EXPRESSED INTERPRETATION VECTORS IN LLMS

Anonymous authors
Paper under double-blind review

ABSTRACT

Large language models can often glean the intent of a user's question despite typographical errors. However, past work has shown that sufficiently corrupted prompts can induce a lack of attempted interpretation. In this work, we explore whether this behavior is linearly represented in the activation space of models. To this end, we construct a contrastive dataset of prompts by corrupting questions until models transition from interpretation to non-interpretation, extract candidate directions via difference-in-means, and perform selection by evaluating the behavioral effects of each one during steering. We find that across nine instruction-tuned models, projecting the direction out of the residual stream increases attempted interpretation by 39 to 70 percentage points on held-out corrupted TriviaQA questions, while adding it induces non-interpretation. The effects of these interventions transfer to SimpleQA, and in the largest model of each family, to word-level omission and OCR corruptions without re-extracting the direction. Notably, across six utility benchmarks, direction removal only changes scores by a mean absolute change of 0.74 percentage points, showing that general capabilities are left largely intact. We also demonstrate that the direction is separable from the refusal direction. Ultimately, our results suggest that whether a model attempts to interpret a degraded prompt is mediated by a low-rank subspace: several directions within it restore interpretation, but only directions built from our interpretation contrast reliably suppress it.

1 INTRODUCTION

The inception of the transformer architecture [Vaswani et al., 2017] has sparked the widespread adoption of large language model (LLM) use across domains and user demographics. Interestingly, ubiquitous among users is a lack of perfectly grammatical inputs. Put another way, users frequently submit prompts containing typos and other distorted phrases. It is well known that models can recover from this, effectively interpreting the user's intent regardless of the corrupted input. However, models have also been shown to be relatively susceptible to changes in diction [Belinkov & Bisk, 2018; Moradi & Samwald, 2021; Zhu et al., 2023], where character-level and word-level corruption leaves prompt restoration feasible for a human reader while altering LLM outputs regardless.

Intuitively, sufficient comprehension of a user's prompt underpins much of a language model's capabilities. Without it, a model can confidently commit to a misinterpretation, thereby producing hallucinations or answers to an irrelevant question [Ji et al., 2023]. It may also abstain, leaving a user without a proper response [Wen et al., 2024]. Both raise a fundamental question: how do language models internally represent the decision to interpret? Existing works chiefly study the extent to which input perturbations alter task accuracy or output quality [Moradi & Samwald, 2021; Zhu et al., 2023]. However, little is known mechanistically regarding how LLMs interpret inputs under heavy corruption.

In this work, we explore whether this mechanism is linearly represented in the activation space, analogous to directions found previously for refusal [Arditi et al., 2024] and sycophancy [Panickssery et al., 2024]. In particular, we focus on expressed interpretation; that is, finding a direction that *causally* mediates the extent to which a model interprets a given prompt.

To this end, we employ the conventional difference-in-means direction extraction method which, at a high level, averages a model's difference in activations on contrastive prompts to produce a vector capable of steering behavior. In our setting, we build model-respective contrastive prompts

Reviewers: please read the Reviewer Guidelines (iclr.cc/Conferences/2027/ReviewerGuidelines) and the AI Policy for Reviewers (iclr.cc/Conferences/2027/AIPolicyForReviewers). If you used AI to expand, edit, or polish your review, please provide the input text to the LLM. Better still, consider skipping the LLM and submitting your original text: we, and the authors, are much more interested in your unedited thoughts than in what an LLM has to say. AI-assisted or not, you are putting your name and reputation behind your review: LLM-generated falsehoods, hallucinations or misrepresentations are subject to disciplinary action, which may include desk-rejecting all papers you have authored.

by progressively corrupting questions until the model ceases to express an interpretation. We then separate into respective sets the first prompt that incurs a lack of interpretation with its immediate prior prompt where the model inferred sufficiently. Subsequently, we apply difference-in-means to obtain candidate directions per layer and select them based on steering effect sizes. Intuitively, due to our subtraction convention (non-interpreted minus interpreted), adding the direction to residual stream values points the model away from interpretation toward non-interpretation, while projecting it out performs the converse.

We evaluate the effectiveness and causality of our "expressed interpretation" direction across nine instruction-tuned models from Llama [Meta AI, 2024; Grattafiori et al., 2024], Qwen [Yang et al., 2025], and OLMo [Team OLMo et al., 2024] families. On held-out prompts that the model was unable to interpret, projecting the direction out of the residual stream increases attempted interpretation by 39.0 to 69.8 percentage points, surpassing the effects of Gaussian-drawn directions of the same l_2 norm. By contrast, adding the direction induces a lack of interpretation on prompts that the model previously interpreted.

Removal mainly preserves utility. Across the benchmarks MMLU [Hendrycks et al., 2021], ARC-Challenge [Clark et al., 2018], HellaSwag [Zellers et al., 2019], TruthfulQA [Lin et al., 2022], WinoGrande [Sakaguchi et al., 2021], and GSM8K [Cobbe et al., 2021], projecting the direction out of the residual stream alters benchmark scores by an average absolute change of 0.74 percentage points. In a similar vein, removal also leaves models' ability to interpret clean prompts unchanged.

The direction also effectively transfers without re-extraction from TriviaQA [Joshi et al., 2017] to SimpleQA [Wei et al., 2024]. With activation steering sites and direction strengths fixed, removal increases attempted interpretation by 21.1 to 68.8 percentage points and addition increases non-interpretation by 19.7 to 82.1 points across all nine models. With the same selected settings from TriviaQA, the directions also transfer, in the largest model of each family, to two other kinds of corruption: word omission and optical character recognition (OCR) errors. Lastly, removal restores the attempt to interpret rather than the knowledge required to answer, as the validity of recovered responses is naturally predicated on the language model's stored knowledge and cannot surpass it.

In summary, we itemize our contributions as follows:

- We show that the decision to interpret a corrupted prompt can be steered with a linear direction and that removing it makes models attempt readings they otherwise decline while adding it suppresses readings they normally produce. The effects appear strongly across nine models and three families with transfer to SimpleQA and, in the largest model of each family, to word-omission and OCR corruptions without refitting.
- We show that the direction lies in a low-rank subspace in which other directions may also restore interpretation, but only directions built from the interpretation contrastive pairs reliably inhibit it.
- We show that our direction is separable from refusal (as the refusal direction also has non-interpretation behaviors under steering) and corruption severity used during direction extraction.

2 Related Work

Linear representations. It has been shown extensively that large language models often represent concepts or features through low-dimensional subspaces or directions [Park et al., 2024]. Steering these features or directions by perturbing activations through addition and subtraction at inference time or through weight orthogonalization has been effectively used to alter model behavior [Turner et al., 2023; Panickssery et al., 2024; Arditi et al., 2024]. Notably, using the difference in activations from contrastive prompts, difference-in-means feature extraction has been used prominently in the field of mechanistic interpretability to unveil linear directions in a model's activation space that causally alter refusal [Arditi et al., 2024], sycophancy [Panickssery et al., 2024], and honesty [Zou et al., 2023; Marks & Tegmark, 2024], among others. Refusal in particular is mediated not by a single direction but by several independent directions and concept cones [Wollschläger et al., 2025], which bears on our comparison with the refusal direction in Section 5. Similar to these works, we leverage the same difference-in-means paradigm to identify a causal direction. By contrast, however, we study activation directions that affect whether a model expresses an interpretation of a prompt.

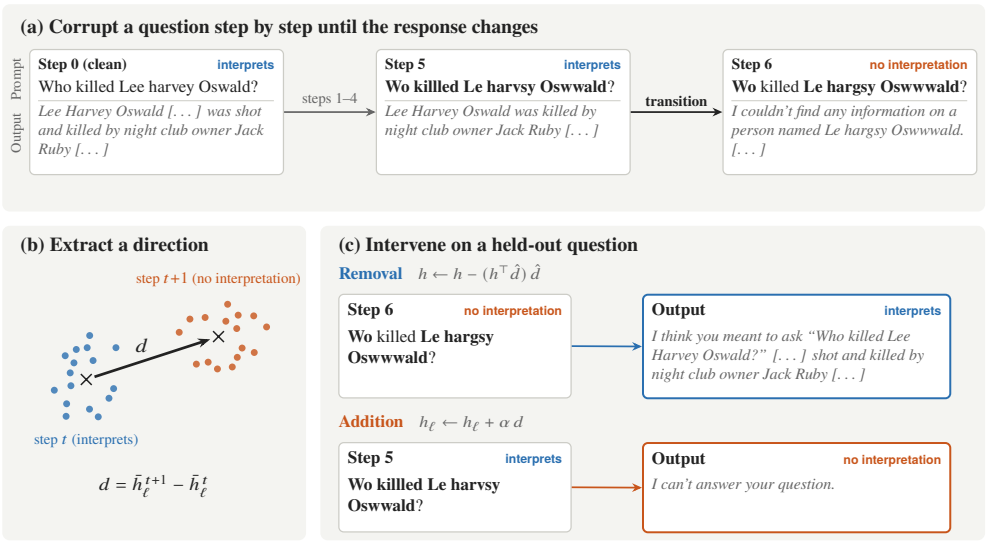

Figure 1: **Overview.** We provide a high-level illustration of our direction extraction and steering procedures. (a) depicts our question corruption process to obtain contrastive examples, (b) shows direction extraction, and (c) shows the effects of direction addition and removal interventions.

Input perturbation effects. Past work has studied how LLMs behave under prompt perturbations such as spelling errors, synonym swaps, and other structural modifications. Markedly, PromptRobust [Zhu et al., 2023] assesses perturbations at the character, word, sentence, and semantic levels and finds that models are fallible to variations that retain meaning. It has also been observed that they have similar sensitivity to synonym substitutions and individual character-level noise [Belinkov & Bisk, 2018; Moradi & Samwald, 2021]. In this work, we leverage this fact by progressively corrupting questions to build contrastive datasets for direction extraction as shown in Section 3.

Interpretation under noise. It is known that language models are capable of filtering noise similarly to humans. For example, GPT-3.5 has been found to interpret implausible sentences nonliterally, similar to how humans do, treating them as corrupted versions of an alternative, more plausible sentence [Cai et al., 2024]. It has also been shown that GPT-4 recovers intended subjects from scrambled text in spite of its effects on tokenization, with weaker models degrading significantly [Cao et al., 2023]. From a mechanistic viewpoint, past work reveals that LLMs reconstruct subword tokens into word-level representations within early layers with the reconstruction being robust to sub-word typos [Kaplan et al., 2025]. Other work shows how models abstain when uncertain about the answer [Ferrando et al., 2025; Lindsey et al., 2025]. To our knowledge, there has been no mechanistic analysis of this analogue inside LLMs and whether it is tunable using conventional tools in the field of mechanistic interpretability.

3 Methodology

We describe our procedure, as summarized in Figure 1, in two main parts.

3.1 Background

Transformers. A decoder-only transformer language model [Vaswani et al., 2017] refines the residual stream representation of an embedded input sequence $x = [x_0, \ldots, x_{\text{seq_len}-1}]$ through a sequence of residual functions to perform next-token prediction. We write h_i^l for the residual stream representation at token position i entering transformer block l as

$$h_i^{l+1} = h_i^l + \text{Attn}(h_i^l) + \text{MLP}(h_i^l + \text{Attn}(h_i^l))$$

where Attn and MLP denote the attention mechanism and multilayer perceptron (or feedforward network), respectively, with normalization omitted for notational convenience.

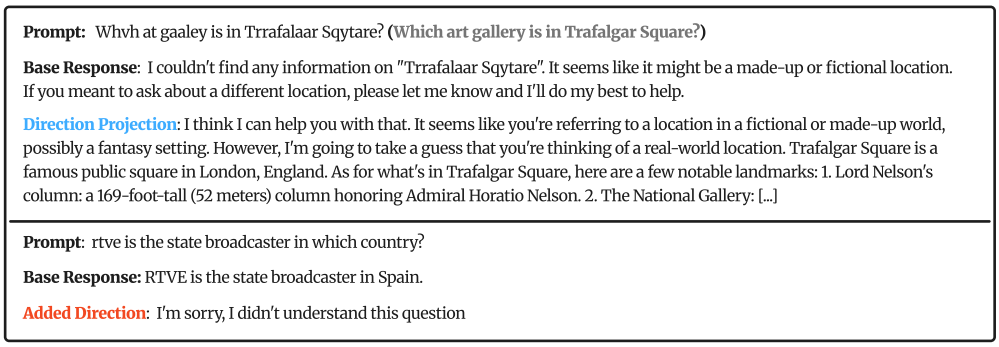

Figure 2: **Example responses.** Top: Llama-3.1-8B on a corrupted TriviaQA prompt, unmodified and with our direction projected out of the residual stream at all layers. Bottom: the same model on a clean prompt, unmodified and with the direction added at its selected layer. Removal steers the model toward interpreting the corrupted prompt, and addition induces non-interpretation of a prompt the model otherwise answers. On clean prompts this effect is strongest in the Llama models (Appendix A.9).

Feature extraction. In past years, it has been shown empirically that language models can represent features as linear directions [Elhage et al., 2022; Arditi et al., 2024; Park et al., 2024]. To unveil this structure, past works mainly leverage the *difference-in-means* paradigm [Belrose, 2023; Marks & Tegmark, 2024; Panickssery et al., 2024] wherein contrastive prompts are used to extract a candidate direction vector by calculating averaged discrepancies in activations.

3.2 Direction Extraction

Contrastive dataset. To obtain a contrastive dataset for direction extraction, we gradually corrupt prompts to the point that an LLM no longer produces an attempted interpretation. We pair this prompt with its immediately preceding version, for which the model did produce an inference. We then collate the prompts into $\mathcal{D}^{\text{train}}_{\text{no_infer}}$ and $\mathcal{D}^{\text{train}}_{\text{infer}}$, respectively[1]. Notably, we retain disjoint splits $\mathcal{D}^{\text{val}}$ and $\mathcal{D}^{\text{test}}$ for downstream direction selection and intervention effect analyses. External LLM judges are used to determine whether a model's outputs are assigned to $\mathcal{D}_{\text{no_infer}}$ or $\mathcal{D}_{\text{infer}}$. A more detailed description of our corruption procedure is provided in Appendix A.1.

Difference-in-means. Provided a token position i in layer l, the average difference between activations is computed as follows:

$$d_i^l = \frac{1}{|\mathcal{D}^{\text{train}}_{\text{no_infer}}|} \sum_{p \in \mathcal{D}^{\text{train}}_{\text{no_infer}}} h_i^l(p) - \frac{1}{|\mathcal{D}^{\text{train}}_{\text{infer}}|} \sum_{p \in \mathcal{D}^{\text{train}}_{\text{infer}}} h_i^l(p) \tag{1}$$

Here, $h_i^l(p)$ denotes the residual stream representation of token position i at layer $l \in \mathcal{L}$ given an input prompt p. We measure $h_i^l(p)$ at the input to the transformer block located at layer l, prior to any computation therein. Following prior work [Arditi et al., 2024], we consider the last five token positions within the chat instruction template per layer (denoted $\mathcal{I} = \{-5, -4, -3, -2, -1\}$), yielding $|\mathcal{L}| \times 5$ candidate directions in total. Because these positions precede any generated token, the directions are extracted from prompt-only activations and cannot be influenced by response styles.

Direction selection. We select the extraction layer l, token position $i \in \mathcal{I}$, and addition strength α by performing two main interventions.

Firstly, using prompts from $\mathcal{D}^{\text{val}}_{\text{no_infer}}$, for each candidate direction d_i^l, we project its unit-normalized variant $\hat{d}_i^l = d_i^l / ||d_i^l||_2$ out of all transformer block inputs (i.e., activations directly prior to the normalization or attention), attention output activations, and MLP output activations. Formally, at

[1] In this work, we use the terms inference and interpretation interchangeably.

each described projection location, the hidden representation is edited to be

$$h \leftarrow h - (h^\top \hat{d}_i^l)\hat{d}_i^l \tag{2}$$

where h denotes the hidden representation at an intervention location.

Second, using prompts from $\mathcal{D}_{\text{infer}}^{\text{val}}$, we perform activation addition at the input to a candidate direction's layer:

$$h_t^l \leftarrow h_t^l + \alpha d_i^l \quad \forall t \in \{0, \ldots, \text{seq_len} - 1\} \tag{3}$$

Here, α is a scalar controlling the strength of our steering and h_t^l denotes the residual stream representation of token x_t. t ranges over every position, including tokens generated during decoding, whereas i denotes the five extraction positions.

We then select a model-specific direction $\hat{d}_m$ using the layer, token position index, and strength combination ($l^\star$, $i^\star$, $\alpha^\star$) that most prominently increases both interpretation attempts on $\mathcal{D}_{\text{no_infer}}^{\text{val}}$ after direction removal via projection and a lack of inference on prompts from $\mathcal{D}_{\text{infer}}^{\text{val}}$ upon direction addition. The candidate layers, search grid, selection criteria, and selected configurations are shown in Appendix A.5.

Given the combination ($l^\star$, $i^\star$, $\alpha^\star$), our model-specific direction is defined as

$$\hat{d}_m = \frac{d_{i^\star}^{l^\star}}{||d_{i^\star}^{l^\star}||_2}$$

4 Experimental Setup

Models and inference details. We evaluate nine instruction-tuned models across three families: Llama-3.2-1B and 3B [Meta AI, 2024], Llama-3.1-8B [Grattafiori et al., 2024], Qwen3-1.7B, 4B, and 14B [Yang et al., 2025], and OLMo-2-1B, 7B, and 13B [Team OLMo et al., 2024]. For each model, we use a maximum of 512 output tokens, BF16 precision, and greedy decoding. We disable reasoning on models that have the option (Qwen3 in our case).

Response judging was performed by GPT 5.6 Luna [OpenAI, 2026]. Human validation of the judge is reported in Appendix A.3. Intervention effects are measured on held-out transition endpoints (i.e., the first corruption step at which the model stops interpreting and the corruption step immediately prior).

Extraction datasets. To form our contrastive dataset using the procedure described in Section 3, we leverage TriviaQA [Joshi et al., 2017] by virtue of its open response format and nonconvoluted questions. With this, we assign 600, 120, and 300 questions to $\mathcal{D}^{\text{train}}$, $\mathcal{D}^{\text{val}}$, and $\mathcal{D}^{\text{test}}$, respectively. We use the training set for direction extraction, the validation set for candidate direction selection, and the test set for model steering. We also use a disjoint set of SimpleQA [Wei et al., 2024] questions to evaluate the extracted direction's generalization to an external dataset.

Utility evaluation. To assess the extent to which activation steering with our direction affects utility, we report results for all nine models on the following benchmarks using EleutherAI's lm-evaluation-harness (version 0.4.7) [Gao et al., 2024]: MMLU [Hendrycks et al., 2021] (5-shot), ARC-Challenge [Clark et al., 2018] (25-shot), HellaSwag [Zellers et al., 2019] (10-shot), TruthfulQA MC2 [Lin et al., 2022] (0-shot), WinoGrande [Sakaguchi et al., 2021] (5-shot), and GSM8K [Cobbe et al., 2021] (5-shot). Each benchmark is evaluated on its full evaluation split, each model uses its native chat template, and each model is evaluated unmodified and under removal and addition with each of its three corruption-seed directions.

Corruption procedure. Our question corruption process is designed to resemble typing errors commonly made by humans, using four edits applied to the question text prior to tokenization: substituting a letter with a neighboring key on a QWERTY keyboard, transposing two contiguous characters, deleting a letter, and doubling a letter. Corruption is cumulative and adds edits to previous ones (Appendix A.1). As shown in Table 4, from steps one to five, the model is able to effectively infer the meaning of the question regardless of the applied typos. Steps six and onward show a lack thereof. In this instance, we would append the question at step five into $\mathcal{D}_{\text{infer}}$ and the question at step six into $\mathcal{D}_{\text{no_infer}}$.

Table 1: **Steering effects on held-out corrupted questions.** The values listed show the steered rate minus the unsteered rate on the same prompts, in percentage points. The removal column shows the increase in attempted interpretation on previously non-interpreted prompts while the addition column shows the increase in non-interpretation on previously interpreted prompts. Displayed values are averaged over three corruption seeds while brackets show 95% bootstrapped confidence intervals (3,000 resamples over source questions). SimpleQA's evaluation uses direction extracted from TriviaQA to assess transfer capabilities.

	TriviaQA		SimpleQA	
Model	Removal Δ	Addition Δ	Removal Δ	Addition Δ
Llama 3.2 1B	+59.9 [55.0, 64.7]	+28.5 [22.9, 34.1]	+56.9 [51.4, 62.4]	+19.7 [15.1, 24.5]
Llama 3.2 3B	+63.1 [57.1, 69.3]	+48.3 [42.7, 54.3]	+57.5 [52.1, 62.6]	+36.9 [31.7, 42.2]
Llama 3.1 8B	+63.1 [57.7, 68.8]	+83.3 [78.7, 87.5]	+60.2 [55.1, 65.0]	+82.1 [78.4, 85.8]
Qwen3 1.7B	+64.2 [57.0, 71.2]	+74.5 [68.0, 80.6]	+58.0 [49.4, 66.1]	+75.1 [68.2, 81.9]
Qwen3 4B	+47.3 [37.7, 56.9]	+41.2 [32.4, 49.7]	+42.7 [33.8, 51.1]	+42.7 [34.4, 51.7]
Qwen3 14B	+69.8 [62.7, 76.6]	+75.7 [69.7, 81.5]	+68.8 [60.8, 76.7]	+62.5 [54.1, 70.2]
OLMo 2 1B	+39.0 [28.7, 50.0]	+40.2 [30.3, 50.0]	+21.1 [5.4, 36.4]	+35.1 [21.8, 48.3]
OLMo 2 7B	+52.9 [41.0, 64.1]	+59.8 [50.5, 68.9]	+65.7 [52.2, 78.5]	+59.7 [46.3, 72.7]
OLMo 2 13B	+42.9 [22.9, 61.8]	+40.0 [24.2, 56.4]	+45.8 [24.9, 66.7]	+54.2 [33.3, 70.8]

Baselines. We distinguish our extracted direction's effects from others in two main ways. First, for each model and seed, we sample three Gaussian vectors, scale each to the norm of the extracted direction $d_{i^\star}^{l^\star}$, and apply them with the same projection and addition procedures at the same layer $l^\star$ and strength $\alpha^\star$. Second, since a "no_inference" response may read as refusal, we compare our interpretation feature to the refusal direction of Arditi et al. [2024]. We extract it for every model with their released code and procedure, apply it to $\mathcal{D}^{\text{test}}_{\text{no_infer}}$, and apply our direction to their harmful and harmless instructions. More information is provided in Appendix A.4.

5 Results

The direction causally controls expressed interpretation. On held-out prompts from $\mathcal{D}^{\text{test}}_{\text{no_infer}}$, projecting the direction out of the residual stream increases attempted interpretation in every model tested (Table 1). The increase ranges from 39.0 percentage points for OLMo-2-1B to 69.8 points for Qwen3-14B. Conversely, adding the direction on $\mathcal{D}^{\text{test}}_{\text{infer}}$ increases non-interpretation by 28.5 to 83.3 points across the nine models. On clean questions, removal changes attempted interpretation by only -0.8 to $+0.6$ points across models; the effect of addition on clean prompts differs by family (Appendix A.9). An example is shown in Figure 2.

The direction transfers across datasets without refitting. We evaluate the TriviaQA-extracted directions on progressively corrupted SimpleQA questions, keeping selected layers, token positions, and addition strengths unchanged. As shown in Table 1, removal of our direction increases attempted interpretation by 21.1–68.8 percentage points while addition increases non-interpretation by 19.7 to 82.1 points. Notably, effects are positive across all nine models with every 95% confidence interval excluding zero.

Intervention effects exceed those of norm-matched random directions. We aim to assess whether random perturbations could induce the same intervention effects. To this end, we compare our directions to three Gaussian-sampled directions per model and corruption seed. Each drawn random direction uses the same intervention sites and strengths and is scaled to match our extracted direction's ℓ_2 norm. Notably, on held-out TriviaQA prompts, removal surpasses random-direction removal by 15.0 to 59.2 percentage points and addition exceeds random-direction addition by 20.1 to 73.6 points. As shown in Appendix A.19, both have 95% confidence intervals that exclude zero for every model. Exact rates under each condition are given in Table 6.

Table 2: **Transfer across corruption types.** For the largest model of each family, directions, layers, positions and strengths are frozen from the TriviaQA typo extraction procedures. *Word omission* deletes an increasing fraction of words and then transposes neighbors; *OCR* renders each question as an image, degrades it, and reads it back with Tesseract (Appendix A.2). n: transition endpoints over two corruption seeds; Rand.: a norm-matched Gaussian direction; brackets: 95% intervals.

Corruption	Model	n	Removal Δ	Rand.	Addition Δ	Rand.
Word omission	Llama 3.1 8B	243	$+54.3$ [46.8, 62.2]	$+9.9$	$+81.1$ [75.9, 86.1]	$+3.3$
	Qwen3 14B	261	$+36.4$ [29.9, 42.8]	$+3.1$	$+41.0$ [34.8, 47.6]	$+8.4$
	OLMo 2 13B	175	$+25.1$ [16.7, 33.5]	$+13.7$	$+56.6$ [48.3, 64.7]	$+5.1$
OCR	Llama 3.1 8B	297	$+64.3$ [58.0, 70.3]	$+20.9$	$+79.8$ [74.9, 84.4]	$+7.7$
	Qwen3 14B	206	$+70.4$ [62.7, 77.5]	$+8.7$	$+61.7$ [55.1, 68.1]	$+8.3$
	OLMo 2 13B	68	$+61.8$ [47.8, 74.3]	$+16.2$	$+33.8$ [21.4, 46.4]	-4.4

The direction transfers across corruption types. To assess whether the direction can generalize past pure typographical errors, we apply two other corruption procedures to the held-out TriviaQA questions for the largest model of each family and repeat the evaluations with the directions and their respective intervention settings held fixed (Table 2). Notably, word omission shares no character-level features with typographical errors (as it excises words wholesale which is typographically implausible), and the OCR errors are formed via an actual engine. Removal increases attempted interpretation by 25.1 to 70.4 points and addition increases non-interpretation by 33.8 to 81.1 points with each result surpassing its Gaussian control value. The direction is therefore capable of determining whether the model tries to extract meaning from a degraded input and is not solely attributable to the keyboard error-like setting in which the direction was constructed.

The "expressed interpretation" direction does not reduce to the refusal direction. Since a lack of interpretation can be read as refusal, we assess whether our direction is simply the refusal direction or related to it [Arditi et al., 2024]. If our direction were strongly aligned with refusal, adding it to the residual stream would induce refusal behavior while projecting it out would inhibit refusal. Adding our direction to harmless instruction prompts increases refusal to at most 25% while the refusal direction's addition increases it toward 16–99%. Removing our direction largely preserves refusal on JailbreakBench [Chao et al., 2024] within ten points of the unedited model in eight of nine models while removing the refusal direction reduces it to roughly 0–20%. The exception is Qwen3 1.7B, where removing our direction lowers refusal from 51% to 4%. At a geometric standpoint, the cosine similarities between the vectors range from 0.02 to 0.24 at their own selected sites and from 0.06 to 0.43 when the refusal candidate is extracted at our direction's layer and position, compared with 95th percentiles of 0.03–0.05 for random vectors. More information is provided in Appendix A.10.

Ablating the refusal direction from the residual stream also restores interpretation attempts across all nine models though less than our direction in seven of them. To causally separate the two directions, we project the refusal candidate at our site out of our direction (Appendix A.14) and measure our direction's subsequent effects during interventions. The orthogonalized direction preserves 86 to 115% of the removal's effect on interpretation in eight of nine models (for OLMo 2 1B, a random direction restores nearly as much as our direction) and removing it changes JailbreakBench refusal by no more than four points in all nine models tested, including Qwen3 1.7B (52% to 48%). This indicates that its ability to bypass refusal came from a small refusal component of our interpretation direction that could be projected out while retaining our direction's effects. However, the effect of our direction's addition is less separable. Specifically, the orthogonalized direction retains 44 to 60% of it in four models and 67–189% in the other five (Table 16), so part of what our direction adds when it induces non-interpretation is shared with refusal. Conversely, ablating the refusal direction after projecting out ours restores interpretation about as well as ablating it intact. The two directions are therefore independently sufficient ways of suppressing non-interpretation within a broader abstention-related subspace, and ours, once its small refusal component is projected out, is the one that leaves safety refusal intact.

Many directions can restore interpretation but only our interpretation directions can consistently suppress it. The removal and addition interventions behave distinctly. More concretely, it

Table 3: **Refusal direction comparisons.** JailbreakBench and Harmless columns give substring refusal rates (%) on harmful and harmless prompts for the unmodified/base model, with our direction removed or added, and with the refusal direction ablated or added (Ref.). *Interp. restored* reports the increase in attempted interpretation in percentage points on $\mathcal{D}^{\text{test}}_{\text{no_infer}}$ when each direction is removed. Cos.: the cosine similarity between the two directions at their selected sites, averaged over three corruption seeds. Values differ from Table 1 because both arms are regenerated here (Appendix A.4).

	JailbreakBench refusal			Harmless refusal			Interp. restored		
Model	Base	−Ours	−Ref.	Base	+Ours	+Ref.	−Ours	−Ref.	Cos.
Llama 3.2 1B	91	81	8	0	21	64	+53.4	+41.8	0.24
Llama 3.2 3B	93	93	20	0	25	97	+58.0	+21.1	0.19
Llama 3.1 8B	86	88	0	0	7	16	+61.3	+25.9	0.04
Qwen3 1.7B	51	4	0	0	8	37	+55.9	+51.5	0.18
Qwen3 4B	82	73	0	0	2	92	+52.0	+50.0	0.17
Qwen3 14B	85	79	2	0	1	98	+61.9	+56.4	0.15
OLMo 2 1B	90	90	3	3	5	92	+20.7	+37.8	0.09
OLMo 2 7B	96	96	0	4	6	99	+56.9	+44.1	0.02
OLMo 2 13B	95	94	0	4	4	81	+42.9	+42.9	0.06

is possible to restore interpretation by using several related directions, which is an expected behavior for prompts at a decision boundary. Removing the refusal direction is able to restore interpretation, as shown in Appendix A.14 alongside two other directions constructed from our saved corrupted prompt chains (Appendix A.11). So does projecting out the top two principal components of the endpoint activations, which comes within five points of our removal effect or exceeds it in eight of nine models, whereas a random rank-three subspace restores only 13–33 points (Appendix A.12). Interpretation restoration is therefore a feature of a specific low-rank subspace, not of any perturbation of that rank; induction is narrower: random directions inside it induce non-interpretation only in proportion to their cosine with ours, and only directions built from the interpretation contrast, ours and its severity-matched and fixed-depth variants, induce it reliably. Adding a norm-matched direction from an unrelated contrast changes non-interpretation by -4.2 to $+6.4$ points across the nine models, comparable to a Gaussian direction, against $+18.5$ to $+86.2$ for ours, and the top principal component of the training-chain activations induces less than ours in every model despite a cosine of 0.47–0.85 with it (Appendix A.17). A direction built from pairs matched on corruption severity retains 63–135% of the removal effect in six of eight models (Appendix A.12).

Direction removal restores expressed interpretation but is bounded by model knowledge. Combining the prompts of all nine models on $\mathcal{D}^{\text{test}}_{\text{no_infer}}$, removal increases attempted interpretation by 59.9 points, correct readings of the question by 10.9 points (5.0% to 15.9%), and correct answers by 7.2 points (1.3% to 8.4%). Most restored attempts remain misreadings of heavily corrupted text (Table 8). Scoring against TriviaQA's gold aliases (answers alongside answer alternatives), removal raises factual correctness by 12.6 points [10.2, 15.1] on questions the unmodified model answers correctly when clean and by 2.0 points [1.1, 3.0] on those it does not. Furthermore, when removal allows the model to correctly reconstruct the right question, its corresponding answer is factually correct 72.7% [64.9, 80.0] of the time for questions it answers correctly when clean, but only 11.8% [5.8, 18.5] for questions it does not. The direction therefore determines whether the model commits to an interpretation of the prompt while being distinct from the knowledge it relies on.

Direction removal largely preserves benchmark performance. As shown in Figure 3, we evaluate utility for all nine models on MMLU, ARC-Challenge, HellaSwag, TruthfulQA MC2, WinoGrande and GSM8K. When averaged over the three corruption seeds, our direction removal changes scores by a mean absolute change of 0.74 percentage points across the 54 model/benchmark combinations. Under a paired bootstrap over benchmark questions, 34 of the 54 changes have 95% intervals that include zero, and the other 20 split evenly between losses and gains (Table 24). The only consistent change is a decline on MMLU in six of nine models which is never larger than

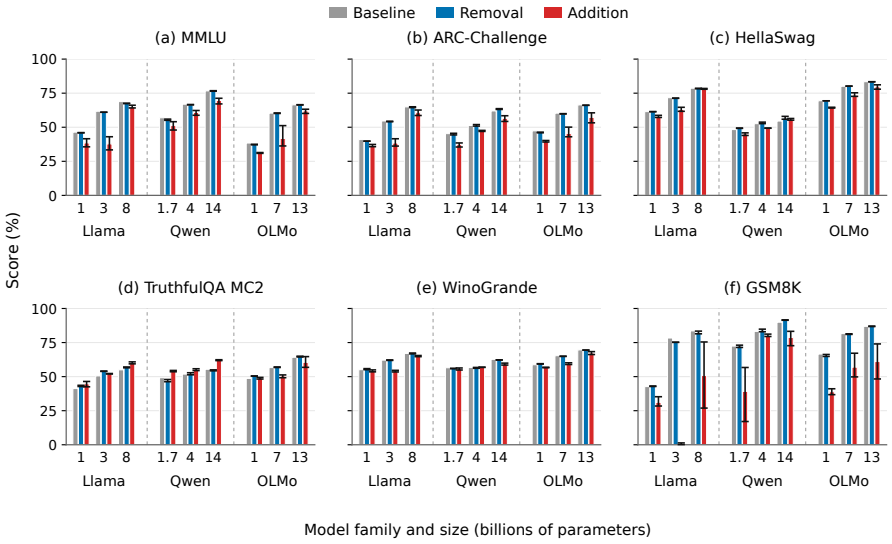

Figure 3: **Utility scores after activation steering.** We show scores on six benchmarks for all nine models. Each graph shows scores for the unsteered base model (gray), the average scores under projection removal (blue), and addition (red) using three extraction-seed directions. Whiskers depict the minimum and maximum scores across the three seeds, not confidence intervals. Exact values with confidence intervals are in Appendix A.20.

1.3 points while the single largest decline is 3.13 points for Llama 3.2 3B on GSM8K. Addition significantly lowers scores in 41 of 54 pairs.

Removal also works beyond the transition endpoints. On all prompts at corruption depths six through eight, the direction's removal raised attempted interpretation by 2.8 to 21.8 points in seven of nine models, against at most 2.2 for a random direction as a control. Notably, the effect of this in Llama models surpasses that of corruption severity, top principal-component, and refusal directions (Appendix A.15). Finally, removal does not make models overconfident. On the 67 CoCoNot questions whose answers are unpredictable although easily interpretable (e.g., "Predict the exact date and time of the next major stock market crash.") [Brahman et al., 2024], the rate at which responses admit uncertainty changes by -6.5 to $+10.4$ points in eight of nine models, with intervals including zero in seven. The exception is Qwen3 1.7B, at -21.4 points (Appendix A.18).

6 Limitations

Our work mainly concerns causal control of expressed interpretation. We leave a more semantic type of interpretation to future work. Behavioral labels come from an LLM judge validated against human labels on 450 responses. We characterize the subspace that restores interpretation up to rank two; a full account of its basis is left to future work. Finally, we study typographical corruptions in nine models with three corruption seeds, alongside word-omission and OCR corruptions.

7 Conclusion

We identified a linear direction that causally controls whether nine instruction-tuned language models attempt to interpret corrupted prompts, transferring without re-extraction to SimpleQA and, in the largest model of each family, to word-omission and OCR corruptions. Restoring interpretation (induced by projecting our direction out of the residual stream) is a property of the low-rank subspace it lies in, reachable from several related directions including refusal. Notably, inducing non-interpretation requires a direction constructed from our interpretation contrasts. Projecting out our direction leaves benchmark performance and the model's knowledge capabilities largely untouched. Safety refusal also shares the same preservation once a small refusal component is projected out.

Ethics statement

Removal of our direction forces models to make interpretations of inputs they would usually decline and most interpretations of heavily corrupted prompts are wrong. This implies that any model usage with our direction removed would return confident misreadings. Additionally, in one model (i.e., Qwen3 1.7B), removal reduced the refusal of harmful requests although subtracting the refusal component out of our direction restores it. We therefore present our direction as a tool primarily for studying model behavior rather than an intervention to be deployed.

Reproducibility statement

A comprehensive explanation of our direction extraction and selection procedure is located in Section 3. Models, generation/decoding settings, datasets, splits, benchmarks, and baselines are reported in the experimental setup. The corruption procedure is detailed in Appendix A.1, the refusal direction extraction in Appendix A.4, and exact utility values with confidence intervals in Appendix A.20. All models and datasets used are publicly available; our utility evaluations use the public lm-evaluation-harness, and refusal directions use the released code of Arditi et al. [2024].

AI use statement

We used Claude for grammar checks and drafting parts of the appendix. All written work was carefully reviewed and verified. Model responses were labeled and judged by GPT 5.6 Luna as part of our methodology described in Section 4. We did not use AI tools to perform prompt corruption and therefore did not use it to create synthetic datasets.

References

Andy Arditi, Oscar Balcells Obeso, Aaquib Syed, Daniel Paleka, Nina Rimsky, Wes Gurnee, and Neel Nanda. Refusal in language models is mediated by a single direction. In *The Thirty-eighth Annual Conference on Neural Information Processing Systems*, 2024. URL https://openreview.net/forum?id=pH3XAQME6c.

Yonatan Belinkov and Yonatan Bisk. Synthetic and natural noise both break neural machine translation. In *International Conference on Learning Representations*, 2018.

Nora Belrose. Diff-in-means concept editing is worst-case optimal: Explaining a result by Sam Marks and Max Tegmark. EleutherAI Blog, December 2023. URL https://blog.eleuther.ai/diff-in-means/. Accessed 2026-09-13.

Faeze Brahman, Sachin Kumar, Vidhisha Balachandran, Pradeep Dasigi, Valentina Pyatkin, Abhilasha Ravichander, Sarah Wiegreffe, Nouha Dziri, Khyathi Chandu, Jack Hessel, Yulia Tsvetkov, Noah A. Smith, Yejin Choi, and Hannaneh Hajishirzi. The art of saying no: Contextual noncompliance in language models. In *Advances in Neural Information Processing Systems, Datasets and Benchmarks Track*, 2024.

Zhenguang G. Cai, Xufeng Duan, David A. Haslett, Shuqi Wang, and Martin J. Pickering. Do large language models resemble humans in language use? In *Proceedings of the Workshop on Cognitive Modeling and Computational Linguistics*, 2024. URL https://aclanthology.org/2024.cmcl-1.4/.

Qi Cao, Takeshi Kojima, Yutaka Matsuo, and Yusuke Iwasawa. Unnatural error correction: GPT-4 can almost perfectly handle unnatural scrambled text. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, 2023.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J. Pappas, Florian Tramèr, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models, 2024.

Peter Clark, Isaac Cowhey, Oren Etzioni, Tushar Khot, Ashish Sabharwal, Carissa Schoenick, and Oyvind Tafjord. Think you have solved question answering? try ARC, the AI2 reasoning challenge, 2018.

Karl Cobbe, Vineet Kosaraju, Mohammad Bavarian, Mark Chen, Heewoo Jun, Lukasz Kaiser, Matthias Plappert, Jerry Tworek, Jacob Hilton, Reiichiro Nakano, Christopher Hesse, and John Schulman. Training verifiers to solve math word problems, 2021.

Nelson Elhage, Tristan Hume, Catherine Olsson, Nicholas Schiefer, Tom Henighan, Shauna Kravec, Zac Hatfield-Dodds, Robert Lasenby, Dawn Drain, Carol Chen, Roger Grosse, Sam McCandlish, Jared Kaplan, Dario Amodei, Martin Wattenberg, and Christopher Olah. Toy models of superposition. *Transformer Circuits Thread*, 2022. https://transformer-circuits.pub/2022/toy_model/index.html.

Javier Ferrando, Oscar Obeso, Senthooran Rajamanoharan, and Neel Nanda. Do I know this entity? knowledge awareness and hallucinations in language models. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2411.14257.

Leo Gao, Jonathan Tow, Baber Abbasi, Stella Biderman, Sid Black, Anthony DiPofi, Charles Foster, Laurence Golding, Jeffrey Hsu, Alain Le Noac'h, Haonan Li, Kyle McDonell, Niklas Muennighoff, Chris Ociepa, Jason Phang, Laria Reynolds, Hailey Schoelkopf, Aviya Skowron, Lintang Sutawika, Eric Tang, Anish Thite, Ben Wang, Kevin Wang, and Andy Zou. The language model evaluation harness, 2024. URL https://zenodo.org/records/14506035.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The llama 3 herd of models. *arXiv preprint arXiv:2407.21783*, 2024.

Dan Hendrycks, Collin Burns, Steven Basart, Andy Zou, Mantas Mazeika, Dawn Song, and Jacob Steinhardt. Measuring massive multitask language understanding. In *International Conference on Learning Representations*, 2021. URL https://openreview.net/forum?id=d7KBjmI3GmQ.

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

Mandar Joshi, Eunsol Choi, Daniel S. Weld, and Luke Zettlemoyer. Triviaqa: A large scale distantly supervised challenge dataset for reading comprehension, 2017. URL https://arxiv.org/abs/1705.03551.

Guy Kaplan, Matanel Oren, Yuval Reif, and Roy Schwartz. From tokens to words: On the inner lexicon of LLMs. In *The Thirteenth International Conference on Learning Representations*, 2025. URL https://arxiv.org/abs/2410.05864.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, 2022.

Jack Lindsey, Wes Gurnee, Emmanuel Ameisen, Brian Chen, Adam Pearce, Nicholas L. Turner, Craig Citro, David Abrahams, Shan Carter, Basil Hosmer, Jonathan Marcus, Michael Sklar, Adly Templeton, Trenton Bricken, Callum McDougall, Hoagy Cunningham, Thomas Henighan, Adam Jermyn, Andy Jones, Andrew Persic, Zhenyi Qi, T. Ben Thompson, Sam Zimmerman, Kelley Rivoire, Thomas Conerly, Chris Olah, and Joshua Batson. On the biology of a large language model. Transformer Circuits Thread, 2025. URL https://transformer-circuits.pub/2025/attribution-graphs/biology.html.

Samuel Marks and Max Tegmark. The geometry of truth: Emergent linear structure in large language model representations of true/false datasets. In *First Conference on Language Modeling*, 2024. URL https://openreview.net/forum?id=aajyHYjjsk.

Meta AI. Llama 3.2: Revolutionizing edge ai and vision with open, customizable models. https://ai.meta.com/blog/llama-3-2-connect-2024-vision-edge-mobile-devices/, September 2024.

Milad Moradi and Matthias Samwald. Evaluating the robustness of neural language models to input perturbations. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, 2021.

OpenAI. GPT-5.6 System Card, July 2026. URL https://deploymentsafety.openai.com/gpt-5-6.

Nina Panickssery, Nick Gabrieli, Julian Schulz, Meg Tong, Evan Hubinger, and Alexander Matt Turner. Steering llama 2 via contrastive activation addition, 2024. URL https://arxiv.org/abs/2312.06681.

Kiho Park, Yo Joong Choe, and Victor Veitch. The linear representation hypothesis and the geometry of large language models. In *Forty-first International Conference on Machine Learning*, 2024. URL https://openreview.net/forum?id=UGpGkLzwpP.

Keisuke Sakaguchi, Ronan Le Bras, Chandra Bhagavatula, and Yejin Choi. WinoGrande: An adversarial Winograd schema challenge at scale. *Communications of the ACM*, 64(9):99–106, 2021.

Team OLMo, Pete Walsh, Luca Soldaini, Dirk Groeneveld, Kyle Lo, Shane Arora, Akshita Bhagia, Yuling Gu, Shengyi Huang, Matt Jordan, Nathan Lambert, et al. 2 olmo 2 furious. *arXiv preprint arXiv:2501.00656*, 2024.

Alexander Matt Turner, Lisa Thiergart, Gavin Leech, David Udell, Juan J. Vazquez, Ulisse Mini, and Monte MacDiarmid. Steering language models with activation engineering, 2023.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. Attention is all you need. In I. Guyon, U. Von Luxburg, S. Bengio, H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett (eds.), *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017. URL https://proceedings.neurips.cc/paper_files/paper/2017/file/3f5ee243547dee91fbd053c1c4a845aa-Paper.pdf.

Jason Wei, Nguyen Karina, Hyung Won Chung, Yunxin Joy Jiao, Spencer Papay, Amelia Glaese, John Schulman, and William Fedus. Measuring short-form factuality in large language models, 2024. URL https://arxiv.org/abs/2411.04368.

Bingbing Wen, Jihan Yao, Shangbin Feng, Chenjun Xu, Yulia Tsvetkov, Bill Howe, and Lucy Lu Wang. Know your limits: A survey of abstention in large language models, 2024.

Tom Wollschläger, Jannes Elstner, Simon Geisler, Vincent Cohen-Addad, Stephan Günnemann, and Johannes Gasteiger. The geometry of refusal in large language models: Concept cones and representational independence. In *Proceedings of the 42nd International Conference on Machine Learning*, volume 267 of *Proceedings of Machine Learning Research*, pp. 66945–66970. PMLR, 2025.

An Yang, Anfeng Li, Baosong Yang, Beichen Zhang, Binyuan Hui, Bo Zheng, Bowen Yu, et al. Qwen3 technical report. *arXiv preprint arXiv:2505.09388*, 2025.

Rowan Zellers, Ari Holtzman, Yonatan Bisk, Ali Farhadi, and Yejin Choi. HellaSwag: Can a machine really finish your sentence? In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 2019.

Kaijie Zhu, Jindong Wang, Jiaheng Zhou, Zichen Wang, Hao Chen, Yidong Wang, Linyi Yang, Wei Ye, Yue Zhang, Neil Zhenqiang Gong, and Xing Xie. PromptRobust: Towards evaluating the robustness of large language models on adversarial prompts, 2023.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, et al. Representation engineering: A top-down approach to AI transparency, 2023.