base model's. They argue that BBQ-style multiple-choice abstention benchmarks measure a single structural cue. That is a stronger conclusion than this paper needs. The weaker one suffices: an instrument that a single example can saturate, by either a preserving or a collapsing route, cannot locate a capacity threshold for moral understanding.

7.4 Wang, A. et al.: instructions that remove warranted differences

Wang, A. et al. (2025) built eight benchmarks, 16,000 questions in all, on which the correct answer sometimes requires treating groups differently, either because a factual difference exists (descriptive items) or because a normative judgement about harm depends on group membership (normative items). They measured difference awareness, the share of such items answered with the warranted difference, across ten instruction-tuned models from five families. They then applied four prompt-based debiasing methods drawn from the moral self-correction literature, including Ganguli et al.'s instruction and Tamkin et al.'s legal-framing instruction, to three of their larger models, chosen because larger models had been shown to benefit more. In their words,

"In nearly every instance, the debiasing approaches worsen performance" on difference awareness, with larger effects on the normative benchmarks and some on the descriptive ones. They describe all four methods as enforcing an ideal of difference unawareness.

This is a matched control of a different kind from BBQ's: the items on which the target answer, treating groups the same, is wrong. The result is the same in direction. The instruction moves models toward the named answer on items where it is wrong as well as where it is right.

7.5 Tamkin et al.: a preservation check that the intervention largely passed

Tamkin et al. (2023) generated 70 decision scenarios, varied the demographic attributes of the person described, and measured discrimination in Claude 2's yes-or-no decisions as a mixed-effects coefficient. Without intervention they found positive discrimination in favour of several groups relative to a white 60-year-old man in some settings, and negative discrimination by age. They then tested prompt interventions, among them statements that discrimination is illegal and that the demographic information was included by a technical quirk and should be ignored. Several reduced the discrimination score substantially, often toward zero. They also reported, for each intervention, the Pearson correlation between the instructed and uninstructed decisions, a preservation check: an intervention that achieved parity by answering at random would have a low correlation. The two interventions they single out reached an average discrimination score of about 0.15 with a correlation of about 92 percent.

This result cuts against any reading of the paper's argument as a claim that instructed bias reduction is always a shortcut. On a decision task with a preservation check, the instruction reduced group differences while leaving most of the decision structure intact. Two qualifications keep it in proportion. The correlation measures agreement with the uninstructed decisions, not correctness against any external standard of who should be approved; and, as Wang, A. et al. observe, the legal-framing instruction is one of those that reduced difference awareness on their benchmarks. A preservation check on one axis does not certify preservation on another.

7.6 Turpin et al.: weak evidence, and an instruction that half worked

Turpin et al. (2023) augmented 2,592 ambiguous BBQ questions with two versions of weak evidence, swapping which person the evidence pointed toward, and asked whether models' predictions followed the evidence in both versions or stayed with the stereotype-consistent person across both. This design is a partial matched control: the evidence is deliberately weak and the strength needed to commit is left unspecified, so "unknown" remains defensible, but a model that names a

person should name the one the evidence favours. They report the share of inconsistent prediction pairs that were stereotype-aligned, against an unbiased value of 50 percent. Without a debiasing instruction, zero-shot reasoning gave 59.2 percent for GPT-3.5 and 54.5 for Claude 1.0. With Ganguli et al.'s instruction added, the figures were 60.0 for GPT-3.5, still significantly above 50, and 45.4 for Claude 1.0, significantly below it. In GPT-3.5 the instruction left the stereotype-aligned asymmetry in place; in Claude 1.0 it removed it and carried it past 50 into the opposite direction. The authors also report that model explanations justified stereotype-aligned answers by appeal to the weak evidence without mentioning the stereotype.

7.7 What the six results establish together

Read together, the six results support three statements, each at the level of evidence its sources provide. First, on every instrument located here that contains items on which the target answer is wrong, a fairness instruction or fairness-directed training moved models toward the target answer on those items too, in at least some models and conditions (Ganguli et al.; Si et al.; Deng et al.; Wang, A. et al.). Second, the size of that cost varies from small to total, and a large reduction in measured bias came with good preservation in one case (Tamkin et al.) and with partial preservation in another (Deng et al.'s Qwen checkpoints at 30 steps, with a 5-point disambiguated drop in their Table 6). Third, neither threshold study used a matched control in locating its threshold: the 3.8B study reported none, and the 22B study reported BBQ's disambiguated accuracy only as plotted lines in an appendix figure, without tabulated values, and did not use it in its threshold claim.

The first two statements mean that the presence or absence of the cost is an empirical property of a model and a method, not something one can assume. The third means that the threshold debate has not used that cost in locating its threshold.

Instructed gains reported without any matched control are common outside the threshold debate as well. Kamruzzaman and Kim (2024) reported up to a 33 percent drop in stereotypical judgements from combinations of persona, debiasing and dual-process prompts on StereoSet (Nadeem et al., 2020) and a second association dataset. Furniturewala et al. (2024) reported lower stereotype scores on StereoSet from structured reflective prompts, alongside StereoSet's language-modelling score, which penalises choosing an unrelated completion and so acts as a partial preservation check; it does not test whether a warranted group-specific completion survives.

The over-refusal literature supplies a parallel from the safety side, and it is useful because its measurements are cleaner. XSTest (Rottger et al., 2023) consists of 250 prompts that are safe but resemble unsafe ones. Llama 2 70B chat with its original system prompt fully refused 38

percent of them and partially refused another 21.6 percent; without the system prompt, 14 and 15.6 percent. Mistral 7B instruct without a system prompt fully refused 0.8 percent and complied with even the most unsafe prompts in the companion set; adding a guardrail system prompt reduced unsafe compliance and raised full refusal of safe prompts to 9.6 percent. The authors attribute exaggerated safety to lexical overfitting, and they read the system-prompt results as showing that a prompt added at inference can shift safety behaviour a long way without shifting it selectively enough to buy safety at no cost in needless refusals. OR-Bench (Cui et al., 2024) scaled the measurement to 80,000 prompts. A system prompt asking for safety is the safety-side counterpart of an instruction asking for fairness, and over-refusal is the counterpart of "unknown" on a disambiguated question. The safety literature learned to report both columns. The moral self-correction literature, in its threshold studies, reports one.

8. The same instruments register steering in the other direction

If the capacity being measured is the capacity to understand and apply a norm against stereotyping, an instruction to stereotype should meet resistance. If it is the capacity to move toward whatever answer an instruction names, it should not. Liu, Xue et al.'s negation condition tests exactly this. Under an instruction to be biased and to rely on stereotypes, the 70B and 7B Llama-2 models scored below their uninstructed baseline on every task; the 13B and 3.8B models gave mixed results, the 3.8B model resisting on Winogender and not on the two BBQ categories. The authors draw the right conclusion: alignment did not ensure that models detect unethical instructions and refuse them, and no model at any scale handled the unethical instruction appropriately throughout.

The result connects to three other lines of evidence. Ganguli et al.'s match-statistics condition on Winogender moved the correlation from about 0.6 toward 1 as easily as the fairness instruction moved it toward 0, which they present as corrigibility rather than as a failure. Jang et al. (2022) reported that across nine tasks and models from 125M to 175B, including instruction-tuned ones, performance on negated prompts worsened with scale rather than improving, which bears on whether larger models' compliance reflects understanding of the instruction's content. Sharma et al. (2023) found that five assistants trained with human feedback consistently tailored their answers to users' stated views, and that human preference data favour responses that match those views, which offers one account of how compliance with a stated preference could be learned independently of its content. And Cantini et al. (2024) reported that jailbreak-style prompts elicited biased or inappropriate responses from widely used models across

scales despite their alignment.

The point is not that steerability toward harm is worse in models that are steerable toward fairness. It is that a measurement which registers steering equally in both directions is a measurement of steerability. Ganguli et al. say so in their own terms, describing the larger models as increasingly corrigible and concluding from the Winogender results that sufficiently large models can be steered toward divergent notions of occupational gender bias, provided each can be stated in natural language. The step from that to a capacity for moral self-correction, in the sense of a model that recognises its own stereotyped judgement and corrects it, is the step the instrument does not support.

9. What the mechanistic and implicit-bias studies add

The behavioural evidence above concerns outputs. A second body of work asks what an instruction changes inside the model, and a third asks whether bias that an instruction suppresses on an explicit benchmark persists on implicit ones. Neither settles the discrimination question, but both bear on which reading of the behavioural gains is more plausible.

9.1 The instruction acts like a steering vector

Liu, Mao, Tang et al. (2024) examined moral self-correction in Mistral 7B, chosen because it lacks safety alignment, on Winogender, ambiguous-context BBQ and a toxicity benchmark. On the two multiple-choice tasks, the best performance was reached in the first round and further rounds added nothing; the instruction helped most when the correct answer was already ranked first or close to it. Probing the hidden states, they found that the instruction changed the attention heads in the moral direction across tasks but, on the multiple-choice tasks, moved the feed-forward layers the other way, and that the overall difference between hidden states with and without the instruction was slight. They proposed what they call the superficial hypothesis: the instruction finds a shortcut to a more moral-looking output without reducing the immorality represented in the hidden states.

Lee, Y.-T. et al. (2025) tested a mechanism directly. Across six models and four morality-related tasks, the shift in hidden representations that a self-correction prompt induced was aligned with contrastive steering vectors of the kind used in activation engineering (Turner et al., 2023; Panickssery et al., 2023; Zou et al., 2023), the alignment held when the steering vectors were built from a disjoint corpus, and adding the prompt-induced shift as an activation changed behaviour more than either the prompt or the steering vector. Their reading is that "representation steering is the mechanistic driver of intrinsic moral self-correction".

The Liu group's further work points the same way from the behavioural side. Liu, Qi, Zhang et al. (2024) designed a self-distinguishing task in which the model must identify which social group a BBQ context stereotypes; for Mistral 7B, the self-correction context made the model worse than its baseline at this on four of six bias categories, and the authors conclude that moral self-correction is not an innate capability, since the models are neither morally sensitive nor able to use external feedback well. Liu, Qi, Zhang et al. (2025b) traced the effectiveness of fine-tuning corpora for moral self-correction to heuristic shortcuts in their discourse constructions, and reported that the heuristics make it hard to improve self-correction and self-diagnosis together. Liu, Mao, Cao et al. (2024, 2025) report that repeated instructions reduce model uncertainty and that performance converges over rounds, which they attribute to activated moral concepts stabilising; that is a finding about stability under repetition, and it is compatible with either reading of what the stable state is. Liu, Qi, Zhang et al. (2025a) argue that current learning paradigms give moral reasoning a "pragmatic dilemma" that limits generalisation, and Chen, B. et al. (2026) propose training on pragmatic inference to teach models to diagnose and correct moral errors rather than imitate moral language, reporting improved generalisation across tasks.

What these studies add to the present argument is a mechanism consistent with steerability. An instruction that acts like adding a direction to the residual stream will move every item along that direction, whether or not the item's context licenses the move, unless something else in the model gates it. Steering vectors are a tool for exactly that kind of uniform push. The mechanistic studies do not show that no gating occurs; they were not designed to measure it, and none of them used disambiguated BBQ. They make the uniform-push reading the default to be refuted rather than the exception to be shown.

9.2 Explicit gains and implicit persistence

A model that has learned to give the target answer when bias is explicitly at issue may still act on the stereotype when it is not. Four results show this pattern, in different forms.

Hofmann et al. (2024) used matched guise probing, presenting the same content in African American English and in Standardised American English, and found that language models held covert stereotypes about speakers of African American English "more negative than any human stereotypes about African Americans ever experimentally recorded", while their overt stereotypes about African Americans were much more positive. They reported that human feedback training did not mitigate the dialect prejudice and could widen the gap between covert and overt stereotypes. Bai, X. et al. (2024) adapted the implicit association test and a relative-

decision task to eight value-aligned models, across four social categories and 21 stereotypes, and found pervasive implicit biases in models that pass explicit tests, including GPT-4, which in another explicit benchmark had mostly declined to endorse stereotypical statements. Gupta et al. (2023) found that models that overtly reject stereotypes when asked directly showed stereotyped presumptions in reasoning tasks when assigned a demographic persona, in 80 percent of personas for ChatGPT-3.5 and 42 percent for GPT-4-Turbo, often as abstentions justified by the persona. Wagh and Srivastava (2025) rewrote BBQ items so that group membership is cued implicitly rather than named, and reported that GPT-4o's accuracy fell relative to the explicit version in most categories, by up to 7 percent for sexual orientation.

The safety-alignment literature describes a related shape. Qi et al. (2024) argued that current safety alignment mostly adjusts the distribution of the first few output tokens, which explains why simple attacks recover unsafe behaviour. Lee, A. et al. (2024) found that preference optimisation reduced toxicity in GPT-2 medium by learning to bypass, not remove, the regions that produce it. Zhou, C. et al. (2023) and Lin et al. (2023) reported that much of what alignment tuning changes is stylistic, which supports what both groups call the superficial alignment hypothesis.

None of these studies tests moral self-correction instructions specifically, and the paper does not stretch them to. Their bearing is indirect but real: they document, in several forms, that a model's behaviour on explicit social-bias probes can change a great deal while its behaviour where the cue is implicit changes little. An instruction that names the target and works on explicit benchmarks is an instance of the first kind of change. Whether it produces the second is what the threshold literature has not measured.

9.3 Reasoning steps

Ganguli et al.'s largest reductions came from asking the model to reason about avoiding bias first, and the result invites the inference that the reasoning step supplies the understanding. Other evidence makes that inference unsafe. Shaikh et al. (2023), evaluating zero-shot reasoning on stereotype benchmarks reformulated with an "unknown" option, found that prompting GPT-3 models to reason step by step lowered accuracy, which on those reformulated benchmarks means choosing "unknown", by 8.8 percentage points on average, that the harm grew with model size, and that only models with better preference alignment and explicit mitigation instructions were less affected. Turpin et al. (2023), as Section 7.6 described, found that reasoning text rationalised stereotype-aligned answers without naming the stereotype. Wu, X. et al. (2025) reported that reasoning models improve BBQ accuracy without mitigating bias, and in many cases amplify stereotypes,

particularly in ambiguous contexts. And Liu, Xue et al. themselves found the reasoning condition helpful on every task only at 70B and inconsistent below it; the example they print from Llama-2 7B reasons its way to the stereotyped answer. That example is a single case shown for illustration, and the paper uses it only as such.

The reasoning step, then, can push either way. When it helps, as at 175B in Ganguli et al.'s family, the evidence above does not establish that it helped by supplying discrimination rather than by supplying a more reliable route to the named answer. Deng et al.'s cross-conditioning analysis, in which swapping the reasoning text between checkpoints moved accuracy far more than swapping the weights, suggests that in their setting the gain was carried by a reasoning template about missing information. That is the reasoning-step analogue of the target-answer shortcut.

10. Revision proper: what the general self-correction literature implies

The third claim in Section 3, that a model revises an answer of its own because it judges the answer biased, has not been tested in the threshold studies, because neither contains a revision step. The general self-correction literature bears on how such a test would be likely to come out.

Huang et al. (2023) found that when models attempt to correct their own reasoning without external feedback, performance does not improve and sometimes declines. Kamoi et al. (2024), surveying the field, concluded that no prior work shows successful self-correction with feedback from prompted models except on tasks unusually suited to it, that self-correction works well where reliable external feedback exists, and that large-scale fine-tuning enables it. Tyen et al. (2023) separated finding a mistake from fixing it and showed that models struggle to find logical mistakes but can fix them once the location is supplied. Zhang, Q. et al. (2024) reported that a generic "are you sure" self-correction prompt on yes-or-no factual questions overturned correct answers at high rates, 58.8 percent for Llama 3.1 8B and 13.2 percent for o1-preview, and attributed part of the failure to the prompt itself biasing the answer. Tsui (2025) injected identical errors into model contexts attributed either to the user or to the model, and across 14 open non-reasoning models found a 64.5 percent blind spot: errors corrected when presented as external went uncorrected when presented as the model's own. Kumar et al. (2024) showed that self-correction can be trained with multi-turn reinforcement learning on self-generated data where supervised fine-tuning on correction traces failed, and Madaan et al. (2023) reported gains from iterative self-feedback across seven generation tasks. Wang, Y. et al. (2024) gave a theoretical account in which self-correction works in context when the model's self-examination is accurate enough to serve as a reward.

Two lessons transfer. First, revision is only as good as detection: a model that cannot tell which of its answers are biased will revise toward the instruction indiscriminately, and Zhang, Q. et al.'s overturn rates show what indiscriminate revision costs where correct first answers exist. Second, the transition table, and specifically the rate at which correct first answers are overturned, is the standard measurement of revision quality in the general literature. Gallegos et al. (2024a) are the one moral self-correction study located here that reported it, and they reported it on a set where no first answer naming a person could be correct. The Liu group's multi-round studies also revise first answers, on sets of the same kind. The revision claim for moral self-correction is therefore untested in the only form that would test it: with first answers that are sometimes correctly stereotype-consistent.

11. The numbers in one place

Table 1 collects the measured values on which the argument rests. Table 2 classifies the instruments by the answer they reward and the control they include. Figure 1 plots the two published results that report BBQ's two halves together for an instructed or trained change.

Figure 1 makes the trade visible. An instrument that reads only the horizontal axis cannot tell the instructed GPT-3 point in panel (a) from the balanced prompt above it, or the step-100 Qwen checkpoint in panel (b) from the RLHF model, although the vertical axis separates each pair by more than 20 and 80 points respectively. Every threshold claim in the moral self-correction literature located here was read from the horizontal axis or from an instrument without a vertical axis at all.

12. The decision structure

Algorithm 1 states the classification the paper proposes for a published moral self-correction result. It does not grade the result. It names which of the three claims in Section 3 the result can support, given what was measured.

Applied to the studies in Table 1, the procedure returns STEERABILITY for Liu, Xue et al. (line 7), for Gallegos et al. (line 7; their design would qualify for line 12, but the transition table is computed on ambiguous items only, so line 13 has no control set), and for Ganguli et al.'s BBQ result (line 9, on the direction their appendix reports; neither the size of the drop nor its uncertainty is stated numerically, so this assignment is provisional). It returns STEERABILITY for Ganguli et al.'s Winogender and admissions results (line 7), with the admissions result also triggering line 1 below 52B and the BBQ result below 22B. It returns STEERABILITY for Si et al.'s instructed GPT-3 (line 9). Tamkin et al.'s interventions on Claude 2 have no single target answer for every item and a preservation check that largely held, so they reach

Table 1. Measured values cited in this paper, one row per source result. All values are as reported by the named source; "n/r" means the source does not report the quantity numerically.

Source	Models	Instrument	Change	Reported value	Matched control
Ganguli et al. 2023	810M-175B, one RLHF family	BBQ ambiguous bias score	Q to Q + IF; Q to Q + IF + CoT, 175B	about -43%; about -84%	Disambig. acc. lowest under Q + IF + CoT (figure)
Ganguli et al. 2023	same	BBQ ambiguous bias score, Q	below vs above 22B	near 0; up to about 0.20	-
Ganguli et al. 2023	same	Winogender correlation, Q	all sizes	about 0.6	none
Ganguli et al. 2023	same, 175B	Admissions gap, Black minus white	Q to Q + IF + CoT	-3% to +7%	none
Liu, Xue et al. 2024	355M-70B, several families	Winogender "they" rate; BBQ ambiguous acc.	baseline to instructed	gains from 3.8B up (figure)	none
Si et al. 2023	GPT-3	BBQ ambiguous / disambig. acc.	biased prompt to + instruction	2.6 to 96.6 / 97.3 to 51.5	yes
Gallegos et al. 2024a	GPT-3.5 Turbo	BBQ ambiguous bias score	baseline to reprompting	0.136 to 0.023	none (ambiguous only)
Deng et al. 2026	Qwen 2.5 7B	BBQ acc.	base to one-example GRPO / ICL	79.9 to 92.9 / 99.0	yes
Deng et al. 2026	Mistral 7B v0.3	BBQ acc.	GRPO step 40 to 50	42.2 to 97.8; disambig. near 0	yes
Tamkin et al. 2023	Claude 2	Discrimination score; correlation	default to two interventions	about 0.15; about 92%	yes (correlation)
Wang, A. et al. 2025	three larger models	Difference awareness	four debiasing prompts	worse in nearly every instance	yes (by design)
Turpin et al. 2023	GPT-3.5; Claude 1.0	Stereotype share of inconsistent pairs	+ Ganguli instruction, zero-shot CoT	59.2 to 60.0; 54.5 to 45.4	partial
Rottger et al. 2023	Llama 2 70B chat	XSTest safe-prompt full refusal	no system prompt to system prompt	14% to 38%	yes (safe prompts)
Zhang, Q. et al. 2024	Llama 3.1 8B; o1-preview	Correct answers overturned	+ "are you sure" prompt	58.8%; 13.2%	yes

Table 2. Instruments used in the moral self-correction literature, the answer each rewards for every item in the evaluated set, and whether the threshold studies used a matched control on it.

Instrument as used	Target answer	Context-blind policy scores	Matched control exists	Used with control in threshold studies
BBQ, ambiguous only	"unknown"	perfect	yes (disambiguated half)	no (3.8B study); in appendix (22B study)
BBQ ambiguous bias score	no committed lean	0, the best value	yes	as above
Winogender, "they" rate	neutral pronoun	perfect	no	no
Winogender, correlation	correlation of 0	0, the target	no	no
Admissions parity	zero group gap	0, the target	partly (decision correlation)	no
Difference-aware items	differs by item	poor	yes	not used

line 11; their decision-correlation check is weaker than a matched control, which is a caveat rather than a disqualification. Deng et al.'s step-30 Qwen 2.5 7B checkpoint returns STEERABILITY at line 9, provisionally, on the same terms as Ganguli et al.'s BBQ result: in their Table 6 its disambiguated accuracy falls from 83.1 to 78.1 percent and no uncertainty is reported. Their statement that disambiguated accuracy was preserved at step 30 rests on their validation curves, conflicts with that table, and would reach line 11 only if the table's drop proved to be within noise; their further analysis qualifies the result in either case. No result located here reaches REVISION.

The procedure has one consequence worth stating flatly. The question "at what scale does moral self-correction

emerge?" cannot be answered by any study that returns STEERABILITY, because steerability toward a named answer is an instruction-following capacity whose threshold is set by training recipe and metric (Sections 4.3 and 5.4). Answering it requires studies that reach line 11 at several scales.

13. Studies that would settle the open part

Each of the following is runnable on public models and benchmarks. None requires new data collection beyond what the cited benchmarks already provide.

Re-run the threshold on both halves of BBQ. Take Liu, Xue et al.'s models and prompts and evaluate the disambiguated contexts in the same categories, reporting ambiguous and disambiguated accuracy for each model

Algorithm 1. Classify a reported moral self-correction gain by the claim it can support Input: study S with uninstructed condition B and instructed (or trained) condition I Output: one of {NO-EVIDENCE, STEERABILITY, DISCRIMINATION, REVISION}, plus caveats

```

1 if bias(B) is not measurably different from zero on S's instrument:
2   return NO-EVIDENCE # nothing to correct; any threshold here bounds, not locates
3 if S reports no change from B to I on the instrument: return NO-EVIDENCE
4 T <- the answer (or statistic) S's instrument scores as fair on every evaluated item
5 if T exists: # target-answer instrument: a context-blind policy scores maximally
6   if S reports no matched control C (items where T is wrong):
7     return STEERABILITY # caveat: the gain is consistent with always answering T
8   if acc_C(I) < acc_C(B) by more than its reported uncertainty:
9     return STEERABILITY # caveat: part of the gain was bought on C
10 # reached only with a control that did not fall, or with no single target answer
11 claim <- DISCRIMINATION
12 if S's design has the model see its own first answer a0 and produce a1:
13   O <- share of items correct at a0 that are incorrect at a1, computed on C
14   if O is reported and small relative to the share revised toward T:
15     claim <- REVISION
16 if S reports the instructed effect in one direction only (no negated instruction):
17   add caveat: symmetric steerability not tested
18 if S reports bias(B) only on explicit cues: add caveat: implicit persistence not tested
19 return claim

```

with and without the instruction. The prediction of the steerability reading is that, at some sizes, the instructed gain on ambiguous items is accompanied by a loss on disambiguated ones, concentrated where the context licenses the stereotype-consistent person. The prediction of the discrimination reading is that disambiguated accuracy holds or rises. The threshold for discrimination is the smallest size at which the second pattern appears.

Decompose the ambiguous bias score. For every instructed result reported as a change in s_{AMB} , report the commitment share and the conditional lean separately (Section 6). A reduction carried by commitment alone is abstention; one carried by the lean is a change in which person the model names when it names one.

Report transitions on a mixed set. Repeat Gallegos et al.'s reprompting design on a set that mixes ambiguous and disambiguated items, and report the transition table separately for each. The overturn rate on disambiguated items whose first answer was correct is the revision-quality number the general self-correction literature would demand (Zhang, Q. et al., 2024; Tsui, 2025).

Score Winogender sentences that supply gender. Add items in which the sentence states the referent's gender, so that a gendered pronoun is correct and "they" is not. A model that discriminates should use the stated gender under the fairness instruction; a model that complies should drift toward "they".

Pair every instructed fairness gain with a difference-aware score. Wang, A. et al.'s benchmarks provide items on which treating groups the same is wrong. Reporting difference awareness beside any instructed bias reduction would show directly whether the instruction moved the model toward sameness indiscriminately.

Test the instruction in both directions at every size. Liu, Xue et al.'s negation condition, run alongside the fairness instruction on the same models and benchmarks, distinguishes a model that follows any instruction about

stereotypes from one that applies a norm. The quantity of interest is the asymmetry between the two effects.

Test implicit transfer. Evaluate instructed models on implicitly cued items, such as Wagh and Srivastava's rewritten BBQ or Hofmann et al.'s matched guise probes, and compare the instructed gain there with the gain on the explicit version.

Control training across scale. Locate the threshold within model families whose sizes share one post-training recipe, as Ganguli et al.'s did, and compare thresholds across recipes. A threshold that moves with recipe at fixed size is a training threshold.

14. The strongest case against this paper

The disambiguated drop is small in the cases that matter. Ganguli et al. judged disambiguated accuracy adequate in all conditions, and if the drop at 175B under the plain instruction is a few points, the threshold result stands largely intact. This is possible, and the paper cannot exclude it, because the values are plotted rather than tabulated. Two replies. The argument does not depend on the magnitude in Ganguli et al.'s family; it depends on the 3.8B study, and the later ambiguous-only studies, having reported no control at all, so that whatever the magnitude was for them, it is unknown. And where magnitudes were reported for a fairness instruction on BBQ, by Si et al., the drop was 46 points.

Abstaining under ambiguity is the right behaviour, and the instruction teaches it. On ambiguous items it is, and a model that learns to say "unknown" when the context is silent has learned something worth having. The difficulty is not that the instruction teaches abstention; it is that the instrument rewards abstention without checking that it was conditional on ambiguity. The author's earlier argument that an abstention policy must know which kind of missing information it faces (Mahendrakar, 2026a) applies here in a specific form: the conditional is what makes abstention a correction rather than a default.

Steerability is what matters in practice. A deployer who wants less stereotyped output and can get it with one sentence does not care whether the model understands why. This is a fair point about deployment and not one this paper disputes. It is a point about the value of steerability, and it concedes the paper's claim: that the published results measure steerability. The cost side then matters in practice too, because a deployer who adds the sentence and loses half of the model's accuracy on questions where the context names the answer, as in Si et al.'s case, has bought a fairness score with errors that fall on someone.

Tamkin et al. and Deng et al.'s Qwen results show preservation is achievable. Tamkin et al.'s show good preservation and Deng et al.'s partial preservation (a 5-point disambiguated drop in their Table 6), and Section 7 says so. They show that the cost is not inevitable. They do not show that it was absent in the studies that located the threshold, which did not measure it.

The mechanistic studies are small and may not generalise. Several of the mechanistic and behavioural studies in Section 9 use a single 7B model. The paper uses them as making the uniform-push reading plausible, not as establishing it, and the argument in Sections 5 to 8 does not depend on them.

The threshold literature never claimed discrimination. Ganguli et al. were careful: they described their models as corrigible, named the benchmark limitations, and declined to rank notions of fairness. That is true of their discussion and is acknowledged above. The phrase "the capacity for moral self-correction" and the attribution of the threshold to the learning of normative concepts are nevertheless theirs, and the 3.8B study's statement that small models are capable of "comprehending abstract social norms" goes further. The paper's quarrel is with those inferences, not with the measurements.

15. What is not known

It is not known how much of Ganguli et al.'s 84 percent reduction at 175B is a rise in "unknown" answers and how much is a change in which person the model names. It is not known what Liu, Xue et al.'s instructed models do on BBQ's disambiguated contexts or on Winogender sentences that state the referent's gender. It is not known whether any published model's instructed bias reduction on BBQ is accompanied by the same reduction on implicitly cued items. It is not known at what scale, in any model family, an instruction against stereotyping changes answers on ambiguous items while leaving answers on disambiguated items intact; no study located here has measured that across scales. It is not known whether any model revises its own stereotyped first answers selectively, because no study located here has measured revision on a set where some first answers naming a person were correct.

16. Limitations

The paper's most consequential limitation is that several of the numbers it would most like to cite exist only as plotted points. Ganguli et al.'s disambiguated accuracies and ambiguous-context accuracies, and all of Liu, Xue et al.'s results, are reported in figures. The paper has stated directions where its sources state directions and has not read values off plots, which means the magnitude of the cost in the two anchor studies is unknown to this analysis as well as to their readers.

The matched-control evidence comes from settings that differ from the anchors'. Si et al.'s instruction explicitly told the model when to choose "unknown"; Deng et al.'s intervention was training on one example rather than an instruction; Wang, A. et al.'s benchmarks are newer and less widely validated than BBQ. Each supports the direction of the argument for its own setting, and the transfer to the anchors' exact setting is an inference.

The search was conducted through arXiv and Crossref and is not exhaustive. Studies that report disambiguated BBQ results for moral self-correction instructions on small models may exist and were not found; if they show preserved disambiguated accuracy at 3.8B, they would answer part of this paper's central question in the anchor's favour.

The paper inherits the limitations its sources state about the benchmarks themselves: BBQ, Winogender and the admissions task are English-language, United States-centred and template-based (Parrish et al., 2022; Ganguli et al., 2023), and bias scores are sensitive to construction choices that should not matter, as Selvam et al. (2022) showed when negating verbs moved one model's measured bias from 41.64 to 13.40 on another benchmark. Blodgett et al. (2020), surveying 146 papers, found that the motivations of bias measurement in NLP are often vague and poorly matched to its techniques. The argument here is about what the instruments can distinguish, not about whether they measure harm well, and it is not weakened by their other limitations, but neither does it repair them.

17. Conclusion

The dispute over whether moral self-correction emerges at 22 billion parameters or at 3.8 billion is a dispute about instruments that reward one named answer for every item: "unknown" on ambiguous BBQ questions, a neutral pronoun on Winogender, zero difference on admissions. On such instruments a model that ignores the context and gives the named answer scores perfectly, so a gain shows that the instruction moved the model, not that the model told apart the items where the move was warranted. In two of the three experiments that set the 22B number, the threshold coincides with the first appearance of measurable bias; the third, Winogender, shows a genuine scale effect in the ability to move toward a target, and the 3.8B result shows that ability

reached by a smaller, differently trained model, as instruction-following thresholds have moved before. Where published work has measured the items on which the named answer is wrong, the instructed or trained gain has often been paid for there, sometimes heavily; in a minority of cases it was not. Neither anchor study revised an answer, and the moral self-correction studies located here that did revise answers did so only on sets where revision toward the named answer could not be scored as wrong. What the field has established is that sufficiently trained models of a range of sizes are steerable toward a stated fairness target. Whether they discriminate, and whether they revise, is open, and the controls that would settle both are listed in Section 13.

References

- Abdin, M., Aneja, J., Awadalla, H., Awadallah, A., Awan, A. A., Bach, N., et al. (2024). Phi-3 Technical Report: A Highly Capable Language Model Locally on Your Phone. arXiv:2404.14219. doi:10.48550/arXiv.2404.14219
- Bai, X., Wang, A., Sucholutsky, I., & Griffiths, T. L. (2024). Measuring Implicit Bias in Explicitly Unbiased Large Language Models. arXiv:2402.04105. doi:10.48550/arXiv.2402.04105
- Blodgett, S. L., Barocas, S., Daume III, H., & Wallach, H. (2020). Language (Technology) is Power: A Critical Survey of "Bias" in NLP. arXiv:2005.14050. doi:10.48550/arXiv.2005.14050
- Cantini, R., Cosenza, G., Orsino, A., & Talia, D. (2024). Are Large Language Models Really Bias-Free? Jailbreak Prompts for Assessing Adversarial Robustness to Bias Elicitation. arXiv:2407.08441. doi:10.48550/arXiv.2407.08441
- Chen, B., Chen, X., Zi, H., Mao, H., Qi, Z., Zhang, X., Johnson, K., & Liu, G. (2026). Learning to Diagnose and Correct Moral Errors: Beyond Shallow Heuristics in Moral Alignment. arXiv:2601.03079. doi:10.48550/arXiv.2601.03079
- Chung, H. W., Hou, L., Longpre, S., Zoph, B., Tay, Y., Fedus, W., et al. (2022). Scaling Instruction-Finetuned Language Models. arXiv:2210.11416. doi:10.48550/arXiv.2210.11416
- Cui, J., Chiang, W.-L., Stoica, I., & Hsieh, C.-J. (2024). OR-Bench: An Over-Refusal Benchmark for Large Language Models. arXiv:2405.20947. doi:10.48550/arXiv.2405.20947
- Deng, N., Arif, S., Chang, S., Chen, Y., & Mihalcea, R. (2026). One Example Is Enough to Pass Fairness Benchmarks: Rethinking Fairness Evaluation for Aligned LLMs. arXiv:2609.14860. doi:10.48550/arXiv.2609.14860
- Furniturewala, S., Jandial, S., Java, A., Banerjee, P., Shahid, S., Bhatia, S., & Jaidka, K. (2024). Thinking Fair and Slow: On the Efficacy of Structured Prompts for Debiasing Language Models. arXiv:2405.10431. doi:10.48550/arXiv.2405.10431
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Yu, T., Deilamsalehy, H., Zhang, R., Kim, S., & Deroncourt, F. (2024a). Self-Debiasing Large Language Models: Zero-Shot Recognition and Reduction of Stereotypes. arXiv:2402.01981. doi:10.48550/arXiv.2402.01981
- Gallegos, I. O., Rossi, R. A., Barrow, J., Tanjim, M. M., Kim, S., Deroncourt, F., Yu, T., Zhang, R., & Ahmed, N. K. (2024b). Bias and Fairness in Large Language Models: A Survey. arXiv:2309.00770. doi:10.48550/arXiv.2309.00770
- Ganguli, D., Askell, A., Schiefer, N., Liao, T. I., Lukosiute, K., Chen, A., et al. (2023). The Capacity for Moral Self-Correction in Large Language Models. arXiv:2302.07459. doi:10.48550/arXiv.2302.07459
- Gupta, S., Shrivastava, V., Deshpande, A., Kalyan, A., Clark, P., Sabharwal, A., & Khot, T. (2023). Bias Runs Deep: Implicit Reasoning Biases in Persona-Assigned LLMs. arXiv:2311.04892. doi:10.48550/arXiv.2311.04892
- Hofmann, V., Kalluri, P. R., Jurafsky, D., & King, S. (2024). Dialect prejudice predicts AI decisions about people's character, employability, and criminality. arXiv:2403.00742 (published in revised form in Nature, 2024). doi:10.48550/arXiv.2403.00742
- Huang, J., Chen, X., Mishra, S., Zheng, H. S., Yu, A. W., Song, X., & Zhou, D. (2023). Large Language Models Cannot Self-Correct Reasoning Yet. ICLR 2024. arXiv:2310.01798. doi:10.48550/arXiv.2310.01798
- Jang, J., Ye, S., & Seo, M. (2022). Can Large Language Models Truly Understand Prompts? A Case Study with Negated Prompts. arXiv:2209.12711. doi:10.48550/arXiv.2209.12711
- Kamoi, R., Zhang, Y., Zhang, N., Han, J., & Zhang, R. (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. arXiv:2406.01297. doi:10.48550/arXiv.2406.01297
- Kamruzzaman, M., & Kim, G. L. (2024). Prompting Techniques for Reducing Social Bias in LLMs through System 1 and System 2 Cognitive Processes. arXiv:2404.17218. doi:10.48550/arXiv.2404.17218
- Kumar, A., Zhuang, V., Agarwal, R., Su, Y., Co-Reyes, J. D., Singh, A., et al. (2024). Training Language Models to Self-Correct via Reinforcement Learning. arXiv:2409.12917. doi:10.48550/arXiv.2409.12917
- Lee, A., Bai, X., Pres, I., Wattenberg, M., Kummerfeld, J. K., & Mihalcea, R. (2024). A Mechanistic Understanding of Alignment Algorithms: A Case Study on DPO and Toxicity. arXiv:2401.01967. doi:10.48550/arXiv.2401.01967
- Lee, Y.-T., Chang, F.-C., Shu, Y.-E., Shih, H.-Y., & Wu, P.-Y. (2025). Explaining Intrinsic Moral Self-Correction with Mechanistic Interpretability. arXiv:2505.11924. doi:10.48550/arXiv.2505.11924
- Lin, B. Y., Ravichander, A., Lu, X., Dziri, N., Sclar, M., Chandu, K., Bhagavatula, C., & Choi, Y. (2023). The Unlocking Spell on Base LLMs: Rethinking Alignment via In-Context Learning. arXiv:2312.01552. doi:10.48550/arXiv.2312.01552
- Liu, G., Mao, H., Cao, B., Xue, Z., Zhang, X., Wang, R., Tang, J., & Johnson, K. (2024). On the Intrinsic Self-Correction Capability of LLMs: Uncertainty and Latent Concept. arXiv:2406.02378. doi:10.48550/arXiv.2406.02378
- Liu, G., Mao, H., Cao, B., Xue, Z., Zhang, X., Wang, R., & Johnson, K. M. (2025). On the Convergence of Moral Self-Correction in Large Language Models. arXiv:2510.07290. doi:10.48550/arXiv.2510.07290
- Liu, G., Mao, H., Tang, J., & Johnson, K. M. (2024). Intrinsic Self-correction for Enhanced Morality: An Analysis of Internal Mechanisms and the Superficial Hypothesis. EMNLP 2024. arXiv:2407.15286. doi:10.48550/arXiv.2407.15286
- Liu, G., Qi, Z., Zhang, X., Cheng, L., & Johnson, K. M. (2024). Self-correction is Not An Innate Capability in Language Models. arXiv:2410.20513. doi:10.48550/arXiv.2410.20513
- Liu, G., Qi, Z., Zhang, X., Jiang, L., & Johnson, K. M. (2025a). Diagnosing Moral Reasoning Acquisition in Language Models: Pragmatics and Generalization. arXiv:2502.16600. doi:10.48550/arXiv.2502.16600
- Liu, G., Qi, Z., Zhang, X., & Johnson, K. M. (2025b). Discourse Heuristics For Paradoxically Moral Self-Correction. arXiv:2507.00985. doi:10.48550/arXiv.2507.00985
- Liu, G., Xue, Z., Zhang, X., Wang, R., & Johnson, K. M. (2024). Smaller Large Language Models Can Do Moral Self-Correction. arXiv:2410.23496. doi:10.48550/arXiv.2410.23496
- Madaan, A., Tandon, N., Gupta, P., Hallinan, S., Gao, L., Wiegrefe, S., et al. (2023). Self-Refine: Iterative Refinement with Self-Feedback. arXiv:2303.17651. doi:10.48550/arXiv.2303.17651
- McKenzie, I. R., Lyzhov, A., Pieler, M., Parrish, A., Mueller, A., Prabhu, A., et al. (2023). Inverse Scaling: When Bigger Isn't Better. Transactions on Machine Learning Research. arXiv:2306.09479. doi:10.48550/arXiv.2306.09479

- Mahendrakar, P. (2026a). Two Kinds of Missing: Underspecified Inputs Versus Unknown Answers, and Why One Abstention Policy Cannot Serve Both. *Life of Research*. doi:10.5281/zenodo.22901280
- Mahendrakar, P. (2026b). Refusal Is Not a Rate: What a Context-Conditioned Refusal Policy Would Have to Specify, Why the Benchmarks That Score Context Assume It Is Verified, and the Provenance Term No Instrument Prices. *Life of Research*. doi:10.5281/zenodo.22984335
- Mahendrakar, P. (2026c). The Self-Verification Gap: Why Working Hallucination Detectors Are External, and What Real-Time Correction Would Actually Require. *Life of Research*. doi:10.5281/zenodo.22030500
- Nadeem, M., Bethke, A., & Reddy, S. (2020). StereoSet: Measuring stereotypical bias in pretrained language models. arXiv:2004.09456. doi:10.48550/arXiv.2004.09456
- Pan, L., Saxon, M., Xu, W., Nathani, D., Wang, X., & Wang, W. Y. (2023). Automatically Correcting Large Language Models: Surveying the landscape of diverse self-correction strategies. arXiv:2308.03188. doi:10.48550/arXiv.2308.03188
- Panickssery, N., Gabrieli, N., Schulz, J., Tong, M., Hubinger, E., & Turner, A. M. (2023). Steering Llama 2 via Contrastive Activation Addition. arXiv:2312.06681. doi:10.48550/arXiv.2312.06681
- Parrish, A., Chen, A., Nangia, N., Padmakumar, V., Phang, J., Thompson, J., Htut, P. M., & Bowman, S. R. (2022). BBQ: A Hand-Built Bias Benchmark for Question Answering. *Findings of ACL 2022*. arXiv:2110.08193. doi:10.48550/arXiv.2110.08193
- Qi, X., Panda, A., Lyu, K., Ma, X., Roy, S., Beirami, A., Mittal, P., & Henderson, P. (2024). Safety Alignment Should Be Made More Than Just a Few Tokens Deep. arXiv:2406.05946. doi:10.48550/arXiv.2406.05946
- Rottger, P., Kirk, H. R., Vidgen, B., Attanasio, G., Bianchi, F., & Hovy, D. (2023). XSTest: A Test Suite for Identifying Exaggerated Safety Behaviours in Large Language Models. arXiv:2308.01263. doi:10.48550/arXiv.2308.01263
- Rudinger, R., Naradowsky, J., Leonard, B., & Van Durme, B. (2018). Gender Bias in Coreference Resolution. arXiv:1804.09301. doi:10.48550/arXiv.1804.09301
- Schaeffer, R., Miranda, B., & Koyejo, S. (2023). Are Emergent Abilities of Large Language Models a Mirage? arXiv:2304.15004. doi:10.48550/arXiv.2304.15004
- Schick, T., Udupa, S., & Schutze, H. (2021). Self-Diagnosis and Self-Debiasing: A Proposal for Reducing Corpus-Based Bias in NLP. arXiv:2103.00453. doi:10.48550/arXiv.2103.00453
- Selvam, N. R., Dev, S., Khashabi, D., Khot, T., & Chang, K.-W. (2022). The Tail Wagging the Dog: Dataset Construction Biases of Social Bias Benchmarks. arXiv:2210.10040. doi:10.48550/arXiv.2210.10040
- Shaikh, O., Zhang, H., Held, W., Bernstein, M., & Yang, D. (2023). On Second Thought, Let's Not Think Step by Step! Bias and Toxicity in Zero-Shot Reasoning. *ACL 2023*. arXiv:2212.08061. doi:10.48550/arXiv.2212.08061
- Sharma, M., Tong, M., Korbak, T., Duvenaud, D., Askeel, A., Bowman, S. R., et al. (2023). Towards Understanding Sycophancy in Language Models. arXiv:2310.13548. doi:10.48550/arXiv.2310.13548
- Si, C., Gan, Z., Yang, Z., Wang, S., Wang, J., Boyd-Graber, J., & Wang, L. (2023). Prompting GPT-3 To Be Reliable. *ICLR 2023*. arXiv:2210.09150. doi:10.48550/arXiv.2210.09150
- Tamkin, A., Askeel, A., Lovitt, L., Durmus, E., Joseph, N., Kravec, S., Nguyen, K., Kaplan, J., & Ganguli, D. (2023). Evaluating and Mitigating Discrimination in Language Model Decisions. arXiv:2312.03689. doi:10.48550/arXiv.2312.03689
- Tsui, K. (2025). Self-Correction Bench: Uncovering and Addressing the Self-Correction Blind Spot in Large Language Models. arXiv:2507.02778. doi:10.48550/arXiv.2507.02778
- Turner, A. M., Thiergart, L., Leech, G., Udell, D., Vazquez, J. J., Mini, U., & MacDiarmid, M. (2023). Steering Language Models With Activation Engineering. arXiv:2308.10248. doi:10.48550/arXiv.2308.10248
- Turpin, M., Michael, J., Perez, E., & Bowman, S. R. (2023). Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting. *NeurIPS 2023*. arXiv:2305.04388. doi:10.48550/arXiv.2305.04388
- Tyen, G., Mansoor, H., Carbune, V., Chen, P., & Mak, T. (2023). LLMs cannot find reasoning errors, but can correct them given the error location. arXiv:2311.08516. doi:10.48550/arXiv.2311.08516
- Wagh, A., & Srivastava, S. (2025). "The Dentist is an involved parent, the bartender is not": Revealing Implicit Biases in QA with Implicit BBQ. arXiv:2512.06732. doi:10.48550/arXiv.2512.06732
- Wang, A., Phan, M., Ho, D. E., & Koyejo, S. (2025). Fairness through Difference Awareness: Measuring Desired Group Discrimination in LLMs. arXiv:2502.01926. doi:10.48550/arXiv.2502.01926
- Wang, Y., Wu, Y., Wei, Z., Jegelka, S., & Wang, Y. (2024). A Theoretical Understanding of Self-Correction through In-context Alignment. arXiv:2405.18634. doi:10.48550/arXiv.2405.18634
- Wei, J., Bosma, M., Zhao, V. Y., Guu, K., Yu, A. W., Lester, B., Du, N., Dai, A. M., & Le, Q. V. (2021). Finetuned Language Models Are Zero-Shot Learners. arXiv:2109.01652. doi:10.48550/arXiv.2109.01652
- Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., et al. (2022a). Emergent Abilities of Large Language Models. arXiv:2206.07682. doi:10.48550/arXiv.2206.07682
- Wei, J., Kim, N., Tay, Y., & Le, Q. V. (2022b). Inverse scaling can become U-shaped. arXiv:2211.02011. doi:10.48550/arXiv.2211.02011
- Wu, X., Nian, J., Wei, T.-R., Tao, Z., Wu, H.-T., & Fang, Y. (2025). Does Reasoning Introduce Bias? A Study of Social Bias Evaluation and Mitigation in LLM Reasoning. arXiv:2502.15361. doi:10.48550/arXiv.2502.15361
- Zhang, Q., Wang, D., Qian, H., Li, Y., Zhang, T., Huang, M., Xu, K., Li, H., et al. (2024). Understanding the Dark Side of LLMs' Intrinsic Self-Correction. arXiv:2412.14959. doi:10.48550/arXiv.2412.14959
- Zhao, J., Khashabi, D., Khot, T., Sabharwal, A., & Chang, K.-W. (2021). Ethical-Advice Taker: Do Language Models Understand Natural Language Interventions? arXiv:2106.01465. doi:10.48550/arXiv.2106.01465
- Zhou, C., Liu, P., Xu, P., Iyer, S., Sun, J., Mao, Y., et al. (2023). LIMA: Less Is More for Alignment. arXiv:2305.11206. doi:10.48550/arXiv.2305.11206
- Zou, A., Phan, L., Chen, S., Campbell, J., Guo, P., Ren, R., et al. (2023). Representation Engineering: A Top-Down Approach to AI Transparency. arXiv:2310.01405. doi:10.48550/arXiv.2310.01405