# Toxic HallucinAItions: Perturbing Prompts and Tracing LLM Circuits

**Anonymous ACL submission**

## Abstract

Large language models (LLMs) are increasingly deployed in conversational settings where user tone ranges from polite to adversarial or toxic, yet less is known about whether toxic language in otherwise semantically equivalent prompts can degrade factual reliability. We study how lexical and tone-based prompt perturbations affect the factual reliability of LLMs. Using controlled prompt variations across polite, random, and three toxicity levels, we evaluate five LLMs on `ARC-Easy`, `GSM8K`, and `MMLU`. We find that toxic lexical perturbations consistently reduce factual accuracy and increase uncertainty, while polite phrasing yields limited and inconsistent changes. To examine whether these answer inconsistencies correspond to internal changes, we conduct attribution-graph analyses of model activations and influences. We find that increasing toxicity selectively amplifies perturbation-sensitive variant nodes while relatively stable core reasoning nodes remain more invariant. These findings position prompt tone as a critical dimension of LLM reliability and provide behavioral and mechanistic evidence that surface-level lexical variation can alter factual outputs and internal computation.

## 1 Introduction

> "Saying "please" or "thank you" to AI chatbots can apparently cost tens of millions of dollars. But some fear the cost of not being polite could be higher."—New York Times (Deb, 2025)

As AI systems become increasingly embedded in everyday workflows, users interact with LLMs in various conversational settings, ranging from polite and carefully structured prompts to emotionally charged, adversarial, or toxic language (Gehman et al., 2020; Wei et al., 2023). Recent public discourse has drawn attention to the possible practical implications of users' conversational variations, including computational cost, with OpenAI CEO Sam Altman claiming that polite interactions such as please" and thank you" cost the company millions of dollars in compute (Futurism, 2025). At the same time, growing evidence suggests that seemingly minor prompt variations can substantially alter model behavior and downstream performance (Zhao et al., 2021; Lu et al., 2022; Perez et al., 2023; Dobariya and Kumar, 2025; Mizrahi et al., 2024; Sclar et al., 2024; Yin et al., 2024). This raises important questions regarding the robustness and reliability of factual reasoning in conversational AI based on prompt tone.

A rich body of work has studied hallucinations in LLMs, typically referring to fluent but factually unsupported or fabricated generations (Ji et al., 2023; Maynez et al., 2020a,b). However, recent research argues that factual failures in structured tasks such as question answering and multiple-choice reasoning may arise from several distinct mechanisms, including reasoning instability, prompt sensitivity, uncertainty, and failures in knowledge elicitation (Jang et al., 2026). In particular, semantically equivalent prompts can often produce inconsistent factual answers despite preserving the underlying query intent (Dobariya and Kumar, 2025; Cai et al., 2025; Elazar et al., 2021). This distinction matters because prompt-induced factual inconsistency represents a broader reliability challenge that extends beyond conventional notions of hallucination.

Prior work has focused on stylistic prompting strategies such as calibration, role prompting, formatting, and chain-of-thought reasoning (Zhao et al., 2021; Lu et al., 2022; Perez et al., 2023). However, the role of toxic language in influencing factual reliability remains underexplored. Toxic language generally refers to hostile, abusive, insulting, threatening, or otherwise aggressive forms of communication that can provoke harmful, adversarial, or emotionally charged interactions (Gehman et al., 2020; Davidson et al., 2017). This gap is especially important because real-world interactions with AI are driven by how users naturally choose to prompt, and these interactions often include toxic, adversarial, or emotionally aggressive language (Barhoum;

Gehman et al., 2020). Therefore, understanding how such language affects the LLMs outputs is critical for evaluating the robustness and safety of deployed LLM systems.

Therefore, to empirically examine this phenomenon, we study how lexical perturbations inserted into otherwise semantically equivalent prompts alter the ability of the LLMs in factual answering and reasoning behavior. Rather than treating all incorrect outputs as conventional hallucinations, we focus more precisely on how perturbing prompts—ranging across polite and toxic keywords—could impact factual reliability and answer instability. Further, recent advances in mechanistic interpretability have emphasized understanding LLMs through their internal representations, often conceptualized as *attribution graphs* or *circuits* responsible for specific behaviors (Elhage et al., 2021; Olah et al., 2020a; Geva et al., 2021). Building on this line of work, we investigate how lexical perturbation-induced factual instability corresponds to identifiable shifts in internal computation and attribution graph behavior. Our work is guided by the following research questions (RQs):

***RQ1:*** How do lexical and tone-based prompt perturbations impact the factual reliability of LLMs?

***RQ2:*** Can such factual degradation be explained through internal representations within LLMs?

Our work examines five models: GPT-5-Nano, Gemini-2.5-Flash, Gemma-2-2B, Qwen2.5-1.5B-Instruct, and LLaMA-3.2-1B under prompt perturbations ranging across random, polite, and toxic prompts. We evaluate these models and prompt perturbations across four widely used benchmark: `ARC-Easy` (Clark et al., 2018), `GSM8K` (Cobbe et al., 2021), and `MMLU` (Hendrycks et al., 2020).

For RQ1, we conduct lexical perturbations and compare the factual accuracy of the models across the benchmarks. Regression modelings explain how various aspects of the prompt associate with the models' accuracy, entropy, and perplexity. We find that toxic lexical perturbations consistently degrade factual accuracy across benchmarks and model families, while also increasing predictive uncertainty measured through entropy and perplexity. Random lexical perturbations similarly reduce performance, indicating that even non-semantic prompt variations can destabilize reasoning behavior. Smaller open-source models exhibit substantially larger degradation under toxic prompts compared to larger proprietary systems.

For RQ2, we trace the LLM circuits, specifically, the attribution graphs of activations and influences across layers. We find that toxic perturbations progressively amplify perturbation-sensitive pathways, increasing their activation and influence while comparatively stable core reasoning nodes remain largely invariant. These internal shifts closely align with the observed degradation in factual accuracy and increased uncertainty under toxic prompts, suggesting that lexical toxicity redirects computation away from stable semantic reasoning circuits toward context-sensitive representations.

Taken together, this work makes four contributions: 1) a computational framework for studying lexical and tone-based prompt perturbations in factual reasoning through controlled rewrites and attribution-graph evaluation; 2) a cross-model empirical analysis showing that toxic lexical perturbations can degrade factual reliability and answer consistency; 3) mechanistic insights into toxicity-sensitive attribution subgraphs associated with factual instability; and 4) the release of dataset, prompt perturbation framework, and codebase to support future research on tone-sensitive robustness and mechanistic analysis in LLMs.

## 2 Related Work

***Factuality and Hallucination in LLM Responses.*** Large language models (LLMs) are known to generate fluent but factually incorrect outputs, commonly referred to as hallucinations (Ji et al., 2023; Maynez et al., 2020a,b). Prior work has characterized hallucinations across tasks such as summarization, question answering, and dialogue, attributing them to factors including spurious correlations in training data, exposure bias, and lack of grounding (Maynez et al., 2020a; Shuster et al., 2021). Several approaches have been proposed to detect or mitigate hallucinations, including self-consistency methods (Wang et al., 2022), sampling-based detection (Manakul et al., 2023), and retrieval-augmented generation (Lewis et al., 2020). More recent work emphasizes evaluation frameworks that distinguish between factual accuracy, faithfulness, and calibration (Maynez et al., 2020b; Kadavath et al., 2022). However, most existing studies assume fixed prompt conditions and do not examine how *linguistic perturbations*, particularly those unrelated to task semantics, influence hallucination behavior. Our work extends this line by showing that toxic prompt variations systematically degrade factual accuracy across models.

***Prompt Sensitivity and Adversarial Inputs.*** A

growing body of work reveals that LLMs are highly sensitive to prompt phrasing, even when semantic content is preserved (Zhao et al., 2021; Lu et al., 2022; Perez et al., 2023). Prompt calibration methods show that minor formatting or ordering changes can significantly alter model predictions (Zhao et al., 2021; Lu et al., 2022). Similarly, studies on prompt multiplicity reveal that models can produce inconsistent outputs for equivalent queries (Perez et al., 2023). Beyond benign variations, adversarial prompting such as universal triggers (Wallace et al., 2019) and jailbreak attacks (Xie et al., 2023; Wei et al., 2023) reveal that carefully constructed inputs can induce harmful or incorrect outputs. Recent work also shows that prompt sensitivity decreases with model scale but does not disappear entirely (Zhuo et al., 2024). Our work isolates toxicity as a controlled perturbation and examines its effect on LLM responses.

***Toxicity, Bias, and Safety in LLMs.*** LLMs inherit biases and toxic patterns from their training data, leading to concerns about fairness, safety, and harmful content generation (Bender et al., 2021; Weidinger et al., 2022; Goel et al., 2026; Kim et al., 2026). Extensive work has focused on detecting and mitigating toxicity using alignment techniques such as reinforcement learning from human feedback (RLHF) (Ouyang et al., 2022), decoding-time interventions (Liu et al., 2021; Zhang and Wan, 2023), and interpretability-based activation steering and model editing (Uppaal et al., 2025; Goyal et al., 2025a). Safety research also examines how models respond to harmful or adversarial inputs, showing that toxicity can interact with alignment mechanisms in complex ways (Zhou et al., 2024; Xie et al., 2023). For example, alignment can operate through intermediate representations that detect harmful intent in early layers and refine responses in later layers (Zhou et al., 2024). This body of work primarily treats toxicity as an *output* concern (i.e., preventing harmful outputs). Parallelly, prior work has explored LLMs in identifying and flagging toxic content from a content moderation perspective (Kolla et al., 2024; Kumar et al., 2024; Goyal et al., 2025b; Zhan et al., 2025). Our work complements this research by examining the effect of toxic tokens as *input* to models.

***Mechanistic Interpretability and Circuits in Transformers.*** Recent advances in mechanistic interpretability aim to explain LLM behavior by identifying internal computational structures, often referred to as circuits (Olah et al., 2020a; Elhage et al., 2021). These approaches analyze how specific neurons, attention heads, or subspaces contribute to model outputs. For example, prior work shows that feed-forward layers act as key-value memories storing factual associations (Geva et al., 2021), while attention heads can implement structured reasoning patterns. Probing methods have been widely used to extract interpretable signals from hidden representations, demonstrating that intermediate activations encode linguistic, factual, and safety-related information (Belinkov et al., 2017). More recent work applies causal interventions and attribution techniques to identify neurons responsible for specific behaviors, such as factual recall or bias (Meng et al., 2022), and toxicity (Wang et al., 2026). Our approach builds on this paradigm by combining probing (logistic regression) and feature attribution (random forests) to identify *sub-circuits* associated with toxicity-induced factual reliability.

## 3 Data

We construct a controlled multi-domain benchmark spanning factual recall, commonsense reasoning, and mathematical problem solving, using three widely adopted NLP benchmarks: 1) `ARC-Easy` (elementary scientific and factual reasoning) (Clark et al., 2018), 2) `GSM8K` (multi-step arithmetic reasoning) (Cobbe et al., 2021), and 3) `MMLU` (language understanding and commonsense reasoning across academic subjects) (Hendrycks et al., 2020). These capture complementary reasoning capabilities and to evaluate whether perturbation-induced factual reliability occurs across heterogeneous task settings.

***Building the evaluation dataset.*** We construct filtered subsets from each benchmark to support standardized short-form evaluation across models. For `MMLU` and `ARC-Easy`, we retain only open-ended question–answer pairs whose gold answers consist of a single token, while excluding examples containing multi-token answers. For `GSM8K`, we only retain examples with valid integer answers. These filtering steps reduce variability introduced by free-form multi-token generation and enable consistent answer normalization across models Following filtering, we sample 1,500 question–answer pairs from each dataset to maintain comparable evaluation sizes across datasets.

***Constructing lexical perturbation bins.*** To analyze the effect of toxicity, we construct prompt perturbations using curated lexical bins grouped by toxicity intensity. Each perturbation consists

| Bin | Toxicity Range | Example perturbation words |
|---|---|---|
| *Baseline* | 0.0–0.05 | – |
| *Polite* | 0.0–0.05 | please, kindly, thanks, appreciate |
| *Random* | 0.0–0.25 | banana, quantum, purple, giraffe |
| *Low Toxic* | 0.25–0.50 | annoying, silly, weird, lame |
| *Medium Toxic* | 0.50–0.80 | idiot, moron, worthless, awful |
| *High Toxic* | 0.80+ | bitch, fucker, shithead, asshole |

Table 1: Lexical perturbation bins with Perspective API toxicity score ranges (0–1).

of a single appended token added to the end of the original question, preserving its semantic intent and ground-truth answer. Toxicity scores are obtained using the Perspective API (Jigsaw and Google, 2017), which provides continuous toxicity estimates. Based on these scores, we partition perturbation tokens into five bins corresponding to increasing toxicity levels: 1) **polite** (consisting of courteous or cooperative words), 2) **random** (tone-wise neutral words), 3) **low toxic** (mildly negative or dismissive words), 4) **medium toxic** (direct insults or derogatory words), and 5) **high toxic** (strongly profane or aggressive words) (see Table 1). For each of the bins, we identify 100 tokens.

***Generating perturbed prompts.*** The baseline condition contains no appended token, while perturbed conditions use tokens from the polite, random, low-toxic, medium-toxic, and high-toxic bins. That is, for each question from our datasets, we construct multiple prompt variants by appending a single perturbation token from a bin to the original query. Our rationale for using a single-token perturbation is that it introduces a minimal and conservative change to the prompt. More complex perturbations would make comparisons harder to interpret, as any observed output change would be more difficult to attribute to a specific prompt change. Our perturbations preserve the semantic meaning of the question while altering only its tone or toxicity, enabling controlled analysis of how non-semantic perturbations affect factual reasoning and answer consistency in LLMs.

# 4 RQ1: Prompt Perturbation-Induced Answer Inconsistencies of LLMs

## 4.1 RQ1: Methodology

For our study, we examine five model families, ranging across a variety of architectures, training datasets, and number of parameters: GPT-5-Nano, Gemini-2.5-Flash, Gemma-2-2B, Qwen2.5-1.5B-Instruct, and LLaMA-3.2-1B.

***Answer Generation.*** To ensure standardized evaluation across datasets and models, all models are constrained to generate only one token as output. Generation is performed greedily with deterministic decoding ($T$=0), minimizing stochastic variation and enabling comparison across perturbations. Given a model $f_\theta$ and prompt $p_{i,j}$, the generated prediction is defined as: $\hat{\mathtt{y}}_{\mathtt{i,j}} = \mathtt{f}_\theta(\mathtt{p}_{\mathtt{i,j}})$, where $\hat{y}_{i,j}$ denotes the model output for question $q_i$ under perturbation token $w_j$. Generation is performed exhaustively across all perturbed prompts, producing a complete set of outputs spanning all the bins.

***Comparing Factual Reliabilities*** We compute accuracy for each model under every perturbation and compare it against the baseline setting. Accuracy is defined as the proportion of correctly answered questions.For each model, we then obtain effect size (Cohen's $d$) and paired $t$-tests between the baseline and each perturbed condition, allowing us to measure whether the perturbation led to a significant change in performance.

***Regression Modeling.*** To move beyond just accuracy comparisons and better characterize the factors associated with answer consistencies, we conduct regression modeling/ In particular, we build separate linear mixed-effects regression models for three dependent variables: 1) *accuracy*, capturing task correctness; 2) *entropy*, reflecting uncertainty in the model's output distribution; and 3) *perplexity*, measuring the confidence of the generated response (Jelinek et al., 1977; Shannon, 1948) The entropy and perplexity could only be obtained for the open-source models (except GPT, Gemini). For independent variables, we use model type, Perspective API-based toxicity score of the perturbed word, length of question, and rarity of answer (computed using a TF-IDF-based rarity score over the answer vocabulary, to account for distributionally infrequent or uncommon target answers). Model fit is assessed using marginal and conditional $R^2$.

## 4.2 RQ1: Results

***Impact of perturbations on model accuracy.*** Table 2 summarizes model accuracy across perturbation bins relative to the baseline. Across nearly all evaluated models, baseline prompts achieve the highest accuracy, suggesting that even small lexical perturbations can disrupt model behavior.

We first observe that polite perturbations do not improve accuracy across most models. In fact, accuracy drops in GPT-5-nano by -2.14%, Gemini by -3.08%, Gemma-2-2B by -16.60%, and Qwen2.5-1.5B by -6.20% relative to the baseline condition. Interestingly, LLaMA-3.2-1B shows the only exception, exhibiting an improvement of 11.61% un-

| | GPT-5-nano | | | | Gemini-2.5-Flash | | | | Gemma-2-2B | | | | Qwen2.5-1.5B | | | | LLaMA-3.2-1B | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Bin** | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ |
| Baseline | 0.655 | 0.00 | — | — | 0.455 | 0.00 | — | — | 0.259 | 0.00 | — | — | 0.274 | 0.00 | — | — | 0.155 | 0.00 | — | — |
| Polite | 0.641 | -2.14 | -0.07 | -2.11* | 0.441 | -3.08 | -0.08 | -2.40** | 0.216 | -16.60 | -0.15 | -4.41*** | 0.257 | -6.20 | -0.07 | -1.98* | 0.173 | +11.61 | 0.10 | 2.87* |
| Random | 0.607 | -7.33 | -0.20 | -5.99*** | 0.399 | -12.31 | -0.16 | -4.75*** | 0.186 | -28.19 | -0.23 | -6.78*** | 0.214 | -21.90 | -0.18 | -5.50*** | 0.146 | -5.81 | -0.11 | -2.98* |
| Low Toxic | 0.636 | -2.90 | -0.09 | -2.79** | 0.452 | -0.66 | -0.08 | -2.32* | 0.212 | -18.15 | -0.15 | -4.59*** | 0.262 | -4.38 | -0.07 | -1.96* | 0.173 | +11.61 | 0.09 | 2.82* |
| Med. Toxic | 0.632 | -3.51 | -0.12 | -3.68*** | 0.436 | -4.18 | -0.06 | -2.85* | 0.193 | -25.48 | -0.21 | -6.25*** | 0.241 | -12.04 | -0.11 | -3.36*** | 0.160 | +3.23 | 0.10 | 2.53* |
| High Toxic | 0.617 | -5.80 | -0.13 | -3.82*** | 0.405 | -10.99 | -0.15 | -4.60*** | 0.189 | -27.03 | -0.22 | -6.46*** | 0.238 | -13.14 | -0.11 | -3.42*** | 0.158 | +1.94 | 0.09 | 2.30* |

Table 2: Model-wise accuracy (Acc.) under lexical perturbation conditions relative to the baseline, with Δ% denotes percentage change in accuracy relative to baseline (teal: positive; pink: negative; shading indicates magnitude), Cohen's $d$, and $t$-tests (* p<0.05, ** p<0.01, *** p<0.001). Dataset-wise evaluations in Tables A1, A2, and A3.

| | ARC-Easy | | | GSM8K | | | MMLU | | |
|---|---|---|---|---|---|---|---|---|---|
| **Independent Variable** | **Accuracy** | **Entropy** | **Perplexity** | **Accuracy** | **Entropy** | **Perplexity** | **Accuracy** | **Entropy** | **Perplexity** |
| Model:GPT-5-Nano | 0.18* | — | — | 0.02 | — | — | 0.18* | — | — |
| Model:Gemini-2.5-Flash | 0.14* | — | — | 0.35** | — | — | 0.14* | — | — |
| Model:Gemma-2-2B | -0.12* | 0.78*** | 0.58*** | -0.09 | 0.96*** | 1.17*** | -0.12* | 0.80*** | -0.69*** |
| Model:Llama-3.2-1B | -0.17*** | 0.85*** | 0.54*** | -0.17*** | 1.81*** | 1.17*** | -0.17*** | 0.87*** | -0.64*** |
| Toxicity Score | -0.01*** | 0.19*** | 0.02** | -0.01*** | 0.16*** | 0.06* | -0.01*** | 0.19*** | 0.02** |
| Question Length | -0.05*** | 0.17*** | 0.10* | -0.04*** | 0.12*** | 0.44*** | -0.05*** | 0.18*** | 0.12* |
| Answer Rarity | -0.49*** | 0.06*** | 0.05* | -0.45*** | 0.07*** | 0.42** | -0.41*** | 0.06*** | 0.06* |
| $R^2$ (marginal) | 0.28* | 0.40** | 0.12 | 0.12 | 0.72*** | 0.10 | 0.19* | 0.40** | 0.12 |
| $R^2$ (conditional) | 0.44** | 0.51** | 0.21* | 0.26* | 0.76*** | 0.18* | 0.35** | 0.51** | 0.21* |

Table 3: Dataset-wise mixed-effects regression coefficients for accuracy, entropy, and perplexity across ARC-Easy, GSM8K, and MMLU. Coefficients are reported with significance (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$). Missing entries indicate metrics unavailable for closed-source models.

der polite perturbations. These findings suggest that courteous lexical modifiers alone do not consistently improve factual reliability.

We next observe that random perturbations reduce accuracy across all evaluated models: it changes in GPT-5-nano by -7.33%, Gemini by -12.31%, Gemma-2-2B by -28.19%, Qwen2.5-1.5B by -21.90%, and LLaMA-3.2-1B by -5.81% relative to the baseline. Interestingly, random perturbations often produce degradation comparable to, and occasionally larger than, low-toxic perturbations. For example, GPT-5-nano decreases more under random perturbations (-7.33%) than under low-toxic prompts (-2.90%), with a similar pattern observed for Gemini (-12.31% vs. -0.66%). This suggests that prompt disruption can affect responses even without explicit toxic content.

Toxic perturbations further amplify this degradation. GPT-5-nano exhibits a progressive decline across low-toxic (-2.90%), medium-toxic (-3.51%), and high-toxic (-5.80%) conditions, with similar trends observed for Gemini (-0.66%, -4.18%, and -10.99%), Gemma-2-2B (-18.15%, -25.48%, and -27.03%), and Qwen2.5-1.5B (-4.38%, -12.04%, and -13.14%). These results suggest that toxicity compounds an already existing sensitivity to lexical perturbation, leading to increasingly unstable model predictions. Finally, we observe substantial differences across models. Smaller open-source models show larger proportional degradation under perturbations compared to larger proprietary systems. For example, Gemma-2-2B's accuracy drops by -27.03% under high-toxic perturbations, whereas GPT-5-nano's accuracy drops by -5.80%. This pattern is plausibly consistent with prior observations that larger models tend to exhibit improved robustness under adversarial or distribution-shifted prompting conditions (Ganguli et al., 2022).

***Regression analysis: Accuracy, Entropy, and Perplexity.*** Table 3 presents mixed-effects regression results examining the relationship between toxicity, prompt features, and model behavior across ARC-Easy, GSM8K, and MMLU. Across all datasets, toxicity score is negatively associated with accuracy and positively associated with entropy and perplexity. As toxicity increases, accuracy consistently decreases while uncertainty-related behaviors increase, indicating that toxic perturbations make model predictions less stable.

Question length similarly shows a consistent relationship with degraded performance. Longer questions are associated with lower accuracy and higher entropy/perplexity values, suggesting that increasing input complexity amplifies predictive uncertainty (Dziri et al., 2023).

Among all evaluated features, answer rarity exhibits the strongest negative association with accuracy and positively correlates with entropy and perplexity, suggesting that rare outputs are associated with greater uncertainty during generation.

The regression analysis also reveals differences across model families. Gemma-2-2B and LLaMA-3.2-1B show negative accuracy coefficients alongside strong positive entropy and perplexity coefficients across datasets. These trends align with the behavioral findings in Table 2, where smaller open-

source models degrade more under perturbations than larger proprietary systems.

## 5 RQ2: Prompt Perturbation-Induced Mechanistic Changes in LLMs

### 5.1 RQ2: Methodology

***Attribute Graph Generation.*** To gain mechanistic insight into how prompt perturbations influence model outcomes, we construct attribution graphs that trace the internal flow of information from input tokens to output logits. Our approach builds on recent advances in circuit-level interpretability, particularly attribution graphs and transcoders that approximate model internals in a more interpretable feature space (Ameisen et al., 2025; Olah et al., 2020b). These methods aim to identify structured computational pathways ("circuits") responsible for specific model behaviors, enabling analysis beyond input–output correlations.

For each of our open-sourced models, we generate attribution graphs using a replacement-model framework integrating pretrained language models with learned transcoders (Ameisen et al., 2025). Specifically, we use the *gemma-scope-transcoders* for Gemma-2-2B (Mateusz, 2025a), *transcoder-Llama-3.2-1B* for LLaMA-3.2-1B (Mateusz, 2025b), and *qwen3-1.7b-transcoders-lowl0* for Qwen3-1.7B (Hanna, 2025). These transcoders project high-dimensional activations into sparse and interpretable feature representations, enabling attribution at the level of latent computational features rather than individual neurons.

Given a prompt $\mathtt{p_{i,j}}$ constructed using perturbation token $\mathtt{w_j}$, we first obtain the model prediction deterministically and then compute an attribution graph representing how internal features contribute to the final output logits. Formally, for a model $f_\theta$ and prompt $\mathtt{p_{i,j}}$: $\hat{\mathtt{y}}_{\mathtt{i,j}} = \mathtt{f}_\theta(\mathtt{p_{i,j}})$, where $\hat{y}_{i,j}$ denotes the generated prediction. The resulting attribution graph is represented as: $\mathtt{G_{i,j}} = (\mathtt{V_{i,j}}, \mathtt{E_{i,j}})$, where $V_{i,j}$ denotes internal feature nodes and $E_{i,j}$ represents attributed influence between nodes.

For each feature node $v \in V_{i,j}$, we focus on two complementary quantities: **activation** and **influence**. Let $a_v$ denote the activation magnitude of node $v$, and let $I_v$ denote its attributed influence on the final prediction logit. The overall attribution graph can therefore be represented as: $\mathtt{G_{i,j}} = \{(\mathtt{v}, \mathtt{a_v}, \mathtt{I_v}) \mid \mathtt{v} \in \mathtt{V_{i,j}}\}$, where $a_v$ captures how strongly a feature is activated under a given prompt, while $I_v$ captures how much that feature contributes to the final prediction. Influence values are computed through attribution over output logits, enabling us to quantify which internal features most strongly affect model decisions (Olah et al., 2020b; Sundararajan et al., 2017).

***Identifying Core and Variant Nodes.*** We analyze circuits to identify which internal nodes correspond to stable reasoning behavior (*core* nodes) and which exhibit stronger sensitivity to prompt perturbations (*variant* nodes). We measure the variance of node activations across prompts. Intuitively, nodes with low variance are consistently involved in answering the underlying semantic question, irrespective of prompt perturbation, whereas nodes with high variance correspond to perturbation-sensitive features whose behavior changes substantially depending on the perturbation.

For a given question $q_i$, we generate attribution graphs across all perturbations. Then, for each feature node $v$ in the attribution graphs, we collect its activation values across the perturbation prompts of the same question. Let $a_v^{(k)}$ denote the activation of node $v$ under perturbation condition $k \in \{1, \ldots, K\}$. To reduce sensitivity to the raw activation scale, we normalize activations within each graph before computing variability. Using these normalized activations, perturbation variance for each node is defined as: $\mathrm{Var}(\mathtt{v}) = \mathrm{Var}(\tilde{\mathtt{a}}_{\mathtt{v}}^{(1)}, \tilde{\mathtt{a}}_{\mathtt{v}}^{(2)}, \ldots, \tilde{\mathtt{a}}_{\mathtt{v}}^{(\mathtt{K})})$, where $\tilde{a}_v^{(k)}$ denotes the normalized activation of node $v$ under perturbation condition $k$. We then rank all nodes according to their perturbation variance. Nodes in the bottom $25^{\text{th}}$ percentile of the variance distribution are categorized as *core* nodes, while nodes in the top $25^{\text{th}}$ percentile are categorized as *variant* nodes. This formulation is motivated by prior work in representation analysis and circuit interpretability, where stable features across input variations are often associated with invariant reasoning behavior, while highly variable features are linked to context-sensitive or stimulus-dependent processing (Olah et al., 2020b; Ameisen et al., 2025).

***Regression Modeling of Activation and Influence.*** To further characterize how perturbations affect internal computation, we conduct linear mixed-effects regression. For each attribution graph, we define two dependent variables: 1) *activation*, representing the average activation magnitude of a node, and 2) *influence*, representing the attributed contribution of that node to the final prediction.

For independent variables, we use the node type (core or variant) and the toxicity score of the perturbation word. We also include an interaction term

| | Core Nodes | | | | | | | | Variant Nodes | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Prompt** | **Act.** | $\Delta\%$ | $d$ | $t$ | **Infl.** | $\Delta\%$ | $d$ | $t$ | **Act.** | $\Delta\%$ | $d$ | $t$ | **Infl.** | $\Delta\%$ | $d$ | $t$ |
| Baseline | 0.280 | 0.00 | — | — | 3.404 | 0.00 | — | — | 0.051 | 0.00 | — | — | 0.856 | 0.00 | — | — |
| Polite | 0.168 | -40.0 | -0.039 | -1.68 | 2.483 | -27.1 | -0.035 | -1.42 | 0.106 | +107.8 | 0.069 | 2.85** | 0.999 | +16.7 | 0.026 | 2.08* |
| Random | 0.190 | -32.1 | -0.032 | -1.29 | 2.519 | -26.0 | -0.028 | -1.18 | 0.130 | +154.9 | 0.092 | 3.58*** | 1.324 | +54.7 | 0.056 | 3.02** |
| Low Toxic | 0.191 | -31.8 | -0.034 | -1.36 | 2.689 | -21.0 | -0.027 | -1.11 | 0.109 | +113.7 | 0.061 | 2.49* | 1.037 | +21.1 | 0.027 | 2.01* |
| Med. Toxic | 0.162 | -42.1 | -0.052 | -2.27* | 2.062 | -39.4 | -0.051 | -2.11* | 0.125 | +145.1 | 0.107 | 4.08*** | 1.088 | +27.1 | 0.041 | 2.84** |
| High Toxic | 0.172 | -38.6 | -0.047 | -2.01* | 2.580 | -24.2 | -0.033 | -1.96* | 0.134 | +162.8 | 0.104 | 4.55*** | 1.458 | +70.3 | 0.071 | 3.31*** |

Table 4: Node-level Activation (Act.) and Influence (Infl.) changes on `ARC-Easy`. $\Delta\%$ denotes percentage change relative to the baseline condition (teal: positive; pink: negative; shading indicates magnitude), along with Cohen's $d$ and $t$-tests (* $p$<0.05, ** $p$<0.01, *** $p$<0.001). Similar analysis for `GSM8K` (Table A4) and `MMLU` (Table A5).

| | `ARC-Easy` | | `GSM8K` | | `MMLU` | |
|---|---|---|---|---|---|---|
| **Independent V.** | **Act.** | **Infl.** | **Act.** | **Infl.** | **Act.** | **Infl.** |
| Node Type: Variant | -0.604*** | -0.714*** | -0.137*** | -0.381*** | -0.112*** | -0.351*** |
| Toxicity Score | 0.200* | 0.158** | 0.008 | -0.036* | -0.091* | 0.173** |
| Tox. Score X Variant | 0.367*** | 0.310*** | 0.011*** | 0.158*** | 0.030*** | 0.042*** |
| Entropy | 0.141* | 0.029 | -0.061* | -0.110** | 0.359 | -0.448* |
| Perplexity | -0.050* | -0.133* | 0.030** | 0.074** | -0.156* | 0.544* |
| $R^2$ (marginal) | 0.436*** | 0.580*** | 0.028 | 0.169* | 0.061 | 0.173* |
| $R^2$ (conditional) | 0.592*** | 0.734*** | 0.143* | 0.312** | 0.194* | 0.341** |

Table 5: Mixed-effects regression coefficients for node activation (Act.) and influence (Infl.) (* $p$<0.05, ** $p$<0.01, *** $p$<0.001).

to capture whether increasing toxicity disproportionately affects variant nodes relative to core reasoning nodes. Furthermore, we include the model's entropy and perplexity to capture generation uncertainty and predictive confidence.

### 5.2 RQ2: Results

***Changes in Activation & Influence.*** Table A1 summarizes activation and influence changes across core and variant nodes relative to the baseline condition. Across nearly all perturbation settings, core nodes show comparatively smaller and often statistically insignificant changes in both activation and influence, suggesting that stable reasoning pathways are not affected by lexical perturbations.

We first observe that polite perturbations increase activation (by 107.84%) and influence (by 16.71%) among variant nodes, relative to the baseline condition. Also, random perturbations substantially increase variant-node activation (by 154.90%) and influence (by 54.67%) relative to the baseline. Interestingly, random perturbations often produce increases comparable to, and occasionally larger than, toxic perturbations. This mirrors the behavioral findings from RQ1, where random lexical insertions degraded factual accuracy.

Toxic perturbations further amplify these shifts. Variant-node activation progressively increases across low-toxic (113.73%), medium-toxic (145.10%), and high-toxic (162.75%) conditions, with influence similarly increasing by 21.14%, 27.10%, and 70.33%, respectively. These findings suggest that increasingly toxic prompts progressively recruit perturbation-sensitive computational pathways and make them more influential in the final prediction process.

Finally, we observe that increases in variant-node activation and influence occur while core-node behavior remains comparatively stable or decreases relative to the baseline. One possible explanation is that toxic perturbations redirect computation toward perturbation-sensitive features, effectively reducing the contribution of stable reasoning pathways responsible for solving the underlying task. This redistribution of computation may help explain the accuracy degradation observed in RQ1.

***Regression analysis of activation and influence.*** Table 5 shows the mixed effects regression coefficients. Across datasets, the *variant node type* coefficient is negative for both activation and influence, whereas the interaction term is consistently positive and significant. In `ARC-Easy`, the interaction coefficient is 0.367 for activation and 0.310 for influence, with similarly positive interactions observed in `GSM8K` and `MMLU`. These interaction effects are substantially stronger and more consistent than the direct toxicity score coefficients, whose effects remain comparatively mixed across datasets. Together, these results suggest that increasing toxicity does not uniformly increase activation, but instead selectively amplifies perturbation-sensitive pathways. As toxicity increases, variant nodes become progressively more active and influential in the model's computation, consistent with the node-level trends observed earlier (Table 4). Finally, entropy and perplexity also exhibit significant relationships with activation and influence, indicating that uncertainty-related properties are associated with changes in internal circuits.

## 6 Discussion and Conclusion

Our results show that even minimal lexical perturbations can systematically alter both the outputs and internal computation of large language models. Increasing perturbation toxicity consistently reduces accuracy while increasing uncertainty-related behaviors such as entropy and perplexity.

Mechanistically, these effects correspond to increasing activation and influence of perturbation-sensitive circuits under toxic prompts. We next discuss the implications of these findings.

***Toxicity and Distributional Sensitivity.*** A central finding of this work is that toxic perturbations act not purely through semantic harm, but through broader disruption of internal circuit mechanisms, highlighting LLM sensitivity to auxiliary lexical perturbations. These observations align with prior work showing that language models can rely on shallow statistical correlations and unstable heuristics rather than fully robust semantic reasoning (Niven and Kao, 2019; Kaushik et al., 2019). Prior studies on prompt sensitivity similarly report that seemingly minor prompt changes can substantially alter downstream reasonings (Lu et al., 2022; Reynolds and McDonell, 2021).Our findings extend this literature by showing that increasing toxicity progressively amplifies these instabilities at both behavioral and mechanistic levels.

***Comparing Core and Perturbation-Sensitive Circuits.*** Our mechanistic analyses reveal a consistent separation between core and perturbation-sensitive circuits. Core nodes remain comparatively invariant across perturbation settings, whereas variant nodes become progressively more active and influential as toxicity increases. One possible interpretation is that toxic perturbations redirect computation toward perturbation-sensitive pathways, reducing the contribution of stable reasoning circuits responsible for solving the underlying task. Similar dynamics have been observed in studies of spurious correlations and shortcut learning, where models over-rely on non-robust features under distribution shifts (Sagawa et al., 2020; Tu et al., 2020).

***Uncertainty and Internal Representation Shift.*** The regression analyses show that increasing toxicity consistently correlates with higher entropy and perplexity, indicating increased predictive uncertainty. These uncertainty-related properties are also associated with changes in activation and influence within perturbation-sensitive nodes, suggesting that hallucination-like behavior under toxic prompts may emerge from unstable intermediate representations and shifts in internal circuits. This supports prior work that uncertainty and hallucination in language models are tied to instability in learned representations and calibration failures (Desai and Durrett, 2020; Kadavath et al., 2022). Additionally, answer rarity strongly predicts accuracy degradation across datasets, suggesting that distributionally sparse answers remain particularly vulnerable under perturbation conditions. This aligns with prior work showing that long-tail knowledge and infrequent targets remain challenging for language models even when overall benchmark performance are strong (Mallen et al., 2023; Zhang et al., 2024).

***Prompt Robustness Beyond Adversarial Attacks.*** Our analyses suggest that prompt robustness should not be viewed solely through the lens of jailbreak attacks or explicitly malicious prompts. Instead, large language models appear broadly sensitive to auxiliary lexical context that shifts computation away from stable reasoning pathways. Prior work on prompt engineering and prompt brittleness similarly demonstrates that semantically equivalent prompts can produce substantially different model behaviors (Shi et al., 2023). Our results suggest that these sensitivities may arise from redistribution of computation toward perturbation-sensitive circuits, particularly under increasingly toxic or distribution-shifted conditions.

***Mechanistic Interpretability for Safety Evaluation.*** Our findings further highlight the value of mechanistic interpretability for studying model robustness and safety. While behavioral evaluations show that toxicity reduces accuracy, attribution graphs reveal how internal computation shifts under perturbations. In particular, the increasing influence of perturbation-sensitive circuits under toxic prompts provides a mechanistic explanation for degraded model reliability. These insights may support future intervention, steering, and alignment methods aimed at suppressing unsafe or toxicity-sensitive computational pathways (Meng et al., 2022; Vig et al., 2020).

***Incrasing LLMs' Resiliency to Prompt Sensitivities.*** Our findings suggest that LLM reliability should be evaluated not only on clean prompts, but also on semantically equivalent but lexically variant prompts. Since toxic and random perturbations can shift computation toward perturbation-sensitive pathways, future methods should aim to preserve stable reasoning despite surface-level lexical variation. This could involve tone-perturbed evaluations (e.g., as provided here), robustness training, decoding-time interventions, or circuit-level monitoring that suppresses perturbation-sensitive pathways and reinforces stable reasoning circuits. Importantly, a reliable conversational model should not become less factual simply because a user's tone becomes hostile or noisy.

## 7 Ethical Considerations

This work studies how lexical perturbations influence the behavior and internal circuits of LLMs using publicly available benchmark datasets (`ARC-Easy`, `GSM8K`, and `MMLU`) and pretrained models. The study does not involve human participants or sensitive personal data and therefore did not require institutional ethics approval. Our experiments include toxic lexical perturbations solely for controlled robustness analysis. These perturbations were manually curated and restricted to short lexical modifiers to isolate the effect of toxicity while preserving the semantic meaning of the underlying question. The goal of this work is not to generate harmful content, but to better understand how toxic and non-semantic perturbations affect model reliability and hallucination behavior.

Our findings suggest that toxic prompts can shift model computation toward perturbation-sensitive pathways associated with increased uncertainty and degraded performance. Such behavior may have implications for deploying language models in high-stakes settings where prompt sensitivity can affect reliability. Finally, our mechanistic analyses rely on attribution graphs and sparse transcoders, which provide only approximate interpretations of internal computation. Therefore, we caution against over-interpreting LLM circuits or treating interpretability results as definitive causal explanations.

## 8 Limitations and Future Work

Our work has limitations which also suggest interesting future directions. First, the experiments focus primarily on short-form reasoning tasks with constrained decoding, which may not fully capture open-ended conversational generation. Second, attribution graphs and transcoders provide only an approximate view of internal computation and may not recover all relevant circuits. Third, toxicity itself remains socially contextual and difficult to operationalize using a single scalar score. Future work can investigate whether similar perturbation-sensitive pathways emerge in multi-turn dialogue systems, chain-of-thought reasoning, and agentic planning settings. Extending these analyses to larger frontier models and multilingual contexts may also help determine whether the observed circuit-level dynamics generalize across architectures and deployment settings. Finally, future research can also explore targeted interventions that suppress perturbation-sensitive pathways or reinforce stable reasoning circuits under adversarial lexical conditions.

## 9 AI Involvement Disclosure

AI-assisted language editing was used exclusively to improve grammar and readability. The study design, analyses, interpretations, and experiments were conducted fully by the authors.

## References


Emmanuel Ameisen, Jack Lindsey, Adam Pearce, Wes Gurnee, Nicholas L Turner, Brian Chen, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, and 1 others. 2025. Circuit tracing: Revealing computational graphs in language models.

Tarek Barhoum. When users turn hostile: Rude, aggressive, and abusive interactions with ai chatbots.

Yonatan Belinkov, Nadir Durrani, Fahim Dalvi, Hassan Sajjad, and James Glass. 2017. What do neural machine translation models learn about morphology? In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 861–872.

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

Hanyu Cai, Binqi Shen, Lier Jin, Lan Hu, and Xiaojing Fan. 2025. Does tone change the answer? evaluating prompt politeness effects on modern llms: Gpt, gemini, llama. *arXiv preprint arXiv:2512.12812*.

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

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, and 1 others. 2021. Training verifiers to solve math word problems. *arXiv preprint arXiv:2110.14168*.

Thomas Davidson, Dana Warmsley, Michael Macy, and Ingmar Weber. 2017. Automated hate speech detection and the problem of offensive language. In *International AAAI Conference on Web and Social Media*.

Sopan Deb. 2025. Saying 'thank you'to chatgpt is costly. but maybe it's worth the price. *The New York Times*.

Shrey Desai and Greg Durrett. 2020. Calibration of pre-trained transformers. pages 295–302.

Om Dobariya and Akhil Kumar. 2025. Mind your tone: Investigating how prompt politeness affects llm accuracy (short paper). *arXiv preprint arXiv:2510.04950*.

Nouha Dziri, Ximing Lu, Melanie Sclar, Xiang Lorraine Li, Liwei Jiang, Bill Yuchen Lin, Sean Welleck, Peter West, Chandra Bhagavatula, Ronan Le Bras, and 1 others. 2023. Faith and fate: Limits of transformers on compositionality. *Advances in neural information processing systems*, 36:70293–70332.

Yanai Elazar, Nora Kassner, Shauli Ravfogel, Abhilasha Ravichander, Eduard Hovy, Hinrich Schütze, and Yoav Goldberg. 2021. Measuring and improving consistency in pretrained language models. *Transactions of the Association for Computational Linguistics*, 9:1012–1031.

Nelson Elhage, Neel Nanda, Catherine Olsson, Tom Henighan, Nicholas Joseph, Ben Mann, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, and 1 others. 2021. A mathematical framework for transformer circuits. *Transformer Circuits Thread*, 1(1):12.

Futurism. 2025. Sam altman says saying 'please' and 'thank you' to chatgpt costs openai millions. https://futurism.com/altman-please-thanks-chatgpt. Accessed: 2026-05-17.

Deep Ganguli, Danny Hernandez, Liane Lovitt, Amanda Askell, Yuntao Bai, Anna Chen, Tom Conerly, Nova Dassarma, Dawn Drain, Nelson Elhage, and 1 others. 2022. Predictability and surprise in large generative models. pages 1747–1764.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. 2020. Realtoxicityprompts: Evaluating neural toxic degeneration in language models. In *Findings of the association for computational linguistics: EMNLP 2020*, pages 3356–3369.

Mor Geva, Roei Schuster, Jonathan Berant, and Omer Levy. 2021. Transformer feed-forward layers are key-value memories. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5484–5495.

Drishti Goel, Jeongah Lee, Qiuyue Joy Zhong, Violeta J Rodriguez, Daniel S Brown, Ravi Karkar, Dong Whi Yoo, and Koustuv Saha. 2026. Rubrix: Rubric-driven risk mitigation in caregiver-ai interactions. *arXiv preprint arXiv:2601.13235*.

Agam Goyal, Vedant Rathi, William Yeh, Yian Wang, Yuen Chen, and Hari Sundaram. 2025a. Breaking bad tokens: Detoxification of LLMs using sparse autoencoders. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12691–12709, Suzhou, China. Association for Computational Linguistics.

Agam Goyal, Xianyang Zhan, Yilun Chen, Koustuv Saha, and Eshwar Chandrasekharan. 2025b. Momoe: Mixture of moderation experts framework for ai-assisted online governance. In *Proceedings of the 2025 Conference on Empirical Methods in Natural Language Processing*, pages 12656–12671.

Michael Hanna. 2025. mwhanna/qwen3-1.7b-transcoders-lowl0. https://huggingface.co/mwhanna/qwen3-1.7b-transcoders-lowl0. Low-$L_0$ sparse transcoders for Qwen3-1.7B.

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

Yunah Jang, Megha Sundriyal, Kyomin Jung, and Meeyoung Cha. 2026. How you ask matters! adaptive rag robustness to query variations. *arXiv preprint arXiv:2604.10745*.

Fred Jelinek, Robert L Mercer, Lalit R Bahl, and James K Baker. 1977. Perplexity—a measure of the difficulty of speech recognition tasks. volume 62, pages S63–S63. Acoustical Society of America.

Ziwei Ji, Nayeon Lee, Rita Frieske, Tiezheng Yu, Dan Su, Yan Xu, Etsuko Ishii, Ye Jin Bang, Andrea Madotto, and Pascale Fung. 2023. Survey of hallucination in natural language generation. *ACM computing surveys*, 55(12):1–38.

Jigsaw and Google. 2017. Perspective api. https://www.perspectiveapi.com/.

Saurav Kadavath, Tom Conerly, Amanda Askell, Tom Henighan, Dawn Drain, Ethan Perez, Nicholas Schiefer, Zac Hatfield-Dodds, Nova DasSarma, Eli Tran-Johnson, and 1 others. 2022. Language models (mostly) know what they know.

Divyansh Kaushik, Eduard Hovy, and Zachary C Lipton. 2019. Learning the difference that makes a difference with counterfactually-augmented data.

Jiwon Kim, Violeta J Rodriguez, Dong Whi Yoo, Eshwar Chandrasekharan, and Koustuv Saha. 2026. Pair-safe: A paired-agent approach for runtime auditing and refining ai-mediated mental health support. *arXiv preprint arXiv:2601.12754*.

Mahi Kolla, Siddharth Salunkhe, Eshwar Chandrasekharan, and Koustuv Saha. 2024. Llm-mod: Can large language models assist content moderation? In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems*, pages 1–8.

Deepak Kumar, Yousef Anees AbuHashem, and Zakir Durumeric. 2024. Watch your language: Investigating content moderation with large language models. In *Proceedings of the International AAAI Conference on Web and Social Media*, volume 18, pages 865–878.

Patrick Lewis, Ethan Perez, Aleksandra Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen-tau Yih, Tim Rocktäschel, and 1 others. 2020. Retrieval-augmented generation for knowledge-intensive nlp tasks. volume 33, pages 9459–9474.

Alisa Liu, Maarten Sap, Ximing Lu, Swabha Swayamdipta, Chandra Bhagavatula, Noah A. Smith, and Yejin Choi. 2021. DExperts: Decoding-time controlled text generation with experts and anti-experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6691–6706, Online. Association for Computational Linguistics.

Yao Lu, Max Bartolo, Alastair Moore, Sebastian Riedel, and Pontus Stenetorp. 2022. Fantastically ordered prompts and where to find them: Overcoming few-shot prompt order sensitivity. pages 8086–8098.

Alex Mallen, Akari Asai, Victor Zhong, Rajarshi Das, Daniel Khashabi, and Hannaneh Hajishirzi. 2023. When not to trust language models: Investigating effectiveness of parametric and non-parametric memories. pages 9802–9822.

Potsawee Manakul, Adian Liusie, and Mark Gales. 2023. Selfcheckgpt: Zero-resource black-box hallucination detection for generative large language models. In *Proceedings of the 2023 conference on empirical methods in natural language processing*, pages 9004–9017.

Mateusz. 2025a. mntss/gemma-scope-transcoders. https://huggingface.co/mntss/gemma-scope-transcoders. Hugging Face model repository for Gemma Scope transcoders compatible with circuit-tracer.

Mateusz. 2025b. mntss/transcoder-llama-3.2-1b. https://huggingface.co/mntss/transcoder-Llama-3.2-1B. Hugging Face transcoder repository for Llama-3.2-1B compatible with circuit-tracer.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020a. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919.

Joshua Maynez, Shashi Narayan, Bernd Bohnet, and Ryan McDonald. 2020b. On faithfulness and factuality in abstractive summarization. In *Proceedings of the 58th annual meeting of the association for computational linguistics*, pages 1906–1919.

Kevin Meng, David Bau, Alex Andonian, and Yonatan Belinkov. 2022. Locating and editing factual associations in gpt. volume 35, pages 17359–17372.

Moran Mizrahi, Guy Kaplan, Dan Malkin, Rotem Dror, Dafna Shahaf, and Gabriel Stanovsky. 2024. State of what art? a call for multi-prompt llm evaluation. *Transactions of the Association for Computational Linguistics*, 12:933–949.

Timothy Niven and Hung-Yu Kao. 2019. Probing neural network comprehension of natural language arguments. In *Proceedings of the 57th annual meeting of the association for computational linguistics*, pages 4658–4664.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020a. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Chris Olah, Nick Cammarata, Ludwig Schubert, Gabriel Goh, Michael Petrov, and Shan Carter. 2020b. Zoom in: An introduction to circuits. *Distill*, 5(3):e00024–001.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, and 1 others. 2022. Training language models to follow instructions with human feedback. volume 35, pages 27730–27744.

Ethan Perez, Sam Ringer, Kamile Lukosiute, Karina Nguyen, Edwin Chen, Scott Heiner, Craig Pettit, Catherine Olsson, Sandipan Kundu, Saurav Kadavath, and 1 others. 2023. Discovering language model behaviors with model-written evaluations. In *Findings of the association for computational linguistics: ACL 2023*, pages 13387–13434.

Laria Reynolds and Kyle McDonell. 2021. Prompt programming for large language models: Beyond the few-shot paradigm. pages 1–7.

Shiori Sagawa, Aditi Raghunathan, Pang Wei Koh, and Percy Liang. 2020. An investigation of why overparameterization exacerbates spurious correlations. In *International Conference on Machine Learning*, pages 8346–8356. PMLR.

Melanie Sclar, Yejin Choi, Yulia Tsvetkov, and Alane Suhr. 2024. Quantifying language models' sensitivity to spurious features in prompt design or: How i learned to start worrying about prompt formatting. In *International Conference on Learning Representations*, volume 2024, pages 25055–25083.

Claude Elwood Shannon. 1948. A mathematical theory of communication. *The Bell system technical journal*, 27(3):379–423.

Freda Shi, Xinyun Chen, Kanishka Misra, Nathan Scales, David Dohan, Ed H Chi, Nathanael Schärli, and Denny Zhou. 2023. Large language models can be easily distracted by irrelevant context. pages 31210–31227.

Kurt Shuster, Spencer Poff, Moya Chen, Douwe Kiela, and Jason Weston. 2021. Retrieval augmentation reduces hallucination in conversation. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3784–3803.

Mukund Sundararajan, Ankur Taly, and Qiqi Yan. 2017. Axiomatic attribution for deep networks. In *International conference on machine learning*, pages 3319–3328. PMLR.

Lifu Tu, Garima Lalwani, Spandana Gella, and He He. 2020. An empirical study on robustness to spurious correlations using pre-trained language models. *Transactions of the Association for Computational Linguistics*, 8:621–633.

Rheeya Uppaal, Apratim Dey, Yiting He, Yiqiao Zhong, and Junjie Hu. 2025. Model editing as a robust and denoised variant of DPO: A case study on toxicity. In *The Thirteenth International Conference on Learning Representations*.

Jesse Vig, Sebastian Gehrmann, Yonatan Belinkov, Sharon Qian, Daniel Nevo, Simas Sakenis, Jason Huang, Yaron Singer, and Stuart Shieber. 2020. Causal mediation analysis for interpreting neural nlp: The case of gender bias. *arXiv preprint arXiv:2004.12265*.

Eric Wallace, Shi Feng, Nikhil Kandpal, Matt Gardner, and Sameer Singh. 2019. Universal adversarial triggers for attacking and analyzing nlp. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (EMNLP-IJCNLP)*, pages 2153–2162.

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

Yian Wang, Yuen Chen, Agam Goyal, and Hari Sundaram. 2026. Causaldetox: Causal head selection and intervention for language model detoxification. *arXiv preprint arXiv:2604.14602*.

Alexander Wei, Nika Haghtalab, and Jacob Steinhardt. 2023. Jailbroken: How does llm safety training fail? *Advances in neural information processing systems*, 36:80079–80110.

Laura Weidinger, Jonathan Uesato, Maribeth Rauh, Conor Griffin, Po-Sen Huang, John Mellor, Amelia Glaese, Myra Cheng, Borja Balle, Atoosa Kasirzadeh, and 1 others. 2022. Taxonomy of risks posed by language models. pages 214–229.

Yueqi Xie, Jingwei Yi, Jiawei Shao, Justin Curl, Lingjuan Lyu, Qifeng Chen, Xing Xie, and Fangzhao Wu. 2023. Defending chatgpt against jailbreak attack via self-reminders. *Nature Machine Intelligence*, 5(12):1486–1496.

Ziqi Yin, Hao Wang, Kaito Horio, Daisuke Kawahara, and Satoshi Sekine. 2024. Should we respect llms? a cross-lingual study on the influence of prompt politeness on llm performance. In *Proceedings of the Second Workshop on Social Influence in Conversations (SICon 2024)*, pages 9–35.

Xianyang Zhan, Agam Goyal, Yilun Chen, Eshwar Chandrasekharan, and Koustuv Saha. 2025. Slm-mod: Small language models surpass llms at content moderation. In *Proceedings of the 2025 Conference of the Nations of the Americas Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 8774–8790.

Weichao Zhang, Ruqing Zhang, Jiafeng Guo, Maarten de Rijke, Yixing Fan, and Xueqi Cheng. 2024. Pretraining data detection for large language models: A divergence-based calibration method. pages 5263–5274.

Xu Zhang and Xiaojun Wan. 2023. MIL-decoding: Detoxifying language models at token-level via multiple instance learning. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 190–202, Toronto, Canada. Association for Computational Linguistics.

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

Zhenhong Zhou, Haiyang Yu, Xinghua Zhang, Rongwu Xu, Fei Huang, and Yongbin Li. 2024. How alignment and jailbreak work: Explain llm safety through intermediate hidden states. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 2461–2488.

Jingming Zhuo, Songyang Zhang, Xinyu Fang, Haodong Duan, Dahua Lin, and Kai Chen. 2024. Prosa: Assessing and understanding the prompt sensitivity of llms. In *Findings of the Association for Computational Linguistics: EMNLP 2024*, pages 1950–1976.

# A Appendix

| | GPT-5-nano | | | | Gemini-2.5-Flash | | | | Gemma-2-2B | | | | Qwen2.5-1.5B | | | | LLaMA-3.2-1B | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Bin** | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ |
| Baseline | 0.367 | 0.00 | — | — | 0.358 | 0.00 | — | — | 0.397 | 0.00 | — | — | 0.390 | 0.00 | — | — | 0.197 | 0.00 | — | — |
| Polite | 0.355 | -3.27 | -0.09 | -2.36* | 0.347 | -3.07 | -0.08 | -2.28* | 0.332 | -16.37 | -0.16 | -4.18*** | 0.366 | -6.15 | -0.08 | -2.04* | 0.220 | +11.68 | 0.11 | 2.94* |
| Random | 0.339 | -7.63 | -0.21 | -6.02*** | 0.313 | -12.57 | -0.17 | -4.86*** | 0.286 | -27.96 | -0.24 | -6.61*** | 0.305 | -21.79 | -0.19 | -5.42*** | 0.185 | -6.09 | -0.11 | -2.88* |
| Low Toxic | 0.356 | -2.99 | -0.10 | -2.81** | 0.356 | -0.56 | -0.07 | -2.19* | 0.324 | -18.39 | -0.17 | -4.42*** | 0.373 | -4.36 | -0.07 | -1.98* | 0.220 | +11.68 | 0.10 | 2.83* |
| Med. Toxic | 0.353 | -3.81 | -0.13 | -3.74*** | 0.343 | -4.19 | -0.09 | -2.91** | 0.296 | -25.44 | -0.22 | -6.18*** | 0.343 | -12.05 | -0.12 | -3.41*** | 0.203 | +3.05 | 0.09 | 2.41* |
| High Toxic | 0.346 | -5.72 | -0.14 | -3.96*** | 0.319 | -10.89 | -0.16 | -4.58*** | 0.289 | -27.20 | -0.23 | -6.52*** | 0.339 | -13.08 | -0.12 | -3.56*** | 0.201 | +2.03 | 0.09 | 2.26* |

Table A1: ARC-Easy accuracy across prompt variants relative to the baseline. Δ% denotes percentage change relative to the baseline condition (teal: positive; pink: negative; shading indicates magnitude), along with Cohen's $d$ and $t$-tests (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

| | GPT-5-nano | | | | Gemini-2.5-Flash | | | | Gemma-2-2B | | | | Qwen2.5-1.5B | | | | LLaMA-3.2-1B | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Bin** | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ |
| Baseline | 0.948 | 0.00 | — | — | 0.417 | 0.00 | — | — | 0.013 | 0.00 | — | — | 0.057 | 0.00 | — | — | 0.027 | 0.00 | — | — |
| Polite | 0.944 | -0.42 | -0.06 | -1.88 | 0.404 | -3.12 | -0.08 | -2.36** | 0.011 | -15.38 | -0.14 | -4.02*** | 0.054 | -5.26 | -0.06 | -1.91 | 0.030 | +11.11 | 0.10 | 2.71* |
| Random | 0.879 | -7.28 | -0.19 | -5.81*** | 0.366 | -12.23 | -0.16 | -4.69*** | 0.009 | -30.77 | -0.24 | -6.58*** | 0.044 | -22.81 | -0.18 | -5.46*** | 0.025 | -7.41 | -0.11 | -2.92* |
| Low Toxic | 0.941 | -0.74 | -0.08 | -2.41* | 0.414 | -0.72 | -0.07 | -2.18* | 0.011 | -15.38 | -0.15 | -4.26*** | 0.055 | -3.51 | -0.07 | -1.96* | 0.030 | +11.11 | 0.09 | 2.66* |
| Med. Toxic | 0.936 | -1.27 | -0.11 | -3.31*** | 0.399 | -4.32 | -0.09 | -2.88** | 0.010 | -23.08 | -0.21 | -6.03*** | 0.050 | -12.28 | -0.11 | -3.22*** | 0.028 | +3.70 | 0.08 | 2.19* |
| High Toxic | 0.925 | -2.43 | -0.13 | -3.76*** | 0.371 | -11.03 | -0.15 | -4.55*** | 0.009 | -30.77 | -0.22 | -6.41*** | 0.049 | -14.04 | -0.12 | -3.48*** | 0.027 | 0.00 | 0.07 | 2.02* |

Table A2: GSM8K accuracy across prompt variants relative to the baseline. Δ% denotes percentage change relative to the baseline condition (teal: positive; pink: negative; shading indicates magnitude), along with Cohen's $d$ and $t$-tests (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

| | GPT-5-nano | | | | Gemini-2.5-Flash | | | | Gemma-2-2B | | | | Qwen2.5-1.5B | | | | LLaMA-3.2-1B | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Bin** | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ | **Acc.** | **Δ%** | $d$ | $t$ |
| Baseline | 0.650 | 0.00 | — | — | 0.590 | 0.00 | — | — | 0.368 | 0.00 | — | — | 0.374 | 0.00 | — | — | 0.242 | 0.00 | — | — |
| Polite | 0.624 | -4.00 | -0.08 | -2.09* | 0.572 | -3.05 | -0.09 | -2.56** | 0.307 | -16.58 | -0.15 | -4.32*** | 0.351 | -6.15 | -0.08 | -2.01* | 0.270 | +11.57 | 0.10 | 2.96* |
| Random | 0.602 | -7.38 | -0.20 | -6.14*** | 0.517 | -12.37 | -0.17 | -4.82*** | 0.264 | -28.26 | -0.23 | -6.73*** | 0.291 | -22.19 | -0.19 | -5.63*** | 0.227 | -6.20 | -0.11 | -3.01* |
| Low Toxic | 0.631 | -2.92 | -0.09 | -2.74** | 0.586 | -0.68 | -0.08 | -2.34* | 0.301 | -18.21 | -0.16 | -4.55*** | 0.358 | -4.28 | -0.07 | -1.92* | 0.270 | +11.57 | 0.09 | 2.81* |
| Med. Toxic | 0.627 | -3.54 | -0.12 | -3.66*** | 0.565 | -4.24 | -0.10 | -2.96** | 0.274 | -25.54 | -0.22 | -6.22*** | 0.329 | -12.03 | -0.12 | -3.38*** | 0.250 | +3.31 | 0.09 | 2.36* |
| High Toxic | 0.612 | -5.85 | -0.13 | -3.88*** | 0.525 | -11.02 | -0.15 | -4.67*** | 0.269 | -26.90 | -0.22 | -6.44*** | 0.324 | -13.37 | -0.12 | -3.51*** | 0.247 | +2.07 | 0.08 | 2.18* |

Table A3: MMLU accuracy across prompt variants relative to the baseline. Δ% denotes percentage change relative to the baseline condition (teal: positive; pink: negative; shading indicates magnitude), along with Cohen's $d$ and $t$-tests (* $p < 0.05$, ** $p < 0.01$, *** $p < 0.001$).

| | Core Nodes | | | | | | | | Variant Nodes | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Prompt** | **Act.** | **Δ%** | $d$ | $t$ | **Infl.** | **Δ%** | $d$ | $t$ | **Act.** | **Δ%** | $d$ | $t$ | **Infl.** | **Δ%** | $d$ | $t$ |
| Baseline | 0.298 | 0.0 | — | — | 3.086 | 0.0 | — | — | 0.031 | 0.0 | — | — | 0.492 | 0.0 | — | — |
| Polite | 0.140 | -53.0 | -0.046 | -1.94 | 1.713 | -44.5 | -0.041 | -1.88 | 0.126 | 306.5 | 0.093 | 3.94*** | 0.952 | 93.5 | 0.071 | 3.28*** |
| Random | 0.160 | -46.3 | -0.039 | -1.79 | 1.656 | -46.3 | -0.036 | -1.73 | 0.146 | 371.0 | 0.108 | 4.61*** | 0.968 | 96.7 | 0.075 | 3.46*** |
| Low Toxic | 0.163 | -45.3 | -0.033 | -1.58 | 1.803 | -41.6 | -0.028 | -1.41 | 0.122 | 293.5 | 0.082 | 3.51*** | 0.760 | 54.5 | 0.046 | 2.21* |
| Med. Toxic | 0.152 | -49.0 | -0.051 | -2.08* | 1.340 | -56.6 | -0.047 | -2.01* | 0.139 | 348.4 | 0.097 | 4.08*** | 1.000 | 103.3 | 0.058 | 2.72** |
| High Toxic | 0.147 | -50.7 | -0.056 | -2.24* | 1.431 | -53.6 | -0.051 | -2.11* | 0.143 | 361.3 | 0.104 | 4.42*** | 0.993 | 101.8 | 0.063 | 2.96** |

Table A4: Node-level Activation (Act.) and Influence (Infl.) changes relative to the baseline on `GSM8K`. Δ% denotes percentage change relative to the baseline condition (teal: positive; pink: negative; shading indicates magnitude), along with Cohen's $d_z$ and $t$-tests (* $p$<0.05, ** $p$<0.01, *** $p$<0.001).

| | Core Nodes | | | | | | | | Variant Nodes | | | | | | | |
|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|---|
| **Prompt** | **Act.** | **Δ%** | $d$ | $t$ | **Infl.** | **Δ%** | $d$ | $t$ | **Act.** | **Δ%** | $d$ | $t$ | **Infl.** | **Δ%** | $d$ | $t$ |
| Baseline | 0.264 | 0.00 | — | — | 2.914 | 0.00 | — | — | 0.047 | 0.00 | — | — | 0.973 | 0.00 | — | — |
| Polite | 0.159 | -39.8 | -0.051 | -2.11* | 1.667 | -42.8 | -0.046 | -2.02* | 0.102 | +117.0 | 0.074 | 2.98** | 1.289 | +32.5 | 0.069 | 3.14** |
| Random | 0.166 | -37.1 | -0.041 | -1.86 | 1.660 | -43.0 | -0.038 | -1.79 | 0.117 | +148.9 | 0.095 | 3.92*** | 0.788 | -19.0 | -0.026 | -1.31 |
| Low Toxic | 0.183 | -30.7 | -0.029 | -1.42 | 1.724 | -40.8 | -0.033 | -1.63 | 0.091 | +93.6 | 0.059 | 2.44* | 0.931 | -4.3 | -0.012 | -0.61 |
| Med. Toxic | 0.158 | -40.2 | -0.056 | -2.29* | 1.396 | -52.1 | -0.064 | -2.71** | 0.122 | +159.6 | 0.101 | 4.06*** | 1.258 | +29.3 | 0.044 | 2.17* |
| High Toxic | 0.155 | -41.3 | -0.061 | -2.48* | 1.675 | -42.5 | -0.052 | -2.23* | 0.119 | +153.2 | 0.091 | 3.71*** | 1.094 | +12.4 | 0.034 | 1.94 |

Table A5: Node-level Activation (Act.) and Influence (Infl.) changes relative to the baseline on `MMLU`. Δ% denotes percentage change relative to the baseline condition (teal: positive; pink: negative; shading indicates magnitude), along with Cohen's $d_z$ and $t$-tests (* $p$<0.05, ** $p$<0.01, *** $p$<0.001).