A Course Correction in Steerability Evaluation: Revealing Miscalibration and Side Effects in LLMs

Trenton Chang[1], **Tobias Schnabel**[2], **Adith Swaminathan**[3], **Jenna Wiens**[1]

[1] University of Michigan
[2]Microsoft Research
[3]Netflix

Abstract

Despite advances in large language models (LLMs) on reasoning and instruction-following tasks, it is unclear whether they can reliably produce outputs aligned with a variety of user goals, a concept called *steerability*. Two gaps in current LLM evaluation impede steerability evaluation: (1) many benchmarks are built with past LLM chats and Internet-scraped text, which may skew towards common requests, and (2) scalar measures of performance common in prior work could conceal behavioral shifts in LLM outputs in open-ended generation. Thus, we introduce a framework based on a multi-dimensional goal-space that models user goals and LLM outputs as vectors with dimensions corresponding to text attributes (e.g., reading difficulty). Applied to a text-rewriting task, we find that current LLMs induce unintended changes or *side effects* to text attributes, impeding steerability. Interventions to improve steerability, such as prompt engineering, best-of-N sampling, and reinforcement learning fine-tuning, have varying effectiveness but side effects remain problematic. Our findings suggest that even strong LLMs struggle with steerability, and existing alignment strategies may be insufficient.

1 Introduction

Large language models (LLMs) continue to advance on reasoning and instruction-following benchmarks (Zhong et al. 2025; Dong et al. 2025). However, these gains may not yield models that reliably satisfy a wide set of specific user goals, a property called *steerability* (Vafa et al. 2025; Li et al. 2024; Miehling et al. 2025). Two fundamental limitations in current benchmark and metric design impede progress in steerability evaluation. First, many benchmarks sample data representative of real-world chat interactions (Köpf et al. 2023; Zhao et al. 2024) or text scraped from the Internet (Raffel et al. 2020). Such data skew toward common requests and miss rarer combinations of goals. For example, making text "more formal and longer" is frequent, whereas "more formal but shorter" is rare; a benchmark that uniformly samples the goal-space can evaluate steering towards both equally. Second, many benchmarks implicitly treat performance as scalar. While potentially suitable for tasks such as instruction-following (Zhou et al. 2023) or question-answering (Hendrycks et al. 2021), single-dimensional metrics cannot measure changes in other dimensions of behavior that may arise in open-ended generation (*e.g.*, (Durmus et al. 2024)). Left unmeasured, such behavioral shifts could conceal harmful behavior (e.g., sycophancy).

We design a steerability evaluation framework to address these gaps, focusing on *steering tasks* in which users aim to *transform* texts in specific ways. We map user goals and LLM outputs into a shared, multi-dimensional *goal-space* with text-to-scalar functions, from which we sample a *steerability probe* comprised of equally-weighted goals. A multi-dimensional goal-space allows us to measure behaviors such as *miscalibration*: too much/too little change along the requested direction, or *side effects*: unintended shifts in dimensions orthogonal to user goals (Amodei et al. 2016). Figure 1 illustrates our framework, showing a user requesting changes in reading level and text length.

Using our proposed framework, we show empirical evidence of challenges in steerability in a text rewriting task. We choose a small number of rule-based goal dimensions to disentangle the accuracy of goal measurement from steerability, and to aid in interpretation of results. First, we find that side effects remain pervasive, even in strong LLMs. As potential mitigation strategies, we find prompt engineering ineffective, while a best-of-N sampling is effective, suggesting that side effects can be mitigated, yet costly due to the need for repeated prompting. A fine-tuning approach (reinforcement learning; RL) improves steerability, but side effects remain. Our results suggests that, while some strategies show promise, they remain insufficient to solve side effects.

In summary, this work makes the following contributions:

- Define steerability as distance in a multi-dimensional goal-space and decompose steering error into miscalibration and side effects (Section 2).
- Build a steerability probe from existing corpora that uniformly samples goals over diverse source texts (Section 3).
- Benchmark LLMs and show that side effects are pervasive across model families (Section 4.1).
- Evaluate inference-time steerability interventions and find prompt engineering ineffective, while best-of-N sampling is effective but costly (Section 4.2).

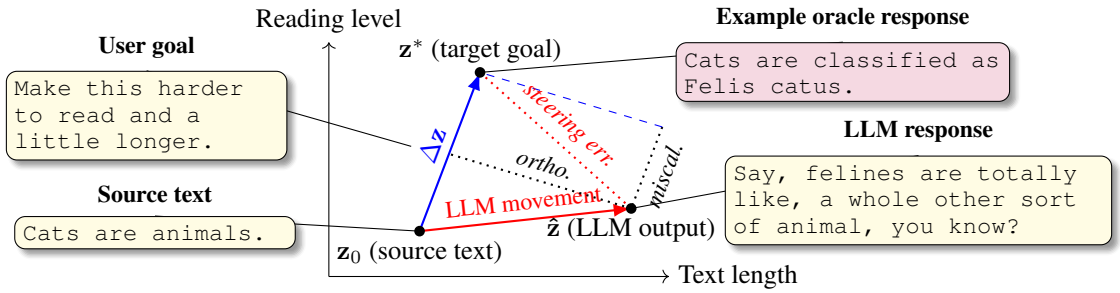

Figure 1: Steerability metrics in 2D goal-space (reading level & text length). A user aims to rewrite text according to some intent, expressed via a prompt (`Make this harder to read...`). The steering error (red dotted line) is the gap between the user's intent (blue) and the LLM's output (red). Miscalibration (miscal.) and orthogonality (ortho.) decompose steering error into components parallel and orthogonal to user intent respectively.

- Fine-tune with RL to reduce steering error, though side effects remain (Section 4.3).

The strength of side effects in LLMs across models highlights a potential gap between current LLM capabilities and steerability. We hope our framework can improve evaluation of LLM alignment with diverse sets of human goals, complementing current measures of LLM capability.

2 A Steerability Measurement Framework

We aim to measure how well a model follows structured, multi-dimensional user goals in a single-turn setting; e.g., text-rewriting. Here, we formalize steerability in the context of LLM evaluation (Section 2.1), and introduce metrics for LLM performance in the space of user goals (Section 2.2).

2.1 Designing a steerability metric

We aim to evaluate the steerability of a conditional generative model f, which produces outputs $y \in \mathcal{Y}$ via sampling $y \sim f(\cdot \mid x)$ given input $x \in \mathcal{X}$. To evaluate f, one generally measures performance over some user goals $\mathbf{z}^* \sim P(\cdot)$, also called *targets*, where users verbalize intents $\mathbf{z}^*$ via $x \sim P(\cdot \mid \mathbf{z}^*)$. Such metrics consist of (i) an aggregation function over intents $\mathbf{z}^*$ and (ii) a loss function $\ell(\cdot, \mathbf{z}^*)$ that captures concordance between outputs and targets $\mathbf{z}^*$:

$$\text{metric}(f) \triangleq \mathbb{E}_{\mathbf{z}^* \sim P(\cdot)} \mathbb{E}_{x \sim P(\cdot|\mathbf{z}^*)} \mathbb{E}_{y \sim f(\cdot|x)} \, \ell(y, \mathbf{z}^*) \quad (1)$$

Prior LLM evaluations choose different aggregation and loss functions, summarized in Table 1. Instruction following tasks often use a binary ℓ (e.g., correctness (Qin et al. 2024; Zhou et al. 2023; He et al. 2024)) and implicitly assume a small set of canonical goals (e.g., instruction types). 1D metrics define a continuous ℓ (e.g., P(desired behavior) (Rimsky et al. 2024; Turner et al. 2023); concept detection "scores" (Wu et al. 2025)), which returns a scalar. Ranking accuracy-based losses (Ouyang et al. 2022; Rafailov et al. 2023) emphasize *relative* rather than absolute response quality. Some evaluations rely on chat log data or web scraping (Köpf et al. 2023; Zhao et al. 2024; Raffel et al. 2020), or are purpose-built to test specific capabilities (Zhou et al. 2023; BIG-Bench contributors 2023; Hendrycks et al. 2021), which may not be representative of potential users and goals.

These shortcomings may be especially pronounced in *steering tasks* where users aim to transform model outputs along multi-dimensional, multi-level dimensions, such as text-rewriting. In particular, steering tasks may contain a wider range of potential user goals than typically seen in benchmarks. Such tasks may also expose miscalibration, as coarse metrics such as binary accuracy/rankings can lead models to score distinct responses identically, flattening different types of deviations from the user's intent. In addition, since steering tasks may include requests for multi-dimensional changes to text, single dimensional metrics may hide unintended side effects in LLM responses.

We contribute a steerability metric that addresses these limitations by (i) aggregating over a *uniform* distribution of goals, allowing us to better identify poor coverage, and (ii) using a loss function ℓ that measures absolute distance between target goals and model outputs in multiple dimensions. Specifically, let $\mathbf{z}_0$ be a source that to be transformed, and let $\hat{\mathbf{z}}$ be the intent satisfied by the LLM output. Recall that $\mathbf{z}^*$ is the user's intent. Treating $\mathbf{z}^*$, $\hat{\mathbf{z}}$ and $\mathbf{z}_0$ as elements of a shared metric space $\mathcal{Z}$ (*e.g.*, Fig. 1), we write:

$$\text{steerability}(f) \triangleq \mathbb{E}_{\mathbf{z}_0, \mathbf{z}^* \sim \mathcal{U}} \, \mathbb{E}_{\hat{\mathbf{z}} \sim f(\cdot|\mathbf{z}_0, \mathbf{z}^*)} [\|\hat{\mathbf{z}} - \mathbf{z}^*\|_2] \quad (2)$$

where $\mathcal{U}$ is a uniform distribution over $\mathbf{z}_0$ and $\mathbf{z}^*$.

2.2 Measuring steerability in practice

Our steerability metric (Eq. 2) puts $\mathbf{z}^*$, $\mathbf{z}_0$, and $\hat{\mathbf{z}}$ in a shared space $\mathcal{Z}$. To define $\mathcal{Z}$, we observe that user goals $\mathbf{z}^*$ for steering tasks often decompose along interpretable dimensions (e.g., "Make this harder to read and a little longer"). Thus, we define $\mathcal{Z}$ to be a set of *dimensions* representing attributes of text (e.g., reading level and length). Formally, define *goal-space* $\mathcal{Z} = [0,1]^{|\mathcal{G}|}$, and functions $g_i : \mathcal{Y} \to [0,1]$ for $i \in 1, \ldots, |\mathcal{G}|$ that translate model outputs $y \sim f(\cdot \mid x)$ into goal-space, where g_i can be based on existing measures of text features (e.g., Flesch-Kincaid grade level (Kincaid et al. 1975), word count). The joint output of all g_i is the *goal-space mapping* of y; *i.e.*, a vector representation of y.

As an example, consider measuring steerability in text-rewriting (Figure 1). A user aims to rewrite a *source* (e.g., "Cats are animals") mapping to $\mathbf{z}_0$ in goal-space. Suppose

Metric	Scoring function (ℓ)	Example evaluation dataset ($P(\mathbf{z}^*, \mathbf{z}_0)$)
Correctness (binary accuracy)	$\mathbf{1}[\hat{\mathbf{z}} = \mathbf{z}^*]$	Instruction-following, reasoning benchmarks (*e.g.*, math)
Ranking accuracy	$\mathbf{1}[R(\hat{\mathbf{z}}_i) > R(\hat{\mathbf{z}}_j)]$ (R: reward)	Pairwise or ordinal preference rankings
Scalar/1D accuracy	$z^* - \hat{z}$; $z^*, \hat{z} \in \mathbb{R}$	Questionnaire-style behavior probes
Steerability (proposed)	$\|\mathbf{z}^* - \hat{\mathbf{z}}\|_2$; $\mathbf{z}^*, \hat{\mathbf{z}} \in \mathbb{R}^n$	Steerability sampled uniformly on $(\mathbf{z}^*, \mathbf{z}_0) \in \mathcal{Z}$

Table 1: Comparison of single-turn LLM evaluation strategies by scoring/loss function and a representative evaluation dataset in terms of LLM output and user goal.

that the user wants a harder-to-read, slightly longer text, which maps to $\mathbf{z}^*$, expressed via a prompt (e.g., "Make this harder to read and a little longer"). We assume $\mathbf{z}^*$ is *feasible*; *i.e.*, it is possible to make the source harder to read and slightly longer. The LLM produces an output (e.g., "Say, felines are totally like...") satisfying intent $\hat{\mathbf{z}}$, which may not match $\mathbf{z}^*$. We quantify the mismatch via *steering error*; *i.e.*, the Euclidean distance between $\mathbf{z}^*$ and $\hat{\mathbf{z}}$ in multi-dimensional goal-space. To ensure *coverage*, we average over a uniform sample of $\mathbf{z}_0$ and $\mathbf{z}^*$, yielding Eq. 2.

However, steering error ($\|\mathbf{z}^* - \hat{\mathbf{z}}\|_2$) does not distinguish miscalibration, or errors in magnitude, from side effects, or errors due to unintended changes. To address this, we write:

$$\|\mathbf{z}^* - \hat{\mathbf{z}}\|_2 = \|(\mathbf{z}^* - \mathbf{z}_0) - (\hat{\mathbf{z}} - \mathbf{z}_0)\|_2. \quad (3)$$

Now consider the orthogonal decomposition of Eq. 3 onto the desired movement vector ($\mathbf{z}^* - \mathbf{z}_0$), yielding $\text{proj}_{\mathbf{z}^* - \mathbf{z}_0}(\mathbf{z}^* - \hat{\mathbf{z}})$ and $\text{proj}^{\perp}_{\mathbf{z}^* - \mathbf{z}_0}(\mathbf{z}^* - \hat{\mathbf{z}})$, respectively. The *scalar* projection (sproj$(\cdot)$), or magnitude of these vectors, correspond to steering error along the direction of the user's intent (*miscalibration*) and the orthogonal error (*orthogonality*), respectively. We normalize the scalar projections to account for the "severity" of the error:

$$\text{miscal}(\mathbf{z}^*, \hat{\mathbf{z}} \mid \mathbf{z}_0) = \text{sproj}_{\mathbf{z}^* - \mathbf{z}_0}(\mathbf{z}^* - \hat{\mathbf{z}}) / \|\mathbf{z}^* - \mathbf{z}_0\|_2 \quad (4)$$

where miscalibration, or over/under-shooting in the direction of the intent, is normalized by requested movement $\|\mathbf{z}^* - \mathbf{z}_0\|_2$. Orthogonality is normalized by observed movement $\|\hat{\mathbf{z}} - \mathbf{z}_0\|_2$:

$$\text{ortho}(\mathbf{z}^*, \hat{\mathbf{z}} \mid \mathbf{z}_0) = \text{sproj}^{\perp}_{\mathbf{z}^* - \mathbf{z}_0}(\mathbf{z}^* - \hat{\mathbf{z}}) / \|\hat{\mathbf{z}} - \mathbf{z}_0\|_2 \quad (5)$$

so that orthogonality corresponds to the proportion of goal-space movement orthogonal to the intent. These normalization steps broadly ensure that errors are penalized in proportion to the amount of requested or observed movement. All of these metrics are non-negative and minimized at zero.

3 Experimental Setup

Steerability probes are benchmarks designed to measure steerability for a steering task. We describe how we construct an example steerability probe for text-rewriting (Section 3.1), candidate steerability interventions evaluated (Section 3.2), and our inference setup (Section 3.3).

3.1 Steerability probe construction

We measure steerability in text-rewriting, a common task likely well-represented in LLM training data. Our probe has two components: (i) goal dimensions defining a goal-space $\mathcal{Z}$ and (ii) a dataset of goals $(\mathbf{z}_0, \mathbf{z}^*) \sim \mathcal{Z}$.

Design principles. Goal-space can be constructed from any set of measurable goals. For this first study, we use goals measured by rule-based evaluators. Rule-based evaluators are deterministic and auditable, which facilitates interpretation of results over learned or model-based evaluators. Otherwise, observed steering error may reflect evaluator error rather than the LLM being tested. However, our choice of evaluators is illustrative, not normative: our framework is modular and can use well-validated learned evaluators without changing the metric definitions. To obtain a diverse sample of source texts, we combine datasets with diverse writing styles, from which a more uniform set can be sampled. We report additional details in the Appendix. [1]

Goal-space. We select reading difficulty (Flesch-Kincaid grade (Kincaid et al. 1975)), formality (Heylighen-Dewaele F-score (Heylighen and Dewaele 1999)), textual lexical diversity (Jarvis and Hashimoto 2021), and text length (word count). Though these dimensions may be correlated in training data, each is independently manipulable in theory (e.g., syllables per word & sentence length affect Flesch-Kincaid; whereas Heylighen-Dewaele measures the part-of-speech distribution). Requests mentioning these dimensions appear in real-world chats (e.g. WildChat/LMSys (Zhao et al. 2024), Appendix). Metric descriptions are in the Appendix. For RL fine-tuning, we focus on 2D goal-space (reading difficulty, formality) to isolate challenges in steerability in a simple setting where goal dimensions are conceptually distinct but likely correlated in real-world text. As a secondary validity check, we verify that LLM-as-judge can detect changes in all chosen goal dimensions (see Appendix).

Source texts and goals. We sample seed texts from news articles (CNN/DailyMail (See, Liu, and Manning 2017)), social media (RedditTIFU (Kim, Kim, and Kim 2019)), English novels (BookSum (Kryściński et al. 2022)), and movie synopses (SummScreenFD (Shaham et al. 2022)), to cover a wide stylistic range (total $N = 8,303$). We compute goal-space mappings for seed texts and min-max scale the empirical middle 95% of each goal dimension to $[0, 1]$, clipping values outside that range, such that goal dimensions are on comparable scales. We then resample $\mathbf{z}_0$ to be uniform over over goal-space $\mathcal{Z}$ via reweighting. For each $\mathbf{z}_0$,

[1] See full technical Appendix: https://arxiv.org/abs/2505.23816.

we choose three *active* goal dimensions at random, and sample $\mathbf{z}^*$ within ± 0.1 to 0.7 of the original value, copying components of $\mathbf{z}_0$ to $\mathbf{z}^*$ for inactive dimensions. Our main probe consists of 64 source texts with 32 goals each ($N = 2,048$). All reported results are statistically significant at level $\alpha = 0.05$ based on a paired, two-sided Wilcoxon rank-signed test, with other tests used as specified.

For RL fine-tuning, our training probe consists of 384 source texts with 16 goals each ($N = 3,072$). We select *one* active goal dimension and report metrics post-RL on 64 held-out source texts with 16 goals each ($N = 1,024$) in 2D goal-space with one active goal dimension unless specified.

Default prompt. To turn $(\mathbf{z}_0, \mathbf{z}^*)$ into prompts, we use a template-based prompt that names active goal dimensions with modifiers "slightly" for changes < 0.2, and "much" when changes are > 0.5, and no modifier otherwise (e.g., "make this [slightly/much] [more/less] formal;" see also Appendix). To avoid penalizing prompt ambiguity rather than steerability failures, we discretize $\mathbf{z}^*$ and $\hat{\mathbf{z}}$ using the same bins implied by the prompt modifiers (cut points at 0, ± 0.2 and ± 0.5) when reporting steerability metrics.

3.2 Candidate steerability interventions

We evaluate common single-turn techniques for influencing model behavior. We choose a set of methods applicable to multi-dimensional, multi-level intents, namely, prompt engineering, best-of-N sampling, and RL fine-tuning.

Prompt engineering. Prompt engineering is the design of a strategy for verbalizing intent $\mathbf{z}^*$, which may span direct instructions (e.g., Figure 1), chain-of-thought (Wei et al. 2022), or negative prompting (e.g., "don't change anything else") (Sanchez et al. 2024). We extend the default prompt by testing the inclusion of negative prompts and specific instructions (e.g., "increase formality by changing X"), a chain-of-thought style directive (e.g., "explain proposed edits"), and an underspecified prompt as a naive upper bound on steering error. While non-exhaustive, this set reflects common strategies proposed in prior work applicable to text rewriting. See the Appendix for examples.

Best-of-N sampling. Best-of-N selects the response with the lowest steering error out of N attempts, assessing whether models are even capable of producing responses with low steering error. To encourage diverse but fluent samples, we use min-p sampling ($p = 0.2$) with temperature 1 and a 0.1 frequency penalty (Minh et al. 2025).

RL fine-tuning. RL fine-tuning optimizes model parameters via online RL, using steering error as the negative reward. Since sampling directly from uniform $\mathcal{U}$ may be infeasible, we reweight training examples from a dataset $\mathcal{D}$ by estimating the density ratio $\mathcal{U}/\mathcal{D}$ via classifier-based methods (Bickel and Scheffer 2009):

$$\min_f \mathbb{E}_{(\mathbf{z}_0,\mathbf{z}^*)\sim\mathcal{D}} \mathbb{E}_{\hat{\mathbf{z}}\sim f(\cdot|\mathbf{z}_0,\mathbf{z}^*)}[\hat{w}(\mathbf{z}_0,\mathbf{z}^*)\cdot\|\mathbf{z}^*-\hat{\mathbf{z}}\|_2^2]. \quad (6)$$

To optimize Eq. 6, we use a policy gradient method based on leave-one-out proximal policy optimization (LOOP) (Chen et al. 2025). We fine-tune a Llama3.1-8B model via rank-stabilized LoRA (Kalajdzievski 2023). We generate rollouts using the same decoding parameters as best-of-N sampling. We discuss modifications to LOOP and hyperparameters in the Appendix.

3.3 LLM inference setup

Models. We evaluate GPT (3.5 turbo, 4 turbo, 4o, 4.1 (OpenAI 2023, 2024b,a)), Llama3 (Llama3 to 3.3, 8B/70B (Meta AI 2024)), Deepseek-R1 variants (8B/70B, distilled (DeepSeek-AI Team 2025)), and Qwen3 (4B/32B/30B-A3B) (Qwen Team 2025), and o1-/o3-mini (OpenAI 2024c, 2025). LLM inference is performed using the OpenAI API (GPT) or vLLM (all others) (Kwon et al. 2023), with greedy sampling and a context length of 32,000 tokens unless specified.

Output post-processing. To ensure metrics are computed over valid rewrites, we post-process responses to remove boilerplate text (e.g., "Sure, here's...") and reasoning tokens (e.g., `<think>` blocks). We also filter refusals, degenerate behavior (e.g., repetitive looping), or rewrites unrelated to the source using LLM-as-judge and manual review of responses flagged by the LLM (see Appendix).[2]

4 Empirical Results

We evaluate steerability in text-rewriting using the proposed metrics. Our results suggest that current LLMs are not steerable, which we largely attribute to side effects. Further analysis suggests goal dimensions may be spuriously entangled (Section 4.1). As candidate interventions, we try prompt engineering, which is ineffective, and best-of-N sampling, which requires extensive sampling (Section 4.2). We then try RL fine-tuning in 2D goal-space, which rivals best-of-128 and disentangles goals, but side effects remain (Section 4.3).

4.1 Large language models are not steerable

Even strong LLMs induce side effects. Neither larger nor newer models meaningfully improve steering error.[3] Median steering error remains high, 0.452 for the largest model (Llama-3.3; Figure 2, left), far from ideal despite outperforming a random baseline (0.770; sampling random goal levels in each dimension). Miscalibration improves (Figure 2, center) with model size (e.g., Llama3.1-8B vs. 70B: $0.667 \rightarrow 0.455$). Some residual miscalibration is expected, since the model may not be calibrated to the magnitude of "slightly/much" in our prompts.

Median orthogonality remains high and skewed towards 1 even as model size increases (Figure 2, right) with Llama3.3-70B performing best with an orthogonality of 0.718. While several pairwise differences are statistically significant, models remain in a high-orthogonality regime on average. Similar trends hold in GPT, Deepseek, Qwen3, and o1/o3 models, where larger/newer models reduce miscalibration but have little effect on orthogonality (see Appendix). Note that, even as miscalibration and orthogonality

[2]Due to filtering, metrics are reported on slightly different response distributions. The effect is negligible: in our main results, rejected responses comprise $\leq$6 ($\approx$0.29%) of outputs in any probe.

[3]Some pairwise tests yield statistical significance, but effect sizes are small.

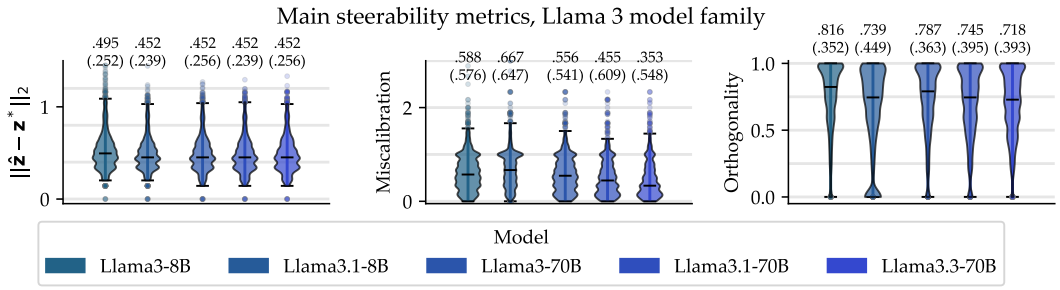

Figure 2: Median (IQR) of steering error (left), miscalibration (middle), and orthogonality, Llama3 family. Caps denote empirical 95% CI with outliers (○) plotted individually. Steering error does not improve with model size (left), but miscalibration does (middle). Orthogonality drops slightly (right), but remains skewed away from 0.

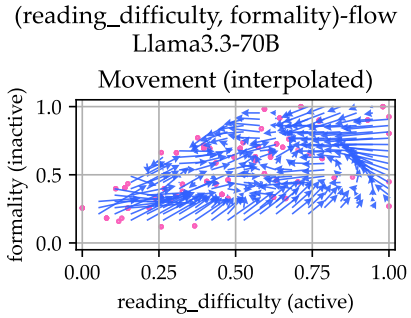

Figure 3: Vector flow of goal-space movement (blue), Llama3.3-70B, in requests to change reading difficulty but not formality. Horizontal movement is desired, but not vertical movement. Source texts in red.

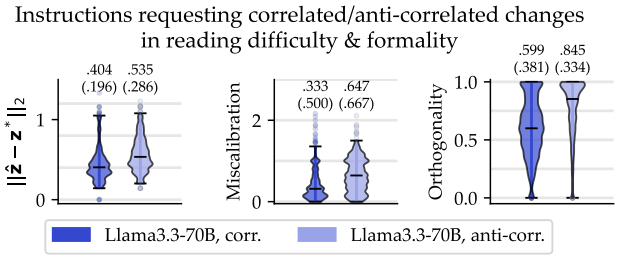

Figure 4: Median and IQR steerability, Llama3.3-70B, in correlated (darker) vs. anti-correlated (lighter) requests for change in reading difficulty and formality. Caps denote empirical 95% CI with outliers (○) plotted individually. Llama3.3-70B struggles more with anti-correlated changes.

decrease, median steering error may not due to normalization (Eq. 4-5; errors are penalized in proportion to the requested/observed movement). To further study side effects, we analyze a 2D goal subspace.

Goal dimensions may be entangled. We investigate side effects in a 2D (reading difficulty, formality) subspace using a vector flow diagram of goal-space movement (Figure 3, Llama3.3-70B, blue vectors). We include instructions requesting changes to reading difficulty (x-axis) but not formality (y-axis), such that vertical movement is a side effect. Figure 3 shows a "current" from the lower left (informal & easy to read) to the top right, suggesting that, when asked to increase reading difficulty without direction on formality, LLMs still increase formality.

The Appendix shows additional movement vectors and flows. We also conduct a preliminary study of coupling between goal dimensions, which suggests that the entanglement is LLM-induced.

While harder-to-read texts are often more formal, they need not be under our chosen measurement functions (Flesch-Kincaid grade, reading difficulty; Heylighen-Dewaele score, formality). LLM behavior appears to reflect this correlation: when stratifying steerability probe results based on whether the prompt requested correlated (e.g., make it harder to read and more formal) vs. anti-correlated changes to reading difficulty and formality (e.g., make it harder to read and *less* formal), Llama3.3-70B is less steerable on anti-correlated requests compared to correlated requests (steering error, 0.535 vs. 0.404; Figure 4, diff.: 0.131, Mann-Whitney $U = 77944.5$), with similar results in other model families (GPT, Deepseek, Qwen3; see Appendix). Thus, side effects may harm steerability in requests running contrary to similar correlations.

On coverage. While our probe is designed to target a uniform distribution of goals, results are similar whether or not we sample source texts uniformly in goal-space (see Appendix). Thus, steerability failures are unlikely to be concentrated in overrepresented goals in our evaluation.

Takeaway #1: side effects impede steerability. Despite progress in LLM reasoning and model capacity, LLMs continue to exhibit side effects. Entanglement between goal dimensions contributes to side effects, limiting steerability for intents that contradict correlations between goal dimensions.

4.2 Inference-time steering is costly

We now study whether inference-time strategies can improve steerability. First, we evaluate whether prompt engineering can elicit responses that satisfy user goals. Second, we leverage best-of-N sampling to test whether such responses are in the support of the model's output distribution.

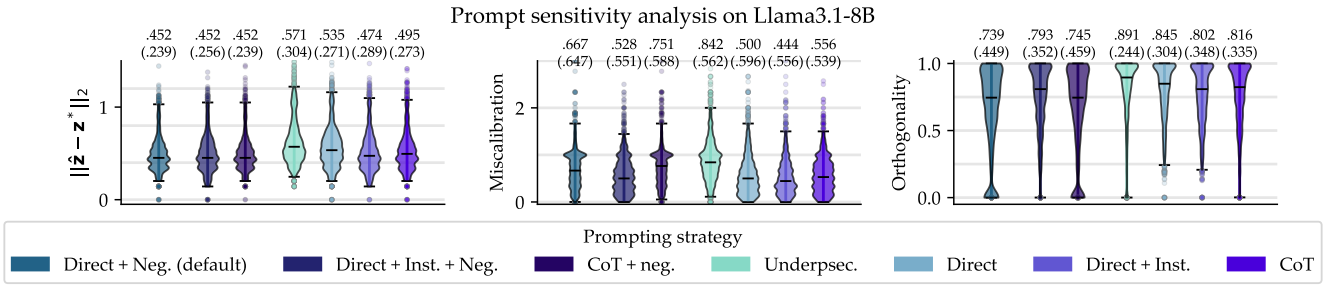

Figure 5: Median and IQR of steering error (left), miscalibration (middle), and orthogonality (right) of Llama3.1-8B across prompting strategies. Caps denote empirical 95% CI with outliers (○) plotted individually. More detailed prompts and removal of the negative prompt marginally improve miscalibration over the default. However, side effects remain severe.

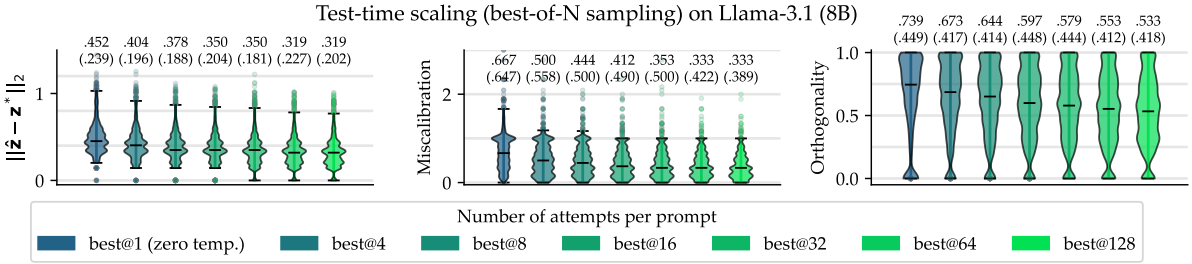

Figure 6: Median and IQR for best-of-$\{4, 8, \ldots, 128\}$ approaches on Llama3.1-8B, with a direct + negative prompt. Caps denote empirical 95% CI with outliers (○) plotted individually. Increasing N improves steerability, but improvements are slow.

Prompt engineering does not solve side effects. More detailed prompting strategies compared to the default (e.g., chain-of-thought style or adding instructions) tend to improve miscalibration, as does removing the negative prompt (median: 0.667 → 0.444, no negative prompt + added instructions; Figure 5, middle). Yet orthogonality remains skewed towards 1 despite improvements under some strategies (e.g., direct + negative prompts; Figure 5, right). Thus, mitigating side effects with prompt engineering alone may be challenging. Results for all strategies are in the Appendix.

Best-of-N sampling is a costly solution. Since side effects remain severe across prompting strategies, we investigate whether responses that reduce side effects exist in the model's sampling distribution via best-of-N sampling. Best-of-4 with Llama3.1-8B lowers steering error (Figure 6, left), outperforming best-of-1 across all prompting strategies and models evaluated (GPT-4.1 vs. Llama3.1-8B: 0.429 → to 0.404, see Appendix). Median orthogonality at best-of-4 also outperforms the top best-of-1 model (GPT-4.1 vs. Llama3.3-70B: 0.718 → 0.673, see Appendix). This improvement with N suggests that responses better-aligned with goals lie within the model's support but are rare in the model's sampling distribution. Best-of-N also scales poorly, lowering median steering error by 0.031 at most when doubling N (Figure 6, left).

Takeaway #2: Inference-time steerability is possible but inefficient. We find that prompt engineering alone may not be powerful enough to surface responses with low steering error. While best-of-N sampling demonstrates the existence of such responses, they remain rare under the base model's output distribution. Our results motivate fine-tuning to increase the likelihood of low steering-error responses.

4.3 RL yields progress towards steerable models

Gains under best-of-N sampling suggest that low steering error responses exist but are rare under an LLM's sampling distribution. We hypothesize RL can shift the output distribution towards such generations. Indeed, RL improves steerability, adopting different rewriting strategies compared to the base model, but does not eliminate side effects.

The post-RL model rivals best-of-128 sampling. In a 2D goal-space (reading difficulty & formality), RL improves steerability in Llama3.1-8B. We report mean and standard deviation to capture improvements in the tails (Table 2). Post-RL steerability rivals best-of-128 sampling in steering error (pre-RL best@128 vs. post-RL mean: 0.210 → 0.119), though orthogonality lags the base model (pre-RL vs. post-RL mean: 0.147 → 0.121). Furthermore, optimizing only miscalibration or orthogonality worsens the other (e.g., RL w/ steering error: 0.294; orthogonality-only: 1.463), which we conjecture may be due to underspecification: flat-reward regions could worsen overfitting (e.g., all formality levels are equal-reward when optimizing miscalibration in reading difficulty only).

RL shifts the model's rewriting strategies. To analyze whether post-RL steerability improvements are meaningful, we examine changes in generation patterns. First, RL mitigates copy-pasting behavior. Before fine-tuning, the base

	Steering error	Miscalibration	Orthogonality
Base model (pre-RL)	0.300 (0.150)	0.986 (0.464)	0.147 (0.328)
Best@128 (pre-RL)	0.210 (0.168)	0.683 (0.539)	0.121 (0.283)
Miscalibration-only reward	0.210 (0.138)	0.542 (0.429)	0.366 (0.395)
Orthogonality-only reward	0.386 (0.248)	1.463 (1.004)	**0.025** (0.134)
Full steering error	**0.119** (0.135)	**0.294** (0.391)	0.160 (0.292)

Table 2: Main results for RL, with an ablation study of the reward model. Mean (std. dev.) of steerability metrics across evaluation probe ($N = 1,024$: 64 held-out source texts; 16 goals each).

Request type	Pre-RL orthogonality	Pre-RL, best@128 orthogonality	Post-RL orthogonality
Corr. requests	0.216 (0.279)	0.238 (0.253)	0.322 (0.283)
Anti-corr. requests	0.330 (0.399)	0.341 (0.320)	0.317 (0.264)
Mean Gap (absolute)	0.114	0.103	0.005

Table 3: Mean (std. dev.) of (from left to right) orthogonality for pre- vs. post-RL model on correlated (top; e.g., increase both dimensions) vs. anti-correlated requests (middle; e.g., change dimensions in opposite directions). RL shrinks the gap in side effects (bottom) between correlated and anti-correlated requests, despite only supervised via 1D instructions.

Model	Sentence BLEU
Pre-RL	0.864 (0.239)
Pre-RL, best@128	0.761 (0.245)
Post-RL	0.529 (0.239)

Table 4: Mean (std. dev.) sentence-level BLEU (original vs. rewrite) by dataset, pre- & post-RL.

model copy-pastes the source text in 135 of 1,024 (13.2%) prompts evaluated, a trivial method to minimize orthogonality. Post-RL, the copy-paste behavior vanishes. BLEU (Papineni et al. 2002) between rewrites and source texts also drops (Table 4; pre-RL vs. post-RL: 0.864 $\rightarrow$ 0.529), suggesting that the post-RL model adopts a less conservative editing strategy to satisfy user goals. Pre- vs. post-RL flow diagrams (see Appendix) support our analysis. Second, RL generalizes to unseen instructions. Despite training with 1D instructions, the post-RL model better handles 2D anti-correlated instructions. We report mean and standard deviation to capture improvements in the tails. The difference in mean orthogonality between correlated vs. anti-correlated requests largely vanishes, dropping from 0.114 pre-RL to 0.005 post-RL (Table 3, right), suggesting improved independence in controlling each goal dimension. We show violin plots summarizing other metrics in the Appendix. Analysis of an anti-correlated rewrite (see Appendix) further illustrates this behavior.

Takeaway #3: RL yields partial progress towards steerability. In a 2D goal-space, we improve the steerability of Llama3.1-8B. These improvements reflect meaningful changes in the model's rewriting strategies, such as reducing copy-paste behavior (lower BLEU score post-RL) and improved disentanglement of goal dimensions (lower orthogonality post-RL in anti-correlated requests). Nonetheless, orthogonality can still be improved, highlighting the need for further work to eliminate side effects.

5 Discussion & Conclusion

We propose a framework for measuring steerability: whether a model can reliably follow diverse, multi-dimensional goals. Existing LLM evaluations directly leverage data from real-world interactions or Internet text, which may not be representative, or use single-dimensional metrics, which do not capture side effects in open-ended generation. Our steerability probe design mitigates these gaps by uniformly sampling goals and measuring multiple dimensions of text. Empirically, LLMs struggle with steerability due to side effects. Inference-time interventions such as prompt engineering and best-of-N sampling offer minor or costly gains. However, RL fine-tuning shows promise as a partial solution. Our work suggests that steerability may be a fundamental challenge for LLM alignment, requiring shifts in model behavior beyond inference-time tweaks. We hope that our framework provides a foundation for measuring LLM alignment with diverse sets of user goals.

Limitations. We focus on steering along verifiable text attributes, leaving goals such as style, to future work. We also only evaluate LLMs in single-turn settings. However, our framework is easily extended to multi-turn settings or generative models beyond text (e.g., multimodal LMs). Our study of interventions is non-exhaustive: we do not vary prompt formatting and apply RL to only an 8B model in 2D goal-space. Larger models may have higher post-RL upside, but optimizing steerability in higher dimensional goal-space could introduce new challenges. Ultimately, our framework is a principled foundation for evaluating LLM steerability, that we hope complements current evaluations of LLM capabilities and improves alignment with diverse human goals.

Acknowledgements

This work was partially done as an intern at Microsoft Research. We thank (in alphabetical order) Donald Lin, Gregory Kondas, Irina Gaynanova, Jennifer Neville, Jung Min Lee, Mahdi Kalayeh, Nathan Kallus, Siddharth Suri, Stephanie Shepard, Wanqiao Xu, Winston Chen, Zhiyi Hu, as well as members of the AI Interaction & Learning Group at Microsoft Research, the Machine Learning & Inference Research team at Netflix, and the NeurIPS 2024 Safe Generative AI Workshop for helpful conversations and feedback on this work. Special thanks to Donna Tjandra, Meera Krishnamoorthy, Michael Ito, Paco Haas, and Sarah Jabbour for their comments on drafts of this work, and to Quentin Gallouédec, the TRL developer community, and the vLLM developer community for their responsiveness on Github issues and engaging in helpful discussions around implementation details.

References

Amodei, D.; Olah, C.; Steinhardt, J.; Christiano, P.; Schulman, J.; and Mané, D. 2016. Concrete problems in AI safety. *arXiv preprint arXiv:1606.06565*.

Bickel, S.; and Scheffer, T. 2009. Discriminative learning under covariate shift. *Journal of Machine Learning Research*, 10: 2137–2155.

BIG-Bench contributors. 2023. Beyond the imitation game: Quantifying and extrapolating the capabilities of language models. *Transactions on Machine Learning Research*.

Chen, K.; Cusumano-Towner, M.; Huval, B.; Petrenko, A.; Hamburger, J.; Koltun, V.; and Krähenbühl, P. 2025. Reinforcement learning for long-horizon interactive LLM agents. *arXiv preprint arXiv:2502.01600*.

DeepSeek-AI Team. 2025. DeepSeek-R1: Incentivizing reasoning capability in LLMs via reinforcement learning. *arXiv preprint arXiv:2501.12948*.

Dong, G.; Lu, K.; Li, C.; Xia, T.; Yu, B.; Zhou, C.; and Zhou, J. 2025. Self-play with execution feedback: Improving instruction-following capabilities of large language models. In *ICLR*.

Durmus, E.; Tamkin, A.; Clark, J.; Wei, J.; Marcus, J.; Batson, J.; Handa, K.; Lovitt, L.; Tong, M.; McCain, M.; Rausch, O.; Huang, S.; Bowman, S.; Ritchie, S.; Henighan, T.; and Ganguli, D. 2024. Evaluating feature steering: A case study in mitigating social biases. https://anthropic.com/research/evaluating-feature-steering.

He, Q.; Zeng, J.; He, Q.; Liang, J.; and Xiao, Y. 2024. From complex to simple: Enhancing multi-constraint complex instruction following ability of large language models. In *EMNLP Findings*, 10864–10882.

Hendrycks, D.; Burns, C.; Basart, S.; Zou, A.; Mazeika, M.; Song, D.; and Steinhardt, J. 2021. Measuring massive multitask language understanding. In *ICLR*.

Heylighen, F.; and Dewaele, J.-M. 1999. Formality of language: Definition, measurement and behavioral determinants. Technical report, Center "Leo Apostel", Vrije Universiteit Brüssel.

Jarvis, S.; and Hashimoto, B. J. 2021. How operationalizations of word types affect measures of lexical diversity. *International Journal of Learner Corpus Research*, 7(1): 163–194.

Kalajdzievski, D. 2023. A rank stabilization scaling factor for fine-tuning with LoRA. *arXiv preprint arXiv:2312.03732*.

Kim, B.; Kim, H.; and Kim, G. 2019. Abstractive summarization of Reddit posts with multi-level memory networks. In *NAACL-HLT*, 2519–2531.

Kincaid, J. P.; Fishburne Jr, R. P.; Rogers, R. L.; and Chissom, B. S. 1975. Derivation of new readability formulas (Automated Readability Index, Fog Count And Flesch Reading Ease Formula) for navy enlisted personnel. Technical report, Naval Technical Training Command Millington TN Research Branch.

Köpf, A.; Kilcher, Y.; von Rütte, D.; Anagnostidis, S.; Tam, Z. R.; Stevens, K.; Barhoum, A.; Nguyen, D.; Stanley, O.; Nagyfi, R.; ES, S.; Suri, S.; Glushkov, D.; Dantuluri, A.; Maguire, A.; Schuhmann, C.; Nguyen, H.; and Mattick, A. 2023. OpenAssistant conversations: Democratizing large language model alignment. In *NeurIPS*, 47669–47681.

Kryściński, W.; Rajani, N.; Agarwal, D.; Xiong, C.; and Radev, D. 2022. BOOKSUM: A collection of datasets for long-form narrative summarization. In *EMNLP Findings*, 6536–6558.

Kwon, W.; Li, Z.; Zhuang, S.; Sheng, Y.; Zheng, L.; Yu, C. H.; Gonzalez, J.; Zhang, H.; and Stoica, I. 2023. Efficient memory management for large language model serving with PagedAttention. In *SOSP*, 611–626.

Li, J.; Peris, C.; Mehrabi, N.; Goyal, P.; Chang, K.-W.; Galstyan, A.; Zemel, R.; and Gupta, R. 2024. The steerability of large language models toward data-driven personas. In *NAACL-HLT*, 7290–7305.

Meta AI. 2024. The Llama 3 herd of models. *arXiv preprint arXiv:2407.21783*.

Miehling, E.; Desmond, M.; Ramamurthy, K. N.; Daly, E. M.; Varshney, K. R.; Farchi, E.; Dognin, P.; Rios, J.; Bouneffouf, D.; Liu, M.; and Sattigeri, P. 2025. Evaluating the prompt steerability of large language models. In *NAACL-HLT*, 7874–7900.

Minh, N. N.; Baker, A.; Neo, C.; Roush, A. G.; Kirsch, A.; and Shwartz-Ziv, R. 2025. Turning up the heat: Min-p sampling for creative and coherent LLM outputs. In *ICLR*.

OpenAI. 2023. GPT-4 technical report. *arXiv preprint arXiv:2303.08774*.

OpenAI. 2024a. GPT-4o system card. *arXiv preprint arXiv:2410.21276*.

OpenAI. 2024b. Learning to reason with LLMs. https://openai.com/index/learning-to-reason-with-llms/.

OpenAI. 2024c. OpenAI o1 system card. *arXiv preprint arXiv:2412.16720*.

OpenAI. 2025. OpenAI o3-mini system card. https://cdn.openai.com/o3-mini-system-card-feb10.pdf.

Ouyang, L.; Wu, J.; Jiang, X.; Almeida, D.; Wainwright, C.; Mishkin, P.; Zhang, C.; Agarwal, S.; Slama, K.; Ray, A.; Schulman, J.; Hilton, J.; Kelton, F.; Miller, L.; Simens, M.; Askell, A.; Welinder, P.; Christiano, P. F.; Leike, J.; and Lowe, R. 2022. Training language models to follow instructions with human feedback. In *NeurIPS*, 27730–27744.

Papineni, K.; Roukos, S.; Ward, T.; and Zhu, W.-J. 2002. BLEU: A method for automatic evaluation of machine translation. In *ACL*, 311–318.

Qin, Y.; Song, K.; Hu, Y.; Yao, W.; Cho, S.; Wang, X.; Wu, X.; Liu, F.; Liu, P.; and Yu, D. 2024. InFoBench: Evaluating instruction following ability in large language models. In *ACL Findings*, 13025–13048.

Qwen Team. 2025. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*.

Rafailov, R.; Sharma, A.; Mitchell, E.; Manning, C. D.; Ermon, S.; and Finn, C. 2023. Direct preference optimization: Your language model is secretly a reward model. In *NeurIPS*, 53728–53741.

Raffel, C.; Shazeer, N.; Roberts, A.; Lee, K.; Narang, S.; Matena, M.; Zhou, Y.; Li, W.; and Liu, P. J. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(140): 1–67.

Rimsky, N.; Gabrieli, N.; Schulz, J.; Tong, M.; Hubinger, E.; and Turner, A. 2024. Steering Llama 2 via contrastive activation addition. In *ACL*, 15504–15522.

Sanchez, G. V.; Spangher, A.; Fan, H.; Levi, E.; and Biderman, S. 2024. Stay on topic with classifier-free guidance. In *ICML*, 43197–43234.

See, A.; Liu, P. J.; and Manning, C. D. 2017. Get to the point: Summarization with pointer-generator networks. In *ACL*, 1073–1083.

Shaham, U.; Segal, E.; Ivgi, M.; Efrat, A.; Yoran, O.; Haviv, A.; Gupta, A.; Xiong, W.; Geva, M.; Berant, J.; and Levy, O. 2022. SCROLLS: Standardized comparison over long language sequences. In *EMNLP*, 12007–12021.

Turner, A. M.; Thiergart, L.; Leech, G.; Udell, D.; Vazquez, J. J.; Mini, U.; and MacDiarmid, M. 2023. Steering language models with activation engineering. *arXiv preprint arXiv:2308.10248*.

Vafa, K.; Bentley, S.; Kleinberg, J.; and Mullainathan, S. 2025. What's producible may not be reachable: Measuring the steerability of generative models. *arXiv preprint arXiv:2503.17482*.

Wei, J.; Wang, X.; Schuurmans, D.; Bosma, M.; ichter, b.; Xia, F.; Chi, E.; Le, Q. V.; and Zhou, D. 2022. Chain-of-thought prompting elicits reasoning in large language models. In *NeurIPS*, 24824–24837.

Wu, Z.; Arora, A.; Geiger, A.; Wang, Z.; Huang, J.; Jurafsky, D.; Manning, C. D.; and Potts, C. 2025. AxBench: Steering LLMs? Even simple baselines outperform sparse autoencoders. In *ICML*.

Zhao, W.; Ren, X.; Hessel, J.; Cardie, C.; Choi, Y.; and Deng, Y. 2024. WildChat: 1M ChatGPT interaction logs in the wild. In *ICLR*.

Zhong, Q.; Wang, K.; Xu, Z.; Liu, J.; Ding, L.; Du, B.; and Tao, D. 2025. Achieving >97% on GSM8k: Deeply understanding the problems makes LLMs perfect reasoners. *Frontiers of Computer Science*, 20(1).

Zhou, J.; Lu, T.; Mishra, S.; Brahma, S.; Basu, S.; Luan, Y.; Zhou, D.; and Hou, L. 2023. Instruction-following evaluation for large language models. *arXiv preprint arXiv:2311.07911*.